Project Description

Objective

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 allows millions of searches in a 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) on that language. Each of these three features will be described individually below.

Motivation

Linguists are working towards understanding what all human languages have in common and, simultaneously, towards understanding the ways in which human languages differ from one another and what the limits on those differences are (see Chomsky 1981, Greenberg 1966). In doing their work, linguists take into account data about the properties of many individual languages. The number of languages under study has been increasing substantially (see Baker 1996, Julien 2002, Kayne 1994, Cinque 1999, Greenberg 1966, Dryer 1992, Haspelmath et. al. 2005). This increase has made it more and more difficult to integrate the data, the descriptions, and the theoretical implications that this ever larger number of languages brings into the field.

The system will be open to anyone for reading but would be curated by authorized linguists worldwide who enter and edit syntactic information and examples from languages they study. The number of languages in the proposed database will increase regularly, as more and more languages from around the world are added. Some of these languages would be relatively well-known ones that have not previously received much attention in the linguistics literature. Others would be lesser-known and endangered languages that linguists from a new generation would have found the means to study in detail. Still others would be what are often called dialects, but which deserve to be studied as separate languages, often with interesting and important syntactic differences relative to their better-known cousins. By having the database open to new dialect distinctions, as well to the entry of previously little-studied languages from all over the world, the number of languages/dialects that the database will contain will be greater than the number 6000-7000 (see Ethnologue 2005) often cited as the number of languages currently spoken.

The database we have in mind will also aim to take into account a far greater number of syntactic properties than has ever been done before. In part, this will simply reflect the knowledge already accumulated, especially over the past 50 years. Technically, this will be made possible by the open-ended character of the database (the property-as-value design described below). Although we plan to "seed" the database with an initial set of properties, we very explicitly intend to allow for the addition to the database of new properties discovered in the future (or currently known to some, but overlooked in the original set). As the fields of comparative syntax and linguistic

typology continue to expand, other properties will be thought of that are of interest and importance. Our database is designed to allow properties to be added without limit.

Design of the Database

On November 9 and 10, 2007, we held a workshop at NYU to investigate the feasibility of such a database. This workshop resulted in a web page summarizing the talks at the workshop (http://www.nyu.edu/gsas/dept/lingu/events/SSWL07/). The consensus of the workshop was that such a project is important and feasible. Given the results of the workshop we implemented a prototype, based on a design of Prof. Dennis Shasha. The prototype is up and running (http://sswl.railsplayground.net/). The basic design of the database is below:

languages(languagename, propertyname, value, contributorname, date, time)
examples(languagename, sentenceid, type, propertyname, value, contributorname, date, time)
properties(propertyname, description, contributorname, date, time)
contributors(contributorname, username, password, affiliation, e-mail, date, time)

The *examples* table contains example sentences and phrases for each language, where each example consists of a line of text (with morpheme boundaries indicated), a gloss and a translation. Each example typically illustrates one or more property-value pairs. An example from the prototype is given below:

Language: Bellinzonese
Example: Al Mario l=è grand asée
Gloss: the Mario he=be.PRS.3SG tall.M.SG enough
Translation: Mario is tall enough.
Contributor: Andrea Cattaneo

The *properties* table gives the definitions of the properties used in the *languages* table. Some examples are given below (underlined text indicates a hyperlink to a definition):

The property Attributive Adjective Agreement has the value "Yes" when there is at least one attributive adjective that shows agreement with (at least some of) the nouns it modifies.

The property Verb Object has the value "yes" when a verb can precede its object in a neutral context. The clause in this property is an active (non-passive) declarative (non-interrogative) clause. The object in this property is a noun phrase (we exclude pronouns). As with all word order properties, we restrict our attention to productive word order patterns.

The current set of properties in the database all have the values “Yes”, “No” and NA (not applicable). For example, English is set as Attributive Adjective Agreement:No. As it turns out, “Yes”, “No” NA values for the word order properties allow for a more fine grained classification of linguistic phenomena and hence make it possible to track

more closely linguistic variation. For example, instead of having a single property "Order of Subject, Verb and Object" (with values SVO, SOV, etc.), we have six properties: "Subject Verb Object" (values Yes, No, or NA), "Subject Object Verb" (values Yes, No, or NA), etc. This allows for an accurate characterization of languages with free or less constrained word order. On the other hand, there is no inherent (or implementation) reason why complex properties (with values other than "Yes", "No" and NA) cannot be allowed. If a property author makes a case for such a value, our model and system can incorporate them seamlessly.

The *contributors* table contains information about who contributes data to the database (where "data" means property definitions, property-value pairs, or examples).

The *languages* table gives the values for each grammatical property. For example, in the prototype there is a property "Attributive Adjective Agreement". The value of this property for French is "Yes" and for English "No". A complete listing of properties with their values and accompanying examples for a language is equivalent to a rough grammatical sketch.

Property-as-Value Design Philosophy

Rather than defining each property as its own column in the languages (or examples) table, a new property for a language can be defined by inserting a new row in that table, e.g. French, Adjective Agreement, Yes, ….. This property-as-value approach is used in e-commerce systems, where any new product may introduce properties held by no other products (e.g., introducing watches to a product line may entail adding the property of wrist size). Prof. Shasha has successfully used this approach in the VirtualPlant system he has designed for plant biologists. Thanks to the property-as-value design, the number of properties in the system may increase continually without requirement for reprogramming.

Because of the property-as-value design, the number of properties in the system can increase without bound. To ensure uniformity and to maintain the quality of the properties, we have tried to fix a number of general characteristics that all properties in the system should have. Each property definition will be illustrated with a number of examples. First, there should be an example from English. Second, there should be examples (from various languages) that exemplify the different values of the property in the definition.

The properties should be defined so as to increase inter-expert agreement. The definitions must be written in such a way that different people, from different backgrounds will set them in the same way for the same language. One way to achieve this goal is to leave out jargon in the definitions. Also, the linguistic terms used in the property definitions should be defined, and the definitions of the terms used should be clear and easy to apply. All general definitions will be placed in the glossary (already implemented), and new property authors will be encouraged to use them. We are looking for ways to test for a given property whether it will give rise to a high degree of inter-expert agreement.

The properties should be formulated so that for any language on earth it will be possible to set them. In other words, it should never be the case that somebody reading the property for some language will simply have no idea how to set it.

Powerful Search Interface

The search interface is designed to exploit the full power of the underlying relational database. At the same time, it has been designed with the working linguist in mind, so that no knowledge of SQL or database terminology or concepts are necessary. A formal specification of the algorithm behind the search interface is provided on the SSWL web site (on the About page). Below, we review the main components already functional in our prototype at http://sswl.railsplayground.net/.

On the Show line, the user may specify some combination of Language, Property, Property-value or Example. For example, clicking on Language and Example in the Show line will produce a listing of all the languages in the database and the examples that have been entered for each of those languages.

Users may constrain the set of results from the Show line using constraints defined in terms of languages, properties, property-values, and/or examples. All these constraint options are found in the boxes under the Show line (see diagram on page 6). For instance, one might ask for all property-values of Dagaare, by clicking on Property_Value in the Show line (meaning: the results will be a list of property-value pairs), and then selecting Dagaare from the list of languages in the language constraint box (to the right of the words "Select Language"). It is possible to specify several different kinds of constraints (from different boxes) in one search.

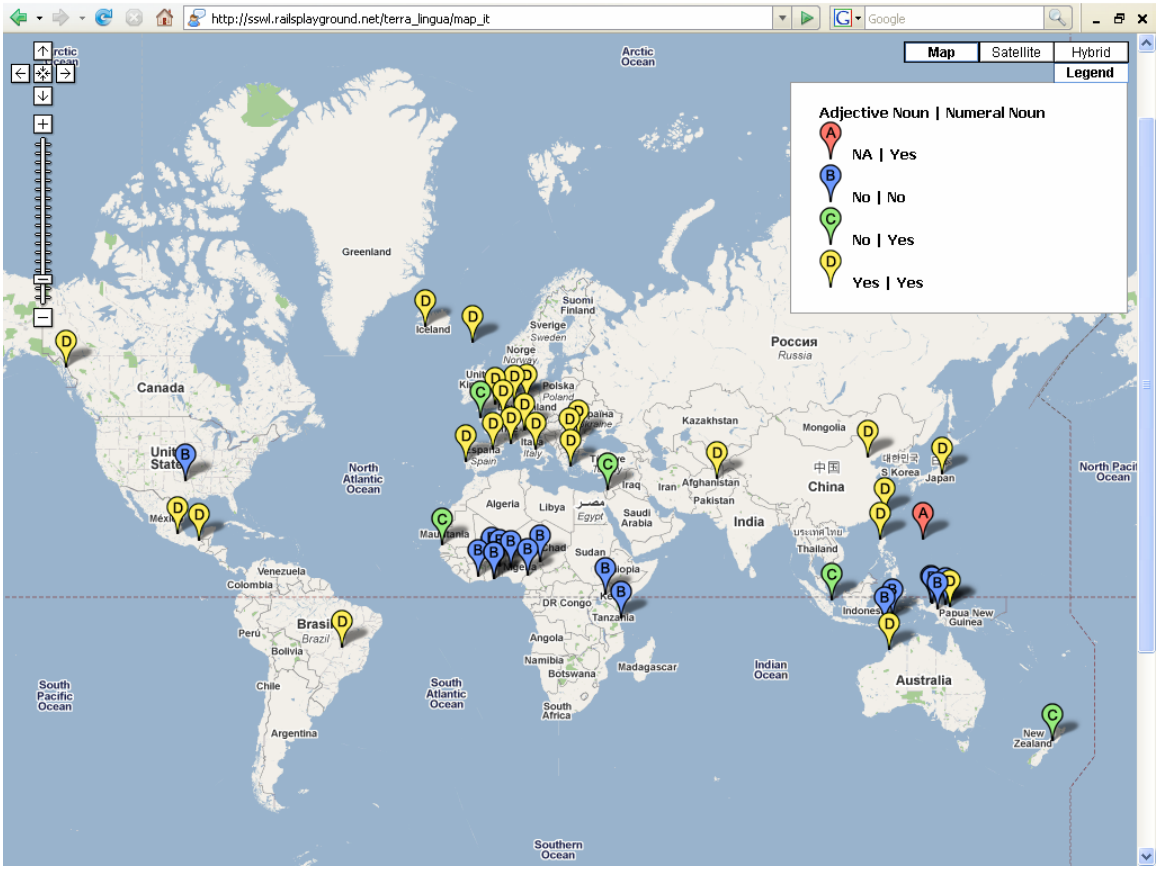
Two particularly useful functions for linguists are the Any and All functions found just under the property constraint box and just under the property-value constraint box. In logical terms, "Any" means disjunction (logical OR) within a constraint. By contrast, "All" means conjunction (logical AND) within a constraint. For example, suppose a user clicks on Language in the Show line and selects the following property-value pairs (in the constraint box): Attributive Adjective Agreement:Yes and Auxiliary Selection:Yes. If All is specified, this search will find the set of languages that have agreement with attributive adjectives AND for which the property of auxiliary selection holds. If one had clicked Any instead, then the search would yield the set of languages that have agreement with attributive adjectives OR for which the property of auxiliary selection holds (or both).

The "Cross" function allows a comparison of a pair of properties on all or a subset of languages. The essential function of Cross is to form tables that are similar to the tetrachoric tables of Greenberg 1963. For example, a cross among Adjective Noun and Numeral Noun, yields the counts and the languages for each combination of Adjective Noun:Yes/No/NA and Numeral Noun:Yes/No/NA. It is also possible to constrain Cross to a particular set of languages (using the language constraint box). The output of Cross for Adjective Noun and Numeral Noun is given below (clicking on the numbers yields a list of the actual languages):

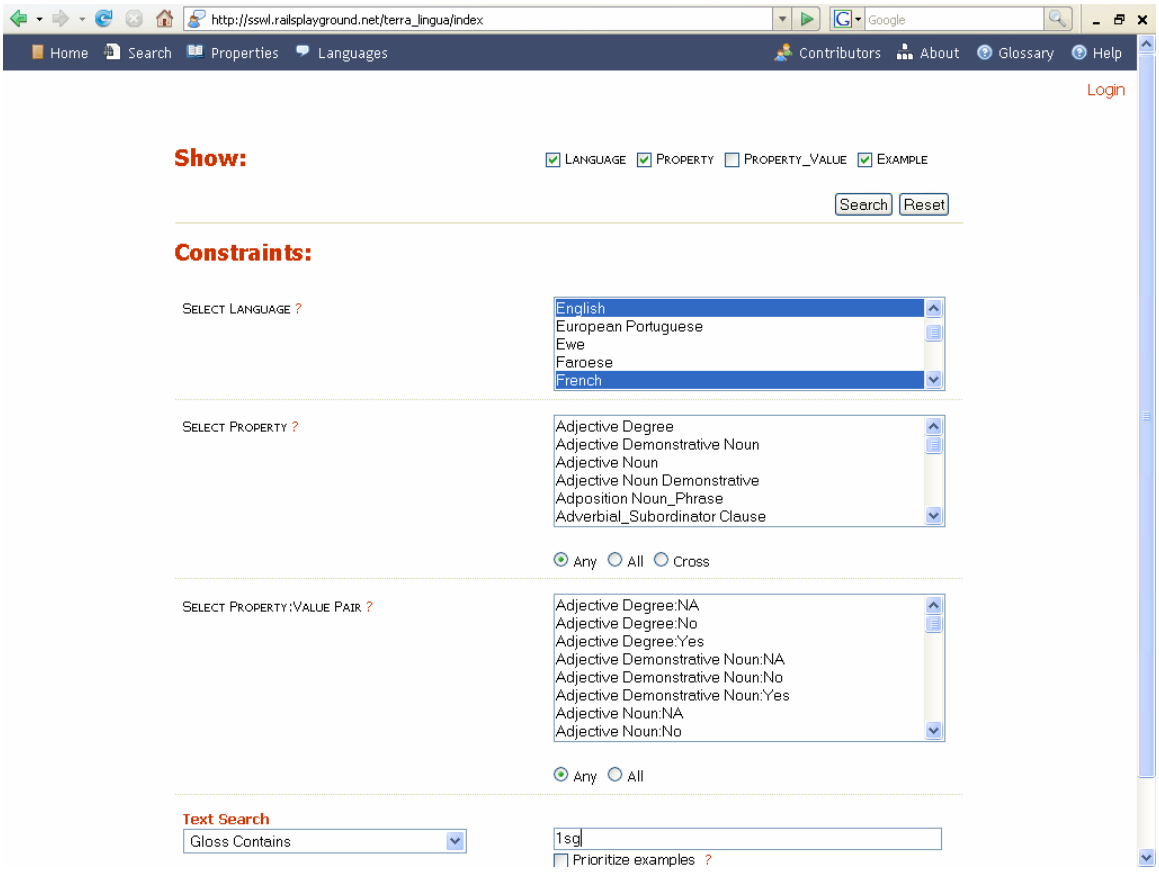
Property 1	Val 1	Property 2	Val 2	# of Langs
Adjective Noun	NA	Numeral Noun	Yes	1
Adjective Noun	No	Numeral Noun	No	17
Adjective Noun	No	Numeral Noun	Yes	5
Adjective Noun	Yes	Numeral Noun	Yes	25
Adjective Noun	NA	Numeral Noun	No	0
Adjective Noun	Yes	Numeral Noun	No	0

Of particular interest here is that there are no languages with the order Adjective Noun but without the order Numeral Noun. This shows that there is some kind of incompatibility between these two properties and it is up to the linguist (whichever framework that they work in) to try to give an explanation.

These results can be mapped (as can all search results):



In the database, an example is represented as a sentence in a target language (with morpheme boundaries indicated), a gloss for that sentence, and a translation into English as well as the grammatical property-value pairs that the example illustrates. It is possible to constrain example searches based on the words and morphemes in the sentence, gloss or translation. For example, if the Show line has Language, Property and Example clicked where language is constrained to French and English and Gloss Contains is set to "1sg", then the search will return all properties having to do with French and English as well as examples of those properties whose gloss contains “1sg”. If the "prioritize example" box is clicked, then only information (in the above example, languages and properties) having a corresponding example will be shown.



Thus, any subset of language, property, property-value, and example can be selected on the Show line and any subset of language, property, property-value or example can be selected in the constraint boxes (with various options for each). A quick calculation shows that the number of basic query types is 1536. Because one can select arbitrary sets of languages, properties, property-value pairs, and text to search, the number of actual queries is already in the millions in the prototype.

In addition to the powerful search interface, we provide a set of browsing pages for both languages and properties. For new users, using the search interface might be intimidating. Browsing, on the other hand is easy, the user simply clicks on a language or property name in order to go deeper into the system. The language browsing page is illustrated below (there is a separate page for properties):

Language Expert Orientation

All the data in the database comes directly from language experts. Each language of the database will have at least one language expert associated with it. To become a language expert, a user must register, and fill out a form that includes a web site which describes their expertise. Language experts are acknowledged on the masthead of the web site (under a link called “Contributors” in the navigation bar on the top of the page). Furthermore, in the near future, we will configure the system so that the contributor name will also appear accompanying every piece of data in the system. For example, the property-value pair **Verb Object:Yes** for Ewe will be listed as having been contributed by Chris Collins.

A language expert will preferably be a native speaker linguist of the language in question. If there are no native speaker linguists available, the language expert can be a linguist with a deep knowledge of the language in question. There is no constraint on the type of theoretical framework adopted by an expert. This project is emphatically meant to be useful to all linguists (formal linguists from different frameworks, linguistic typologists, field workers, etc.).

The language expert will be in charge of setting the property values for their language. Once they have set these values, the project coordinator (Chris Collins) and language coordinator (Ken Hiraiwa) will look them over to see if there are any obvious errors or inconsistencies. It will also be possible for other language experts to evaluate these property-value settings and provide feedback to the contributor through:

commenting (not yet implemented, see Proposed Work), the SSWL Google Group (already set up), and direct e-mail.

The language expert will also be in charge of adding examples to illustrate the property-value pairs. These examples should conform to the Leipzig glossing rules (see http://www.eva.mpg.de/lingua/resources/glossing-rules.php). For each language in the database, each property-value pair should have at least one illustrating example. All examples in the database will be checked by the example editor, whose job it is to ensure that example follow the Leipzig conventions and to check for consistency between the entered examples and the property-value pairs for a language. We are in the process of searching for a linguist to fill the role of example editor.

Concerning consistency, the database allows only one value for each property for each language. So there is no way that for French the property Attributive Adjective Agreement could have two contradictory values. The property value can be changed by the project coordinator or the language coordinator. Similarly, language experts will soon be able to edit their own data. But in no case can one language expert change a value set by another. Conflicts will be resolved in multiple steps. First, it will be necessary to find out whether the difference in opinion is related to some lack of information or ambiguity in the property definition. If so, that can be resolved by adding more information or further examples to the property definition (in a way that does not alter its content). Second, it is important to find out whether the conflict is due to a dialectal difference. If so, then the dialect difference can be registered in the commenting system, or if the dialect is different enough, another language entry in the database can be started.

The language experts may wish to work in a team. For example, the language expert for Russian may choose to ask other linguists if they want to help out with Russian. The anticipated amount of time that a language expert must commit to the system will be on the average two hours (or more) a month, checking in at least once a month to see if there are new properties. Language experts may rotate from time to time.

Just as all the data in the database will come from language experts, the properties will be written by property experts. For example, a set of properties on anaphora will be written by somebody who has an extensive publication record working on anaphora. Property authors will be acknowledged on the masthead of the web site (under a link called “Contributors” in the navigation bar of website). Their name will also appear on the property definitions that they write.

Once a set of definitions is written, the property author will send them to the administrators (linguisticexplorer@gmail.com) who will send them out to two or more language experts for review. The language experts will be asked to make sure that it is clear how the proposed properties would be set for their particular languages. This is not a formal review process since no attempt will be made to keep the identities of the language experts secret. Furthermore, the language experts will not have the job of rejecting or accepting the definitions. At the same time as the property definitions are being evaluated by the language experts, they will also be posted to the Google Group sswl.linguistics, and a discussion thread will be opened for them. Once posted, the property definitions will be open to the public for feedback. All property definitions will remain open for feedback for two months. The property authors will be encouraged to send drafts of their properties to colleagues to help sharpen definitions and provide examples. The administrators will review the property definitions for consistency and

redundancy with pre-existing property definitions. The administrators will also check that the name of the property follows the database naming conventions. The property authors will then incorporate the feedback from all the above sources (colleagues, language experts, the Google Group sswl.linguistics, and administrators) and submit the final property definitions. They will then be uploaded to the database.

Pilot Launch

In order to test the feasability of the system and whether there will be interest in the project by the wider linguistic community we did a pilot launch. The NYU internal launch (involving NYU grads) was done during April 2009, and the public launch was done during June 2009. During the NYU internal launch, Norman King, an NYU undergraduate doing an independent study project, supervised various grads in the NYU Department of Linguistics entering data about their native languages. These sessions were filmed with Silverback usability testing software and formed the basis of a report written by Normal King on the usability of SSWL (the report is available on the SSWL Google Group). The public launch was inaugurated with an announcement on the LinguistList. Here are the statistics (noted on July 2, 2009) after these two launches (from the About page on the SSWL site):

Number of Contributors:	61
Number of Languages:	70
Number of Languages at 90% (or over):	40
Number of Properties:	36
Number of Examples:	484
Number of Property-Value Pairs:	1895

Another way to guage interest in the site is to look at the Google Analytics statistics for the month of June (June 2 to July 1, 2009). Here are some figures:

Total number of site visits:	2,651
Total number of pages viewed:	15,351
Number of countries represented:	79

Based on these statistics, we can make the following inferences: There is a significant interest in SSWL, and people are motivated to contribute. Assuming a (conservative) rate of 20 new languages a month, for 12 months a year, we should be able to reach 520 languages after two years (520 languages = 40 initial languages + 12 months x 2 years x 20 languages).

The pilot launch has also taught us some important lessons about design, content and workload management. First, the pilot launch has taught us that it is important that all data entry fields be validated (meaning that they should come with restrictions on their form and content). Second, the pilot launch has made us realize the importance of implementing contributor editing (where contributors can edit their own mistakes). People often make mistakes, and want to change them right away. Third, we now realize that we need mechanisms for language experts to communicate with one another (e.g., a commenting system). It is often the case that an expert will write us and point out

problems with the data (in every case so far, the contributor has been happy to make the change).

Review of Existing Linguistic Databases

This short review is limited to web-accessible cross-linguistic databases of morpho-syntactic phenomena. The database that is most closely related to ours in scope is WALS (World Atlas of Linguistic Structures). SSWL and WALS differ in several important ways however. SSWL is a language expert oriented system, and much of its character falls out directly from that basic design feature. Concretely, this means that all the data on a particular language come from a native speaker linguist (or a linguist with a deep knowledge of the language). WALS, on the other hand, is a property author oriented system. The data in WALS was entered by the authors of the properties. So the two systems are complementary (each with their own strengths and weaknesses), not in competition, and there is plenty of room for collaboration. For example, property definitions tested out on SSWL could be candidates WALS property definitions.

Some other databases on the internet hold archives of responses to a common questionnaire. These include the Syntactic Atlas of Northern Italy and the Variation in Control Structures. Hence it is not possible to do sophisticated searches over this data.

Yet other databases support database queries and have example sentences, but are currently limited to a single area of morphology or syntax (e.g., The African Anaphora Database, The Anaphora Typology Database, Graz Database on Reduplication, Berlin-Utrecht Reciprocals Survey), and usually contain data on a very small number of languages. In contrast, our database will cover all areas of syntax, and will contain a large number of languages (an estimated 520 languages after two years of grant).

Lastly, TDS (Typological Database System) is not itself a data collection, but only contains data from other databases; so it's not the same kind of thing as the SSWL at all.

We would like to establish collaborative relationships with all existing linguistic databases. For example, we have been in communication with members of the WALS team about formulating the word order properties (Matthew Dryer) and finding property authors for SSWL (Martin Haspelmath). We have also written letters and received responses from the following people who have expressed interest in collaboration: Dik Bakker (Agreement Database), Dunstan Brown (Surrey Morphology Group), Balthasar Bickel (AUTOTYPE). Alexis Dimitriadis (TDS) has offered to include SSWL in TDS, once SSWL stabilizes a bit. He has also provided technical advice of various kinds. Gary Simons (Ethnologue) did a full review of our initial prototype, and helped us obtain the Ethnologue genetic classification data. Maria Polinsky (Variation in Control) has offered to write a set of properties for SSWL. A team working on ODIN (see description under Proposed Work) is interested in figuring out ways for SSWL and ODIN to share examples and other kinds of data.

Proposed Work

Although a prototype of the system is up and running, the three main features of our database (property-as-value design, powerful search interface, and language expert orientation), make it possible to extend it in many ways. In the remainder of the project

description, we will outline the work we would like to undertake on the database with NSF funding for the next two years.

Developing New Properties: A priority of the next phase of development of SSW is to obtain more properties. The immediate target areas for new properties are case/alignment (nominative/accusative versus ergative/absolutive) and anaphora (e.g., types of reflexive pronouns). We already have several commitments for new properties from Eric Reuland (anaphora), Maria Polinsky (control and raising), and Ljuba Veselinova (negation of copula and existential verbs). We are eager to get typologists, formalists and field workers involved in writing these property definitions (and even to have collaborations among these groups). Our goal for the two years is to obtain 200-400 properties (in approximately 10-20 property areas like word order or anaphora). Our main approach to finding property authors will be to contact known experts individually.

Co-PI Collins and co-PI Kayne with the help of two graduate students (Andrea Cattaneo and Jim Wood) wrote the initial set of word order properties for the database. They will continue to write property definitions where needed, but the main responsability for generating new property definitions should fall into the hands of the wider linguistic community.

Examples: ODIN (Online Database of Interlinear Text) is an online database of IGT (Interlinear Glossed Texts). In collaboration with teams at Simon Fraser University (Professors Chung-Hye Han and Anoop Sarkar) and at the University of Washington at Seattle (Professor Fei Xia), we plan to integrate access to ODIN into the output of SSWL searches. For example, if a user does a search for all the examples illustrating Object Verb word order for a particular language, he or she will be given the option of "Get More Examples from ODIN". If the user clicks on this, more examples from ODIN will be found and listed. We are jointly investigating more sophisticated ways of sharing information between the two systems. The ODIN team has supplied a letter of collaboration.

Interactivity: We will add editing forms so that a language expert can edit any data that they have entered (including examples and property-value pairs). Language experts will be able to comment on the data that other language experts have entered. We will implement a comment architecture to support threaded discussion, for all three types of data: property definitons, property-value pairs, and examples. Lastly, we will also add a references table in order to give potential bibliographic support to any asserted fact in the database. A reference id column in the languages and examples table will be linked to a references table and will be accessible via a hyperlink.

Incorporate SIL Language Classification Data: We have obtained official permission from SIL Ethnologue to use their language classification data in our search interface. This will allow such searches as: "Find all Germanic languages with Attributive Adjective Agreement."

Displaying Results with Sungear: Sungear is a system originally developed for bioinformatics to compare the results of different experiments. Like a Venn diagram, Sungear visually shows the size of the intersections of various sets. Unlike Venn diagrams, Sungear extends to more than three sets very naturally. For linguistics applications, the sets would be the sets of languages corresponding to different property-value pairs (e.g., all the languages with Attributive Adjective Agreement:Yes). Intersections of those sets would be languages that agree on two or more property-value

pairs. Visualization of these linguistic interactions could in principle open up a whole new qualitative way of analyzing linguistic data. A preliminary demo of the concept for SSWL can be found here: http://cs.nyu.edu/~crispy/sswlsungear/

Chaining Searches: The new query interface will allow the results of one query to be reused as the input to another query. For example, suppose I run the query "Find all languages with attributive adjective agreement." This will yield a list of languages, which we can label Q1 (the list of previous queries will be displayed in a small window in the query interface). Now it will be possible to take Q1, and run the following query "Find all languages of Q1 which are spoken in Africa." This will yield another list of languages, which we can label Q2, and so on.

Language Comparisons: It will be possible to find the set of languages that are at least x% (e.g. 90%) identical to a given language. As with all searches, it will be possible to relativize this search to particular property sets and particular sets of languages. So queries such as the following will be possible: Find all the African languages that are 90% similar to Ewe with respect to word order properties.

Complex Searches over Properties : A core feature of our database, which will distinguish it from every linguistic database that we have found, will be the ability to search over properties that bear certain relations to other properties. For example, it will be possible to search for all the implicational universals (Greenberg 1966) in the database: Find all the properties P1 and P2, such that whenever P1 has the value "yes" in some language, P2 also has the value "yes". It will be possible to limit this search to particular sets of languages (e.g., Romance languages), and particular sets of properties. A variant on this theme will be to find all the zeros: all the P1:V1 (property-value pair) and P2:V2 such that there are no languages that instantiate P1:V1 and P2:V2 (e.g., no languages that have both Noun Numeral:Yes and Noun Adjective:No). These functions, Implicational Universals and Zeros, will be present right on the search interface.

Scalability and Hierarchies: When we have thousands of properties and thousands of languages, the current data entry interface will require enhancement. In order to face this problem, languages will be searchable using a keyword search typed into a keyword window (to the left of the constraint boxes). For example, suppose the user wants to find information about languages of Togo. It will be possible to type "Togo" into the keyword window, and all the languages of Togo will move to the top of the choice list. Then the user will be able to choose any one of those languages to work with. Similarly, in order to search for a particular property or property-value pair, it will be possible to enter a keyword into the keyword window, and all properties incorporating that keyword will be listed. We prefer this to creating a global hierarchy for the properties in the database, because there is no agreed-upon property hierarchy.

Achieving Interoperability: The database will be designed to be maximally interoperable with other databases and projects that exist on the internet. For all the data in the database, we will adhere to standards in the field. One example is to use the ISO 639-3 codes for languages. Furthermore, we will attempt to follow the recommendations outlined in Bird and Simons (2003) to the greatest extent possible. We will also pursue interoperability with OLAC (Open Language Archives Community), interoperability with GOLD (General Ontology for Linguistic Description) (Farrar and Langendoen 2003) and interoperability other linguistic databases on the internet, especially for examples in IGT format.

Testing: We will do usability testing with Silverback software through the duration of the grant. We have already begun this kind of testing with a report written by Norman King as an undergraduate independent study project. We will also start to do testing on inter-expert agreement in setting property values. When a set of properties has been written by a property author, we propose to launch a test trial, where several different language experts for a single language set the property values. We will then look at the inter-expert agreement. If certain properties have low levels of inter-expert agreement (under 95%), we will reformulate the property definitions to try to bring up the levels.

Mode of Work and Responsibility of Participants

The work in the project will be centered around a series of weekly meetings involving all the participants (Co-PIs, grads, undergrads, people from other schools). During these meetings, the relevant participants will give progress reports of their activities. This way there will be maximal collaboration in the project.

Co-PI Collins is the project coordinator. He will be in charge of overall coordination of the computational and linguistic efforts. For the duration of the grant, PI Collins will serve as the highest level administrator of the database.

Co-PI Shasha will be in charge of all architectural issues involving the structure of the database, queries, visualization, and data input. He will also be in charge of issues involving the statistical analysis of the data in the database.

Co-PI Richard Kayne will be in charge of higher level linguistic considerations (identifying useful features for the database, and areas where the database could prove useful for research).

Prof. Ken Hiraiwa (Department of English Literature, Meiji Gakuin University) is the language coordinator. He is responsible for contacting language experts and inviting them to participate. He is also responsible for checking over property-value pairs for obvious errors.

The programming team will pursue development using Ruby on Rails and MySql. Prof. Shasha will provide technical direction to the team.

The linguistics graduate student will be in charge of the interface (using HTML/XML/CSS), adding content to the glossary, the help files, and the About page. He or she will also help to develop property definitions with Professors Collins and Kayne. The linguistics grad will be expected to have a particularly close collaboration with the programming team.

In addition to this, there will be various other linguists, graduate students and undergraduate students who will be participating in the project as members of the research team.

Timeline (Past and Present)

Nov. 9, 2007:	Workshop on the Feasibility of a Web-Based Database of the Syntactic Structures of the World's Languages, held at NYU. (http://www.nyu.edu/gsas/dept/lingu/events/SSWL07/)
June 2, 2009:	Public launch of SSWL on the LinguistList.
May 1, 2010:	Proposed starting date for NSF grant.
April 30, 2011:	End of Year 1

Preliminary implementation of all proposed additions.

April 30, 2012: End of Year 2

The database will contain at least 520 languages and 200-400 properties.

Preferably all the languages will be valued for all properties.

Broader Impacts

The database will become a tool that is a frequent presence in any classroom where syntax, semantics or morphology is taught (at any level, graduate or undergraduate). If a professor or a student has a question about a certain linguistic property (which languages have it, how is it defined, how it relates to other properties), he or she will be able to immediately access the database and project the results onto a screen during class.

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 phonology/phonetics and sociolinguistics, and potentially transform all of linguistics.

A central concern of the database will be the issue of interoperability (which is related to the issue of dissemination). 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 and even outside of academia.

The grant will be instrumental in training linguistics and computer science students, both graduate and undergraduate. These students will learn to work in an interdisciplinary team with active user feedback.

There have been a total of five undergraduates involved in the project, two of whom completed successful indpendent study projects. PI Collins has also advertised on the NYU Database of Undergraduate Research Opportunities for undergradutes who would like to become involved in the project (doing honors theses or independent projects): https://www.nyu.edu/cas/ugresearch/index.php

This project offers a convenient forum for people working on endangered and less studied languages to make their data public, and just as importantly, get their data integrated into current theoretical discussions. On the highest level, the benefit of the proposed project to society is to enable the study of linguistic diversity (and hence cultural diversity) found on earth.

Results from Prior NSF Support

Title: SGER: Prototype and Specifications for a Web-based Database of the Syntactic Structures of the World's Languages (SSWL).

Proposal Number: 0817202

PIs: Chris Collins and Richard Kayne

Amount: $44,663.00 (supplemental support: 23,470)

Period of support: 05/01/08-04/30/09 (extended to: 4/20/2010)

Summary of the results:

The SGER NSF grant resulted in a working prototype discussed above. There have also been two conference presentations:

Collins, Chris. 2009. *A Database of the Syntactic Structures of the World's Languages.* Presented at: Interfaces syntaxe-sémantique-pragmatique, Leysin, Switzerland, March 23-26.

Taylor, Michael. 2009. *The database of Syntactic Structures of the World's Languages: Progress report and challenges to date.* Presented at: Small Tools for Cross-Linguistic Research, June 15-16, Utrecht.

Title: Conceptual Data Integration for the Virtual Plant
PIs: Gloria Coruzzi (FAS-Bio) and Dennis Shasha
Period of support: 6/1/2005 - 5/31/2008
Award #: DBI-0445666
Amount: $1,592,964
Summary of results:

In that proposal, we have used a similar database schema (the property-as-value schema) in order to ensure that new properties could be entered flexibly. The result has been software with a small (roughly 80) user community (www.virtualplant.org).

The work has supported one PhD student, has resulted in the training in bioinformatics of two computer science master's students, and the training in bioinformatics of roughly eight biology students at NYU alone.

Publications:

Karen E Thum, Michael J Shin, Rodrigo Gutierrez, Indrani Mukherjee, Manpreet S Katari, Damion Nero, Dennis Shasha and Gloria M Coruzzi. "An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis." BMC Systems Biology 2008, 2:31 (04 Apr 2008)

Rodrigo A. Gutierrez, Miriam L. Gifford, Chris Poultney, Rongchen Wang, Dennis E. Shasha, Gloria M. Coruzzi and Nigel M. Crawford. "Insights into the genomic nitrate response using genetics and the Sungear software system." JXB Advance Access published online on April 29, 2007 Journal of Experimental Botany, doi:10.1093/jxb/erm079

Christopher S. Poultney, Rodrigo A. Gutirrez, Manpreet S. Katari, Miriam L. Gifford, W. Bradford Paley, Gloria M. Coruzzi and Dennis E. Shasha "Sungear: Interactive visualization and functional analysis of genomic datasets" Bioinformatics, 2007; Jan 15;23(2):259-61 doi: 10.1093/bioinformatics/btl496