Une approche par acceptions pour les bases lexicales multilingues

Gilles Sérasset, Etienne Blanc

Acception Based Approach for Multilingual Lexical Databases

Gilles Sérasset Étienne Blanc

GETA, IMAG-campus
(UJF & CNRS)
BP 53, 38041 Grenoble Cedex 9, France
Gilles.Serasset@imag.fr, Etienne.Blanc@imag.fr

À paraître dans les actes de T-TA-TAO 93, Montréal, 30/9/-2/10/93.
Collection "Actualités Scientifiques"

Abstract

Many projects are conducted to develop multilingual lexical databases. Some of these projects use an interlingual approach (KBMT-89, EDR, …), where others choose a bilingual approach (Multilex, …).

GETA uses an interlingual approach based on acceptions (word-senses) to develop its multilingual lexical database management system: NADIA. With this approach, the interlingual lexicon is the union of the acceptions of the languages that appear in the database.

The interlingual set of relations is freely defined by the linguist, with the exception of the predefined "isa" relation, which is necessary in any case because acceptions of terms from one language don't always have lexical equivalents in another language.

This lexical architecture has no influence on the linguistic content of the monolingual dictionaries, neither on the format adopted by the linguist to code the information.

The NADIA system is designed to be multilingual (the system handles multilingual databases), application independent (databases can be used for several purposes), generic (the linguistic structures are defined by the linguist) and theory independent (many computational formalisms can be used to define the linguistic structures). The system provides many tools (browser, editor, defaulter, coherence checker, …) to simplify the management of a multilingual lexical database. Moreover, it handles the management of the interlingual dictionary as much as possible.

EDR and KBMT-89 projects (which use a knowledge based approach) are faced with theoretical and methodological problems (concept discrimination, knowledge representation, concept classification, …). The acception based approach is a good choice to avoid the complexity introduced by knowledge representation problems and to keep advantages of the interlingual approach.

Multilex provides a way for the linguist to define the different linguistic structures, however, the linguist can only use typed feature structures to code his linguistic theory. At GETA we have chosen to provide the linguist with different formalisms (trees, graphs, typed feature structures, sets, …).

Keywords

Multilinguism, Lexical Databases Management System, Lexical Databases, Interlingua, Acceptions.

Une approche par acceptions pour les bases lexicales multilingues

Gilles Sérasset Étienne Blanc

GETA, IMAG-campus
(UJF & CNRS)
BP 53, 38041 Grenoble Cedex 9, France
Gilles.Serasset@imag.fr, Etienne.Blanc@imag.fr

Résumé

De nombreux projets sont menés pour développer des bases lexicales multilingues. Certains utilisent une approche interlingue (KBMT-89, EDR, …), alors que d'autres ont choisi une approche bilingue (Multilex, …).

Le GETA utilise une approche interlingue basée sur les acceptions (sens de mots) pour développer son système de gestion de bases de données lexicales: NADIA. Selon cette approche, le lexique interlingue est l'union des acceptions des langues qui apparaissent dans la base.

L'ensemble des relations et attributs interlingue est défini librement par le linguiste, à part la relation prédéfinie "isa" (relation de sous-acception à sur-acception), qui est nécessaire dans tous les cas, car les acceptions des termes d'une langue n'ont pas toujours d'équivalent lexical dans une autre langue.

Cette architecture lexicale n'a pas d'influence sur le contenu linguistique des dictionnaires monolingues, pas plus que sur le format adopté par le linguiste pour coder l'information.

Le système NADIA est conçu pour être multilingue (le système gère des bases multilingues), indépendant des applications (les bases peuvent être utilisées dans diffèrent buts), générique (les structures linguistiques sont définie par un linguiste) et indépendant de la théorie (de nombreux formalismes informatiques peuvent être utilisés pour coder les structures linguistiques). Le système propose de nombreux outils (navigateur, éditeur, défauteur, vérificateur de cohérence, …) pour simplifier la gestion d'une base multilingue. De plus, il prend en charge, autant que possible, la gestion du dictionnaire interlingue.

Les projets EDR et KBMT-89 (qui utilisent une approche basée sur la connaissance) sont confrontés à des problèmes théoriques et méthodologiques (raffinement des concepts, représentation de la connaissance, classification des concepts, …). L'approche basée sur les acceptions est un bon choix pour éviter la complexité introduite par la représentation des connaissances, tout en gardant l'avantage d'une approche interlingue.

Multilex donne un moyen au linguiste de définir différentes structures linguistiques, mais celui-ci ne peut utiliser que les structures de traits typées pour coder sa théorie linguistique. Au GETA, nous avons choisi de proposer au linguistes de nombreux formalismes (arbres, structures de traits typées, graphes, ensembles, …).

Mots clés

Multilinguisme, Systèmes de gestion de bases lexicales, Bases lexicales, Interlingua, Acceptions.

Introduction

Les besoins en ressources lexicales de grande taille pour le Traitement Automatique des Langues Naturelles (TALN) en général et la Traduction Automatique (TA) en particulier, augmentent chaque jour. On considère que ces ressources représentent la partie la plus coûteuse d'un système de TALN. Pour cette raison, on observe un intérêt croissant pour le développement de dictionnaires réutilisables.

Pour développer une base de données lexicale multilingue, deux approches peuvent être utilisées. En premier lieu, *l'approche par transfert,* où les liens entre les langues, se font au travers de dictionnaires bilingues et unidirectionnels. Cette approche est utilisée par de nombreux systèmes de TA, ainsi que par certains projets de bases de données, comme notamment, les projets Acquilex et Multilex. En second lieu, *l'approche interlingue,* où les liens entre les langues se font au travers d'un dictionnaire interlingue unique. Une telle approche a été adoptée aux USA par le projet KBMT-89 (Knowledge Based Machine Translation) à l'université Carnegie Mellon et par le projet japonais EDR (Electronic Dictionary Research).

Le laboratoire GETA (Groupe d'Étude pour la Traduction Automatique) s'intéresse aux problèmes posés par la construction et l'utilisation de bases de données lexicales multilingues, indépendantes d'une théorie linguistique et indépendantes d'une application. Pour cela, le GETA a choisi de développer un système de gestion de bases de données lexicales, le système NADIA. Ce système est basé sur une approche interlingue. Comme unités interlingues, nous avons choisi d'utiliser les *acceptions* (sens généralement reconnus d'un mot). Le système NADIA fournit de nombreux outils pour la gestion d'une base lexicale multilingue. De plus, ce système laisse libre champ au linguiste quant aux structures linguistiques des entrées.

Cet article présente le projet NADIA. Après une étude générale de l'approche interlingue, nous présenterons le cadre dans lequel s'inscrit ce système, puis, nous donnerons une vue détaillée de l'approche par acceptions. Ensuite, nous étudierons le projet en présentant brièvement Parax, une étude de faisabilité, et l'architecture du système NADIA. Finalement, nous étudierons certains projets de bases de données lexicales. Nous verrons les choix fait par les responsables de ces projets et les problèmes auxquels ils ont été confrontés. Nous terminerons par une justification de nos choix, au vu de ces projets.

I. Approche par acceptions

1. Approche interlingue

L'approche interlingue utilise un langage artificiel intermédiaire (appelé *interlangue* et employé comme langage pivot) pour réaliser le lien entre les langues.

Les énoncés de toutes les langues considérées peuvent être représentés par cette interlangue (indépendamment de la langue de l'énoncé). Aussi, une interlangue doit-elle avoir son propre lexique et son propre ensemble d'attributs et de relations.

Une interlangue doit être définie en référence à un certain ensemble de langues naturelles, à moins qu'un univers de référence fixe (ontologie) ne soit représenté de manière autonome par la machine.

Une interlangue consiste en deux parties distinctes : un lexique et un ensemble d'attributs et de relations.

1.1. Lexique interlingue

La première partie d'une interlangue est le lexique. Celui-ci doit être suffisamment complet pour représenter les différents sens des mots trouvés dans l'ensemble des langues considérées.

Ainsi, un lexique interlingue doit contenir au moins autant de sens de mots que chaque dictionnaire monolingue.

Comme une interlangue est définie pour établir un lien entre les langues, ce lexique interlingue doit fournir un lien lexical entre les mots dans différentes langues. Aussi, deux sens équivalents de différentes langues doivent-ils être reliés à une seule unité interlingue.

Hélas, il n'y a pas nécessairement correspondance directe entre les sens des mots de différentes langues. Prenons l'exemple des mots français "fleuve" et "rivière" (dans leur sens concret le plus commun). Ces deux mots sont traduits en anglais par le mot "river" (dans son sens le plus commun). Les deux mots français ont deux sens différents[1]. L'anglais ne distingue pas ces deux sens. Un lien doit donc être établi entre ces sens dans le lexique interlingue. Par contre, cette distinction n'est pertinente que si l'on va de l'anglais vers le français. Dans un contexte de traduction anglais-japonais, cette distinction n'a pas lieu d'être, puisque le mot japonais "kawa" recouvre le même sens que le mot anglais "river".

Si l'interlangue est définie via un univers de référence fixe (ontologie), une description des différents sens des mots de chaque langue devra être donné dans cet univers.

Dans ce cas, des sens équivalents de différentes langues devront avoir des descriptions identiques. De plus, des sens "proches" (comme "rivière" et "fleuve") devront avoir des description "proches". Dans un contexte de traduction automatique, cette "distance" entre les descriptions devrait être automatiquement reconnue dans le cas où l'on n'a pas équivalence directe.

1.2. Attributs et relations interlingues

La seconde partie de l'interlingue est l'ensemble de ses attributs et relations. Cet ensemble d'attributs et de relations doit être suffisamment complet pour permettre de coder les aspects linguistiques de toutes les langues considérées.

Cette partie n'est pas simple à définir, même si des études linguistiques fondamentales produisent de plus en plus de "microthéories" interlingues ou universelles (selon les termes de [Nirenburg et Defrise 1990]) pour des phénomènes linguistiques, tels que l'aspect, le temps, la modalité, etc. qui, 20 ans plus tôt, semblaient ne pouvoir être décrits que par référence à une langue.

1.3. Projets utilisant l'approche interlingue

Certains projets ont adopté l'approche interlingue. Parmi ces projets, considérons l'américain KBMT-89 et le japonais EDR. Ces projets ont des buts différents. Le premier développe des dictionnaires pour une application particulière (un système de traduction basé sur la représentation des connaissances), alors que le second développe des bases lexicales pour différents systèmes de traduction automatique.

---

[1] Une "rivière" est un cours d'eau se jetant dans un autre cours d'eau. Un "fleuve" est un grand cours d'eau se jetant dans la mer.

Le projet KBMT-89 [Gates et al. 1989, Nirenburg 1989] a défini et implémenté un système de traduction basé sur la connaissance du monde. Pour cela, un dictionnaire de 900 unités lexicales pour l'anglais et de 800 unités lexicales pour le japonais (représentant environ 1500 concepts) a été développé.

Le projet EDR [EDR 1988, EDR 1990] vise au développement de ressources de grandes tailles. EDR a développé de grands dictionnaires d'environ 300 000 mots en anglais et 300 000 mots en japonais (200 000 mots de vocabulaire général et 100 000 en vocabulaire terminologique), ainsi qu'un dictionnaire de 400 000 concepts, un dictionnaire de cooccurrences (en anglais et japonais) ainsi que des dictionnaires bilingues anglais-japonais et japonais-anglais (300 000 entrées chacun).

2. Cadre du projet NADIA

Notre projet, NADIA (Neutral Advanced Dictionaries by Interlingual Acceptions) est un projet visant au développement d'un système de gestion de bases de données lexicales multilingues.

Nous visons 4 objectifs principaux :

- **Multilinguisme** : le système gère des bases de données multilingues. Aussi devrons-nous prendre en compte les différents systèmes d'écriture et les différentes procédures de tri.
- **Indépendance vis-à-vis des applications** : le système n'introduit pas de restriction sur les applications qui utiliseront les bases définies. Toute utilisation des bases est possible (pour traduction, correction, apprentissage des langues par l'homme), pourvu que l'information linguistique nécessaire soit présente.
- **Généricité** : les structures linguistiques utilisées par les bases seront définies par un linguiste, via un langage spécialisé.
- **Indépendance vis-à-vis des théories** : le système ne doit pas introduire de restrictions sur la théorie linguistique sous-jacente à une base. Il doit au contraire permettre l'utilisation de nombreuses théories linguistiques qui mettent en œuvre de nombreux formalismes informatiques.

Nous avons choisi une approche interlingue car nous pensons ainsi apporter la meilleure solution au critère d'indépendance vis-à-vis des applications. Une telle approche assure la compatibilité avec une application interlingue. De plus, il est possible de générer des dictionnaires bilingues pour tout couple de langue présent dans la base.

Afin de réduire les problèmes de l'approche interlingue, nous avons choisi d'utiliser des acceptions comme unités interlingues. Une acception d'une langue est le sens particulier d'un mot, admis et reconnu par l'usage. Une acception, en tant qu'unité interlingue est une acception d'une des langues de la base.

Le projet ULTRA du CRL (Computing Research Laboratory, New Mexico state university) utilise une approche analogue [Farwell et al. 1992].

3. L'interlangue vue par NADIA

Afin de pouvoir fournir des outils utiles et puissants, le système de gestion de bases lexicales introduit des restrictions sur les bases elles-mêmes : les bases lexicales multilingues devront être basées sur l'approche interlingue par acceptions.

3.1. Architecture lexicale

L'architecture lexicale régit l'organisation des différents dictionnaires et leur relations à l'intérieur d'une base lexicale.

Nous insistons sur le fait que cette architecture n'a pas d'influence sur le contenu linguistique des dictionnaires de la base, pas plus que sur le format que le linguiste adopte pour coder l'information.

Une base lexicale sous NADIA est composée de deux types de dictionnaires. Le premier type ne comprend qu'un dictionnaire : le dictionnaire interlingue. Le second type regroupe les dictionnaires monolingues.

Le dictionnaire interlingue contient les acceptions des différentes langues de la base.

Les dictionnaires monolingues contiennent l'information linguistique des différentes entrées lexicales. Le linguiste est libre de définir, pour chaque langue, ses entrées, ses unités lexicales et leurs informations associées, pourvu que le dictionnaire monolingue fasse le lien entre entrées et acceptions. Un dictionnaire monolingue est généralement divisé en deux parties. La première regroupe les acceptions de la langue du dictionnaire (partie purement monolingue). La seconde regroupe les acceptions d'autres langues qui n'ont pas d'équivalent dans la langue du dictionnaire (partie contrastive).

3.2. Lexique interlingue

L'utilisation des acceptions nous permet d'éviter les problèmes du raffinement des sens. Pour chaque langue, on peut choisir un dictionnaire existant comme référence. Les acceptions de la langue seront les acceptions trouvées dans ce dictionnaire de référence.

Le dictionnaire interlingue d'acceptions consiste en l'union ensembliste des acceptions des langues de la base. S'il y a correspondance directe entre deux acceptions de deux langues différentes, celles-ci sont confondues en une seule acception interlingue. S'il n'y a pas correspondance directe, les acceptions d'une langue sont conservées dans le dictionnaire interlingue.

La gestion d'un lexique interlingue est une tâche complexe. Aussi, cette gestion est assurée par la machine. Les modifications sur le lexique interlingue seront effectuées lorsque la machine détectera un problème.

3.3. Attributs et relations de l'interlangue

L'ensemble des attributs et des relations de l'interlangue n'est pas fixé. Cet ensemble est défini par un linguiste pour chaque base.

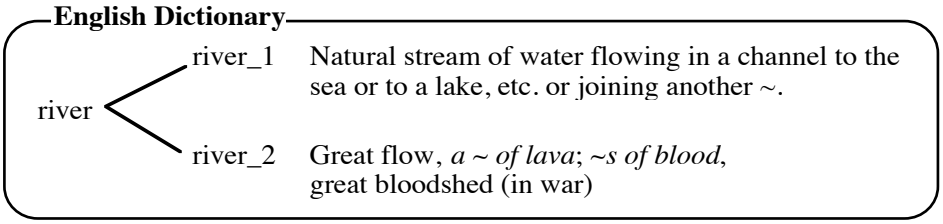
Par contre, afin de gérer les équivalences indirectes (comme dans l'exemple de "rivière", "fleuve" et "river" vu plus haut), la relation entre acceptions "isa" (relation de sur à sous acception) est prédéfinie. Ainsi, par cette relation, on code que les acceptions usuelles de "rivière" et de "fleuve" sont des sous-acceptions de l'acception usuelle de "river".

Cette relation sera toujours définie dans le dictionnaire interlingue d'acceptions. Elle constitue l'ensemble minimal des relations et attributs interlingues.

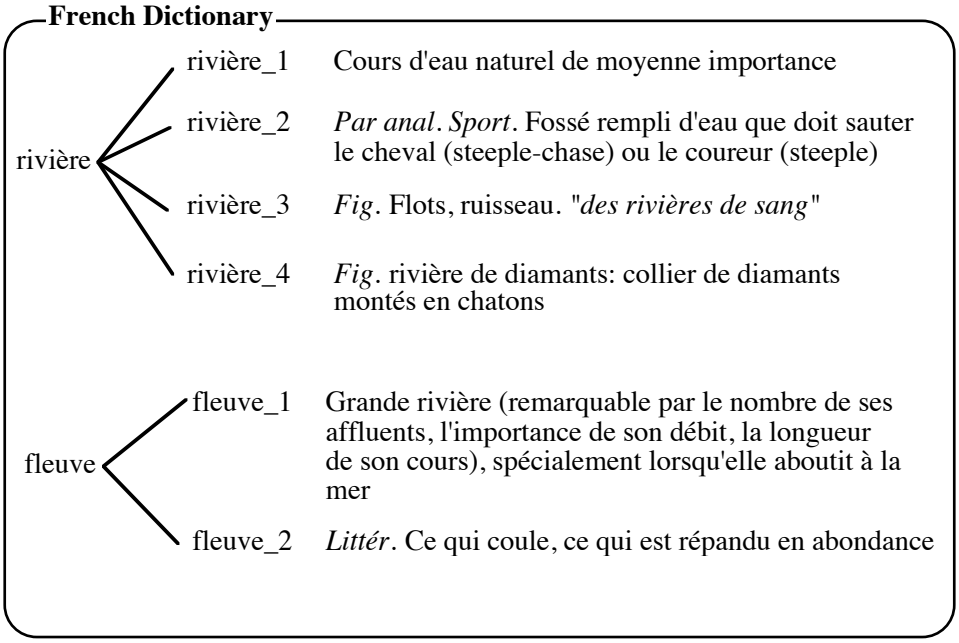
3.4. Un exemple

Reprenons l'exemple précédent plus en détail.

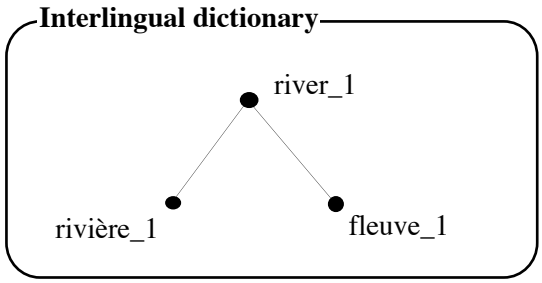
Dans la partie purement monolingue du dictionnaire anglais, l'entrée "river" a deux acceptions :



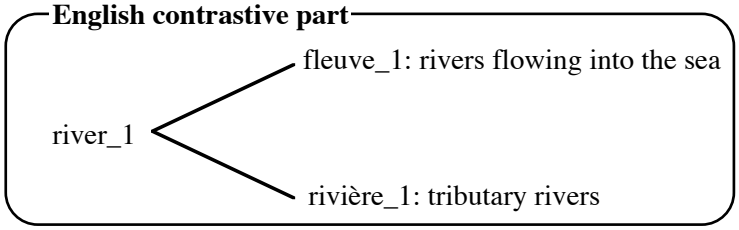
La partie purement monolingue du dictionnaire français contient les acceptions suivantes pour "rivière" et "fleuve".



Comme le mot anglais "river" peut être traduit par "rivière" ou par "fleuve" suivant ce qu'il représente, le dictionnaire interlingue doit établir que les acceptions "rivière_1" et "fleuve_1" sont des sous-acceptions de l'acception "river_1".



La partie contrastive du dictionnaire anglais contient les informations permettant de distinguer les deux acceptions.



II. Le projet NADIA

1. Parax

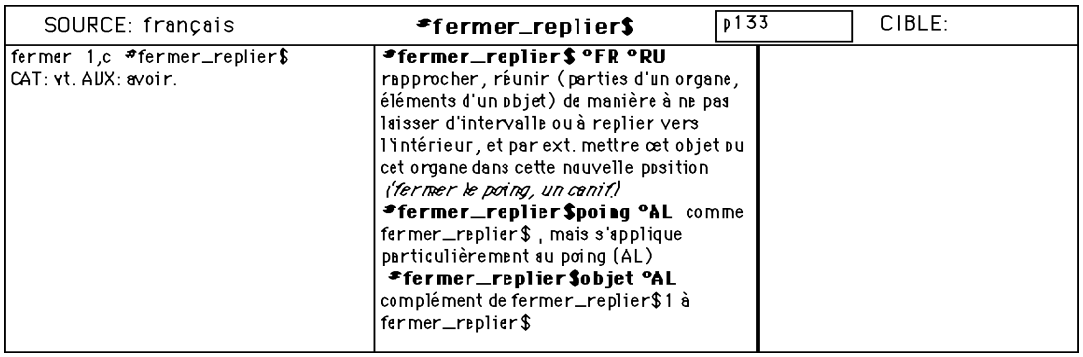
Parax est une maquette implémentée sous HyperCard™ par Étienne Blanc pour étudier la faisabilité d'une telle approche. Cette maquette comporte pour l'instant 5 langues : français, anglais, allemand, russe et chinois.

La structure linguistique des dictionnaires monolingues est inspirée des dictionnaires du générateur de systèmes de traduction automatique ARIANE du GETA [Boitet et al. 1982, Boitet et al. 1985]. Cette structure linguistique est fixe.

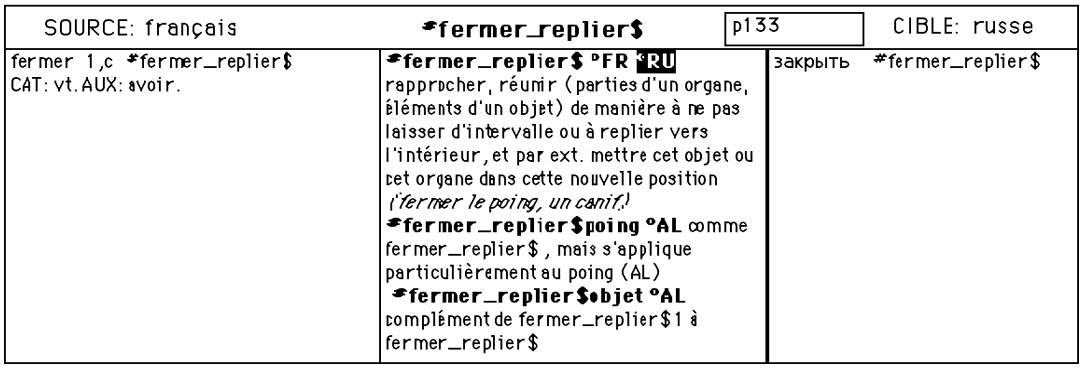
Cette maquette illustre le choix de l'interlangue par acceptions. L'utilisateur peut sélectionner un mot d'une des langues de la base. Puis, Parax lui propose une liste d'acceptions correspondant au mot choisi. Chaque acception est accompagnée d'une définition en français. Dans une première approche, les définitions sont stockées avec les acceptions dans la base interlingue (et non dans chacune des bases monolingues). Elle n'apparaissent donc que dans la langue de référence de la base : le français.

	fermer_1
fermer 1,a #fermer_fermeture$ CAT: vt. AUX: avoir.	**fermer_fermeture$ °AL °AN °FR °RU** appliquer une partie mobile pour boucher un passage, une ouverture *(fermer la porte, les rideaux,)* **fermer_fermeture$organe °CH** fermer *(bouche, yeux,....)* **fermer_fermeture$norg °CH** fermer *(sauf bouche, yeux....)*
fermer 1,b #fermer_lieu CAT: vt. AUX: avoir.	**fermer_lieu °AL °AN °FR °RU** priver de communication avec l'extérieur, par la mise en place d'un élément mobile; interdire le passage *(fermer sa chambre, une valise, une route,)*
fermer 1,c #fermer_replier$ CAT: vt. AUX: avoir.	**fermer_replier$ °FR °RU** rapprocher, réunir (parties d'un organe, éléments d'un objet) de manière à ne pas laisser d'intervalle ou à replier vers l'intérieur, et par ext. mettre cet objet ou cet organe dans cette nouvelle position *(fermer le poing, un canif,)* **fermer_replier$poing °AL** comme fermer_replier$, mais s'applique particulièrement au poing (AL) **fermer_replier$objet °AL** complément de fermer_replier$1 à fermer_replier$

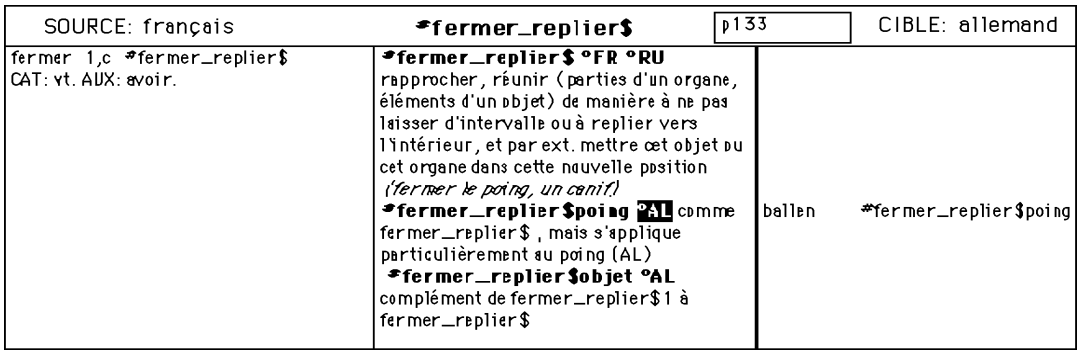
L'utilisateur peut ensuite choisir une des acceptions. Dans l'exemple fourni, on trouve un problème contrastif entre le français et l'allemand. Il existe deux acceptions allemandes différentes correspondant à l'acception française sélectionnée : "fermer_replier". Si l'objet du verbe "fermer" est un poing, la traduction allemande sera "ballen", sinon, elle sera "zuklappen".



Si l'utilisateur veut un équivalent russe, il n'a qu'a choisir le russe comme langue cible.



Par contre, s'il désire la traduction allemande, il lui faudra sélectionner la sous-acception voulue, puis la langue cible.



L'utilisateur peut ainsi voir le terme français et son équivalent allemand (pour l'acception donnée).

2. NADIA : Architecture logicielle

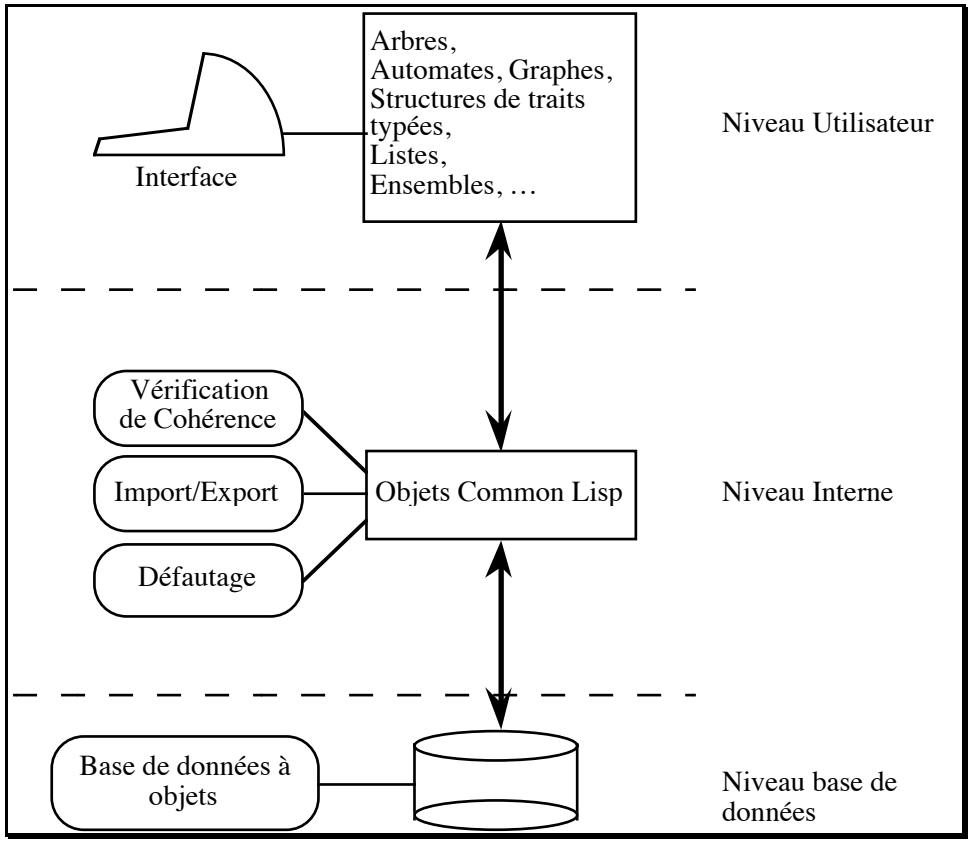
La maquette Parax n'est qu'une illustration de l'approche interlingue que nous avons choisie. Elle n'autorise aucune indépendance vis à vis d'une théorie linguistique. Elle ne fournit pas non plus d'outils spécialisé pour la gestion de bases lexicales multilingues.

Le projet NADIA est la seconde étape dans le développement d'un système de gestion de bases lexicales. Ce projet vise à la construction d'un prototype aussi complet que possible d'un tel système. Ce prototype inclut tous les outils prévus pour la gestion de bases lexicales. Il est construit sans souci de performance en terme de nombre d'unités lexicales.

Ce prototype est développé sur Macintosh™ , avec MCL (Macintosh Common Lisp) et CLOS (Common Lisp Object System). Il utilise les techniques de programmation par objets.

L'architecture logicielle de NADIA se compose de 3 parties :

- *Le niveau base de données* est le niveau le plus bas. C'est à ce niveau que les objets seront archivés ou retrouvés. Le prototype utilise une base de données à objets du domaine public écrite en MCL : WOOD[2]. Ce niveau est complètement transparent aux utilisateurs du système (lexicographes, linguistes,…).
- *Le niveau interne* est le niveau où travaillent les différents outils linguistiques. Les unités linguistiques sont représentées comme des objets de Common Lisp. Ce niveau est transparent à l'utilisateur.
- *Le niveau utilisateur* est le niveau abstrait où travaille l'utilisateur. À ce niveau, l'utilisateur manipule des objets linguistiques (arbres, structures de traits typées, listes, automates, graphes, ensembles, …). Les différents outils communiqueront à ce niveau.



Le fonctionnement de cette architecture est basé sur l'aller et retour entre les niveaux utilisateur, interne et base de données.

La définition de la structure des entrées d'un dictionnaire est faite au niveau utilisateur en terme de structures linguistiques usuelles (arbres, structures de traits, graphes, …). Cette définition est ensuite traduite en terme d'objets Common Lisp.

---

[2] WOOD signifie "William's Object Oriented Database". Il s'agit d'un programme permettant de manipuler des objets persistants en MCL. Il est écrit par Bill St. Clair (bill@cambridge.apple.com).

La définition des règles de cohérence et d'intégrité est faite par le linguiste au niveau utilisateur. Ces règles sont "compilées" et utilisées au niveau interne. Elle sont stockées dans la base.

Les requêtes d'interrogation de la base sont exprimées au niveau utilisateur. Elles sont traduites et évaluées au niveau interne. Le résultat est présenté au niveau utilisateur.

L'utilisateur peut exporter et importer des entrées ou parties d'entrées. La définition des règles d'import et d'export est faite au niveau utilisateur, les règles sont compilées et effectivement appliquées au niveau interne. Le résultat des procédures d'export est stocké sous un format SGML (Standard Generalised Markup Language) reflétant les structures linguistiques (les conventions de TEI (Text Encoding Initiative) seront suivies autant que possible).

3. Gestion des dictionnaires

On trouve deux sortes de dictionnaires dans une base lexicale multilingue NADIA :

- *Les dictionnaires monolingues* sont divisés en deux parties :
  - Une partie purement monolingue qui contient les acceptions de la langue et leurs informations linguistiques associées,
  - Une partie contrastive qui contient les acceptions existant dans d'autres langues, mais pas dans la langue du dictionnaire, ainsi que les informations associées.
- *Le dictionnaire interlingue* contient les acceptions interlingues et leur relations.

Le linguiste définit les structures linguistiques des entrées de chaque dictionnaire monolingue via un langage spécialisé. Il est aussi possible de définir des contraintes de cohérence et d'intégrité sur ces structures, ainsi que des règles de valeurs par défaut. Le linguiste gère les informations des dictionnaires monolingues, à l'aide d'outils de NADIA (éditeur, défauteur, vérificateur de cohérence, …)

Le dictionnaire interlingue est difficile à gérer. Une acception interlingue est créée lorsqu'une nouvelle acception apparaît dans une langue. Le lexique interlingue doit fournir des liens lexicaux entre les différentes langues. Pour cela, les acceptions équivalentes de deux langues différentes doivent être réunies en une seule acception interlingue. Aussi, lorsqu'il ajoute une nouvelle acception dans un dictionnaire, le lexicographe doit vérifier si une acception interlingue équivalente existe dans la base interlingue. Si oui, il va lier cette nouvelle acception à l'acception interlingue. Si non, il doit créer une nouvelle acception interlingue. Une telle gestion de la base interlingue suppose que :

- le lexique interlingue fournisse suffisamment d'informations pour que le lexicographe puisse vérifier l'existence d'une acception interlingue. Cette information doit définir chaque acception de manière non ambiguë.
- cette information soit comprise par tous les lexicographes. Elle doit donc être fournie dans une langue commune.
- le lexicographe qui crée une nouvelle acception fournisse cette information dans la langue commune (qui n'est pas nécessairement sa langue maternelle).

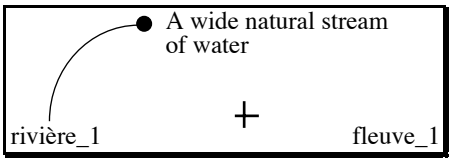
Pour ces raisons, nous avons choisi de confier la gestion de la base interlingue au système. Celui-ci crée une nouvelle acception dans le dictionnaire interlingue lorsqu'une nouvelle acception apparaît dans un dictionnaire monolingue. Il lie un nouveau terme avec une acception interlingue déjà existante lorsque cela est possible.

La détermination de l'existence d'une acception interlingue ou du besoin d'en créer une nouvelle se fait par un "effet de bord". On demande au lexicographe de donner

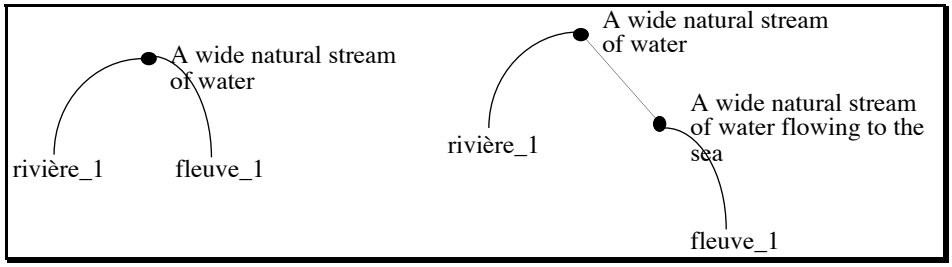
l'équivalent du terme en cours dans une des langues qu'il connaît le mieux (parmi celles de la base). Le système sait ainsi si l'acception interlingue correspondante existe ou non, et crée une nouvelle acception si nécessaire.

Pour illustrer les principes de la gestion du dictionnaire interlingue, prenons l'exemple de l'indexage des deux mots français "rivière" et "fleuve". Supposons que la base lexicale contient déjà un dictionnaire anglais.

Le lexicographe indexe le mot français "rivière". Il sélectionne l'acception qui convient parmi les différentes acceptions de l'équivalent anglais : "river" et la lie au nouveau mot français. Quand il indexera le mot "fleuve", il retrouvera la même acception de "river".

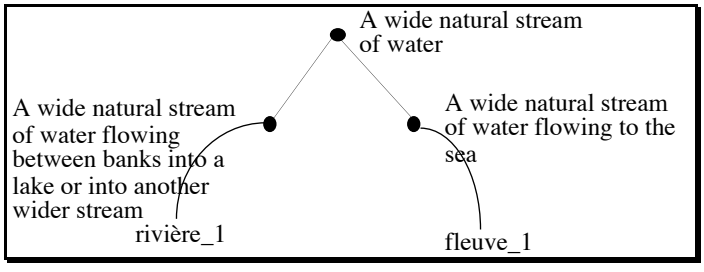


Devant ce problème, le lexicographe pourrait adopter plusieurs choix : lier cette acception au mot "fleuve" ou créer une nouvelle acception pour "fleuve" et la lier à l'acception de "river".



Le premier choix est incorrect, puisque l'acception est déjà liée à "rivière" et que "rivière" n'est pas synonyme de "fleuve". Le second l'est aussi, puisque "fleuve" ne représente pas un sous sens de "rivière".

Le choix correct pour résoudre ce problème est de modifier le lien entre "rivière" et l'acception originale (celle qui correspond à "river") en créant deux nouvelles acceptions (une pour "rivière", l'autre pour "fleuve"). Ces deux nouvelles acceptions sont des sous-acceptions de l'acception originale.



Afin d'assurer une cohérence dans la gestion du dictionnaire monolingue, et pour alléger le travail du lexicographe, le système prend en charge cette gestion en détectant automatiquement les différents problèmes qui peuvent se poser et en proposant des solutions au lexicographe.

III. Justification des choix

1. Analyse de projets existants

1.1. KBMT-89

KBMT-89 (Knowledge Based Machine Translation) était un projet de recherche du Carnegie Mellon University's Center for Machine Translation. Le but de ce projet est de construire un système de Traduction Automatique utilisant une approche basée sur la connaissance. Le système KBMT-89 utilise une représentation de la connaissance du domaine en tant qu'interlangue.

La connaissance acquise pour ce système inclut une ontologie (environ 1500 concepts). Les lexiques contiennent environ 800 unités lexicales japonaises et 900 unités anglaises. On trouvera une description plus détaillée dans [Meyer et al. 1990, Nirenburg 1989].

Les dictionnaires de KBMT-89 sont composés de :

- une base des connaissances (ontologie),
- des lexiques monolingues (japonais et anglais).

La base des connaissances contient un ensemble de concepts, classés en une hiérarchie. Ces concepts servent d'interlangue.

Les lexiques contiennent un ensemble de "super-entrées" (lemmes) décomposés en "entrées" (sens). À chacun de ces sens sont associées des informations linguistiques (catégorie, orthographe, phonologie, morphologie, traits syntaxiques, structures syntaxiques, sémantiques, relations lexicales et pragmatiques).

Chaque sens d'un mot est généralement relié à un concept de la base de connaissance. Un sens peut être relié à une structure interlingue qui n'est pas reflétée par un concept de l'ontologie (une attitude ou une relation).

1.2. EDR

EDR (Electronic Dictionary Research) est un projet japonais visant à la construction d'une base lexicale de grande taille pour différents systèmes de Traduction Automatique. EDR a développé des dictionnaires de mots, des dictionnaires de concepts, des dictionnaires bilingues et des dictionnaires de cooccurrences pour l'anglais et le japonais.

Le dictionnaire de concepts de EDR contient une description des concepts. Cette description se fait au travers de relations entre concepts. EDR utilise 32 types de relations et 5 types d'attributs pour cette description. Les concepts sont classifiés dans une hiérarchie.

Le dictionnaire de concepts est utilisé comme un dictionnaire interlingue. Le dictionnaire de mots fait le lien entre mot et concept, et contient les informations linguistiques associées à chaque mot.

1.3. Multilex

Le projet Européen Multilex (projet DG XIII - ESPRIT) vise à définir des standards pour la construction de base lexicales multilingues. Les langues considérées sont celles de la Communauté Européenne.

Multilex utilise une approche par transfert. Une base lexicale Multilex comporte deux sortes de dictionnaires : des dictionnaires monolingues et des dictionnaires bilingues.

Les dictionnaires monolingues contiennent des unités lexicales (sens) ainsi que l'information linguistique associée. Un langage a été développé pour que le linguiste

puisse définir les structures utilisées. Ces structures doivent être codées sous forme de structures de traits typés.

Les dictionnaires bilingues sont unidirectionnels. Chaque unité bilingue est composée d'une unité lexicale source, d'une unité lexicale cible, d'une condition d'application et d'une règle de transformation lorsque nécessaire.

2. Problèmes rencontrés

Il est intéressant d'analyser les principaux problèmes rencontrés par les projets de bases lexicales multilingues. Tous rencontrent des problèmes liés au codage des informations linguistiques, et les projets utilisant un vocabulaire interlingue butent sur le raffinement des concepts. Enfin, nous parlerons des problèmes introduits par la représentation des connaissances, puis de ceux introduits par une classification des concepts.

2.1. Représentation des informations linguistiques

Le projet EDR représente les informations linguistiques sous forme d'attributs. Les attributs possibles sont définis a priori et ne peuvent être modifiés. L'information linguistique n'est pas très détaillée. Mais les informations présentes sont suffisantes pour être utilisées par des systèmes de Traduction Automatique.

Le projet Multilex est plus souple puisqu'il permet au linguiste de définir la structure linguistique des dictionnaires. Par contre, celui-ci est obligé d'utiliser des structures de traits typées pour coder ses informations lexicales. Bien adaptées pour certains problèmes, les structures de traits typées le sont moins pour coder certaines structures linguistiques telles que les arbres ou les automates.

2.2. Raffinement des concepts

Pour utiliser des concepts dans l'interlangue, on doit répondre à la question : "qu'est-ce qu'un concept ?". Prenons par exemple les trois phrases suivantes :

- Un éléphant apparaît.
- Un éléphant est un animal intelligent.
- L'éléphant est une espèce en danger.

Pour "éléphant", on doit savoir si l'on a affaire à trois concepts différents ou à trois réalisations différentes d'un même concept.

EDR considère que les trois phrases contiennent 3 réalisations différentes d'un même concept (la première phrase parle d'un éléphant en tant qu'individu, la seconde en tant qu'éléphant typique et la troisième, en tant qu'espèce animale).

KBMT-89 a défini quatre critères pour raffiner les concepts. L'un de ces critères déclare que si une unité lexicale a deux ensembles d'attribut grammaticaux incompatibles, alors deux unités lexicales doivent être créées. Les attributs grammaticaux peuvent être morphologiques (ex : nom discret ou non), syntaxiques ou lexicaux (ex : collocations). Mais ce critère n'est pas absolu : il dépend de la taille et de la systématicité des différences entre les deux ensembles.

Dans l'exemple précité, il n'est pas aisé de voir si l'on a affaire à un ou plusieurs concepts. Dans la première phrase, le nom "éléphant" est un nom discret (il est possible de dire "3 éléphants apparaissent"). Dans la seconde et la troisième phrases, le nom éléphant relève du non discret (ex : dire "3 éléphants sont des animaux intelligents" n'aurait pas la même signification). De plus dans la troisième phrase, on ne peut pas utiliser l'article indéfini.

On doit faire un choix entre créer 3 concepts ou un seul. Lorsque le choix est fait (quel qu'il soit), le même choix doit être fait à chaque fois qu'un cas semblable se présente. Lorsque l'on construit une base de taille réelle, plusieurs lexicographes doivent intervenir. Assurer la cohérence des choix devient alors un problème méthodologique complexe.

2.3. Problèmes introduits par la représentation des connaissances

EDR et KBMT-89 ont choisi d'utiliser une représentation des connaissances en tant qu'interlangue.

Ce choix introduit des problèmes méthodologiques lorsque l'on veut construire des bases de taille réelle.

Premièrement, il est coûteux et difficile de décrire les concepts pour une base à grande échelle (pour EDR, qui a décrit avec succès 400000 concepts, il a fallu 1200 hommes-années). Pour être réellement envisageable, une telle description devra se faire par une extraction automatique ou semi-automatique des concepts. EDR a largement utilisé une telle approche (à partir d'un grand corpus).

Lors de cette extraction des connaissances, EDR a utilisé un grand corpus de textes. Pour des raisons de rapidité et de complexité lors de l'acquisition, EDR a dû faire des généralisations hâtives [H. Suzuki, communication personelle]. Prenons par exemple la phrase "un éléphant mange une pomme". Cette phrase dit qu'un certain éléphant est en train de manger une certaine pomme. EDR en extrait 2 relations pour le dictionnaire de concepts :

- Tout éléphant peut être agent du verbe "manger",
- Toute pomme peut être l'objet de l'action "manger".

2.4. Problèmes introduits par une classification des concepts

EDR et KBMT-89 réduisent le nombre de relations entre concepts en les factorisant. Cette factorisation se fait à l'aide d'une hiérarchie et d'un mécanisme d'héritage.

Ainsi, par exemple, dans la hiérarchie de EDR, le fait qu'un oiseau peut voler est codé. Le mécanisme d'héritage permet donc de savoir qu'un moineau, qu'une alouette, qu'une hirondelle, … peut voler.

Par contre cette classification des concepts peut poser des problèmes théoriques et méthodologiques.

Il faut tout d'abord gérer les exceptions. Comment doit-on représenter le fait qu'une autruche est un oiseau et qu'il ne peut voler ? EDR a choisi de déclarer explicitement qu'une autruche ne peut voler via une relation négative [H. Suzuki, communication personnelle]. Cette assertion remplace l'assertion héritée. Hélas, de telle relations négatives doivent être ajoutées avec soin et ne peuvent pas être insérées automatiquement puisqu'elles peuvent introduire des incohérences dans la base.

La modification d'une telle hiérarchie est très délicate, puisque tous les sous-concepts du concept modifié seront affectés.

L'ajout d'une relation à un concept dans la hiérarchie peut s'avérer délicat si celui-ci a de nombreux sous-concepts. En effet, cette relation sera héritée par l'ensemble des sous-concepts. Il faudra donc vérifier qu'il n'y a pas de nouvelles exceptions parmi les sous-concepts.

3. Justification des choix

Les choix que nous avons faits pour l'organisation générale du projet NADIA sont justifiés tant par notre désir d'assurer à terme une certaine compatibilité (au moins en termes d'échange de données) avec les projets précités que par les problèmes qu'ils ont rencontrés.

Multilex utilise une approche par transfert, alors que EDR et KBMT-89 ont choisi une approche interlingue. Afin de pouvoir être compatible avec chacune de ces approches, nous avons choisi l'approche interlingue. Avec une telle approche, il est possible de générer des dictionnaires de transfert, et donc de disposer de dictionnaires compatibles avec ceux de Multilex.

Le projet EDR utilise une structure linguistique relativement figée. Multilex est plus souple, mais ne permet d'utiliser que des structures de traits typés. Pensant que la souplesse est un atout important, nous avons choisi de donner au linguiste la possibilité de définir ses structures linguistiques. Pour augmenter encore cette souplesse , nous n'imposons pas une structure informatique de base pour coder les structures linguistiques. Le système NADIA permet d'utiliser la plupart des structures informatiques les plus utilisées à l'heure actuelle en TALN. Il est aussi possible de mélanger ces différentes structures dans une même entrée, afin de disposer d'une représentation adaptée à chaque problème linguistique. C'est ainsi que nous garantissons une indépendance du système vis-à-vis de la théorie linguistique choisie.

Les deux projets interlingues étudiés utilisent une approche par représentation des connaissances. Nous avons vu que cette approche pose certains problèmes théoriques et méthodologiques. Pour tenter d'apporter une solution à certains de ces problèmes, nous avons choisi l'approche par acceptions interlingues, qui permet d'éviter le problème du raffinement des unités interlingues (ce raffinement devient systématique, puisqu'on utilise des dictionnaires existants comme référence). Elle nous permet aussi de ne pas représenter les connaissances du monde et de réduire ainsi considérablement le coût de la construction d'une base.

Le projet EDR utilise une classification de concepts pour réduire le nombre de relations de description. Comme nous n'utilisons pas de description de concepts, cette classification devient inutile. Nous n'utilisons qu'une sorte de relation entre acceptions : une relation de sur-acception à sous-acception. Celle ci nous permet de coder les phénomènes contrastifs. Il n'est cependant pas question de développer une hiérarchie complète d'acceptions par le biais de cette relation. Celle-ci ne sera utilisée que localement, en cas de problème contrastif.

Conclusion

Nous avons présenté ici les grandes lignes de notre projet NADIA, qui vise à la construction d'un système de gestion de bases de données lexicales multilingues permettant aux utilisateurs de définir des bases multilingues, indépendamment des applications qui utiliseront les données.

Le projet NADIA utilise une approche interlingue originale : l'interlingue par acceptions. Cette approche permet de s'affranchir des problèmes de représentation de connaissances que l'on rencontre souvent dans les projets interlingues.

En donnant la possibilité au linguiste de choisir la (ou les) structure(s) linguistiques et informatiques qu'il désire, nous garantissons une certaine dépendance par rapport à la théorie linguistique sous-jacente de chacune des bases. Nous effectuons ainsi un

nouveau pas vers le partage des données linguistiques en permettant à différentes théories linguistiques de cohabiter sur une seule et même plateforme.

Le prototype de NADIA décrit ici est en cours de développement. Après ce prototype, plusieurs voies s'offriront à nous. Il nous sera possible d'améliorer les différents outils, et notamment les outils d'import/export qui permettront encore une fois un partage de données linguistiques entre différents systèmes.

Une nouvelle voie de recherche se dessine dès à présent. Nous faisons l'analogie entre un dictionnaire et un document structuré. De la même manière qu'un document est une suite de chapitres (ayant un titre) composés de parties contenant elles-mêmes différents paragraphes, un dictionnaire est une suite d'articles (ayant une forme d'entrée) composés de différents sens, contenant eux-mêmes différentes informations linguistiques.

Une telle analogie nous permet d'envisager, comme pour un document structuré, de définir différentes vues d'une base lexicale. Ainsi, un lexique quadrilingue en colonnes pourra être une vue d'une base lexicale quadrilingue. Un fichier SGML pourra aussi être une vue d'une base lexicale. Cela nous permettra de confondre les problèmes d'interface et d'import/export. Nous pourrons ainsi étudier comment définir différentes théories linguistiques en tant que différentes vues d'une seule et même structure.

Nous espérons ainsi faciliter de plus en plus, non pas seulement le partage de données linguistiques, mais aussi la communication entre différentes "écoles" linguistiques, qui pourraient mettre en commun, sur une seule base, les différents aspects qui les intéressent, dans leurs codages préférés.

Références

Boitet C., Guillaume P. et Quezel-Ambrunaz M. (1982). *ARIANE-78 : an integrated environment for automatic translation and human revision*, COLING-82, Juillet 1982 : pp. 19-27.

Boitet C., Guillaume P. et Quezel-Ambrunaz M. (1985). *A case study in software evolution : from ARIANE-78.4 to ARIANE-85*, Theoretical and Methodological Issues in Machine Translation of Natural Languages, 14-16 Août 1985**, vol. 1/1** : pp. 27-58.

EDR (1988). *Electronic Dictionary Project*, Japan Electronic Dictionary research institute Ltd, novembre 1988.

EDR (1990). *EDR Technical Reports, an overview of the electronic dictionaries*, EDR, japan electronic dictionaries research institute Ltd, Technical reports n° TR-024, TR-025, TR-026, TR-027, TR-029, 176 p.

Farwell D., Guthrie L. et Wilks Y. (1992). *The Automatic Creation of Lexical Entries for a Multilingual MT system*, COLING-92, 20-28 Juillet 1992**, vol. 2/4** : pp. 532-538.

Gates D. et al. (1989). *Lexicons,* Machine Translation**, vol. 4**(1) : pp. 67-112.

Meyer I., Onyshkevych B. et Carlson L. (1990). *Lexicographic Principles and Design for Knowledge-Based Machine Translation*, Carnegie Mellon University, Technical Report n° CMU-CMT-90-118, 13 Août 1990, 66 p.

Nirenburg S. (1989). *Knowledge-based machine translation,* Machine Translation**, vol. 4**(1) : pp. 5-24.

Nirenburg S. et Defrise C. (1990). *Lexical and Conceptual Structure for Knowledge-Based Machine Translation*, ROCLING III, 20-22 Août 1990**, vol. 1/1** : pp. 105-130.

-o-o-o-o-o-o-o-o-o-o-