Structuration de terminologies à l'aide d'outils de TAL avec TERMINAE

Sylvie Szulman* — Brigitte Biébow* — Nathalie Aussenac-Gilles**

** Université de Paris-Nord, Laboratoire d'Informatique de Paris-Nord (LIPN)*
Av. J.B. Clément, F93430 Villetaneuse
{Brigitte.Biebow,Sylvie.Szulman}@lipn.univ-paris13.fr
*** Université Toulouse 3, Institut de Recherche en Informatique de Toulouse (IRIT)*
118 route de Narbonne, F31062 Toulouse Cedex 4

RÉSUMÉ. *Cet article présente les derniers développements de* TERMINAE, *qui en font une plate-forme adaptée à la construction de diverses ressources terminologiques à partir de textes, y compris d'ontologies. Désormais,* TERMINAE *aide à élaborer aussi bien des fiches terminologiques, qu'un réseau conceptuel structurant les concepts associés aux termes ou encore une ontologie formelle sur laquelle des inférences peuvent être effectuées. La démarche méthodologique et l'utilisation de l'outil sont détaillés sur un exemple de construction d'un réseau conceptuel portant sur la fabrication du verre à partir d'un texte décrivant cette fabrication. L'accent est mis sur les modalités d'exploration des textes à l'aide des outils de traitement du langage naturel disponibles sur cette plate-forme.*

ABSTRACT. *This article presents* TERMINAE *last developments which is now a platform suited to build terminological resources from texts, including ontologies.* TERMINAE *may be used to build terminological forms or a conceptual network which structures concepts issued from terms or to build a formal ontology upon which inferences may be made. The methodology and the tool are detailed on an example about glass manufacture, from a text describing the process. We point out the way to explore text with the help of NLP tools available on the platform.*

MOTS-CLÉS : *acquisition de terminologie à partir de textes, ontologie, terminologie, plate-forme, extraction de termes, extraction de relation.*

KEYWORDS: *terminological acquisition from texts, ontology, terminology, platform, term extractor, relation extractor.*

1. Introduction

Afin de répondre à des besoins variés, les terminologies comportent aujourd'hui des données plus riches que de simples listes de termes d'un domaine spécialisé. Elles sont structurées à l'aide de relations : les termes sont regroupés en familles de synonymes, les notions sous-jacentes sont structurées en hiérarchies et reliées par des relations sémantiques. Les terminologues travaillent de plus en plus à partir de textes sur support informatique à l'aide de logiciels de traitement du langage naturel (TAL). Ces outils rendent plus efficace et plus systématique le repérage des termes. Mais il reste encore beaucoup à faire avant que ces nouvelles pratiques ne se généralisent. En particulier, il est indispensable d'avancer dans la mise à disposition d'environnements de représentation, gestion et consultation de terminologies, et dans l'intégration de ces environnements avec les logiciels de TAL utilisables pour construire des terminologies à partir de corpus.

Dans la lignée des travaux du groupe TIA, entre autres [BAC 95], [SLO 95], nous avons déjà développé des environnements de construction de bases de connaissances terminologiques à partir de textes, Géditerm [AUS 99], et de construction d'ontologies à partir de textes, TERMINAE [BIE 00]. Nous avons également expérimenté des logiciels d'extraction de termes (Lexter [BOU 94] et Nomino [DAV 90]) et de recherche de relations, Caméléon [SEG 01].

A partir de ces travaux, nous présentons aujourd'hui une nouvelle version de TERMINAE adaptée à la construction de terminologies et offrant une meilleure intégration de logiciels de TAL. En particulier, TERMINAE permet de dépouiller les résultats de l'extracteur de candidats-termes, Lexter, et d'explorer un corpus à l'aide de motifs de relations dans le module Linguae. Une terminologie dans TERMINAE est composée d'un ensemble de fiches terminologiques et d'un réseau conceptuel, constitué de concepts issus des fiches terminologiques et des relations sémantiques entre eux. Un cadre méthodologique, associé à ce logiciel, définit les étapes qui vont de l'exploration des textes à la structuration d'une terminologie sous la forme d'un réseau conceptuel.

Dans cet article, nous situons d'abord notre travail par rapport aux diverses approches en terminologie. Nous présentons ensuite la méthode de modélisation associée (section 2) à TERMINAE pour la construction d'ontologie. Nous dressons un état de l'art d'outils d'aide à la construction de ressources terminologiques (section 3). Le cœur de l'article est consacré à une description des fonctionnalités d'analyse de texte et des évolutions récentes apportées TERMINAE pour l'adapter à la structuration de terminologie (section 4). Nous illustrons son utilisation et la méthode de construction de terminologie à partir d'un texte sur la fabrication du verre [LAJ 62].

Termes, notions, concepts

Quelques éclaircissements nous semblent nécessaires pour bien situer notre travail dans une approche de la terminologie partant de l'étude des textes. Soulignons

toutefois qu'en tant que cogniticiennes, nous nous intéressons moins à la théorie linguistique qu'à ses résultats pratiques et aux outils d'acquisition de connaissances à partir de textes en découlant.

En terminologie classique, une *notion* est définie comme « unité de pensée constituée d'un ensemble de caractères attribués à un objet ou à une classe d'objets, qui peut s'exprimer par un terme ou par un symbole »(définition de l'Office de la langue française du Québec, conforme à celle de l'ISO). Les notions d'un domaine scientifique ou technique s'organisent en un ensemble structuré par des relations non linguistiques, le système notionnel. Le travail habituel d'un terminologue est alors d'étudier quels termes correspondent à une notion donnée, suivant ainsi une démarche onomasiologique en partant du concept pour rechercher les signes linguistiques qui lui correspondent. Notion et concept sont équivalent, et correspondent selon [RAS 91] au signifié, normé par la discipline scientifique ou technique.

Notre approche, à l'opposé, part des mots retenus dans les textes en tant que candidats-termes et étudie leur comportement linguistique en corpus pour élaborer un *concept*. Ce concept est construit parce qu'il a de la pertinence au sein d'un modèle défini pour une application donnée. Par exemple, il fera partie d'un thésaurus afin de faciliter la recherche documentaire dans un corpus, ou bien il sera structuré au sein d'une ontologie formelle pour améliorer les recherches sur le web. L'ensemble des mots désignant le concept prend le statut de *terme*. C'est donc un point de vue sémasiologique que nous suivons, partant du mot et de son usage dans les textes pour déterminer le concept.

Dans la première version de TERMINAE, nous avions choisi le mot « notion »pour désigner le sens d'un terme car le mot « concept »désignait pour nous le concept formel habituel en intelligence artificielle. Nous passions d'une notion à un concept formel. Mais à la réflexion, ce choix masque l'élaboration d'une définition à partir de l'analyse des occurrences de mots en contexte et des besoins applicatifs. Nous référant à [RAS 95], nous préférons maintenant parler de *concept* (comme dans Géditerm) pour désigner un « sens normé ».

Selon Rastier, le sens d'un mot résulte de son interprétation en contexte. Lorsqu'un mot est choisi pour devenir l'objet d'une fiche terminologique, et donc un terme, nous étudions ses différents sens présents dans le corpus et répartissons les occurrences entre ces sens. Ce sens est construit à partir de l'interprétation des occurrences du mot en contexte, puis normé, c'est-à-dire restreint pour l'application, dans une définition. C'est ce sens normé qui correspond au *concept* associé au terme. Puisqu'il est créé à partir d'une interprétation des mots en contexte, il ne s'agit pas d'un signifié normé.

2. TERMINAE, outil et méthode de construction des terminologies et ontologies

2.1. *Présentation générale*

TERMINAE est à l'origine un outil d'aide à la construction d'ontologie qui répond à des besoins liés à la modélisation à partir de textes. Il repose sur des méthodes linguistiquement fondées et accompagnées d'outils de TAL, comme l'étude terminologique sur un corpus, pour justifier et aider la modélisation. Il offre une traçabilité totale des textes vers l'ontologie et *vice versa*, pour que l'utilisateur puisse vérifier l'adéquation entre la définition des concepts et leur interprétation linguistique ; ceci est nécessaire pour comprendre l'ontologie, l'utiliser et la maintenir. Il permet la mise en évidence des choix de modélisation, par l'inclusion de commentaires et la définition d'une typologie de concepts, toujours pour faciliter la compréhension. Enfin, il inclut un langage de description formel, ce qui garantit la validité logique de l'ontologie et aide le cogniticien en lui signalant incohérences et redondances. TERMINAE repose sur des bases théoriques issues de la linguistique et de la représentation des connaissances, il intègre dans un même environnement des outils d'ingénierie linguistique et d'ingénierie des connaissances.

2.2. *La méthode sous-jacente à* TERMINAE

La méthode de construction d'ontologie à partir de textes est résumée dans la figure 1 et explicitée dans [AUS 00]. Après la constitution du corpus et une étude linguistique, menée à l'aide d'outils de TAL, les connaissances sont normalisées puis formalisées. Pour conduire ces étapes qui sont au cœur du processus de modélisation, TERMINAE fait des propositions précises et originales.

Normalisation. Ce processus particulier de conceptualisation, fondé sur l'analyse de corpus, consiste en deux parties : la première demeure dans le domaine du traitement lexical et exploite les données retenues par l'étude linguistique, elle conduit à une fiche terminologique ; la seconde, se démarquant de toute approche linguistique, porte sur l'interprétation sémantique et la structuration des concepts et des relations sémantiques ; elle mène à une fiche de modélisation dans le cas d'une ontologie, à un réseau conceptuel dans le cas d'une terminologie.

Les candidats-termes et les relations lexicales déterminés par les outils sont associés à leurs occurrences dans le corpus ; un premier travail consiste à distinguer pour chaque candidat-terme et chaque hypothèse de relation lexicale s'ils conduisent à une ou plusieurs interprétations dans le domaine. En cas de polysémie, il faut décider quels sens parmi ceux présents dans le corpus sont à retenir car pertinents pour la modélisation. Parmi l'ensemble des termes et relations lexicales potentiels, le cogniticien choisit ceux dont il va poursuivre l'analyse. Ce sont les termes qui, à la fois, ont du sens en corpus et qui présentent un intérêt par rapport aux objectifs du modèle. Pour chacun d'eux, il étudie ses contextes d'occurrence afin d'en donner une définition en

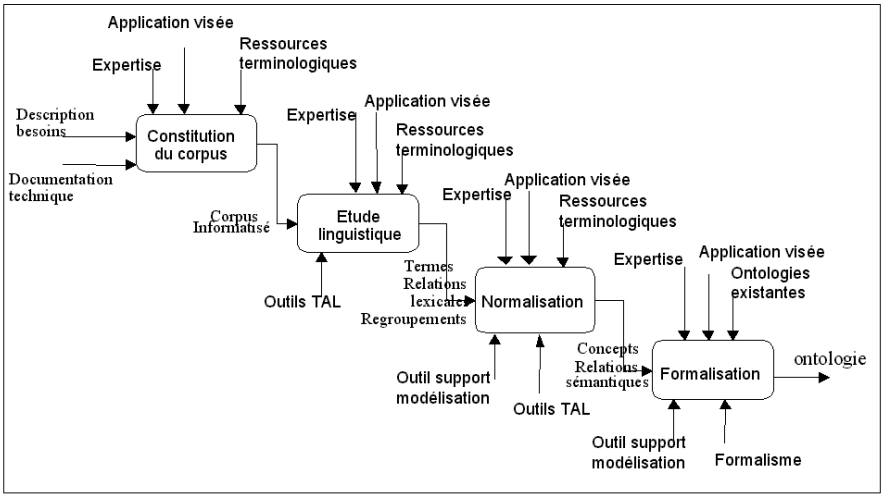

Figure 1. *Le processus de construction d'ontologie de* TERMINAE

langage naturel non contraint, qui rende compte du contenu des textes. Ces éléments sont consignés dans des fiches « terminologiques ».

La deuxième étape de la normalisation consiste à définir des concepts et des relations sémantiques à partir des termes et des relations lexicales précédentes. Il faut en donner une description normalisée, reprenant les étiquettes de concepts et de relations déjà définis, et pertinente par rapport à la tâche pour laquelle le modèle est construit.
Dans le cas d'une ontologie, les descriptions sont consignées dans des fiches de modélisation semi-formelles, au sens où seule la rigueur du travail garantit la cohérence du modèle. Elles comportent des relations prédéfinies, dont l'outil fournit la liste et la signature, afin d'en contrôler la définition et d'en limiter le nombre. De nouvelles relations peuvent être définies selon les besoins en précisant leurs caractéristiques (domaine, valeur, place dans la hiérarchie).
Dans le cas d'une terminologie, les descriptions sont consignées et structurées dans un réseau conceptuel lui aussi semi-formel. Cette structuration nécessite souvent la création de nouveaux concepts et relations. Il n'y a ni validation ni classification puisque l'étape suivante n'est pas effectuée.

Formalisation. La formalisation comprend l'élaboration et la validation de l'ontologie. Des ontologies existantes, générales ou proches du domaine, ou même un glossaire, peuvent aider à définir les couches hautes de la base de connaissances en larges sous-domaines. Les concepts et relations sémantiques des fiches de modélisation sont traduits en concepts et rôles formels dans le langage de description de l'ontologie. Ils sont dits terminologiques car issus de termes.

Le langage de description de TERMINAE correspond au langage terminologique d'une logique de descriptions de type $\mathcal{ALNR}$ [WOO 92]. Les concepts y sont décrits par des conditions nécessaires et suffisantes, afin de les organiser en une taxonomie de subsomption le long de laquelle les propriétés sont héritées. Les concepts peuvent ensuite être classifiés dans cette hiérarchie d'après leurs propriétés. La classification n'intervient qu'à la fin du processus de modélisation, lorsque la structure de la taxonomie n'est plus remise en cause et que les concepts sont considérés comme définis correctement par rapport au domaine et à l'application. Pendant la construction de l'ontologie, des algorithmes de calcul et de comparaison de formes normales sont utilisés pour rechercher l'existence de concepts similaires à celui décrit et pour vérifier la validité de l'insertion ou de la modification d'un concept. L'insertion de concepts et de relations provoque parfois la remise en question de la structure existante, car elle doit prendre en compte la correction de l'héritage des propriétés des concepts. Il est alors souvent nécessaire ou utile de rajouter des concepts.

2.3. *Retour d'expérience*

En tant que logiciel de construction d'ontologie, TERMINAE met l'accent sur la formalisation et sur le passage d'informations trouvées dans des textes à leur représentation en logique de description. Or notre expérience de modélisation d'une ontologie des outils de l'ingénierie des connaissances nous a montré les limites d'une approche trop formelle [AUS 00]. Il faut des étapes et des structures intermédiaires entre les termes trouvés dans les textes et la représentation des objets qu'ils dénotent, comme proposé dans [NOB 00]. De plus, il est difficile d'identifier d'emblée un bon noyau de primitives conceptuelles. Cela suppose une démarche descendante, partant de la racine pour décrire des concepts primitifs, et définir ensuite des concepts plus spécifiques. La pratique a montré que, pour cela, une première structuration d'ensemble des concepts formels devait être stabilisée.

Ces éléments nous ont incitées, d'une part, à améliorer les capacités de recherche dans les textes de TERMINAE et, d'autre part, à définir un réseau conceptuel intermédiaire, utilisant les structures formelles mais non interprété. En effet, il est nécessaire de pouvoir modifier la structuration de l'ontologie avant d'atteindre un état stable, et donc de disposer d'une version de travail informelle dont on puisse repérer les parties non validées. Si l'application nécessite des capacités inférentielles en plus d'une structuration de connaissances, le passage à la formalisation reste toujours possible. TERMINAE devient ainsi un logiciel adapté à la structuration de terminologies.

3. Des logiciels pour assister la construction de terminologies à partir de textes

Parmi les logiciels permettant de construire des terminologies, nous écartons d'emblée les bases de données terminologiques et ne considérons que les bases de connaissances terminologiques ou BCT. Une BCT comporte une *terminologie*, ensemble des descriptions des termes désignant des concepts, un *réseau conceptuel* qui précise le

sens de ces concepts et des *textes* dont ces connaissances sont tirées. Les textes justifient la définition des termes, et donc la structuration des connaissances retenues dans le réseau conceptuel ainsi que l'association terme/concept. Nous présentons ci-dessous plusieurs environnements de contruction de terminologies, et plus particulièrement Géditerm dont nous nous sommes inspirées pour faire évoluer TERMINAE. Nous les différencions des environnements de construction d'ontologies que nous présentons ensuite, et pour lesquels nous soulignons en quoi ils sont mal adaptés à la construction de terminologies à partir de textes. Enfin, nous dressons un panorama d'outils de TAL qu'il nous semble pertinent d'intégrer dans un processus et dans une plate-forme de construction de ressources terminologiques.

3.1. *Des environnements pour structurer des terminologies*

Plusieurs caractéristiques permettent de confronter logiciels et langages de construction de BCT [CON 00] : les utilisateurs qu'ils ciblent (linguistes ou cogniticiens), le rôle qu'ils donnent à l'application visée (prise en compte ou non dans le choix et la structuration des données), la place accordée au corpus étudié (présent ou non dans la BCT), la richesse de leur modèle de données et son degré de formalisation et, enfin, la part du processus de modélisation qui est couverte.

Ainsi, CODE4 [SKU 94] et son successeur DockMan [SKU 98] sont des outils de construction de terminologies formelles, ou d'ontologies dans le cas de CODE4, s'appuyant sur un langage logique. Destinés à des cogniticiens, ils permettent avant tout de structurer et de décrire un réseau conceptuel. CODE4 accordait peu de place aux textes, alors qu'ils sont fondamentaux dans DockMan, puisque l'objectif affiché est que les connaissances modélisées à partir des textes facilitent ensuite la recherche d'information dans ces textes.

D'autres environnements utilisent aussi des langages formels (frames, langage à objets, logique de descriptions ou graphes conceptuels) pour représenter les structures d'une BCT. Parmi ceux-ci, citons Hytropes [EUZ 96] et la BCT de N. Capponi [CAP 96] qui sont des éditeurs de modèles, ou encore CGKAT [MAR 95] qui vise la recherche d'information. Ces langages ne sont utiles qu'à la fin du processus de modélisation, lorsqu'ont été déterminés les concepts pertinents pour l'application. Ils doivent être choisis en fonction des types de traitements à effectuer sur les données : classification de nouveaux concepts, recherche de connaissances à l'aide de requêtes, selon des critères sémantiques ou structurels. Les textes et la terminologie sont alors des ressources intermédiaires et non des résultats en soi, l'objectif essentiel étant le modèle formel.

Au contraire, certains outils sont plus focalisés sur les premières phases de l'analyse de texte, sur le repérage de concepts et de relations. Généralement destinés aux

linguistes, ils accordent plus de place à l'analyse linguistique. Ainsi, HTL [BOU 96], interface de validation et d'organisation des résultats de l'extracteur de candidats-termes Lexter, permet de regrouper des termes synonymes, de définir des concepts et de les relier entre eux, tout en gérant les occurrences des termes associés. Plus qu'une simple interface de saisie, ACI [ASS 98] propose des outils d'analyse des sorties de Lexter, comme Lexiclass, pour aider à dégager des familles de concepts, à les structurer d'abord localement puis au sein d'une ontologie. Cette plate-forme de modélisation autorise un suivi depuis l'analyse des textes jusqu'à la formalisation, mettant l'accent sur la validité locale des résultats ainsi établis.

Un des environnements les mieux adaptés à la représentation d'une BCT non formelle à partir d'une analyse linguistique est SystemQuirk [AHM 00]. Cette plate-forme offre un concordancier et un extracteur de candidats-termes pour repérer des termes, les décrire et alimenter ensuite la BCT. L'accent est mis sur l'organisation des termes et la définition des concepts, peu sur le réseau conceptuel. Bien qu'affichés comme indépendants de toute application, les modèles obtenus avec SystemQuirk conviennent mieux pour la traduction depuis ou vers l'anglais.

Finalement, c'est Géditerm, environnement de gestion de BCT proche de SystemQuirk, qui a le plus influencé nos choix pour spécifier une évolution de TERMINAE. Développé en étroite collaboration avec des linguistes, ce logiciel cherche à répondre aux besoins d'un terminologue qui veut structurer une terminologie sur support informatique à partir d'analyses linguistiques d'un corpus [AUS 99]. Il intègre toutes les composantes de la BCT, y compris le corpus. Ainsi, à partir de l'étude de l'usage des mots en corpus, sont définis des termes, des concepts et des relations entre concepts. Les termes sont décrits sous forme de fiches à l'aide d'informations classiques en terminologie et paramétrables par l'utilisateur (langue, catégorie grammaticale, ellipses possibles, autorisées ou interdites par certaines normes, abréviations, etc.). Des expériences menées avec Géditerm ont montré qu'une BCT doit s'appuyer sur une structure de donnée non formelle [CON 00]. En fait, le réseau conceptuel de la terminologie est un résultat à part entière, même s'il peut ensuite être repris pour une éventuelle formalisation.

3.2. *Des environnements pour construire des ontologies*

De nombreux outils d'aide à la création et à la mise au point d'ontologies existent, la plupart accessibles publiquement sur le web et non commercialisés. Ces outils reposent sur des langages de représentation plus ou moins puissants, qui reprennent pour la plupart les acquis des langages de frames et des logiques de description.

Initialement, le problème de la construction des ontologies a été abordé dans l'optique de favoriser leur réutilisation, et, étudié par des informaticiens, il a été ramené à un problème d'interopérabilité, de langage et de format d'échange. C'est ainsi qu'Ontolingua, historiquement le premier outil dédié à la construction et à l'échange d'on-

tologies, est orienté réutilisation, par fusion et extension, d'ontologies existantes disponibles dans une bibliothèque, et autorise l'exportation d'ontologies dans différents formats. Il permet à un utilisateur, ou groupe d'utilisateurs, de visualiser des ontologies existantes et de construire coopérativement de nouvelles ontologies. Il s'appuie sur un langage compatible avec le format d'échange KIF, ce qui est supposé assurer une facilité de réutilisation des ontologies, pour lequel il offre des interfaces de saisie et d'organisation des éléments de l'ontologie.

Dans ce même esprit, toute une lignée de langages et d'environnements assistent la création d'ontologies avec la même gamme de propositions :

– un langage formel plus ou moins expressif et en général d'autant moins puissant pour réaliser des inférences qu'il est plus expressif ;

– des interfaces de saisie de concepts, relations, axiomes et heuristiques selon le langage ;

– des moyens de vérifier l'organisation des connaissances au sein de l'ontologie : visulation graphique du réseau conceptuel ou des concepts en hiérarchie ; gestion des listes par type d'objets, etc. ;

– des moyens pour vérifier formellement la définition de l'ontologie conformément à la sémantique du langage : par classification de concepts, par vérification de l'unicité des concepts ou de la complétude du modèle.

Pour un état de l'art exhaustif, voir [COR 00] sur les langages de descriptions d'ontologies et [DUI 99] sur les environnements de développement qui font un état de l'art des travaux sur les ontologies. Dans le cadre du projet Européen OntoWeb, un inventaire des outils et méthodes pour la construction d'ontologies est en cours ([URLc]).

Du côté des langages de modélisation d'ontologie, la volonté de standardisation et les convergences des différentes approches en cours ont débouché sur le choix de OIL ([FEN 00] et [URLb]) comme standard. Ce langage, orienté vers la mise à disposition de connaissances sur le web, combine les points forts des frames, des logiques de descriptions et de XML. Son environnement, OilEd, est un simple éditeur d'ontologie, interfacé avec un classifieur en logique de description FaCT. Il permet la traduction d'une ontologie OIL en RDF et XML.

En se focalisant sur la réutilisation, ces outils nous semblent mettre de côté les vrais problèmes relatifs à la construction d'une ontologie : avant même de savoir comment les formaliser, le cogniticien doit d'abord trouver les bons concepts à conserver dans l'ontologie et leurs propriétés ou relations, et cela à partir des entretiens avec les experts et utilisateurs, ou, de manière complémentaire, par l'étude de documents. Il doit pouvoir vérifier qu'ils ont du sens pour des utilisateurs ou experts du domaine. Enfin, il doit pouvoir juger de la bonne adéquation de cette ontologie à l'application visée.

Or le choix des concepts et leur description s'effectuent au niveau des connaissances. La formalisation n'est ici d'aucune aide. Un environnement de construction d'ontologie devrait donc prendre aussi en charge la spécification de ce modèle avant sa formalisation. Une des équipes à s'intéresser aux ontologies au niveau conceptuel,

et à la définition d'une vraie méthodologie, est celle de A. Gomez-Pérez [BLÁ 98] avec le projet Methontology. Leur environnement offre une interface présentant les concepts et leurs propriétés sous forme de tables, puis permettant de saisir des instances de concepts selon ce modèle. Enfin, le DFKI [MAE 00] propose un environnement pour construire des ontologies par apprentissage à partir de textes étiquetés à l'aide d'analyseurs syntaxiques et lexicaux. L'environnement repose sur le langage et l'interface de structuration d'ontologie OntoEdit. Plusieurs outils d'analyse de textes et d'apprentissage peuvent être choisis et combinés en fonction de l'application visée. Le résultat obtenu au terme d'un processus interactif de validation des connaissances acquises est une ontologie formelle de laquelle il n'est pas possible de revenir au texte.

On voit donc clairement que, pas plus que la version initiale de TERMINAE, la plupart des environnements de construction d'ontologies ne peuvent répondre directement aux besoins liés à la construction de terminologies à partir de texte, essentiellement pour deux raisons :

- ils n'accordent pas une place suffisante aux textes, soit comme source de connaissances, soit comme composante de la structure de donnée finale ;

- ils se focalisent trop sur la formalisation alors que le niveau conceptuel convient mieux pour une terminologie.

3.3. *Des logiciels de TAL pour faciliter le repérage de connaissances*

L'étude des textes pour y repérer des connaissances à l'aide de logiciels de traitement automatique des langues connait aujourd'hui un essort important car ces outils semblent parvenus à une maturité suffisante pour être utilisés par d'autres communautés que les chercheurs en TAL. En effet, certains logiciels conduisent à des résultats presque directement exploitables comme des listes de groupes de mots ou des propositions de relations. Cet essor s'explique aussi par une convergence de vues entre les outils disponibles, leurs possibilité d'adaptation et les besoins en modélisation. Plusieurs types d'applications (recherche d'information, aide à la rédaction, etc.) requièrent l'utilisation combinée de ces logiciels. Bien que l'objectif final soit d'informatiser le plus possible, à l'aide de techniques d'apprentissage en particulier, la plupart des approches actuelles envisagent la modélisation comme un processus interactif et cyclique conduit par un analyste qui tient compte des indications fournies par différents logiciels pour décider des connaissances à modéliser.

Les logiciels disponibles aujourd'hui pour dépouiller des textes sont de plus en plus diversifiés, et certains d'entre eux intègrent les acquis et techniques des précédents pour faire des propositions spécialisées, toujours mieux adaptées à la structuration de terminologies et d'ontologies. Plutôt qu'un inventaire exhaustif, nous mentionnons ci-dessous les types de logiciels les plus communément utilisés :

– les analyseurs syntaxiques (comme Cordial Université [1]) permettent de préparer les textes pour les exploiter ensuite à l'aide d'autres logiciels pour y rechercher des

1. Distribué par la société Synapse Developpement.

informations particulières ; le fait que ces outils élémentaires soient plus performants et plus facilement disponibles depuis quelques années ouvre de nouveaux horizons ;

– les extracteurs de candidats-termes, comme Lexter [BOU 94] ou Nomino [DAV 90], permettent désormais de lister non seulement les groupes nominaux mais aussi les verbes et groupes verbaux ;

– les concordanciers, qui facilitent la recherche de patrons lexico-syntaxiques (Sato [DAO 92]), guident peu les utilisateurs mais sont utiles pour mettre au point des moyens précis de recherche ;

– les extracteurs de relations : certains, comme Caméléon [SEG 01] ou Prométhée [MOR 99], s'appuient sur des marqueurs de relations généraux ou spécifiques ; ils font appel alors à des fonctions proches de celles des concordanciers ; d'autres, comme Likes [URLa] font des recherches plus statistiques de segments répétés par exemple ;

– les outils coopératifs de construction de classes de mots, comme ASIUM [FAU 00], exploitent les régularités de construction de phrases dans les textes pour proposer des patrons de fouille du texte ; ces patrons peuvent suggérer une structuration des concepts, et facilitent la recherche d'instances dans les textes ;

– les logiciels de classification deviennent également de plus en plus précis dans la mesure où certains s'appuient désormais non seulement sur les cooccurrences, mais aussi sur les rôles syntaxiques des mots regroupés, pour suggérer des classes ;

– des logiciels plus sophistiqués combinant plusieurs de ces techniques pour fournir des indicateurs de classes et de relations. Ainsi, Syntex [BOU 00] s'appuie à la fois sur une analyse linguistique (syntaxique et lexicale) et sur une analyse distributionnelle pour suggérer des termes, des classes de termes et des relations entre eux.

Le pas à franchir aujourd'hui du côté de l'ingénierie des connaissances est d'évaluer les apports complémentaires de ces outils, d'imaginer des scénarios d'utilisation optimale en fonction des applications, et surtout de mieux les intégrer pour favoriser le dépouillement conjoint (et non indépendant) de leurs résultats. C'est dans cet esprit que nous avons intégré dans TERMINAE la possibilité de dépouiller des résultats de Lexter appliqué à un corpus d'étude, et un module de recherche de motif dans ce corpus, Linguae.

4. TERMINAE, plate-forme pour la gestion de terminologie

Nous tirons de ces expériences quelques conclusions sur ce type d'environnement. Tout d'abord, les rubriques des structures de données, en particulier celles décrivant les termes, doivent pouvoir être adaptées à chaque application. Ensuite, la formalisation des données doit intervenir dans une phase finale, après leur structuration rigoureuse en fonction des besoins de l'application. La trace des choix de modélisation doit être conservée sous plusieurs formes : à l'aide de commentaires dans chaque structure, grâce au lien vers des occurrences des termes dans les textes et enfin grâce au lien entre les structures elles-mêmes. Enfin, le logiciel doit permettre d'exploiter facilement les résultats d'outils d'analyse de textes et d'adapter leur utilisation à chaque type d'application.

TERMINAE a donc évolué afin de prendre en compte les résultats d'expérience décrits ci-dessus, et permet maintenant de décrire des terminologies. Le logiciel intègre des outils linguistiques en offrant quelques mécanismes simples de traitement de corpus par la recherche de motifs (Linguae). De plus, il autorise une structuration des concepts dans un réseau conceptuel. Cette représentation des connaissances s'avère moins contraignante que le processus de modélisation formelle prévu dans les fiches de modélisation : au lieu de commencer par définir formellement un concept et ses rôles puis de le classer, le cogniticien doit situer un concept par rapport aux autres concepts de la hiérarchie et peut se contenter (provisoirement) de définitions locales des rôles. La figure 2 rend compte du processus de constitution d'une terminologie et définit ce qu'est une terminologie dans TERMINAE.

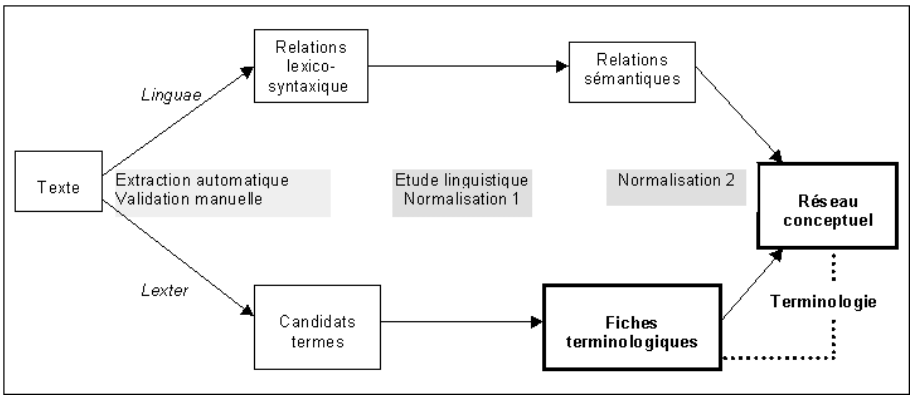

Figure 2. *Processus de constitution d'une terminologie dans* TERMINAE

Pour illustrer la méthode proposée par TERMINAE et l'utilisation du logiciel, nous prenons comme exemple la constitution d'une terminologie à partir d'un texte portant sur la fabrication du verre. Ce texte court (3 pages, 1 425 mots) est extrait d'un manuel pédagogique destiné à des élèves du secondaire [LAJ 62]. Le manuel a été rédigé par un expert d'une entreprise verrière pour présenter « les bases technologiques de l'industrie et des applications du verre ». L'étude du texte sert de point de départ pour organiser en une terminologie structurée des connaissances sur la nature des matières vitreuses et sur la fabrication du verre. La terminologie rend compte des connaissances contenues dans le texte.

4.1. *Utilisation de Lexter*

TERMINAE permet de travailler directement sur les résultats bruts fournis par Lexter pour en dégager des termes pouvant donner lieu à la définition de concepts. Les résultats de Lexter sont regroupés selon trois fichiers dont l'un contient les occurrences des candidats-termes et, pour chacune d'elles, la séquence du corpus l'incluant, un autre contenant la liste des candidats-termes. TERMINAE permet de supprimer auto-

matiquement de ces fichiers les candidats-termes non pertinents à partir d'indications fournies par l'utilisateur : liste de mots, catégorie grammaticale (adverbe, déterminant, etc.), ou caractéristiques particulières (les nombres, les mots comportant des caractères spéciaux, les mots en majuscules, etc.). L'interface permet aussi d'explorer la liste des candidats-termes et de leurs occurrences à l'aide de requêtes pour éliminer manuellement des erreurs.

A partir des fichiers ainsi nettoyés, un expert concerné par cette terminologie peut valider les candidats-termes, afin de ne retenir que ceux pertinents pour le domaine considéré. Le travail, manuel, consiste à :

– retenir les candidats-termes pertinents pour le domaine considéré,

– regrouper des candidats-termes, en particulier les synonymes syntaxiques,

– supprimer des candidats-termes.

A partir du texte sur le verre, 472 candidats-termes (correspondant à 790 occurrences dans le texte) ont été extraits. Parmi les plus fréquents, on retrouve des mots simples très significatifs du contenu du texte : *verre*, *température*, *liquide*, *oxyde*, *élément*, *atome*, *solide*... La fenêtre de la figure 3 montre les occurrences du candidat-terme *liquide* dans le texte.

Les groupes nominaux constituent en général d'excellents points d'entrée pour étudier des textes et repérer des concepts. Dans le cas de cette étude, la petite taille du texte fait ressortir les candidats-termes composés avec une faible fréquence (la plupart ont pour fréquence 1). Il ne faut pas pour autant les négliger car, étant moins génériques, ils sont souvent plus révélateurs de concepts que les candidats-termes simples. Parmi les groupes nominaux extraits, on trouve *état vitreux*, *température de liquidus*, *oxyde formateur*, *solide rigide*, *verre courant*, *composition chimique*, *matières minérales* ou encore *composé cristallin*.

TERMINAE crée ensuite le fichier des candidats-termes validés et de leurs occurrences. Une fenêtre spécifique permet à l'utilisateur de sélectionner, parmi les candidats-termes simples et composés, ceux pour lesquels il souhaite créer une fiche terminologique. Sur cette fenêtre (figure 3), les candidats-termes validés apparaissent triés par ordre de fréquence décroissant. Sélectionner un candidat-terme permet de voir l'ensemble de ses occurrences (partie droite de la fenêtre). La fenêtre de la figure 3 montre les occurrences du candidat-terme *liquide* dans le texte.

4.2. *Elaboration des fiches terminologiques*

4.2.1. *La fiche terminologique*

Un candidat-terme validé accède au statut de terme lorsqu'une fiche terminologique est créée pour le décrire. Reprenant les choix de Géditerm, la fiche terminologique rend compte de phénomènes comme la synonymie (deux termes associés au

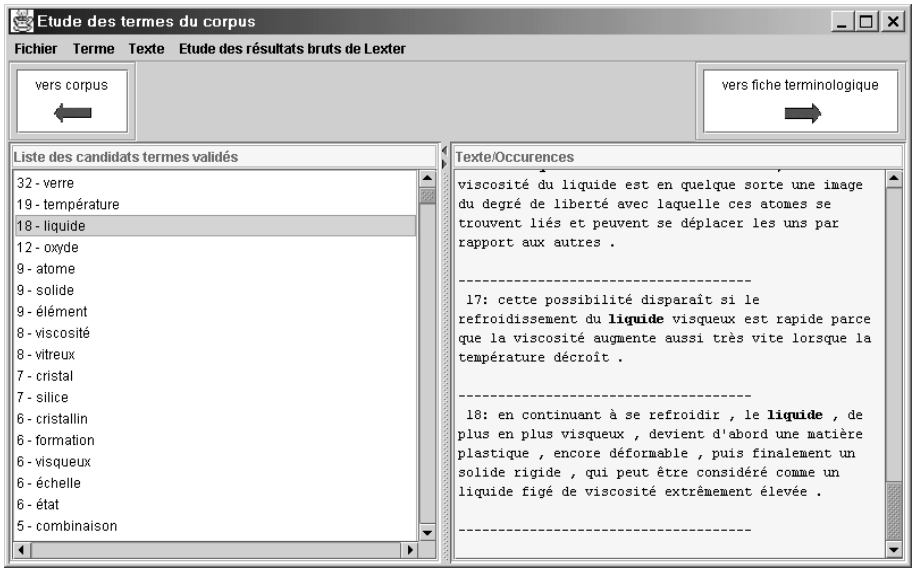

Figure 3. *Fenêtre d'étude des termes du corpus*

même concept) et la polysémie (un même terme associé à deux concepts). Elle décrit un terme, ses concepts associés, leurs synonymes. La figure 4 présente la fiche du terme *liquide* qui a un seul concept associé (étiqueté *corpsLiquide*).

Cette fiche comporte des informations qu'il est classique de trouver dans une fiche terminologique : des informations relatives à sa création et ses mises à jour (auteur et date de création sur la figure 4) et des informations lexicales associées au terme sous forme de rubriques. Le choix des rubriques est fait en début de modélisation en fonction de l'utilisation qui sera faite de la terminologie. Cependant, l'utilisateur peut toujours ajouter ou supprimer une rubrique lexicale. Sur la figure 4, il s'agit de la langue, de la catégorie grammaticale et du genre.

Sur la fiche, chaque concept associé au terme concerné peut être visualisé en sélectionnant ce concept dans la sous-fenêtre Concepts. Les informations suivantes relatives au concept sont affichées et peuvent être modifiées :

– un ensemble d'occurrences dans le corpus (celles du terme qui le désigne et de ses synonymes),

– une définition en langage naturel,

– les synonymes,

– les termes proches (rubrique « Voir aussi ») du terme qui le désigne (*température de liquidus*, *viscosité*).

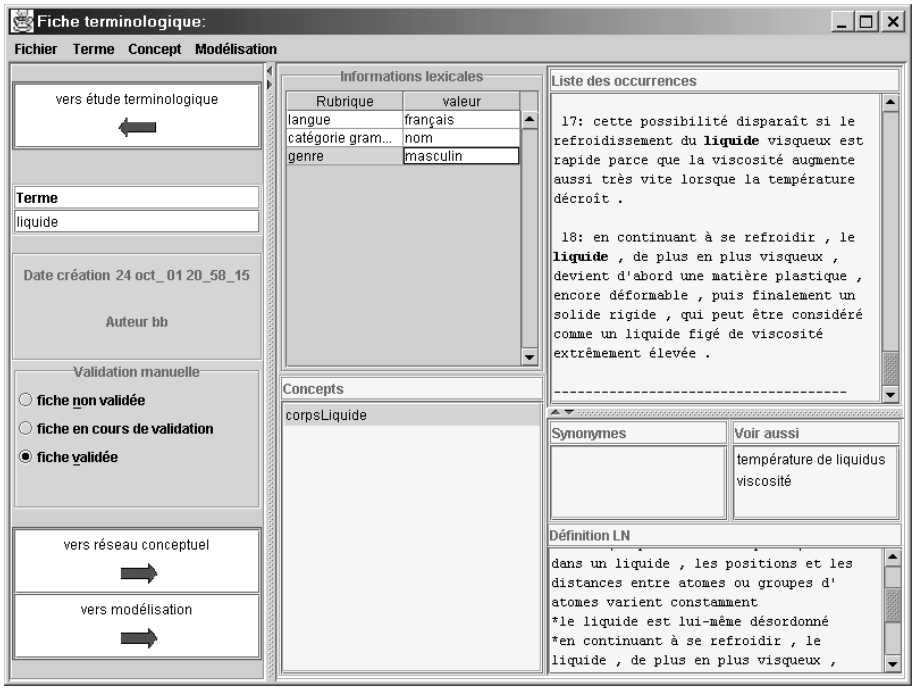

Figure 4. *Fiche terminologique du terme liquide*

Un synonyme est un terme dont au moins un sens est identique au concept étudié. Ce terme fait lui aussi l'objet d'une fiche terminologique. Le sens commun à ces termes sera représenté par un seul concept, qui sera étiqueté par un seul des termes (le terme « vedette »).

4.2.2. *Construction de la fiche terminologique*

Une fiche terminologique rend compte de la première partie du processus de normalisation (cf. paragraphe 2.2 et Normalisation 1^{re} étape sur la figure 2).

Pour décider de la création d'une fiche terminologique, l'un des points de départ possibles consiste à considérer d'abord les candidats-termes les plus fréquents (*verre*, *température*, *liquide*, *oxyde* dans notre cas). Un autre pourrait être de se focaliser sur un candidat-terme renvoyant à un concept central, comme *état vitreux* ou *viscosité*. Il est aussi intéressant de partir des relations trouvées dans le texte à l'aide de Linguae et de considérer les candidats-termes participant à une relation.

Lorsque l'on part d'un candidat-terme, la fenêtre de la figure 3 permet de visualiser les occurrences des groupes nominaux dont fait partie le candidat-terme. Par exemple

pour *verre*, l'observation des occurrences fait ressortir les expressions suivantes dont le terme *verre* est « en tête »(au sens de [BOU 94]) :
verre de silice, *verre d'oxydes*, *verre courant*
ou encore des verbes dont *verre* est sujet :
`verre peut contenir des formateurs`,
`verre peut contenir des modificateurs fondants`,
`verre peut contenir des modificateurs stabilisants`
Enfin, on peut lister des actions possibles sur le verre :
`réchauffer un verre`,
`former des verres`
ainsi que des caractérisations relatives au verre que l'on va pouvoir trouver dans le texte :
`les compositions du verre`,
`les propriétés du verre`,
`les applications du verre`.
L'importance du candidat-terme dans le texte justifie la création d'une fiche terminologique.

Une fois la fiche créée, on remplit les rubriques lexicales et informationnelles. On détermine s'il y a un ou plusieurs concepts et on distribue les occurrences sur les différents concepts. Dans notre corpus, le terme *cristal* n'a pas la même signification sous les formes « cristal »et « cristaux », le premier désignant un type de solide, le second désignant les éléments constituant les solides. Deux concepts vont être créés dans la fiche *cristal*, chacun avec ses occurrences et sa définition.

Cette définition textuelle est construite à partir de l'interprétation des occurrences. Pour *liquide*, un seul sens est présent dans le texte. L'étiquette du concept est proposée automatiquement, elle peut être modifiée lors de la structuration du réseau, comme ce sera le cas pour *liquide*.

L'analyse des occurences de *liquide* montre une régularité d'apparition des termes *viscosité* et *température de liquidus* avec *liquide*, ce qui conduit à indiquer ces termes dans la rubrique « Voir aussi ». Aucun synonyme n'ayant été trouvé, la rubrique « Synonymes »reste vide. Les synonymes peuvent être repérés par l'application sur le corpus de marqueurs de synonymie grâce au module Linguae puis par l'étude des occurrences de ces marqueurs.

Pour ensuite structurer les concepts dans le réseau conceptuel, il est nécessaire d'étudier les relations entre les concepts, *via* la recherche de relations entre les termes grâce au module Linguae.

4.3. *Linguae*

Ce module offre la possibilité de travailler sur un corpus étiqueté syntaxiquement par l'outil Cordial Université. Linguae permet de projeter des marqueurs lexico-

syntaxiques sur un corpus et de déterminer ainsi des relations entre les termes. Dans Linguae, un marqueur est décrit par un ou plusieurs motifs, selon sa complexité et sa variabilité.

Un motif est constitué d'éléments pris dans la liste suivante : mot, forme lemmatisée d'un mot, type grammatical (adapté de ceux proposés par l'analyseur syntaxique), séparateur lexicographique, nombre de mots fixé ou quelconque.

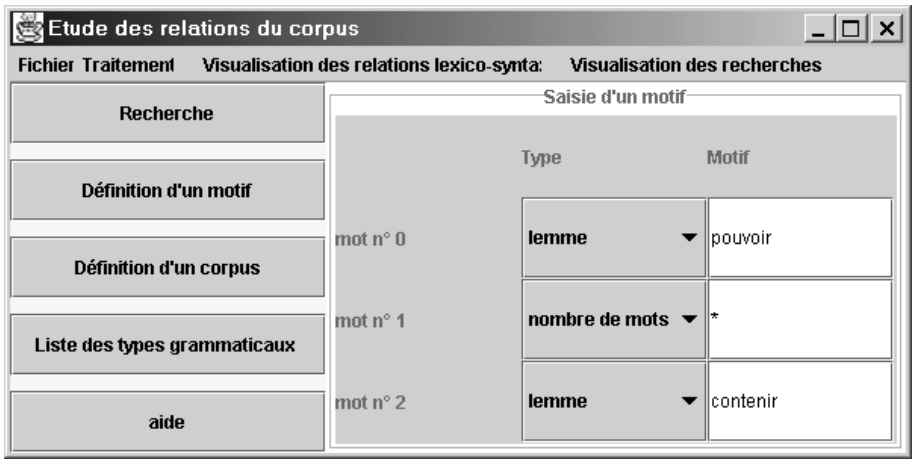

Figure 5. *Définition d'un motif*

La figure 5 présente la définition du motif « pouvoir * contenir », spécifique à ce corpus et révélateur d'une relation de composant/composé. Ce motif est constitué de trois éléments : lemme = pouvoir, nombre_de_mots = *, lemme = contenir. Il permet de retrouver dans le corpus les phrases suivantes :

- Les verres *peuvent contenir* plusieurs formateurs et plusieurs modificateurs ....
- ...les verres *peuvent contenir* en proportions quelconques les différents oxydes dont ils sont constitués.
- La plupart des verres *peuvent généralement contenir* [...] des composés qui ne sont que des oxydes ...

Une relation entre termes, par exemple l'hyperonymie, correspond souvent à un ensemble de marqueurs. Ces marqueurs peuvent être prédéfinis ou spécifiques au corpus. Les relations elles-mêmes peuvent être prédéfinies (hyperonymie, meronymie) ou spécifiques comme les propriétés d'un corps (viscosité d'un corps, température d'un corps). Linguae offre un jeu de relations prédéfinies (hyperonymie, meronymie, synonymie) auxquelles sont associés des marqueurs connus ; il est possible de rajouter des relations spécifiques et des marqueurs supplémentaires. L'ensemble des motifs associés aux marqueurs d'une relation peut être appliqué automatiquement, et

les résultats sauvegardés pour une interprétation ultérieure. Par exemple, dans ce corpus, la relation composant/composé (forme particulière de la relation de méronymie) a au moins deux marqueurs spécifiques : « pouvoir contenir »présenté ci-dessus, et « être_constitué_de ». Le motif de ce deuxième marqueur est le suivant : lemme=être, lemme=constituer, typeGram=PREP.

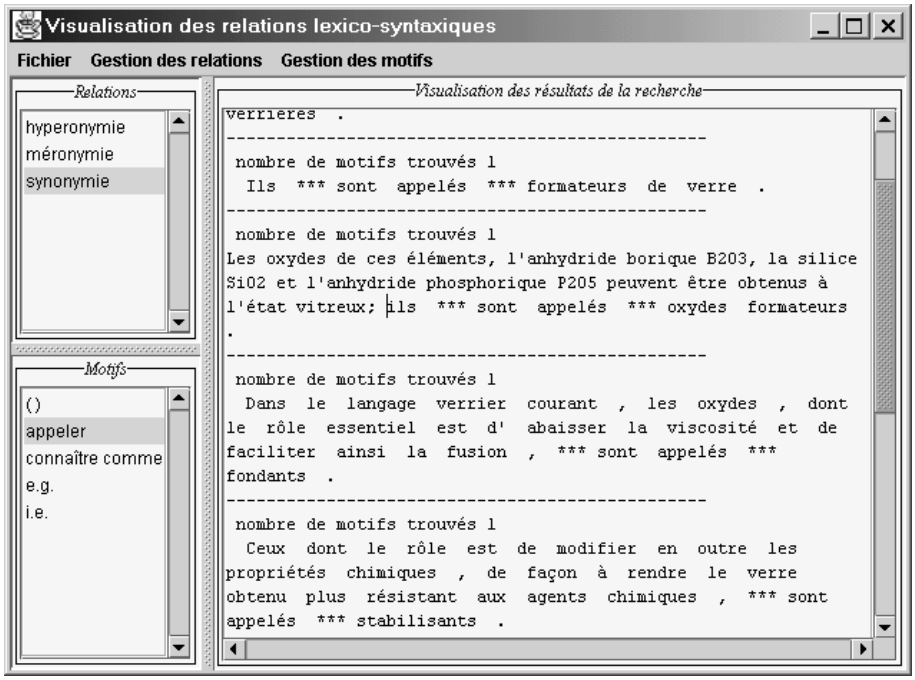

Figure 6. *Fenêtre des relations lexico-syntaxiques*

Linguae permet de visualiser les résultats et de passer des relations entre termes à des relations sémantiques entre concepts (figure 7). L'outil dissocie les deux types de relations, une relation lexico-syntaxique pouvant être associée à une ou plusieurs relations sémantiques. En effet, les résultats de la recherche correspondant à une relation lexico-syntaxique peuvent être distribués sur les différentes relations sémantiques. Il arrive que les résultats d'une relation lexico-syntaxique conduisent à une interprétation sémantique ne relevant pas de la relation de départ.

Par exemple, l'application des marqueurs génériques de synonymie est un cas intéressant où le lien entre une relation lexico-syntaxique et un marqueur doit être adapté au corpus. « appelé »est un marqueur de synonymie ([PEA 98] productif dans le corpus qui fait ressortir 7 occurrences (figure 6). Comme le note Hamon dans [HAM 00], ce marqueur peut correspondre à d'autres relations que la synonymie. Ainsi, dans notre corpus, les relations mises en évidence sont des relations d'hyperonymie.

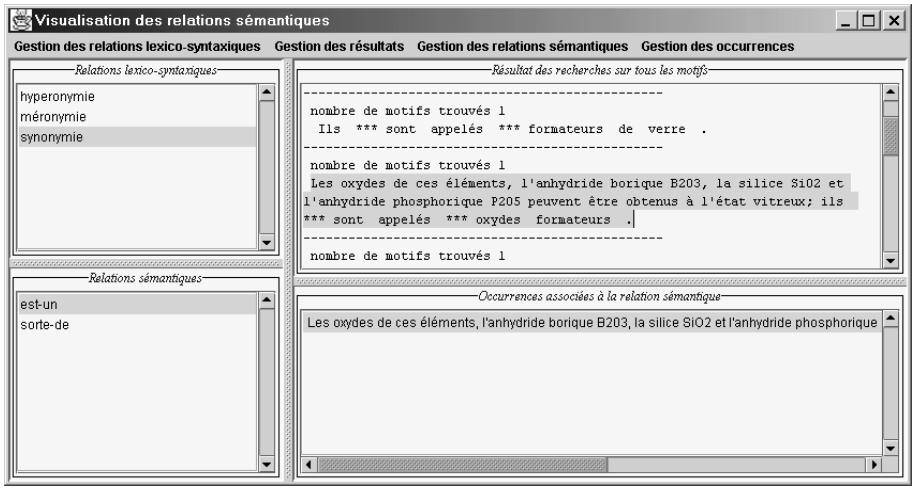

Figure 7. *Fenêtre d'étude des relations sémantiques*

L'une des occurrences du marqueur
`- le produit ainsi obtenu est` *`appelé`* `vitrocéramique`
traduit que la vitrocéramique est une *sous-classe* de produit ; cette relation est interprétée en la relation sémantique classe_sous-classe, nommée SORTE-DE dans le réseau conceptuel.
Une autre occurrence
`- Les oxydes de ces éléments, l'anhydride borique B203, la silice Si02 et l'anhydride phosphorique P205 peuvent être obtenus à l'état vitreux ; ils sont` *`appelés oxydes formateurs`*`.`
met en évidence une relation sémantique classe_instance, nommée EST-UN dans le réseau conceptuel. Cette occurrence justifiera la relation EST-UN entre les concepts individuels ANHYDRIDEBORIQUE, SILICE, ANHYDRIDEPHOSPHORIQUE et le concept générique OXYDEFORMATEUR.

D'autres relations spécifiques sont trouvées en cherchant des cooccurrences. Par exemple, la relation entre un liquide et sa viscosité, qui marque une propriété d'un liquide, apparaît sous la forme *liquide visqueux, liquide peu visqueux, liquide de viscosité élevée*... L'analyse des occurrences montre aussi qu'il y a un lien entre la température et la viscosité d'un liquide : plus la température augmente, plus la viscosité diminue (occurrence 17 sur la figure 4). Des relations sémantiques VISCOSITÉ et TEMPÉRATURE vont être définies sur le concept LIQUIDE et liées à des occurrences les mettant en évidence.

Ce travail complète la première étape de normalisation présentée au paragraphe 2.2. L'étape suivante consiste à passer des fiches terminologiques et des relations sé-

mantiques à un réseau conceptuel. Dans le réseau conceptuel, le concept LIQUIDE possédera deux propriétés désignant la température et la viscosité. La relation liant température et viscosité d'un corps apparaîtra sous forme de commentaire. Les termes *température* et *viscosité* vont chacun produire un concept et une relation dans le réseau. Le concept sera associé à la fiche terminologique et donc aux occurrences du terme ; la relation sera associée aux occurrences du terme cooccurrant avec *liquide*. Les liens des concepts et relations vers les occurrences des termes dans les textes permettent de retrouver l'interprétation qui a conduit à l'organisation de la terminologie.

4.4. *Le réseau conceptuel*

Le réseau conceptuel, moins formel et moins contraint qu'une ontologie, permet ensuite de structurer les concepts correspondant aux termes, appelés « concepts terminologiques », en utilisant les relations mises à jour et les définitions des fiches terminologiques. Le réseau reste à un niveau semi-formel, au sens où l'on utilise les structures du langage formel sans les interpréter systématiquement.

La deuxième étape du processus de normalisation consiste à structurer l'ensemble des concepts et des relations terminologiques définis, que l'outil présente sous forme de listes alphabétiques et de hiérarchie. Cette étape nécessite de fréquents aller-retour avec l'étape précédente, elle est réalisée de façon incrémentale en considérant d'abord des sous-ensembles du modèle puis en les réunissant. La structuration va faire apparaître de nouveaux concepts et relations, et conduire au renommage de certains pour faciliter l'interprétation du réseau.

Par exemple, les concepts FUSION et SOLIDIFICATION sont plus compréhensibles s'ils sont regroupés sous un concept qui exprime la notion de processus chimique. Un retour à l'étape d'analyse du corpus montre qu'aucun terme correspondant n'existe dans le corpus ; le concept PROCESSUSCHIMIQUE sera donc un concept non terminologique de regroupement, qui sert à organiser le réseau. Ce concept est créé en même temps que les concepts FUSION et SOLIDIFICATION.

Lors de l'étude des concepts, il apparaît que la normalisation de LIQUIDE ne peut être dissociée de SOLIDE, car les deux termes jouent des rôles similaires. Le mot *solide* est utilisé sous les formes nominales et adjectivales avec deux sens différents : un corps solide (composé de cristaux) et l'état solide d'un corps qui caractérise une propriété « état »d'un corps. Le terme *corps solide* existe dans le corpus, mais pas *corps liquide*. Le concept LIQUIDE va donc être regroupé avec CORPSSOLIDE sous CORPS, et renommé CORPSLIQUIDE par symétrie avec CORPSSOLIDE. Le concept CORPSLIQUIDE reste terminologique et renvoie à la fiche de *liquide*.

Un ensemble de propriétés des CORPS est créé, sous la forme d'un concept PROPRIÉTÉCORPS regroupant entre autres la température, l'état, le degré d'organisation... Le concept VISCOSITÉ est alors renommé en ÉTATLIQUIDE, par opposition à ÉTATSOLIDE créé pour représenter l'adjectif solide. Garder le nom viscosité nuirait à la lisibilité du réseau, mais le concept ÉTATLIQUIDE renvoie à la fiche terminologique

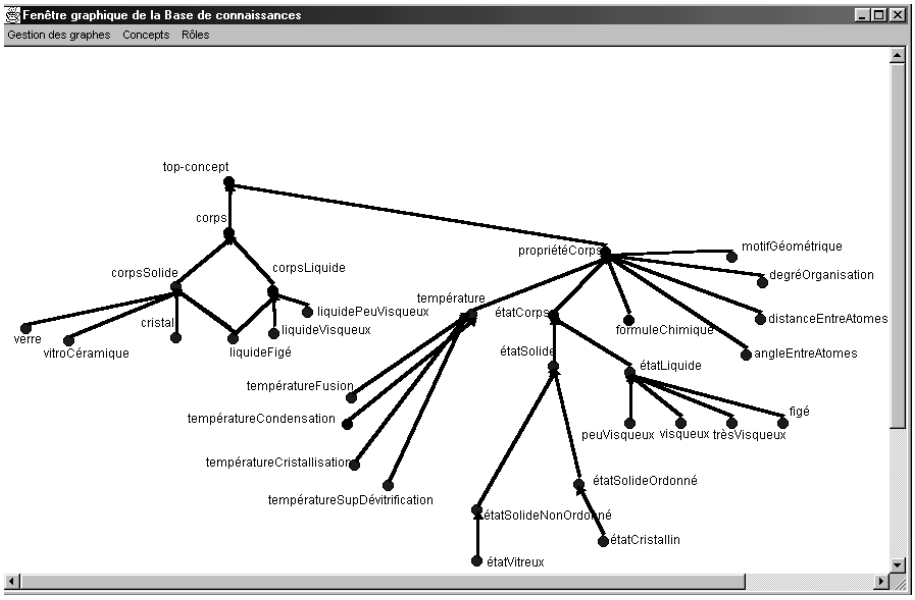

Figure 8. *Hiérarchie des concepts*

viscosité. Les relations attribuant les propriétés aux corps sont créées : A-POUR-TEMPÉRATURE, A-POUR-ÉTAT, A-POUR-DEGRÉ-ORGANISATION, A-POUR-FORMULE...

La relation trouvée dans le corpus exprimant un lien entre les propriétés de viscosité et de température d'un corps n'est pas exprimable dans le langage du réseau conceptuel, elle doit être donnée en commentaire. Ce commentaire portera sur le concept CORPS plutôt que sur CORPSLIQUIDE, car il n'y a pas de distinction absolue entre CORPSSOLIDE et CORPSLIQUIDE. Le corpus (figure 4) explique en effet clairement qu'un liquide de plus en plus visqueux devient un solide rigide, c'est-à-dire un liquide figé de viscosité extrêmement élevée :

```
- 18 : en continuant à se refroidir, le liquide, de plus en plus
visqueux, devient d'abord une matière plastique, encore défor-
mable, puis finalement un solide rigide, qui peut être considéré
comme un liquide figé de viscosité extrêmement élevée.
```

On remarquera donc sur la figure 8 la présence d'un double héritage du concept LIQUIDEFIGÉ vers CORPSSOLIDE et CORPSLIQUIDE. Ce double héritage interdit toute relation d'exclusion entre CORPSSOLIDE et CORPSLIQUIDE.

La figure 9 présente le réseau conceptuel dans lequel est intégré le concept normalisé CORPSLIQUIDE associé au terme *liquide*. La hiérarchie des fils du concept

courant (ici, CORPSLIQUIDE) est affichée au centre, à côté des informations propres à ce concept (rôles et commentaires) et de la liste complète des concepts du réseau. Environ 50 d'entre eux sont terminologiques ainsi qu'une douzaine de relations (rôles).

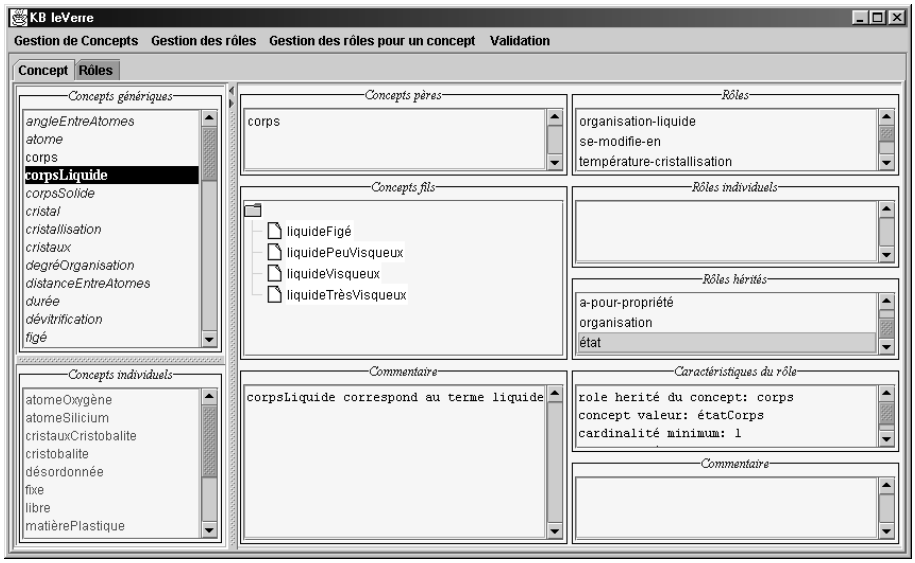

Figure 9. *Visualisation des éléments du réseau conceptuel*

Construire ce réseau revient à structurer hiérarchiquement les concepts normalisés, sans distinguer forcément leurs conditions nécessaires et suffisantes des conditions non définitoires, et sans contraindre le réseau à une validité formelle. La relation hiérarchique est une relation d'héritage, les autres relations étant exprimées sous forme de rôles. La seule vérification porte sur la signature des relations, qui correspond aux caractéristiques du rôle : un concept ne peut avoir deux rôles de même nom, à moins que l'un ne restreigne l'autre. Cela évite de définir un rôle avec des significations incohérentes. D'ailleurs, la liste des rôles déjà définis est disponible à tout moment. De même, le nom d'un concept doit être unique.

Tant que le réseau conceptuel n'est pas figé, la position des concepts et les rôles associés peuvent évoluer. Ensuite, il est beaucoup plus facile de passer à une ontologie formelle si l'application nécessite des déductions. En effet, la vue d'ensemble du réseau conceptuel permet d'abord de choisir les concepts et rôles primitifs, puis de construire l'ontologie selon une démarche descendante en validant incrémentalement chaque définition de concept formel décrite dans une fiche de modélisation.

5. Conclusion

Grâce à ses développements récents, TERMINAE devient une plate-forme pour la construction de ressources terminologiques variées. Il est maintenant possible d'élaborer aussi bien des fiches terminologiques proches des fiches papier classiques, qu'un réseau conceptuel structurant les concepts associés aux termes de manière souple ou une ontologie formelle sur laquelle des inférences peuvent être effectuées. Chaque résultat est visualisable dans l'outil et peut être obtenu sous forme textuelle.

Ainsi, la nouvelle version de TERMINAE se distingue par les évolutions suivantes :

– un positionnement plus clair dans la terminologie sur corpus, en clarifiant le vocabulaire (terme, notion et concept) ;

– une meilleure intégration des outils linguistiques, avec la possibilité de travailler directement sur les résultats de Lexter et la disponibilité d'un module de recherche de motifs lexico-syntaxiques, Linguae ;

– l'ajout d'informations utiles à la pratique terminologique dans les fiches terminologiques ;

– une meilleure adéquation à la modélisation conceptuelle grâce à un outil qui structure les concepts sans imposer les contraintes d'un formalisme trop puissant, et qui permet d'alterner raffinements locaux et restructurations globales ;

– une ouverture vers les autres plates-formes de développement d'ontologies *via* l'intégration de formats standard.

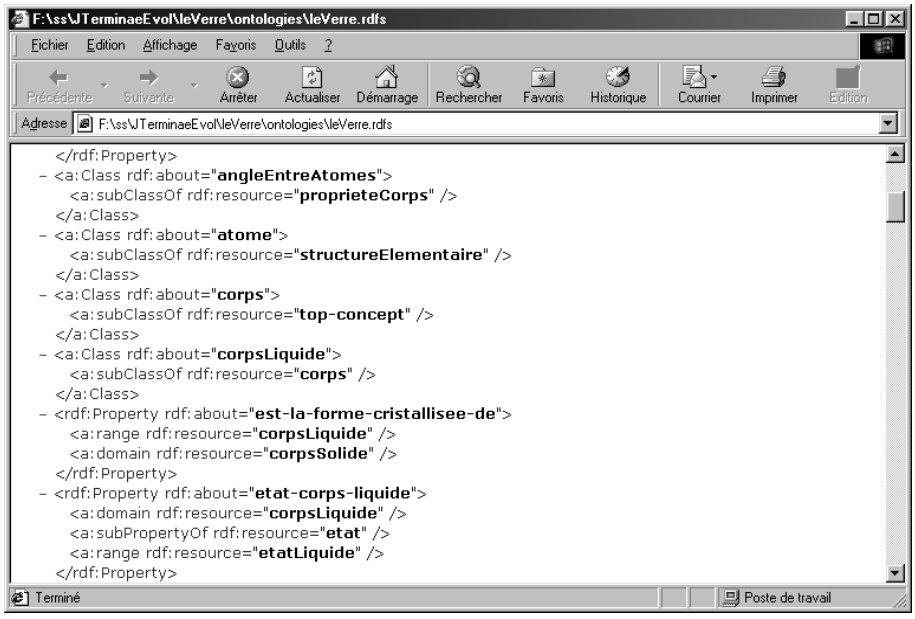

Figure 10. *Le réseau conceptuel en format RDF*

En effet, TERMINAE permet maintenant de générer le réseau conceptuel sous différents formats comme OIL, RDF (figure 10), et également de récupérer des ontologies en provenance de ces formats. Il devient possible d'échanger des ontologies, d'utiliser des logiques de description plus riches que celle de TERMINAE pour enrichir une ontologie, de compléter une ontologie existante d'éléments textuels facilitant une interprétation à partir de textes.

6. Bibliographie

[AHM 00] AHMAD K., « System Quirk », University of Surrey, UK, http ://www.computing.surrey.ac.uk/ai/SystemQ/, 2000.

[ASS 98] ASSADI H., « Construction d'ontologies régionales à partir de textes techniques », Thèse d'informatique, Université Paris 6, Paris, France, 1998.

[AUS 99] AUSSENAC-GILLES N., « GEDITERM, un logiciel de gestion de bases de connaissances terminologiques », *Terminologies Nouvelles*, vol. 19, 1999, p. 111-123.

[AUS 00] AUSSENAC-GILLES N., BIÉBOW B., SZULMAN S., « Revisiting Ontology Design : a methodology based on corpus analysis », DIENG R., CORBY O., Eds., *Knowledge Engineering and Knowledge Management : Methods, Models, and Tools. Proc. of the 12th International Conference, (EKAW'2000)*, LNAI 1937, Springer-Verlag, 2000, p. 172-188.

[BAC 95] BACHIMONT B., « Ontologie régionale et terminologie : quelques remarques méthodologiques et critiques », *La Banque des Mots*, vol. 7/95, 1995, p. 65-86, Conseil International de la Langue Française.

[BIE 00] BIEBOW B., SZULMAN S., « TERMINAE : une approche terminologique pour la construction d'ontologies du domaine à partir de textes », *Proc. of Reconnaissance des Formes et Intelligence Artificielle (RFIA'2000)*, vol. II, 2000, p. 81-90.

[BLÁ 98] BLÁZQUEZ M., FERNÁNDEZ M., GARCÍA-PINAR J., GÓMEZ-PÉREZ A., « Building Ontologies at the Knowledge Level using the Ontology Design Environment », *Proc. of the 11th Knowledge Acquisition Workshop (KAW'98)*, Banff, Canada, 1998.

[BOU 94] BOURIGAULT D., « LEXTER, un Logiciel d'EXtraction de TERminologie, Application à l'acquisition des connaissances à partir de textes », Thèse d'informatique, Ecole des Hautes Etudes en Sciences Sociales, Paris, France, 1994.

[BOU 96] BOURIGAULT D., GONZALEZ-MULLIEZ I., GROS C., « LEXTER, a Natural Language Processing Tool for Terminology Extraction », *Proc. of the 7th International Congress EURALEX*, Göteborg, Suède, 1996, p. 771-779.

[BOU 00] BOURIGAULT D., FABRE C., « Approche linguistique pour l'analyse syntaxique de corpus », *Cahier de Grammaire , Numéro spécial "sémantique et corpus"*, vol. 25, 2000, p. 131-151, Presses Universitaires du Mirail.

[CAP 96] CAPPONI N., « Modélisation d'une base de connaissances terminologiques », Rapport de recherche, 1996, Université de Nancy I, CRIN/LORIA.

[CON 00] CONDAMINES A., AUSSENAC-GILLES N., « Entre textes et ontologies formelles : les bases de connaissances terminologiques », ZACKLAD M., GRUNDSTEIN M., Eds., *Capitalisation des connaissances*, Paris : Hermès, Sciences Publications, Traités IC2, 2000.

[COR 00] CORCHO O., GÓMEZ-PÉREZ A., « A Roadmap to Ontology Specification Languages », DIENG R., CORBY O., Eds., *Proc. of the 12th International Conference*

on Knowledge Engineering and Knowledge Management (EKAW'2000), LNAI 1937, Springer-Verlag, 2000, p. 80-96.

[DAO 92] DAOUST F., « Système d'Analyse de Textes par Ordinateur », Centre ATO, Université du Québec à Montréal, 1992.

[DAV 90] DAVID S., PLANTE P., « Termino version 1.0 », Centre d'Analyse de Textes par Ordinateur, Montréal, Canada, 1990.

[DUI 99] DUINEVELD A., STUDER R., WEIDEN M., KENEPA B., BENJAMINS V., « WonderTools ? A comparative study of ontological engineering tools », *Proc. of the Workshop on Knowledge Acquisition, Modelling and Management (KAW'99)*, 1999.

[EUZ 96] EUZENAT J., « HYTROPES : a www front-end to an object management system », *Proc. of Knowledge Acquisition Workshop (KAW'96)*, 1996.

[FAU 00] FAURE D., « Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM », Thèse d'informatique, Université Paris 11, Orsay, France, 2000.

[FEN 00] FENSEL D., HORROCKS I., VAN HARMELEN F., DECKER S., ERDMANN M., KLEIN M., « OIL in a nutshell », R. DIENG-KUNTZ O. C., Ed., *Knowledge Engineering and Knowledge Management : Methods, Models, and Tools. Proc. of the 12th International Conference, (EKAW'2000)*, LNAI 1937, Springer-Verlag, 2000, p. 1-16.

[HAM 00] HAMON T., « Vérification sémantique en corpus spécialisé : acquisition de relations de synonymie à partir de ressources lexicales », Thèse d'informatique, Université Paris 13, Villetaneuse, France, 2000.

[LAJ 62] DE LAJARTE D., « Monographie de technologie verrière à l'usage de l'enseignement secondaire technique », Saint Gobain Recherche, 1962.

[MAE 00] MAEDCHE A., STAAB S., « Semi-automatic Engineering of Ontologies from Text », *Proceedings of the Twelfth Inernational Conference on Software Engineering and Knowledge Engineering (SEKE'2000)*, 2000.

[MAR 95] MARTIN P., « Knowledge Acquisition using Documents, Conceptual Graphs and a Semantically Structured Dictionnary », *Proc. of the 9th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW'95)*, Banff, Canada, 1995.

[MOR 99] MORIN E., « Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique », *Traitement Automatique des Langues*, vol. 40/1, 1999, p. 143-166.

[NOB 00] NOBÉCOURT J., BIÉBOW B., « Mdos : a modelling language to build a formal ontology in either Description Logics or Conceptual Graphs », DIENG R., CORBY O., Eds., *Knowledge Engineering and Knowledge Management : Methods, Models, and Tools. Proc. of the 12th International Conference, (EKAW'2000)*, LNAI 1937, Springer-Verlag, 2000, p. 57-64.

[PEA 98] PEARSON J., *Terms in Context*, vol. 1 de *Studies in Corpus Linguistics*, John Benjamins, Amsteram/Philadelphia, 1998.

[RAS 91] RASTIER F., *Sémantique et recherches cognitives*, PUF, Paris, France, 1991.

[RAS 95] RASTIER F., « Le terme : entre ontologie et linguistique », *La Banque des Mots*, vol. 7/95, 1995, p. 35-65, Conseil International de la Langue Française.

[SEG 01] SEGUÉLA P., « Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques », Thèse d'informatique, Université Toulouse III, Toulouse, France, 2001.

[SKU 94] SKUCE D., LETHBRIDGE T., « CODE4 : a multi-functional knowledge management system », GAINES B., Ed., *8th Knowledge Acquisition Workshop (KAW'94)*, Banff, Canada, 1994.

[SKU 98] SKUCE D., « Intelligent Knowledge Management : Integration of Documents, Knowledge Bases, Databases and Linguistic Knowledge », *Proc. of the 10th Knowledge Acquisition and Management Workshop (KAW'98)*, Univ. of Calgary, Banff, Canada, 1998.

[SLO 95] SLODZIAN M., « Comment revisiter la doctrine terminologique aujourd'hui ? », *La Banque des Mots*, vol. 7/95, 1995, p. 11-18, Conseil International de la Langue Française.

[URLa] « URL : http ://www-ensais.u-strasbg/LIIA/likes/likes.html ».

[URLb] « URL : http ://www.ontoknowledge.org/oil ».

[URLc] « URL : http ://www.ontoweb.org/ ».

[WOO 92] WOODS W. A., SCHMOLZE J. G., « The KL-ONE family », *Computers Mathematical Applications*, vol. 23, 1992, p. 133-177.