Samovar ou comment capitaliser les traces de la coopération entre acteurs d'un projet-véhicule avec une mémoire de projet

Utilisation des ontologies pour la recherche d'information dans le domaine de l'automobile

Joanna Golebiowska* — Rose Dieng-Kuntz* — Olivier Corby*
Didier Mousseau**

** Inria, Projet Acacia 2004, Route des Lucioles BP 93, F-06902 Sophia Antipolis*

{Joanna.Golebiowska, Rose.Dieng,Olivier.Corby}@sophia.inria.fr

*** Renault, TPZ D12 138, sce 12113, 860, quai de Stalingrad, F-92109 Boulogne*

RÉSUMÉ. *Dans cet article, nous présentons Samovar (système d'analyse et de modélisation des validations des automobiles renault), un outil de capitalisation de connaissances dans le domaine de l'automobile. Basé sur la construction et l'exploitation d'une mémoire de projet (en particulier la mémoire des problèmes rencontrés au cours d'un projet), Samovar repose sur un ensemble d'ontologies qui structurent les connaissances et guident la recherche d'informations dans la mémoire de projet par le moteur de recherche Corese. Ces ontologies ont été construites en exploitant un outil de traitement linguistique sur un corpus textuel et en définissant des règles heuristiques sur les candidats termes obtenus grâce à cet outil. Nous généralisons notre approche à la construction d'une mémoire de projet de conception d'un système complexe quelconque.*

ABSTRACT. *This paper describes Samovar (Systems Analysis of Modelling and Validation of Renault Automobiles), aiming at preserving the memory of the problems encountered during a project in automobile design so as to exploit them in new projects. Samovar relies on (1) the semi-automatic building of ontologies by using a linguistic tool on a textual corpus, (2) the «semantic» annotations of the problem descriptions relatively to these ontologies, (3) the formalisation of the ontologies and annotations in RDF(S) so as to integrate in Samovar the tool CORESE that enables an ontology-guided search in the base of problem descriptions.*

MOTS-CLÉS : *mémoire de projet, ontologie,* Text-mining*, outils linguistiques, TALN, annotations sémantiques, recherche d'information, base de données, RDF(S).*

KEYWORDS : *Project memory, Ontology, Text-mining, Linguistic Tool, Natural Language Processing, Semantic Annotation, Information retrieval, Database, RDF(S).*

Document numérique. Volume 5 – n° 3-4/2001, pages 173 à 192

1. Introduction

Le domaine de Samovar est le processus des validations des prototypes au cours d'un projet-véhicule. Ce processus complexe soulève un grand nombre de problèmes constituant un frein au raccourcissement du cycle, du fait des validations à refaire. Il en résulte une augmentation des délais et des coûts. L'observation des validations amène à croire qu'une partie des échecs est due à la perte d'informations et d'expériences produites. Samovar vise à améliorer l'exploitation de ces informations et de les rendre disponibles pour les projets à venir.

1.1. *Contexte*

Le cycle de développement des produits automobiles se décompose en de nombreux sous-cycles itératifs (conception/réalisation/validations, de courte et/ou longue durée). L'ensemble de ces cycles est ponctué par des jalons et vagues prototypes qui marquent les sorties des versions successives des maquettes et prototypes. Au cours d'un projet-véhicule, sont effectuées des validations pendant lesquelles les services d'essai vérifient la conformité d'une pièce ou d'une fonctionnalité par rapport aux exigences précisées dans le cahier des charges : *i.e.* finesse du grain de la planche de bord, bruit que fait une portière en se fermant, comportement de la voiture roulant sur des petits pavés, sa résistance à des hautes ou basses températures… Ces validations sont réparties tout au long du projet-véhicule et effectuées successivement par des services d'essai, en commençant par les fonctionnalités les plus élémentaires pour aboutir aux tests de synthèse. Le projet commence par des tests avec les services d'essais liés aux bureaux d'étude en rapport avec les pièces validées, et se finit par des tests de performance, de vitesse et de crash.

Les validations permettent de relever des problèmes de non-conformité. Ces problèmes sont documentés dans un système de gestion unique – système de gestion des problèmes (noté SGPb), depuis leur détection jusqu'à leur résolution. Ce système inclut une base de données (BD), avec les informations nécessaires au processus de gestion des problèmes : informations sur les acteurs intervenants, et surtout descriptions et commentaires sur les problèmes apparus.

1.2. *Problématique abordée*

L'apparition des problèmes engendrant des coûts supplémentaires et allongeant les délais, une solution envisagée vise à exploiter l'information contenue dans la base SGPb, pour l'utiliser non seulement comme un système de gestion de problèmes mais aussi comme un véritable gisement d'informations : ses champs textuels sont extrêmement riches et pas suffisamment exploités. Les acteurs s'y expriment librement, en décrivant les problèmes, les différentes solutions proposées

ou les contraintes pour la mise en œuvre d'une solution. Il existe aussi d'autres sources d'informations, tel un référentiel officiel de l'entreprise ou de nombreuses bases locales des services d'essai : on pourrait recouper leurs informations avec le contenu de la base SGPb. Il s'agira de proposer des moyens pour extraire, structurer et rendre réutilisable par des projets cette grande quantité d'informations. Les projets-véhicules en cours ont ainsi exprimé des besoins liés à la recherche d'informations au cours des validations : *i.e.* similitudes entre les incidents, détection des corrélations avec d'autres incidents, afin de pouvoir s'inspirer des solutions existantes au sein d'un même projet, voire entre des projets différents.

1.2.1. *Dimension coopérative de Samovar*

1.2.1.1. Diversité de l'information et des sources d'information

Il est important de souligner la grande diversité et hétérogénéité des sources d'informations existantes. Le processus des validations des projets-véhicule est complexe. Il met en œuvre une grande quantité d'informations reparties à travers des systèmes d'information divers et hétérogènes. Chacun des systèmes est dédié à un segment déterminé des validations (en amont du cycle, au cours de la vague Pre-Série, etc.), et concerne une population spécifique (bureaux d'étude, le Centre des Prototypes, l'usine, etc.). De plus, il existe de véritables bases locales où les experts stockent leur propre expérience. L'information est donc de nature diverse (documentation, BD, experts) et distribuée dans différents systèmes qui sont gérés par des services (acteurs) différents. Samovar vise à corréler les démarches et les informations indépendantes, localisées dans différents systèmes, éloignées les unes des autres, et dédiées à des tâches particulières, à travers une démarche fédératrice et coopérative orientée *retour d'expérience* dans un objectif de mise à disposition aux acteurs des validations.

Nous nous sommes intéressés en particulier au référentiel contenant toute la nomenclature de l'entreprise, les composants cités dans le SGPb étant souvent identifiés avec les noms officiels du référentiel.

1.2.1.2. Diversité des acteurs

Notre vision des validations est celle d'un système dynamique fondé sur des échanges. Capitaliser un savoir-faire d'acteurs dans un processus d'échanges techniques, concernant essentiellement des problèmes/solutions, revient à identifier et à capturer les échanges entre les acteurs de ce processus. L'identification des processus d'échanges permet d'amorcer le domaine de capitalisation que l'on veut étudier et modéliser. Les échanges fournissent non seulement l'information sur l'aspect dynamique du système (flux d'information ou *workflow*) mais indiquent aussi les populations qui coopèrent et les systèmes support de cette coopération. Un échange est en premier lieu caractérisé par un émetteur et un récepteur. L'émetteur et le récepteur appartiennent à une organisation hiérarchique de l'entreprise – ils sont ainsi définis par le rôle qu'ils jouent dans l'entreprise. Le service auquel chacun

d'eux appartient fournit les premières indications sur le périmètre fonctionnel (ou compositionnel) et le périmètre de prestation.

Une solide analyse du processus des validations est donc indispensable, notamment afin de définir les périmètres techniques et fonctionnels de chacun des acteurs d'entreprise (responsable pièce, service essai, architectes), coopérant au cours des validations, les périmètres représentant les zones de responsabilité, respectivement en termes de pièces et de prestations.

1.2.1.3. Diversité du vocabulaire

La gestion des problèmes dans la base SGPb se déroule selon les phases de *relevé*, *analyse*, *pilotage* et *résolution* d'un problème. Selon l'avancement, différentes acteurs interviennent : service d'essai, animateur, responsable pièce, architecte. Autrement dit, les directions Etudes, Méthodes, Achats, Qualité, Après-vente, etc. De plus, la base SGPb est utilisée sur différents sites géographiques (directions ou usines) : Flins, Douai, Lardy, Aubevoie. Le système se présente donc comme un espace de collaboration et coopération des entités diverses et variées, reparties géographiquement, intervenant à des moments différents et employant souvent un vocabulaire différent. Divers métiers participent au processus des validations : l'acoustique, l'aérodynamique, l'ergonomie, la géométrie, etc. Chaque métier emploie son propre vocabulaire, avec des termes spécifiques pour qualifier les problèmes et les solutions : par exemple dans le cas d'un problème de bruit sur la planche de bord, deux personnes de métiers différents – le métier de *planche de bord* et celui de *l'acoustique* – sont amenées à collaborer. Chacune d'elles possède un vocabulaire spécifique. A travers le dialogue existant entre ces métiers, le SGPb constitue un espace de gestion coopérative entre ces différentes populations.

L'hétérogénéité du vocabulaire constitue un problème pour le cogniticien qui ne possède aucune connaissance du domaine, ni le vocabulaire propre à l'entreprise. On peut parler d'un « technolecte », vocabulaire de l'entreprise, et puis d'un « idio-technolecte », celui qui est propre à l'expert. Mais la variation du vocabulaire d'un service à l'autre est surtout un problème reconnu au sein de l'entreprise même, et elle est à l'origine de nombreux problèmes de communication. Ce vocabulaire peut varier selon le contexte (par le contexte il faut entendre tout ce qui évolue autour de l'activité d'essai). Le vocabulaire peut donc varier en fonction du service d'essai (synthèse, design, montage, responsable pièce, architecte), de l'endroit où il est effectué (plateau, Centre de réalisation des prototypes, usine, service d'essai), de l'objet de l'essai (maquette, prototype, banc d'essai, l'ensemble des prestations), des critères d'essai (prestations fonctionnelles objectives/subjectives, techniques), de l'envergure de l'essai (plateau – sur un prototype ou sur un sous-ensemble des pièces) ou du type de problème.

1.2.2. *Dimension TALN de Samovar*

La diversité et la richesse du vocabulaire va conditionner les techniques et outils à employer. De plus, si certaines informations sont assez simples à extraire du SGPb, il n'en est pas de même pour les données textuelles de SGPb. Comme il a été mentionné précédemment, il est fréquent qu'un même terme dénotant une pièce (existant dans le référentiel officiel), ou un problème, aient des appellations différentes selon le service, voire selon l'avancement dans le projet. Par exemple, on peut observer qu'un terme peut être utilisé à la place d'un autre dans le cas d'une variation terminologique, ou encore, dans le cas où il existe une relation sémantique entre les concepts qu'il désigne (méronymie, proximité géographique, etc.). Ainsi un *Climatiseur* peut avoir diverses variantes sémantiques méronymiques ou holonymiques : *conditionnement d'air*, *fonction chauffage*, *commande de répartition de chauffage*, *commande de répartition de ventilation*, *boîtier bloc chauffage*, *capteur de température raccord ventilation*, etc. L'objectif linguistique sera donc de détecter un terme sémantiquement intéressant, de le classer par rapport au processus des validations, et d'y attacher toutes les variations rencontrées. La problématique concerne ici l'extraction des termes (puis, en deuxième temps, des relations) et leur structuration. Elle nécessite l'emploi d'outils et de techniques adaptés, en l'occurrence ceux de traitement automatique des langues naturelles (TALN), qui se prêtent bien au traitement des matériaux textuels où les connaissances sont exprimées dans une langue naturelle.

Une synthèse sur les outils TALN dédiés à l'extraction des termes et relations à partir des textes est proposée dans (Aussenac-Gilles *et al.*, 2000). Il existe plusieurs outils d'extraction des candidats termes : Lexter (Bourigault, 1994), Nomino (http://www.ling.uqam.ca/nomino), Ana (Enguehard, 1992). Pour l'acquisition de relations sémantiques, il existe des approches basées sur l'exploitation des contextes syntaxiques (Grefenstette, 1994), ou sur l'utilisation des patrons lexico-syntaxiques (Hearst 1992 ; Jouis 1993) ainsi que quelques outils, Coatis (Garcia 1998) pour des relations causales, Cameleon (Seguela 1999 ; Seguela et Aussenac-Gilles 1999, Seguela 2001) pour des hyponymes et méronymes.

Samovar (Golebiowska, 2000) permet de structurer les connaissances des champs textuels de la base SGPb, et d'y effectuer des recherches afin de trouver des questions similaires. A partir des diverses BDs de l'entreprise, nous avons construit une ontologie offrant plusieurs vues sur le processus des validations : problème, projet, prestations, pièces. Après avoir ainsi amorcé manuellement l'ontologie, nous l'avons complétée progressivement avec les éléments issus des données textuelles de SGPb à l'aide de techniques et outils linguistiques. Cette étape est automatique, mais l'aide de l'expert est nécessaire tout au long du processus. Puis les problèmes de la base SGPb ont été annotés automatiquement avec les concepts de ces ontologies. En utilisant le moteur de recherche Corese intégré dans Samovar, l'utilisateur peut alors accéder à la base des problèmes pour effectuer des recherches guidées par l'ontologie.

2. L'ontologie de Samovar

L'ontologie de Samovar prend en compte les 4 composantes du processus des validations :

– composante *Pièce* : basée sur le référentiel officiel d'entreprise, et correspondant au découpage fonctionnel d'un véhicule en sous-composants ;

– composante *Problème* : contient une typologie de problèmes reflétant les différents types de problèmes décrits dans le système de gestion des problèmes SGPb ;

– composante *Prestations* : correspond aux prestations recoupées avec le découpage organisationnel de l'entreprise ;

– composante *Projet* : reflète la structure d'un projet et est construite avec des connaissances acquises sur le fonctionnement d'un projet-véhicule, suite aux interviews que nous avons menées auprès des différents acteurs du projet.

Chaque sous-ontologie se présente comme une hiérarchie de spécialisation de *n* niveaux. Chaque sous-ontologie (ou ses composantes) reflète l'un des points de vue principaux sur le processus des validations. Ainsi la composante *Projet* est représentative du cycle d'étude d'un véhicule, de son jalonnement, des options du véhicule, etc. Ce point de vue permet de comparer l'évolution d'un projet par rapport à un autre *via* les problèmes sur la même vague, même tronçon ou même support.

La composante *Projet* contient, entre autres, les références aux différents acteurs comme les responsables pièces ou encore les architectes. Les deux permettent des points de vue correspondant aux découpages possibles des pièces d'un véhicule : point de vue par zone d'architecte et point de vue par zone de responsable pièces. Par exemple, grâce aux références au responsable pièces *Planche de Bord* il sera possible de connaître les pièces du périmètre *Planche de Bord* : tableau de bord, console, aérateur, etc.

La composante *Prestation* reflète un service rendu au client et correspondant à ses attentes. Par exemple, l'aspect ergonomique d'un véhicule : le confort d'un siège, l'acoustique du poste de conduite, les performances d'endurance ou son comportement dans des conditions extrêmes : hautes ou basses températures, fort ensoleillement, etc. Les prestations reflètent, entre autres, l'organisation métier des services d'essai (endurance, bruit, ergonomie, etc.).

Ainsi, les différentes composantes de l'ontologie de Samovar fournissent les points de vus spécifiques pour l'ensemble du processus des validations. Ces points des vue permettent de gérer et d'accéder à l'information de façon sélective, facilitant par là leur exploitation. Ils facilitent aussi l'identification et la détection du vocabulaire propre à chaque métier. Ils améliorent ainsi l'interaction entre les différents acteurs du processus des validations au cours de la gestion d'un problème.

2.1. *Construction de l'ontologie*

La construction de l'ontologie se passe en deux temps, en fonction du type de données et des moyens impliqués :

– une première extraction des informations contenues dans des bases de données,

– une seconde extraction avec des techniques et outils linguistiques, pour découvrir les informations « cachées » dans les textes.

Soulignons qu'avant la première phase il est nécessaire d'établir un modèle préliminaire du domaine (1) qui va guider la construction de l'ontologie.

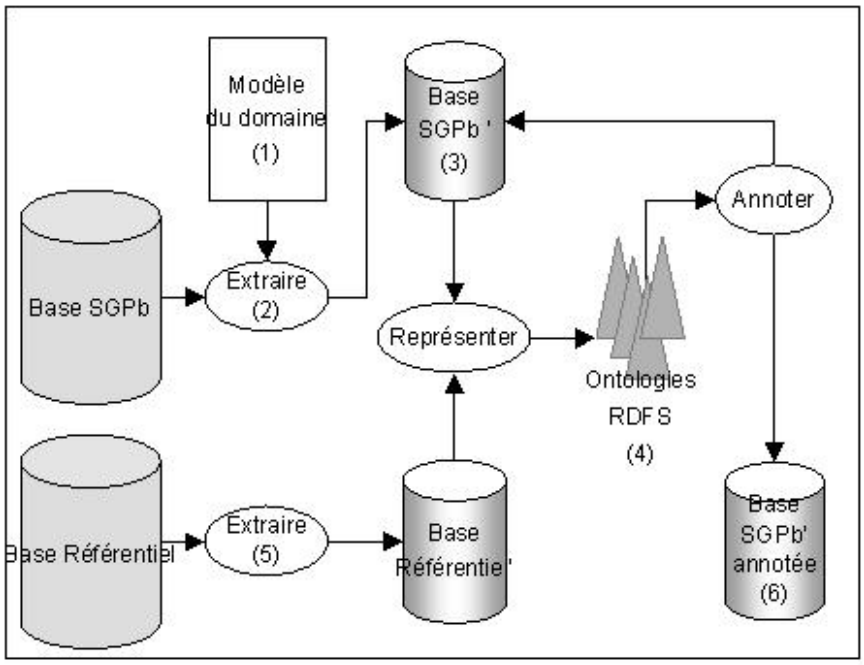

Figure 1. *Construction des ontologies pour Samovar*

L'« ossature » de l'ontologie est amorcée avec les données des bases existantes (cf. figure 3) :

Une première extraction des données d'origine (2) est effectuée suivant le modèle du domaine. Elle fournit ainsi un sous-ensemble de la base d'origine (3) que l'on traduit ensuite sous forme d'ontologie (4), en respectant le format RDFS (prévu pour Corese).

En parallèle, une autre extraction est faite à partir du référentiel Pièce (5), afin de compléter les données précédentes avec les informations supplémentaires. Une fois l'ontologie construite, elle sert pour annoter les données extraites (3) avec les éléments de l'ontologies (4). On obtient ainsi la base initiale annotée (6).

Un deuxième processus traite les données textuelles (l'objectif final étant d'enrichir le résultat de la première extraction avec l'information issue des textes). Ce processus exploite les sorties obtenues après application de l'outil linguistique Nomino sur les données textuelles issues des commentaires textuels dans la base SGPb. Nomino[1] est un outil d'extraction des groupes nominaux (GN) d'un corpus représentatif d'un domaine. Nomino prend un corpus textuel en entrée et produit en sortie un ensemble de « lexiques » – listes de noms, unités complexes nominales (UCN), unités complexes nominales additionnelles (UCNA), verbes, adjectifs, adverbes. Les UCN(A) correspondent aux groupes prépositionnels (GP) ou groupes nominaux (GN). Les lexiques des UCN(A) sont accessibles sous forme de graphes qui illustrent les dépendances existantes pour un GP ou GN.

Nous utilisons les lexiques et graphes produits par Nomino afin de :

– détecter les termes significatifs (*i.e.* correspondant à des points sensibles des validations),

– enrichir l'ontologie Problème à l'aide des graphes de Nomino, en exploitant la régularité de leurs structures.

2.1.1. *Détection des termes sensibles*

Dans un premier temps, nous avons analysé les lexiques produits par Nomino afin de détecter les termes les plus fréquents que l'on peut supposer être les termes plus représentatifs du domaine (*câblage, montage, tuyau, fixation, centrage, pièce, mise en place, conformité, dérivation, trou, agrafe, vis, contact, maintien, serrage, patte, position, géométrie, écrou, visser, hygiène, raccordement, etc.*). Ces termes structurés nous ont permis d'amorcer l'ontologie *Problème*. Sa structuration initiale (cf. figure 2) reposait sur la validation par les experts.

Les termes retenus pour l'amorce sont tous ceux qui sont exploitables en tant qu'indices sémantiques pour un problème : par exemple un problème de *Centrage* va être détecté grâce à la présence d'indices comme *indexage*, *coaxialité*, *entraxe*, etc. La validité des termes a été confirmée par des experts.

1. http://www.ling.uqam.ca/nomino

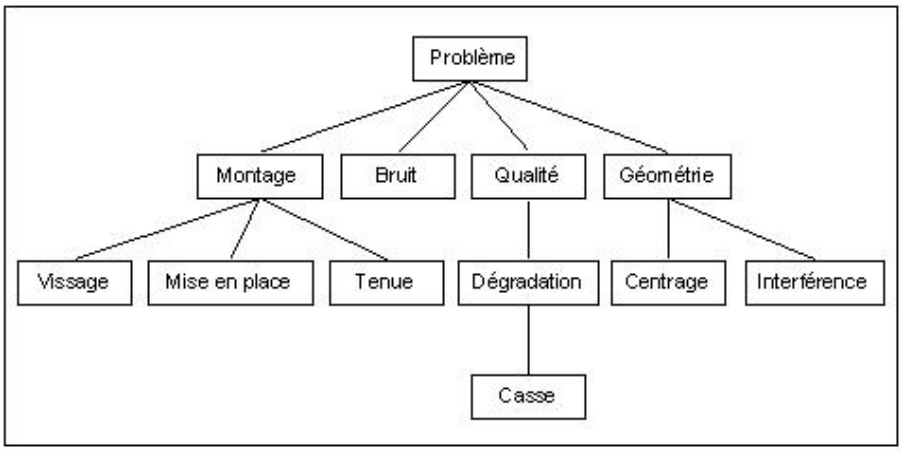

Figure 2. *Extrait de l'ontologie Problème*

L'amorce de l'ontologie étant ainsi constituée, pour l'enrichir, nous avons utilisé les groupes prépositionnels issus de Nomino.

2.1.2. *Exploitation des graphes de Nomino pour l'enrichissement de l'ontologie Problème*

En plus de substantifs, Nomino produit des groupes nominaux et prépositionnels. Nous avons exploité les structures des cas les plus fréquents :

clipsage_impossible clipsage_difficile clipsage_inefficace difficulte_de_mep difficulte_de_mise_en_place difficulte_d'alignement difficulte_de_chaussage difficulte_d'emmanchement difficulte_d'accostage difficulte_d'agraphage difficulte_d'acces difficulte_de_vissage difficulte_de_serrage difficulte_de_emmanchement difficulte_de_montage durete_de_clipsage durete_de_connexion, durete_de_mep eclairage_insuffisant effort_de_clipsage effort_de_declipsage effort_de_mep effort_de_montage effort_de_positionnement effort_de_raccordement effort_de_serrage effort_de_sertissage effort_de_sertissage_insuffisant effort_d_encliquetage effort_de_branchement effort_de_chaussage deterioration_de_connecteur deterioriation_lecheur_ext detrompage_insuffisant gene_pour_clipsage gene_pour_fixation gene_pour_la_fixation_du_presseur gene_pour_la_mep_du_monogramme impossibilite_de_clipsage_du_tuyau impossibilite_de_mep_du_protecteur mauvais_centrage mep_difficile manque_boutonniere probleme_de_clipsage probleme_de_fixation

Figure 3. *Extrait des UCN produits par Nomino*

L'analyse de ces UCN nous a fourni les structures que nous avons exploitées pour construire les règles heuristiques. Ces structures sont composées d'un terme `Indicateur de probleme`, indiquant l'existence d'un problème (*difficulté, effort, impossibilité*), suivi par ce problème (*chaussage, clipsage*). Ainsi on peut trouver des cas comme :

difficulté de … (par exemple difficulté de mise en place)
dureté de … (par exemple difficulté de connexion)
effort de … (par exemple effort de clipsage)
impossibilité de … (par exemple impossibilité de mep)
gêne pour … (par exemple gêne pour fixation)
mauvais … (par exemple mauvais centrage)
difficile … (par exemple mep difficile)

Nous avons exploité ces régularités structurelles pour construire des règles heuristiques qui vont permettre d'alimenter de façon semi-automatique l'ontologie. Ces règles vont refléter les structures existantes dans le corpus, et une fois activées, vont permettre d'enrichir l'ontologie *Problème* en y attachant chaque nouveau terme.

Les règles correspondent donc aux combinatoires possibles entre les éléments des ontologies *Pièce* et *Problème*, et attestées dans les textes. Une règle se présente comme une suite de catégories, chacune pouvant être décorée avec un ensemble de traits (par exemple `type=Problème` pour signifier que l'élément fait partie de l'ontologie *Problème*, `type=Pièce` pour un élément de l'ontologie *Pièce*, etc.).

Prenons comme exemple, la règle heuristique R1. Elle précise que si l'on rencontre l'expression `<terme1> de <terme2>` où `<terme1>` est déjà connu comme correspondant à un concept `concept1` de l'ontologie Problème, alors le système peut suggérer de rattacher un concept dénoté par `<terme2>` comme un fils de `concept1` dans l'ontologie Problème.

R1 : Nom[type= Problème,n=i] Prep[lemme= « de »] **Nom{type → Problème, n=i+1}** ;

Cette règle appliquée à l'UCN(A) BRUIT DE FROTTEMENT DU VOLANT PENDANT SON REGLAGE EN HAUTEUR permettra de proposer un concept « *Frottement* » comme fils du concept « *Bruit* » déjà connu de l'ontologie Problème.

REMARQUE. — une notation différente (accolades et flèche →) permet de distinguer la partie déduction de la règle.

R2 : **Nom{type → Problème}** Prep[lemme= « de »|| lemme= « sur »|| lemme= « sous »] Nom[type= Pièce] ;

La règle heuristique R2 précise que si le système rencontre l'expression <terme$_1$> de <terme$_2$> où <terme$_2$> est déjà connu comme dénotant un concept concept$_2$ de l'ontologie Pièce, alors il suggérera de rattacher un concept désigné par <terme$_1$> comme un concept dans l'ontologie Problème.

Cette règle appliquée au UCN(A) BROUTEMENT DU BRAS-BALAIS AR SUR PPP3 permet de découvrir le concept « *Broutement* », connaissant la pièce « *Bras-balai AR* ».

R3 : Nom[lemme= « difficulte »|| lemme= « effort »|| lemme= « durete »|| lemme= « risque »] Prep[lemme= « de »] **Nom{type → Problème}**

La règle heuristique R3 précise que si le système rencontre l'expression <terme$_1$> de <terme$_2$> où <terme$_1$> est soit *difficulté* soit *effort* soit *dureté*, soit *risque*, alors il suggérera de rattacher un concept désigné par <terme$_2$> comme un concept dans l'ontologie Problème.

Cette règle appliquée à l'UCN(A) DIFFICULTE D'ENCLIQUETAGE TUYAUX CARBURANT SUR RESERVOIR permet de découvrir le concept « *encliquetage* ».

Ces trois types de règles reflètent trois façons d'enrichir l'ontologie :

– l'ontologie peut être enrichie à partir d'elle-même,

– elle peut l'être en exploitant l'ontologie *Pièce*,

– elle peut l'être en exploitant les indicateurs de problèmes.

2.1.3. *Cinématique du processus*

Comme le résume la figure 5, après avoir amorcé l'ontologie *Problème*, nous l'avons complétée progressivement : prenant en entrée les sorties Nomino, l'ontologie *Pièce* et la base d'heuristiques, le système analyse les GN correspondant aux sorties Nomino, pour voir avec quelle règle chacun peut s'apparier.

En sortie, on obtient des candidats-problèmes à placer dans l'ontologie *Problème*. L'utilisateur valide chaque candidat et décide si l'endroit proposé pour l'insérer dans la hiérarchie existante est correct : le terme est alors inséré dans l'ontologie et acquiert le statut de concept.

Pour formaliser l'ontologie, nous avons choisi le langage RDF Schema (RDFS) proposé par le W3C pour décrire l'ontologie à laquelle sont relatives les annotations RDF décrivant des ressources accessibles par le web. Or de telles annotations RDF sont aussi pertinentes pour décrire les ressources faisant partie de la mémoire d'un projet-véhicule, comme les descriptions des problèmes rencontrés dans ce projet.

De ce fait, nous avons développé un parseur qui, à la fin du processus, génère une version de l'ontologie dans RDF Schema. Après la génération de la version RDF(S) de l'ontologie, les annotations des problèmes de la base SGPb sont mises à jour automatiquement.

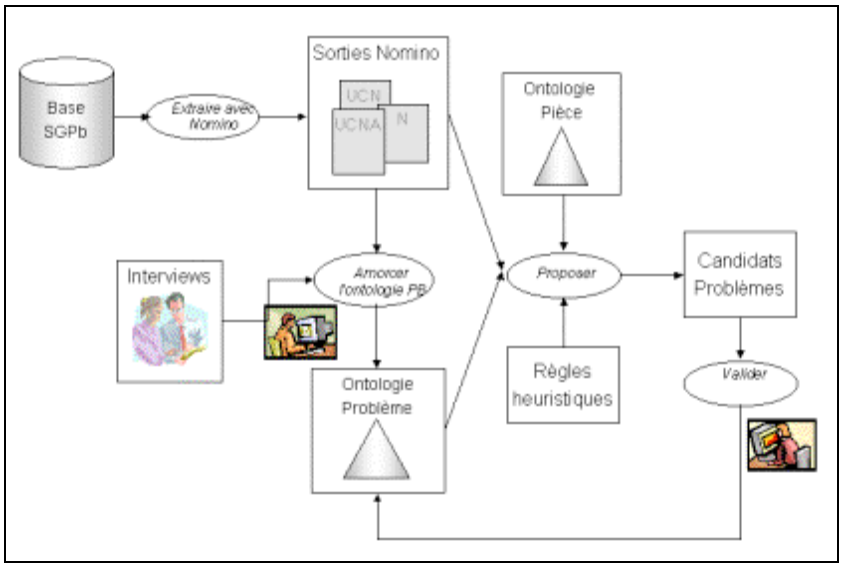

Figure 4. *Processus d'enrichissement de l'ontologie Problème*

2.2. *Exploitation de Corese*

Les ontologies construites servent à l'annotation des documents pour le moteur de recherche Corese (Corby *et al.*, 2000). Celui-ci implémente un processeur RDF/RDFS basé sur les graphes conceptuels (GC) : il utilise en entrée et en sortie les langages RDF et RDFS pour exprimer et échanger des métadonnées sur des documents, et il propose un mécanisme de requêtes et d'inférences basé sur le formalisme des GC. On peut le comparer à un moteur de recherche permettant des inférences sur des annotations RDF, grâce à leur traduction en GC et grâce à la traduction RDF des GC réponses aux requêtes des utilisateurs.

Comme on peut voir sur la figure 5, la démarche de Corese repose sur la correspondance entre deux formats : RDF(S) et les GC, les RDF Schema étant traduits dans le support de GC et les annotations RDF dans la base des GC. Une requête se présente comme un énoncé RDF, traduit en un graphe requête qui est ensuite projeté sur la base des GC afin d'isoler les graphe s'appariant. Les graphes résultats sont ensuite traduits en retour vers RDF. Le mécanisme de la projection prend en compte les hiérarchies et les relations de spécialisation décrites dans les schémas RDF et traduites en GC.

Pour pouvoir exploiter Corese, nous avons traduit automatiquement l'ontologie de Samovar dans le langage RDFS et annoté automatiquement en RDF les problèmes de la base SGPb avec des concepts de ces ontologies. Après ces 2 étapes,

l'utilisateur peut faire des recherches (cf. figure 6) : les réponses aux requêtes prendront en compte les liens modélisés dans l'ontologie de Samovar.

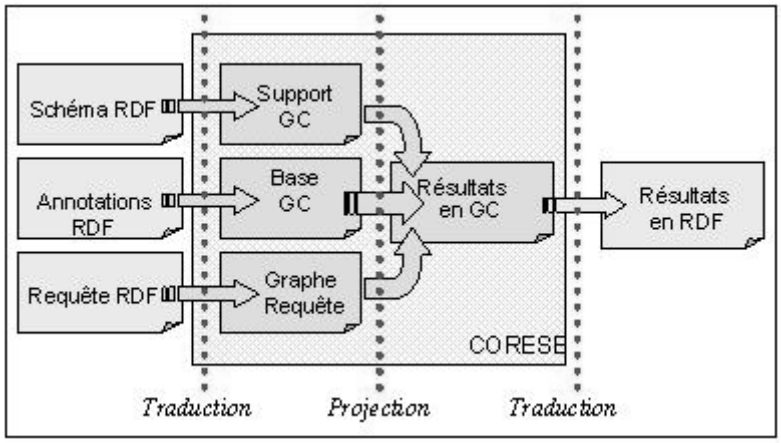

Figure 5. *Démarche de Corese*

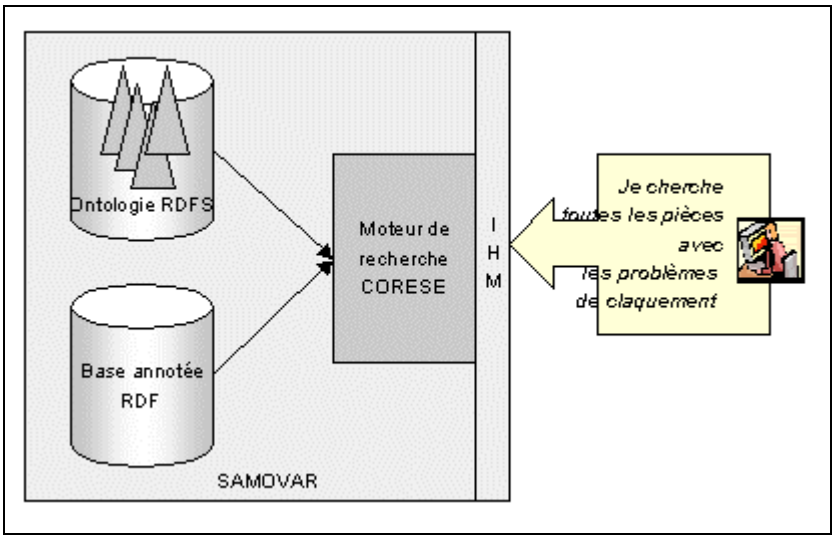

Figure 6. *Organisation de Samovar*

2.2.1. *Exemples de requêtes*

Les 2 exemples suivants montrent que les problèmes issus des textes et structurés avec des liens hiérarchiques permettent de retrouver des doublons de descriptions-de-problèmes.

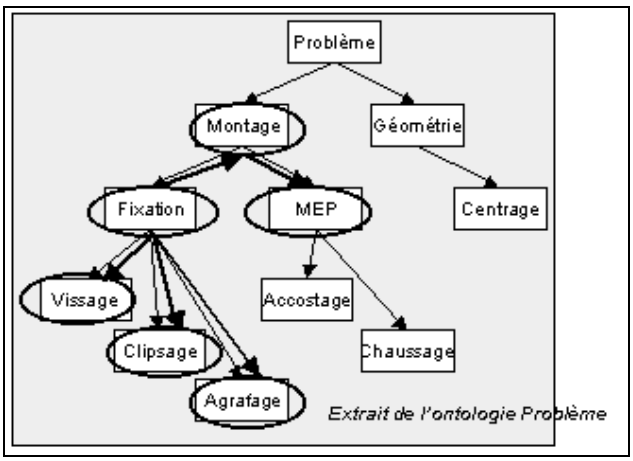

Figure 7. *Parcours des ontologies pour l'extraction de l'information* : *extrait de l'ontologie Pièce*

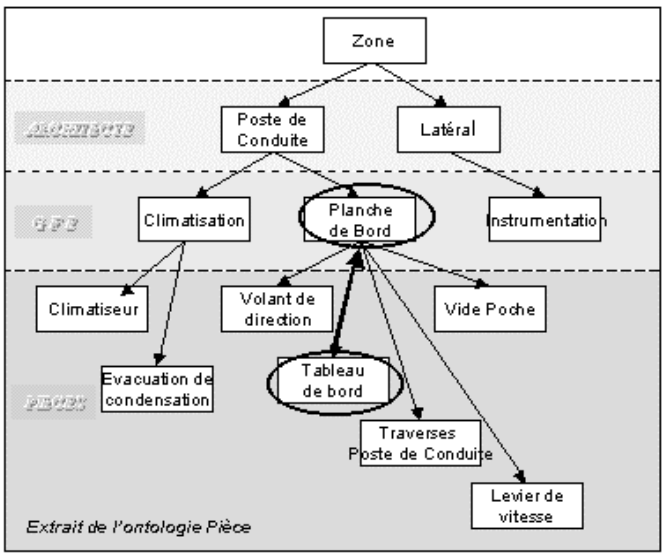

Figure 8. *Parcours des ontologies pour l'extraction de l'information* : *extrait de l'ontologie Pièce*

Soit l'extrait de l'ontologie Pièce illustré par la figure 7, et l'ontologie Problème illustré par la figure 8.

Dans le premier exemple, l'utilisateur cherche à retrouver les `problèmes de fixation sur le soufflet de levier de vitesse`. Une seule réponse est obtenue :

Pb_Fixation rdf :about=http://host.renault.fr/SAMOVARXML/MO-02057.xml
libelle DIAMETRE DU SOUFFLET AU NIVEAU DU BOUTON PRESSION NON EN CONCORDANCE AVEC LE DIAMETRE DU POMMEAU DU SELECTEUR DE VITESSE
(**VOIR PS-00193**)
piece SOUFFLET_DE_LEVIER_DE_VITESSE

En revanche, en étendant la requête aux termes plus généraux, suivant les liens des ontologies (dans notre cas, *Montage*), on retrouve un 2ème cas, et il s'agit bien d'une description-de-problème similaire :

Pb_Montage rdf :about=http://host.renault.fr/SAMOVARXML/PS-00193.xml
libelle BOUTON PRESSION DU SOUFFLET DE LEVIER DE VITESSE IMMONTABLE
(**GEREE PAR MO-02057**)
piece SOUFFLET_DE_LEVIER_DE_VITESSE

Ainsi par un parcours successif des ontologies dans le sens de la généralisation ou de la spécialisation, l'utilisateur peut étendre la requête aux concepts subsumant et aux concepts « frères ». Dans l'exemple, il peut explorer, niveau par niveau, les problèmes de soufflet de levier de vitesse : partant de *Fixation*, remonter vers le concept-père (*Montage*) et redescendre ensuite sur les autres fils (*MEP*) (cf. figure 7).

Dans le second exemple, l'utilisateur aimerait retrouver les problèmes de `centrage sur les traverses de poste de conduite`. Le système retourne 3 cas dont 2 s'avèrent être des descriptions-de-problèmes pointant l'une sur l'autre :

Pb_Centrage rdf :about=http://host.renault.fr/SAMOVARXML/MO-00403.xm
libelle FIXATIONS PDB : FIXATIONS LATERALE G ET COMPTEUR DECENTRE SUR TRAVERSE.
piece TRAVERSE_DE_POSTE_DE_CONDUITE

Pb_Centrage rdf :about=http://host.renault.fr/SAMOVARXML/MO-02071.xml
libelle FIXATION : SUPPORT CARMINAT SUR TRAVERSE DECENTREE. (**VOIR PS-00023**)
piece TRAVERSE_DE_POSTE_DE_CONDUITE

Pb_Centrage rdf :about=http://host.renault.fr/SAMOVARXML/PSXj2-00023.xml
libelle NON COAXIALITE DES TROUS DE FIXATION SUPPORT CALCULATEUR CARMINAT SUR TRAVERSE.(**GEREE PAR MO-02071**)
piece TRAVERSE_DE_POSTE_DE_CONDUITE

Si on n'obtient pas de réponses quant au problème de centrage, il est possible donc d'étendre le résultat aux concepts pères : *planche de bord* ou *montage* pour le premier cas, ou *géométrie* pour le second. Le parcours de l'ontologie permet donc de se déplacer dans l'ensemble de la base, suivant les axes sémantiques modélisés sous forme de liens dans les ontologies, et de retrouver des descriptions-de-problèmes similaires.

De plus, *via* le lien hypertexte (par exemple *Pb_Centrage rdf* :*about*= http*://host.renault.fr/SAMOVARXML/MO-00403.xml*), l'utilisateur peut accéder au contenu intégral d'un problème, c'est-à-dire au texte d'origine.

Mentionnons que la gestion de synonymes et de termes sémantiquement proches constitue une solution intéressante face au vocabulaire hétérogène, rendant ainsi possible l'exploitation du système par les acteurs de métier ou culture différents.

2.2.2. *Evaluation des ontologies pour la recherche des descriptions-de-problèmes similaires*

Les tests ont été effectués sur des ontologies Pièce et Problème couvrant le corpus correspondant à un extrait de la base SGPb d'un projet-véhicule : (1) un premier lot concernait un périmètre spécifique (Planche de Bord) pour 2 jalons, (2) un second lot traitait la totalité de la base du projet. Nous avons construit ces ontologies en prenant en compte les différentes sources des données (référentiel officiel recoupé avec les pièces issues de la base des problèmes). En termes métier, le domaine correspond au processus de montage. En l'état actuel, le périmètre Planche de Bord contient 118 concepts et 3 relations, dont les 22 pièces sont structurées en 6 zones d'architectes, 12 tronçons, et 3 niveaux de granularité reflétant le référentiel officiel ; l'ontologie Problème comporte environ 43 types de problèmes ; l'ontologie Prestation comporte 9 prestations extraites automatiquement de la base ; ces ontologies ont servi à annoter 351 descriptions-de-problèmes. La base entière contient 792 concepts et 4 relations dont les 467 pièces sont structurées de la même manière, plus une typologie en 39 responsables pièces ; l'ontologie Problème complète comporte environ 75 types de problèmes ; l'ontologie Prestation comporte 38 prestations extraites automatiquement de la base ; ces ontologies ont servi à annoter 4 483 descriptions-de-problèmes.

Les recherches des descriptions-de-problèmes similaires effectuées pour le premier lot démontrent que la démarche est intéressante : les descriptions-de-problèmes ayant des pointeurs réciproques ont toutes été retrouvées (pour tous les cas où les descriptions-de-problèmes pointées par un lien faisaient partie du périmètre couvert – les tests étant effectués pour le périmètre de Planche de Bord, les descriptions-de-problèmes traitées peuvent pointer vers d'autres appartenant au même périmètre ou bien à un périmètre différent). La validité de la démarche a été démontrée ainsi : les utilisateurs courants de la base SGPb marquaient les doublons par un code spécifique dans un des champs de la base d'une part, et d'autre part de façon explicite dans le texte du commentaire. Au cours de nos tests, toutes les

descriptions-de-problèmes ayant ce marquage ont été retrouvées. De plus, le nombre de réponses s'en trouve automatiquement réduit. L'exploitation immédiate de la recherche des doublons de validations a fortement intéressé les responsables de projets-véhicule. L'annotation des descriptions-de-problèmes avec les problèmes issus des textes et leur structuration en ontologies que l'on parcourt en cours de généralisation des requêtes, donne donc des résultats intéressants.

On peut constater aussi que la modélisation de l'ontologie est essentielle dans cette démarche. Des modifications dans l'ontologie Problème, faites à titre de tests, ont eu des répercussions plus ou moins heureuses sur les résultats. Il est important de s'assurer de la validité de l'ontologie auprès des experts.

Plus généralement, la démarche est fortement dépendante du corpus du domaine traité : dans le cas de sa réutilisation pour un autre domaine, il sera probablement nécessaire d'actualiser les règles qui permettent d'extraire de nouveaux concepts afin de couvrir les structures non traitées. Les règles sont en effet très liées au domaine.

D'autres « réglages » nécessaires sont apparus au cours du processus. Par exemple les annotations avec des problèmes se font actuellement par appariement – l'annotation avec une étiquette est déclenchée sur la présence d'indices (par exemple *Centrage* va être détecté grâce à la présence des indices comme *indexage*, *coaxialité*, *entraxe*. Selon l'ordre de déclenchement, une description-de-problèmes peut se trouver annotée avec des termes différents : il serait intéressant d'ordonner leur déclenchement. Par ailleurs, il semble qu'adjoindre d'autres outils TALN (tagger et analyseurs) raffinerait davantage les résultats dans la construction de l'ontologie *Problème*.

3. Méthode de construction d'une mémoire de projet

En retrouvant des informations sur les validations en double, Samovar a amorcé un processus de capitalisation qu'on pourrait étendre à plus large échelle – exploiter les incidents et les solutions existantes entre les différents projets-véhicule, voire au sein de la même gamme ou même projet, et à plus long terme, exploiter cette capitalisation pour détecter les problèmes récurrents d'une entreprise en faisant remonter les points sensibles « générateurs de problèmes » aux bureaux d'étude. On peut envisager d'exploiter le principe pour d'autres projets, à condition de faire des « ajustements » notamment au niveau des ontologies. On peut ainsi généraliser notre approche à d'autres domaines, par exemple pour construire et exploiter une mémoire des problèmes rencontrés dans un projet de conception d'un système complexe, (*i.e.* incidents rencontrés pour la conception d'un avion, d'un satellite, voire d'une centrale électrique, etc.). Nous proposons une méthode basée sur les étapes suivantes :

– analyser le processus visé (dans notre cas : les validations d'un projet-véhicule), afin d'établir le modèle du domaine analysé. Ce modèle va guider la construction des ontologies ;

– s'il existe une BD ou une nomenclature décrivant les composants de ce système, l'exploiter pour construire semi-automatiquement une ontologie *Composant* (*i.e.* notre ontologie *Pièce*). Sinon, utiliser des outils linguistiques et une méthode telle que celle décrite dans (Aussenac-Gilles *et al.*, 2000) pour construire cette ontologie *Composant* ;

– de même, s'il existe une description des caractéristiques d'un projet dans l'entreprise (resp. une description de la structure organisationnelle de l'entreprise), l'exploiter pour construire une ontologie *Projet* (resp. *Organisation* – cf. notre ontologie *Prestation*) ;

Dans les deux cas, exploiter le modèle conceptuel construit préalablement, afin de restreindre l'extraction à l'information pertinente ;

– constituer un corpus de textes décrivant les problèmes rencontrés lors d'un ou plusieurs projets existants, et provenant de documents textuels ou de champs textuels de BDs ;

– exploiter un outil linguistique permettant l'extraction de candidats-termes (tel que Lexter et Nomino pour des textes en français) ;

– analyser les régularités dans les candidats-termes susceptibles de décrire des types de problèmes et écrire des règles heuristiques permettant, à partir de l'observation de ces régularités et des ontologies *Composant* et *Projet*, de proposer des termes à intégrer sous forme de concepts dans l'ontologie *Problème* et de proposer leur position dans cette ontologie. Cette phase n'exige pas de compétences particulières, mais peut être facilitée par de l'expérience dans le domaine linguistique ;

– valider par un expert humain les propositions du système ;

– annoter automatiquement par les concepts des ontologies *Problème*, *Composant*, *Projet* et *Organisation* les descriptions élémentaires de problèmes (dans le corpus textuel) ;

– exploiter un traducteur en RDFS des ontologies et un traducteur en RDF des annotations, de façon à pouvoir utiliser le moteur de recherche Corese pour interroger la base ainsi annotée de descriptions de problèmes.

4. Conclusions

La démarche de Samovar est une démarche de construction de mémoire de projets axée sur les problèmes de validations. Elle se base sur l'extraction et l'acquisition des connaissances à partir des sources sélectionnées (système SGPb essentiellement), et leur modélisation sous forme d'ontologie multicomponentielle. Quatre composantes constituent l'ontologie : Problème, Pièce, Prestation et Projet,

offrant ainsi quatre points de vue différents sur le processus des validations. Ces quatre points de vue constituent quatre points d'accès possibles à la masse d'informations, facilitant par là une lecture et une exploitation de ces informations. Samovar permet donc une navigation « profilée », guidée *via* les thèmes et les axes d'intérêt choisis par les utilisateurs.

Samovar peut se comparer à plusieurs approches de construction d'ontologies exploitant des outils TALN à partir de corpus textuels. Terminae (Biébow et Szulman, 1999) propose une méthodologie et un environnement pour la constitution d'une ontologie à partir d'un corpus de textes. La méthode est basée sur l'étude des occurrences des termes dans un corpus afin d'un extraire la définition conceptuelle, et l'environnement aide l'utilisateur dans sa modélisation. Lexiclass (Assadi, 1998) permet de construire une ontologie régionale à partir d'une documentation technique, grâce à un outil classant, en fonction de leur contexte terminologique, les syntagmes extraits du corpus et aidant le cogniticien à découvrir les champs conceptuels importants du domaine. (Aussenac-Gilles *et al.*, 2000) décrivent une méthodologie générale de construction d'ontologies à partir des textes en combinant différents outils TALN existants (*i.e.*, Lexter pour l'extraction des candidats termes, et Cameleon pour l'extraction des relations, Terminae aidant à la construction de la hiérarchie). Notre démarche s'en rapproche : nous utilisons aussi différents outils spécifiques à chaque étape du processus, mais avec un corpus d'origines diverses (des interviews et des données textuelles issues des BDs existantes). (Maedche et Staab, 2000) proposent aussi une architecture générale pour la construction d'une ontologie à partir des textes : ils exploitent (a) différents outils TALN pour construire une taxonomie de concepts et (b) des techniques d'apprentissage pour extraire d'autres relations des textes. Enfin, notons que Samovar permet la recherche des problèmes similaires comme un système de raisonnement à partir de cas (Moussavi, 1999).

Remerciements

Nous remercions nos collègues pour leurs conseils judicieux au cours des nos travaux et leurs relectures de cet article, ainsi que Renault qui a financé ces travaux.

5. Bibilographie

ASSADI H., Construction d'ontologies à partir de textes techniques, Application aux systèmes documentaires, Thèse de doctorat, Université Paris 6, 1998.

AUSSENAC-GILLES N., BIÉBOW B., SZULMAN S., « Revisiting Ontology Design : a Method Based on Corpus Analysis », *EKAW 2000*, Juan-les-Pins, 2-6 October, p. 172-188.

BIÉBOW B., SZULMAN S., « Terminae : a linguistics-based tool for building of a domain ontology », *EKAW 1999*, LNAI 1621, Springer-Verlag, 1999.

BOURIGAULT D., Lexter, un Logiciel d'Extraction de TERminologie, Application à l'acquisition des connaissances à partir de textes, Thèse de doctorat, E.H.E.S.S, Paris 1994.

CORBY O., DIENG R., HEBERT C., « A Conceptual Graph Model for W3C Resource Description Framework », *ICCS 2000*, Springer-Verlag, Darmstadt, August 2000, p. 468-482.

ENGUEHARD C., ANA, Apprentissage Naturel Automatique d'un réseau sémantique, Thèse de doctorat, UTC, 1992.

GARCIA D., Analyse automatique des textes pour l'organisation causale des actions. Réalisation du système informatique COATIS, Thèse de doctorat, Université de Paris IV, 1998.

GOLEBIOWSKA J. (2000), « SAMOVAR – Setting up and Exploitation of Ontologies for capitalizing on Vehicle Project Knowledge », *EKAW 2000, Workshop Ontologies and Texts*, Juan-les-Pins, p. 79-90.

GOLEBIOWSKA J. (2002), Exploitation des ontologies pour la mémoire d'un projet-véhicule, Méthode et outil SAMOVAR, Thèse de doctorat, UNSA, 2002.

GREFENSTETTE G., *Explorations in automatic thesaurus discovery*, Kluwer Academic Publishers,1994 Boston.

HEARST M., « Automatic Acquisition of Hyponyms from Large Text Corpora », *COLING 1992*, Nantes, July 25-28.

JOUIS C., Contribution à la conceptualisation et à la Modélisation des connaissances à partir d'un analyse linguistique de textes. Réalisation d'un prototype : le système SEEK, Thèse EHESS, 1993.

MAEDCHE A., STAAB S., « Mining Ontologies from Texts », *EKAW 2000*, p. 189-202.

MORIN E., « Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique », *TAL 1999*.

MOUSSAVI M., « A Case-Based Approach to Knowledge Management », *AAAI 1999 Workshop Exploring Synergies of KM and CBR*, Orlando, FL. AAAI Press 1999, Technical Report WS-99-10.

SÉGUÉLA P., AUSSENAC-GILLES N., « Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine », *IC 1999*, p. 79-88, Paris.

SÉGUELA P., « Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés », *Terminologies Nouvelles*, 1999, 19 :52-60.

SÉGUELA P., Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques, Thèse de doctorat, Université Toulouse III, 2001.