Romanica Cracoviensia
11 / 2011

Sonia Szramek-Karcz

Université de Silésie
de Katowice

LA DESCRIPTION DES NOMS DANS L'EUROWORDNET ET L'APPROCHE ORIENTÉE OBJETS

0. INTRODUCTION

Contrairement à ce que nous pensions en entreprenant nos recherches sur la comparaison des deux bases de données lexicales susmentionnées, il s'est avéré que la question ne porte pas sur le *comment*, mais bien sur le *si* l'on peut comparer l'Approche Orientée Objets (Banyś 2002a, 2002b), et l'EuroWordNet (http://www.illc.uva.nl/uroWordNet), vu les nombreuses différences qui les séparent. C'est bien cette interrogation qui sous-tendra notre propos illustré par des exemples de la description des noms proposés dans les deux bases.

Nous tenterons d'y répondre en trois points: premièrement en analysant les origines et les buts assignés lors de la création des deux bases de données, deuxièmement en comparant leurs unités de description et troisièmement en observant la méthode choisie pour atteindre le cap de la bonne traduction (de l'équivalent adéquat) en langue cible.

Ces trois points vont nous permettre d'étayer l'hypothèse selon laquelle la comparaison entre l'Approche Orientée Objets (désormais : AOO) et l'EuroWordNet (désormais : EWN) présente quelques difficultés ce qui ouvre la discussion sur leur utilité pour la traduction automatique.

Précisons que par l'AOO nous comprenons la base des données lexicales réunies et organisées selon les principes de l'AOO, de plus par EWN nous entendons les données contenues dans la base lexicale créées selon le projet de l'EWN.

1. LA CRÉATION DES DEUX BASES DES DONNÉES

Premièrement, l'AOO et l'EWN ont des origines différentes. Les buts qui leur ont été assignés lors de leurs créations diffèrent également.

L'AOO a été conçue pour servir à la TA. C'est une proposition de stockage des informations lexicales permettant de créer une base de données lexico-sémantiques dont l'architecture simple, flexible et ouverte aux modifications serait aisément réutilisable et enrichie dans l'avenir sans nuire à son intégrité. À l'exception du lien dialectique entre les unités (Banyś 2002a, 2002b.) vu à la surface et rendu par les opérateurs et les attributs de l'objet en question, l'approche prend aussi en compte les informa-

tions sémantiques contenues entre autres dans la relation partie-tout, l'héritage sémantiques montré dans les hiérarchies des classes et des domaines. Mais il ne s'agit pas seulement de contenir dans une base de données les informations explicites et implicites (ce dernier point étant un défi et sujet à moult discussions, recherches et publications) mais aussi de les présenter dans une structure flexible au point de les rendre utiles et réutilisables, aptes aux modifications à effectuer dans un champs donné sans détruire la construction de la base de données, ni fausser la toile des relations et des dépendances des unités qui la construisent. L'objet, comme le nom de l'approche l'indique, constitue l'unité de description dont la définition sera présentée dans la section suivante.

L'EuroWordNet est un projet qui s'est étalé sur 36 mois, financé par la Commission Européenne (Vossen, Díez-Orzas, Peters 1997 ; EuroWordNet 2001 http://www.illc.uva.nl/uroWordNet, 20.06.2010) qui a envisagé de créer des WordNets nationaux contenant des relations sémantiques de base pour plusieurs langues européennes (anglais, néerlandais, espagnol et italien et en 1998 allemand, suédois, français, tchèque, estonien) dans le but de créer une base de données lexicales multilingue. Même si le financement du projet européen est terminé, les Wordnets nationaux sont aujourd'hui élaborés pour une cinquantaine de langues dont la liste mise à jour par la Global WordNet Organization se trouve à l'adresse suivante : http://www.globalwordnet.org/gwa/wordnet_table.htm (20.06.2010). L'idée de connecter les langues afin de permettre une traduction des termes d'une langue à l'autre est née avec le projet lui-même.

Le projet l'EWN se base sur le WN américain et nous devons nous souvenir que cette base électronique de données lexicales organisée en un ensemble de réseaux sémantiques n'est plus aujourd'hui ce qu'elle était en 1985, date officielle de la création de Wordnet. À l'origine, l'idée n'était pas de construire un lexique complet mais d'identifier les principaux nœuds lexicaux et d'explorer les différentes sortes de relations sémantiques qu'ils entretiennent. La théorie testée, stipulait que s'il y avait un schéma correct des relations lexicales, la définition en découlerait. Il semblait alors redondant d'introduire les définitions dans le réseau des relations sémantiques. Mais d'un simple moteur de recherche, WordNet est devenu une base lexicale autonome à laquelle les définitions ont été ajoutées pour faciliter le travail des linguistes et agrandir les possibilités d'exploitation de la base (aujourd'hui, les travaux sur ImaggeNet se poursuivent (http://www.image-net.org/ [20.06.2010], Deng at all. 2009)). Depuis, il ne cesse de se développer, sa version 3.0 contient environ 80,000 synsets de noms, c'est-à-dire d'ensembles de synonymes. Le WordNet a été divisé en quatre réseaux sémantiques distincts pour les noms, verbes, adjectifs et adverbes (Fellbaum 1998). C'est la description des noms qui nous intéresse aujourd'hui car elle permet une comparaison aisée avec la proposition d'une base de données orientée objets.

L'AOO a commencé ses premières descriptions en 2001, les travaux sur le EWN dans le cadre du projet de la Commission Européenne concernent les années 1996––1999 mais les principes remontent au WN donc aux années 80 du siècle précédant. L'écart dans le temps dans la création et des buts assignés à des deux bases de données est plus important que cela ne peut paraître.

2. LES UNITÉS DE LA DESCRIPTION

Les unités de la description dans l'AOO et le EWN sont respectivement l'objet et le synset.

Dans l'optique orientée objets, la direction d'analyse est inversée par rapport à la description des sens des mots appliquée dans le cadre des structures prédicats-arguments où l'on part de la fonction propositionnelle (prédicat) pour arriver à ses arguments (objets) qui saturent les positions ouvertes par ces premiers (Karolak 1984, Banyś 1981, 1983, 1984) ; autrement dit nous partons d'un objet et cherchons les prédicats qui peuvent lui être assignés. Ces prédicats sont répartis en attributs (les adjectifs et les constructions N(Prép)N) et en opérations (les verbes) qui pour des raisons d'organisation descriptive des objets se trouvent divisés en opérateurs constructeurs, manipulateurs et accesseurs. Les constructeurs comme leur nom l'indique, construisent la classe d'objets en question ou construisent la situation où la classe d'objets n'apparaît pas, les accesseurs fournissent les informations sur le comportement et la structure de la classe, les manipulateurs par contre effectuent toutes sortes d'opérations sur la classe d'objets ou que la classe d'objets peut effectuer. La description dans les deux sens est importante mais il fallait bien choisir une des options possibles et ce sont les objets qui ont été choisis comme le pivot du système.

Dans dans l'AOO, l'objet est défini par ses opérations et attributs, son statut est fonctionnel et ses caractéristiques ontologiques ne sont pas prises en compte comme c'est le cas de la classification dans EuroWordNet.

Les objets sont ensuite regroupés en classes d'objets, ce qui veut dire que la classe d'objets est un ensemble d'objets qui partagent les mêmes opérations et (ou) les mêmes attributs. La notion se rapproche, sans être identique, des classes d'objets de G. Gross (Gross 1992, 1994a,b, 1995a,b ; Le Pesant, Mathieu-Colas 1998 : 7–33). Une classe d'objets est définie par les opérations qu'elle admet et une même opération peut bien constituer un élément de faisceau définitionnel de classes d'objets différentes. Il est facile de s'imaginer que l'opérateur « travailler » va appartenir aux classes différentes suivant les exemples de son emploi: machine travaille, vin travaille, argent travaille, goutte travaille, etc.

Dans l'AOO on distingue autant de classes d'objets que d'ensembles d'opérations et d'attributs. Dans le cas de <gendarme>, nous avons affaire à trois classes différentes : personne chargée du maintien de l'ordre et de la sécurité publique <professions>, épointement rocheux situé sur une arête, pouvant constituer un obstacle à la progression des alpinistes <obstacles>, saucisse séchée et fumée, vendue par paire, de couleur brune dorée <aliments>.

Les objets constituent une classe d'objets, mais la réalisation concrète d'un objet <chirurgien> comme par exemple *Monsieur Duval* est son instance, une instance d'objet. Nous retrouvons les instances des objets, parmi les hyponymes du bas de la hiérarchie du EWN.

Dans l'AOO les classes d'objets sont agencées dans une structure où chacune d'elle possèdent sa super-classe et sa sous-classe. Au somment de cette hiérarchie se trouvent les classes conceptuelles de WordNet appelés « unique beginners », ce qui a permis de diviser les tâches des lexicographes et d'organiser le travail de description des objets

dans l'AOO. C'est le seul point commun entre les unique beginners et les classes d'objets de l'AOO.

Les classes hypo et hypéronymes dans l'AOO (comme les synsets dans le WN) sont désignées suivant la relation X EST-UNE SORTE DE Y. Les hiérarchies lexicales de type IS-A-KIND-OF sont largement utilisées pour représenter le savoir (Sowa 2000). Elles permettent d'éviter des cercles vicieux « vacuous circles » (Miller et al. 1990 : 247), définir idem per idem n'est plus possible. L'avantage de l'application des hiérarchies lexicales (dans le WN comme dans l'AOO) réside principalement dans l'économie de la description garantie par le système d'héritage sémantique où les sous-classes héritent de tous les attributs et les opérateurs de leurs classes hyperonymes.

La description du lexique dans l'EWN (comme celle de WN bien évidemment) s'articule autour des SYNSETS. Les synsets (S:) (le nom vient de « synonym set ») sont des ensembles de synonymes et constituent le noyau de la construction de cette base de données. Précisons et soulignons que dans l'EWN on traite des synonymes des mots qui peuvent s'interchanger dans certains (mais pas tous) les contextes.

Dans l'EWN, avant de constituer une liste non-structurée des synsets (Inter-Lingual--Index) chacun des WordNets nationaux possède sa propre hiérarchie lexicale gérée par les relations de synonymie, hypéronymie, hyponymie où quoique ce soit la synonymie la relation sémantique de base entre les mots, dans l'organisation des synsets c'est la relation de subordination – appelée dans ce contexte l'hyponymie – qui importe. Par exemple le nom nurse a un hyponyme (subordonné) du nom caregiver, ou inversement, caregiver est un hypéronyme (super ordonné) du nom nurse. C'est cette relation sémantique qui organise les noms dans une hiérarchie lexicale wordnetienne. L'hypéronymie est une relation entre des sens particuliers de mots. Nous avons donc une relation de hypéronymie distincte pour chaque sens du mot de WordNet. Les hyperonymes de l'infirmière (nurse) sont les suivants : {nurse} @→ {health professional, health care provider, caregiver} @→ {professional, professional person} @→ {adult, grownup} @→ {person, individual, someone, somebody, mortal, soul (a human being)} @→ {organism, being} @→ {living thing, animate thing} @→ {object, physical object} @→ {physical entity} @→ {entity}. Les parenthèses indiquent un synset, et @→ est une relation sémantique à lire « est une sorte de ». Ce dernier « entity » est le « unique beginner » pour tous les noms dans la base.

La différence entre la hiérarchie lexicale du WordNet et celle adoptée dans l'AOO est que la première est ontologique et la deuxième linguistique. Elles se recouvrent partiellement car la langue nome la réalité et chirurgien 'surgeon' et le nom désignant la profession 'professional, professional person' du point de vue ontologique et lexicale, il ne doit pas pourtant être adulte de point de vue de la langue, même si dans la réalité c'est ce qui a lieu dans la majorité des noms de professions. Les différences sont plus évidentes si l'on compare le classement des noms comme : missionnaire, prostitué, cordiste, brancardier, prêtre etc. (Szramek-Karcz 2011, à paraîte), car si la différence dans la classification ontologique (WN) et linguistique (AOO) au niveau de chirurgien consiste en « être ou ne pas être adulte », ces deux classements (de WN et de l'AOO) divergent dans la description des activités humaines charnières, celles sur lesquelles on pourrait discuter pour savoir si elles appartiennent aux professions ou pas. Le Wordnet,

comme toutes les ontologies, présente une sorte de hiérarchie des relations entre les nœuds qui des fois, varie considérablement des classements linguistiques.

Dans le projet de l'EWN les types de relations internes à la langue ont été largement agrandis, des équivalents-complexes ont été introduits, comme par exemple eq_near_synonyme, has_eq_hyperonym ou has-Eq_hyponym, car l'EWN doit faire face aux problèmes de la connexion de plusieurs WNs, ce qu'il effectue par intermédiaire de Inter-Lingual-Index.

3. L'ÉQUIVALENT EN LANGUE CIBLE

Troisièmement, la recherche de l'équivalent en langue cible s'effectue différemment.

L'AOO part du principe que les données lexicales devraient être décrites de manière à faciliter leur réutilisation et leur adaptation dans un autre projet qui, vu le progrès prodigieux des sciences informatiques, pourrait voir le jour dans un avenir très proche. Le mot d'ordre est l'architecture modulaire et qui dit architecture modulaire (cf. Meyer 1988) dit des modules autonomes et organisés dans une structure cohérente. Sans rendre compte de la complexité de chacun des modules et de leur nombre, largement supérieurs à ce que l'on présente, le « problème » de « présenter son intervention lors de la rencontre des romanistes polonais à Cracovie » sous forme rudimentaire, pourrait être décomposé comme suit : « présenter son intervention lors de la rencontre des romanistes polonais à Cracovie » – « mener des recherches », « préparer son intervention », « se déplacer » etc.

Nous pouvons jongler avec les modules, les modifier, en ajouter ou en supprimer, sans nuire à la structure de la base de données ni à la traduction des modules car ils sont bilingues et n'ont rien en commun avec les modules dans l'EWN.

Dans le projet EWN le mot « modulaire » apparaît dans un autre contexte. La structure modulaire de l'EWN (Vossen et. al. 1997 : 1) est résumée comme suit : premièrement il y a des modules de langue (ang. « language modules ») qui contiennent le lexique (« lexicon ») conceptuel de chacune des langues, deuxièmement il y a le module indépendant de la langue « language independent module » qui comprend l'Inter-Lingual-Index (l'index inter linguale abrégé comme ILI), le domaine ontologique « Domain Ontology » et l'ontologie Top-Concept « Top-Concept Ontology ».

L'endroit où les wordnets nationaux se rencontrent est l'Inter-Lingual-Index. Inter--Lingual-Index est une liste des synsets du WN anglais version 1.5 qui est agrandie si besoin est (cf. Vossen et al. 1999). Ainsi chacun des synsets du WordNet monolingue (ou national) retrouve au moins un équivalent parmi les synsets du Inter-Lingual-Index. Les problèmes des équivalents au niveau de l'Inter-Lingual-Index, vu la spécificité de chacun des WordNets, sont faciles à imaginer et ont été montrés à maintes reprises dans les publications traitant le sujet (Dorr, Martí, Castellón 1998 ; Ide, Véronis 1998).

La création des Wordnets bilingues nous parait plus adéquate au besoin de la TA comme le proposent dans son article intitulé « A prototype English-Arabic dictionary Based od WordNet » W.J. Black et S. El-Kateb (2004).

Les modules dans l'AOO sont en deux langues, c'est une approche bilingue dont la description vise en premier lieu la TA. Les correspondances entre la langue de départ

et la langue d'arrivée sont inhérentes à l'architecture de la base ce qui entraîne la traduction immédiate de chacun des modules tandis que dans l'EWN, la traduction n'est qu'un résultat de l'équivalence au niveau de l'ILI. Les synsets des Wordnets monolingues retrouvent (ou ne retrouvent pas) leurs équivalents en langue cible avec un réseau de relations d'équivalence qui les connecte.

CONCLUSION

En guise de conclusion, soulignons que notre interrogation de départ portait sur les possibilités de comparer l'AOO et l'EWN et notre démarche à montré que des différences majeures caractérisaient ces deux approches. Dès lors, nos trois orientations nous ont laissé penser que la comparaison devient hasardeuse et perd de sa pertinence dans la mesure où les différences entre l'AOO et l'EWN sont bien trop grandes. La question reste à savoir laquelle de ces deux bases de données s'adapte le mieux à la traduction automatique (Szramek-Karcz 2011).

BIBLIOGRAPHIE

BANYŚ W., 1981, Description indéfinies : arguments ou prédicats en position d'arguments ?, *Linguistica Silesiana* 4.

BANYŚ W., 1983, *Ambiguïté réfrentielle des phrases à descriptions indéfinies en français*, Katowice : Wyd. Uniwersytetu Śląskiego.

BANYŚ W., 1984, Sémantique, structure, syntaxe et lexique, *Cahiers de Lexicologie*, 45.

BANYŚ W., 2002a, Bases de données lexicales électroniques – une approche orientée objets. Partie I : Question de modularité, *Neophilologica* 15, 7–29.

BANYŚ W., 2002b, Bases de données lexicales électroniques – une approche orientée objets. Partie II : Question de description, *Neophilologica* 15, 206–249.

BLACK W.J., EL-KATEB S., 2004, A prototype English-Arabic Dictionary Based on WordNet, (in:) *GWC 2004 Proceedings*, P. Sojka, K. Pala, Smrž, Chr. Fellbaum, P. Vossen (eds.), Brno : Masary University, 67–74.

BOGACKI K., KAROLAK S., 1991, Fondements d'une grammaire à base sémantique, *Lingua e Style* XXVI, 3, 309–345.

DORR B.J., MARTÍ A., CASTELLÓN I., 1998, Evaluation of EuroWordNet and LCS-Based Lexical Resources for Machine Translation, (in:) *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, pp. 393–397.

FELLBAUM Chr., (eds.) 1998, *WordNet : An Electronic Lexical Database*, Cambridge, Mass.–London : The MIT Press.

GROSS G., 1992, Forme d'un dictionnaire électronique, (in :) *L'environnement traductionnel, actes du colloque de Mons, 25–27 avril 1991*, Sillery (Canada) : Presses de l'Université du Québec, 255–271.

GROSS G., 1994a, Classes d'objets et synonymie, *Supports, opérateurs, durées. Annales Littéraires de l'Université de Besançon* 516, 93–102.

GROSS G., 1994b, Classes d'objets et descriptions des verbes, *Langages* 115, 15–30.

GROSS G., 1995a, A propos de la notion d'humain, *Lexiques-grammaires comparés en français : actes du Colloque international de Montréal (3–5 juin 1992*), *Lingvisticae Investigationes Suplementa*, XVII John Benjamins Publishing Company, 71–80.

GROSS G., 1995b, Une sémantique nouvelle pour la traduction automatique – les classes d'objets, *La Tribunes des industries de la langue et de l'information électronique*, 16–19.

IDE N., VÉRONIS J., 1998, Word Sense Disambiguation : The State of the Art, *Computational Linguistics* 24(1), 1–40.

KAROLAK S., 1984, Składnia wyrażeń predykatywnych, (in :) *Gramatyka współczesnego języka polskiego : Składnia*, Z. Topolińska (red.), Warszawa, PWN.

LE PESANT D., MATHIEU-COLAS M., 1998, Introduction aux classes d'objets, *Langages* 131, 6–32.

MEYER B., 1988, *Object Oriented Software Construction*, Upper Saddle River (N.J.) : Prentice Hall International Series in computer Science.

MILLER G.A., BECKWITH R., FELLBAUM Chr., GROSS D., MILLER K., 1990, Introduction to WordNet : An On-line Lexical Database, *International Journal of Lexicography* 3/4, 235–244.

SOWA J.F., 2000, *Knowledge representation: logical, philosophical, and computational foundations*, Pacific Grove, CA : Brooks/Cole Publishing Co.

SZRAMEK-KARCZ S. 2011a, Comparaison de l'Approche Orientée Objets avec l'EuroWordNet pour la traduction automatique. *Neophilologica*, à paraître.

SZRAMEK-KARCZ S., 2011b, Description des noms désignant les activités humaines, Katowice : Wyd. Uniwersytetu Śląskiego, à paraître.

SZRAMEK-KARCZ S., 2006, Description des professions dans l'approche orientée objets, (in :) *Proceedings of the International Conference : Semantic Relations in Language and Culture, Białystok 24–26 october 2005*, K. Bogacki & A. Miatluk (eds.), Białystok : Wyd. Uniwersytetu w Białymstoku, 300–308.

VOSSEN P. (ed), 1999, *Euro WordNet: Multilingual database with lexical semantic networks*, Dordrecht : Kluwer Academic Publishers.

VOSSEN P., BLOKSMA L., PETERS W., KUNZE C., WAGNER A., PALA K., VIDER K., BERTAGNA F., 1999, *Extending the Inter-Lingual-Index with new Concepts, EuroWordNet*, Amsterdam : University of Amsterdam.

VOSSEN P., DÍEZ-ORZAS P., PETERS W., 1997 Multilingual Design of EuroWordNet, (in :) *Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for Natural Language Processing Applications*, (Madrid, July 1997), P. Vossen, N. Calzolaris, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.), Amsterdam : Vrije Universitet, 1–8.

Summary

Nouns description in EuroWordNet and Object Oriented Approach

The present article is a description of nouns in lexical database EuroWordNet and database created in Object Oriented Approach. It broaches the following subjects: the purpose of creating lexical databases (historical outline), the presentation of basic units in both databases and methods of selecting a target language equivalent. The below piece of writing is a preface to discussion concerning the utility of the above mentioned databases in machine translation.

Streszczenie

Opis rzeczowników w EuroWordNet i w ujęciu zorientowanym obiektowo

Artykuł traktuje o opisie rzeczowników w leksykalnej bazie EurWordNet i bazie stworzonej na podstawie ujęcia zorientowanego obiektowo. Tekst porusza tematykę: celowości tworzenia baz leksykalnych (rys historyczny), przedstawienie jednostek podstawowych w obu bazach oraz metody wyszukiwania (dotarcia) ekwiwalentu w języku docelowym. Stanowi zalążek dyskusji nad użytecznością wymienionych wyżej baz dla tłumaczenia automatycznego.