Gestion des terminologies riches : L'exemple des acronymes

Ying Zhang, Mathieu Mangeot

Gestion des terminologies riches : l'exemple des acronymes

Ying ZHANG[1] *et Mathieu MANGEOT*[1]
(1) GETALP-LIG, 41, rue des Mathématiques BP53 38041 Grenoble Cedex 9
ying.zhang@imag.fr, mathieu.mangeot@imag.fr

RÉSUMÉ

La gestion des terminologies pose encore des problèmes, en particulier pour des constructions complexes comme les acronymes. Dans cet article, nous proposons une solution en reliant plusieurs termes différents à un seul référent via les notions de pivot et de prolexème. Ces notions permettent par exemple de faire le lien entre plusieurs termes qui désignent un même et unique référent : Nations Unies, ONU, Organisation des Nations Unies et onusien. Il existe Jibiki, une plate-forme générique de gestion de bases lexicales permettant de gérer n'importe quel type de structure (macro et microstructure). Nous avons implémenté une nouvelle macrostructure de ProAxie dans la plate-forme Jibiki pour réaliser la gestion des acronymes.

ABSTRACT

Complex terminologies management – the case of acronyms

Terminology management is still problematic, especially for complex constructions such as acronyms. In this paper, we propose a solution to connect several different terms with a single referent through using the concepts of pivot and prolexeme. These concepts allow for example to link several terms for the same referent: Nations Unies, ONU, Organisation des Nations Unies and onusien. Jibiki is a generic platform for lexical database management, allowing the representation of any type of structure (macro and microstructure). We have implemented a new macrostructure ProAxie in the Jibiki platform to achieve acronym management.

MOTS-CLÉS : base lexicale multilingue, macrostructure, Jibiki, Common Dictionary Markup, Proaxie, Prolèxeme
KEYWORDS : multilingual lexical database, macrostructure, Jibiki, Common Dictionary Markup, Proaxie, Prolexeme

1 Introduction

Cet article concerne la gestion de terminologies multilingues. Le problème abordé dans cet article est celui de l'association de plusieurs termes d'une même langue à un même référent : « Jean-Paul II » et « Karol Jozef Wojtyla » en français, ou en anglais « John Paul II » et « Karol Jozef Wojtyla ». De même, certains liens évoluent avec le temps : le pape désignait « Jean-Paul II » en 2004 et « Benoît XVI » en 2012. Des pays parlant la même langue (p. ex : France et Suisse romande) peuvent également utiliser des mots différents pour le même concept. Par exemple, « chien renifleur » et « chien drogue ».

Inversement, le même terme peut désigner des concepts différents : dans la province de langue allemande de Bolzano en Italie, le « Landeshauptmann » est le président du conseil provincial, avec des compétences beaucoup plus limitées que le « Landeshauptmann » autrichien, qui est à la tête de l'un des États (Land) de la fédération autrichienne. Pour la gestion des acronymes, un terme et son acronyme peuvent par exemple désigner le même référent. Dans un contexte multilingue, la difficulté est d'établir une correspondance entre ces termes. L'article introduisant la notion de prolexème [Tran, 2006] présente le problème des termes ayant des acronymes dans certaines langues, mais pas dans d'autres. Dans le projet Prolexbase , Tran [Tran, 2006] considère le prolexème comme le regroupement de lemmes associés aux différentes formes d'un nom propre qui apparaissent dans les différents textes d'une langue donnée. Par exemple, en français, Prolexbase regroupe dans le même prolexème « organisation des nations unies »[1], « Nations unies », « ONU » et « onusien »[2]. En anglais, Prolexbase regroupe « United Nations » et son acronyme « UN ».

Quelles solutions mettre en place de façon à choisir, pour un terme donné dans une langue donnée, le meilleur équivalent dans une langue cible ? Cette recherche est motivée par un besoin réel d'une entreprise dans la gestion de sa terminologie multilingue.

Le but principal de notre travail a été de définir un cadre théorique composé d'une nouvelle macrostructure basée sur des concepts existants et sur la définition de nouveaux concepts. Ce cadre a ensuite été validé par une expérimentation pratique à l'aide d'un outil générique de gestion de bases lexicales.

Cet article est organisé de la façon suivante. Dans la section 2, nous présentons les macrostructures préconisées pour les données. La section 3 présente les outils utilisés et l'implémentation de la macrostructure. La section 4 présente les résultats de l'implémentation et de l'utilisation. Enfin, nous concluons et donnons quelques perspectives pour la gestion de terminologies riches.

2 Données : choix de la macrostructure

Lors de toute discussion scientifique, il est primordial de bien s'entendre sur les termes utilisés. C'est pourquoi nous commencerons par définir les termes et concepts principaux que nous utiliserons par la suite.

Un *dictionnaire* est composé d'un ou plusieurs *volumes* reliés entre eux par des *liens* qui sont le plus souvent des *liens de traduction*. Un volume est un ensemble *d'articles* comportant des *mots-vedettes* de la même langue. Un article comporte au moins un mot-

1 Nous avons repris exactement la terminologie française de Prolexbase et les concepts.

2 Mettre « onusien » dans ce groupe est sans doute une erreur.

vedette et le plus souvent d'autres informations (prononciation, classe grammaticale, définition, exemples, etc.). La structure des articles est appelée *microstructure*. L'organisation des volumes qui composent la structure d'un dictionnaire est appelée *macrostructure*. La macrostructure la plus simple est celle d'un dictionnaire monolingue ne comportant qu'un seul volume. Pour les dictionnaires bilingues langue A (LgA) ⟷ langue B (LgB), on trouve souvent des macrostructures avec deux volumes : un volume LgA ⟶ LgB et un volume miroir LgB ⟶ LgA. Ces macrostructures constituent l'essentiel des dictionnaires imprimés. L'avènement de l'outil électronique permet de s'abstraire des contraintes liées à l'impression, notamment la représentation restreinte à deux dimensions. On peut maintenant concevoir des macrostructures plus complexes utilisant par exemple des volumes pivot. Le dictionnaire devient alors une base lexicale à plusieurs dimensions d'où il est possible d'extraire des vues spécifiques permettant de retrouver le format initial des dictionnaires imprimés.

Nous détaillerons dans la suite deux macrostructures de ce type.

2.1 Macrostructure pivot

Le projet Papillon, lancé en 2000 [Tomokiyom et al., 2001], a eu pour but de construire une ressource lexicale pour plusieurs langues dont au moins l'anglais, le français et le japonais. Les macrostructures bilingues traditionnelles obligeant à construire un dictionnaire par couple de langues, le nombre de dictionnaires croît de manière triangulaire par rapport au nombre de langues en présence. Cette solution devient rapidement ingérable. Il fallait donc en trouver une nouvelle, un *dictionnaire multilingue à structure pivot* : un volume monolingue pour chaque langue et un volume pivot (ou volume interlingue) au centre regroupant les liens entre les articles [Sérasset et Mangeot, 2001]. La microstructure des articles monolingues est basée sur le concept de *lexie* défini dans la lexicographie explicative et combinatoire [MEL'ČUK et al., 1995] issue de la théorie sens-texte. Chaque article décrit une lexie. Une lexie est une unité lexicale (sens de mot) qui est représentée soit par un lexème (regroupement de mots-forme), soit par une locution nominale.

Chaque lexie est reliée par un lien interlingue à une *axie* (ou *acception interlingue*). Les axies sont contenues dans le volume pivot. Chaque axie regroupe les équivalents dans plusieurs langues d'une même lexie (ou sens de mot).

Les concepts d'axie et de structure pivot ont été définis pour le projet Papillon et ensuite repris dans la norme Lexical Markup Framework [Francopoulo et al., 2009].

2.2 Macrostructure ProAxie

Cette macrostructure a pour but de résoudre le problème de relier plusieurs termes qui désignent un même et unique référent. Pour la gestion des acronymes, les liens riches sont plus répandus et plus complexes. Pour implémenter la gestion des acronymes, nous proposons une nouvelle macrostructure avec deux notions: *proaxie*[3] et *prolexème* [Tran, 2006]. Nous avons également besoin des concepts d'axie et de lexie, qui ont été

3 ProAxie (A en majuscule) est le nom de macrostructure. Proaxie (a en minuscule) est le nom de concept.

présentés au §2.1. Voir la figure 1.

2.2.1 Prolexème

Il y a un seul volume de prolexèmes pour chaque langue. Dans ce volume, les prolexèmes regroupent les lexies qui représentent le même sens sémantique mais dont la réalisation syntaxique est différente (forme de surface, classe grammaticale, etc.). Les liens bidirectionnels entre les lexies et leurs prolexèmes sont marqués avec une étiquette (alias, acronyme, dérivation, définition, etc.). Par exemple, l'entrée de type prolexème « fra.organisation_des_nations_unies.1 » est reliée à l'entrée de type lexie « ONU » par un lien étiqueté « acronyme », à « nations unies » par un lien étiqueté « alias », à « onusien » par un lien étiqueté « dérivation », et à « organisation des nations unies » par un lien étiqueté « définition ». Ce lien n'est pas la définition lexicographique du prolexème, mais caractérise seulement le terme préféré pour le décrire.

2.2.2 Proaxie

Il y a un seul volume de proaxies dans un dictionnaire. Les proaxies relient les prolexèmes de langues différentes qui ont le même sens. Prenons l'exemple d'un dictionnaire trilingue : français, anglais et chinois. L'entrée de type proaxie « proaxie.united_nations.1 » relie l'entrée « fra.organisation_des_nations_unies.1 » du volume des prolexèmes français, l'entrée « eng.united_nations.1 » du volume des prolexèmes anglais, et l'entrée « zho.联合国.1 » du volume des prolexèmes chinois. Les liens entre l'entrée de proaxie et les entrées de prolexèmes sont bidirectionnels.

2.2.3 Conception globale

Dans cette macrostructure, nous avons deux couches : une couche de base et une couche « Pro ». Dans la couche de base, nous avons deux types de volume : volume des

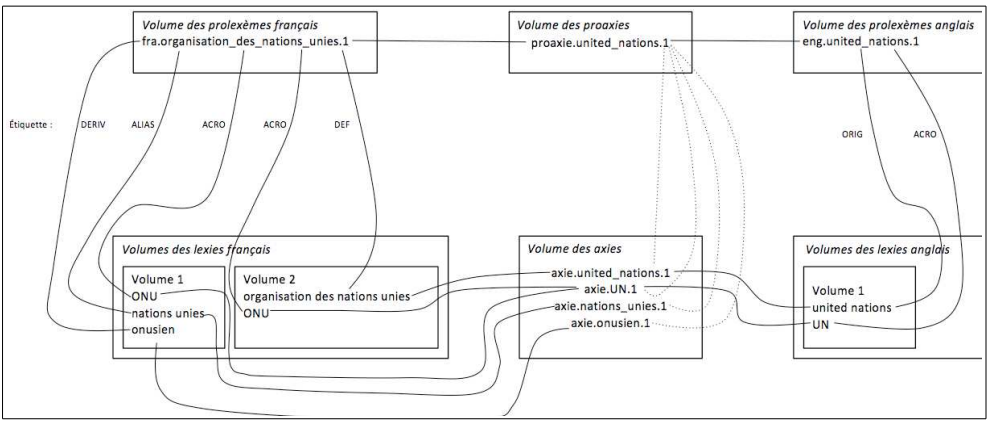

Figure 1 : Exemple de ProAxie

lexies et volume des axies. Dans la couche « Pro », nous avons également deux types de volume : volume des prolexèmes et volume des proaxies.

Grâce au volume d'axies, nous pouvons relier les lexies qui se correspondent

exactement, comme l'acronyme français « ONU » relié avec l'acronyme anglais « UN ». Grâce à la couche « Pro », nous pouvons proposer en traduction des lexies des langues cibles de même sens, comme indiqué dans la figure 1.

Les étiquettes portées par les liens ont pour but de permettre de proposer les meilleures traductions. Par exemple, le japonais « 国際連合 » est la lexie de même sens que « Organisation des Nations Unies », son acronyme est « 国連 ». Cet acronyme utilise le premier et le troisième kanji, ce qui est différent des initiales de la lexie de définition. Il existe peut-être une langue qui a deux acronymes, l'un correspondant à l'acronyme des initiales, l'autre correspondant à une sélection de caractères ou de mots. Donc, nous avons décidé de ne pas lier ces deux types d'acronymes avec une même axie, voir la figure 2.

Considérons les liens de la lexie « ONU » du français vers l'anglais, vers le japonais et vers le chinois. Nous proposons trois niveaux de traduction classés selon la précision obtenue :

- Vers l'anglais : « ONU »→« UN ». Le système trouve une lexie directe par le volume des axies. C'est le premier niveau de traduction et le plus précis.
- Vers le japonais : « ONU »→«国連». Le système cherche le lien dans le volume des prolexèmes français avec l'étiquette « Acro ». Puis il trouve le lien dans les proaxies, ensuite il suit le lien de prolexème japonais, et enfin il arrive au volume des lexies japonaises, et trouve une lexie en suivant un lien étiqueté « Acro ». Donc la lexie proposée au deuxième niveau de la langue cible est cet acronyme. Le deuxième niveau de traduction comprend toujours le premier niveau de traduction. C'est-à-dire que « ONU » et « UN » sont accédés avec la même étiquette « Acro », donc le lien « ONU »→ « UN » correspond également au deuxième niveau de traduction.
- Vers le chinois : « ONU »→« 联合国 ». Le système trouve les lexies par prolexème et proaxie sans étiquette correspondante. Ces lexies proposées constituent le troisième niveau, le moins précis. Le troisième niveau de traduction comprend les niveaux précédents.

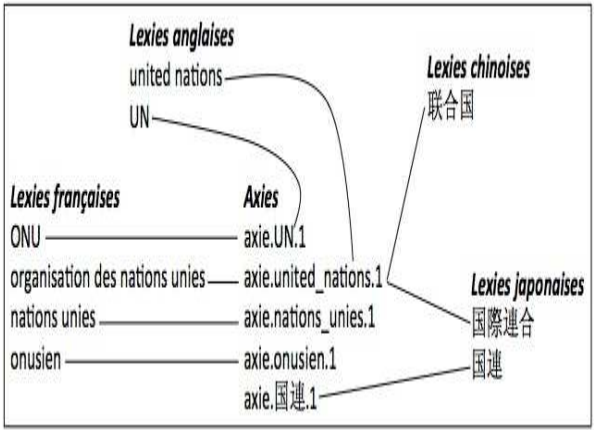

Figure 2 : Liens entre les lexies et les axies de ProAxie

La quantité de lexies de résultat augmente suivant les niveaux de traduction, du

premier vers le troisième. Par exemple, on traduit le terme « ONU » vers l'anglais, le chinois et le japonais. Le premier niveau de traduction est la lexie anglaise « UN ». Le deuxième niveau de traduction est la lexie anglaise « UN » et la lexie japonaise « 国連 ». Le troisième niveau de traduction comprend les lexies anglaises « UN » et « United Nations », les lexies japonaises « 国連 » et « 国際連合 », et la lexie chinoise « 联合国 ».

Dans certaines situations, une base lexicale (un dictionnaire) a plusieurs volumes pour une seule langue. Par exemple, lorsqu'il y a plusieurs versions d'édition ou que la ressource lexicale est créée par un système de traduction automatique, on trouvera un volume provenant de Systran, un volume de Google, un volume d'IATE, etc. Notre macrostructure permet de gérer plusieurs volumes dans une même langue, (voir la figure 3). Étant donnée une langue, il existe un ou plusieurs volumes de lexies, mais un seul volume de prolexèmes. Pour un dictionnaire, il y a un seul volume de proaxies et un seul volume d'axies. Les entrées des lexies sont reliées aux entrées de type prolexème et de type axie. De plus, les prolexèmes sont reliées aux proaxies, et vice-versa.

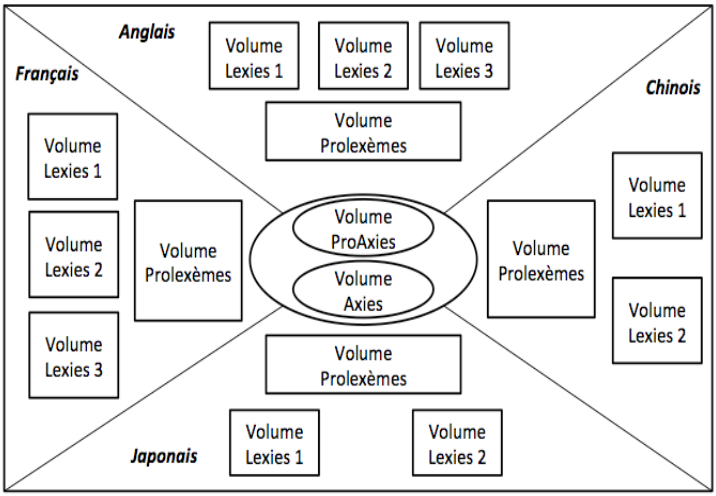

Figure 3 : Macrostructure de ProAxie

3 Outils nécessaires : plates-formes de manipulation

3.1 Plate-forme Jibiki version 1

Pour implémenter la macrostructure de ProAxie, nous avons utilisé la plate-forme Jibiki. Elle permet la construction de sites Web contributifs dédiés à la construction de bases lexicales multilingues. Cette plate-forme a été développée principalement par Mathieu Mangeot [Mangeot et Chalvin, 2006] et Gilles Sérasset [Sérasset et Mangeot, 2001]. Elle a été utilisée dans divers projets (projet LexALP, projet Papillon, projet GDEF, etc.). Le code est disponible en source ouvert et téléchargeable gratuitement par SVN sur ligforge.imag.fr. Avec cette plate-forme, on peut faire les manipulations d'import, export, édition, modification et recherche dans des bases lexicales. On peut aussi gérer les contributions.

Jibiki est une plate-forme générique, elle permet de traiter presque toutes les ressources lexicales de type XML en utilisant différentes microstructures et macrostructures. On utilise des pointeurs CDM (Common Dictionary Markup) [Mangeot, 2002] pour gérer n'importe quel type de microstructure sans la modifier. Les pointeurs sont utilisés également pour indexer des parties d'information spécifiques et permettre ensuite une recherche multi-critères. Cette structure est stockée dans un fichier de métadonnées sous forme XML. Pour chaque pointeur CDM, on indique le chemin XPath vers l'élément correspondant dans la microstructure XML. Les liens de traduction sont à ce stade traités comme des pointeurs CDM classiques.

La version 1 de Jibiki présentait plusieurs limitations. Les liens de traduction étaient traités avec des pointeurs CDM, comme des éléments d'information classiques. Ces liens étaient simples. Il n'y avait pas de possibilité de décrire des liens entre plusieurs volumes différents. Il n'était pas non plus possible d'ajouter des attributs (poids, étiquette, volume cible, etc.) sur les liens. Nous avons remedié à ces defauts dans Jibiki-2. Jibiki est utilisé pour plusieurs macrostructures. Pour chaque macrostructure, il a été nécessaire de recoder une partie du programme pour réaliser les différents types de liens.

3.2 Gestion des liens riches : Jibiki-2/Pivax

La gestion des liens riches correspond aux liens avec des attributs, comme volume cible, poids, type, langue, étiquette libre, etc. Pour réaliser l'implémentation de liens riches, nous avons séparé le module de traitement des liens de celui des autres pointeurs CDM.

La réalisation informatique est basée sur deux algorithmes. Le premier collecte les liens, le deuxième construit le résultat. Plus précisément, le premier recherche tous les liens possibles dans l'ensemble des liens riches de tous les volumes pour une entrée recherchée. Le deuxième algorithme réalise les étapes suivantes : (1) chercher les liens vers les axies puis vers les lexies cibles ; (2) chercher les liens vers les prolexèmes de la langue source puis vers les proaxies, vers les prolexèmes des langues cible, et à la fin vers les lexies cibles ; (3) traiter l'étiquette ; (4) trier et afficher.

4 Résultats préliminaires

Nous avons séparé les trois niveaux de traduction pour afficher les résultats de recherche dans Jibiki : (1) traduction directe par axie, (2) traduction par prolexème et proaxie avec la même étiquette, (3) traduction par prolexème et proaxie sans étiquette.

Pour faciliter la lecture, nous affichons l'étiquette, la langue et le mot-vedette au 1er et au 2e niveau. Nous affichons tous les détails (phrases exemples, définitions, partie du discours, etc.) au 3e niveau, y compris les lexies du même prolexème de la langue source. Enfin, nous n’affichons pas la traduction au 2e niveau si elle a déjà été trouvée et est déjà affichée au 1er niveau.

4.1 Scénario 1 : terme « UN » de l'anglais vers toutes les langues

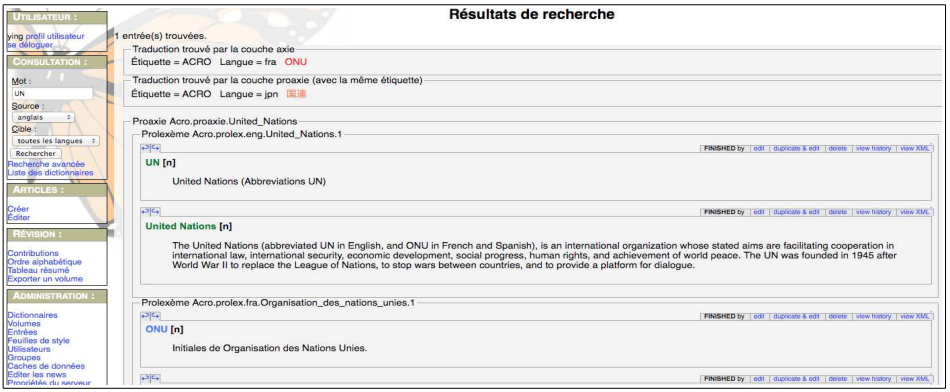

FIGURE 4 - terme « UN » de l'anglais vers toutes les langues

Lexies trouvés en théorie : le premier niveau de traduction est la lexie française « ONU ». Le deuxième niveau de traduction comprend la lexie française « ONU » et la lexie japonaise « 国 連 ». Le troisième niveau de traduction comprend les lexies françaises « ONU », « Nations Unies », « onusien » et « Organisation des Nations Unies », la lexie chinoise « 联合国 » et les lexies japonaises « 国際連合 » et « 国連 ».

Lexies affichées par l'interface : Le premier niveau de traduction est la lexie française « ONU ». Le deuxième niveau de traduction est la lexie japonaise « 国連 ». Le troisième niveau de traduction comprend toutes les lexies : les lexies françaises « ONU », « Nations unies », « onusien » et « Organisation des Nations Unies », la lexie chinoise « 联合国 », les lexies japonaises « 国際連合 » et « 国連 », et les lexies anglaises « UN » et « United nations ».

4.2 Scénario 2 : terme « onusien » du français vers toutes les langues

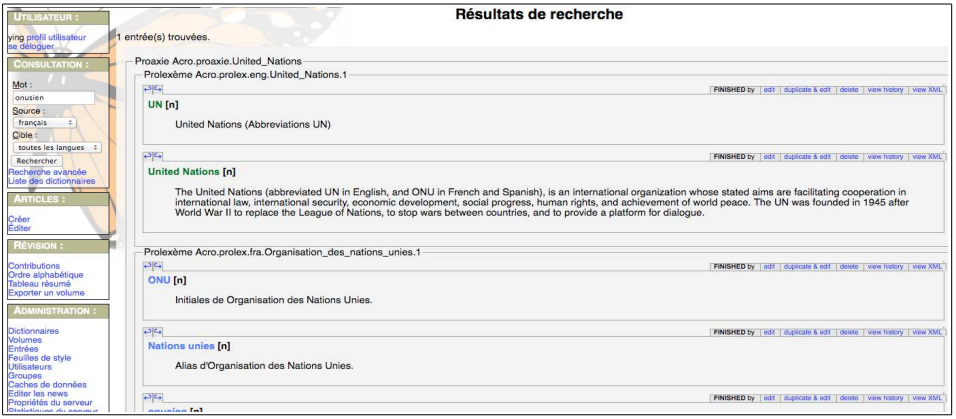

FIGURE 5 - terme « onusien » du français vers toutes les langues

Lexies trouvés en théorie : le premier et le deuxième niveau de traduction sont vides. Le troisième niveau comprend les lexies anglaises « UN » et « United nations », la lexie chinoise « 联合国 » et les lexies japonaises « 国際連合 » et « 国連 ».

Lexies affichées par l'interface : le premier et le deuxième niveau de traduction sont vides. Le troisième niveau comprend toutes les lexies.

5 Conclusion et perspectives

Nous avons présenté la gestion des terminologies avec liens riches en utilisant un exemple d'acronyme (ONU) de nom propre. Nous avons repris les concepts de lexie, d'axie, de prolexème, et introduit le concept de proaxie pour produire la macrostructure de ProAxie. Dans cette macrostructure, une étiquette est utilisée pour relier les lexies et leurs variantes. Nous avons implémenté la solution de la macrostructure ProAxie dans la plateforme Jibiki en utilisant la nouvelle Jibiki-2/Pivax, et créé trois niveaux de traduction en théorie et en affichage.

Concernant les données, la base actuelle est une preuve de concept qui comporte quelques exemples issus de ProlexBase. Nous souhaitons tester cette solution en passant à l'échelle sur de grosses bases telles que la CJK (chinois, japonais, coréen, arabe) avec 24 millions d'entrées ou l'Unified Medical Language System[4] avec 5 millions de termes.

Dans l'avenir, nous souhaitons faire évoluer la macrostructure de ProAxie pour prendre en compte d'autres types de synonymie, et transposer le concept de prolexème pour que cette solution puisse être utilisée dans un autre domaine linguistique. Par exemple, pour une ressource lexicale comprenant des textos, en français « A+ » correspond à « À plus » avec une étiquette « texto », et en anglais « L8R » correspond à « later » avec l'étiquette « texto ».

Nous prévoyons de prendre en compte également les quatre variations du diasystème basé essentiellement sur ce que Eugenio Coseriu propose [Tran, 2006] : diachronique (variété dans le temps), diaphasique (variété concernant les finalités de l'emploi), diatopique (variété dans l'espace) et diastratique (variété relative à la stratification socio-culturelle). Nous voudrions enrichir la notion d'étiquette selon cette théorie.

6 Références

TRAN, M. (2006). Prolexbase : Un dictionnaire relationnel multilingue de noms propres : conception, implémentation et gestion en ligne. *Thèse de doctorat d'informatique*, Tours, pages 54-57.

SÉRASSET, G. et MANGEOT, M. (2001). Papillon Lexical Database Project: Monolingual Dictionaries and Interlingual Links. *In Proc. Of NLPRS 2011*, Tokyo, pages 119-125.

TOMOKIYOM, M., MANGEOT, M. et PLANAS, E., (2001). Papillon: a Project of Lexical Database for English, French and Japanese, using Interlingual Links. *In Actes de JST 2001*, Tokyo, 3 p.

4! http://www.cjk.org; http://www.nlm.nih.gov/research/umls/

MEL'ČUK, I., CLAS, A. et POLGUÈRE, A. (1995). Introduction à la lexicologie explicative et combinatoire. *Livre,* 256 p.

FRANCOPOULO, G., BEL, N., GEORGE, M., CALZOLARI, N., MONACHINI. M., PET, M. et SORIA, C. (2009). Multilingual resources for NLP in the lexical markup framework (LMF). *In journal de Language Resources and Evaluation, March 2009, Volume 43,* pages 55-57.

MANGEOT, M., et CHALVIN, A. (2006). Dictionary Building with the Jibiki Platform : the GDEF case. *In Actes de LREC 2006*, Genoa, pages 1666-1669.

MANGEOT, M. (2002). An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language. *In Actes de LREC 2002*, pages 37-44.