MED-TYP: A Typological Database for Mediterranean Languages

Andrea Sansò

Dipartimento di Linguistica, Università di Pavia
Strada Nuova 65, I-27100, Pavia
sanso@humnet.unipi.it

Abstract

Electronic databases are increasingly popular tools in typological research. Despite the advantages of such tools, there are problems connected both with their construction and with their standardization. For instance, there is generally a considerable gap between the information stored in typological databases and primary data: primary morphosyntactic data are much more difficult to handle computationally than typological generalizations. Moreover, the need for standardization has led typologists to develop highly refined glossing practices and guidelines for collecting data, but there are still too few initiatives to increase standardization in typological databases. The aim of this paper is to suggest a radically new approach to the storage of data for typological analysis. The Med-Typ Database, which is currently being developed at the University of Pavia, has been providing us with concrete experience of the problems that need to be addressed when creating typological databases. This database uses XML annotation and aims to be both a collection of data for future analyses of areal distribution of features within the Mediterranean area and a tool for systematic analysis of the range of variation found in various typological domains.

Introduction: the MED-TYP project

The aim of this paper is twofold, and can be summarized as follows:

a) firstly, we aim to describe the on-going experience of the *Med(iterranean)-Typ(ology) Database*, which is currently being developed at the University of Pavia. The data included in the Med-Typ Database have been providing us with concrete experience of the problems that need to be addressed when creating typological databases;
b) secondly, we aim to suggest a radically new approach to the storage of data for typological analysis, and to discuss the advantages of such an approach.

The Med-Typ project was launched in 1997 and concluded in 2000.[1] Its basic assumption was that some structural features of Mediterranean languages have been significantly influenced by the fact that these languages have been in contact for several centuries. The major aim of the project was to outline a typology of Mediterranean languages, and to describe the distribution of various structural traits within this area so as to uncover possible phenomena of areal convergence. To do so, it has been necessary to plot the features of Mediterranean languages against the universal tendencies ascertained in the world's languages with respect to the phenomena taken into account: this made it possible to distinguish true areal features of the Mediterranean area, derived from language contact, from the results of universal typological tendencies.

The research was based on a language sample including both languages in the Mediterranean area (Catalan, Spanish, French, Provençal, Italian, Sardinian, Friulan, Slovene, Serbo-Croatian, Albanian, Modern Greek, Turkish, Maltese, Modern Hebrew, Modern Standard Arabic, Arabic dialects, Berber) and languages that do not belong, strictly speaking, to the Mediterranean area, but are historically connected to it (Portuguese, Basque, Macedonian, Bulgarian, Romanian). The analysis was based mainly on synchronic data, but diachronic investigation was not excluded. The selected research topics were:

(i) the expression of possession
(ii) relative clause formation
(iii) subordination strategies
(iv) noun phrases and pronominal clitics
(v) volitional constructions
(vi) spatial deixis
(vii) indefinite and negative quantification
(viii) evaluative morphology
(ix) intensifiers and reflexives
(x) yes-no questions
(xi) converbs

The results of this project can be summarized as follows:

a) if linguistic area is to be intended as a group of languages sharing a significant number of features by virtue of contiguity, there is no Mediterranean area as such;
b) much in the spirit of Dahl (2001), however, the areal dimension in the study of Mediterranean languages has revealed a number of unexpected contact phenomena which are significant irrespective of whether they can be described in terms of linguistic areas in the traditional sense. Thus, "area" has turned out to be a significant notion when examining the distribution of typological features in Mediterranean languages, in comparison to those of neighboring European languages and, more generally, to universal typological tendencies concerning the phenomena taken into account.

The creation of an electronic database of linguistic phenomena in the Mediterranean area was not among the aims of the Med-Typ project. A new three-year project on

[1] The project (extended title: "Languages in the Mediterranean area: typology and convergence") has been financed by the Italian National Research Council (CNR). Researchers from the following universities took part in it: Università di Pavia, Università di Pisa, Università per Stranieri di Perugia, Università per Stranieri di Siena, Università di Trieste, Università della Tuscia (Viterbo). The reader is referred to Cristofaro & Putzu (2000), Ramat & Stolz (2002), Ramat (2003), and Stolz & Sansò (forthcoming).

the typology of languages in the Mediterranean area and their relation to European languages has been launched by the Italian Ministry of Education (*FIRB – Fondo per gli Investimenti della Ricerca di Base*) in February 2003: the research program of this project explicitly involves the creation of a typological database, based on both the data collected during the Med-Typ project and new data that will be collected in the years 2003-2006.

The database

The Med-Typ database aims to be both a collection of data for future analyses of possible typological implications and areal distribution (*quantitative typology*) and a tool for systematic analysis of the range of variation found in various typological domains (*qualitative typology*).
The first draft of the database is expected to be published on the web later this year. It will include the following topics: (1) relative clause formation strategies; (2) noun phrases and pronominal clitics; (3) evaluative morphology.

Motivation and aims

Typological data are computationally treatable, and hence electronic databases are increasingly popular tools in typological research. Despite their popularity, there are only a few initiatives towards standardization (Monachesi et al 2002, Dimitriadis & Monachesi 2002), and available databases can be quite heterogeneous. On the other hand, the need for standards has led typologists to develop highly refined glossing practices (see for instance Lehmann et al 1994; Bickel et al 2004; Lehmann 2004a) and guidelines for collecting data and language documentation (Lehmann 2001; Lehmann 2004b), so that the time is now ripe to address the issue of standardization in typological databases.
The language sample of the Med-Typ database is sufficiently large to include various morphosyntactic types for each phenomenon in question, but not large enough to allow an approach in terms of a relational database structure, which is by and large the most common solution adopted when designing typological databases (see for instance the typological database of agreement, cf. Tiberius et al 2002, or the typological database of intensifiers and reflexives developed at the Free University of Berlin, cf. König et al 2003).
It should be added that the primary data contained in our database are more difficult to handle computationally than typological generalizations. Moreover, we did not want to be dependent on proprietary solutions. These considerations led us to design a database with XML tagging (Sansò 2003). The use of XML as a mark-up language has many well-known advantages (XML makes it possible to exchange complex data between systems that use different formats, it is based on the "single-source/multiple-output" principle, and is also more longeval than the applications used in the creation of typological databases). The most striking advantage is the possibility of storing a huge amount of pieces of information as attributes of elements: these pieces of information are not displayed but may be searched. But what has been crucial to this choice is the awareness that a high degree of interoperability when creating linguistic resources is essential. The Semantic Web is going to crucially determine the shape of linguistic resources of the future, consistently with the vision of an open space of shareable knowledge available on the web for processing. The need of ever growing language resources for effective content processing requires a change in the paradigm, and the design of a new generation of language resources, based on open-content interoperability standards.
In our view, interoperability can only be achieved by associating openly available XML schemas with the documents, or by deriving from those schemas ideas and solutions that can help in the mark-up of linguistic information that is not under the form of a text. This is not at all a trivial and uncontroversial task.

Architecture

The primary data collected in the Med-Typ database are mainly samples of clauses/sentences (or list of words) drawn from grammars/dictionaries or elicited through questionnaires distributed to native speakers. Both types of data are provided with morphological glosses and the exact reference of the source from which the examples are taken is given, in order to ensure that all the information stored in the database can be traced back to its original source. The abbreviations used in the glosses follow the list established by Bickel et al (2004). The main task of annotators thus consists in creating a uniform, possibly theory-neutral annotation scheme for this kind of data. The range of phenomena to be annotated poses a considerable challenge to any attempt to adapt existing annotation practices, predominantly designed for annotating written texts or dialogues.

XML files

We make use of stand-off annotation, i.e. annotations are stored in documents separate from primary data. There are some obvious reasons to do so, especially when the base material is large (i.e. series of clauses or sentences) and the markup involves multiple overlapping hierarchies. Here is an example of stand-off annotation for relative clauses in Catalan.

File: ca_relcl.xml

```
<CLAUSES xml:lang="ca">
<CLAUSE id="001">
<w id="w_001">L'</w> <w id="w_002">home</w>
<w id="w_003">que</w> <w id="w_004">soparà</w>
<w id="w_005">aquí</w> <w id="w_006">és</w>
<w id="w_007">el</w> <w id="w_008">meu</w>
<w id="w_009">germà</w>
</CLAUSE></CLAUSES>
```

File: ca_relcl_annotation.xml

```
<ANNOTATION href="ca_relcl.xml">
<ITEM id="ann_001"
href="ca_relcl.xml#CLAUSES/CLAUSE[@id(001)]"
source="Cesar Montoliu">
<HEAD
href="ca_relcl.xml#CLAUSES/CLAUSE/w[@id(01)..id(02)]
" function="SBJ" referential="yes" animate="yes"
definite="yes" />
<RELATIVE_CLAUSE href="
ca_relcl.xml#CLAUSES/CLAUSE/w[@id(03)..id(06)]"
tense="PST" />
<RELATIVIZER href="
ca_relcl.xml#CLAUSES/CLAUSE/w[@id(03)]" func-
tion="SBJ" /> </ITEM> </ANNOTATION>
```

Figure 1: Stand-off annotation for relative clauses

DTDs

For each linguistic phenomenon contained in the first draft of the database a DTD has been created that includes the whole set of tags used in the annotation of that phenomenon. In the following figure the DTD for a class of XML documents (namely, "relcl") is displayed as exemplification of the kind of linguistic traits that are relevant to this phenomenon.

relative_clauses.dtd

```
<!ELEMENT RELATIVE_CLAUSES (ITEM+)>
<!ATTLIST RELATIVE_CLAUSES language CDATA #REQUIRED>
<!ELEMENT ITEM (MAIN_CLAUSE, TRANSLATION)>
<!ATTLIST ITEM variety CDATA #IMPLIED>
<!ATTLIST ITEM informant CDATA #REQUIRED>
<!ELEMENT MAIN_CLAUSE (#PCDATA|HEAD|RELATIVE_CLAUSE)*>
<!ELEMENT HEAD (#PCDATA)>
<!ATTLIST HEAD function (SBJ|OBJ|IND_OBJ|POSS|OBL) #RE-
QUIRED>
<!ATTLIST HEAD referential (yes|no) #REQUIRED>
<!ATTLIST HEAD animate (yes|no) #REQUIRED>
<!ATTLIST HEAD definite (yes|no) #REQUIRED>
<!ELEMENT RELATIVE_CLAUSE (#PCDATA|RELATIVIZER)*>
<!ELEMENT RELATIVIZER (#PCDATA)>
<!ATTLIST RELATIVIZER function (SBJ|OBJ|IND_OBJ|POSS|
OBL) #REQUIRED>
<!ELEMENT TRANSLATION (#PCDATA)>
```

Figure 2: DTD for the class of XML documents named "relcl" (i.e., "relative clauses"; first draft).

In designing the DTDs for the first draft of the database the following requirements have been identified:

- be robust and wide-coverage;
- be flexible, customizable and usable for practical applications;
- be modular (to allow for partial instantiations of the scheme);
- be suitable for multilingual application[2]

Queries

When selecting a language, the user gains access to the full set of primary data, which are displayed on the screen as a triple, *example* + *gloss* + *translation*, i.e. without including any judgment or typological generalization. A "readme" file will instruct the user to perform queries based on the linguistic traits characterizing primary data. Queries are made possible through the use of the XSLT language (Clark 1999; Kay 2003). XSLT is a functional programming language optimized for parsing and generating XML documents. An XSLT program (or stylesheet) takes one or more XML files as its input and transforms them into one or more files in HTML or XML. The following properties make XSLT an ideal candidate for our queries:

a) all the files of our database are in XML format; they are grouped together to form families of similar documents meeting the requirements of the same DTD(s); XSLT allows the programmer to ride easily through families of similar documents and to extract relevant information from them;
b) in XSLT, the programmer specifies what output should be produced when particular patterns occur in the input. This makes it relatively easy to translate simple queries based on certain properties of the primary data into XSLT code.

In the transformation process, XSLT uses XPath to define parts of the source document that match one or more predefined templates. When a match is found, XSLT will transform the matching part of the source document into the result document. The parts of the source document that do not match a template will end up unmodified in the result document. Here is an example of the code associated with the query *display examples of subject relativization*:

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:template match="/"> <xsl:variable
name="clauses" select="document(//file[1]/@href)" />
<xsl:variable name="annotation" se-
lect="document(//file[2]/@href)" /> <xsl:variable
name="glosses" select="document(//file[3]/@href)" />
<xsl:for-each select="$annotation/ANNOTATION/
ITEM/RELATIVIZER[@function='SBJ']"> … <xsl:value-of
select="$clauses/CLAUSES/@language" /> … Example:
<xsl:value-of select="$clauses/CLAUSES/CLAUSE/@id"
/> Source: questionnaire (source of informa-
tion:<xsl:value-of select="$annotation/ANNOTA-
TION/ITEM/@source" />) … <xsl:for-each se-
lect="$clauses/CLAUSES/CLAUSE/w"> … <xsl:value-of
select="." /> … </xsl:for-each> … <xsl:for-each se-
lect="$glosses/GLOSSES/GLOSS/g"> … <xsl:value-of
select="." /> … </xsl:for-each> … Translation:
<xsl:value-of se-
lect="$annotation/ANNOTATION/ITEM/TRANSLATION" /> …
</xsl:for-each> … </xsl:template> </xsl:stylesheet>
```

Figure 3: A fragment of XSLT code (*display examples of subject relativization*)

And here is the output of the query:[3]

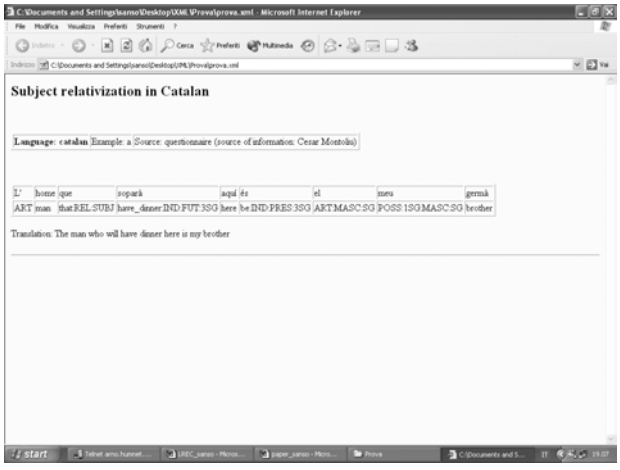

Figure 4: Output of the query "display examples of subject relativization"

[2] The current DTDs present slightly different annotation choices, depending on language-specific phenomena. For example, while for some languages the morphological function of a suffix is easily identifiable, it is not clear how tenable the same claim is for languages such as Arabic dialects, which typically exhibit complex cases of discontinuous morphology.

[3] The output is a simple HTML page; the HTML code has been removed from Figure 3 for the sake of clarity.

The database is queried by using pull-down menus. Each option in the menus makes reference to an .xsl file. In figure 4 the language Catalan is being selected:

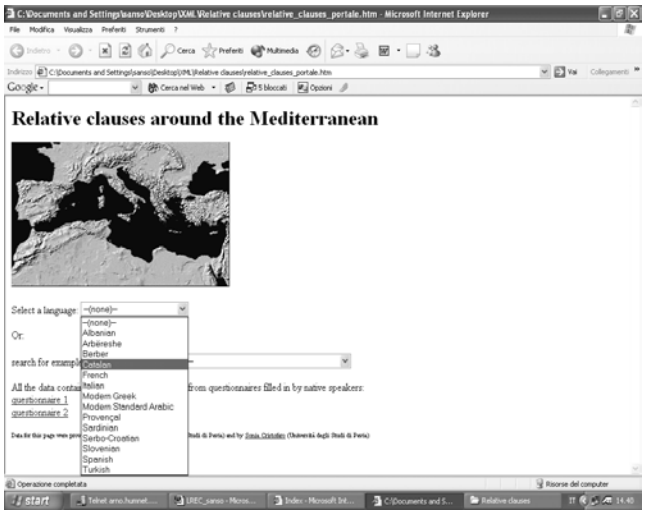

Figure 5: Selecting a language

Advantages over traditional typological databases

The database described in this paper is an easily browsable resource, which contains several interesting advantages over traditional typological databases. As already underscored, it is not dependent on proprietary solutions. The main advantage, however, is that it can be easily integrated into larger-scale resources such as annotated corpora, thus maximizing reusability and interoperability. In fact, the Med-Typ database is meant to become a subpart of a larger linguistic resource to be developed at the University of Pavia, namely an annotated corpus of Mediterranean languages designed for typological research. More generally, the flexibility of annotation schemes makes it possible to integrate in the same database both data drawn from grammars/questionnaires, and data from existing or new corpora.

To sum up, one of the main aims of the research program described in this paper is to put forward a concrete proposal for a best practice in the annotation of linguistic information for typological research. We hope that the results of this project will be of particular benefit to developers of typological databases aiming at maximal user-friendliness, descriptive appropriateness, interoperability, reusability, and computational efficiency.

Acknowledgements

The research reported here is supported by the Italian Ministry of Education (FIRB – Fondo per gli Investimenti della Ricerca di Base – Research program *Europa e Mediterraneo dal punto di vista linguistico: storia e prospettive* [code: RBNE01X7E7]). This support is gratefully acknowledged.

References

Bickel, B., Comrie, B. & Haspelmath, M. (2004). The Leipzig glossing rules. Conventions for interlinear morpheme-by-morpheme glosses. Leipzig: Max-Planck-Institut für Evolutionäre Anthropologie. (http://www.eva.mpg.de/lingua/files/morpheme.html)

Clark, J. (Ed.) (1999). XSL Transformations (XSLT) Version 1.0. W3C Recommendation 16 November 1999. (http://www.w3.org/TR/xslt)

Cristofaro, S. & Putzu, I. (Eds.) (2000). Languages in the Mediterranean area. Typology and convergence. Il progetto MEDTYP: studio dell'area linguistica mediterranea. Milano: FrancoAngeli.

Dahl, Ö. (2001). Principles of Areal Typology. In Haspelmath, M., König, E., Österreicher, W. & Raible, W. (Eds.), Language typology and language universals: an international handbook (pp. 1456--70). Berlin-New York: de Gruyter.

Dimitriadis, A. & Monachesi, P. (2002). Integrating different data types in a Typological Database System. In Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spain, 27 May – 2 June 2002.

Kay, M. (Ed.) (2003). XSL Transformations (XSLT) Version 2.0. W3C Working Draft 12 November 2003. (http://www.w3.org/TR/xslt20/)

König, E., Gast, V., Hole, D., Siemund, P. & Töpper, S. (2003). Typological database of intensifiers and reflexives. Berlin: Freie Universität. (http://noam.philologie.fu-berlin.de/~gast/tdir)

Lehmann, Ch. & Bakker, D. & Dahl, Ö. & Siewierska, A. (1994). EUROTYP Guidelines. Strasbourg: European Science Foundation (EUROTYP Working Papers) (2. ed.).

Lehmann, Ch (2001). Language documentation: a program. In Bisang, W. (Ed.), Aspects of typology and universals (pp. 83--97). Berlin: Akademie Verlag.

Lehmann, Ch. (2004a). Interlinear morphemic glossing. In Lehmann, Ch. et al. (Eds.), Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2. Halbband. Berlin-New York: de Gruyter.

Lehmann, Ch. (2004b). Data in linguistics. Linguistic Review 21 (3-4).

Monachesi, P., Dimitriadis, A., Goedemans, R. & Mineur, A.-M. (2002). A unified system for accessing typological databases. In Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spain, 27 May – 2 June 2002 (pp. 1029--1035).

Ramat, P. (2003). Il sardo tra le lingue del Mediterraneo. In Loi Corvetto, I. (Ed.), Dalla linguistica areale alla tipologia linguistica, Atti del Convegno della Società Italiana di Glottologia, Cagliari, 27-29 settembre 2001 (pp. 15--33). Roma: Il Calamo.

Ramat, P. & Stolz, Th. (Eds.) (2002). Mediterranean languages. Papers from the MEDTYP workshop, Tirrenia, June 2000. Bochum: Brockmeyer.

Sansò, A. (2003). Typological databases: A new approach. In Hajičová, E., Kotěšovcová, A., Mírovský, J. (Eds.), Proceedings of CIL17, CD-ROM. Prague: Matfyzpress, MFF UK. ISBN: 80-86732-21-5.

Stolz, Th. & Sansò, A. (forthcoming). The Mediterranean area revisited. Word-iteration as a potential Mediterraneanism. In McMahon, A., Vincent, N., & Matras, Y. (Eds.), Language contact and areal linguistics.

Tiberius, C., Brown, D. & Corbett, G. (2002). A typological database of agreement. In Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spain, 27 May – 2 June 2002 (pp. 1843--1846).