Les ressources dictionnairiques

Introduction

Le dictionnaire, outil dans la diffusion du savoir de notre société, est le résultat d'un long processus de développement et de représentation de notre connaissance des langues.

Les premiers dictionnaires sont apparus dans l'Antiquité sous forme de listes de mots, comme les listes bilingues akkadien-sumérien (vers 2400 av. J.-C.), les listes de mots de la Grèce antique, dont par exemple celle de Protagoras d'Abdère contenant des mots difficiles extraits des poèmes d'Homère (vers le V^e siècle av. J.-C.), ou encore les dictionnaires chinois (II^e siècle av. J.-C.).

C'est seulement vers 1502 qu'Ambrogio Calepino[1] va publier le Dictionarium (dictionnaire bilingue latin-italien), qui au fil de ses éditions va devenir le tout premier dictionnaire multilingue avec onze langues (latin, grec, italien, espagnol, français, allemand, hébreu, flamand, anglais, polonais et hongrois) en 1588.

Aujourd'hui, les dictionnaires papier tels que le Larousse, le Robert ou bien d'autres encore font partie intégrante de notre vie quotidienne. Avec le développement de l'informatique, la plupart des dictionnaires existant sur support papier ont été mis sur support électronique et commercialisés sur CD-ROM, sur DVD-ROM ou bien sont accessibles sur Internet. Il s'agit d'un nouveau type de dictionnaire, que nous allons appeler dictionnaire informatisé.

Depuis une vingtaine d'années, de nombreux chercheurs ont développé un grand nombre de modèles de bases de données lexicales ou dictionnaires électroniques formalisés, que nous appellerons dictionnaires électroniques. Les dictionnaires électroniques comportent des données spécifiques destinées à l'analyse automatique des langues. Nous pouvons distinguer deux types d'usage d'un dictionnaire électronique : usage humain ou usage automatique. Un dictionnaire électronique à usage humain contient souvent des informations implicites qui nécessitent une connaissance de la part du lecteur et qui ne sont pas adaptées aux machines. Un dictionnaire électronique servant de données pour des programmes de TAL a besoin d'informations explicites et non ambiguës.

Dans cette partie, nous allons présenter uniquement les projets qui nous ont inspiré dans la construction de notre dictionnaire. Nous avons utilisé les travaux sur les codes flexionnels des dictionnaires DELA. Les projets EuroWordNet et Papillon montrent la nécessité d'utiliser une approche par pivot dans la structure d'un dictionnaire multilingue. Nous découvrirons une stratégie de peuplement de base lexicale à travers le projet Papillon. Nous avons étudié les relations sémantiques de WordNet et du DEC pour définir les relations

[1] voir l'article *dictionnaire* de l'Encyclopédie Hachette Multimédia

spécifiques aux noms propres.

2.1 Travaux du LADL

Sous la direction de Maurice Gross, le Laboratoire d'Automatique Documentaire et Linguistique (LADL) de l'université de Paris VII a développé plusieurs dictionnaires électroniques, qui peuvent être regroupés en deux catégories. La première catégorie comporte les dictionnaires de formes non fléchies : le DELAS [Courtois, 1992] pour les mots monolexicaux, le DELAP [Laporte, 1990] pour la phonémisation des mots monolexicaux et le DELAC [Silberztein, 1990] pour les mots polylexicaux. La seconde catégorie regroupe les dictionnaires de formes fléchies : le DELAF, le DELAPF et le DELACF.

[Courtois, 1992] définit ainsi l'objectif des dictionnaires du LADL :

> *Un objectif des dictionnaires électroniques est de construire des structures où sont répertoriées les unités de la langue, avec un certain nombre de propriétés nécessaires au traitement automatique.*

Le DELAS, ou Dictionnaire Électronique du LADL de formes simples, pour le français comporte environ 80 000 entrées de mots monolexicaux, c'est-à-dire des séquences de lettres. Une entrée du DELA se présente sous la forme suivante :

abacule,N1+z3

abajoue,N21+z3

cheval,N4+Anl+z1

où le mot *abacule* correspond à la forme canonique. Le code *N1* indique que ce mot est un nom qui suit la classe morphologique numéro 1 : (0,-,s,-) (voir Annexe B) ; *z3* est un code sémantique permettant de préciser que le mot *abacule* appartient à un langage spécialisé, contrairement au mot *cheval*. La figure 2.1 et la figure 2.2 présentent les codes grammaticaux et les codes sémantiques du DELAS [Paumier, 2006].

En appliquant les règles de flexion sur le DELAS, nous obtenons le Dictionnaire Électronique du LADL de formes fléchies ou DELAF, constitué d'environ 900.000 formes fléchies. Une entrée du DELAF se présente sous la forme suivante :

mercantiles,mercantile.A+z1:mp:fp

glace,.N+z1:fs

où *mercantiles* correspond à la forme fléchie et *mercantile* à la forme canonique (ou lemme). *A+z1* précise que ce mot est un adjectif appartenant au langage courant. *mp* et *fp* indiquent que *mercantiles* est la forme du masculin pluriel et aussi la forme du féminin pluriel de la forme canonique *mercantile*.

La structure du DELAC (Dictionnaire Électronique du LADL de mots composés) et celle du DELACF (Dictionnaires Électronique du LADL de mots composés fléchis) sont identiques aux deux dictionnaires précédents. Le DELACF est constitué de plus de 100 000 mots composés (90 000 noms, 15 000 constructions être Prép N, 8 000 adverbes et 500 conjonctions).

Dans le dictionnaire DELAP (Dictionnaire phonémique) et DELAPF (Dictionnaire phonémique de formes fléchies), chaque entrée comporte en plus une représentation phonémique de sa prononciation. Le DELAPF contient environ 620 000 entrées. Voici un exemple du DELAP :

phonémique,fonemik,.A31

Code	Signification
A	adjectif
ADV	adverbe
CONJC	conjonction de coordination
CONJS	conjonction de subordination
DET	déterminant
INTJ	interjection
N	nom
PREP	préposition
PRO	pronom
V	verbe

FIG. 2.1 – Codes grammaticaux du DELAS.

Code	Signification
z1	mot courant
z2	mot rare
z3	mot technique
Abst	abstrait
Anl	animal
AnlColl	animal collectif
Conc	concret
ConcColl	concret collectif
Hum	humain
HumColl	humain collectif
t	verbe transitif
i	verbe intransitif
en	particule pré-verbale (PPV) obligatoire
se	verbe pronominal
ne	verbe à négation obligatoire

FIG. 2.2 – Traits du DELAS.

Il existe aussi des dictionnaires du LADL pour l'allemand, l'anglais, le coréen, l'espagnol, le grec, l'italien, le norvégien, le portugais, le serbe et le thaïlandais.

Multiflex [Savary, 2006] est un programme qui permet de fléchir des mots polylexicaux [Savary, 2000] à partir de leur lemme. Pour cela, un formalisme [Savary, 2005] permettant de décrire la création des formes fléchies a été mis en place.

Des données spécifiques pour chaque langue sont nécessaires. Voici un exemple pour le polonais :

Polish
<CATEGORIES>
Nb : sing, pl
Case : Nom, Gen, Dat, Acc, Inst, Loc, Voc
Gen : masc_pers, masc_anim, masc_inanim, fem, neu
<CLASSES>
noun : (Nb,<var>),(Case,<var>),(Gen,<fixed>)

adj :(Nb,<var>),(Case,<var>),(Gen,<var>)
adv :

La première partie de ce fichier décrit les catégories grammaticales (nombre, cas, genre) qui existent en polonais. La deuxième partie précise pour chaque classe grammaticale si celle-ci varie suivant le nombre, le cas ou le genre. En lisant ce fichier, on constate que le genre des noms polonais est toujours fixe et qu'ils varient suivant le nombre et le cas, tandis que les adverbes sont invariables.

Les mots polylexicaux sont découpés en unités et chaque unité est associée à une variable ($1, $2...). Par exemple, le mot *Athens '04* est décomposé en cinq unités :

$1=Athens
$2=<espace>
$3='
$4=0
$5=4

Chaque unité est associée à un code flexionnel, sauf si celle-ci est invariable. Par exemple :

avant-garde(garde.N21:fs)

Dans cet exemple, le code *N21* indique que l'unité *garde* suit la règle morphologique : (-,0,-,s). Le code *fs* signifie féminin singulier.

On attribue à chaque mot polylexical un code flexionnel :

avant-garde(garde.N21:fs),NC_XXN

Les codes flexionnels de Multiflex peuvent être représentés sous la forme d'un graphe (figure 2.3). Pour notre exemple, on aura donc dans la variable $1 le mot *avant*, dans $2 le trait d'union et dans $3 le mot *garde*. L'expression *Gen==$g* signifie que le genre est fixe et qu'il correspond au genre de la troisième unité, c'est-à-dire féminin. L'expression *Nb=$n* indique que le nombre peut être variable et prendre toutes les valeurs de sa catégorie, à savoir singulier et pluriel. *<Gen=$g ;Nb=$n>* précise le genre et le nombre du résultat qui sont déterminés par l'unification et qu'ils s'accordent avec le genre et le nombre de la 3ème unité. En appliquant le programme Multiflex, on obtiendra le résultat suivant :

- *avant-garde,avant-garde.NC_XXN:fs*
- *avant-gardes,avant-garde.NC_XXN:fp*

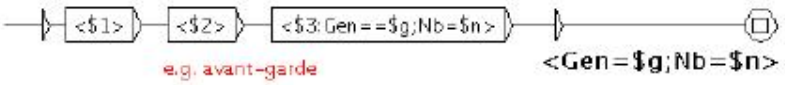

FIG. 2.3 – Code flexionnel NC_XXN.

Discussion

A partir des travaux du LADL et de Multiflex, nous retiendrons la nécessité d'utiliser des codes flexionnels qu'il faudra associer à chaque lemme afin de générer automatiquement toutes les formes fléchies d'un nom propre (voir sections 5.2.6 et 5.2.7 du chapitre 5 page 99). Le but de notre thèse n'étant pas de développer un autre système de codes flexionnels, nous utilisons, pour le cas du français et du serbe, les codes du DELAS pour les noms

propres monolexicaux (voir Annexe B) et envisageons d'utiliser les codes de Multiflex pour les noms propres polylexicaux.

De plus, nous avons prévu, dans le cadre de travaux futurs, de développer un système identique pour la génération d'alias et de dérivés de noms propres (voir sections 3.2.1 et 3.2.2 du chapitre 3 pages 57 et 61).

2.2 EuroWordNet

Avant de présenter la base de données lexicale multilingue EuroWordNet, il nous paraît indispensable de commencer par une description du projet WordNet, qui constitue sans doute une référence indispensable à connaître dans le monde des dictionnaires électroniques et qui sert de point de départ à EuroWordNet.

2.2.1 WordNet

Développé en 1985 par des linguistes du Laboratoire des Sciences Cognitives de l'Université de Princeton, sous la direction de G. A. Miller, WordNet [Miller, 1995] est une base de données lexicales anglaises dont la conception a été inspirée des théories psycholinguistiques et informatiques sur la mémoire lexicale humaine. L'objectif de ce projet est de lister, de classifier et d'établir des relations entre le contenu lexical et le contenu sémantique de la langue anglaise. La version actuelle de WordNet (2.1), consultable sur le site www.cogsci.princeton.edu, comporte plus de 150 000 mots.

WordNet est un réseau lexical où chaque nœud correspond à un synset et chaque arc est formé par les relations entre synsets. Le synset (ou *synonym set*) est défini comme un ensemble de mots interchangeables, représentant un sens particulier. Par exemple, le nom propre anglais *Paris* (figure 2.4) appartient à quatre synsets différents.

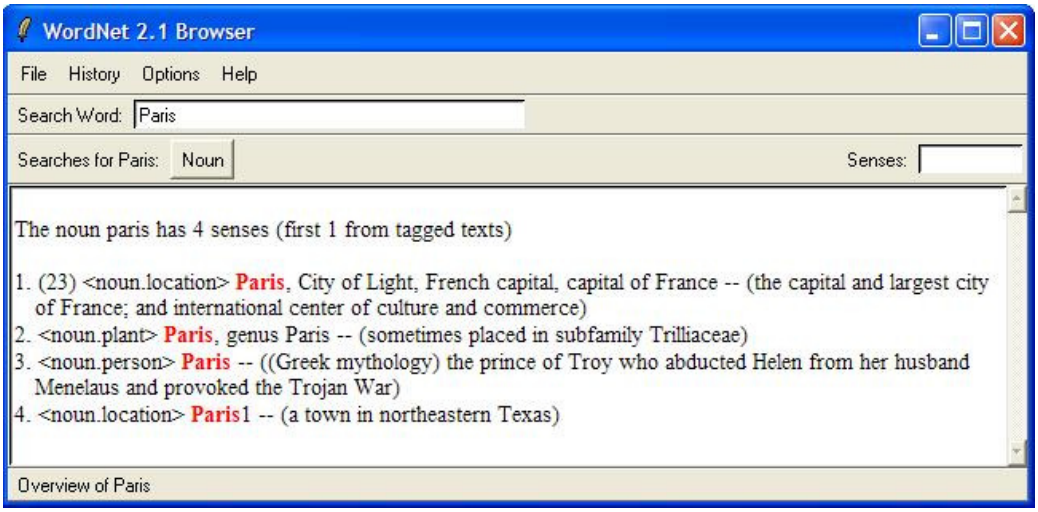

FIG. 2.4 – Recherche du nom propre *Paris* dans WordNet.

Dans WordNet, le lexique est partitionné en quatre catégories syntaxiques : nom, verbe, adjectif, adverbe (figure 2.5). Ce découpage est basé sur une hypothèse cognitive, selon laquelle les mots dans notre mental sont classés en fonction de leur catégorie syntaxique. Chaque catégorie syntaxique possède sa propre hiérarchie de classes sémantiques et ses

POS	Unique Strings	Synsets	Word-Sense Pairs
Noun	117 097	81 426	145 104
Verb	11 488	13 650	24 890
Adjective	22 141	18 877	31 302
Adverb	4 601	3 644	5 720
Totals	155 327	117 597	207 016

FIG. 2.5 – Nombre de mots et de concepts dans WordNet 2.1.

propres relations sémantiques. Il n'existe aucune relation entre des unités lexicales de catégories syntaxiques différentes.

Les noms sont regroupés dans vingt-cinq classes :

- act, action, activity
- attribute, property
- quantity, amount
- natural object
- plant, flora
- event, happening
- animal, fauna
- body, corpus
- relation
- natural phenomenon
- possession
- food
- artifact
- process
- group, collection
- person, human being
- communication
- substance
- location, place
- time
- motive
- shape
- state, condition
- cognition, knowledge
- feeling, emotion

Les verbes sont regroupés en quinze familles :

- body : verbs of grooming, dressing and bodily care.
- change : verbs of change of size, temperature, intensity, etc.
- cognition : verbs of thinking, judging, analyzing, doubting, etc.
- communication : verbs of telling, asking, ordering, singing, etc.
- competition : verbs of fighting, athletic activities, etc.
- consumption : verbs of eating and drinking.
- contact : verbs of touching, hitting, tying, digging, etc.
- creation : verbs of sewing, baking, painting, performing, etc.
- emotion : verbs of feeling.
- motion : verbs of walking, flying, swimming, etc.
- perception : verbs of seeing, hearing, feeling, etc.
- possession : verbs of buying, selling, owning, and transfer.
- social : verbs of political and social activities and events.
- stative : verbs of being, having, spatial relations.
- weather : verbs of raining, snowing, thawing, thundering, etc.

Les adjectifs sont divisés en deux classes :

- adjectifs descriptifs (*big*, *interesting*)
- adjectifs relationnels, qui sont des dérivés de noms (*fraternal*, *presidential*)

Les adverbes ne possèdent aucune structure hiérarchique dans WordNet.

WordNet est construit autour de deux relations principales : la synonymie, qui est modélisée à travers le concept de synset, et l'hyperonymie (figure 2.6), une relation transitive permettant de construire une hiérarchie entre les synsets.

Autour des synsets, WordNet a défini d'autres relations sémantiques (figure 2.7). La méronymie, relation inverse de l'holonymie, permet de spécifier si un synset est une partie

=> entity, something
=> object, physical object
=> artifact, artefact
=> instrumentality, instrumentation
=> conveyance, transport
=> vehicle
=> motor vehicle, automotive vehicle
car, auto, automobile, machine, motorcar

FIG. 2.6 – Exemple de relation d'hyperonymie dans WordNet.

d'un autre synset. L'antonymie exprime les sens opposés entre les synsets. La relation d'implication (*entailment*) s'applique uniquement pour les verbes.

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless
Hyponymy (subordinate)	N	sugar maple, maple maple, tree
Meronymy (part)	N	brim, hat gin, martini
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

FIG. 2.7 – Relations sémantiques dans WordNet.

2.2.2 EuroWordNet

Le projet européen EuroWordNet [Vossen, 1998], coordonné par P. Vossen de l'université d'Amsterdam, a été lancé en 1996. L'objectif d'EuroWordNet est de construire une base de données lexicales multilingue contenant plusieurs langues européennes. Comportant au départ seulement quatre langues (néerlandais, italien, espagnol et anglais), EuroWordNet s'est achevé pendant l'été 1999 avec quatre langues de plus (allemand, français, estonien et tchèque).

Selon [Vossen et al., 1997], il existe plusieurs manières de développer une base de données multilingue :

– La première solution, sans doute la plus coûteuse, consiste à créer des liens par paire de langues. Pour une base de données multilingue contenant quatre langues, il faudrait 12 liens interlingues différents (néerlandais → italien, italien → néerlandais, néerlandais → espagnol, espagnol → néerlandais, néerlandais → anglais, anglais → néerlandais, italien → espagnol, espagnol → italien, italien → anglais, anglais → italien, espagnol → anglais, anglais → espagnol). L'ajout d'une nouvelle langue peut s'avérer très compliqué. La complexité du problème augmente avec le nombre de langues.

– Une deuxième solution consiste à créer une langue artificielle structurée qui va servir d'interlangue. La mise en place d'une langue artificielle nécessite de résoudre plusieurs difficultés. Le lexique doit être précis et assez large pour pouvoir englober les lexiques des différentes langues. L'ajout d'une nouvelle entrée dans une langue peut parfois amener à revoir et améliorer la langue artificielle.
– Une autre solution serait de prendre une des langues comme pivot. Mais cela rend le modèle dépendant de la structure de la langue servant de pivot. Si un sens donné d'un mot est absent dans la langue pivot alors qu'il existe dans une autre langue, cela peut aussi être gênant pour le modèle.
– Une quatrième solution, celle qui a été adoptée par les concepteurs d'EuroWordNet, envisage d'utiliser un ensemble de concepts non structurés, qui servent de liens interlingues. L'avantage d'une telle solution est que cette liste d'index non structurée ne doit respecter aucune théorie linguistique ou cognitive, car elle contiendra simplement des numéros d'identité uniques et ne possédera pas de structure interne. De plus l'ajout d'une nouvelle langue ne remettra pas en cause la totalité de l'index ou les relations que les wordnets entretiennent déjà avec l'index, mais seulement une petite partie de celui-ci.

L'architecture globale d'EuroWordNet [Vossen, 1999] [Jansen, 2004] (figure 2.8) est formée de trois niveaux. Le premier niveau comprend les différentes bases de données lexicales monolingues, qui ont été développées suivant le modèle de WordNet 1.5. Le deuxième niveau, indépendant des langues, comprend un *Inter-Lingual-Index* (ILI). Les synsets de wordnets monolingues ayant été reliés à un même élément de l'ILI (*enregistrement-ILI*) seront considérés comme des concepts équivalents. L'ensemble des synsets de WordNet 1.5 a servi de point de départ à l'ILI d'EuroWordNet. Le dernier niveau contient une ontologie de domaine (*Domain Ontology*) et une ontologie supérieure (*Top Ontology*) (figure 2.9) [Vossen et al., 1998]. L'ontologie supérieure fournit une hiérarchie sémantique des différents enregistrements-ILI et l'ontologie de domaine permet de répartir les enregistrements-ILI selon des thèmes (sport, hôpital, restaurant, trafic aérien, etc.).

L'ontologie supérieure se décompose en trois parties :
– Entité du premier ordre (*1stOrderEntity*) : entité concrète de notre environnement. Par exemple : *Comestible (Function)*, *Living (Natural, Origin)*, etc.
– Entité du deuxième ordre (*2ndOrderEntity*) : situation statique ou dynamique. Par exemple : *length (Property)*, *day (Time)*, etc.
– Entité du troisième ordre (*3rdOrderEntity*) : entité non observable. Par exemple : *idea*, *thought*, *information*, *theory*, *plan*, etc.

Contrairement à WordNet, EuroWordNet autorise des relations entre les différentes catégories syntaxiques. Dans le projet EuroWordNet, il existe deux types de relations : les relations internes d'une langue entre les synsets (figure 2.10) et les relations entre les synsets et les enregistrements-ILI.

Voici les relations les plus importantes entre les enregistrements-ILI et les synsets d'EuroWordNet :
– EQ_SYNONYM : si le synset correspond à un seul et unique enregistrement-ILI (synset : diventare IT / enregistrement-ILI : to become).
– EQ_NEAR_SYNONYM : si un synset correspond à plusieurs ILI-records, si plusieurs synsets correspondent à un même enregistrement-ILI, ou encore s'il y a des doutes sur le choix de l'enregistrement-ILI.
– EQ_HAS_HYPERONYM : si un synset est plus spécifique que les enregistrements-ILI disponibles (synset : kunstproduct NL (artifact substance) / enregistrements-ILI : artifact ; product).

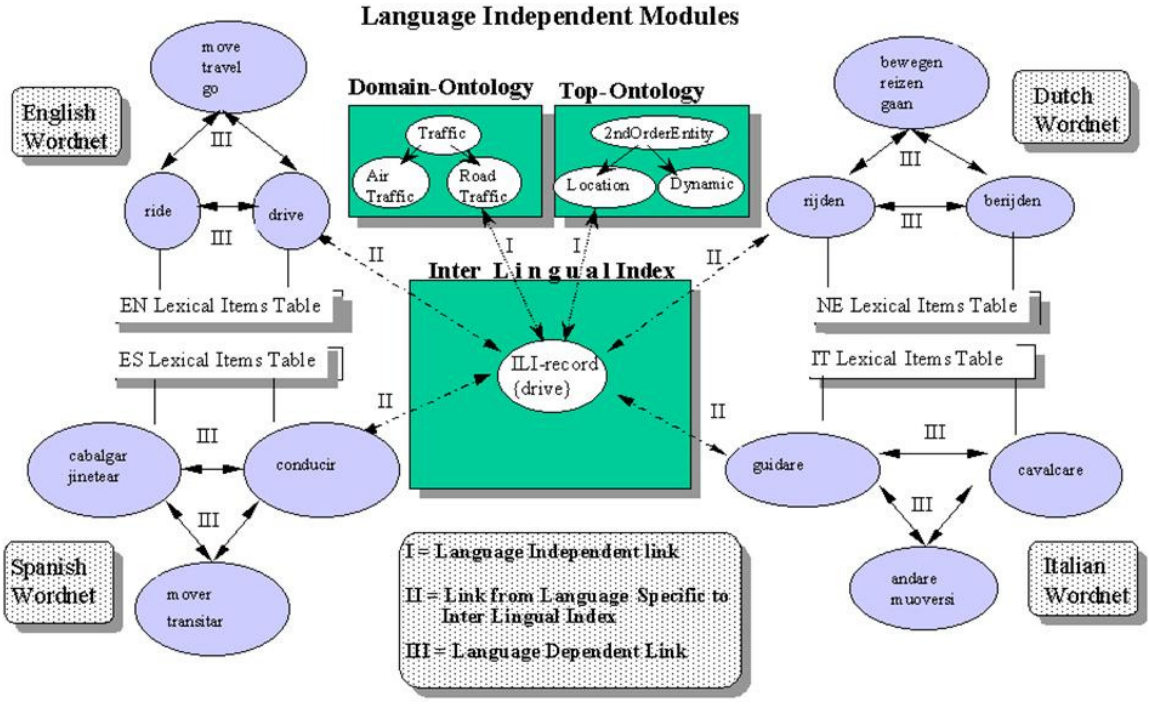

FIG. 2.8 – Architecture d'EuroWordNet

– EQ_HAS_HYPONYM : si un synset peut être associé à plusieurs enregistrements-ILI (synset : dedo ES (a finger or toe) / enregistrements-ILI : toe ; finger).

2.2.3 Balkanet : une extension d'EuroWordnet

Le projet Balkanet [Tufis et al., 2004] [Krstev et al., 2004] est une extension d'EuroWordNet appliquée aux langues des Balkans et à quelques autres langues européennes : bulgare, grec, roumain, serbe, turc et tchèque. Ce projet a débuté en septembre 2001 et s'est achevé en août 2004.

L'objectif du projet Balkanet est de traduire un ensemble de 8 000 concepts dans les six langues du projet pour produire des ressources lexicales et des outils pour le TAL qui soient assez flexibles et réutilisables par différentes applications. L'ILI de Balkanet (BILI) est le résultat de deux sélections sur l'ILI d'EuroWordNet. Le premier ensemble de concepts sélectionné par Balkanet correspond aux concepts de base (Base Concept) d'EuroWordNet, qui regroupent les concepts les plus utilisés et qui constituent une référence pour d'autres concepts. Le dernier ensemble est formé des enregistrements-ILI qui ont été utilisés par la plupart des langues d'EuroWordNet.

Les concepteurs de Balkanet se sont vite aperçus que l'utilisation des enregistrements-ILI d'EuroWordNet posait certains problèmes :

– la difficulté à trouver une traduction juste pour chaque enregistrement-ILI.
– le manque d'information sur les enregistrements-ILI pour pouvoir affecter les liens entre chaque synset des WordNets de Balkanet et les enregistrements-ILI.
– la non structuration des ILI, qui peut nuire à l'application du projet dans des systèmes d'extraction d'information.

En raison de ces problèmes, ils ont décidé de remplacer l'ensemble des enregistrements-ILI d'EuroWordNet par WordNet 1.7 et de considérer l'anglais comme langue pivot.

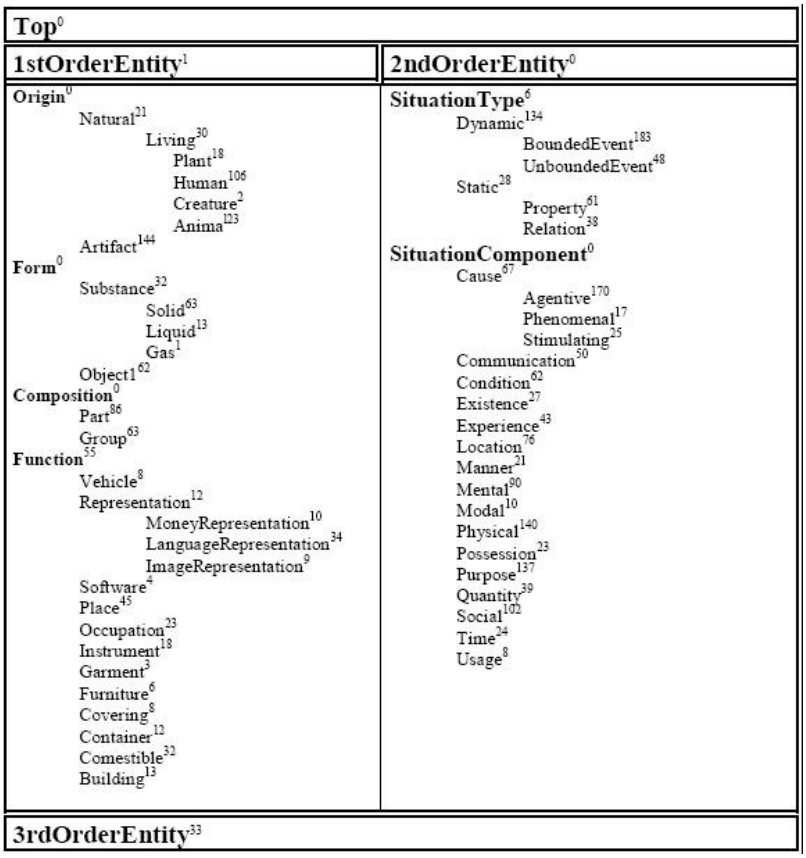

FIG. 2.9 – EuroWordNet Top-Ontology

2.2.4 Discussion

L'étude du projet WordNet et de ses extensions EuroWordNet et Balkanet nous a permis de constater l'importance de la relation de synonymie par rapport aux autres relations sémantiques. Comme dans ces projets, nous nous sommes inspiré de la relation de synonymie pour développer nos différents concepts du domaine des noms propres (voir chapitre 3 page 53). Cette étude nous a aussi permis de connaître les relations sémantiques qui peuvent exister entre différents concepts. Pour le cas des noms propres, nous avons retenu de WordNet les relations de méronymie, de synonymie et d'hyperonymie.

On retrouve dans ces projets de nombreux noms propres. Il s'agit essentiellement de noms propres les plus connus, comme *Victor Hugo*, *Paris*, *Europe*, etc.

Les projets EuroWordNet et Balkanet utilisent un niveau interlingue, basé sur la notion de pivot, qui s'avère indispensable pour créer une base de données lexicale multilingue. Les éléments qui sont reliés à un même pivot correspondent entre eux à une traduction d'une langue vers l'autre. Nous ferons de même (voir section 3.1.1 chapitre 3 page 54).

Dans ces projets, on ne trouve pas d'informations syntaxiques et flexionnelles associées à chaque entrée. De plus, ils ne précisent pas si une entrée correspond à la forme dérivée d'une autre entrée. Nous n'avons pas d'informations sur le contexte d'une relation de synonymie et pourtant ce contexte (politique, savant, familier, etc.) peut s'avérer utile dans l'aide à la traduction. Par exemple, il serait incorrect de traduire *j'ai passé mes vacances sur les plages de France* par *I spent my holidays on the beaches of the French Republic*. Il serait

SEMANTIC RELATION	EXAMPLE
near_synonym	tools <> instrument
xpos_near_synonym	movement <> move
has_hyperonym	mercedes > car
has_hyponym	car > mercedes
has_xpos_hyperonym	election > to vote
has_xpos_hyponym	to fear > paranoia
has_holonym	
has_holo_part	wheel > car
has_holo_member	player > team
has_holo_portion	liquid > drop
has_ holo_made_of	wood > stick
has_holo_location	center > city
has_meronym	
has_mero_part	car > wheel
has_mero_member	team > player
has_mero_portion	drop > liquid
has_mero_made_of	stick > wood
has_mero_location	city > center
antonym	man > woman
near_antonym	to give > to take
xpos_near_antonym	to love > hate
causes	to try > to succeed
is_caused_by	to succeed > to try
has_subevent	to sleep > to snore
is_subevent_of	to pay > to buy
role	hammer > to hammer
role_agent	dog > to bark
role_instrument	sail > to sail
role_patient	learner > to teach
role_location	school > to teach
role_direction	
role_source_direction	ship > disembark
role_target_direction	casa > rincasarse
role_result	ice > to freeze
role_manner	loudly > shout
involved	to hammer > hammer
involved_agent	to teach > teacher
involved_patient	to learn > learner
involved_instrument	to hammer > hammer
involved_location	to teach > school
involved_direction	to pass > place
involved_source_direction	to race > the start
involved_target_direction	to collapse > ground
involved_result	to crystalize > crystal
...	

FIG. 2.10 – Relations internes d'une langue entre les synsets dans EuroWordNet.

indispensable dans le cas des noms propres d'avoir une relation sémantique qui puisse relier un auteur à ses œuvres, une capitale à un pays, etc.[2]

L'utilisation des synsets pour modéliser les noms propres présente quelques inconvénients. Un synset regroupe un ensemble de mots ou de groupes de mots qui sont en relation de synonymie. On ne peut associer directement à un synset des informations spécifiques à un élément du synset (la flexion, les dérivés, la phonétique, les règles de création d'alias ou de dérivés, etc.). Voici quelques exemples de synsets :

{Paris, City of Light, French capital, capital of France} (1)
{Musset, Alfred de Musset, Louis Charles Alfred de Musset} (2)
{France, French Republic} (3)

Nous ne pouvons associer le dérivé *Parisian* au synset (1) car *Parisian* est uniquement le dérivé de *Paris* et non le dérivé de *French capital*. Il faudra préciser que cette information sera uniquement liée à l'élément *Paris* de ce synset. De plus, si l'on souhaite associer la relation d'accessibilité *Paris* est la capitale de la *France*, cela risque de poser un problème. Cette relation devra s'appliquer à tous les élements du synset. Ainsi, on aurait la relation *City of Light* est la capitale de la *France*. Cette relation est peut-être sémantiquement correcte mais elle n'apparaîtra pas dans la plupart des textes.

Peut-on dire que *Paris* et *French Capital* sont des synonymes ? Si jamais le pays change de capitale cela risque de ne plus être vrai. Cela devient compliqué pour le cas du *Jammu-et-Cachemire* qui possède deux capitales : *Srinagar* en été et *Jammu* en hiver.

A cause de ces raisons, nous avons décidé de séparer chaque élément d'un synset en plusieurs éléments différents. Parmi les relations de synonymie (voir section 3.3.1 chapitre 3 page 63), on distingue deux types de synonymie : *Musset* et *Alfred de Musset* versus *Paris* et *City of Light*.

En anglais, dans le cas de l'exemple (1), nous allons créer deux prolexèmes différents : un pour le nom propre *Paris* et un pour le nom propre *City of Light*. Cela permettra d'associer le dérivé *Parisian* au prolexème *Paris*. Celui-ci sera en relation de synonymie avec un contexte stylistique avec le prolexème *City of Light*. La relation de synonymie entre *Paris* et *capital of France* ou *French capital* sera modélisée sous la forme d'une relation d'accessibilité et d'une relation d'expansion classifiante (voir section 3.3.4 chapitre 3 page 69).

Dans le cas de l'exemple (2), les noms propres *Musset* et *Alfred de Musset* sont des variantes (voir alias section 3.2.1 page 57) de la forme vedette *Louis Charles Alfred de Musset*. Nous créerons un prolexème *Louis Charles Alfred de Musset* et nous associerons ces deux alias à ce prolexème, soit directement, soit par règles.

2.3 Le Trésor de la Langue Française informatisé

Le Trésor de la Langue Française informatisé (TLFi) est un dictionnaire monolingue français principalement destiné aux humains. Il est accessible gratuitement sur Internet à l'adresse suivante : `http://atilf.atilf.fr`.

Ce dictionnaire est une adaptation électronique des 16 volumes du Trésor de la Langue Française (TLF) [Dendien and Pierrel, 2003], qui a débuté en 1993 dans les locaux du laboratoire CNRS de l'Institut National de la Langue Française (INALF), puis s'est poursuivie dans le laboratoire d'Analyse et Traitement Informatique de la Langue Française (ATILF) de Nancy. Entièrement encodé dans un format XML (figure 2.11), le TLFi comporte environ

[2] voir relation d'accessibilité section 3.3.3 page 66.

100 000 mots vedettes, 270 000 définitions et 430 000 exemples extraits de la base textuelle Frantext[3].

<art><ved><mot>RÉMITTENT, -ENTE, </mot><cod>adj.</cod></ved>
<sync><H><paramage/><B><dom> MÉD. </dom><cro>[En parlant d'une affection, d'un trouble, d'un symptôme] </cro><def n="t"> Qui présente des poussées et des atténuations successives. </def> <exe n="e"> On a décrit un tétanos discontinu ou rémittent (<aut> CAMUS, GOURNAY </aut><tit> ds Nouv. Traité Méd. <ct> fasc. 2 1928 </ct></tit><loc> , p. 803 </loc><dum>). </dum></exe><syntita n="i"> Psychose rémittente (<so> POINSO-GORI 1972 </so><dum>). </dum></syntita></B><H>
...
<rbbg> BBG. ARVEILLER (R.). Doc. lexicogr. tirés des dict. In : [Mél. Wartburg (W. von)]. Tübingen, 1968, p. 268. QUEM. DDL t. 9.</rbbg>
</art>

FIG. 2.11 – Extrait de l'article *RÉMITTENT* du TLFi.

Discussion

Le TLFi ne comporte pas de noms propres. Nous retiendrons essentiellement l'idée d'une interface de consultation performante, adaptée à un large public (voir section 7.2 chapitre 7 page 118) et celle d'associer une transcription phonétique à chaque nom propre (voir section 5.2.5 chapitre 5 page 97). Notre but n'étant pas de créer une encyclopédie sur les noms propres mais de créer une ressource linguistique destinée à des applications du TAL, nous n'allons pas associer à chaque nom propre une définition, des commentaires ou des exemples comme le ferait le TLFi.

Nous envisageons cependant de faire apparaître des liens vers des encyclopédies, comme Wikipédia, ou d'autres ressources lexicales, comme EuroWordNet et Framenet (voir section 3.3.4 chapitre 3 page 69) en utilisant la relation d'expansion classifiante.

2.4 Dictionnaire Explicatif et Combinatoire

Le Dictionnaire Explicatif et Combinatoire (DEC) est un dictionnaire développé par [Mel'čuk, 1999] pour le russe et le français. Chaque article de ce dictionnaire est élaboré suivant les méthodes définies dans la Lexicologie Explicative et Combinatoire, issue de la théorie Sens-Texte [Mel'čuk et al., 1995].

Le terme explicatif dans le nom du dictionnaire insiste sur le fait que chaque article du dictionnaire est décomposé suivant ses différents sens et selon une méthode rigoureuse. Le DEC est un dictionnaire combinatoire car les combinatoires lexicales et syntaxiques de chaque unité lexicale sont exhaustivement détaillées. Ce dictionnaire contient environ 558 vocables répartis sur quatre volumes.

La construction d'une entrée du DEC doit obligatoirement respecter à la fois une certaine microstructure, organisant la structure des articles (définition, connotations, régimes, etc.), et une certaine macrostructure, régissant l'ensemble des articles.

La macrostructure du DEC s'articule autour de deux notions : la lexie et le vocable. Une lexie ou unité lexicale est définie soit comme un sens particulier d'un mot (lexème),

[3]Frantext est un corpus constitué de plus de 3 600 œuvres littéraires datant du XIX^{e} jusqu'au XX^{e}.

soit comme une locution (phrasème), alors qu'un vocable correspond à un regroupement de lexies. Voici l'exemple d'un vocable du DEC :

> **MÉPRIS**, nom, masc.
> I. Attitude émotionnelle défavorable...[*le mépris pour ce corrupteur*]
> II. Opinion selon laquelle quelque chose n'a pas d'importance...[*le mépris du danger des convenances*]

La première ligne indique la forme graphique du vocable suivie de sa morphologie (catégorie, genre et/ou nombre). En dessous du vocable, une liste de lexies est présentée sous forme d'arborescence.

Chaque article est structuré en trois zones : la zone sémantique, la zone combinatoire et la zone phraséologique.

La zone sémantique, dont l'objectif est de fournir une définition du contenu sémantique d'une lexie, est elle-même divisée en deux parties. La première partie donne une définition lexicographique, dont voici un exemple :

> I. *Mépris de X envers Y pour Z* = Attitude émotionnelle défavorable de X à l'égard de Y causée par le fait suivant : X croit que les actions, l'état ou les propriétés Z de Y causent que Y n'a pas de valeur morale ou sociale ; cette attitude est celle qu'on a normalement dans de pareilles situations.

Le défini, généralement en italique, est un phrasème contenant la lexie vedette et ses actants sémantiques, c'est-à-dire les arguments de prédicat sous forme de variables (X, Y, Z). A droite de l'égalité (=) se trouve le définissant, qui explicite, en utilisant les actants sémantiques, le sens de la lexie.

Après la définition lexicographique, une liste de connotations, qui parfois peut être absente, est donnée :

Connotations

> 1) Cœur I.1a est le siège des sentiments [voir CŒUR I.4a].
> 2) Cœur I.1a est le siège de l'intuition [voir CŒUR I.4b].
> 3) Cœur I.1a qui bat représente la vie [voir les phrasèmes correspondants dans le CŒUR I.1.a].

La zone combinatoire, elle-même divisée en deux parties, renseigne le lecteur sur toutes les combinaisons de syntaxe possibles d'une lexie et sur les liens qu'entretient cette lexie avec d'autres lexies.

La première partie, appelée zone de combinatoire syntaxique, se présente sous la forme d'un tableau de régime, où les colonnes représentent les actants sémantiques et les lignes listent les valeurs possibles que peuvent prendre ces actants, et une liste de contraintes, dont voici un extrait :

Régime

1=X	2=Y	3=Z
1. *de* N 2. A_{poss} 3 A	1. *de* N 2. *pour* N 3. *envers* N 4. *à l'égard de* N	1. *pour* N

$\left.\begin{array}{l} 1) C_3 \textit{ sans } C_2 \\ 2) C_{1.1} + C_{2.1} \\ 3) C_{1.2} + C_{3.1} \end{array}\right\}$: impossible

$4) C_{1.2} + C_{2.1}$: impossible si $C_{2.1}$ désigne une personne

$5) C_{1.3} + C_{2.1}$: non souhaitable

C_1 : *le mépris de Paul, son mépris, le mépris populaire*

C_2 : *le mépris de pour, envers, à l'égard de ce collègue son hypocrisie*

$C_1 + C_2 + C_3$: *le mépris de Paul, son mépris, le mépris populaire envers à l'égard de ce ministre pour son hypocrisie ses propos diffamatoires, son mépris de l'art pour son inefficacité*

La dernière partie de la zone combinatoire, appelée zone de combinatoire lexicale, liste les fonctions lexicales qui peuvent être appliquées sur cette lexie. Une fonction lexicale est indépendante des langues et se présente sous la forme suivante : f(x) = $\{y_1, y_2...y_k\}$. y_i est une valeur de f sur x. Par exemple, *Magn(fièvre)={forte, élevée ; de cheval}*. Voici un extrait d'une liste de fonctions lexicales et de leurs résultats pour la lexie *MÉPRIS* :

Syn : dédain, irrespect, condescendance, arrogance, hauteur II, morgue, **litt** mésestime, **litt** superbe
Anti : respect I
Anti : considération, égard ; différence, estime
Magn : grand, profond, absolu, souverain, sans bornes ; hautain, froid

L'article se termine par des exemples extraits de textes littéraires contenant la lexie vedette :

> *L'Anglaise reconnut sa rivale et fut glorieusement anglaise ; elle nous enveloppa d'un grand regard plein de son mépris anglais et disparut dans la bruyère avec la rapidité d'une flèche [H. de Balzac]. Je vais peut-être vous paraître vieux jeu, mais j'ai un mépris sans bornes pour ces femmes qui vont d'amant en amant, le plus souvent sans amour, pour des raisons de prestige ou de caractère [A. Maurois]. Rien ne m'a donné un absolu mépris du succès que de considérer à quel prix on l'obtient [G. Flaubert]. Le mépris âcre et froid des passants lui pénétrait dans la chair et dans l'âme comme la bise.*

lexie vocable	lexie num	lexie cgs	lexie formuleEtiquette	FL formuleFL	FL lexie	phraseme phraseme	exemple exemple
LION	I	nom, masc	animal sauvage ou espèce animale	QSyn	_Roi des animaux_	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.
LION	I	nom, masc	animal sauvage ou espèce animale	Gener	fauve	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.
LION	I	nom, masc	animal sauvage ou espèce animale	Gener	félin	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.
LION	I	nom, masc	animal sauvage ou espèce animale	Gener	carnivore	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.
LION	I	nom, masc	animal sauvage ou espèce animale	Femelle du L.	lionne	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.
LION	I	nom, masc	animal sauvage ou espèce animale	Petit du L.	lionceau	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.

FIG. 2.12 – Dico

Le projet Dico (Dictionnaire combinatoire) (figure 2.12), dont l'objectif est de construire une base lexicale pour le français comprenant 3 000 vocables, est une simplification du DEC. La finalité de ce projet est de pouvoir fournir des lexiques destinés au TAL et de produire un dictionnaire accessible au grand public (`http://www.olst.umontreal.ca/`).

Discussion

Le DEC ne possède pas de noms propres parmi ses entrées. Bien que la syntaxe soit un sujet intéressant, nous n'avons pas l'ambition de faire la même chose. Nous avons retenu du DEC l'association possible d'une entrée avec un phrasème (voir la relation d'éponymie dans la section 3.3.5 chapitre 3 page 69). Nous nous sommes inspiré du DEC pour définir la relation d'accessibilité[4].

2.5 Papillon

Le projet Papillon [Mangeot-Lerebours et al., 2003] [Mangeot-Lerebours, 2001] est né en 2000 à la suite d'une collaboration entre le GETA-CLIPS et le NII (National Institute of Informatics) de Tokyo. L'objectif du projet est de créer une base lexicale multilingue à usage humain et pour des agents logiciels. Destiné au départ au français et au japonais, ce projet s'est étendu à d'autres langues, comme l'allemand, l'anglais, le malais, le laotien, le thaï, le vietnamien et le chinois.

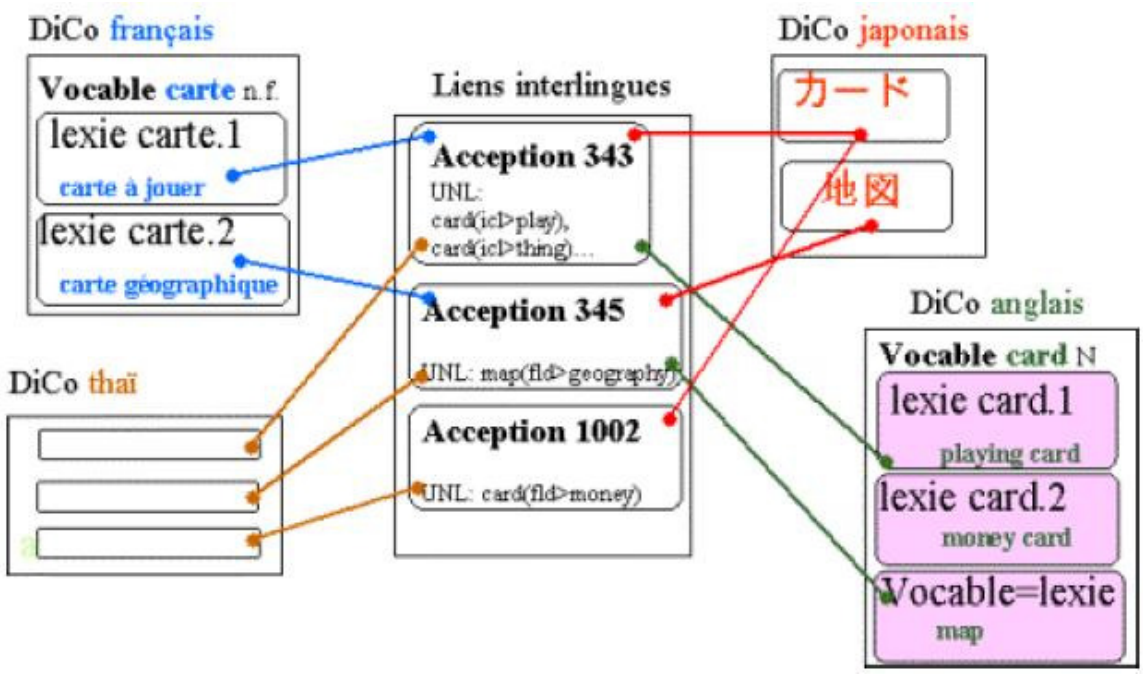

FIG. 2.13 – Macrostructure de Papillon

La macrostructure de la base Papillon (figure 2.13) repose sur la notion d'axie, c'est à dire d'acception interlingue. Dans sa thèse, [Sérasset, 1994] définit cette notion ainsi :

> *Une acception monolingue est une unité sémantique d'une langue. Elle est locale à une langue de la base. [...]*
> *Le but essentiel de la base lexicale est de fournir un lien entre les acceptions monolingues des différents dictionnaires. Pour cela, nous définissons l'ensemble des acceptions interlingues comme étant l'union des ensembles d'acceptions monolingues de différents dictionnaires de la base.*

[4]L'origine de cette dernière était la relation de chef de [Mel'čuk, 1999] ; voir section 3.3.3 du chapitre 3 page 66.

Dans la figure 2.13, le vocable anglais *card* possède trois lexies : *playing card*, *money card* et *map*. Ces trois lexies sont reliées à des acceptions différentes dans le niveau interlingue. Ces relations permettent de trouver la traduction d'une lexie dans une autre langue. Ainsi, la lexie *map* sera traduit en français par la lexie *carte géographique*.

La microstructure de la base lexicale Papillon (figure 2.14) s'est inspirée de celle utilisée dans la base DiCo.

```
<element name="lexie">
<complexType>
<sequence>
<element ref="d:headword" minOccurs="1" maxOccurs="1"/>
<element ref="d:writing" minOccurs="0" maxOccurs="1"/>
<element ref="d:reading" minOccurs="0" maxOccurs="1"/>
<element ref="d:prononiation" minOccurs="0" maxOccurs="1"/>
<element ref="d:pos" minOccurs="1" maxOccurs="1"/>
<element ref="d:langage-level" minOccurs="0" maxOccurs="1"/>
<element ref="d:semantic-formula" minOccurs="1" maxOccurs="1"/>
<element ref="d:government-pattern" minOccurs="0" maxOccurs="1"/>
<element ref="d:lexical-functions" minOccurs="0" maxOccurs="1"/>
<element ref="d:examples" minOccurs="0" maxOccurs="1"/>
<element ref="d:full-idioms" minOccurs="0" maxOccurs="1"/>
<element ref="d:more-info" minOccurs="0" maxOccurs="1"/>
</sequence>
<attribute ref="d:id" use="required"/>
</complexType>
<element>
```

FIG. 2.14 – Schéma XML des lexies.

Papillon-CMD comporte plus d'un million d'entrées dans huit langues différentes comprenant à la fois des noms communs et des noms propres. Papillon-NADIA, basé sur le format DiCo, contient toutes les fonctions lexicosémantiques de la lexicographie explicative et combinatoire entre les entrées de dictionnaires monolingues et n'est pas limité aux noms communs.

La stratégie de construction de Papillon passe par deux étapes. La première consiste en une construction automatique au cours de laquelle des dictionnaires ou ressources existantes sont récupérés et intégrés dans Papillon. Au cours de la deuxième étape, des contributeurs pourront travailler à partir des entrées obtenues lors de l'étape précédente.

Une interface de consultation est proposée aux internautes. Elle permet aussi à toute personne extérieure au projet de contribuer au développement de la base lexicale. Une fois validées par des experts, leurs contributions pourront être intégrées définitivement dans la base Papillon. Elle est accessible à l'adresse suivante : `http://www.papillon-dictionary.org/Home.po`.

Discussion

La construction d'un dictionnaire peut se faire suivant plusieurs stratégies. Il peut s'agir d'une construction manuelle (comme le DEC), automatique ou mixte (à la fois automatique et manuelle, comme dans le projet Papillon). La construction manuelle nécessite gé-

néralement un temps de construction plus long et un coût plus cher qu'une construction automatique.

Notre stratégie de peuplement de Prolexbase consiste dans un premier temps à récupérer manuellement les noms propres d'un dictionnaire papier (voir section 8.3 du chapitre 8 page 141) et des listes de toponymes des précédents travaux du projet Prolex. Il s'agit ensuite de les convertir suivant un format spécifique (voir section 7.18 du chapitre 7 page 127) et de les intégrer dans notre dictionnaire. Nous avons prévu d'utiliser le programme d'extraction automatique de noms propres de [Friburger, 2002] pour remplir automatiquement notre dictionnaire.

La construction d'un dictionnaire électronique multilingue nécessite un travail collaboratif. Pour cela, il est nécessaire de posséder une interface permettant de remplir la base de données (voir section 7.3 chapitre 7 page 119). Étant donné le manque de moyens et de personnes, nous n'avons pas d'experts pour vérifier chaque donnée rentrée dans notre base de données. De ce fait, dans notre projet, seuls des experts peuvent contribuer au développement de notre dictionnaire électronique.

Nous constatons aussi qu'un niveau interlingue est indispensable pour un dictionnaire multilingue. Nous n'avons pas retenu l'idée d'une acception par langue car notre nom propre conceptuel ne correspond pas à un sens d'un nom propre mais à un certain point de vue sur le référent de ce nom propre suivant un diasystème (voir chapitre 3 page 53).

Conclusion

L'étude des différents modèles de dictionnaires électroniques a été très enrichissante et nous a permis d'observer les différentes stratégies mises en place dans la conception de leur structure. Il existe de nombreux autres modèles de dictionnaires électroniques, mais nous n'avons présenté dans cette partie que ceux qui nous ont paru les plus intéressants pour notre projet.

La conception d'un dictionnaire électronique nécessite de spécifier au départ plusieurs paramètres. En fonction de ces paramètres, l'architecture à adopter pour construire le dictionnaire peut varier énormément.

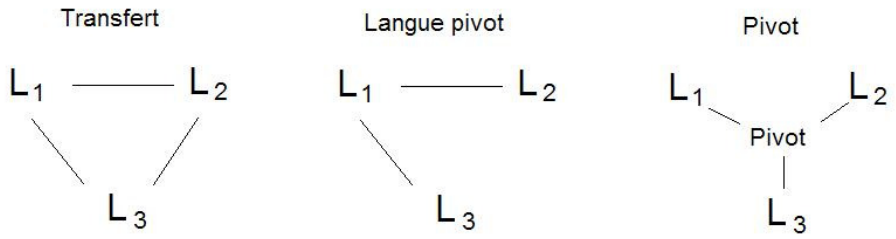

Fig. 2.15 – Différentes architectures

Le premier paramètre, sans doute le plus important, est le nombre de langues que l'on compte inclure dans le dictionnaire électronique, car, selon que le dictionnaire soit monolingue, bilingue ou multilingue, son architecture ne sera pas la même. La macrostructure d'un dictionnaire monolingue se présente sous la forme d'un unique volume, tandis que celle d'un dictionnaire bilingue nécessite au moins deux volumes, un contenant les entrées d'une langue avec les liens vers l'autre langue et vice-versa. La figure 2.15 présente les différentes architectures que l'on peut utiliser pour développer un dictionnaire contenant plus de deux

langues. L'approche multilingue par pivot a été mise en œuvre dans le projet EuroWordNet et dans le projet Papillon sous la forme d'axie. L'approche par transfert a été utilisée dans le projet Eurotra pour la traduction automatique [Danlos, 1989]. L'approche par langue pivot a été utilisé dans le projet Balkanet. Cette approche est recommandée par l'Afnor [Francopoulo, 2003] :

> *Dans un dictionnaire bilingue, nous avons besoin d'un lien pour traduire un sens en un autre et on pourrait imaginer qu'il suffit d'un simple lien entre deux sens [...]. Si cette stratégie est viable pour deux langues, elle est intenable pour un nombre de langues plus important [...]. Nous représentons les traductions via un objet intermédiaire [CN RNIL N 7 : 2003-11-25].*

Il faut aussi définir le format des entrées et les informations linguistiques que l'on souhaite intégrer dans notre dictionnaire. Selon l'utilisation envisagée, des informations syntaxiques, sémantiques ou morphologiques peuvent se révéler indispensables.