RÉCITAL 2002, Nancy, 24–27 juin 2002

Conceptualisation d'un système d'informations lexicales, une interface paramétrable pour le T.A.L

Djamé Seddah, Evelyne Jacquey*

Laboratoire LORIA, Equipe Langue et Dialogue,
Campus Scientifique, BP 239
F-54506 Vandœuvre-lès-Nancy Cedex
jacquey@loria.fr, seddah@loria.fr

Mots-clefs – Keywords

Bases de données, ressources lexicales
Data Bases, Lexical Ressources

Résumé - Abstract

La nécessité de ressources lexicales normalisées et publiques est avérée dans le domaine du TAL. Cet article vise à montrer comment, sur la base d'une partie du lexique MULTEXT disponible sur le serveur ABU, il serait possible de construire une architecture permettant tout à la fois l'accès aux ressources avec des attentes différentes (lemmatiseur, parseur, extraction d'informations, prédiction, etc.) et la mise à jour par un groupe restreint de ces ressources. Cette mise à jour consistant en l'intégration et la modification, automatique ou manuelle, de données existantes. Pour ce faire, nous cherchons à prendre en compte à la fois les besoins et les données accessibles. Ce modèle est évalué conceptuellement dans un premier temps en fonction des systèmes utilisés dans notre équipe : un analyseur TAG, un constructeur de grammaires TAGs, un extracteur d'information.

Lexical ressources which would be normalized and freely accessible is a major issue in the NLP research area. This article aims to show how to built an information system which allow (1) a freely access for distinct NLP systems (tagging, parsing, information extraction, etc.) and (2) an easy update of data by a restricted team of researchers, this update being manual or computed. Starting with a a subset of the MULTEXT lexicon which is accessible from the server ABU, we aim to take into account the various needs and the variability of accessible lexical data. Our modelisation is evaluated with three existing systems of our team : EGAL (parsing), a builder of Tag grammars and VULCAIN (information extraction).

*Cet article a été rédigé à parts égales par les deux auteurs. Djamé Seddah est actuellement doctorant et Evelyne Jacquey a soutenu sa thèse en décembre 2001.

1 Introduction et problématique

Les travaux théoriques des linguistes sont le plus souvent fondés sur un lexique, supposé préexistant et pertinent pour le ou les phénomènes étudiés. Les faits ne valident cependant pas ce présupposé. Dans le traitement automatique des langues, l'un des problèmes récurrents est la non-réutilisabilité des lexiques : ils sont le plus souvent construits empiriquement et liés irrémmediablement à l'application visée, ils ne sont que très rarement disponibles gratuitement et suffisamment standardisés. Le Lexique-Grammaire (LADL), n'est accessible gratuitement qu'en consultation. Il existe un certain nombre de lexiques morpho-syntaxiques publics, tel MULTEXT[1], mais aucun ne contient des informations de valence ou de pronominalisation, des informations sémantiques comme les restrictions sélectionnelles, les rôles sémantiques, des informations conceptuelles interfacées avec des hiérarchies conceptuelles comme EuroWordnet ou avec des thesauri.

Une solution possible serait de disposer d'une base de connaissances lexicales normalisées avec trois objectifs principaux : (1) synthétiser les ressources lexicales disponibles, les faire collaborer et les enrichir, (2) éviter de construire un nouveau lexique dédié chaque fois qu'un nouveau système est construit, et (3), permettre à des applications différentes mais complémentaires de collaborer pour mettre en œuvre une tâche multimodulaire.

Nous nous proposons d'amorcer ici les spécifications d'un système d'informations lexicales en précisant les axes qui ont permis sa modélisation tout en veillant à sous-spécifier ce système pour qu'un processus d'échange avec la communauté puisse l'enrichir sans en bouleverser les fondements. Notre contribution se situe donc sur le premier axe de la réflexion, celui de la définition d'un format abstrait et d'une organisation générique des données lexicales correspondant aux besoins décelés dans la section (2). La méthodologie developpée (section 3) est celle de la conception de bases de données relationnelle (Merise) qui sera ensuite étendue à une famille de schémas XML compatibles dans nos futurs travaux (Romary, 2001). La validation conceptuelle du système proposé se fera en la confrontant à trois types d'applications (un analyseur TAG, une méta-grammaire et un système d'extraction d'information) en section (4).

2 Des besoins multiples

Avant d'envisager une modélisation, nous résumons, en l'illustrant par quelques exemples, notre analyse des besoins.

L'analyseur de grammaire d'arbres adjoints : La méthode usuelle d'analyse des parseurs TAG[2] consiste à préselectionner les arbres de la grammaire dont les ancres correspondent aux

[1]MULTEXT (Ide & Véronis, 1994), (Véronis & Khouri, 1995), est le plus gros projet d'envergure européenne visant à l'avènement d'une base linguistique commune. Le succès de ce projet est venu de la simplicité des objectifs affichés : standardiser des ressources textuelles et des données linguistiques, créer des ressources linguistiques informatisées, monolingues et multilingues, et, créer des outils génériques pour l'annotation et l'exploitation de corpus. La mise au point des étiquettes s'est fait dans une optique de simplicité et de compromis visant à leur réutilisabilité évidente à travers une dénomination linguistique quasiment universelle au sein des langues indo-européennes (Nom, Adverbe, Verbes...). L'un des aspects les plus intéressant est la sous-spécification de chaque langue par rapport au lexique initial qu'elle intègre.

[2]Les grammaires d'arbres Adjoints (Joshi *et al.*, 1975) sont un formalisme syntaxique où l'on associe, à la substitution classique des grammaires hors contexte, une opération appellée adjonction consistant en l'insertion

items lexicaux présents dans la phrase à analyser, (Vijay-Shanker & Joshi, 1985). Un des problèmes connus du formalisme LTAG est l'explosion combinatoire de la quantité d'arbre dûe à la lexicalisation de la grammaire[3]. Une solution repose sur l'élaboration d'une méta-grammaire TAG (Candito, 1999) qui, à partir d'un système de règles, génère des schèmes (arbres non instanciés) ancrés par le lexique via une procédure basée sur le filtrage de structures de traits. Bien que cette approche simplifie le processus, une modification des structures de traits lexicaux ne se propage pas sur la méta-grammaire sans une intervention humaine conséquente.

L'idéal serait de pouvoir modifier les entrées lexicales sans pour autant devoir modifier l'interface d'accès à ces mêmes données. Un système répondant dynamiquement aux requêtes d'une méta-grammaire à travers un métalangage de descripteur serait une solution envisageable.

L'étiqueteur de E.Brill : Cet outil, servant à attribuer une étiquette morpho-syntaxique à un mot d'un corpus, est l'une des applications les plus courantes dans le domaine du TAL. Un processus relativement simple mais fastidieux permet de créer le lexique permettant le traitement d'un corpus (Seddah, 1998). Néanmoins le lexique ainsi créé n'est que très difficilement réutilisable, c'est pourquoi l'étiquetage, et les opérations qu'il englobe, seraient éminemment simplifiés si on disposait en amont d'un lexique suffisamment conséquent pour couvrir la plupart des mots du corpus tout en permettant de faire varier la granularité des informations présentes dans les étiquettes morpho-syntaxiques.

Notre système, pour être réellement utilisable, doit pouvoir générer le plus grand lexique possible et éventuellement permettre de changer le jeu d'étiquette dynamiquement (en remplaçant éventuellement les étiquettes morpho-syntaxiques par des arbres TAG) sans modifier le système d'inférence du tagger (Brill, 1992; Deloupy, 1995).

Les systèmes de génération de textes : Le principe de la génération de textes est l'inverse de celui de l'analyse : partir d'une représentation conceptuelle pour aboutir au texte contenant l'information à communiquer, (Reiter & Dale, 2000), (Gardent, 2002). Les besoins vis-à-vis d'un lexique sont inversés : ils vont des concepts aux lexèmes.

En prenant pour élément central, une représentation conceptuelle du prédicat, le lexique doit permettre d'accéder à plusieurs informations : (1) les définitions sémantiques correspondant aux représentations conceptuelles du prédicat et des autres concepts mis-en-jeu, et (2), pour chaque définition sémantique, les rôles associés aux participants ; rôles thématiques, positions dans les LCS (*Lexical Conceptual Structures*) de Jackendoff (Jackendoff, 1990). Muni de ces informations, le système d'informations lexicales doit enfin fournir la ou les liste(s) de lexèmes à utiliser avec, de préférence, des informations statistiques de pertinence.

L'extraction d'informations : Le système VULCAIN (Todirascu & Romary, parution en mai 2002) a pour objectif d'extraire de l'information sur un domaine limité, celui de textes sur la sécurité des systèmes informatiques. Les données d'entrée sont des textes. A partir des résultats partiels d'un analyseur LTAG, des listes de groupes significatifs sont extraits, principalement des groupes nominaux simples de la forme Nom-Nom, Nom-Adjectif, GroupeNominal-Préposition-GroupeNominal. En sortie de ce système d'extraction d'information, doivent appa-

d'un sous-arbre dans une branche d'un autre arbre.

[3]La grammaire FTAG (Abeillé, 1991), retranscrite en XML, devient trop lourde pour des systèmes classiques : dans sa version actuelle, elle correspond à une dizaine de milliers de schèmes.

raître des listes d'entités identifiées (instances et catégories) et validées sur l'ontologie.

Les informations nécessaires au bon fonctionnement de ce système sont, outre l'analyseur LTAG, un lexique TAG au format TAGML et des liens entre les entrées lexicales de celui-ci et l'ontologie du domaine. La particularité du lexique TAG attendu repose sur les liens entre lexique et ontologie, qui sont représentés par des correspondances entre les arbres élémentaires et les structures conceptuelles associées aux concepts de l'ontologie du domaine. De plus, ces structures conceptuelles sont écrites dans le formalisme des logiques de description.

L'interface syntaxe - sémantique : Les travaux actuellement en cours dans notre équipe de recherche consistent, à partir des données sémantiques de la grammaire et de l'analyse syntaxique d'une phrase donnée, à enrichir l'analyse syntaxique issue d'un parseur LTAG (op cité) d'un calcul de la forme logique correspondant à la forêt partagée résultat. Les données nécessaires se synthétisent principalement par l'intersection de plusieurs domaines : les informations sémantiques propres à l'ancre lexicale et la structure argumentale du schème que l'entrée lexicale instancie. Or pour récupérer ces données il est nécessaire de travailler sur la réalisation de ces intersections paradigmatiques, le lexique. Notre système d'information va donc proposer un filtrage des informations issues de la méta-grammaire et va enrichir cette description d'un lexique logique tout en inférant une sous-spécifications de ces types via le cadre de sous-catégorisation de l'entrée lexicale.

Synthèse des types de données : Les exemples que nous venons d'illustrer brièvement montrent l'étendue et la diversité des types de données attendus suivant les applications. Au vu de ces quelques exemples, un système d'informations lexicales normalisées et partagées doit se caractériser par plusieurs propriétés majeures :

- La modularité de l'information : les différents types d'informations lexicales ne peuvent plus être simplement organisés autour de la classification reposant sur les niveaux linguistiques de description (phonétique, morphologie, syntaxe, sémantique). Les méthodes de conception des bases de données, qu'elles soient relationnelles ou objet, nous semblent appropriées dans ce cas pour plusieurs raisons :
  - Via leurs différents niveaux de représentations schématiques, elles permettent d'avoir une vue globale du système ainsi que des vues plus partielles selon les besoins,
  - Les représentations schématiques, si elles sont rigoureusement documentées, permettent des échanges avec des chercheurs de toute culture scientifique, ce qui permet une meilleure adéquation avec les besoins de la communauté et une mise à jour régulière de ceux-ci,
  - Elles reposent sur des meta-langages formels dont les propriétés ont été démontrées.
- La diversité des formats de sortie des différents types de données attendus. L'information sémantico-conceptuelle est attendue sous la forme de λ-termes typés ou de DRS (*Discourse Representation Structure*), sous la forme de structures de traits typées particulières (elles sont destinées au filtrage des schèmes pertinents) avec les méta-grammaires, sous la forme d'expressions en logique de description avec les systèmes d'extraction d'information comme VULCAIN, et enfin, sous la forme de structures conceptuelles conformes au formalisme de Jackendoff dans le cadre d'un processus de génération comme celui qui sera défendu par le projet GENI.

Il nous semble donc important d'envisager le processus de réponses aux attentes des applications en deux temps : retour du résultat sous une forme standard définie dans le système d'informations lexicales lui-même, puis, traduction de ce format dans celui demandé par l'application client.

3 Conception d'un système d'informations lexicales

Le système d'informations lexicales dont nous proposons une modélisation dans le formalisme des bases de données relationnelles comporte, comme nous l'avons vu plus haut, un nombre conséquent de types d'informations. L'accès à l'information devant être modulaire, dans l'ensemble des types d'informations, nous avons différencié sept sortes principales : les restrictions sélectionnelles, les représentations conceptuelles, les représentations sémantiques, les réalisations syntaxiques par argument, les informations flexionnelles, les informations concernant les familles de constructions syntaxiques, et par la suite, les informations dérivationnelles. Nous donnerons dans un premier temps le schéma entité-association des données du système envisagé ainsi que le schéma relationnel correspondant en 3FN (troisième forme normale).

Dans un deuxième temps, nous aborderons les actions subies par le système ou produites par lui. Pour synthétiser les choses, nous proposerons un schéma conceptuel des traitements pour les actions principales.

3.1 Schéma conceptuel et relationnel des données

Le schéma conceptuel des données du système d'informations lexicales proposé est représenté dans la figure (1)[4]. Ce schéma comporte une entité par type d'informations identifié ci-dessus.

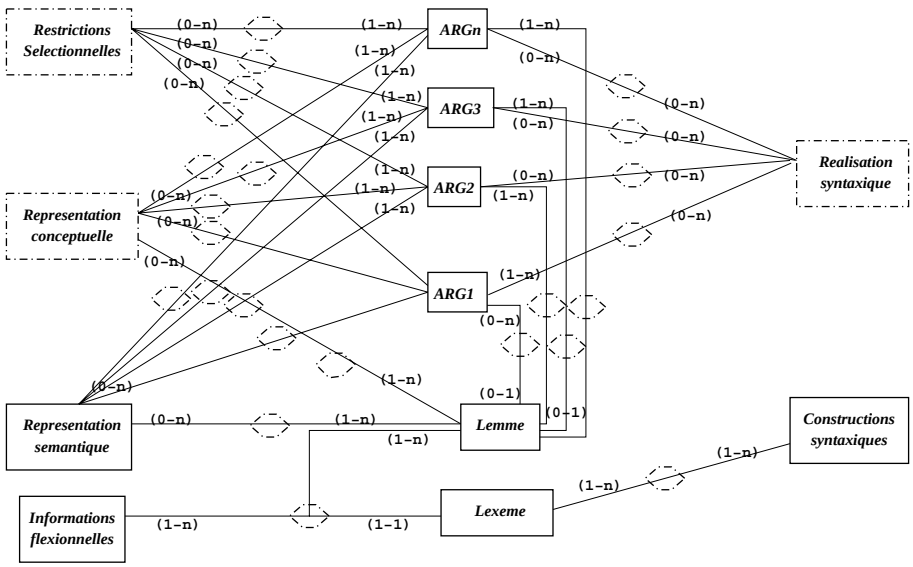

Figure 1: Le schéma entité association des données du système

[4]Dans ce schéma, toutes les entités entourées de rectangles en pointillés représentent des données non encore accessibles.

Par ailleurs, en termes d'organisation des informations lexicales classiques, nous avons insisté sur plusieurs points : **La variabilité sémantique et conceptuelle** d'un même lemme[5] est représentée par les cardinalités (N-M) des associations entre ces deux types d'informations et l'entité LEMME.

La variabilité des réalisations syntaxiques des arguments est représentée par des associations de type (N-M) entre les entités ARGUMENT et l'entité RÉALISATION SYNTAXIQUE.

L'accès direct aux informations propres à chaque argument d'un lemme prédicatif est rendu possible via la décomposition passant par les entités ARGUMENT.

Chaque forme fléchie, c'est-à-dire chaque instance de l'entité LEXÈME est supposée correspondre à au plus une paire construite à partir d'un lemme et d'une forme de flexion.

Le reste des associations du schéma peut être classé selon le type de cardinalité.

Les associations ((0-n) - (1-n)) caractérisent tout d'abord les liens entre les restrictions sélectionnelles, les représentations conceptuelles, les représentations sémantiques, les réalisations syntaxiques, et les arguments. Cela signifie que, pour tout argument, il y a au moins une information de chacun des types cités ci-dessus et il peut y en avoir plusieurs. En revanche, certaines instances des types d'informations citées peuvent ne pas caractériser une instance d'un argument donné, ou peuvent en caractériser plusieurs.

La même type d'association caractérise le lien entre une instance de lemme et une instance de CONSTRUCTIONS SYNTAXIQUES.

Les associations ((0-n) - (0-1)) caractérisent le lien entre une instance de lemme et des instances d'arguments. Cela signifie que pour une instance de lemme, il peut y avoir au plus une instance de chaque type d'argument interne.

L'association ((0-n) - (1-1)) caractérise uniquement le lien entre une instance de lemme et une instance d'argument externe.

Enfin, chacun des types d'entités du schéma en (1) comporte au moins un attribut IDENTIFIANT et un attribut CONTENU dont la valeur est une chaîne de caractères. Pour tous les types d'informations à valeur non atomique, nous créons un attribut RESSOURCES dont la valeur est le chemin du fichier contenant l'information, par exemple une description XML d'un schème. A terme, ces fichiers contiendront l'information dans un format commun à l'ensemble du système, mais qui reste à définir. Le format de cette information sera ensuite traduit par un ou plusieurs module(s) particuliers dans le format requis par l'application client.

Le schéma entité-association ci-dessus peut être transformé en schéma relationnel en 3FN (troisième forme normale). Etant données les cardinalités des associations, le schéma relationnel comporte treize tables correspondant aux entités et vingt tables correspondant aux associations de type (N-M).

[5] Dans le cas de mots homonymes, nous supposerons deux lemmes différents, par exemple *fraise*1 pour le fruit et *fraise*2 pour l'outil. Dans le cas de mots lexicalement ambigu, un même lemme correspondra à deux n-uplets distincts de l'entité REPRÉSENTATION SÉMANTIQUE, ce sera le cas de *plateau* par exemple.

3.2 Les différentes actions possibles

Comme le montre la figure (2), l'ensemble des traitements des données peut être schématisé de manière globale. Ce schéma est représenté selon la méthode de conception MERISE[6].

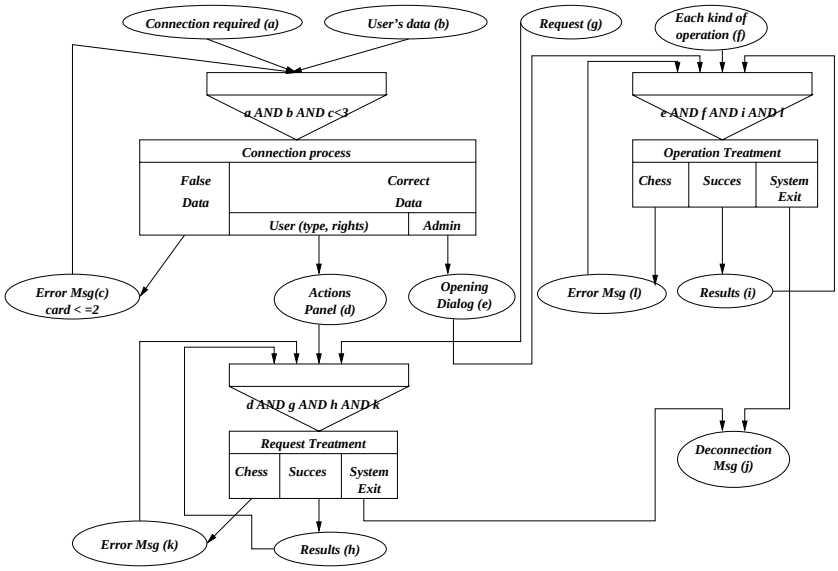

Figure 2: Le schéma conceptuel des traitements du système

On notera plusieurs points particuliers. Tout d'abord, lors de la procédure de connexion, le système est supposé reconnaître l'utilisateur désirant se connecter, et en particulier son type : "utilisateur" ou "administrateur". S'il est du type "utilisateur", le système est supposé détenir les informations décrivant ses droits. Cela signifie donc que le système dispose d'une base contenant ces données. Nous ne l'avons pas représentée dans la mesure où elle ne nous semble pas centrale pour notre sujet, l'organisation des informations lexicales.

Ensuite, l'ensemble des traitements du système est subdivisé en deux sous-ensembles en fonction du type d'utilisateur : "utilisateur" ou "administrateur". En particulier, les réponses du système varient d'un type à l'autre : plus conviviales et plus limitées (résultats d'interrogation ou messages d'erreurs aussi explicites et informatifs que possible) avec le type "utilisateur", pas nécessairement conviviales et plus étendues avec le type "administrateur" dans la mesure où celui-ci peut procéder à tout type d'opérations sur le système (interrogation, mise à jour, insertion, suppression). On pourrait encore imaginer un troisième type, le type "contributeur". Les contributions seraient vérifiées et intégrées ensuite par un administrateur.

Enfin, la procédure de connexion est autorisée à échouer deux fois seulement, ce qui est représenté dans la condition de synchronisation "s1" par la condition "c<3".

[6]Les ronds représentent des données ou "événements", les rectangles représentent des opérations avec des données d'entrée et des données de sortie, et les triangles représentent des conditions de synchronisation entre les données d'entrée pour le déclenchement d'une opération.

4 Validation conceptuelle de la modélisation proposée

Analyse syntaxique : L'analyseur dont nous avons étudié les besoins est un analyseur TAG (Lopez, 1999). Comme nous l'avons vu en (2), l'une des difficultés avec ce type d'analyseur est l'explosion du nombre d'arbres formant le lexique, explosion qui rend leur gestion difficile. Le système d'informations lexicales proposé peut fournir une aide importante via l'organisation des données qu'il suppose. Ainsi, pour instancier les arbres décrivant la construction passive avec sujet nominal, une solution serait de sélectionner tous les lexèmes étant au participe passé avec l'auxiliaire *être* et ayant un objet direct nominal. Une fois cette sélection faite, il suffit d'associer à chaque instance un arbre décrivant la construction passive en question. Pour ce faire, une requête appropriée en algèbre relationnelle sera appliquée à l'ensemble des relations LEMME(id_lem, contenu, id_arg1, id_arg2, id_arg3, ..., id_argn), ARG2-RSYNT(id_arg2, id_rsynt), RSYNT(id_rsynt, contenu), LEXEME(id_lex, contenu, id_lem, id_flex) et FLEXIONS(id_flex, contenu). L'ensemble des lemmes ayant un objet direct nominal est égal au résultat de la requête
$R1 = \Pi_{id\_lem}(\bowtie_{id\_arg2} (\bowtie_{id\_rsynt} ((\sigma_{(contenu='np')}(\text{RSYNT})), \text{ARG2-RSYNT}), \text{LEMME}))$.
L'ensemble des lemmes ayant une forme passive et un objet direct nominal est égal au résultat de la requête
$R2 = \Pi_{id\_lex}(\bowtie_{id\_flex} (\sigma_{(contenu='ppassé-passif')}(\text{FLEXIONS}), \bowtie_{id\_lem} (R1, \text{LEXEME})))$.
Il suffit alors de modifier la table LEXÈME - CONSTRUCTIONS SYNTAXIQUES pour toutes les instances différentes de l'attribut `id_lex` et de modifier la table CONSTRUCTIONS SYNTAXIQUES en y insérant un nouvel n-uplet de la forme (csynt_0001, 'constr. passive en *par*', file=ArbrePassifPar).

Méta-grammaires : Telle qu'elle est construite actuellement, la méta-grammaire de B. Gaiffe et de B. Crabbé[7], basée sur (Candito, 1999), produit des schèmes non instanciés décrits par des structures de traits (cf. section 2.1). Ces structures de traits décrivent les propriétés des schèmes, ces propriétés n'étant pas forcément structurées comme le sont les informations lexicales. Les schèmes décrivant les différentes structures passives par exemple vont au moins contenir la paire attribut-valeur [PASSIF = +].

Cette information n'étant pas directement accessible dans le système d'informations décrit en section 3, deux solutions sont possibles. La première consiste à demander à l'utilisateur de traduire les schèmes en question sous la forme de requêtes interprétables par le système d'informations. Cette première solution n'est cependant pas conforme à nos objectifs, ni même avec le schéma conceptuel de traitement (section 3.2), car l'utilisateur n'est pas supposé connaître la structure des données du système. Une seconde solution consiste donc à construire un traducteur pour les structures de traits décrivant les schèmes. Ce traducteur sera intégré au système d'informations sous la forme d'un module indépendant, activé uniquement lorsque l'application client sera une méta-grammaire.

La conception d'un tel traducteur demande néanmoins la synthèse entre les informations lexicales, plus ou moins accessibles comme la valence, la réalisation des arguments, l'existence d'une forme participe passé avec auxiliaire *être* par exemple, et le comportement syntaxique des lexèmes.

[7] Travail de doctorat en cours et non encore publié.

Extraction d'informations : Le système qui nous sert actuellement de "laboratoire d'expériences" est le système VULCAIN (cf. section 2.4). L'une de ses particularités est le lien existant entre les entrées lexicales et les concepts de l'ontologie du domaine.

Une fois ces liens définis dans (Todirascu & Romary, parution en mai 2002), il sera possible de les intégrer sous la forme d'instances de la table LEXÈME - REPRÉSENTATION CONCEPTUELLE. Le système VULCAIN pourra ainsi se décharger de cette information qui lui sera rendue dynamiquement par le système d'informations lexicales.

Une telle perspective demande cependant là encore un coût de traduction, et cela pour deux raisons. Tout d'abord, le lexique de VULCAIN est décrit en TAGML. L'extraction du lexème de chaque entrée lexicale demande donc une première traduction. Ensuite, VULCAIN attend une représentation conceptuelle exprimée en logique de description, ce qui demande une seconde traduction.

5 Conclusion et perspectives

Au vu du besoin croissant dans le domaine du TAL de ressources lexicales normalisées et librement accessibles, et dans la suite du lexique MULTEXT, nous avons jeté les bases d'un système d'informations lexicales destiné à répondre aux besoins des applications, différents par leur nature et par le format des données attendues.

Les schémas conceptuels des données et des traitements peuvent être vus comme une traduction, dans le cadre des bases de données relationnelles, d'un format abstrait et d'une organisation des données lexicales. Ces schémas, quoique non suffisants, ont permis d'approcher plus précisément le problème complexe que constitue un tel projet. En particulier, les quelques exemples d'applications pris en compte ont montré la diversité des données attendues, tant pour ce qui de leur organisation que pour ce qui est de leur format.

Plusieurs directions de recherche nous semblent dont importantes. Tout d'abord, redéfinir un format abstrait des données qui permettent une traduction aussi peu coûteuse que possible dans les différents formats de données attendus (structures de traits, arbres TAG et TAGML, logique de description, logique des prédicats, λ-calcul). D'autre part, définir plusieurs formats d'échanges entre le système d'informations lexicales d'un côté, et certaines applications caractéristiques du TAL. Enfin, mais cela restera une tâche permanente, il nous semble indispensable d'intégrer les résultats d'études lexicales existantes et exploitables pour enrichir les données actuelles.

Références

ABEILLÉ A. (1991). *Une grammaire lexicalisée d'arbres adjoints pour le français.* PhD thesis, Paris 7.

BRILL E. (1992). A simple rule-based part of speech tagger. In *Proc. of the Third Conference on Applied Natural Language Processing*, p. 152–155, Trento, Italy.

CANDITO M.-H. (1999). *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées, Application au français et à l'italien.* PhD thesis, Université de Paris 7.

DELOUPY C. (1995). La méthode d'étiquetage d'Éric brill. *Traitement Automatique des Langues.*

GARDENT C. (2002). *Proposition d'Action Concertée de Recherche INRIA pour le projet GENI*. Rapport interne, INRIA.

IDE N. & VÉRONIS J. (1994). Multext (multilingual tools and corpora). In *14th Conference on Computational Linguistics (COLING'94), Kyoto, Japan*.

JACKENDOFF R. (1990). *Semantic Structures*. MIT Press, Cambridge, MA.

JOSHI A., LEVI L. & TAKAHASHI M. (1975). Tree adjunct grammars. *Journal of the Computer and System Sciences*.

LOPEZ P. (1999). *Analyse d'énoncés oraux pour le dialogue homme machine à l'aide de Grammaires Lexicalisées d'arbres*. PhD thesis, Université Henry Poincaré Nancy 1.

REITER E. & DALE R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

ROMARY L. (2001). Towards an abstract representation of terminological data collections - the tmf model. In *TAMA 2001, Terminology in Advanced Microcomputer Application*.

SEDDAH D. (1998). Mesure des poids et distances pour l'inférence de grammaires stochastiques. Master's thesis, Université Paris 7.

TODIRASCU A. & ROMARY L. (parution en mai 2002). *Le projet VULCAIN, rapport final*. Rapport interne, LORIA.

VIJAY-SHANKER K. & JOSHI A. K. (1985). Some computational properties of tree adjoinig grammars. In *COLING*, Chicago, USA.

VÉRONIS J. & KHOURI L. (1995). Etiquetage grammatical multilingue: le projet multext. *Traitement Automatique des Langues*.