

ALEXANDRIA

A POWERFUL MULTILINGUAL RESOURCE FOR WEB

Dominique Dutoit

MEMODATA & CRISCO (CNRS)
17 rue Dumont d'urville – 14000 CAEN - FRANCE
d.dutoit@memodata.com

Abstract

This paper is dealing with a new web interface to display linguistic data on the web. This new web interface is a general proposal for the web. Its present name is 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 two years. If you want to read the HTML page, for instance <http://www.memodata.com>, 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. If not, your browser is not in French. Then, you have to use the menu to modify the target language and choice the French between 22 languages.

1. The contextual interface

Alexandria's resources come from different projects:

- the activity of the company since 1988, concerning TID, a conceptual dictionary
- three European projects:
 - o CRISTAL (conceptual indexing)
 - o EuroWordNet
 - o Balkanet
- one Asian project, the project AlexKor managed by the University of Pusan (South Korea),
- 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 but active reading; finally for learning.

1.1. Level 1 : understanding

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. A typical Alexandria webpage is similar to:

```
<html> <head>
<title>my website</title>
<!-- Alexandria IS DISTRIBUTED "AS IS". NO WARRANTY OF ANY KIND IS EXPRESSED OR IMPLIED-->
<!-- YOU USE AT YOUR OWN RISK. THE AUTHOR WILL NOT BE LIABLE FOR DATA LOSS, DAMAGES, LOSS OF -->
<!-- PROFITS OR ANY OTHER KIND OF LOSS WHILE USING OR MISUSING THIS SOFTWARE-->
<!-- Alexandria's parameters for Webmasters -->
<!-- Download and place on your site the "alexandria.wm.js" to update your special terms-->
<!-- Customization is only possible for registered customers-->
<script type="text/javascript" language="JavaScript1.2" src="/alexandria-memodata/alexandria.wm.fr.js"></script>
<!-- Copy the lines below in each page you want Alexandria's functionalities to be available. -->
<script type="text/javascript" language="JavaScript1.2"
src="http://www.sensagent.com/alexandria/scripts/alexandria.main.js"></script>
<link rel=StyleSheet href="http://www.sensagent.com/alexandria/css/alexandria.window.css" type="text/css">
<!--End of the code-->
</head> <body>
<p>Do you know the meaning of <b lang="fr">ramage</b>?</p>
<p>If not, double click on the word. A window will appear with the answer.</p>
And, do you understand <b lang="cn"> 正确地做</b>. Please, select and click again. Last, what about isotopy, in
linguistics. You can try.
</body> </html>
```

Figure 1: How to install Alexandria on a web page

Let's take a concrete example. In the following snapshot, Alexandria suggests some translations for the word *rassembler* found in one web page of TV5.org. It was a page of news. With Alexandria, it is a problem to distinguish a reader and a learner. One click and user goes from one language to the next. Many English people who learned French in school understand the general meaning of

sentences. However, they 'block' on some words. Now, all they have to do is double-click on the word to get its translations into their mother tongue. Readers who are proficient in French can access a definition in French. The basic idea is to provide ancillary tools for easier reading. Doing this, contrarily to the translation software, the system transforms a reader or a player to a learner.

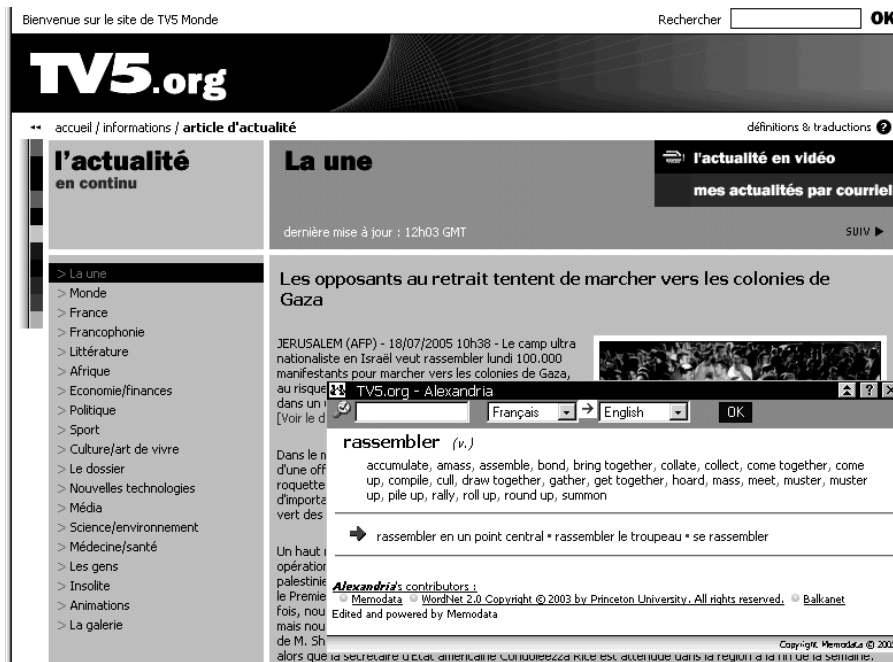



















Figure 2: Alexandria on TV5.org

1.2. Level 2: discovery

Sometimes, a correct but short answer of a dictionary is disappointing. So we could be happy to discover behind each Alexandria's entry a new content: an access to the semantic nets. Two main kind of semantic nets are compiled and shared. We will not discuss these nets here but let us say that one of these nets is TID, a dictionary based on "concepts". The other one is WordNet. Let's suppose that we try to retrieve the name of the *hat of the Pope*. We can search the word *hat* (*chapeau* in French). Following the more specific concepts, *\chapeau d'homme*, *\chapeau rigide d'homme*, the right word will appear soon. It is also possible to start from the result (*tiare*, *tiara*) to search for the decisions of the Pope. In TID, a typical concept, called *theme*, near to the *domain* in Wordnet, will reveal the *small (lexical) world* of the Pope:

-  *prélat* — prelate[Classe]
-  *souverain (monarque)*[Classe]
-  *papauté* — papacy, pontificate[membre]
-  *leader spirituel* — spiritual leader • *catholique* — Catholic[Hyper.]
-  *chef de l'Église, évêque de Rome, évêque universel, le Pape, pape, pasteur suprême, patriarche d'Occident, pontife, S.S, Saint-Père, Sa Sainteté, serviteur des serviteurs du Christ, souverain pontife, successeur de saint Pierre, Très Saint-Père, vicaire de Dieu, vicaire de Jésus-Christ, vicaire de saint Pierre, Votre Sainteté* — Bishop of Rome, Catholic Pope, Holy Father, Pontiff, Pope, Roman Catholic Pope, Vicar of Christ
-  *envoyé pontifical*[Classe]
-  *lettre (décision) émise par le Pape*[Classe]
-  *ornement papal*[Classe]
-  *assemblée convoquée des évêques* — council[Classe]
-  *mission évangélique*[Thème]
-  *droit canonique* — forgiveness;grace;indulgence;mercy;pardon[Thème]
-  *vêtement du Pape*[Thème]
-  *cour papale*[Thème]
-  *le Pape*[termes liés]
-  Alexander VI, Borgia, Pope Alexander VI, Rodrigo Borgia • Alfonso Borgia, Borgia, Calixtus III • Gregory, Gregory I, Gregory the Great, Saint Gregory I, St. Gregory I • Gregory, Gregory VII, Hildebrand • Gregory, Gregory XIII, Ugo Buoncompagni • Innocent III, Lotario di Segni • Albino Luciano, John Paul I • *Jean-Paul II, Karol Wojtyła* — John Paul II, Karol Wojtyła • Leo I, Leo the Great, St. Leo I • Leo III • Giovanni de Medici, Leo X • Aeneas Silvius, Enea Silvio Piccolomini, Pius II • Antonio Ghislieri, Pius V • Luigi Barnaba Gregorio Chiaramonti, Pius VII • Giuseppe Melchiorre Sarto, Pius X • Achille Ratti, Ambrogio Damiano Achille Ratti, Pius XI • Eugenio Pacelli, Pius XII • Odo of Lagery, Urban II • Bartolomeo Prignano, Urban VI[Spéc.]
-  *papesse* • *papal* — apostolic, apostolical, papal, pontifical[Dérivé]
-  *pontifical* — pontifical • *papal* — apostolic, apostolical, papal[Rel.Pr.]

1.3. Level 2: other services and ranking

According to the 22 languages, the needs and the material, several services are provided by the server: spell checker, management of the morphology,

tokenization, detection of compounds, projection of synonyms etc. These services can also be called for research activities, such as query expansion, detection of topics... About 2 millions of root forms are accessible, organized according to various principle and ready to use.

Despite these services, Alexandria is still too recent to be a very famous service on the web. With a constant growth of 20% by month, since the beginning, the various services provide up to 70.000 queries each days. We can make the judgment that this result is very small in comparison to the needs. But an URL allows us to measure the expansion. This URL displays the web pages that use Alexandria. Concerning March 2006, we could anticipate that about 50.000 pages will have installed this agent. The URL is: <http://www.memodata.com/awstats/awstats.pl?config=sensagent>.

It is a pleasure to scan this list. If you do it, we will discover that every part of the society is concerned, without any differences in term of age, instruction, social class etc. We could find:

- medical websites
- research websites in any domain
- cultural websites in any domain
- scholar and academic websites,
- but also
- jew websites
- muslim websites
- christian websites
- Buddhist websites
- atheistic websites
- but also
- politician websites
- non governmental websites
- worker websites
- entrepreneur websites
- and
- homosexual websites
- heterosexual websites
- and also
- game websites
- family websites
- personal websites
- ...

Alexandria concerns everybody.

2. The contextual meaning and conclusion

2.1. Multilingualism and multiculturalism

Let us consider a very general claim.

The BalkaNet project adopted the Princeton WordNet structure and concepts (links between synsets) as the model for the development of wordnets for five Balkan languages and Czech. However, the development of these wordnets showed that mirroring Princeton WordNet synsets and the relations among them to Balkan languages is neither the simplest nor the most appropriate solution. Its rationale could be found principally in the necessity of obtaining a coherent multilingual lexical database. Namely, many Princeton WordNet synsets do not have the lexicalisation in one or several Balkan languages. However, the opposite is also true.

This very common truth was the main source of interest of our projects: when you use semantic nets you are not obliged to find matches between languages when they missed. In fact, with these architectures any users (and computers) could try to understand by the conceptual contexts.

2.2. Non linguistic content

The great usefulness of Alexandria has already been proven by its numerous users. We see, however, that much work still has to be done in order to improve its performance. In order to achieve a really wide coverage of languages, special attention has to be paid on the so called less studied languages, for which many important and extensive resources have been developed, but whose incorporation in the system is not straightforward. In the future, the incorporation of the resources dealing with the encyclopedic knowledge (such as described in Maurel 2002) will be examined. But, Alexandria has not the goal to become an encyclopedia. Dealing with languages and not with the concrete reality the most probable evolution of the system will be more linguistic and symbolic. For instance, works on cognitive linguistics and at the opposite on formal ontology has yet begun.

2.3. Less studied languages and new content policy

The goal of our team is to integrate in Alexandria new languages (the third edition included 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. 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.

3. References

- Dutoit D, Nugues P. (2002). A lexical network and an algorithm to find words from definitions, in *acte de European Conference on Artificial Intelligence*, ECAI, LYON.
- Dutoit, D. (1992). A set theoretical approach to lexical semantics, in *Coling*.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge: MIT press.
- Mel'cuk, I (1992). *Dictionnaire Explicatif et combinatoire du français contemporain*. Montreal: Presses de l'Université.
- Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit, Maria Grigoriadou (2002). *Balkanet: A multilingual Semantic Network for Balkan Languages*, In *Proceedings of the First International WordNet Conference, Mysore India*.

Vossen, P. (1988), ed.: *EuroWordNet: a multilingual database with lexical semantic network*, Kluwer academics publisher,