Documenting variation across Europe and the Mediterranean:

The Pavia Typological Database

Andrea Sansò

Dipartimento di Linguistica Teorica e Applicata – Università di Pavia
Corso Strada Nuova 65, I-27100, Pavia, Italy
asanso@gmail.com

Abstract

This paper describes the Pavia Typological Database (PTD), a follow-up to the MED-TYP database (Sansò 2004). The PTD is an ever-growing repository of primary linguistic data (words, clauses, sentences) documenting a number of morphosyntactic phenomena in the languages of Europe (and including in some cases languages from the Mediterranean area). Its prospective users are typologists wanting to access primary, typologically uninterpreted (but glossed) data, but also anyone interested in linguistic variation on a continental scale. The paper discusses the background and motivation for the creation of the PTD, its present coverage, the techniques used to annotate the primary data, and the general architecture of the database.

1. Background: Areal typology and Standard Average European

The last three decades have witnessed an increasing popularity of areal studies within linguistic typology (see Campbell et al. 1986; Nichols 1992; Haspelmath 1998, 2001; Dahl 2001; Koptjevskaja-Tamm and Wälchli 2002; Ramat and Stolz 2002; Haspelmath et al. 2005, among others). The primary object of these studies is not (or, perhaps, no longer) the identification of *Sprachbünde* (or *linguistic areas*, as they are often referred to in English). As Dahl (2001: 1456) puts it, "areal patterns [of distribution of linguistic features] are of interest irrespective of whether they can be described in terms of linguistic areas in the traditional sense": in fact, hardly any typological variable is evenly distributed over the world, and most reveal systematic areal skewings. For example, inclusive/exclusive distinctions, or numeral classifiers show frequency peaks around the Pacific, clicks are found in substantial frequency only in Africa, relative pronouns only in Europe, etc. As a result of this renewed interest in areal issues, a brand-new discipline has come to the fore, which combines insights from both traditional typological studies and areal linguistics (i.e., the discipline concerned with similarities between geographically contiguous languages which has its roots in traditional dialectology and sociolinguistics). This new discipline, *areal typology*, can be defined as the study of the areal distribution of typologically relevant features of languages.

Many European languages (Romance, Germanic, Balto-Slavic languages, but also, more marginally, the westernmost Finno-Ugrian languages, as well as Maltese) share a number of structural features which give these languages a very peculiar profile and make them stand out among the world's languages. This "exotic" character of European languages is a relatively new insight, as only recently "linguists realized how peculiar the core European languages are in some ways when seen in the world-wide context" (Haspelmath 2001: 1492). The EUROTYP research program, launched in the early nineties by the European Science Foundation, has produced a critical mass of typological and areal studies dealing with dozens of grammatical features that are characteristic of the core European languages, and that together define the Standard Average European area.

2. The Pavia Typological Database (PTD)

Such typological findings have renewed a strong interest in the study of Europe as a linguistic area. In 2003, 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*), was launched under the direction of Paolo Ramat (Università di Pavia). The main aim of this research program has been 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 both micro- and macro-areal convergence. This research program is a follow-up to another research program, called MEDTYP and sponsored by the Italian National Research Council (1997-2000), which was concerned almost exclusively with languages in the Mediterranean area.

The PTD has been created within the framework of the FIRB research program. Its main goal is to make the amount of data collected within this and the previous research program available to the community of typologists. These data were only partly in electronic form, and were at risk of evaporating in the short term.

2.1. Typological databases

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 (see below, §4). 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.

Over the last few years, it has become increasingly clear that the task of typological databases is not only the electronic storage and automatic retrieval of the data, but also the determination of the range of potential variation

within the area surveyed in the database, both outlining the 'nonesuches' of language structure in that area and identifying common tendencies in the languages surveyed.

Two kinds of typological databases can be distinguished according to their structural properties. Firstly, there are databases that collect and document primary language data (e.g. the *Agreement Database*, Tiberius et al. 2002; the *Typological Database of Intensifiers and Reflexives*, König et al. 2003). Secondly, there are databases collecting secondary language data, such as the Universals Archive and the *grammatisches Raritätenkabinett*, both developed at the University of Konstanz (**http://ling.uni-konstanz.de**), and the World Atlas of Language Structures (WALS, Haspelmath et al. 2005). The PTD belongs to the first kind to full right. It provides primary information on a number of morphosyntactic phenomena without including any typological generalization. The primary data collected in the database are mainly samples of clauses/sentences (or lists 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 to 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 following (non-exhaustive) list provides some possible questions that a potential user should like to ask when accessing the PTD:

(i)   is there any information on linguistic feature/phenomenon x available in a language L?
(ii)   what languages have/do not have $x_1, x_2…$?
(iii)   which languages have x but not y? and vice versa
(iv)   which languages have neither x, nor y?
(v)   when a L has x_1, does it always have x_2?
(vi)   how often x occurs in a family F?
(vii)   what is the areal distribution of x?

Question (vii) cannot be directly answered by the PTD, although the raw data allowing areal generalizations can be easily plotted onto a map. This is a motivated choice, because in our view a typological database "of the first kind" should contain as few typological and areal generalizations as possible (see below, §3).

2.2. Present coverage

In this section, I will illustrate the present coverage of the PTD, both in terms of languages already included in the database, and in terms of modules of the database.

2.2.1. Languages

Ideally, a typological database documenting linguistic variation in Europe should contain (at the very least) information on the national languages and the major non-national varieties spoken throughout Europe. Of course, the degree of granularity of language documentation is a debated and sensible issue in any areal-typological enterprise. In order to maximize completeness, a two-step procedure has been adopted. In the first stage, we aimed at providing complete data and information on the languages contained in the following box.

Languages in the PTD: The nucleus
Albanian, Basque, Belarusian, Breton, Bulgarian, Catalan, Czech, Danish, English, Estonian, Finnish, French, Galician, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Macedonian, Maltese, Manx, Norwegian, Polish, Portuguese, Rumanian, Russian, Sardinian, Serbo-Croatian, Slovenian, Slowak, Sorbian, Spanish, Swedish, Turkish, Welsh.

In the second stage, data on a number of circum-European languages have been included whenever available. Part of them are consistently and massively represented in the MED-TYP database (Sansò 2004), from which the PTD stems out, and were migrated from there into the PTD. Some cases in points are Arabic, Arabic varieties, and Modern Hebrew. This fact should ideally make the database a complete source of documentation on the whole Euro-Mediterranean domain. Other languages occasionally included are languages of the Caucasus, as well as any other extra-European languages for which relevant data were available (e.g. Georgian, Japanese, Korean, and so on). Data on non-standard, regional and dialectal varieties are generally stored under their national variety, but their peculiar character is always explicitly signalled when the data are displayed on the screen.

2.2.2. Morphosyntactic phenomena

The following morphosyntactic modules are already available on the home page of the PTD:

(a) **Relative Clauses**: this module provides information on strategies of relative clause formation in 20 languages (Albanian, Catalan, Finnish, French, Galician, German, Greek, Italian, Maltese, Sardinian, Serbo-Croatian, Slovenian, Spanish, Swedish, Arabic and Arabic varieties, Modern Hebrew); other languages (Irish, Manx, Welsh, Czech, Bulgarian, etc.) are being currently added;

(b) **Action Nominal Constructions**: this module contains information about the morphology and the subcategorization properties of Action Nominals in a number of European languages. Data are still being inserted, but a demo of this module is already available. This demo directly operates on the XML file containing the primary data, so that it progresses at the same rate as the insertion of new data;

(c) **Possessive Noun Phrases**: this module contains information about the morphosyntactic and semantic properties of possessive noun phrases in a number of European languages. The insertion of data, collected from both questionnaires and grammars, is almost complete.

(d) **Coordinating constructions**: the module contains primary data gathered through a 35 question questionnaire on three types of coordinating constructions: conjunctive constructions (e.g. *A and B*), disjunctive constructions (*A or B*), and adversative constructions (*A but B*).

(e) **Deictic elements**: in this module real language examples of deictic elements (adverbs and demonstratives) are stored. This module has the shape of a parallel corpus, and it is based on the

translations of *Harry Potter and the Chamber of Secrets* in many European languages (cf. Da Milano, to appear). Two query forms are available: the first one allows the user to select some semantic parameters (e.g. distance from the speaker, distance from the hearer, and so on), while the second one allows the user to get the translated examples starting from the original English source.

2.2.3. Accessibility

Though still in progress, the PTD is already available on the web, at the URL **www.unipv.it/paviatyp**. Any update is carefully signalled in a log. This is necessary for an electronic resource which is still in progress: the potential user is always aware of any insertion of new data and can run new queries as soon as new data are available or old data are corrected and refined. The website also contains a large array of documentation files explaining and documenting transcription and annotation practices, as well as a contact form which enables the user to interact directly with the creators of the database, to signal any error or mis-interpretation of data, and to provide new examples.

3.   Architecture of the database

3.1.   Design

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 portability (in the sense of Bird and Simons 2003) when creating linguistic resources is essential. In our view, portability 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 consisted in creating a uniform, possibly theory-neutral annotation scheme for this kind of data. The range of phenomena annotated posed a considerable challenge to any attempt to adapt existing annotation practices, predominantly designed for annotating written texts or dialogues.

The technicalities of the annotation procedure are discussed in further detail in Sansò (2003; 2004) and Ramat and Sansò (to appear), and will not be resumed. Suffice it to list here some basic features of the annotation practice followed in the creation of the database:

(i)   each word of the examples is enclosed between **<w>** tags and has a unique identifier (encoded by means of the attribute **id**);

(ii)   a number of upper-level units are singled out by means of specific labels (e.g. **<head>**, **<relativizer>**, **<deixis>**, **<PNP>**, **<co-ordinating_construction>**, etc.);

(iii)   any of these upper-level units is provided with a number of attributes (e.g. **definite[y|n]**, **variable[y|n]**, **animate[y|n]**, etc.) conveying grammatical and semantic information;

(iv)   the most important and pervasive of these attributes is **gl**, i.e. the morphemic gloss of the word, compliant with the Leipzig Guidelines (cf. Bickel et al. 2004); this gloss guarantees full readability of the data also by non-typologists;

(v)   the English translation of the clause is enclosed between **<translation>** tags;

(vi)   the label **<comment>** provides any additional information concerning the clause in question (e.g., comments provided by the source of information) which cannot be represented in terms of yes-no feature-value pairs;

(vii)   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 DTDs are publicly available on the PTD website.

3.2.   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.

By way of illustration, let us consider the first module of the database (relative clauses): it is queried by using pulldown menus for four 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 the Catalan relative clauses in which the head is definite.

Figure 1. *Selecting a language and a type of head.*

Once one has created his 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 one selects 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, and the translation. 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. The same layout is used in all the modules of the database.

3.3. Executing queries

Queries are executed by means of JavaServer Pages technology (JSP). JSP is a Java technology that allows software developers to dynamically generate HTML, XML or other types of documents in response to a Web client request. The technology allows Java code and certain pre-defined actions to be embedded into static content. JSP pages easily combine static templates, including HTML or XML fragments, with code that generates dynamic content. More concretely, a JSP page is simply an HTML web page that contains additional bits of code that execute application logic to generate dynamic content. For example, a JSP page may contain HTML code that displays static text and graphics, as well as calls to objects that access a database; when the page is displayed in a user's browser, it will contain both the static HTML content and dynamic information retrieved from the database. The separation of user interface and program logic in a JSP page allows for a very convenient delegation of tasks between web content authors and developers. It also allows developers to create flexible code that can easily be updated and reused. Because JSP pages are automatically compiled as needed, web authors can make changes to the presentation code without recompiling the application logic, and vice versa. This makes JSP a very flexible method of generating dynamic web content.

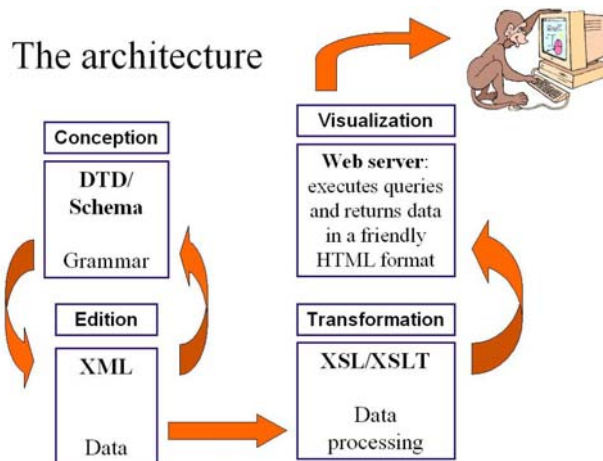The general architecture of the database is schematized in the following figure:

Figure 3. *The architecture of the Pavia Typological Database.*

4. Future work: Standardization, diffusion, evaluation

Despite their popularity, there are only a few initiatives towards standardization of typological databases (Monachesi et al. 2002, Dimitriadis and Monachesi 2002), and available databases can be quite heterogeneous. XML is increasingly used instead of proprietary software in the creation of typological databases of the first kind (cf., for instance, the database on Information Structure described by Dipper et al. 2006), although this still appears to be a minority choice.

All in all, the typologist wanting to access typological databases can now be helped in his/her search by some useful interfaces such as the Typological Database System, which is designed as an online portal to multiple independently developed typological databases (cf. Saulwick et al. 2006; the working prototype of this system can be accessed at `languagelink.let.uu.nl/tds`). On the other hand, the need for standards which 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) makes the time particularly ripe to address the urgent issue of standardization in typological databases.

According to the classification of archiving practices proposed by the E-MELD project (cf. Dry and Simons 2006, and **http://emeld.org**), an archiving practice is **good** if it creates a resource with the potential for long term preservation, **better** if it also promotes easy interpretation by future generations, and **best** if it facilitates automated interoperation with resources developed by other linguists. The PTD has been created following a "better" practice: XML guarantees longevity, and the amount of documentation available on its website makes it easily interpretable over the years. We also believe that our archiving practice has the potential to become "best": the use and adaptation of XML available standards to encode the information stored in the database will no doubt facilitate the interoperation of the PTD with similar resources that have been independently archived.

In the meantime, the PTD has become partner of the Language Typology Resource Center (LTRC), based in Utrecht. This center is intended to become a resource pool for language typology. It is being created by the joint effort of a range of members of the European typological community. The goal is to develop a web-accessible archive with materials and links that are helpful to typologists. The LTRC network has also started an e-mail list for the discussion and development of meta-data standards for typological resources, especially electronically accessible ones.

In addition to this, the PTD is joining the Open Archive Initiative (OAI), and the OLAC-metadata for this resource are being compiled. These include a set of standards that are currently followed by the OLAC archives and services (Simons and Bird 2003), and which guarantee wide accessibility and localizability to language resources.

The evaluation work on the PTD is still in its infancy. The database has been extensively used and queried by the restricted community of the FIRB research program, but no doubt its usage within a larger community will underscore new problems and drawbacks of this resource. The lack of evaluation procedures and benchmarks for typological databases is a stumbling block to the task of evaluating the PTD, but only to a limited extent. The main efforts of the typological community in Pavia are currently being devoted to a serious scrutiny of the potentialities of this resource, and to the enhancement of the portability characteristics that have been singled out as crucial for other types of language archives and resources (Bird and Simons 2003).

To conclude, we hope that the results of our work will be of particular benefit to developers of typological databases aiming at maximal userfriendliness, descriptive appropriateness, interoperability, reusability, and computational efficiency, and that our experience could usefully contribute to the discussion on practices of language archiving and documentation.

6. References

Bickel, B., B. Comrie, and M. Haspelmath (2004). The Leipzig glossing rules. Conventions for interlinear morpheme-by-morpheme glosses. Leipzig: Max-Planck-Institut for Evolutionary Anthropology. (**www.eva.mpg.de/lingua/files/morpheme.htm**).

Bird, S., and G. Simons (2003). Seven dimensions of portability for language documentation and description. *Language* 79 (3), pp. 557-582.

Campbell, L., T. Kaufman, and Th. C. Smith-Stark (1986). Meso-America as a linguistic area. *Language* 62, pp. 530-570.

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

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*, Volume 2, pp. 1456-70. Berlin-New York: de Gruyter.

Da Milano, F. (to appear). Demonstratives in parallel texts: A case study. *Sprachtypologie und Universalienforschung* [special issue on parallel corpora and typology, guest editor Michael Cysouw].

Dimitriadis, A., and P. Monachesi (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.

Dipper S., M. Götze, and S. Skopeteas (2006). A typological language archive for researching information structure. Paper presented at the Workshop on Language Archives, Standards, Creation, and Access, 28th Annual Meeting of the *Deutsche Gesellschaft für Sprachwissenschaft*, Bielefeld (Germany), 22-24 February 2006.

Dry, H. and G. Simons (2006). The experience of the E-MELD Project in developing and evaluating digital archiving recommendations. Paper presented at the Workshop on Language Archives, Standards, Creation, and Access, 28th Annual Meeting of the *Deutsche Gesellschaft für Sprachwissenschaft*, Bielefeld (Germany), 22-24 February 2006.

Haspelmath, M. (1998). How young is Standard Average European? *Language Sciences* 20 (Special issue on Areal Typology, guest editor Paolo Ramat), pp. 271-287.

Haspelmath, M. (2001). The European linguistic area: Standard Average European. In: M. Haspelmath, E.

König, W. Österreicher, and W. Raible (eds.), *Language Typology and Language Universals. An International Handbook*, Volume 2, pp. 1492-1510. Berlin-New York: de Gruyter.

Haspelmath, M., M.S. Dryer, D. Gil, and B. Comrie (eds.) (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press.

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

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

Koptjevskaja-Tamm, M., and B. Wälchli (2002). Circum-Baltic languages: An areal-typological approach. In: Ö. Dahl, and M. Koptjevskaja-Tamm (eds.), *Circum-Baltic Languages*, Vol. I, pp. 615-750. Amsterdam-Philadelphia: John Benjamins.

Lehmann, Ch. (2001). Language documentation: a program. In: W. Bisang (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, pp.1834-1857. Berlin-New York: de Gruyter.

Lehmann, Ch. (2004b). Data in linguistics. *Linguistic Review* 21 (3-4), pp. 275-310.

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

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. Chicago: The University of Chicago Press.

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.

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

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)*, pp. 1157-1160. Lisbon: Porto Editora.

Saulwick, A., M. Windhouwer, R. Goedemans, A. Dimitradis (2006). Architecture and procedures for the integration of linguistic databases in the Typological Database System. Paper presented at the Workshop on Language Archives, Standards, Creation, and Access, 28th Annual Meeting of the *Deutsche Gesellschaft für Sprachwissenschaft*, Bielefeld (Germany), 22-24 February 2006.

Simons, G., and S. Bird (eds.) (2003). OLAC Metadata. 8 December 2003. (`www.language-archives.org/OLAC/metadata.html`)

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*, pp. 1843-1846.