

The Pavia Typological Database

First module (1.0): Relative clauses

Andrea Sansò

1. Background

The **Pavia Typological Database** has been created within the framework of the research program *Europa e Mediterraneo dal punto di vista linguistico: Storia e prospettive* (“Europe and the Mediterranean from a linguistic point of view: History and perspectives”), sponsored by the Italian Ministry of Education (FIRB – *Fondo per gli Investimenti della Ricerca di Base*), and launched in 2003 under the direction of **Paolo Ramat** (Università di Pavia). The main aim of this research program is the typological documentation of morphosyntactic phenomena in languages belonging to the Euro-Mediterranean area, in order to describe the distribution of various structural traits within this area and to uncover phenomena of areal convergence. Researchers from the Universities of Pavia, Pisa, Rome (La Sapienza) and Siena (Università per Stranieri) take part in the program. The research program is a follow-up to another research program, called MEDTYP and sponsored by the Italian National Research Council (1997-2000; coordinator of the Programme, Prof. Romano Lazzeroni, University of Pisa). The MEDTYP project was concerned almost exclusively with languages in the Mediterranean area: the exploration of 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 neighbouring 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 MEDTYP project. On the contrary, the research agenda of the FIRB project explicitly involves the creation of a typological database, which contains both data coming from the MEDTYP project and new data collected from 2003 onwards.

The **first module** of the database (**1.0**) provides information on **strategies of relative clause formation** in 15 languages belonging to the Euro-Mediterranean area. The data in the database are samples of clauses/sentences elicited through questionnaires distributed to native speakers, and have been collected by **Sonia Cristofaro** and **Anna Giacalone** (Università di Pavia). (See section 2 “Theoretical assumptions”.)

Other modules are already planned: the second release (**version 2.0**) will contain information on the structure of noun phrases in the Mediterranean languages; subsequent modules will deal with a number of morphosyntactic domains (e.g., evaluative morphology, coordinating constructions, etc.), and will be

published on the web in the near future. The Pavia Typological Database aims to become both a collection of typological data of Euro-Mediterranean languages for future analyses of possible typological implications and areal distribution (quantitative typology) and a tool for the systematic analysis of the range of variation found in various typological domains (qualitative typology).

At the same time, the first module will be constantly updated with new data. As it is customary in the field of language resources, **any major update of the first module will be signalled through the use of a different serial number** (1.1, 1.2, and so on).

Suggestions concerning both the data and the make-up of the database are appreciated, and can be sent to andrea.sanso@unipv.it. Any piece of helpful information will be gratefully acknowledged.

2. Theoretical assumptions

The first module of the database records examples of relative clauses in 15 languages belonging to the Euro-Mediterranean area, including both standard and colloquial varieties. The typological framework used is basically the one by Keenan and Comrie (1977), but the classification of relative clauses benefits from recent advances in the understanding of their typology (e.g. Comrie 2002, Cristofaro and Giacalone Ramat 2002). The description of the typological features of relative clauses in languages of Europe and the Mediterranean is meant to show to what extent they reflect the typological tendencies ascertained for relative clauses crosslinguistically and, conversely, to what extent they can contribute to a more precise definition of such tendencies.

The definition of a relative clause is essentially semantic, and in line with the one by Keenan and Comrie (1977):

a syntactic object is a **relative clause** “if it specifies a set of objects (perhaps a one-member set) in two steps: a larger set is specified, called the **domain of relativization**, and then restricted to some subset of which a certain sentence, the **restricting sentence**, is true” (Keenan and Comrie 1977: 63-64).

Admittedly, this definition does not exclude from the collection of relevant data structures that do not strictly speaking qualify as relative clauses, but perform the same function as relative clauses and are sometimes produced by informants as a reply to our questionnaires. Two such examples are the German and Turkish structures below:

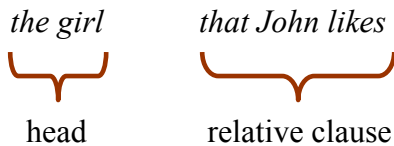
<i>Er</i>	<i>sah</i>	<i>den</i>	<i>in</i>	<i>seinem</i>	<i>Büro</i>	<i>arbeitenden</i>	<i>Mann</i>
he	see.IND.PST.3SG	ART.ACC	in	his	office	work.PTCP.PRS	man

‘He saw the man who works in his office.’

<i>Yaz-diğ-im</i>	<i>mektup</i>	<i>nerede?</i>
write-INF-1SG	letter	where

‘Where is the letter I wrote?’

The domain of relativization is expressed in surface structure by the **head NP** and the restricting sentence by the **relative clause** proper:



Languages differ with respect to the way RCs are formed: it is often the case that there is more than one distinct type of RC even within a single language. The questionnaires have been designed with a view to maximising variation in the semantic and grammatical features of both heads and relative clauses, in order to obtain a fully accurate picture of the relativization strategies for each language of the sample.

The relevant **dimensions of variation for heads** are the following:

1. indefinite vs. definite head (*I was told this by **a girl** who had dinner here* vs. *I was told this by **the girl** you are looking for*)
2. inanimate vs. animate head (*I found the solution in a **book** the teacher recommended yesterday* vs. *The **man** who had dinner here is my brother*)
3. nonreferential vs. referential head (*I am looking for a **woman** that has already done this job* vs. *I'm looking for **the woman** that you met at the party*)
4. head in core role vs. head in oblique role (***On each of the days** we met it rained* vs. ***The day** he told me that **was very important for both of us***)

The **main dimension of variation of relative clauses** is the role of the relativized element. Different relativization strategies differ with regard to which NPs positions they can relativize. For instance, the first relative clause below relativizes a subject (i.e., the head NP functions as the subject of the main verb of the restricting clause), whereas the second one relativizes an indirect object (i.e., the head NP functions as the indirect object of the main verb of the restricting clause):

German:

Er sah den Mann, der in seinem Büro arbeitet
 He see.IND.PST.3SG ART.ACC man who in his office work.IND.PRS.3SG
 'He saw the man who is working in his office.'

Italian:

Conosco la persona a cui mio fratello ha venduto il libro
 know.IND.PRS.1SG ART person to whom my brother have[AUX].IND.PRS.3SG
 sell.PTCP.PST ART book
 'I know the person to whom my brother sold the book.'

The sentences in the questionnaire involve a number of different syntactic functions for both the head and the relativized element. The main syntactic functions of the relativized element are:

- (1) **subject** (*The man **who had dinner here** is my brother*)
- (2) **object** (*The man **you met at the party** is my brother*)
- (3) **indirect object** (*The girl **you sold the book to** is John's sister*)
- (4) **possessor** (*The man **my sister met the brother of** is a very important person*)
- (5) **oblique**; oblique relativized elements are further subdivided according to the semantic role of the relativized element:

time (*On the day we met it rained*)

location (*A house where nobody can live is purposeless*)

instrument (*The knife my brother usually cuts the bread with is in the left drawer*)

companion (*The man my sister and I travelled with is a very important person*)

goal (*Every place we go to is crowded with people*)

manner (*The way in which one does it is very important*)

reason (*The reason why I did that is irrelevant to the present discussion*)

A major distinction within the inventory of relativization strategies (both within a single language and across languages) is the distinction between [+case] and [-case] strategies. A strategy is [+case] if it “presents a nominal element in the restricting clause that unequivocally expresses which NP position is being relativized. For instance, the English strategy that forms *the girl who John likes* (substituting the by now almost abandoned form *the girl whom...*) is not case-coding since *who*, the only relevant particle in the restricting clause, can be used as well if the role of the head NP in the restricting clause is different, e.g. *the girl who likes John*” (Keenan and Comrie 1977: 64). The following Russian example, in which there is a relative pronoun clearly encoding the role of the relativized element, is an example of [+case] strategy:

devuška, kotoruju Džon ljubit
girl who.ACC John likes
'the girl who John likes'

Comrie (2002) classifies this and similar strategies under the rubric of **relative pronoun strategy**. Another relativization strategy widespread across languages makes use of an invariable relative marker. By **invariable relative marker** is meant here a relative element that does not specify the role of the relativized item: under this definition, markers such as Italian *cui*, Portuguese *quem*, Spanish *quien*, French *qui*, and Greek *pou* are counted as invariable relative markers (just like Italian *che*), because they can be used to relativize several syntactic roles without changing their form (although they can be combined with adpositions indicating the role of the relativized item – this strategy is indicated in the pull-down menu as **adposition + invariable relative marker**, and is considered to be a [+case] strategy).

In addition to the use of relative pronouns and invariable relative markers, other strategies are possible among the languages of the sample. One of them is the so-called **resumptive pronoun strategy** (or **pronoun retention strategy**, cf. Comrie 2002), in which a personal pronoun refers back to the head, as in the colloquial English example *a girl that her eighteenth birthday was on that day*: this is a very diagrammatic strategy used in many substandard varieties (cf. Fiorentino 1999). Another type is the **gap strategy** (or **gapping**), where there is no explicit relative marker, and no personal pronouns within the restrictive clause, as in the following Tunisian Arabic example:

Bi'-t le-ktāb l-tufla ti-xdem hūni
sell-PFV.1SG ART-book to-girl IPV.3SG.F-work here
'I sold the book to a girl who is working here.'

There are some complex cases in which the relativization strategy is more difficult to classify. One such example is the so-called **relative pronoun strategy** in Modern Standard Arabic:

Aš-šaxs-u al-laḏī 'iltaqā-hu ax-ī bilamsi
ART-person-NOM ART-REL_PRO.NOM.SG.M meet.PFV.3SG.M-him brother-my yesterday

ṣadīqun *lī*
 friend my
 ‘The person my brother met yesterday is a friend of mine.’

As Comrie (2002: 89ff.) notes, the nominative case of the relative pronoun does not encode the role of the head in the relative clause (one would expect accusative, not nominative case). The nominative case on the pronoun is by agreement with the head, and the relative clause contains a resumptive pronoun, so that the clause is an instance of the *pronoun retention* (or *resumptive pronoun*) *strategy* (cf. also Cristofaro and Giacalone Ramat [2002: 100], who talk of a *non-resumptive relative pronoun*).

Another important feature of the database is that it collects **a significant amount of examples of relativization of circumstantials**. As Cristofaro and Giacalone Ramat (2002) show, in a number of Mediterranean languages, [-case] strategies are frequently used for the relativization of circumstantials such as time, and occasionally place, manner, and reason. Moreover, there appears to be a crucial correlation between the use of [-case] strategies for the relativization of time circumstantials and the semantics of head nouns: [-case] strategies “seem to be favored whenever the head noun is less specific and less referential” (Cristofaro and Giacalone Ramat 2002: 105). The database is able to provide a fully-fledged description of the relativization strategies used for the relativization of circumstantials by combining information about the head nouns and the role of relativized elements.

3. Architecture of the database

There are two kinds of typological databases. Firstly, there are databases that collect and document primary language data (e.g. the agreement database, cf. Tiberius et al. 2002; the Typological Database of Intensifiers and Reflexives, cf. König et al. 2003). Secondly, there are databases collecting secondary language data, such as the Universals Archive developed at the University of Konstanz (<http://ling.uni-konstanz.de>).

The Pavia Typological Database belongs to the first class to full right. It provides primary information on relative clauses without including any typological generalization. The primary data collected in the 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 primary data contained in the 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, 2004). 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 enterprise 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 solutions with the documents, or by deriving from those solutions ideas 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.

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. This is an example of a Slovenian relative clause annotated with basic grammatical and semantic information:

```
<ITEM id="slo_030" source="Janez Orešnik">
<MAIN_CLAUSE tense="PST">
<w id="slo_030_01" gl="book.ACC">Knjigo </w>
<w id="slo_030_02" gl="be[AUX].IND.PRS.1SG">sem </w>
<w id="slo_030_03" gl="sell.PTCP.SG.M">prodal </w>
<HEAD syn_function="IO" referential="y" animate="y" definite="n">
<w id="slo_030_04" gl="girl.DAT">deklici, </w>
</HEAD>
<RELATIVE_CLAUSE tense="PST" relative_pronoun="n" case="n" resumptive_pronoun="y"
complementizer="y">
<RELATIVIZER syn_function="OBJ" variable="n">
<w id="slo_030_05" gl="REL">ki </w>
</RELATIVIZER>
<w id="slo_030_06" gl="PRO.ACC.3SG.F">jo </w>
<w id="slo_030_07" gl="be[AUX].IND.PRS.3SG">je </w>
<w id="slo_030_08" gl="POSS.1SG">moj </w>
<w id="slo_030_09" gl="brother">brat </w>
<w id="slo_030_10" gl="meet.PTCP.M.SG">srečal </w>
<w id="slo_030_11" gl="at">na </w>
<w id="slo_030_12" gl="party">zabavi </w>
</RELATIVE_CLAUSE>
</MAIN_CLAUSE>
<TRANSLATION>I sold the book to a girl that my brother met at the
party</TRANSLATION>
<COMMENT>This clause has a variant (slo_030_a) with the relative pronoun in the ac-
cusative form</COMMENT>
</ITEM>
```

The technicalities of the annotation procedure are discussed in further detail in Sansò (2003; 2004) and Ramat and Sansò (to appear). Suffice it to list here some basic features of the annotation practice followed in the creation of the database:

- 1) Each word of the clause is enclosed between <w> tags and has a unique identifier (encoded by means of the attribute id);
- 2) each <w> element is provided with a morpheme-by morpheme gloss;
- 3) a number of upper-level units are singled out by means of specific labels: these are the <HEAD>, the <RELATIVE_CLAUSE>, and the <RELATIVIZER>;
- 4) any of these upper-level units is provided with a number of attributes conveying grammatical and semantic information;
- 5) the English translation of the clause is enclosed between <TRANSLATION> tags;
- 6) the label <COMMENT> provides any additional information concerning the clause in question (e.g., comments provided by the source of information).

For each linguistic phenomenon contained in the first draft of the database a Document Type Declaration (DTD) has been created that includes the whole set of tags used in the annotation of that phenomenon. The DTD for the class of XML documents collecting relative clauses is displayed below as exemplification of the kind of linguistic traits that are relevant to this phenomenon:

```
<!ELEMENT RELATIVE_CLAUSES (ITEM+)>
<!ATTLIST RELATIVE_CLAUSES language CDATA #REQUIRED>
<!ELEMENT ITEM (MAIN_CLAUSE, COMMENT*, TRANSLATION)>
<!ATTLIST ITEM id CDATA #REQUIRED>
<!ATTLIST ITEM source CDATA #REQUIRED>
<!ELEMENT MAIN_CLAUSE (w+|HEAD|RELATIVE_CLAUSE)*>
<!ATTLIST MAIN_CLAUSE tense (PRS|PST|FUT) #REQUIRED>
<!ELEMENT w (#PCDATA)>
<!ATTLIST w id CDATA #REQUIRED>
<!ATTLIST w gl CDATA #REQUIRED>
<!ELEMENT HEAD (w+)>
<!ATTLIST HEAD syn_function (SBJ|OBJ|IO|POSS|OBL) #REQUIRED>
<!ATTLIST HEAD referential (y|n) #REQUIRED>
<!ATTLIST HEAD animate (y|n) #REQUIRED>
<!ATTLIST HEAD definite (y|n) #REQUIRED>
<!ELEMENT RELATIVE_CLAUSE (w+|RELATIVIZER)*>
<!ATTLIST RELATIVE_CLAUSE tense (PRS|PST|FUT) #REQUIRED>
<!ATTLIST RELATIVE_CLAUSE relative_pronoun (y|n) #REQUIRED>
<!ATTLIST RELATIVE_CLAUSE case (y|n) #REQUIRED>
<!ATTLIST RELATIVE_CLAUSE resumptive_pronoun (y|n) #REQUIRED>
<!ATTLIST RELATIVE_CLAUSE complementizer (y|n) #REQUIRED>
<!ELEMENT RELATIVIZER (w*)>
<!ATTLIST RELATIVIZER syn_function (SBJ|OBJ|IO|POSS|OBL) #REQUIRED>
<!ATTLIST RELATIVIZER role (instrument|location|time|companion|goal|reason|manner|possessor)
#IMPLIED>
<!ATTLIST RELATIVIZER variable (y|n) #IMPLIED>
<!ELEMENT TRANSLATION (#PCDATA)>
<!ELEMENT COMMENT (#PCDATA)>
```

4. Understanding and making queries

Queries are made possible through the use of the XSLT language (eXtensible Stylesheet Language Transformations Clark 1999; Kay 2003). XSLT is a functional programming language optimized for parsing and generating XML documents. An XSL 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.

The first module of the database (relative clauses) is queried by using pulldown menus for 4 different fields (Language, Head, Relativization strategy, and Role of the relativized element). Once the appropriate value has been chosen, clicking on the “Send query” button will output all records in the database which match the selected values. In the particular case displayed in figure 1, we will obtain all records for Catalan in which the head is definite.

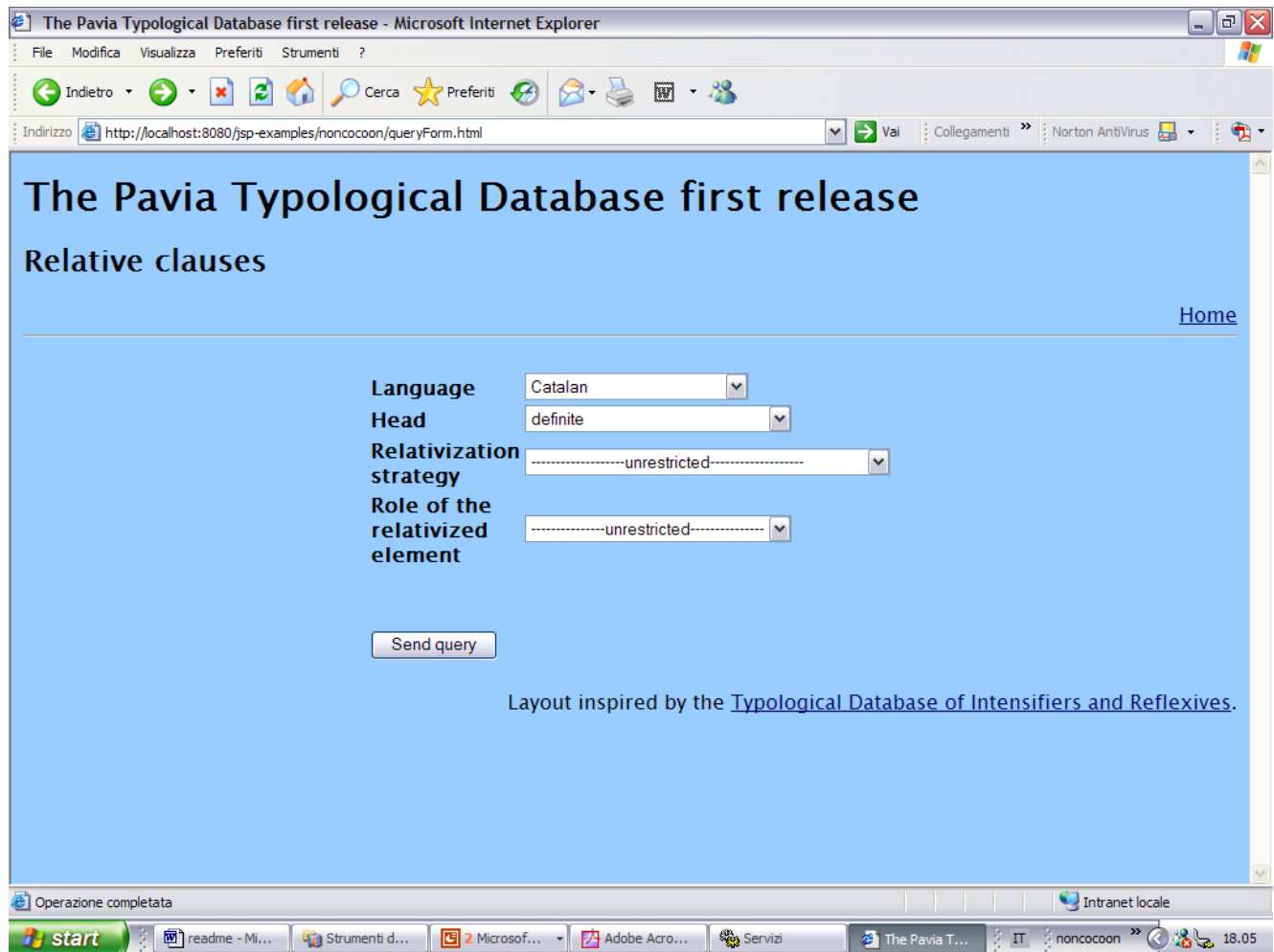


Figure 1: Selecting a language and a type of head

Once you have created your query and clicked on the ‘Send query’ button, the database will return all records which match the query. The number of records returned will obviously depend on which value you select and/or on how specific the query is. The data are displayed on the screen as a triple, *example + gloss + translation*, i.e. without including any judgment or typological generalization:

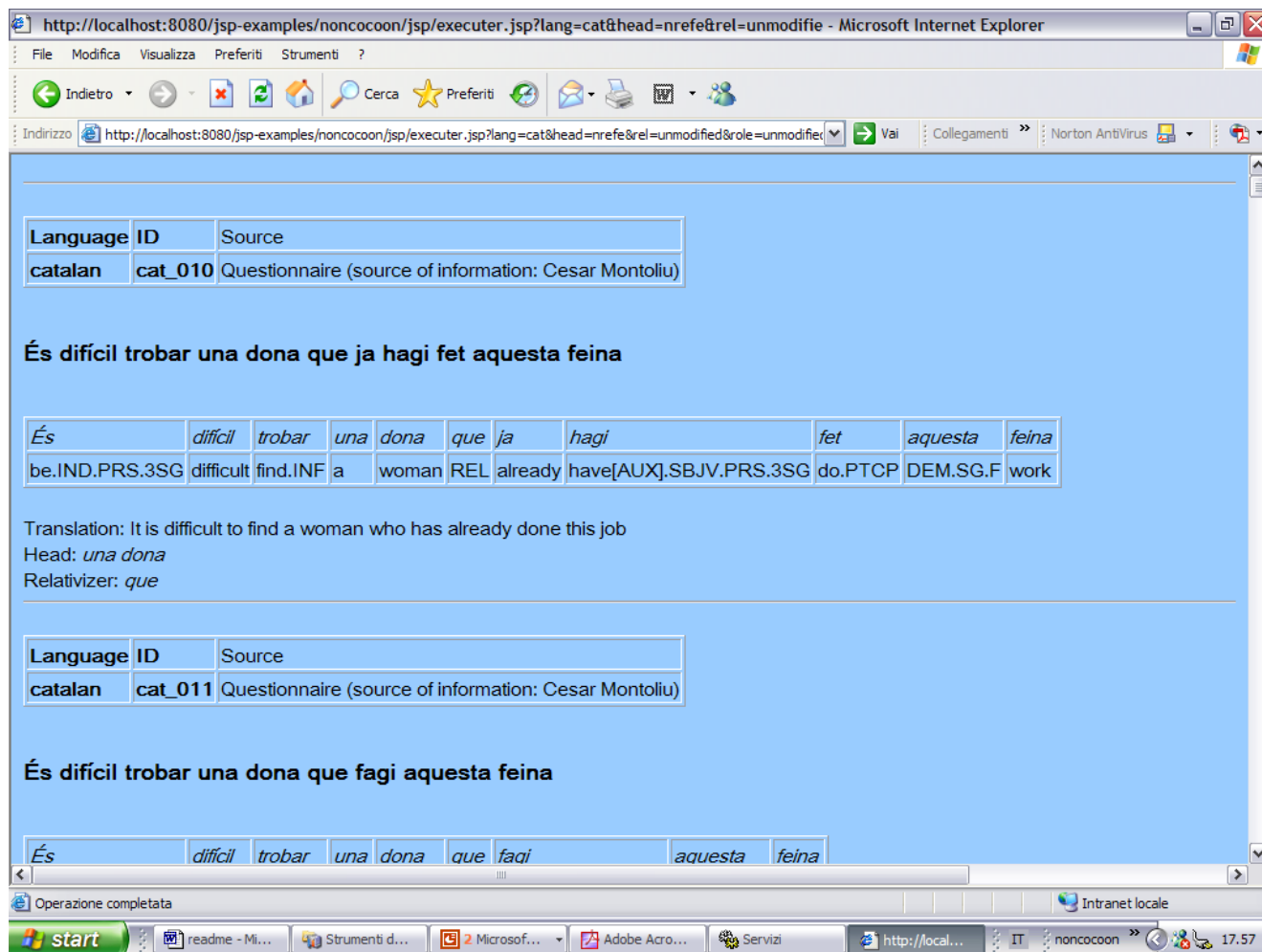


Figure 2: The output

The first two lines of each returned record contain three fields, which are fairly self-explanatory: the language, the ID of the record (a unique identifier), and the source from which the example is taken (grammar, questionnaire, etc.). Then come the clause (in bold), a table containing the interlinear morphemic gloss, the translation, and information concerning the head and the relativizer. If there are some comments available for a specific record, they are displayed at the end of each record. A horizontal line separates examples from one another.

References

- Bickel, B., B. Comrie and M. Haspelmath. 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>).
- Comrie, B. 2002. "Rethinking relative clause types: the Mediterranean area". In P. Ramat and Th. Stolz (eds.), *Mediterranean languages: Papers from the MEDTYP workshop, Tirrenia, June 2000*, 87-98. Bochum: Universitätsverlag Dr. N. Brockmeyer.

- Cristofaro, S. and A. Giacalone Ramat. 2002. "Relativization patterns in Mediterranean languages, with particular referene to the relativization of time circumstantials". In P. Ramat and Th. Stolz (eds.), *Mediterranean languages: Papers from the MEDTYP workshop, Tirrenia, June 2000*, 99-112. Bochum: Universitätsverlag Dr. N. Brockmeyer.
- Dahl, Ö. 2001. "Principles of areal typology". In M. Haspelmath, E. König, W. Österreicher and W. Raible (eds.), *Language typology and language universals: An international handbook*, 1456-70. Berlin-New York: de Gruyter.
- Fiorentino, G. 1999. *Relativa debole. Sintassi, uso, storia in italiano*. Materiali Linguistici 24. Milan: Francoangeli.
- Kay, M. (ed.). 2003. XSL Transformations (XSLT) Version 2.0. W3C Working Draft, 12 November 2003 (<http://www.w3.org/TR/xslt20/>).
- Keenan, E.L. and B. Comrie. 1977. "Noun phrase accessibility and universal grammar". *Linguistic Inquiry* 8: 63-99.
- König, E., V. Gast, D. Hole, P. Siemund, and S. Töpper. 2003. Typological database of intensifiers and reflexives. Berlin: Freie Universität (<http://noam.philologie.fu-berlin.de/~gast/tdir>).
- Ramat, P. and A. Sansò. To appear. "Per una definizione dell'area linguistica mediterranea". In V. Orioles and F. Toso (eds.), *Atti del Convegno Internazionale "Mediterraneo Plurilingue"*, Genova, 13-15 maggio 2004.
- Sansò, A. 2003. "Typological databases: A new approach". In E. Hajičová, A. Kotěšovcová and J. Mírovský (eds.), *Proceedings of CIL17*, CD-ROM. Prague: Matfyzpress, MFF UK. ISBN: 80-86732-21-5.
- Sansò, A. 2004. "MED-TYP: A Typological Database for Mediterranean Languages". In M.T. Lino et al. (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 1157-1160. Lisbon: Porto Editora.
- Tiberius, C., D. Brown and G. Corbett. 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, 1843-1846.