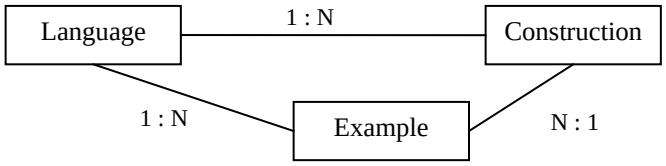
Oral presentation

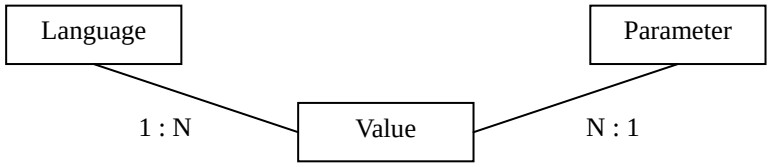
An extensible design for linguistic survey databases

Typology is a data-intensive discipline, and database software has become an important tool for managing, analyzing and sharing one's collection of research data. But databases for linguistic research have a number of properties that distinguish them from the business-oriented databases that inform much of the theory and practice in the field of databases. Most crucially, the properties, categories and values that should be recorded in a research-oriented database are typically not known at the start of data collection. This frequently leads to problems, especially for multi-person projects, since revising a database that is already in use is a much more complicated proposition than including the same features in the original design. The result is that databases are underused: many researchers are reluctant to rely on them, or limit their use to presenting their data on the web after the research activities of the project have been concluded.

This talk will present a database design that, among other features tailored to linguistic databases, supports the easy modification and addition of linguistic attributes and category values. Our database is particularly tailored to the type of cross-linguistic survey that investigates instances of some construction or phenomenon in diverse languages. It includes a core triad of "entities" *Language, Construction and (Example) Text;* the research group can define linguistic properties of interest for each of them, which are then completed for each language, construction or sentence in the course of data collection. The database design also includes facilities for managing lists of "enumerated" possible values (they are all stored in a single table). The design is sufficiently flexible that the software is now being used, with only superficial changes, for five other survey projects by unrelated research groups.



A research-oriented database will need to evolve over the course of a project, as already mentioned. Some typological databases are added to over a long period of time, and can come to define hundreds of properties of a language (or other unit of description). Creating and revising forms and queries for so many parameters is complicated and error-prone. Our solution is to store the names, types and documentation of linguistic parameters in their own table; the value they take for each language (or construction, morpheme, etc.) is given in a separate table, which relates languages with parameters. In different terms, Parameters ("questions") are entities in a many-to-many relationship with Languages; at the intersection of the two is the Value of a parameter for some language (the "answer").



The approach allows new parameters to be added at any time without modifying the relational design of the database, and allows the new parameters to be displayed in existing forms. It also has the benefit of easily allowing a property to take multiple values per language (but only if this is desirable), and provides a natural place to store the documentation of each parameter, in the Parameter table itself.

Our database software consists of a web interface written in php, and a MySQL database back end. While it is a first implementation and incomplete in various ways, it has proved useful to diverse projects and is being further developed. It is also freely available to interested linguists who might want to use it as the basis for their own database. The underlying design is suitable for an even wider variety of research needs.