Conceptual maps and Computational Linguistics: the Italian ALTI project

Francesco Di Maio[1], *Johanna Monti*[2]

[1] Dipartimento di Scienze della Comunicazione, Università degli Studi di Salerno - Italy
[2] Dipartimento di Studi del Mondo Classico e del Mediterraneo Antico, Università degli Studi di Napoli "L'Orientale" - Italy

dimaio@unisa.it, jmonti@unior.it

Abstract

ALTI linguistic multifunctional databases are the result of a project started in 1998 when the research interests of different Italian universities (Università degli Studi di Napoli L'Orientale, Università di Pisa, Università degli Studi di Salerno, Università degli Studi di Roma "Tor Vergata", Università degli Studi di Perugia) came together in one research project of national interest under the coordination of prof. Domenico Silvestri of the Università degli Studi di Napoli L'Orientale.

ALTI stands for **Atlanti Linguistici Tematici Informatici (**Electronic Thematic Linguistic Atlases), which represent a new typology with respect not only to traditional lexicography, but also to computational linguistics, since they put together the characteristics of traditional dictionaries and terminological collections with a conceptual map, which highlights the conceptual relation among terms. The word *atlas* is used to underline that it is not a simple dictionary but a collection of maps, organized as multimedia hyper-textual atlases which collect linguistic data belonging to specialized linguistic areas (such as onomatology, food-terminology, numerals, linguistic activities, metalanguage of linguistics, lexical-grammar) in a multilingual and interlinguistic perspective.

The Atlases describe the phenomenology of specific language areas by linking definitions and usage as given in conventional dictionaries to specific cognitive categories which create conceptual networks, several sets of maps (one for each atlas) or cognitive ellipses. These conceptual networks allow to navigate and to explore the relations between concepts inside a specialized linguistic area.

On the one hand, they are meta-dictionaries, since they refer to other dictionaries, lexical and terminological resources already available in traditional or electronic format, but they also add new information by following original research perspectives, and, on the other hand, they offer multimedia information such as graphs, photos, films which represent very useful tools for the users.

The languages investigated are: ancient and modern Indo-European and non-Indo-European languages; ancient and modern Celtic languages; Latin and Italy's ancient languages; major modern languages.

In conclusion the Atlases are an open work since it is always possible to modify and update them with new contents and so achieve rich virtual cognitive universes.

In our contribution we will describe the main features of the project, the research methodologies, the structure of the Atlases and of the lexical entries, the results achieved until now and the future aims.

Index Terms: conceptual maps, Electronic Thematic Linguistic Atlases, computational linguistics.

1. Introduction

In this contribution we describe the general outline of the Italian ALTI project. The acronym ALTI stands for **Atlanti Linguistici Tematici Informatici (**Electronic Thematic Linguistic Atlases), which are the result of the efforts of different Italian research groups (Università degli Studi di Napoli L'Orientale, Università di Pisa, Università degli Studi di Salerno, Università degli Studi di Roma "Tor Vergata", Università degli Studi di Perugia) coordinated by prof. Domenico Silvestri of the University of Naples L'Orientale. The Atlases are a collection of data belonging to specialized linguistic areas (such as onomatology, food-terminology, numerals, linguistic activities, metalanguage of linguistics, lexicon-grammar) organized inside conceptual maps or cognitive ellipses.

The main aim of the ALTI project is to implement multimedia hyper-textual atlases, which allow to navigate and explore synchronically and diachronically inside specialized thematic lexica, built mainly on the basis of existing lexical resources. They fulfill completely the main requirements of hypertexts such as reticularity, iconicity, non-finiteness and interactivity. On the one side, they are meta-dictionaries since they refer to other dictionaries or terminological and lexical collections, to which they add new information according to original research perspectives, and, on the other side, they contain multimedia information, such as graphs, photos, movies and other useful disambiguation tools for the final user. It is an open work, since every node of the map can be linked to any other node of the same atlas or different atlases. It is always possible to increase and modify the content of the atlases, which become in this way very rich virtual cognitive universes. Interactivity is of course another feature of these hyper-textual atlases, providing the user with a personal reading of the information or with several different readings of the same concept.

2. The Atlases

The ALTI project is composed of six different Atlases:

- the DETIA (Dizionario degli Etnici e Toponimi dell'Italia Antica) has the main aim to create a repository of ethnics and toponyms of Ancient Italy according to the Augustan *discriptio* in *regiones*.
- the AGAM (Atlante Generale dell'Alimentazione Mediterranea) collects all the terms connected with food in the Mediterranean area together with information regarding their preparation and contextual specification.
- the AULIL (Atlante Universale dei Logonimi e delle Istanze Logonimiche) has the main aim of creating a database of all the terms connected with the speech acts.

- the AUNIN (Atlante Universale dei Numerali e delle Istanze di Numerazione) is focused on the numerals which are described using information concerning their linguistic and semantic features.
- the DLM (Dizionario del Lessico Metalinguistico) collects all the terms which describe and explain the phenomenology of languages together with their usage, the different definitions they were given to in different periods of time, authors, works and theoretical schools. It contains 30,000 lemmas, in 31 languages, extracted from approximately 80 texts, with 10.000 authors' quotations.
- The DICOMP (Dizionario delle parole composte) collects all the compound words of a specific terminological field. In particular for this project 30,000 lemmas concerning economy, which are the results of one of the research activities carried out at the University of Salerno under the direction of prof. Annibale Elia, are made available.

As in all computational linguistic and lexicological researches, the structure of the data and of the databases, the query modalities and the updating procedures were changed during the project.

At the beginning the design of the database started from the creation and formalization of tables, then lexicographical workplaces were designed and released in order to carry out the lexicographical work using database facilities. In this way different lexicographical workplaces were implemented for each linguistic area (numerals, food-terminology, …). Once the databases had been created, a Web interface was designed in order to put all data on the Internet and let users access them with dynamic query modalities. To sum up, the ALTI system foresees: (a) different lexicographical workplaces, used by the researchers to store the entries together with relevant information in specific databases, (b) a Web interface composed of dynamic web pages implemented using the ASP language (Active Server Pages).

3. Sources and research methodology

The main sources of information of the researches carried out are:

- dictionaries, glossaries, lexicographical and/or terminological collections, both on paper or in electronic format, off-line and on-line;
- multimedia information for the description of lemmas (graphs, photo, movies, and so on);
- knowledge of the researchers who analyze the sources in order to identify the lemmas.

The researchers on the basis of the sources at their disposal, identify, organize and store:

- lemmas extracted from the sources adding the relevant bibliographical information;
- lexicological and lexicographical information (also in multimedia format) organized according to the descriptors of the cognitive maps of the different atlases;
- possible relations between lemmas inside one or more atlases.

All this information is stored inside the databases, which are updated and managed using lexicographical workplaces for each atlas typology.

These databases represent a set of structured knowledge which allows to have at disposal:

- a set of thesauri organized on a thematic basis for specific lexical areas, with advanced query possibilities;
- a documentation about specific lexical areas which allows to understand their historical evolution and their multilingual dimension;
- a set of tools for updating, managing and querying linguistic data.

4. Basic entry features

The lemmas updated in the atlases can be composed of single words or compound words. All the lemmas are updated in their canonical form, i.e. nouns in the singular form, verbs in the infinitive tense, and so on and are always lowercase, except for the spelling conventions of a specific language.

The lemmas are updated with relevant information connected to their peculiar nature. For instance, for the numerals the following information is required: the function they can have inside the numeral system of a given language, the grammatical category, morphology, structuring procedures, internal and external syntax, etymology, possible gestural and graphic codification and cultural implications.

5. Conceptual maps

For each atlas a cognitive map or ellipsis was created together with a series of electronic data concerning the lemmas.

Conceptual maps were used since they:

- are a graphical representation of knowledge where concepts are organized according to geometrical forms known as nodes and the links between them represent their relations;
- describe the structure of knowledge of the domain under investigation
- represent the main result of the activity of knowledge analysis and modelling.

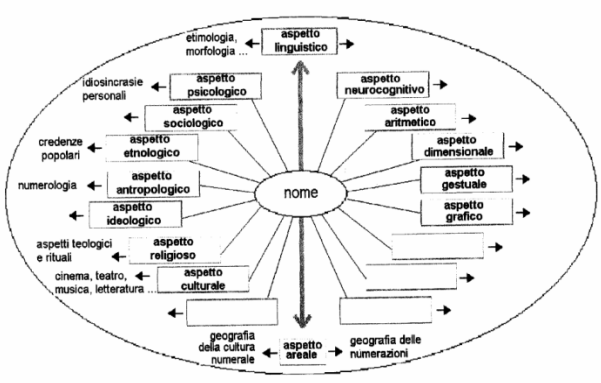

Figure 1: *Example of conceptual map or cognitive ellipsis.*

The maps allow the user to navigate inside one or more atlases.

The focus of each map is given by the lemma. Each map is divided up in two distinct zones: on the left of the map there

is the zone of subjective cultural data and on the right the one concerning objective cultural data.
The subjective cultural data are in common to all the different maps and are represented by psychological, sociological, ethnological, anthropological, ideological, religious and cultural aspects. Of course these aspects vary according to the specific linguistic area investigated. For instance if we take into account the ethnological aspect, this may concern the popular beliefs connected with a specific number (for instance 17 in Italy) in the AUNIN, whereas it may concern the raw/cooked categories in the AGAM.
The objective cultural data are specific of each atlas. For instance, in the AUNIN this zone of the map foresees the neuro-cognitive, arithmetic, dimensional, gestural and graphic aspects, whereas in the AGAM we find gastronomic, economic, technological, dietetic and medical aspects.
In each map, on the border between these two zones there are: on the top of the map the linguistic aspect (represented by etymology, morphology, …) and at the bottom the areal aspect, which is foreseen for the creation of specific geo-linguistic maps.

6. The ALTI Lexicographical workplaces

In the first phase of the project a lexicographical workplace for each atlas was created for the entry of lexical data. All these data will join in the near future in one database managed locally and on the WEB. The software was implemented in Windows, and based on MSAccess tables with SQL queries.
The databases are structured according to the requirements of the different typology of linguistic data. For instance, in the AUNIN, the data are input using a lexicographical workplace in a database composed of three different archives: bibliography, numeral systems and lemmas. Therefore the compilation of three different cards, one for each archive, is foreseen:

- System card
- Numerals card
- Bibliography card

Figure 2: *AUNIN Lexicographical workplace: the numerals card*

Whereas in the AULIL, the lexicographical workplace foresees the compilation of two different cards which correspond to two different databases: bibliography and logonyms.

7. The ALTI Web site

In the second phase of the project, once the databases had been designed and implemented, the next step was the distribution of the ALTI data on the Internet. Therefore, WEB interface pages were designed to handle data dynamically using the ASP language (Active Server Pages). In addition to pure HTML code, this language creates several scripts which generate the page code to be sent to the user browser. In this way the dynamic contents (that is, the contents extracted by the database located on the web server) can be displayed and their appearance changed according to the rules coded in the scripts, without sending the code to the final user program. Only the result is sent, with a significant saving of waiting times on the Internet. This language interfaces with different database types (like Access) with no need of converting data, that would need further programming and indexing operations with the risk of possible errors. The Web site design, which handles the contents of the different linguistic databases (AULIL, AUNIN, etc.), was performed through distinct steps. First of all, the query criteria of the database (the following figure shows the search page designed for the AULIL database) had to be decided.

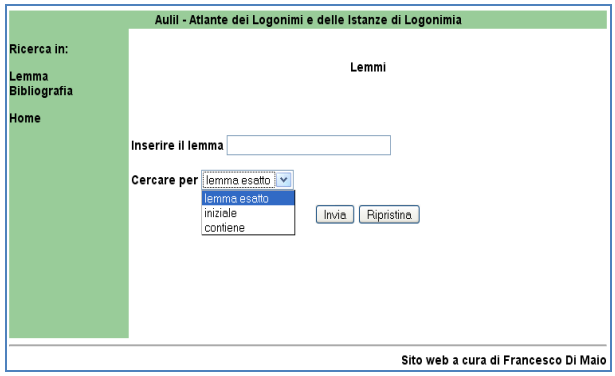

Figure 3: *The AULIL lemma query interface*

The figure shows an additional parameter which can be used in combination with the field **lemma** to perform the query. The input string of the lemma field can be combined along with the three options in the **Cercare per** (Search for) drop-down box:

- **Lemma esatto** (exact lemma): it queries the database looking for the exact string as lemma
- **Iniziale** (initial): it queries the database looking for the string as initial part of a lemma
- **Contiene** (contains): it queries the database looking for the string as part of one or more lemmas.

The input and selection/combination of these parameters in the available fields determine the transfer of parameters and the generation of results through a query to the database in SQL language. For example, searching for the **ne** string as initial string of lemmas in the database and combining the **iniziale** parameter of **Cercare per**, the parameters will be transferred to the query engine of the database when the user clicks the **Invia** (Enter) key :

http://www.alti.unisa.it/alti/aulil/risultlemma.asp?filtrolemma=ne&select=1&Submit=Invia

which for the user means “look for all the lemmas that begin with ne”. For more expert users we could split the string as follows:

- *filtrolemma* : the variable to be input in the query field **lemma**
- *select*: any of the three options of **Cercare per**, whose parameters are: 1=iniziale (initial), 2=lemma esatto (exact lemma) and 3=contiene (contains)
- *Submit*: the Invia (Enter) key, which is the transmission command of these parameters to the server hosting our database. The required data will be searched for, and, if available, they will be displayed in the risultlemma.asp page, a dynamic page specifically created. The following example shows the query used for this search:

```
<%
sel=request("select")
If (Request.QueryString("filtrolemma") <> "") and sel="1" Then
stringasql="SELECT lemma, fonte, pagine FROM logonimi GROUP
BY lemma, fonte, pagine HAVING lemma LIKE '" + filtrolemma +
"%" + "' ORDER BY lemma ASC"
end if
%>
```

If the result of the query is successful, the data will be displayed as in the figure below:

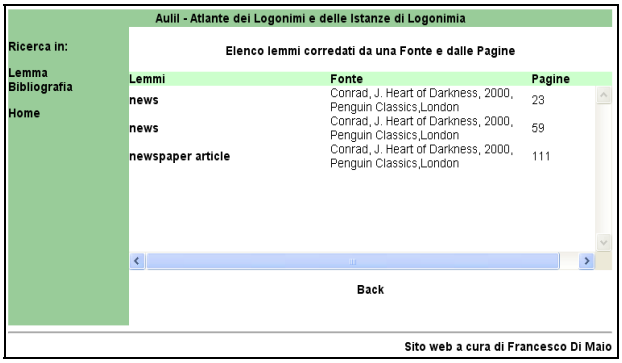

Figure 3: *The AULIL lemma query result interface*

The query result is given by the lemmas with additional information about the source and the pages of the text/dictionary which they were extracted from.
Regarding hyper-textuality, we can connect to another page containing other related information in this page, if we click on a lemma in the list, such as *newspaper article*, we can read the catalogue details of the single lemma from the table which contains them using a query, and we can display them in a different page:

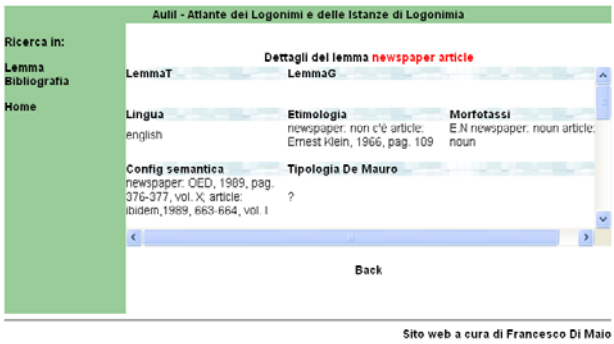

Figure 3: *The AULIL lemma catalogue details interface*

The page for this data was designed in scroll mode so that it can display the information it contains. The user can read the information using the scrollbar on the right side of the page as displayed in the above figure.
The criteria used for the bibliographical search page are similar to the ones used for the lemma search page. In this example, a combination of the search fields **autore (**author**)** and **titolo (**title**)** was used so that the user can search any of the items or both of them at the same time.

Figure 3: *The AULIL bibliography query interface*

Moreover, in this example, you can use the criteria previously described for the search page of lemmas , i.e. search by exact string, initial string and "contained in" string.

8. Conclusions

We have provided an overview of the Italian ALTI project, and in particular we have described the different components and the results so far achieved.
What we have described is only the beginning of a research project which should be considered as an "open" work, that is a work in progress since new languages and new data can be and will be added.
We believe that the framework we have designed can prove to be very useful for both theoretical research and computational applications and thus could become a model for similar lexical resources.
The results of the ALTI research project, in particular the Web site and all the linguistic data, will be available on the Internet in the very near future.

9. Note

Abstract, paragraphs 1,2,3,4,5 and conclusions are by Johanna Monti, whereas paragraphs 6 and 7 are by Francesco Di Maio.

10. References

Atti del Convegno su Numeri e istanze di numerazione tra preistoria e protostoria linguistica del mondo antico – Napoli 1-2 dicembre 1995 - AIΩN (Annali del Dipartimento di Studi del Mondo Classico e del Mediterraneo Antico – Sezione Linguistica, direttore: Domenico Silvestri) 17, 1995.

Chiari, I., Introduzione alla linguistica computazionale, Gius. Laterza & figli, Roma-Bari, 2007.

Di Maio, F., Monti, J., Gli atlanti tematici informatici: oltre il dizionario elettronico. Esemplificazioni dalla versione tedesca dell' Atlante Universale dei Numerali e

delle Istanze di Numerazione (AUNIN) e dell'Atlante Universale dei Logonimi e delle Istanze Logonimiche (AULIL) (in press)

Pannain, R., "Numerali ed istanze di numerazione: note per un progetto di tipologia areale dei numerali", AIΩN (Annali del Dipartimento di Studi del Mondo Classico e del Mediterraneo Antico – Sezione Linguistica, direttore: Domenico Silvestri) 22, 63-105, 2000.

Silvestri, D. "Logos e logonimi", Vallini C. (ed.) , Le parole per le parole. I logonimi nelle lingue e nel metalinguaggio, Atti del convegno Istituto Universitario Orientale Napoli 18-20 dicembre 1997, Il Calamo, Roma, 21-37, 2000

Silvestri, D., "From the eloquence of light to the splendor of the word", Semiotica Special issue, Signs and Light: Illuminating Paths in the Semiotic Web 136 -1/4, , Guest editor : Susan Petrilli: 117-132

Silvestri, D., "I lessici tematici tra lingua standard e lessici scientifici" AIΩN (Annali del Dipartimento di Studi del Mondo Classico e del Mediterraneo Antico – Sezione Linguistica, direttore: Domenico Silvestri) 24, 11-30, 2002

Vallini, C. (ed.), Le parole per le parole. I logonimi nelle lingue e nel metalinguaggio. Atti del convegno Istituto Universitario Orientale Napoli 18-20 dicembre 1997, "Lingue, Linguaggi, Metalinguaggi" 1, Collana diretta da C. Vallini e V. Orioles, Il Calamo, Roma, 2000.