Modélisation du domaine par une méthode fondée sur l'analyse de corpus

Nathalie Aussenac-Gilles, Brigitte Biébow, Sylvie Szulman

Modélisation du domaine par une méthode fondée sur l'analyse de corpus

Nathalie Aussenac-Gilles(*), Brigitte Biébow (**) et Sylvie Szulman (**)

(*)IRIT, Université Toulouse 3, 118, route de Narbonne, 31062 TOULOUSE Cedex 4,
http://www.irit.fr, Nathalie.Aussenac-Gilles@irit.fr
(**)LIPN, Université Paris 13, Av. J.B Clément, 93430 VILLETANEUSE, http://www.lipn.univ-paris13.fr,
{Brigitte.Biebow, Sylvie.Szulman}@lipn.univ-paris13.fr

Résumé

Les nombreux travaux actuels sur les ontologies et modèles de domaines, justifiés par la perspective de leur réutilisabilité, proposent très peu de solutions aux problèmes pratiques de recueil et de structuration de ces connaissances. Cet article propose une méthode de construction de modèles de domaine ou d'ontologies, dont l'originalité est de se fonder sur l'analyse de corpus en utilisant ses principes linguistiques et ses logiciels de traitement automatique de la langue. Cette démarche se veut un complément efficace et précis aux méthodes classiques de modélisation du domaine à partir d'expertises individuelles.

Mots clés : Construction d'ontologies, acquisition de connaissances à partir de textes, terminologie.

1 Introduction

La modélisation de connaissances du domaine d'une application a fait l'objet de nombreux travaux pendant les dix dernières années à travers les recherches sur les ontologies. Plus récemment, ces études ont pris un tournant en s'orientant soit vers la réutilisation comme solution au problème de la construction de ces ontologies, soit vers l'acquisition automatique de connaissances selon des techniques d'apprentissage et de fouille de données, ou encore en se focalisant sur l'intégration des ontologies avec les méthodes de résolution de problème. Toutes les difficultés liées à la construction de ces modèles sont cependant loin d'être résolues.

La plupart des articles faisant un point sur la conception d'ontologies rapportent les mêmes points faciles [19] [29]. Les concepteurs reproduisent les erreurs classiques du développement des premiers systèmes experts. Ils adoptent des principes de modélisation et des démarches spécifiques à leur équipe, et manquent de repères précis. Pire, ils ont l'habitude de passer directement des connaissances brutes à leur implémentation, sans prendre le soin de construire un modèle intermédiaire. De ce fait, les résultats obtenus sont difficiles à interpréter et à maintenir, et biaisés par le formalisme utilisé. Même s'il y en a, les propositions méthodologiques concrètes sont assez rares ou alors jugées caricaturales [18].

A notre connaissance, les problèmes difficiles comme la sélection des concepts, le choix de leurs propriétés et de leurs relations, leur regroupement, l'influence de l'application dans ces choix ou encore la gestion de la masse des connaissances sont peu mentionnés, alors qu'ils sont loin d'être résolus.

Sans apporter de solution universelle et définitive, nous défendons une approche différente reposant sur la linguistique pour aider le concepteur. Cette méthode cherche à réduire plusieurs difficultés en s'appuyant sur des principes novateurs, représentatifs du courant français de travaux à la convergence entre terminologie, linguistique, ingénierie des connaissances (IC) et intelligence artificielle. Au sein du PRC-I3 et de l'AFIA, le groupe TIA, dont les auteurs font partie, anime la communauté française sur ce thème. Ces principes peuvent s'exprimer comme suit :

- partir des textes comme sources de connaissances : ils constituent un support tangible, rassemblant des connaissances stabilisées qui servent de référence ; leur utilisation améliore la qualité du modèle final,
- enrichir le modèle conceptuel d'une composante linguistique : l'accès aux termes et textes qui justifient la définition des concepts garantit une meilleure compréhension du modèle ;
- utiliser des techniques et outils de Traitement Automatique des Langues (TAL) basés sur des travaux linguistiques : ces outils permettent l'exploitation systématique des textes et leurs résultats facilitent la modélisation.

Dans cet article, nous exposons notre méthode de modélisation de connaissances du domaine à partir de textes. Nous situons d'abord (partie 2) cette démarche et ses caractéristiques par rapport aux différents courants de l'ingénierie des connaissances. Nous la présentons ensuite en insistant sur les données produites et les étapes méthodologiques (partie 3). Enfin, nous rapportons une expérience récente de mise en œuvre d'une partie de cette méthode dans le domaine de l'ingénierie des connaissances, en soulignant les traitements linguistiques effectués et leur apport à la modélisation (partie 4).

2 Notre approche dans l'IC

Afin de justifier notre proposition, nous situons la place possible de l'analyse de textes dans la modélisation de connaissances. Nous présentons ensuite quelques approches représentatives de différentes manières d'aborder la modélisation du domaine en IC.

1.1 Modélisation du domaine et cycle de vie

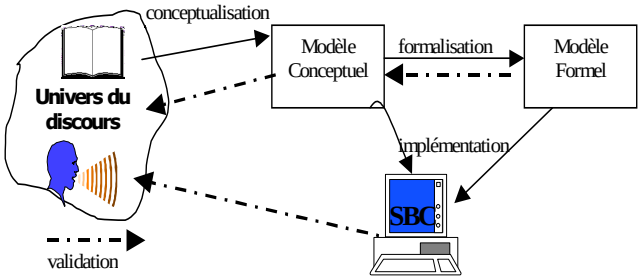

Fig. 1 *Cycle de vie d'un SBC*

La modélisation est l'activité centrale en ingénierie des connaissances. Ce modèle est le support pour sélectionner, recueillir, organiser des connaissances avant de décider comment les intégrer dans l'application finale. Il est élaboré par le cogniticien, aidé d'un expert ou de spécialistes du domaine, à partir d'une étude du besoin auprès d'utilisateurs d'une part, et d'observations, de verbalisations d'experts et/ou de textes d'autre part. Dans le cas de systèmes à base de connaissances (SBC), ce modèle spécifie l'activité du futur système dans lequel il sera implémenté (Fig 1.). D'autres types d'applications utilisent les connaissances contenues dans le modèle sans qu'elles soient implémentées. Le modèle est alors consulté directement en fonction de son organisation (via un hypertexte par exemple), ou bien il sert de ressource directe via une interface, par exemple pour définir des index, des glossaires, des thésaurus. Bien sûr, le type d'application visée influence directement son contenu.

Les sources de connaissances, ou univers du discours au sens des systèmes d'information, comportent des connaissances implicites (savoir-faire, attitudes, etc.) et des connaissances explicites ou explicitées (notes d'observation, récapitulatifs d'entretiens, textes rédigés). A partir de cet univers, le cogniticien définit un modèle conceptuel, qu'il fait valider en fonction des besoins de l'application et de l'expertise [11]. Après cette étape de conceptualisation où objets et tâches sont définis, la formalisation, facultative, consiste à transcrire ce modèle dans un formalisme normalisant la description, afin de vérifier sa cohérence et sa complétude, ou de préparer l'implémentation. Lorsqu'elle est réalisée, l'implémentation a pour but de rendre utilisable, ou opérationnel, le modèle formel en le traduisant dans un langage de programmation. La validation du système passe alors par son utilisation. Chaque étape peut être plus ou moins développée suivant l'application visée. Les retours vers les étapes précédentes sont indispensables pour vérifier la correction de l'étape en cours.

Pour faciliter le repérage et l'analyse des connaissances, il est habituel de distinguer dans un modèle conceptuel les connaissances de résolution de problème des connaissances du domaine. Notre approche se concentre sur les connaissances du domaine, qui se retrouvent également dans les ontologies, sans préjuger de l'intérêt des textes pour modéliser les méthodes de résolution de problème.

1.2 Importance des textes dans le processus de modélisation

Au sein du processus de modélisation, les textes ont été longtemps sous-utilisés, alors qu'ils constituent, lorsqu'ils existent, une source importante de connaissances. Le renouveau récent de l'acquisition de connaissances à partir de textes s'explique par un triple constat :

- Les connaissances stabilisées, qui constituent une part significative des modèles conceptuels, sont souvent décrites dans des textes ;
- Les textes sont de plus en plus sur support informatique, ce qui les rend très accessibles ;
- Il existe de nouveaux outils de Traitement Automatique de la Langue (TAL) parvenus à une maturité suffisante pour être utilisés dans le contexte de l'acquisition de connaissances. Ces outils ont bénéficié de l'augmentation exponentielle des capacités de mémoire et de traitement des ordinateurs, tout autant que de nouveaux résultats en linguistique de corpus.

Il faut maintenant définir en ingénierie des connaissances des outils et des méthodes prenant en compte l'analyse automatique des textes, et utiliser les techniques mises au point en linguistique pour l'étude des corpus. Ce type de méthode s'appuie sur un cadre de référence théorique novateur en linguistique [28]. Ce cadre se démarque d'une linguistique conceptuelle qui fait l'hypothèse (sous jacente à la plupart des travaux sur les ontologies formelles) que le sens des termes peut s'étudier dans l'absolu, comme si le lien terme concept était un lien de référence, unique et figé. Les travaux actuels de la linguistique de corpus, les hypothèses de la sémantique référentielle ou encore l'étude des langues spécialisées offrent des alternatives mieux adaptées à la modélisation de connaissances.

La constitution d'un modèle à partir de textes présente plusieurs avantages. Elle apporte une meilleure lisibilité au modèle et facilite ainsi sa maintenance, à condition bien sûr de conserver un lien entre le texte et le modèle ainsi que des traces des choix de modélisation. En effet, lorsque le modèle doit être interprété, le lecteur fait appel à toutes les connaissances implicites évoquées par le nom des primitives, qui dépassent largement le sens décrit formellement. Le nom d'un concept joue non seulement sur la compréhension de l'utilisateur lorsqu'il interprète des résultats fournis par le système, mais aussi sur celle du cogniticien en cours de modélisation lorsqu'il doit remettre en cause ses choix, ou enfin lors de la maintenance du système en cas de correction ou de

modification du modèle. Ainsi, les textes associés à un modèle rendent explicites une partie des choix de modélisation, et tout particulièrement le choix des étiquettes des concepts et des relations, tout en justifiant leur définition. Ils permettent en fait, d'une part, d'expliquer le sens du terme dans le domaine (son interprétation référentielle) et, d'autre part, de reconstituer la transformation qui a mené au sens formel (restreint mais justifié par l'application) donné à la primitive portant cette étiquette dans l'ontologie.

1.3 Approches classiques pour la construction d'ontologies

Depuis plusieurs années, l'ingénierie des connaissances a élargi sa problématique. Elle vise des applications assistant intelligemment un opérateur, et non plus seulement des systèmes capables de raisonner intelligemment. La place de l'expert et des textes comme sources de connaissances en ont été significativement modifiées [11].

- En se focalisant sur l'expert comme individu ayant acquis, par son expérience, une plus grande compétence qu'il est seul à détenir, on recueille des connaissances individuelles, des savoir-faire à faire expliciter. Une approche basée sur des entretiens et des techniques inspirées de la psychologie est généralement proposée.
- En se focalisant sur les textes et l'intertextualité [28], on privilégie les connaissances stabilisées, les savoir déjà partagés et explicités, comme le vocabulaire et des connaissances ontologiques. Des approches linguistiques sont alors possibles.

En sélectionnant les quelques approches présentées par la suite, nous voulons souligner l'évolution actuelle concernant les sources de connaissances et les techniques utilisées en ingénierie des connaissances. Nous différencions ainsi des approches plus « classiques » (KOD, CommonKADS et les textes sur le construction d'ontologies formelles) presque exclusivement tournées vers les experts, d'approches plus récentes s'appuyant presque uniquement sur les textes (BCT, approche différentielle). Ce cadre nous permettra ensuite de développer notre point de vue qui, refusant d'opposer ces sources de connaissances, cherche à en tirer la complémentarité en fonction de l'application visée.

- KOD [30] : Les textes, essentiellement des retranscriptions d'entretiens avec un expert, ont un rôle central dans KOD. L'exploitation du texte est systématique mais manuelle pour ce qui est du repérage des connaissances (la K-Station permet ensuite de gérer dans une base les connaissances identifiées et les textes dont elles sont issues). KOD s'appuie sur des principes linguistiques et terminologiques. Les groupes nominaux sont mis en évidence pour trouver les concepts, les verbes pour décrire les activités. Ce travail est très coûteux et produit une quantité importante de connaissances, pas toujours pertinentes pour l'application.
- Common-KADS et assimilées : Les textes y tiennent une place marginale. Il s'agit soit de retranscriptions d'entretiens avec l'expert, soit de textes techniques relatifs à l'application. L'exploitation de ces textes est orientée par la tâche que doit réaliser l'application. On y cherche les concepts qui jouent les rôles spécifiques à la méthode de résolution de problème que l'on a associée à cette tâche. Sur le point précis de l'analyse de retranscriptions, les recommandations actuelles de CommonKADS sont tout à fait semblables à celles suggérées dans KOD, mais elles ne sont étayées par aucune référence linguistique. L'exploitation du texte est complémentaire de l'utilisation d'ontologies du domaine quand elles existent.
- Approche ontologique [19][18]: Dans la plupart des travaux sur les ontologies, les textes sont très peu utilisés. Ils sont généralement exploités manuellement de manière arbitraire, soit pour construire les ontologies, soit pour les adapter à une application particulière. La réutilisabilité du résultat est privilégiée, souvent au détriment de sa qualité et donc de son utilisabilité même.

- Approche BCT : Cette approche est purement linguistique [2][21][13]. Son origine provient de la terminologie. Elle s'appuie sur la constitution d'un corpus en fonction de l'application visée, à partir duquel les termes du domaine sont extraits selon des critères linguistiques et à l'interprétation des occurrences. L'étude des occurrences de ces termes et des relations lexicales conduit à définir un réseau de concepts qui les décrit. Les termes étant considérés comme des parties du texte, ils peuvent être polysémiques et c'est leur étude qui permet de dégager leurs différents sens. Des travaux ont fait l'hypothèse que ce réseau pouvait être une étape intermédiaire pour constituer un modèle du domaine[4]. Or la pratique a montré que la BCT devait être fortement remaniée avant d'y parvenir : des connaissances, pertinentes linguistiquement, sont inadéquates pour l'application alors que d'autres doivent être rajoutées [13][14].
- Approche différentielle : Nous appelons ainsi un courant issu des propositions de B. Bachimont et qui a influencé les réflexions issues du groupe TIA [8]. Les textes y sont considérés comme la source presque exclusive de connaissances. Le problème de l'IC est décrit comme celui de la construction de modèles à partir de l'expression linguistique de connaissances. Il s'agit d'assurer une continuité sémantique entre l'interprétation des expressions lexicales, leur organisation structurelle dans le modèle conceptuel et, éventuellement, leur représentation opérationnelle dans le système informatique cible. Plusieurs étapes sont proposées dans ce cheminent, dont la *normalisation* basée sur des principes de différenciation explicites, qui fait du modèle conceptuel une ontologie appelée « ontologie régionale ». Dans cette ontologie, les termes sont des étiquettes

de concepts, non polysémiques puisqu'ils résultent de choix effectués au cours de l'interprétation et de la normatlisation. Une différenciation formelle permet ensuite de construire des concepts formels pour lesquels on peut penser qu'ils seront utilisés dans le système en adéquation avec le sens qui leur est attribué dans l'ontologie régionale.

1.4 Une méthode s'appuyant sur les textes

Issue de l'approche des BCT, notre méthode en reprend les étapes et les techniques linguistiques. Le modèle conceptuel est d'ailleurs organisé comme une BCT, avec une composante linguistique pour conserver termes et textes associés au réseau conceptuel. A la différence des BCT, ce modèle doit être directement adapté à l'application et utilisable. Or, du point de vue méthodologique, l'approche purement linguistique ne propose pas suffisamment de critères pour conduire l'analyse des corpus, pour réduire la quantité d'informations à exploiter et pour décider de la façon de les représenter dans le modèle cible.

De tels critères sont énoncés dans la démarche différentielle, dont nous reprenons ainsi le principe de normalisation [8]. D'autres sont relatifs à la tâche, qui est donc prise en compte pour filtrer les informations à chercher dans les textes, celles à retenir dans le modèle puis pour orienter leur structuration. Comme dans CommonKADS, il est possible, voir recommandé, de mener l'analyse des textes en fonction d'une MRP et des caractéristiques de la tâche de l'application visée.

Par contre, comme dans KOD, les textes sont exploités systématiquement selon des résultats et des techniques linguistiques. La recherche de définitions de concepts s'appuie par exemple des marqueurs linguistiques de relations conceptuelles, Nous préconisons aussi l'utilisation systématique d'outils logiciel automatisant l'application de ces techniques, ces outils étant actuellement disponibles, alors qu'ils n'existaient pas lorsque KOD a été définie (1988). A la différence de KOD, nous sélectionnons si possible des textes techniques contenant des connaissances stabilisées qui décrivent à la fois le domaine et l'application.

Comme dans l'approche ontologique, nous proposons de découper le domaine en grands sous domaines, et de réutiliser les ressources existantes, ontologies de modèle (primitives de modélisation) ou ontologies de domaine, glossaires, index … pour amorcer la structuration. Notre approche va également jusqu'à un modèle formel, pour vérifier la cohérence et la correction de la structuration. Par contre, notre priorité est de fournir un modèle le plus pertinent possible pour l'application et donc complètement ad hoc le plus souvent. Sa réutilisation requiert des modifications non triviales.

Pour terminer cette caractérisation, soulignons que notre position revient à tirer le meilleur profit des textes tout en ne les considérant pas comme la seule source de connaissances. Suivant les applications, le rôle des experts peut aller de la constitution du corpus et de l'aide à la validation du modèle jusqu'à un rôle primordial pour fournir des connaissances que lui seul détient.

3 Notre méthode

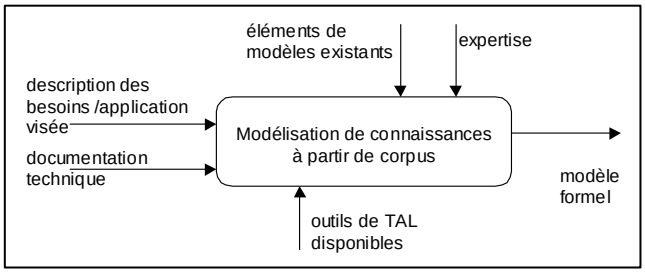

Fig. 2 : *Vue globale de la méthode*

La méthode présentée est assez générale. Les choix méthodologiques et techniques restent ouverts. Ils dépendent de différents facteurs :

- la description des besoins relatifs à l'application ;
- la documentation technique à disposition ;
- les éléments de modèles déjà existant (glossaires, terminologies, ontologies, etc.) ;
- l'expertise disponible ;
- les outils d'analyse linguistique disponibles.

La méthode est mise en œuvre par un ou plusieurs analystes ayant plus ou moins de compétences en linguistique, en modélisation et en formalisation, que nous appellerons "le cogniticien" par la suite. Celui-ci décide à chaque étape de la méthode, des techniques qu'il va utiliser en fonction des facteurs précédents et de ses propres compétences. Il est clair que l'application pratique de la méthode ne peut s'imaginer sans un logiciel adéquat qui permette à la fois de gérer la grande masse d'informations (termes, concepts et relations), de les décrire, de les organiser puis de les représenter formellement. Ce type d'environnement doit permettre de confronter facilement les termes et relations lexicales, le texte d'où ils sont tirés et le modèle dans lequel on va les intégrer.

Nous avons déjà développé les outils Terminae [7] et Géditerm [3] dans le même esprit. Terminae permet de consulter un corpus, d'intégrer les résultats de l'extracteur de candidats-termes Lexter pour en retenir un certain nombre de termes. Les termes peuvent être polysémiques. Le cogniticien associe à chaque sens d'un terme, une notion. Ensuite, ces notions sont structurées et différenciées, puis formalisées sous forme de concepts étiquetés par le terme, dans un langage proche d'une logique de descriptions. Les liens entre les termes et leurs occurrences dans le corpus, la notion et le concept formel résultant sont sauvegardés. Géditerm insiste davantage sur les phases initiales de repérage des termes et de l'association terme/concept justifiée par les occurrences des termes en corpus. Géditerm ne distingue pas notion et concept. Un concept dans Géditerm correspond

à une notion dans Terminae. Comme Terminae, Géditerm accepte en entrée des listes de candidats termes produites par Lexter. Géditerm ne permet pas de formaliser le réseau conceptuel obtenu mais permet mieux de gérer la structuration en amont de la formalisation et sa visualisation graphique.

Pour présenter ce cadre méthodologique, nous précisons dans une première partie la nature des données utilisées et produites tout au long du processus, c'est-à-dire lors du passage du corpus à un modèle du domaine. Ensuite, nous détaillons les étapes de la démarche.

1.5 Des textes à un modèle formel

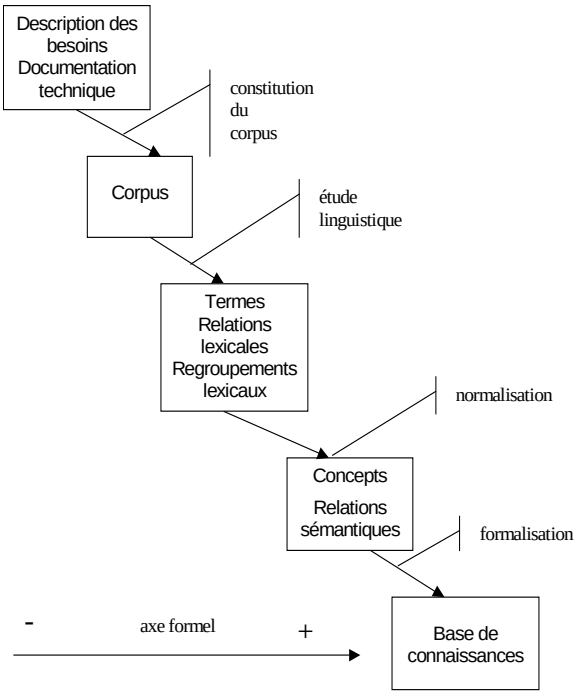

Fig. 3 : *Les différents types de données lors du passage de textes à un modèle du domaine*

Nous présentons d'abord ici une vision axée sur les objets mis en jeu (Fig. 3). La méthode part des textes constituant la documentation technique pour aboutir à une modélisation formelle du domaine. Elle distingue les termes des concepts et les relations lexicales des relations sémantiques. Les termes et les relations lexicales correspondent à des syntagmes présents dans le corpus et considérés comme caractéristiques du domaine. Les regroupements lexicaux rassemblent des syntagmes apparaissant dans des contextes analogues. Les syntagmes sont interprétés en contexte local (la phrase ou le paragraphe) puis global (le texte ou le corpus). Lorsqu'ils sont attestés, ils donnent lieu à la création de concepts et relations sémantiques, dont ils sont les étiquettes. L'ensemble des concepts et relations forme un réseau sémantique, non formel mais compréhensible par le concepteur. Les concepts et relations étant extraits du corpus et contraints par l'application, ce réseau forme une ontologie régionale au sens de [8].

Dans le modèle formel, concepts et relations sont formalisés dans un langage terminologique assimilable à une logique de descriptions, sous forme de concepts et de rôles organisés en une hiérarchie d'héritage. Les concepts sont caractérisés selon deux dimensions, l'une linguistique exprimant s'ils correspondent ou non à un syntagme du corpus, l'autre de structuration indiquant la motivation ayant conduit à intégrer ce concept dans le modèle formel. Ces informations facilitent la maintenance et la compréhension du modèle et de la base de connaissances [6].

1.6 Description détaillée des étapes

Le processus de modélisation est détaillé ici, de la constitution du corpus à celle du modèle formel (Fig. 4).

1.6.1 Constitution du corpus

A partir de la description des besoins expliquant quels sont les objectifs de développement du modèle, le cogniticien choisit dans la documentation technique à sa disposition les textes à inclure dans le corpus. Il peut s'agir de textes didactiques, de spécifications techniques, de normes, de compte-rendu d'expériences, d'articles scientifiques… Le corpus doit couvrir complètement le domaine requis par l'application. Le choix nécessite une expertise des textes du domaine afin de caractériser leur type et la couverture du domaine. Un glossaire sur le domaine est utile pour déterminer les sous-domaines à explorer et vérifier qu'ils sont tous couverts. Le corpus est ensuite mis sur support informatique s'il ne l'était pas. Le début de la modélisation peut conduire à revoir le contenu du corpus.

1.6.2 Utilisation d'outils de TAL

L'étude linguistique est menée à l'aide d'outils de TAL avec l'objectif de déterminer les termes et les relations lexicales qui seront éventuellement modélisés. Nous différentions les outils dédiés à l'acquisition de connaissances terminologiques comme les extracteurs terminologiques, de ceux spécialisés dans la modélisation conceptuelle ou les outils linguistiques classiques.

Les extracteurs de candidats-termes fournissent un grand nombre de données et nécessitent une sélection des termes longue et fastidieuse qui requiert une bonne expertise du domaine. Cette solution est à retenir si l'on dispose d'une main d'œuvre disponible et compétente dans le domaine. Les extracteurs reposent sur des principes d'analyse statistique Ana [17], Startex [25], syntaxique Lexter[9], Nomino [16], ou mixte. Ils sont généralement associés à un environnement d'analyse plus complet, pour la consultation et la validation de leurs résultats. Leur application nécessite peu de compétence linguistique.

Les extracteurs de relations à partir de marqueurs linguistiques Startex, Prométhé [22], Caméléon [26], nécessitent de définir d'abord des marqueurs de relation pour ensuite les appliquer au corpus et ramener les termes en relation. En partant de marqueurs prédéfinis, il est possible de déterminer des termes que l'on réutilise pour créer de nouveaux marqueurs et définir ainsi des relations lexicales et des termes. Ces outils sont très séduisants mais nécessitent de bonnes compétences linguistiques. Ils sont complémentaires des extracteurs

de candidats termes, en contribuant à la sélection et à la définition des termes.

Les outils de regroupement conceptuel associent les syntagmes nominaux qui partagent des relations de dépendances syntaxiques. Les regroupements proposés peuvent ensuite être analysés manuellement pour constituer des classes sémantiques comme avec Zellig [20] ou Lexiclass [1]. Les résultats sont difficilement interprétables, mais ils permettent de structurer les termes pour définir les concepts. Plus original mais ponctuellement utile pour traiter des instances de certaines classes, Lexis [23] permet de repérer des noms propres dans un corpus.

Les concordanciers (outil d'alignement sur un mot donné ou un groupe de mots) Sato [15] sont utilisables lorsque l'on dispose déjà des termes, pour les étudier en contexte en facilitant leur visualisation. Ils sont d'utilisation simple.

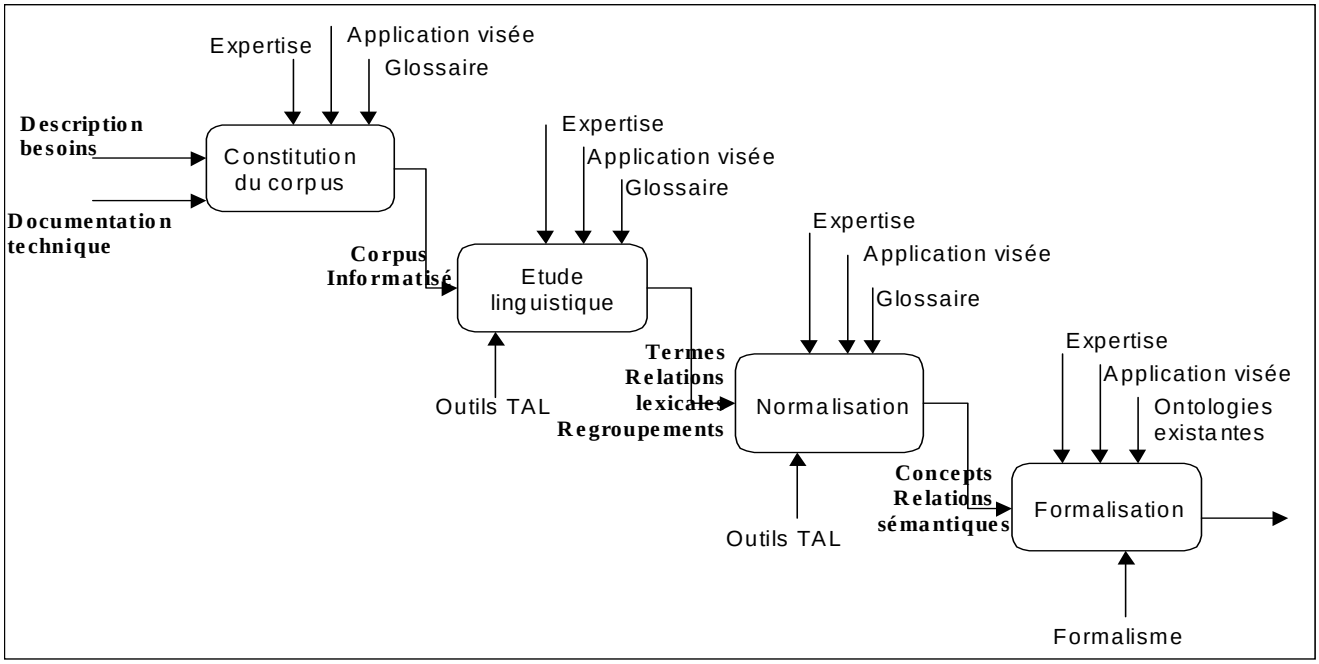

Fig. 4 : *Etapes du processus de modélisation à partir de textes selon notre approche*

1.6.3 Normalisation

La normalisation est un processus particulier de conceptualisation fondé sur l'analyse de corpus en suivant [24] et [8]. La normalisation consiste en deux parties : la première reste dans le domaine du traitement lexical et exploite les données retenues par l'étape antérieure ; la seconde partie porte sur l'interprétation sémantique et la structuration des concepts et des relations sémantiques. Au cours de la normalisation, la masse de données à considérer est peu à peu restreinte.

Les 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 terme et chaque 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, le cogniticien doit choisir 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. Puis, il étudie chaque syntagme d'après ses contextes d'occurrences afin d'en donner une définition en langage naturel non contraint, qui rende compte du contenu des textes.

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. Cette description normalisée n'a pas été obtenue en utilisant le paradigme différentiel comme dans [8]. L'interprétation de la description est contrainte par le corpus dont elle est issue et l'application.

Ces descriptions amorcent une structuration du domaine sous forme de réseau sémantique. Elles restent toutefois semi-formelles, au sens où seule la rigueur du travail garantit la cohérence du modèle.

1.6.4 Formalisation

La formalisation comprend l'élaboration et la validation de la base de connaissances. Des ontologies existantes, générales ou proches du domaine, ou même un glossaire, peuvent permettre de définir les couches hautes de la base de connaissances en larges sous-domaines. Ensuite, les concepts et relations sémantiques provenant de l'étude linguistique doivent être traduits en

concepts et rôles dans le langage de la base de connaissances, puis il faut les insérer dans le modèle. Cette insertion des concepts et rôles terminologiques induit parfois une remise en question de la structure existante, car elle doit prendre en compte la correction de l'héritage des caractéristiques (rôles) des concepts. Il est souvent nécessaire ou utile de rajouter des concepts pour améliorer la structuration de la base. Lors de l'insertion d'un nouveau concept, une vérification locale est effectuée, qui garantit la correction syntaxique de la description ajoutée. Une validation complète du modèle doit être réalisée lorsque la base atteint un état stable, pour vérifier la cohérence du modèle.

4 Application sur un exemple : les outils de l'IC

Dans cet article, nous nous restreindrons à la mise en œuvre de la méthode présentée ci-dessus pour modéliser le sous-domaine des outils de l'Ingénierie des Connaissances mentionnés dans un corpus. L'objectif de cette ontologie est de permettre à des chercheurs du domaine de décrire leurs propres outils et de les comparer à ceux qui existent déjà.

Cet objectif s'intègre dans un projet de plus grande envergure que nous décrivons ci-dessous, qui définit l'application et le corpus. Puis, nous détaillons les différentes phases de la méthode pour identifier et structurer les concepts du sous-domaine des outils de l'Ingénierie des Connaissances. Nous terminons par une présentation des premiers résultats.

1.7 Contexte de l'expérience

1.7.1 Le projet global

Le groupe TIA a voulu confronter ses propositions théoriques, techniques et méthodologiques dans le cadre d'un projet. L'objectif de ce projet est l'élaboration d'un thesaurus en français du domaine de l'ingénierie des connaissances pour permettre d'indexer les pages Web des chercheurs. Le parti a été pris d'appliquer directement les méthodes et les outils du groupe sans prendre en compte au départ d'autres ontologies sur ce domaine, comme celle du projet (KA^2) développée pour un objectif analogue mais en langue anglaise pour la communauté internationale [5]. Cependant, il existe des différences importantes entre ce projet global et $(KA)^2$:

- Le domaine couvert est plus large que les activités des chercheurs, leurs productions et organisation.
- Les sources de la connaissances sont principalement des textes et quelques experts du domaine tandis que pour (KA)2 seuls des experts du domaine ont défini et structuré l'ontologie.
- La méthode utilisée est fondée sur l'analyse de corpus et non sur l'introspection.
- Le résultat final est un thésaurus et non une ontologie considérée comme formelle.

La validation de l'approche se fera par l'utilisabilité des résultats, sachant que pour le moment il existe peu de pages Web de chercheurs français dans ce domaine. Cette expérimentation permettra d'affiner la nature de la tâche d'indexation et d'améliorer le thésaurus obtenu.

1.7.2 Notre expérience

L'expérience décrite dans ce qui suit, peut être considérée comme un produit dérivé du projet global. Nous avons décidé de construire une ontologie sans connaître son impact sur la création du thésaurus. Nous pensons qu'une ontologie des outils de l'IC peut être utile pour les chercheurs de ce domaine. Cette ontologie permettra de situer des nouveaux outils par rapport à ceux qui existent déjà. Le domaine est suffisamment restreint pour espérer obtenir un résultat dans un temps raisonnable et ainsi évaluer la méthode.

1.7.3 Le corpus

Pour constituer un corpus de référence, le groupe TIA n'a pas retenu les pages Web existantes, du fait de leur nombre faible et de leur contenu très spécifique à une personne ou une équipe. Le corpus choisi, décrit dans l'article [10] comporte 34 articles scientifiques publiés au cours des trois dernières années à la conférence annuelle de IC et rassemblés dans un ouvrage de synthèse [12]. Ce sont des articles techniques destinés à des chercheurs du domaine et donc assez peu didactiques. Une première étude des termes issus de ce corpus a montré qu'il ne couvrait pas très bien la totalité du domaine et qu'il contenait peu de définitions de concepts. Pour pallier ces insuffisances, le groupe TIA a rajouté quatre textes décrivant le domaine de façon plus globale, augmentant ainsi de 30% la taille du corpus. Le nombre de mots est passé de 160 000 à 207 000. Nous avons travaillé pour notre expérience sur la totalité de ce corpus.

1.7.4 Les outils utilisés et les cogniticiens

Les outils linguistiques mis en œuvre sont un extracteur de candidats termes (Lexter) et un extracteur de relations (Caméléon). Caméléon se focalise sur l'étude des relations sémantiques identifiées dans des textes pour enrichir incrémentalement un modèle conceptuel [27]. Nous avons utilisé certains modules de Terminae pour valider et visualiser les résultats de Lexter et construire le réseau conceptuel. Terminae possède un gestionnaire de base de connaissances qui permet de décrire un ensemble structuré de concepts. Un classifieur teste la validité de l'insertion d'un concept et informe l'utilisateur de la détection d'incohérences ou de redondances. De plus, les concepts possèdent une caractéristique linguistique. Un concept est dit *terminologique* s'il vient de la liste des termes du corpus, *terminologique non attesté* s'il n'est pas dans le corpus mais est considéré comme un terme du domaine, *non terminologique* s'il est introduit sans lien même indirect avec le domaine. Les concepts sont aussi différenciés selon leur rôle dans la structuration du modèle : un concept est dit de *structuration ascendante* s'il est introduit pour structurer les concepts de plus bas niveaux, *descendante* s'il

spécifie un concept de niveau élevé ; la limite entre les deux voies de structuration est imprécise.

Dans le cas de cette expérience, comme les propriétés sur les concepts ne sont pas définies formellement, les concepts sont tous *primitifs*, et il n'y a pas de *concepts de regroupement* (qui correspondent à un concept créé pour factoriser une propriété commune à un ensemble de concepts).

Les concepteurs du modèle sont eux-mêmes des experts du domaine étudié mais ne sont pas linguistes.

1.8 L'étude linguistique

Deux façons d'amorcer le travail sont possibles. L'une consiste à se focaliser en priorité sur les termes, puis de rechercher les relations qui les associent et leurs occurrences. L'autre privilégie les relations lexicales, comme autant d'indicateurs de contextes riches en connaissances. C'est à partir des occurrences de relations que l'on repère des termes du domaine et que l'on amorce leur définition.

Dans le cadre de notre expérience, nous avons tout d'abord privilégié une démarche centrée sur les termes, puisqu'un terme "outil" est donné dès le départ comme le type d'objets à décrire. L'étude des relations doit compléter ce travail sur les termes.

1.8.1 Démarches pour l'étude des termes

Dans notre cas particulier, les outils logiciels ont presque tous un nom propre. Or, dans la perspective de l'indexation de travaux de recherche sur le Web, il est intéressant de répertorier les différents noms d'outils et de méthodes. Comme les noms propres peuvent être identifiés en tant que tels, nous avions le choix encore entre deux démarches d'analyse des termes, selon la place accordée aux noms propres.

La première approche consiste à rechercher le terme « outil » dans les résultats de Lexter et à considérer tous les termes comportant « outil » en tête Ceux-ci désignent potentiellement des sous-classes du concept OUTIL associé à « outil » . On recherche ensuite les relations lexicales mettant en jeu chacune de ces sous classes avec Caméléon, ce qui permet de regrouper des sous-classes et de les structurer. Enfin, on recherche les noms des outils en tant que noms propres co-occurrents avec un des termes trouvés. Une fois ces noms d'outils obtenus, l'analyse plus précise de leurs contextes d'apparition permet de les classer parmi les différents types d'outils ou bien de constater qu'ils n'apparaissent pas comme spécialisant « outil » mais un autre terme (« système » par exemple) et de devoir considérer alors la relation entre « outil » et ce terme.

La deuxième approche considère d'abord tous les noms propres du corpus, qui désignent des auteurs, des acronymes divers ou des institutions. L'examen de leur contexte permet d'éliminer rapidement tous les termes qui n'ont aucune proximité avec « outil ». Une étude plus fine du contexte permet de trouver un hyperonyme à chacun des noms propres. Les termes «outil », « système », « projet », « algorithme », « méthode » apparaissent. La recherche de relations entre ces termes doit permettre ensuite de les organiser.

Idéalement, les deux méthodes peuvent être croisées, ce qui garantit que tous les termes relevant d'« outil » présents dans le corpus ont bien été considérés.

1.8.2 Validation de candidats termes

C'est la première approche qui a été appliquée au corpus dans le cadre de notre expérience. 109 occurrences de candidats termes comportant « outil » en tête ont été trouvées par Lexter. L'analyse des contextes de ces occurrences a mené à en éliminer un certain nombre comme n'étant pas pertinents : les termes désignant des outils en dehors du domaine comme « outil de préformage de la semelle » qui relève d'une application spécifique ; « outil de support du processus expérimental » ou « outil de travail » sont trop généraux. Simultanément, l'étude des occurrences permet de filtrer les occurrences pertinentes et de rejeter celles qui n'apportent aucune connaissance utile à la définition de termes. Par exemple, les termes « outil de GL du projet » et « outil de génie logiciel du projet » sont des exemples de synonymie syntaxique en contexte de « outil de génie logiciel ». Seul, ce dernier terme et ses occurrences sont pertinents pour notre application. Ainsi, 67 termes ont été retenus.

1.8.3 Démarches pour l'étude des relations

Beaucoup de relations peuvent être facilement identifiées par la lecture des occurrences des termes. Cependant, cette méthode est longue. Le nombre d'occurrences à lire a été réduit en utilisant Caméléon [26]. La méthode associée à Caméléon préconise tout d'abord de mettre au point une base de marqueurs adaptés au corpus et auxquels on a associé le type de relation sémantique qu'ils permettent de repérer. Pour aider dans cette tâche, Caméléon fournit comme point de départ une base de marqueurs génériques associés aux relations classiques d'hypéronymie et de méronymie. Peu de variations par rapport aux cas généraux ont été trouvées. Pour chaque concept dans le modèle, nous avons appliqué les hypothèses de Caméléon sur les occurrences des termes correspondants. Caméléon a trouvé de 2 à 10 relations pour les concepts sous OUTIL, et environ 50 relations pour OUTIL, le tiers de ces relations a été validé. Une utilisation ultérieure de Caméléon devrait permettre d'identifier de nouveaux patrons de relation en projetant les couples de concepts en relation sur le corpus.

1.9 La normalisation

A ce stade, nous avons une liste de termes à partir de laquelle nous devons créer un ensemble structuré de concepts. La normalisation va consister à éliminer des termes et structurer hiérarchiquement ceux qui restent, en particulier grâce à leurs relations.

1.9.1 Elimination et regroupement

La validation précédente, réalisée rapidement, nécessite peu d'expertise. L'étape suivante, menée avec une expertise plus fine, nécessite un retour au texte avec une réflexion approfondie pour éliminer ou regrouper des termes. Cela permet de réduire le nombre de termes et de commencer à structurer leur ensemble. La réduction consiste, soit à éliminer simplement des termes, soit à regrouper sous la même étiquette des termes jugés synonymes dans le corpus, soit à regrouper sous la même étiquette des termes que l'on ne veut pas distinguer pour l'application. La distinction entre termes synonymes et termes assimilés est consignée.

Ainsi, des termes très spécifiques comme « outil anthropotechnique » ou « outil de déploiement » vont être assimilés respectivement à « outil de génie cognitif » et « outil de génie logiciel », car ce degré de granularité est considéré comme trop fin, l'objectif étant de décrire les outils d'IC. Par contre, « outil de cartographie » va être simplement supprimé.

Des termes sont synonymes lorsque des auteurs utilisent des termes différents pour des notions identiques. Nous avons été capables de décider de ces relations de synonymie car nous sommes à la fois cogniticiens et compétents dans le domaine. Dans le cas où le cogniticien découvre le domaine, il ne peut s'appuyer que sur des critères linguistiques et des similarités d'usage de ces termes dans leurs différentes occurrences. Des compétences linguistiques sont ici très utiles pour juger de cette synonymie en contexte. Ainsi l'analyse des contextes d'occurrences d' « outil textuel », « outil d'analyse de textes », « outil d'analyse de corpus », « outil linguistique » permet d'identifier les trois termes, sous l'étiquette « outil d'analyse de corpus ». Il faut cependant faire attention à ne pas assimiler des termes généraux et des termes plus précis qu'il est important de différencier pour l'application. Par exemple, « outil terminologique » est plus spécifique que « outil d'analyse de corpus » et les deux seront retenus. Par contre, « outil d'exploration de corpus » comme « outil de fouille de corpus » seront ici assimilés à « outil d'analyse de corpus ». Identifier « outil d'extraction de candidats termes » et « outil d'extraction de terminologie » nécessite un retour au texte et conduit à définir soigneusement ce qui les différencie : l'extraction de terminologie comporte à la fois l'extraction de candidats termes et celle de relations candidates. En même temps, les relations d'hyperonymie apparaissent et sont soigneusement enregistrées.

Cette étape réduit l'ensemble des 67 concepts à 46.

1.9.2 Noms propres et noms proches

Au fur et à mesure de l'analyse des candidats termes, le cogniticien recueille d'une part les noms propres qui sont des noms d'outils ou de systèmes (30 ont été relevés), et d'autre part des termes dont le sens apparaît proche de « outil » (18 termes, comme «système», « projet »,«atelier », «collecticiel », « concordancier ». Ces derniers correspondent soit à des synonymes soit à des termes proches qui pourraient mener à d'autres noms d'outils et intervenir dans la structuration. Les termes recueillis sont organisés sous forme de listes. La liste des outils existants (noms propres) comporte l'indication du type d'outil générique dont ils relèvent et de leur créateur, lorsque l'information est présente dans le texte ou connue du cogniticien.

1.9.3 Structuration

Comme nous nous intéressons aux outils de l'IC, nous avons défini le concept OUTIL, puis nous avons structuré les concepts autour d'OUTIL.

Pour définir le concept OUTIL, nous avons étudié tous les termes proches d'outil. Le choix de ces termes a été guidé par notre expertise et certains noms propres d'outils. Nous avons étudié les relations de ces termes avec le terme « outil » dans le corpus. Ainsi, nous avons étudié les relations des termes « méthode », « algorithme », « formalisme », « système », « atelier » avec le terme « outil ». De l'étude des relations, nous avons distingué les outils conceptuels (« méthode », « algorithme », « formalisme », « modèle ») des outils logiciels (« outil »). En utilisant notre expertise et les occurrences de certains outils, nous distinguons deux sortes de logiciels : ceux qui sont développés selon un processus d'ingénierie et ceux utilisés dans ce processus, qui peuvent être soit des outils d'ingénierie des connaissances (OUTIL D'INGENIERIE DES CONNAIS-SANCES), soit des outils de génie logiciel (OUTIL DE GENIE LOGICIEL). Comme il y a une grande variété de systèmes développés qui peuvent être considérés comme suivant un processus d'ingénierie des connaissances, nous ne les avons pas détaillés.

1.10 Premiers résultats

Notre travail en est à ses débuts. Nous avons surtout étudié la relation d'hyperonymie. Nous devons étudier d'autres relations comme la méronymie, la relation « sert-à » ou « utilise » pour continuer la structuration. Pour chaque outil spécifique, nous avons défini son ou ses auteur(s), un commentaire et les occurrences contenant le nom de l'outil dans le corpus. La construction d'une ontologie avec Terminae permet de tester la cohérence du réseau sémantique construit. Nous montrons dans la suite les éléments principaux de la structuration.

Nous avons créé sous OUTIL deux concepts terminologiques OUTILLOGICIEL et OUTILCONCEPTUEL. La figure 5 montre la hiérarchie sous OUTILLOGICIEL et OUTILCONCEPTUEL. Dans toutes les figures montrant une partie de l'ontologie, les concepts terminologiques sont en italiques, les concepts individuels qui correspondent à un outil spécifique commencent par une majuscule. Seulement OUTILINGENIERIECONNAISSANCES et OUTILVALIDATION ne sont pas terminologiques, ils sont terminologiques non attestés (TNA). Le concept terminologique OUTILAIDE regroupe tous les concepts correspondant aux termes composés qui ont « outil d'aide » en tête. Nous avons travaillé sur les outils linguistiques décrits dans le corpus.

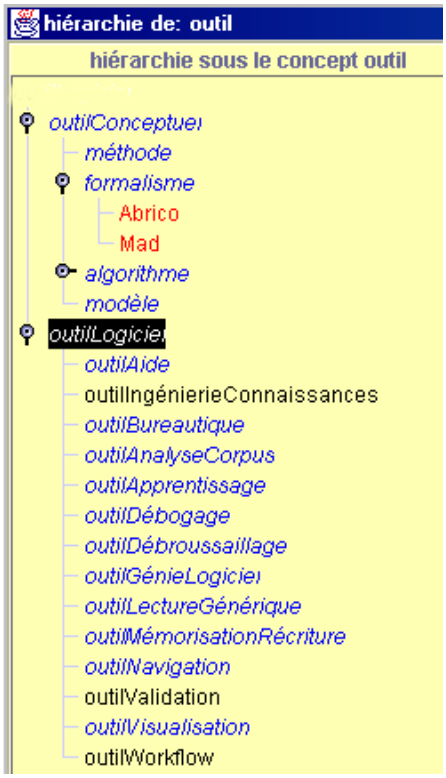

Fig. 5 : *schéma sous* OUTIL

La figure 6 présente les concepts sous OUTILANALYSECORPUS. Nous avons trouvé OUTILEXTRACTIONTERMESCANDIDATS et OUTILREPERAGERELATIONS que l'on a regroupé sous OUTILTERMINOLOGIQUE.

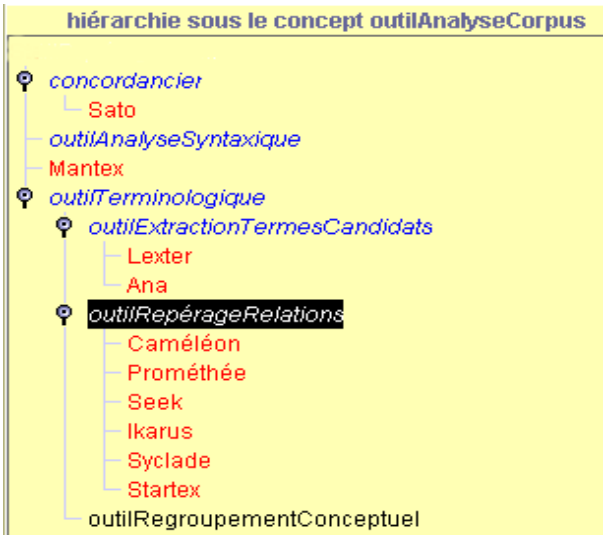

Fig. 6 : *Schéma sous OUTIL D'ANALYSE DE CORPUS*

1.11 Conclusion de l'expérience

Une des conclusions de ce travail est d'avoir mis en évidence l'importance de prendre encompte conjointement plusieurs critères pour retenir les connaissances à modéliser. Quel que soit le type de données (concepts, termes, rôles ou relations) que nous devons sélectionner et quelle que soit la manière de les trouver (lecture de liste ou de textes, manuellement ou en utilisant des outils linguistiques), leur sélection et l'organisation des modèles dépendent de quatre critères :

1. notre expertise sur le domaine
2. l'application visée
3. les textes (les contextes des informations, les auteurs et la nature des textes)
4. d'autres informations apportées par les outils linguistiques : les termes co-occurents, les termes en relations lexicales, la fréquence de certains termes.

Ces critères peuvent produire des contradictions. La priorité est donnée à ce qui paraît pertinent pour l'application compte tenu de l'expertise, ce qui peut être en désaccord avec ce qui est dit dans les textes. Par exemple, l'outil ASTREE est décrit dans le corpus comme un outil d'IC, ce qui manque de précision lorsqu'on veut créer une ontologie qui caractérise les outils de l'IC. L'étude approfondie du corpus montre qu'ASTREE est un outil d'aide à la création de modèles conceptuels. Aussi, nous avons décidé de définir le concept OUTIL D'AIDE A LA MODELISATION et ASTREE comme un concept fils.

Nos conclusions comportent également des résultats sur la manière d'utiliser les principes et logiciels que nous avons retenus, les données qu'ils permettent de trouver ainsi que le moment où il vaut mieux les utiliser. L'expérience doit être menée plus loin pour déboucher sur des propositions vraiment précises et bien organisées au sein de notre méthode, et sur une véritable ontologie comme résultat. Enfin, nous envisageons une validation de l'ontologie obtenue par rapport au domaine, qui consisterait à essayer de décrire de nouveaux outils d'IC au sein de cette ontologie.

5 Conclusion

Le travail présenté dans cet article a été initié par les recherches récentes des membres du groupe TIA. Il s'agit pour nous d'établir les étapes, outils et méthodes à appliquer pour dégager un modèle conceptuel du domaine à partir de l'analyse d'un corpus, en utilisant des outils de traitement automatique des langues. Ce projet n'en est qu'à ses débuts. L'objectif de réaliser un thésaurus en langue française dans le domaine de l'ingénierie des connaissances, à partir d'un corpus et en s'appuyant sur des techniques linguistiques pour mener la modélisation, est d'une importance au moins équivalente au projet européen KA². Il s'apparente au projet EuroKnowledge [29] consistant à inventorier la terminologie anglaise de la modélisation au niveau connaissance au sein d'un ouvrage didactique de référence.

Pour expérimenter et affiner notre méthode, nous nous sommes donné comme objectif de construire une ontologie du sous-domaine des outils de l'Ingénierie des Connaissances mentionnes dans le corpus. Ce travail de modélisation est particulièrement difficile car il porte sur un domaine de recherche dont les termes sont évidemment en constante évolution. Le vocabulaire n'est donc pas figé, chaque auteur d'article usant d'un voca-

bulaire spécifique dont il est lui-même expert. Le nombre de termes qui ne sont présents que dans un seul document est très important. Cette modélisation nécessiterait donc d'être négociée avec la communauté, dans un but d'éclaircissement des concepts en jeu. Il s'agit bien d'une volonté descriptive et non pas normative.

Cependant, la mise en œuvre de cette méthode sur une application est une illustration passionnante de tout son potentiel mais aussi de toutes les questions pratiques, méthodologiques et même théoriques qu'il reste à traiter. Il est clair que nous sommes loin d'avoir exploité au mieux la complémentarité des différents types d'analyse possibles du corpus. Nous n'avons pas pu utiliser toute la gamme d'outils qui existent aujourd'hui à notre disposition. Or les résultats obtenus rapidement sont déjà de bonne qualité et prometteurs. Ils seront complétés et évalués lors d'une prochaine étape du projet.

Références

[1] ASSADI H., Construction d'ontologies à partir de textes techniques : Application aux systèmes documentaires. Thèse de l'Université Paris 6. 1998.

[2] AUSSENAC-GILLES N. et CONDAMINES A., Bases de connaissances terminologiques : enjeux pour la consultation documentaire, J.Maniez et W.Mustapha El Hadi (eds), *Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information*, Villeneuve d'Asq : Univ. Charles de Gaulle, pp. 71-88. 1999.

[3] AUSSENAC-GILLES N., GEDITERM, un logiciel de gestion de bases de connaissances terminologiques, *in* Actes des Journées Terminologie et Intelligence Artificielle (TIA'99), Nantes, *Terminologies Nouvelles* n°19, pp 111-123. 1999

[4] AUSSENAC-Gilles N., BOURIGAULT D., CONDAMINES A. et GROS C., How can knowledge acquisition benefit from terminology ? *Proc. of the 9th Knowledge Acquisition for Knowledge Based Systems Workshop*, Banff (CAN), 1995.

[5] BENJAMINS R., FENSEL D., DECKER D. et GOMEZ PEREZ A., (KA)² : building ontologies for the internet : a mid-term report. In *Proc. of the international workshop on ontological engineering on the global information infrastructure* (EKAW'99). pp 1-24, 1999.

[6] BIEBOW B., SZULMAN S., TERMINAE : A linguistic-based tool for the building of a domain ontology, *11th European Workshop, Knowledge Acquisition, Modeling and Management (EKAW 99)*, Dagstuhl Castle, Germany, pp 49-66. 1999.

[7] BIEBOW B., SZULMAN S., Terminae : une approche terminologique pour la construction d'ontologies du domaine à partir de textes. *Actes de RFIA2000, Reconnaissances des Formes et Intelligence Artificielle*, Paris (F), 2000.

[8] BACHIMONT B. : Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In *: Ingénierie des Connaissances, évolutions récentes et nouveaux défis*. Paris:Eyrolles, 2000.

[9] BOURIGAULT D. : Lexter, un Logiciel d'Extraction de TERminologie . Application à l'extraction des connaissances à partir de textes, Thèse en Mathématiques, Informatique appliquée aux sciences de l'homme. EHESS, Paris, 1994.

[10] BOURIGAULT D., CHARLET J., Construction d'un index thématique de l'Ingénierie des Connaissances. *Actes de Ingénierie des Connaissances IC'99 (Paris)*, 107-118, 1999.

[11] CHARLET J., REYNAUD C. et TEULIER R. Ingénierie des connaissances pour les systèmes d'information. *Conception des Systèmes d'Information*, ed. C. Cauvet, Traité IC2, Paris: Hermès. 2000.

[12] CHARLET J., ZACKLAD M., KASSEL G. et BOURIGAULT D. (eds.) *Ingénierie des Connaissances, évolutions récentes et nouveaux défis*. Paris : Eyrolles, 2000.

[13] CONDAMINES A. et AUSSENAC-GILLES N., Entre textes et ontologies formelles : les bases de connaissances terminologiques. I*n Capitalisation des connaissances*. Zacklad M. Grundstein M. (Eds.). Paris : Hermès. Traités IC2. 2000.

[14] CONDAMINES A., REBEYROLLE J., Construction d'une base de connaissances terminologiques à partir de textes: expérimentation et définition d'une méthode. in J. CHARLET, M. ZACKLAD, G. KASSEL, D. BOURIGAULT (eds.) *: Ingénierie des Connaissances, évolutions récentes et nouveaux défis*. Paris:Eyrolles, 2000.

[15] DAOUST F., *SATO (Système d'Analyse de Textes par Ordinateur)* version 3.6, Manuel de référence, Centre ATO Université du Québec à Montréal, 1992.

[16] DAVID S. et PLANTE P., *Termino version 1.0*, Rapport du Centre d'Analyse de Textes par Ordinateur. Université du Québec à Montréal, 1990.

[17] ENGUEHARD C. et PANTERA L., Automatic natural acquisition of terminology *Journal of Quantitative Linguistics*, vol.2, n°1, pp 27-32, 1995.

[18] FRIDMAN NOY N. et HAFNER C. The state of the Art in Ontology Design : a Survey and Comparative Review. *Artificial Intelligence Magazine*. pp 53-74, Fall 1997.

[19] GOMEZ-PEREZ A., "Développements récents en matières de conception, de maintenance et d'utilisation des ontologies". in ENGUEHARD C. et CONDAMINES A. (Eds.) : *actes des 3es Rencontres "Terminologie et intelligence artificielle"* (Nantes), dans *Terminologies Nouvelles*. (19). Bruxelles. pp 9-20. 1999

[20] HABERT B., NAULLEAU E., NAZARENKO A. Symbolic word clustering for medium-size corpora . *16th International Conference on Computational Linguistics*, Copenhagen, Danemark, pp 490-495, 1996.

[21] MEYER I., SKUCE D., BOWKER L. et ECK K. : Towards a new generation of terminological ressources : an experiment in building a terminological knowledge base. *Proceedings COLING'92*, Nantes, pp 956-960, 1992.

[22] MORIN E., Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique, *TAL* (*Traitement Automatique des Langues*), vol.40, n°1, Paris : Université Paris VII, pp 143-166, 1999.

[23] POIBEAU T., Repérage des entités nommées : un enjeu pour les système de veille, in *Actes de TIA'99 (Terminologie et Intelligence Artificielle)*, Nantes, *Terminologies Nouvelles* n°19, pp 43-51, 1999.

[24] RASTIER F., Le terme : entre ontologie et linguistique. *La banque des mots*, n° spécial 7/95, pp 35-65, Paris : CLIF, 1995.

[25] ROUSSELOT F., FRATH P. et OUESLATI R., Extracting Concepts and relations from corpora. *Proceedings ECAI'96, 12th European Conference on Artificial Intelligence*, 1996.

[26] SEGUELA P., Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus

spécialisés,in *Actes de TIA'99 (Terminologie et Intelligence Artificielle)*, Nantes, *Terminologies Nouvelles* n°19, pp 52-60, 1999.

[27] SEGUELA P. et AUSSENAC-GILLES N., Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine, *Actes de IC'99 (Ingénierie des Connaissances)*, pp 79-88, Paris, 1999.

[28] SLODZIAN M., Comment revisiter la doctrine terminologique aujourd'hui ? *La banque des mots*, n° spécial 7/95, pp 11-18, Paris : CLIF, 1995.

[29] USCHOLD M., Knowledge level Modelling : concepts and terminology. *The knowledge engineering review*, Vol. 13:1,pp 5-29, 1998.

[30] VOGEL C., *Génie cognitif*. Paris : Masson, 1988.