

Project Summary

Linguists are working toward understanding what all human languages have in common and, simultaneously, toward understanding the ways in which human languages differ from one another, and what the limits on those differences are (see Chomsky 1981, Greenberg 1966). The goal of this project is to build a web database of the syntactic structures of the world's languages (SSWL) to enable linguists and the linguistically curious to explore the connections among the grammatical systems of the world's languages.

SSWL has three features which when combined make it powerful and unique: First, SSWL is based on a property-as-value design, which allows an unlimited number of syntactic properties to be added. Second, the query interface exploits the full potential of the underlying relational database, allowing millions of searches in linguist-friendly format (no knowledge of SQL or databases required). Third, SSWL is language expert oriented, in the sense that all the data in the database on a particular language will come from experts (preferably native speaker linguists) of that language.

This database will be based on a working prototype, developed during Fall 2008 under NSF SGER 0817202 (Co-PIs: Prof. Chris Collins and Prof. Richard Kayne in collaboration with Prof. Dennis Shasha). Currently the prototype allows a wide range of searches over languages, properties, property-value pairs, and examples (actual sentences from languages represented in interlinear glossed form). In the proposed work, a series of much more sophisticated and powerful search functions will be added, including a function which finds the implicational universals for a given set of properties and languages, a function yielding the set of languages similar to a given language (at a given level of similarity, e.g., 90%), the ability to extend the SSWL search results with ODIN IGTs (Interlinear Glossed Texts) from the internet, the ability to incorporate genetic classification data into searches (e.g., "Find all German languages such that..."), visualization of search results using Suneat, and the ability to chain searches, so the output of one search can be used as the input of another. Lastly, the number of grammatical properties in the database will be expanded considerably, and a reference table with an interface will be added. Usability testing will be done on each new feature.

Intellectual Merit: the open-endedness and sophisticated queries of SSWL have the potential of completely changing the way people do comparative syntax and linguistic typology. SSWL furthermore has the potential of accelerating our rate of acquisition of knowledge about language considerably.

Broader Impacts: Since the project is not oriented towards any specific linguistic framework, it will be maximally inclusive, allowing people of all syntactic frameworks to participate. If successful, such a model of an open-ended database of linguistic knowledge could spread to other areas of linguistics, including semantics, phonology, phonetics and morphology, transforming the field of linguistics. A central concern of the database will be the issue of interoperability. If our project is successful it could contribute toward the goal of making linguistic databases interoperable with other (linguistic and non-linguistic) databases, and hence make the results of our project widely available outside of the linguistics community. Lastly, the grant will be instrumental in training two graduate students, one from Computer Science and the other from Linguistics, and undergraduates doing independent study projects.