Alexandria as a Result of the Integration of WordNet and LDI

Dominique Dutoit
Memodata
17 rue Dumont d’Urville
14000 CAEN (France)
d.dutoit@memodata.com

Ournia Papadima
University of Stendhal, Grenoble (France)
Département Informatique Pédagogique
rania4gr@yahoo.fr

Abstract
This paper is dealing with several problems related to the integration of the two following resources, Wordnet (Fellbaum 1998) and Le Dictionnaire Integral – LDI (Dutoit 1992). The LDI project began in 1988 in France and is based on different principles and goals from Wordnet. Nevertheless, LDI and WordNet have much in common and their integration is possible and highly expected so as to share the information contained in these two models. In the first section we will see in what extent the two linguistic models differ and we will mention some of the reasons of their differences. The two implemented technical data models are schematised and compared. In a second section we explain the integration tasks of the two models and the actual state of the integration process. In the conclusion, we see how it will soon be possible to consult the resulting merge and how to update the data on the Internet.

WordNet and LDI
Many issues are treated differently in WordNet and in LDI. These issues will be introduced through very simple examples and polysemy will be deliberately avoided. The first problem concerns the number of hypernyms or other descriptors needed to describe what is called a word-meaning. Should the word yen be classified in the Japan area or in the monetary unit area? Making it possible to access to a word-meaning unit from any semantic compounds corresponding to its meaning was a major goal of LDI. Therefore in LDI a word-meaning unit like yen stands in the monetary unit class as well as in the Japan domain. More precisely, in LDI no entry is monohierarchical. Concerning domains, no entry is monoaxial.

Another issue is the number of hypernyms for one word-meaning. According to us one word-meaning may have several hypernyms. On the contrary, Wordnet provides most of the times only one direct synset of hypernyms:

-warrior, for the synset of samurai

Where LDI suggests several classes
In Wordnet, *nobleman* or *Japanese* are not hypernyms of the word *samurai*. In LDI we consider that *Japanese* or *nobleman* can be used to lexically define *samurai* or, in a text, to furnish the lexicon with one anaphora. So, there are considered as generic for *samurai* as *warrior*. Finally, we obtain that *samurai* has:

- for generic and for hypernym *warrior*
- for generics *Japanese*, *nobleman* etc.

Remark: in LDI terminology, the term *generic* means a relation between a *concept* (not a synset) and a *word*. This term is opposite to *specific*. A concept groups all the words referring to itself. A generic word for a concept of class (a part of taxonomy for instance) is a word that evokes exactly the concept. A specific word for a concept adds information to this concept (these principles are based on differential semantics).

The second issue is about the use of glosses. For many years we have not been using any defining gloss since the graph itself had to explain the meaning of the words. Instead we have been using concepts, which are some kinds of semantic compounds, used to analyse and integrate the glosses. For the gloss *money of Japan*, we had only two disambiguated concepts:
are considered as potentially referring to the same notion when used in a particular context. The relevance of topology in LDI compared to Wordnet is well exemplified through the two differing treatments of any word whose gloss begins with the words informal terms for. In WordNet the gloss for the synset containing the word cabbage is informal term(s) for money. This synset is linked to the synset money (the most common medium of exchange) by a hypernym relation. In the case of LDI the levels of discourse are not discriminating in determining the reference of the word. The topology alone, which is a set of relations, defines the notion of money and the modifier informal term does not change the topology of the identifier of the notion. Unfortunately this small difference prevents us from a correct mapping between WordNet and LDI and it could be easily solved by a minor change in WordNet.

**Integrating WordNet and LDI**

The two following schemes are representing the two lexical resources we have just described earlier:

![Diagram of LDI and WordNet data structures](image_url)

**Caption:** In LDI:
- concept\_i is a label of a concept such as person, sell, place. There are many types of concepts. Some concepts are able to group any related terms to them. There are two main types of concepts. The class records objects with the same part of speech. The theme records any objects defined by the theme without any constraint of POS. For example, the theme concept Japan records various things like money (yen), food (sushi), people (samurai), places (Tokyo) etc. Another theme concept, the verbal theme hunt groups any words inherently participating to the process.

\( CF_{\text{mn}} \) (Conceptual Function) gives the label of a relation between two concepts or between one word-meaning and a concept.

\( LF_{\text{mn}} \) (Lexical Function) gives the label of a relation between two words. Most of these relations are directly inspired by the meaning-text theory (Melčuk 1992) and they are more precise and delimited than the eng-derivative relation.

In Wordnet:
- * the grey ellipse represents a synset with its gloss and its literals.
- \( SR_{\text{mn}} \) (Synset Relation) gives the label of a wordnet relation between two synsets.
The synthesis of the two models is formally given below:

![Diagram showing the synthesis of two models]

**Fig 2: Integration of the two data structures**

Even in the case of Wordnet 2.0, which introduces the notion of Domain, it’s not possible to merge the two graphs. As it is explained in the documentation explains it, the notion of domain is represented by a set of relations:

- Added topical classification (domain) for some synsets. This relation is represented by the "−c", "−e", "−r", "−s", "−t", "−u" pointer characters. In the index files, the list of pointer types contains a "−c" if the term is a domain, and a "−e" if the term is assigned to a domain.

As it’s not easy to fusion a subset (the way to represent the domains in LDI) and a relation (the way to represent a domain in Wordnet), the two models can not be integrated directly: There is no formal redundancy.

Thus, to integrate the two lexical resources into one container, sharing the information, the modifications concern only LDI:
- integrating the notion of synsets in LDI.
- mapping the French EurowordNet synsets to the relevant word-meaning in LDI.
- adding the glosses.

The most important modification of the data model is not visible on the graph. The question is to count how many graphs result from the merge when we share the two data models. It is right to say that in LDI we had only one graph. But, in WordNet there are two type of graphs. Eventually, the number of graphs is defined by the number of outlines in the figure and this number is 3:

- the fact is that WordNet contains two logics for graphs. One type of graph is defined by the notion of synset with the following content:
  - one gloss for one language
  - one or more literal for one language

The second graph is defined by Wordnet considered as a whole.
The third graph is defined by LDI considered as a whole.

To record this, we had to modify the database format. Instead of having a table of relations with only two nodes and one relation, we have a new table of relations with three nodes and one relation: Item 1, Item 2, Relation between item 1 and item 2, location where this relation exists.

Today we use this generalisation to record number of different graphs.

Also, a new web service called Alexandria displays a part of the resulting merge. Moreover, Alexandria displays parts of results of many projects (Balkanet, Eurowordnet, AlexKor, due to collaboration with the University of Pusan,
Prof. Aesun Yoon). For each of these resources, links are made and the complete integrity of each work is given by the hypergraph defining the Alexandria database. Below, we can see one Alexandria window found by double-click on the word rassembler in the website www.tv5.org.

Content and availability of tools and lexicon

Finally, we would like to give further information about several points. The first point concerns the degree of achievement of the French WordNet and the merging between WordNet and LDI. The second point concerns the availability of an interface to store new data, in WordNet, in LDI or in the two lexical resources. The last point concerns the Alexandria project.

The French WordNet and the merging between WordNet and LDI

The French WordNet is a project beginning during the second phase of the European project EuroWordNet (Vossen 1988). We were one of the two teams implicated in the project for French. The other team was the University of Avignon with Marc El-Bèze. In GWC 2004, Yablonsky (2004) remembered that two complementary approaches were devised in Eurowordnet to build local Wordnet from scratch: 1) the merge approach: building taxonomies from monolingual lexical resources and then, making a mapping process using bilingual dictionaries 2) the expand approach: mapping directly local words to English synsets using bilingual dictionaries.

In our case, in 1998, we had apparently the choice: 1) with LDI, we had something, huge, and very close to WordNet, 2) we had multilingual dictionaries (used in Alexandria). Due to the difference of nature of the two graphs and due to the fact that in 1998 it was not feasible to make an enormous single and complex database (together WordNet, LDI and an hypergraph management system) to prepare
the work, we chose to use our multilingual dictionary. Unfortunately, in that resource we had no gloss for French but only the concepts of LDI. We are now giving below the results:

- **French Gloss**: 45230
- **Gloss still to add**: 23458
- **Other Glosses displayed**: 47235
- **WN synset merged to LDI synset**: 38251
- **WordNet synset to merge to LDI**: 27553
- **Total LDI synset**: 180000
- **Wordnet synset with French**: 43000

Making benefits from WordNet is commonly expected from this nice international collaboration. Our contribution should be the availability of LDI containing all the concepts also labelled in English:

- LDI concepts with English label: 6756
- LDI concepts to label in English: 33693

We are expecting to begin this task as soon as possible.

**The availability of an interface able to store new data, in WordNet, in LDI or in the two lexical resources**

Technically, this interface is already ready to use on Internet. The following snapshot gives an example of the interface:


**Fig 4: immunochemistry synset in the java webstart application Lexidion**

This entry shows a successful merging between LDI and WordNet for the word immunochemistry (immunochimie), with glosses in three languages (English, French and Greek) and concepts (represented as folders) coming from LDI and synsets (represented by the *). Concerning lexical functions, in the left of the window, WordNet and the hybrid LDI are comparable.

So, the relations are added between the synsets immunochemistry and immunochimie. Issued by Wordnet, we see the relation derived (in the subspace MWG) and issued by LDI we see the relation N rel: related to, in the subspace MA0).

Lexidion is a Java web start application called. It is a software maintained by the company since 1988. The two last releases (2003 and 2004) of Lexidion provide: the full support of UNICODE UTF-8, Internet and Hypergraph Management System (HMS).

The HMS is able to store, display, edit, browse etc. any kind of graph. It was used for the integration presented in this paper but we use it also to store, edit, display etc. verbnet frames, patterns, various ontologies etc.
Concerning trees and oriented graphs like LDI or WordNet, Lexidiom allows since many years the following features: 1) make relation between words, concepts, POS, level of discourse, 2) edit, copy, logically delete etc objects. Today Lexidiom contains up to 1,900,000 relations and 780,000 nodes in 24 languages. Most of these objects are not yet displayed by Alexandria. In fact, the version available on the web displays only 450,000 relations and 325,000 lemmas and glosses. Before making Lexidiom accessible to the net to selected teams, we await that improvements will be able to furnish Alexandria content dynamically and directly from Lexidiom. The real-time limit is strong and we need to serve up to 10 queries by second.

**Alexandria**

Alexandria is an amazing tool that can be downloaded free of charge, under certain conditions. Although the initial idea was hatched six or seven years ago, its technical realization has only been feasible for the past several months. For instance, if you want to browse a website, a daily's website for instance you must go to the first page, an ordinary HTML page, double click on any word at random and you'll see a window open with a definition of the word followed by a list of synonyms and expressions using the word. Alexandria is already available in fifteen languages. Alexandria's resources come from different projects (the activity of the company since 1988, three European projects CRISTAL, EuroWordNet and Balkanet, the project AlexKor managed by the University of Pusan, and of course WordNet).

Alexandria has several goals. Many French people who learned English in school understand the general meaning of sentences. However, they 'block' on some words. Now, all they have to do is double-click on words to get translations into their mother tongue. Readers who are proficient in English can have an access to a definition in English. The basic idea is to provide ancillary tools for easier reading. If you are a website author, you can install the software free of charge and make it available to your readers. All you have to do is to go to a website having installed the code and copy it your web pages.

The goal of our team is to integrate in Alexandria new languages (the third edition will add some Chinese, Japanese, Arabian, Polish for up to 25,000 synsets). As regards existing languages, we are pleased to add domain oriented resources such as controlled vocabularies, thesaurus and terminology. Improving the level of knowledge of the entities of the semantic web could be also a nice output. As for the education in foreign languages, Alexandria has to deal with part of the grammar: understanding and use of grammatical terms, such as pronouns or articles. For instance, we take the initiative of to augment the Greek lexicon (provided by Balkanet) with words not described in Wordnet. Almost five hundred pronouns and determiners were added, with their cases (genitive, accusative etc.) and with further explanations. Finally, we hope to develop many partnerships and associations with professional groups etc. We feel that Alexandria can improve many proposals concerned by the development of lexicon and shows its essential uses (education, research, commerce) in the context of the globalization.

**References**


Dutoit D., Nugues P.: « A lexical network and an algorithm to find words from definitions », *Acte de European Conference on Artificial Intelligence, ECAI 2002, Lyon*.


