Analyse de textes spécialisés pour le recueil de données terminologiques

Les nouveaux besoins en terminologie demandent que soient définis de nouveaux modèles et de nouvelles méthodologies. Trois caractéristiques principales que devraient respecter ces nouveaux modèles de données sont présentées : distinction des données linguistiques et conceptuelles, mise en réseau des relations conceptuelles, renseignements précis sur l'usage. La meilleure façon d'obtenir ces informations consiste à travailler sur des corpus. Il faut définir des méthodes assisées par des outils informatiques pour repérer ces données dans les textes. L'article propose une première description des tâches à effectuer et des outils permettant de les mener à bien.

Termes-clés : analyse de corpus, bases de connaissances terminologiques, outils informatiques, sémantique textuelle.

1 Introduction

De nombreux projets associés à l'amélioration de la communication en entreprise (cf. *Terminologies nouvelles,* n°13) ou à la veille technologique font apparaître de nouveaux besoins concernant la terminologie. Dans un article à paraître dans *Terminogramme,* nous avons défini trois classes d'éléments qui devraient caractériser les besoins pour les futures bases de données terminologiques. La prise en compte de ces éléments permet de constituer ce qu'on appelle désormais les *Bases de connaissances terminologiques (BCT)* (cette appellation venant se substituer à celle de *Base de données terminologiques (BDT)*). Le recueil de ces nouvelles données terminologiques nécessite la mise en œuvre de nouvelles méthodes de travail et même d'analyse des données. En particulier, afin de rendre compte au plus juste du fonctionnement de la terminologie, il est indispensable de travailler sur des textes réellement produits dans les entreprises, condidérés comme représentatifs de la langue de l'entreprise. Le travail à effectuer se rapproche alors de celui de l'analyse de corpus, mais les tâches et les moyens de les effectuer, en particulier en s'aidant d'outils informatiques, restent encore à définir.

L'article présente le modèle de Base de connaissances terminologiques (BCT) que nous avons défini. Puis, il propose le détail des tâches à mener à bien pour recueillir les données terminologiques dans les textes. Enfin il s'essaie à une première description de la mise en œuvre d'outils pour assister ces tâches.

2 Modèle de BCT

D'abord dans le cadre d'Aramiihs (Action recherche et application Matra-Irit en interface homme-système) puis dans celui de l'ERSS (Équipe de recherche en syntaxe et sémantique), nous avons mis au point un modèle simple de BCT qui a été mis en œuvre pour différents types d'application, d'abord à Matra Marconi Space (MMS) puis à Aérospatiale et au Centre national d'études spatiales (CNES). Nous reprenons les grandes lignes de ce modèle ici. Pour plus de précisions, voir Condamines & Amsili (1993) et Condamines (1994).

2.1 Présentation du modèle

Le modèle est organisé en trois champs:
– Un champ «terme» (T) qui contient les informations linguistiques (formes du signifiant, nature, genre, type de complémentation...);
– Un champ «lien terme/concept» (L T/C) qui concerne les conditions d'usage, c'est-à-dire les éléments qui, dans la situation de communication, influent sur l'établissement de(s) lien(s) entre un (des) terme(s) et un (des) concept(s);

– Un champ «concept» qui permet d'enregistrer les données concernant le concept dénommé par le terme; ces concepts sont mis en réseau par l'intermédiaire de relations qui ont comme particularité d'être consensuelles (partagées par l'ensemble des locuteurs) et réutilisables quelles que soient les applications ultérieures.

Les relations entre termes sont calculées. Ainsi, deux termes seront totalement équivalents si :
– Ils sont reliés au même concept;
– Ils ont la même nature grammaticale; en effet, nous considérons par exemple que, du point de vue conceptuel, un verbe et un nom, ou un adjectif et un nom... peuvent désigner le même concept, par exemple, *gel d'un test* et *geler un test* renvoient au même concept; mais, si on parle d'équivalence linguistique, alors, l'appartenance à la même catégorie grammaticale est requise;
– Les valeurs respectives des attributs sont identiques; si ces valeurs ne sont pas identiques, l'équivalence n'est pas totale.

Dans le schéma ci-dessous, T1 et T2 (qui renvoient au même concept) sont identiques si la catégorie grammaticale de T1 est la même que la catégorie grammaticale de T2 et que LT1/T2 et LT2/C1 sont identiques.

On peut, de la même façon, calculer l'équivalence interlinguistique en ajoutant une restriction sur la langue.

Les concepts C1, C2, C3, C4 et C5 sont reliés par des relations conceptuelles «est-un», «partie-de», «éventuellement cause», «conséquence»... Chaque concept se manifeste par une réalisation linguistique (T) ou par plusieurs comme le concept C1. Les liens LT/C rendent compte des conditions d'usage.

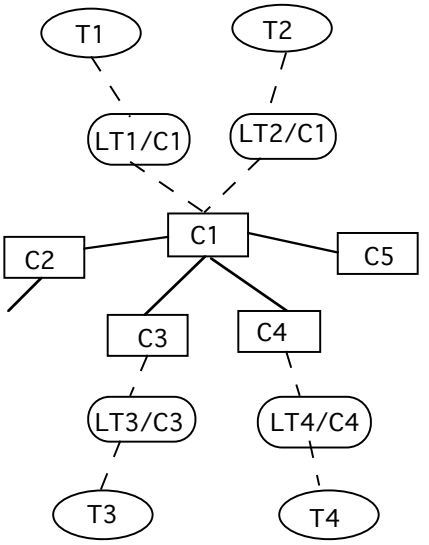


figure 1 : exemple de BCT

2.2 Les trois éléments principaux

Trois caractéristiques de ce modèle nous semblent indispensables pour répondre aux nouveaux besoins en matière de terminologie :

2.2.1 Distinction du champ concernant les données linguistiques et du champ concernant les données conceptuelles

On sait que le terme fonctionne à la fois comme signe linguistique et comme dénomination de concept. Les deux types d'informations, linguistiques et conceptuelles, sont généralement enregistrées dans les bases de données terminologiques (BDT) mais elles ne sont pas forcément bien identifiées et distinguées. Or, il convient de ne pas confondre ces données, d'une part parce que, théoriquement, la confusion des deux niveaux conduit à une identification du sémantique avec le conceptuel, ce qui est erroné; d'autre part, en fonction des utilisations, il se peut que les besoins d'enrichissement concernent soit les données linguistiques, soit les données conceptuelles. En effet, si la BCT est conçue comme une ressource de base, consensuelle et réutilisable, sa mise en œuvre va nécessiter, dans la plupart des cas, que les données soient complétées. Une utilisation de type Traitement automatique de la langue entraînera un enrichissement des données linguistiques et une utilisation de type Représentation des connaissances, un enrichissement des données conceptuelles.

2.2.2 Mise en réseau des concepts

Les données conceptuelles apparaissent généralement dans les BDT sous une forme rédigée, dans la définition. Il est clair que parce qu'il s'agit de termes, la définition sert ici à situer l'élément étudié dans le système conceptuel (contrairement à un mot que la définition devrait servir à situer dans le système linguistique). Or, s'il y a système conceptuel, il doit être possible de définir, à partir des formes linguistiques présentes dans les définitions, quelles relations conceptuelles se manifestent. Les recherches n'étant pas encore suffisamment avancées, il semble prématuré, à l'heure actuelle, de représenter uniquement sous forme de réseau toutes les définitions. Pour les relations les plus fréquentes, cela est toutefois possible tout en conservant, pour le moment, la définition rédigée. Dans le cas où il n'y a pas de définition déjà faite, il est préférable de chercher directement dans les corpus ces relations conceptuelles. (cf. 3.3).

2.2.3 Informations précises sur l'usage

Pour les besoins qui commencent à se faire sentir en entreprise, ce sont là des informations qui jouent un rôle très important. En effet, les entreprises qui se lancent dans la maîtrise de leur

communication en interne ont besoin non seulement d'informations sur la terminologie du domaine dans lequel elles évoluent mais aussi d'informations sur l'usage réel dans leur entreprise. Pour reprendre la comparaison avec la langue générale, les entreprises ont besoin d'informations non seulement sur la compétence mais aussi sur la performance. Qui utilise quel terme pour parler de quel élément ? Telles sont les questions qui doivent recevoir une réponse grâce au travail du terminologue-linguiste. Jusqu'à présent, pour des raisons évidentes de coût, ces informations étaient peu recherchées. Or, à la fois parce que les besoins évoluent et parce que le travail est de plus en plus assisté par des machines, on peut envisager de donner des réponses à ces questions et les intégrer dans la BCT.

Comment ces différents types d'informations peuvent être retrouvés dans les corpus et comment ce travail peut être assisté par des outils ? Nous essayons de répondre à ces questions dans les prochains paragraphes.

3 Les tâches à effectuer pour retrouver les données terminologiques dans les corpus

Différentes tâches, toutes basées sur l'analyse de corpus, doivent être effectuées pour déterminer les données terminologiques correspondant au modèle défini en 2.

3.1 Analyse de corpus

L'analyse de corpus représentatifs du domaine (textes écrits ou retranscriptions d'entretiens, d'exposés...) permet de rendre compte de données conformes à l'usage réel qui est fait par les locuteurs du domaine voire d'une entreprise du domaine. En réalité, il s'agit de mettre en évidence les caractéristiques non seulement de la langue spécialisée utilisée mais aussi des discours spécialisés, c'est-à-dire, de rendre compte des compétences du locuteur manifestées à travers l'actualisation de la langue spécialisée dans des discours, «car c'est bien dans le texte que la signification s'inscrit dans la réalité de la communication» (Descamps 1992 : 334).

Il s'agit de caractériser le comportement linguistique des formes à l'étude, à partir de l'examen de leurs occurrences en corpus. L'analyse s'appuie sur les contextes d'apparition de chaque forme afin de déterminer les parentés, le plus souvent sémantiques, qui peuvent être établies par généralisation à partir de manifestations lexico-sémantiques différentes. L'objectif est d'utiliser des connaissances que l'on a sur le fonctionnement de la langue générale afin d'opérer des classifications de contextes en langue spécialisée. Si l'analyse est suffisamment poussée, on peut définir des patrons lexico-syntaxiques qui définissent ces classes de fonctionnement des contextes (cf. exemples en 3.3).

Notons que nous faisons l'hypothèse, que nous discutons ailleurs (à paraître dans Condamines 1995), qu'une continuité s'établit entre signifié (auquel nous accédons par l'analyse linguistique) et concept. Nous parlerons désormais de concept en lieu et place de signifié et de relations conceptuelles en lieu et place de relations sémantiques.

Prenons le cas où deux termes X et Y ont été repérés et pour lesquels on souhaite définir les données terminologiques conformément au modèle décrit en 2.

Trois séries de tâches sont à effectuer : les premières sont centrées sur les variantes de formes linguistiques (terme (T)), les deuxièmes sur les relations terme/concept (lien terme/concept (L T/C)); les troisièmes sur les relations entre concepts (C).

3.2 Tâches centrées sur le repérage des variantes de signifiant, analyse des variantes d'une même forme linguistique

Des formes morpho-lexicalement apparentées à X ou Y apparaissent dans des contextes sémantiquement identiques à ceux dans lesquels apparaissent X ou Y. L'analyse des contextes ne faisant apparaître qu'une seule classe sémantique, on peut en déduire qu'il n'y a qu'un concept identifiable sous des formes différentes. Ces formes sont soit des sigles – *Scao* pour *Système de contrôle d'attitude et d'orbite* – soit des variantes linguistiques (ellipses ou autres) – par exemple *segment moyen de l'IVA* peut apparaître sous la forme *IVA moyenne*.

3.3 Tâches centrées sur la relation terme/concept

3.3.1 Analyse des dénominations alternatives : 2 termes (non apparentés morphologiquement)/un seul concept

On a affaire à une seule classe de contextes, c'est-à-dire un seul type de fonctionnement sémantique, entourant des formes n'ayant pas de parentés morphologiques. Comme dans le cas précédent, il n'y a qu'un seul concept mais, cette fois, deux signifiants non apparentés morphologiquement. Par exemple, dans la terminologie de Matra Marconi Space, *senseur* et *capteur* sont équivalents. Dans ce cas, deux termes sont créés, chacun étant en relation avec le même concept.

3.3.2 Analyse des polyacceptions : 1 seul terme/plusieurs classes de contextes

Une même forme linguistique apparaît dans des contextes qui se généralisent en au moins deux classes. Dans ce cas, au moins deux cas de figure peuvent se présenter. Soit la forme est homonymique, soit la forme est polysémique.

Si la forme est homonymique, deux ou plusieurs classes de contextes apparaissent sans parentés sémantiques; deux termes sont créés, chacun étant en relation avec un concept particulier. Si la forme est polysémique, deux ou plusieurs classes de contextes apparaissent qui entretiennent des parentés sémantiques évidentes; un seul terme est créé, qui est en relation avec plusieurs concepts, eux-mêmes entretenant des relations. Ce cas de figure apparaît souvent avec les déverbaux (les noms reliés à un verbe), fréquents dans les langues de spécialité; ces noms dénomment fréquemment soit l'action, soit le résultat de l'action : un achat peut correspondre soit à l'action d'acheter («au moment de l'achat de sa voiture...») soit au résultat de cette action («pose tes achats sur la table»).

Dans certains corpus, la polyacception (un terme/plusieurs concepts), concerne un seul référent (au sens d'élément de la réalité), central pour le domaine. Dans ce cas-là, nous considérons que nous avons affaire à plusieurs «points de vue» sur un objet. On peut souvent relier un point de vue à un métier, une compétence particulière, c'est-à-dire que l'on trouve un point de vue dominant (une classe de contextes dominante) dans les textes issus d'un seul et même métier. Ainsi, une étude réalisée au CNES et rapportée dans Rébeyrolle (1995) a mis en évidence sept acceptions pour le terme *satellite*, chaque acception étant caractéristique d'un métier concernant le produit satellite. Si nous ne considérons que deux de ces acceptions, le satellite vu comme un «corps artificiel» et le satellite vu comme un «véhicule», on peut montrer que le premier point de vue est repérable grâce aux contextes correspondant au patron :

V de type *projeter dans l'espace* + dét + N1 (1)
[1] *lancer des satellites en orbite*

V support de type *faire* + dét + (déverbal) DVB de type *projeter dans l'espace* + *de* + dét + N1
[2] *effectuer le lancement du satellite*
[3] *effectuer le tir de SPOT4*
et la seconde acception grâce aux contextes correspondant au patron :

V de type *placer* + dét + N2 + *sur* + dét + N1
[4] *placer un système sur le satellite*
[5] *placer un senseur stellaire sur le satellite*
[6] *poser des fixations sur le satellite*

V support de type *faire* + dét + DVB *placer* + *de* + N2 + *sur* + dét + N1
[7] *réaliser l'embarquement du prototype sur le satellite SPOT4*
[8] *faire l'intégration des instruments sur les satellites porteurs*
[9] *faire l'intégration des essais électriques, essais mécaniques sur le satellite*
[10] *réaliser l'embarquement du navigateur autonome sur le satellite*

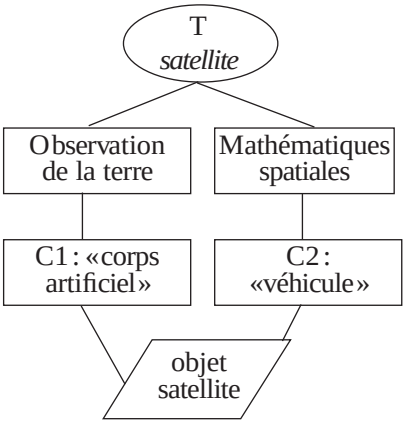
Le premier type d'acception apparaît surtout dans le département Observation de la terre et le second dans le département Mathématiques spatiales. Cette différence d'acception en fonction du métier (qui manifeste des points de vue différents) est à situer dans la question de l'usage dont nous traitons ci-dessous.

(1) Remarque : N1 = *satellite* ou *satellite* + adjectif.

3.3.3 Analyse de l'usage en fonction de la provenance du document : influence sur le choix de l'une ou l'autre forme linguistique ou influence sur le lien terme/concept

Deux cas se présentent. Dans le premier, nous avons opté pour consigner l'information dans la partie Terme; il s'agit des cas où une forme, sigle ou variante, est préférée dans l'un ou l'autre métier, il peut être alors indiqué, dans le champ Terme : «forme préférée dans tel métier».

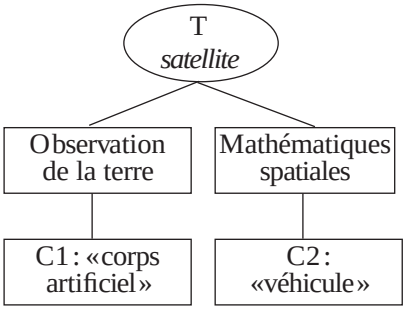
Dans le deuxième cas, l'information est indiquée dans le champ terme/concept. Il s'agit des cas où le métier a une influence sur le choix d'un terme pour dénommer un concept (cas des termes équivalents : plusieurs termes pour un seul concept) ou bien sur le point de vue (cas où un terme dénomme plusieurs concepts apparentés : un terme/plusieurs concepts). Par exemple, les deux points de vue évoqués dans l'étude du CNES pourraient être représentés de l'une des façons suivantes :



Première façon de représenter deux points de vue pour *satellite*

Dans ce cas, il y a création d'un nouveau champ qui permet d'enregistrer l'existence d'un référent unique. Notons que «corps artificiel» et «véhicule» ne sont ici que des

étiquettes de concepts. Le premier concept devrait être en relation avec le terme *corps artificiel* et le second avec le terme *véhicule*. Le champ lien terme/concept contient ici l'information concernant le métier dans lequel est privilégié un usage de satellite.



Deuxième façon de représenter deux points de vue pour *satellite*

Dans ce cas, l'existence d'une parenté entre les concepts, ici étiquetés «corps artificiel» et «véhicule», est indiquée par la mise en relation de ces deux concepts.

3.4 Tâches centrées sur l'analyse des relations entre concepts

3.4.1 Déterminations de paradigmes taxonomiques : plusieurs termes/plusieurs concepts taxonomiquement reliés

Cette étape a pour objet d'aider l'étape suivante d'identification des relations conceptuelles.

En effet, des termes représentant des concepts taxonomiquement apparentés à X ou Y peuvent apparaître en position sémantique identique à celle de X ou Y dans les phrases, c'est-à-dire peuvent permettre de repérer une relation conceptuelle «basique» entre X et Y. En effet, si on a :

X *est-un* Z et Z *R* U, on peut avoir X *R* U.

Or, dans les cas où il y a peu d'occurrences de X, il peut être intéressant de tester la relation sur Z et de faire l'hypothèse qu'elle est valable aussi pour X, puisqu'en principe un élément hérite des traits de fonctionnement des éléments supérieurs dans la taxonomie. Par exemple, si Z *a-pour-symptôme* U et si X *est-un* Z, on peut penser que X *a-pour-symptôme* U; Si une infection *a-pour-symptôme* de la fièvre et si la grippe *est-une* infection alors on peut penser que la grippe *a-pour-symptôme* de la fièvre. Dans les corpus, au moment de travailler sur les relations qu'entretient l'hyperonyme (*infection* dans l'exemple), on pourra aussi travailler à partir des hyponymes (*grippe* dans l'exemple), en particulier dans les cas où la forme étudiée apparaît peu fréquemment.

Cette étape constitue un premier type d'analyse de relations. Il faut chercher les contextes dans lesquels X et Y sont cooccurrents et apparaissent avec des marqueurs de relations taxonomiques, par exemple un contexte comme :

«de tous les X, Y est le seul qui...»

qui permet de repérer que Y *est-un* X..

3.4.2 Analyse de la (des) relation(s) conceptuelle(s) entre X et Y

L'hypothèse de départ la plus simple est qu'il y a une seule relation sémantique entre X et Y qui s'actualise par des manifestations lexico-syntaxiques différentes. S'il y a plusieurs relations, il faut évaluer quels sont les différents points de vue mis en jeu.

Pour effectuer l'analyse en corpus de la façon la plus efficace, on peut mettre en œuvre une méthodologie en plusieurs étapes:

• Étape A

Pour chaque terme (pour nous, X et Y), constitution de classes :

– Classe 1 des variantes, sigles ou variantes de formes;

– Classe 2 des équivalents (dénominations alternatives);

– Classe 3 des hyponymes.

• Étape B

Recherche des cooccurrences de X et Y apparaissant dans le même contexte(2), en faisant intervenir, selon les besoins (cooccurrences insuffisantes ou au contraire trop nombreuses) les classes définies en A.

• Étape C

Interprétation des résultats;

Soit il n'y a aucun contexte d'apparition simultanée de X (ou ses différentes classes) et de Y (ou ses différentes classes). Dans ce cas, ou bien le corpus n'est pas représentatif du domaine, ou bien, il n'existe pas de relation entre les concepts dénommés par X et par Y.

Soit il y a un contexte d'apparition simultanée de X (ou ses classes) et de Y (ou ses classes). Dans ce cas, on peut retrouver plusieurs cas de figure.

– Premier cas de figure.

Le contexte peut être insignifiant, c'est-à-dire qu'on ne peut pas lui attribuer une pertinence par rapport à une relation conceptuelle. Cette insignifiance peut être due au fait que la relation qui existe est trop implicite pour être mise au jour dans le contexte à l'étude; elle peut être due au fait que la relation conceptuelle qui existe n'est pas pertinente dans l'assertion linguistique qui sous-tend le contexte; elle peut être due enfin au fait que l'interprétant (le linguiste-terminologue) manque de compétences linguistiques (il ne connaît pas certains mots ou tournures apparaissant dans le

(2) La notion de contexte mériterait ici un plus long développement. En effet, l'expérience nous a montré que se contenter d'une phrase comme contexte était insuffisant. Nous avons décidé de limiter arbitrairement le contexte de façon numérique et avons opté pour un contexte de 35 chaînes de caractères.

contexte) ou techniques (il ne connaît pas la nature de la relation conceptuelle qui apparaît).
– Deuxième cas de figure.
Le contexte est signifiant d'une seule façon: on peut trouver, à travers les divers contextes une régularité sémantique qui manifeste l'existence d'une relation conceptuelle unique.
– Troisième cas de figure.
Les divers contextes dans lesquels apparaissent X et Y permettent d'identifier plusieurs classes de fonctionnement sémantique, donc plusieurs relations conceptuelles; ce cas semble associé aux cas de polyacceptions en lien avec des points de vue en fonction de types de raisonnement différents (par exemple, les relations cause et conséquence peuvent être inversées).

L'approche que nous préconisons pour repérer les relations conceptuelles constitue une abstraction par rapport à ce qui est réellement fait. En effet, on ne teste pas tous les couples de termes, ce qui serait extrêmement long et fastidieux mais seulement les couples dont nous pensons, pour diverses raisons (compétence linguistique, intervention d'un expert), qu'ils peuvent entretenir des relations. Par ailleurs, au fur et à mesure que le travail progresse, les résultats sur les relations conceptuelles se cumulent et rendent la recherche plus efficace.

Il est clair que les différentes étapes que nous proposons font appel à un examen minutieux des corpus qui demande des connaissances sur le fonctionnement de la langue et une grande capacité à généraliser. Par ailleurs, les résultats que nous obtenons constituent des hypothèses linguistiques qui doivent être soumises à un, voire plusieurs, expert(s). Le fait que chaque résultat puisse être argumenté par des portions de textes rend la discussion à la fois plus distanciée (puisqu'elle se base sur un texte écrit par un tiers) et plus fructueuse.

Ce type d'approche, qu'il paraissait inconcevable de mettre en œuvre il y a quelques années deviendra sans doute incontournable. En effet, d'une part, les besoins en données terminologiques qui rendent compte au plus près du fonctionnement linguistique à l'intérieur des entreprises sont de plus en plus pressants; d'autre part, et heureusement, de plus en plus d'outils viennent faciliter ce travail d'analyse de corpus. Dans le prochain paragraphe, nous essayons de faire un panorama rapide des types d'outils qui pourraient être utilisés pour ces tâches.

4 Quels outils d'aide pour ces tâches

La rapide présentation que nous proposons ici a essentiellement pour objectif de montrer combien les besoins émergents en ce qui concerne l'analyse de corpus, pour en extraire des données terminologiques, vont de pair avec la définition de nouveaux outils d'analyse de textes, pas toujours dédiés à la terminologie mais qui pourraient fort bien être utilisés dans ce cadre. Nous n'avons utilisé que certains des outils que nous évoquons; pour les autres, nous avons assisté à une démonstration ou bien nous n'en avons connaissance que par les articles qui les décrivent. Nous avons voulu nous placer du côté de l'utilisateur potentiel qui, ayant défini ses besoins, souhaiterait voir son travail assisté par des outils.

4.1 Outils de recherche de concordances

Les concordanciers permettent d'accéder à toutes les occurrences de formes et à leur contexte d'apparition. Cet accès peut être affiné si, comme dans *Sato* (Daoust 1990), une base de données lexicale permet d'attribuer à chaque forme sa ou ses catégories grammaticales. Il reste que s'il n'y a pas d'analyse syntaxique, une forme est traitée comme ayant toutes les valeurs qui lui sont attribuées dans la base lexicale. Ainsi, *porte* est à la fois verbe et nom et conserve cette étiquette quel que soit son contexte d'apparition. Il est évident que le couplage avec un analyseur syntaxique rendrait ces outils beaucoup plus efficaces. *Sato* permet aussi d'attribuer des propriétés aux chaînes de caractères. Il est possible d'utiliser cette fonctionnalité pour constituer les classes de variantes ou d'hyponymes telles que nous les avons présentées en 3.3. Ce type d'outils est l'outil de base pour l'exploration des corpus.

4.2 Outils d'extraction de termes et de relations candidats

De nombreux laboratoires de recherche français (nous connaissons moins bien la situation ailleurs) travaillent à la définition d'outils d'extraction de termes candidats ou de relations candidates.

Les outils d'extraction de termes fonctionnent sur deux principes. Soit ils fonctionnent sur un principe statistique (Enguehard et Pantéra 1995), soit ils fonctionnent à partir d'une analyse linguistique poussée (*cf.* *Lexter* : Bourigault 1994 et *Termino*: David & Plante 1990), soit ils fonctionnent sur un principe mixte couplant une approche statistique et linguistique (Daille 1995).

Dans tous les cas, ces outils fournissent une liste de termes candidats, à charge pour le linguiste/terminologue d'effectuer un tri.

Les outils d'extraction de relations candidates ne sont pas conçus, dans un premier temps pour assister le travail du terminologue, mais celui du cogniticien, c'est-à-dire de la personne chargée de recueillir les connaissances (en l'occurrence à partir

de corpus) qui vont lui permettre de construire un système à base de connaissances. Or, ainsi que nous l'avons montré aillleurs (Condamines 1995), les relations mises en place dans un système à base de connaissances sont proches des relations conceptuelles qu'on peut trouver dans une terminologie. Il est donc possible d'utiliser ces outils dans le cadre de la terminologie. Deux types d'outils peuvent exister en fonction du principe de base mis en œuvre. Dans le premier, on utilise des connaissances sur la langue afin de repérer des relations conceptuelles dans les langues de spécialité; par exemple pour *Seek* (Jouis 1995) une liste des marqueurs linguistiques de la relation *est-un* et *partie-de* (est constitué de, a pour partie...) a été faite. Lorsque le système rencontre ces formes dans le corpus, il suggère à l'utilisateur l'existence d'une relation conceptuelle. Dans le deuxième type d'outils (Frath *et al.* 1995), le système, après avoir lemmatisé le corpus, repère des formes récurrentes apparaissant entre deux formes dont il présume qu'elles sont des termes. L'idée est ici qu'il peut y avoir des marqueurs de relations voire des relations qui sont dépendants du domaine. Lorsque l'utilisateur est mis en présence de ces contextes récurrents, il peut travailler à leur généralisation pour définir la nature de la relation mise en œuvre.

4.3 Outils d'analyse statistique

Nous ne parlerons ici que de *Alceste* (Reinert 1991), que nous connaissons mais il est vraisemblable que tous les outils à base de calculs statistiques un peu poussés permettraient le même type d'utilisation.

Ces outils visent à donner du texte une classification en termes de contenu sur la base de calculs d'occurrences lexicales. Un texte se trouve ainsi organisé en plusieurs classes auxquelles l'utilisateur donne une interprétation en fonction de sa problématique. Par exemple, on devrait pouvoir identifier des «points de vue» émergents à partir des résultats fournis par un tel outil.

L'expérimentation que nous avons faite de *Alceste* (Rébeyrolle 1995), nous a permis de mettre au point une méthodologie qui rend compte de classes de fonctionnement de contextes autour d'un terme. Or, nous l'avons vu, la mise en évidence de plusieurs classes de contextes nous permet de présumer un phénomène de type homonymique ou polysémique. L'utilisation d'un tel outil a donc présenté un grand avantage pour nous (cf. Condamines & Rébeyrolle, à paraître) puisqu'il nous a permis de travailler à la mise en évidence de points de vue à travers l'étude de contextes à partir des résultats fournis par *Alceste*.

4.4 Quelles perspectives pour l'utilisation de ces outils

Si le travail sur corpus devient un des éléments importants de la terminologie – mais aussi d'autres types de travaux comme l'extraction de connaissances à partir de textes – on devrait assister à un développement important des outils. Mieux encore, ces outils devraient être intégrés pour constituer de véritables stations d'analyse de textes. C'est déjà ce qui se met en place dans certaines grandes entreprises comme à IBM ou à EDF (cf. Herviou *et al.* 1995).

5 Conclusion

La terminologie est actuellement le champ de changements importants tant théoriques qu'appliqués. Les besoins, en particulier dans les entreprises, ont évolué; les modèles de données se sont adaptés; les méthodologies, grâce au développement d'outils informatiques, convergent vers une prise en compte plus grande des corpus qui, seuls, manifestent l'usage réel des termes par une communauté de locuteurs. L'analyse de corpus est actuellement faite de façon souvent intuitive. Il convient de définir des méthodes adaptées aux besoins spécifiques de la terminologie, aussi bien en ce qui concerne la façon d'utiliser les outils et, plus particulièrement d'utiliser les résultats fournis par un outil en entrée d'un autre outil, que sur l'interprétation que l'on peut faire des résultats obtenus. Afin de rendre ces méthodes les plus efficaces et reproductibles possibles, nous avons proposé d'organiser le recueil de données terminologiques en tâches que nous avons définies par rapport à un modèle de BCT. Ce travail devrait se poursuivre pour permettre un affinement de la méthodologie et de la prise en compte d'outils.

Anne Condamines,
Équipe de recherches en syntaxe et sémantique, Ura 1033, CNRS, Université de Toulouse le Mirail, France.

Bibliographie

Auger (Pierre), 1994: «Les outils de la terminotique», dans *Terminologies nouvelles* , n° 11, p. 46-52.

Bourigault (Didier), 1995: «Conception et exploitation d'un logiciel d'extraction de termes en entreprise. Problèmes méthodologiques et théoriques», dans Clas (André) *Actes des 4es Journées scientifiques du réseau Lexicologie, terminologie, traduction*, Lyon 28-30 septembre 1995, à paraître.

Bourigault (Didier) et Condamines (Anne), 1995 : «Réflexions sur le concept de base de connaissances

terminologiques», dans Actes des 5 Journées du PRC IA, Nancy 1-3 février 1995, Toulouse, Teknea, p.425-444.

Condamines (Anne) et Amsili (Pascal), 1993 : «Terminology between Language and Knowledge : An example of Terminological Knowledge Base», dans Schmitz (Klaus-Dirk), TKE93 *(Terminology and Knowledge Engineering)*, Frankfurt, Indeks Verlag, p. 316-323.

Condamines (Anne), 1995: «Terminology: new needs, new perspectives», dans *Terminology*, vol. 2, n°2, p.219-238.

David (Sophie), Plante (Pierre), 1990: *Termino version 1.0*, Rapport du Centre Ato, *Analyse de textes par ordinateur*, Université du Québec à Montréal.

Daille (Béatrice), 1995: «ACTA: Une maquette d'aide à la construction automatique de banques terminologiques monolingues et bilingues», dans Clas (André), *Actes des IVes Journées Scientifiques du réseau LTT : lexicomatique et dictionnairiques*, Lyon, 28-30 septembre 1995, à paraître.

Daoust (François), 1990 : «L'informaticien, le lecteur et le texte. L'approche Sato», dans *ICO Québec* (Intelligence artificielle et sciences cognitives au Québec), vol. 2, n°3, p. 55-60.

Descamps (J.L), Mochet (M.A), Lewin (T), Lamizet (B) et Costes (D.), 1992: *Sémantique et concordances*, Paris, Klincksieck, (Publication de l'Inalf, Collection «St-Cloud»).

Enguehard (Chantal) et Pantéra (Laurent), 1995: «Automatic Natural Acquisition of a Terminology», dans *Journal of quantitative linguistics*, vol. 2, n°1, p. 27-32.

Frath (Pierre), Oueslati (Rochdi) et Rousselot (François), 1995: «Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques». dans *Actes des Journées Acquisition, validation, apprentissage du PRC-GDR-IA du CNRS*, Grenoble: 5-7 avril, Le Chesnay, France, Inria, p.173-186.

Herviou (Marie-Luce), Ogonowski (Antoine) et Dauphin (Eva), 1994: «Tools for extracting and structuring Knowledge from Texts», dans *Proceedings of Computational Linguistics, Coling*, Japan : Kyoto, vol. 2, p. 1049-1053.

Jacquemin (Christian), 1994 : «Quelques mécanismes spécifiques d'une grammaire d'unification adaptée à l'extraction terminologique», dans *Actes du 9e congrès Reconnaissance des formes et intelligence artificielle* (RFIA'94), Paris, AFCET, p. 385-396.

Jouis (Christophe), 1995 : «*Seek*, un logiciel d'acquisition des connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe», dans *Actes des Journées Acquisition, validation, apprentissage du PRC-GDR-IA du CNRS*, Grenoble : 5-7 avril, Le Chesnay, France, Inria, p. 159-172.

Kocourek (Rostislav), 1991 : «Textes et termes», dans *Meta*, vol. 36, n°1, p. 71-76.

Meyer (Ingrid), Bowker (Lyne) et Eck (Karen), 1992 : «Cogniterm: An experiment in Building a Terminological Knowledge Base»; dans Tomola (H.), Varantola (K.) et Salimi-Tolonen (T.), *Euralex'92 Proceedings I-II*, Tampere, Finland : University of Tampere, p. 159-172.

Rastier (François), 1991 : *Sémantique et recherche cognitive*, Paris, Puf.

Rebeyrolle (Josette), 1995 : *Apports de la terminologie à l'étude des points de vue*, Mémoire de fin de stage de DEA au CNES, Université Toulouse le Mirail, Équipe de recherches en syntaxe et sémantique.

Reinert (Max), 1991 : «Les mondes lexicaux et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars», dans *Langage et société*, n°66, p. 5-39.