

Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



Managing differences: The TDS approach

Alexis Dimitriadis
Utrecht institute of Linguistics OTS

On behalf of the TDS project: Alexis Dimitriadis, Menzo Windhouwer, Adam Saulwick,
Rob Goedemans, Tamas Biro.

Integration of linguistic data (or indeed any kind of data) is commonly based on identifying correspondences in content among the component resources and transforming, to the extent possible, variant data into a target schema. This assumes that a target schema can be designed that can capture all, or nearly all, of the information contained in the schemas of the component resources. What cannot be expressed in the new schema is discarded, or perhaps retained with secondary status.

The Typological Database System project (<http://language.link.let.uu.nl/tds/>) provides a unified query interface to multiple, independently developed typological databases. Our approach is based on the conclusion that no single schema could capture the content of the component schemas with negligible information loss: Instead, the TDS system is designed around the task of organizing a collection of semantically heterogeneous, and incommensurable, resources.

On the one hand, there are non-semantic sources of difference that usually CAN be resolved. The approach of the TDS is to compensate, wherever possible, for purely notational variation such as different design decisions, software platforms, document or font encodings, abbreviations for values, etc. Theoretically contentful differences, however, are preserved and presented to the end-user. Each data field and value is accompanied by descriptive documentation, which allows users to understand what they are looking at, and potentially to rely on common sense or linguistic judgement in evaluating relationships between resource schemas.

Semantically related fields are linked with the aid of an integrated ontology, developed as part of the project. This ontology too is based on the principle of incommensurability: It is what we call an "inclusive" ontology, allowing multiple formalizations of the same phenomenon.



The approach is appropriate for any type of resource that involves linguistic analysis: language descriptions, theoretical claims, or texts with any sort of annotation; it is particularly suited to integrating a combination of such resource types. (The TDS currently includes several databases consisting entirely of ``analytical" parameters characterizing language as a whole, e.g., ``basic word order=SOV" or ``copular adjectives are preverbal"; other databases contain a mix of analytical parameters, parameters describing several instantiations of some construction in a language such as intensifier constructions, and example sentences and their descriptive properties).

To the extent that annotations depend on one's linguistic analysis of the construction being annotated, no common method of annotating a phenomenon could be satisfactory for everyone: Different theories organize the world in different ways. Even if a ``consensus" analysis can be devised for some domain, the open-ended nature of linguistic research would probably make it a short-lived victory. Any system of annotation is liable to become inadequate as field workers are confronted with unexpected phenomena in new languages, or as researchers develop new analyses.

One COULD agree to use some common terms with a common meaning; but even with these, linguists invariably differ in the details. The best way to achieve data integration without information loss is to allow each linguist to explain what they mean when they use a term or make an assertion. And this means accepting that there will be a multitude of incompatible analyses, and undertaking to manage them instead of trying to map them to a common standard.

In short:

- * Differences between linguistic resources are part of their core content: They cannot be removed without significant loss.
- * The question of what can be discarded during data transformation depends on the purposes of the end-user: It cannot be decided at the point of data integration into an archive.
- * Careful documentation of data and encoding is essential for any kind of information sharing.
- * If such documentation is in machine-readable format, it can support some kinds of automated transformation.
- * But the documentation must also address subtleties that only a human reader can address; therefore, fully machine-useable documentation is not possible.

