# Integrating different data types in a Typological Database System

**Alexis Dimitriadis, Paola Monachesi**

Utrecht University, UiL-OTS
Trans 10, 3512 JK Utrecht, The Netherlands
{*Alexis.Dimitriadis,Paola.Monachesi*}*@let.uu.nl*

**Abstract**

The aim of the *Typological Database System* project is the creation of a unified interface to numerous independently developed typological databases, which will allow the user to simultaneously query them from a single gateway. The main challenge behind the project lies in the great variability of the included data. In order to provide a unified interface the system will rely on detailed metadata, which will describe the content of each component database in terms of a common description framework. The common framework will be organized into an ontology of linguistic terms and notions, including alternative definitions, glossing standards, and database specific notions.

## 1. Introduction

Typology, the study of the variation that language exhibits, is one of the most important and interesting fields of linguistics. Typological databases are a valuable tool for this enterprise; a number of them have been developed by researchers in the field, often for personal or small-group use. Increasingly, these databases are being made available to the linguistic community over the Internet, providing the potential for enormous increases in the power of exploratory typological investigation.

However, these databases can be quite heterogeneous. A typologist seeking information on a particular subject may find parts of it scattered over several databases, organized in various forms and expressed in ways that reflect different research traditions. As the number of potentially relevant databases increases, so does the amount of effort required by a user to locate them, understand how they are organized, figure out the query system and perform a query, and interpret the results.

The aim of the *Typological Database System (TDS)* project (Monachesi et al., 2002) is to facilitate this process, by developing a software system that allows a user to simultaneously query many different typological databases through a single interface. This system, which is currently under development, will reside on a server computer that a user can query over the internet by use of a standard web browser. The component databases can in principle reside in separate, remote servers, although for performance reasons the server may need to maintain local copies of some or all of them. By accessing the single gateway site, the user will gain access to the data contained in all the databases participating in the project. The goal is for the system to behave as much as possible like a single, *virtual* database.

The TDS project (*http://www-uilots.let.uu.nl/td/*) is being carried out by a research group in the Netherlands Graduate School of Linguistics (LOT), with members representing the Universities of Amsterdam, Leiden, Nijmegen and Utrecht. A number of typological databases developed by participating researchers constitute the initial components of the TDS, and have been providing us with concrete experience on the problems that need to be addressed.

## 2. Obstacles to combining the databases

The aim of the TDS project is to combine diverse databases and present them to users as a unified virtual database. The challenge lies in the great heterogeneity of the included data, which can have several sources.

**Diversity of content type** Typological databases usually consist of logical variables describing each language as a whole. Example 1 illustrates a fragment of the *Typological Database Nijmegen*, which employs variables that encode information about various language phenomena. For example, variable *V456* states whether there is agreement between the subject and the verb. However, several of the component databases in this project contain example sentences with detailed annotations, as can be seen in example 2, a sample from the *Spinoza database* with several levels of description.

The ultimate goal of the project is to integrate different types of content so that, for example, a single query could return both examples and logical variables as an answer.

**Diversity of theoretical commitments** Because there is no single, universally accepted, and exhaustive linguistic theory, the information in the various databases reflects the analytical and theoretical commitments of its creators. A linguist can recognize the descriptive content of a statement based on identifiable assumptions, and the TDS project will place a high priority on preserving and making visible the framework of assumptions that will allow a (linguist) user to properly interpret any data extracted from a component database. Provided that this is ensured, it can be very useful to return information that only approximately matches the theoretical framework of the user.

**Diversity in form** In many cases the different databases use equivalent, or near-equivalent, ways of describing data. One obvious example is the use of different abbreviations for broadly accepted linguistic notions, such as *accusative* or *plural.*

There is also great variation in the choice of abbreviations used to label properties such as part of speech, gender and agreement features, etc. It is generally easy to reconcile purely notational differences, but the definitions of such notions can also differ in their details. It is thus necessary to establish guidelines for distinguishing notational variation

| | |
|---|---|
| $V27$ | PRED ADJ AGR |
| Predicative adjs agree with the subj in nb and/or gender | |
| $V106$ | ATTR ADJ AGR CASE |
| Attributive adjs agree with their nominal heads in case | |
| $V456$ | VERB FLEX SUBJ |
| Finite verbs agree with their subjects | |
| $V469$ | FLEX ORDER = VERB-TMA-X |
| In a V the morpheme order is Stem-Tense/Mood/Aspect-Agr | |
| $V475$ | DEF ART |
| The definite article is obligatory | |

Example 1: Typological Database Nijemegen

| | |
|---|---|
| Text line: | 1 |
| Orthographic | Njadi nuna jàka na-laku-ka i Umbu Ndilu nàhu la woka |
| Phonemic | ndj\adi nuna dj\Aka na=laku=ka=l Umbu Ndilu nAhu la wOka |
| Morphological | njadi nu-na jàka na=laku=ka i Umbu Ndilu nàhu la woka |
| Gloss | thus DEM-3s if/when 3sNom=go=PFV DEF lord male.name LOC garden |
| Idiomatic | So, that one, when Umbu Ndilu goes to the garden |

Example 2: Spinoza Database

from theoretically important differences, and to normalize the data with respect to the former but not the latter.

Consultations with the community of prospective users have established that the preservation of the specific claims made by the creators of the individual databases is of the utmost importance. Extensive normalization of the collected data into a common form would result in unacceptable distortion. Therefore a major focus of research will be the question of how to best strike a balance between improving usability and preserving the reliability of the collected information.

This paper focuses on the task of representing and recognizing relationships between the contents of different databases, especially in cases where the correspondences are only partial.

## 3. Integrating databases: a pilot study

In order to investigate the problems related to the integration of the various databases, a pilot study has been carried out dealing with merging the portions of the databases that are related to *agreement*. Only two of the databases participating in the project contain variables explicitly concerned with agreement: The *Person Agreement Database (PAD,)* and the *Typological Database Nijmegen (TDN).*

### 3.1. Agreement in the PAD

The PAD contains the following variables on agreement:

1. Two overview variables: ExistenceOfAgreement (Boolean), TypeOfAgreement.
2. Up to three "alignment types" (AlignA, AlignB, AlignC).
3. Four identical blocks of eight variables, defining clause-level agreement with various types of controllers.[1]

The categories *subject* and *object* are not used. Instead, each block of variables describes agreement with one of the following types of controllers:

S: Sole argument of an intransitive verb (Vars. B6–B13)

A: Agent-like argument of a transitive verb (Vars. B14–B21)

P: Patient-like argument of a transitive verb (Vars. B22–B29)

R: Recipient-like argument of a transitive verb (Vars. B30–37)

Each block of variables, describing agreement with one of above controllers, consists of the following eight variables; they are listed with their possible values.

1. Type of agreement: *grammatical, anaphoric,* or *both*
2. Target of agreement: *V, Aux,* etc.
3. Form of agreement marker: *prefix, suffix, proclitic,* etc.
4. Person agreement paradigm: A listing of the paradigm, or *None.*
5. Number agreement paradigm: A listing of the paradigm, or *None.*
6. Gender agreement paradigm: A listing of the paradigm, or *None.*
7. Paradigm for the inclusive/exclusive distinction: A listing of the paradigm, or *None.*
8. The third-person form is null: *No, singular only, yes*

### 3.2. Agreement in the TDN

The TDN includes twelve variables relating to agreement. Of these, seven contain information not included in the PAD; we only describe the other five, which contain information overlapping with the content of the PAD, since they are the ones that present challenges to merging the databases:

1. Two variables on the existence of subject- and object-agreement on finite verbs.
2. Two variables on whether the subject and object markers are identical to the possessive morpheme.
3. One variable on whether the third-person singular subject marker is null.

Note that the information provided by the last-mentioned variable of the TDN can also be found in the PAD (element 8); however, the two variables are structured quite differently and a non-trivial amount of logic is required to relate their values, even in prose description. This is an instance of "diversity in form", which space does not permit us to discuss here.

[1]The element carrying the agreement morpheme is called the *target*, and the element with which the target agrees is called the *controller.* Thus, in subject-verb agreement the subject is the controller and the verb is the target.

### 3.3. The notion *subject*

On the basis of the information provided by the two databases, it is possible to perform various types of combined queries. But even for a simple query such as *which languages have subject-verb agreement?,* providing the appropriate answer is not a trivial task.

The TDN contains a boolean variable answering exactly this question. The PAD includes a block of variables giving more information about subject-verb agreement, if it exists. But there is a complication: the notion *subject* is not a primitive in the PAD. Instead, it recognizes the four different types of controllers *S, A, P, R.* How can this classification be used to get information on subject-verb agreement? We must define a query in terms of the available categories.

Because descriptions such like "agent-like argument" are subject to many possible interpretations, additional information must be obtained from the creators of the database in order to clarify the relationship of these categories to the traditional category *subject;* on the basis of the available descriptions, we hypothesize that the common notion *subject* includes all controllers of type S, all or most controllers of type A, and perhaps some controllers of type P and R. The most useful strategy, then, is probably to search for data on S and A controllers. However, since the correspondence between the category *subject* and the available categories is clearly imperfect, the user must be warned about the situation so that he or she can properly interpret the results of the query.

The system must either have this solution in its information store or it must be able to derive it on the basis of available information, using general inference procedures.

Let us now consider the reverse situation: a linguist who subscribes to the PAD's classification of arguments wants to know which languages have agreement with the patient-like argument. Since one of the component databases does provide this information, the unified interface should make it available. In addition, a user who submits this query should be informed that the TDN includes information about *object agreement*, which includes most instances of agreement with the P (patient-like) argument. The user interface might search the TDN for information on object agreement and present the results, but it would probably be more useful if it presented the user with information on query terms similar to the user's desired query, and allowed the user to manually refine the final query. A user could then decide to rely exclusively on the PAD, or to additionally look up information on object agreement on the TDN, later refining the answer by consulting off-line grammars or by other means.

## 4. The linguistic knowledge base

The real example discussed in the previous section reveals that even simple queries cannot receive an appropriate answer without resolving terminological and conceptual differences. In order to achieve this, the system will rely on a knowledge base of linguistic terminology, glossing standards and database specific terminology, which will allow correspondences to be mapped out at the linguistic level.

The knowledge base must provide information about all the terms that are necessary for the successful use of the system. Note that there are many different sorts of terms: we can begin with *linguistic objects* such as sentence, verb phrase, noun, verb, word, morpheme, segment; *properties* of linguistic objects such as number, case, part of speech, etc.; *values* for such properties, such as singular, accusative, noun; *relations* between linguistic objects, such as precedence, containment, c-command, agreement, etc.; *roles* of the objects participating in a relation, for example in an agreement relation we speak of a target, a controller, and a property (such as *gender*) for which the agreement holds. Relations can be of various arities (i.e, they can relate various numbers of objects), and as the last example showed, they can involve formal properties as well as linguistic objects. There are various containment and mutual exclusion relationships within each of these sorts; for example a morpheme is a part of a word. Different types of relationships hold between objects of different sorts; for example, *accusative* is a value of the property *case.*

The knowledge base should contain information about all the linguistic objects and phenomena addressed in the databases: agreement, word order, anaphora, aspect, stress, inflection, to name just a few. Each linguistic topic should also be linked to a general, human-readable definition. For example, in the case of agreement a possible definition provided by (Steele, 1978): *The term agreement commonly refers to some systematic covariance between a semantic of formal property of one element and a formal property of another. For example, adjectives may take some formal indication of the number and gender of the noun they modify*.

Obviously, creating and managing such complex information is a considerable undertaking. Fortunately, much of it need only be undertaken once, since a taxonomy of linguistic notions and objects is valid independently of any particular database. (Of course, a taxonomy cannot be directly used to describe a particular database unless it includes the terms and notions the database uses). There are initiatives currently underway to create linguistic *ontologies*, described in the following section, which organize and cross-classify the sort of information just described. (Lewis et al., 2001) are developing an ontology of morphosyntactic terms with multiple inheritance and a variety of relations holding among the terms. The TDS can benefit from such initiatives, extending the framework of information they provide with the more specialized information needed for the purposes of the TDS project.

### 4.1. Ontologies

Researchers in artificial intelligence developed ontologies to facilitate knowledge sharing and reuse. More recently, the notion of ontology is becoming widespread in various fields from information retrieval to e-commerce and knowledge management. The interest in ontologies is due to their potential for facilitating communication across people and application systems.

An ontology, in the present sense, is a formal, explicit specification of a shared conceptualization (Fensel et al., 2001), where *conceptualization* refers to an abstract model of some phenomenon, *explicit* means that the type of concepts used and the constraints on their use are explicitly defined, and *formal* means that the ontology should be ma-

chine understandable.

Ontologies are at the basis of the Semantic Web. As described by (Berners-Lee et al., 2001), the most typical kind of ontology for the Web has a taxonomy and a set of inference rules. The taxonomy defines classes of objects and relations among them. Ontology pages on the Web can provide the solution to terminology problems by providing equivalence relations. Furthermore, ontologies can be defined as extensions of (i.e., additions to) other, existing ontologies, and should be able to evolve.

## 5. The integration strategy

Ontologies of linguistic notions can assist us in solving the data integration problems described earlier. To return to the example of a query on subject-verb agreement described in section 3.3., our knowledge base should contain the information that, for example, the category *S* as used in the PAD is a proper subcategory of the category *Subject*, while the category *A* is *mostly* a proper subcategory of *Subject*—meaning that it includes a small proportion of cases to which the label *Subject* is not applicable. A query-generating procedure should then generate a search for subjects by searching for all terms that are completely or mostly subcategories of *Subject;* that is, for *A* and *S*. If some terms are "mostly" subcategories, then the system must also generate a warning to that effect.

As additional component databases are integrated to the TDS, their contents will be mapped to the notions provided by the common ontology. When they rely on notions that are not precisely described in the common ontology, component-specific extension ontologies must be created that define terms with only partial correspondence to the terms in the default ontology. This would be the case with the *Subject* versus *A, S* example if both sets of notions were not already in the the common ontology.

Purely notational variation between databases ("diversity in form") is not as troublesome. For example, variation in the choice of abbreviations used in glossing agreement features and the like can be addressed by simply choosing one abbreviation as the standard, with no injury done to any of the component databases (as long as the variation is only notational, of course). Wherever possible the TDS selects an existing system of annotation guidelines and adopts it as the standard. Adopted standards include the Eurotyp standard for morphological annotation (Bakker et al., 1993), and the Ethnologue (Grimes, 2000) language names and codes as the canonical identifiers for languages in the component databases. Where applicable, the TDS project will also strive for compatibility with the Dublin Core Metadata Element Set, the OLAC extensions, and the ISLE initiative.

The names and abbreviations selected as standard will form part of the *uniform terminology* of the TDS, a distinguished subset of the terms included in the knowledge base. The uniform terminology will be used in descriptions and definitions provided by the unified interface, the names of linguistic terms displayed in predefined query screens, etc. It will also be the preferred vocabulary to use, when possible, for the metadata describing the content of the component databases.

## 6. Conclusions

The aim of the TDS project is to combine various databases and present them to users as a unified virtual database. The challenge of the project lies in the diversity of the included data. We have centered our discussion on our strategy for addressing one of the major sources for diversity: variation in the theoretical commitments of the creators of the various component databases. We have described a solution that makes extensive use of an ontology of linguistic notions.

A further challenge is the integration of different types of content; e.g., integration of the example-based Spinoza database with the variable-based PAD and TDN. In future research, we intend to explore the application to this problem of the approaches discussed in this paper.

## 7. References


Dik Bakker, Östen Dahl, Martin Haspelmath, Maria Koptjevskaja-Tamm, Christian Lehmann, and Anna Siewierska. 1993. EUROTYP guidelines. Technical report, European Science Foundation Programme in Language Typology.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific American*, 284(5):34–43.

Dieter Fensel, Frank van Harmelen, Ian Horrocks, Deborah L. McGuiness, and Peter F. Patel-Schneider. 2001. OIL: An ontology infrastructure for the semantic web. *IEEE Intelligent Systems*, 16(2):38–45.

Barbara F. Grimes, editor. 2000. *Ethnologue*. SIL, Dallas, 14th edition. Online version: *http://www.ethnologue.com/*.

William Lewis, Scott Farrar, and Terry Langendoen. 2001. Building a knowledge base of morphosyntactic terminology. In Steven Bird, Peter Buneman, and Mark Liberman, editors, *Proceedings of the IRCS Workshop on Linguistic Databases*, pages 150–156.

Paola Monachesi, Alexis Dimitriadis, Rob Goedemans, Anne-Marie Mineur, and Manuela Pinto. 2002. A unified system for accessing typological databases. In *Proceedings of the Third International Conference on Language Resources and Evaluation*.

S. Steele. 1978. Word order variation: A typological study. In Joseph H. Greenberg, C. A. Ferguson, and E. A. Moravcsik, editors, *Universals of Human Language, vol. 4: Syntax*, volume 4: Syntax. Stanford University Press, Stanford, CA.