Letter of Intent for the Digging into Data Challenge 2009

Chung-hye Han, Anoop Sarkar ({*chunghye,anoop*}*@sfu.ca*) and Fei Xia (*fxia@u.washington.edu*)

In this letter of intent for the Digging into Data Challenge 2009, we propose a research project that will allow complex structured searches of language data from a thousand different languages. It will enable a sophisticated but easy to use search capability over all linguistic articles available on the web. This will transform the data collection process for linguists by providing easy access to data through the use of precise searches for linguistic phenomena from a large set of papers that no one person could read through.

Our team is comprised of members from two participating countries: the Simon Fraser University (SFU) team in Canada led by Chung-hye Han (PI) and Anoop Sarkar (Co-PI), and the University of Washington (UW) team in US led by Fei Xia. Each team will include research assistants, some hourly students and/or contractors.

1 The significance of the project

Linguistics is the scientific study of language, and the object of study, the language data, is presented in research papers using *interlinear gloss text* (IGT) which conventionally consists of three lines: a language line for the language in question, a gloss line that contains a word-by-word or morpheme-by-morpheme gloss, and a translation line, usually in English. An example is shown in (1).

(1) Rhoddodd yr athro lyfr i'r bachgen ddoe
gave-3sg the teacher book to-the boy yesterday
"The teacher gave a book to the boy yesterday" (Bailyn, 2001)

One of the major obstacles that linguists encounter is finding data relevant to their research. Typically, to access the published data, linguists do a search using keywords on electronic databases such as Linguistics and Language Behavior Abstracts (LLBA) or Modern Language Association (MLA). This will return a list of articles that may be relevant for the research topic in question. Linguists then must read through the articles to look for the type of examples they need for their research. While this strategy may work for small amounts of data, it does not scale well. Validating or augmenting key components of a linguistic theory requires analyzing data from a large number of languages. As data availability increases, this strategy becomes time-consuming and inefficient. A search may not return many articles containing relevant examples simply because they are not tagged with the keywords used to do the search. Linguists thus often resort to compiling their own data, duplicating effort and resources.

There has been a lot of effort by the linguistic community to address the issue. For instance, *The Linguist List* compiles a long list of electronically available linguistic resources. The *Open Language Archives Community* (OLAC) acts as an online virtual library of language resources, and provides a search tool that searches several dozen online linguistic resources. While such resources help linguists find examples in a particular language or a language family, they do not allow complex queries for the linguistic phenomena that are the focus of research. Such queries would involve concepts that cannot be captured using a predefined set of keywords. For instance, no search engine currently allows a linguist to precisely specify a search in structural terms that would find, for instance, long distance scrambling examples, or examples with multiple wh-fronting, or other unforeseen queries for new phenomena in multiple languages. The search will be able to find examples in a paper which does not even contain the keywords that would be used to describe these structure-based searches.

The goal of the proposed work is to provide a facility that allows these kinds of queries; which we call *structural queries* in the rest of the paper. The work can be divided into three main components:

- Study queries and define a query language: We plan to conduct a survey and identify the types of queries that linguists would like to use to retrieve data relevant to their research. From the survey data, we will define a query language that is sufficient to representing the queries. The full proposal will contain a description of the type of query language we plan to use, its theoretical and practical properties and its effectiveness in search over this dataset.

- Data enrichment: As the query language becomes more sophisticated we will have to enrich the underlying database of examples to enable structural queries. We have already shown that we can bootstrap the structure needed using the fact that the gloss is provided in a language for which we have a parser that assigns structure. Our aim is to have a database containing all language data in the form of IGTs from all the linguistic papers available on the web.

- Build a web-based interface that will allow researchers to enter structural queries. Linguists will be able to browse the IGTs from the search and refine the queries to obtain exactly the group of linguistic examples that will be useful to their research. Query results will be organized by language, language family, positive vs. negative examples, and other useful traits. Popular queries will be shared between users to help create a community of data-driven scholarship in linguistics.

The proposed project will allow linguists to directly search for the examples that illustrate the linguistic phenomena that are the topic of their research in multiple languages through structural queries. For instance, structural queries on example sentences enriched with syntactic parses will ensure that all the examples that meet the search criteria will be returned, whether the example is tagged with the relevant keywords or not. The proposed project will facilitate web-based data-driven scholarship in linguistics, as our database can contain examples from both published articles and manuscripts, as long as they are published on the web. Being able to search across many different languages will facilitate typological research and advance the theory on cross-linguistic variation, a major research theme in theoretical linguistics.

2 The dataset

We plan to use Online Database of INterlinear Text (ODIN), which currently contains about two hundred thousand IGT instances from close to a thousand languages and is in the process of being expanded in terms of instances and languages. We choose this database for several reasons. First, the database was built and maintained by the UW team, so we have easy access to all the data in the database including the IGT instances, the linguistic documents from which IGTs have been extracted, and all the intermediate results produced by various components of the ODIN system. Second, the IGT data has been manually checked with respect to the content and the language code; therefore, the dataset is very clean. Third, it is one of the largest datasets of linguistic examples in the linguistics field.

Since our data creation programs and methodology have been thoroughly tested on other sources, our plan is to contact the owners of other online linguistic resources such as Project Muse (muse.jhu.edu), LingBuzz (ling.auf.net/lingbuzz), Semantics archive (semanticsarchive.net), and Rutgers Optimality Archive (roa.rutgers.edu), and to use the linguistic papers stored in those archives. We expect the size of ODIN to increase dramatically in the near future as more documents are processed and added to the database.

3 Prior work

3.1 The SFU team

The SFU team includes Chung-hye Han (PI), Anoop Sarkar (co-PI), and two research assistants.

Chung-hye Han is an Associate Professor at Simon Fraser University in the Department of Linguistics and the director of the Experimental Syntax Laboratory. Her main areas of research are syntax and semantics and their interface in natural language, and computational applications of linguistic theories. She currently holds a SSHRC Standard Research Grant to investigate grammar competition in language acquisition and an NSERC Discovery Grant on computational semantics. She has extensive research experience with online corpora in her linguistics and computational work. In her linguistics work, she has used the Penn-Helsinki Parsed Corpus of Middle English (www.ling.upenn.edu/hist-corpora/) to make generalizations about clause structure in English, and Sejong corpus (www.sejong.or.kr) to extract supporting data for her semantics work on binding in Korean. In her computational work, she has led a project that developed a Korean/English parallel Treebank, corpora annotated with rich linguistic information, which

was released to the public through Linguistic Data Consortium. Using the Treebank, she has built computational resources and tools for the analysis of Korean, including a Korean morphological tagger and a Korean parser. Her expertise in syntax and corpus-based methodology will be helpful in studying the type of structural queries useful to linguists and in defining the query language usable by linguists. More information can be obtained at her home page (www.sfu.ca/∼chunghye).

Anoop Sarkar is an Assistant Professor at Simon Fraser University in the School of Computing Science and co-director of the Natural Language Laboratory. His research is focused on the application of machine learning methods to natural language processing (NLP). In particular, he has contributed to the study of semi-supervised learning for NLP where partially annotated data is augmented by using learning methods on unstructured data. He holds an NSERC Discovery Grant, an NSERC Equipment Grant, an IBM Faculty Award, and MITACS awards for research into this topic. Since we can treat the interlinear text as providing partial information about various languages, semi-supervised learning methods will be useful in adding relevant structural information into the dataset in order to provide a flexible search infrastructure in our project. His expertise is also in the area of natural language parsing and stochastic grammar formalisms which will be useful in the definition and implementation of the notion of structural queries over interlinear text. Anoop Sarkar has over fifty publications in the field of computational linguistics and has published extensively on the topic of semi-supervised learning in NLP. More information is available on his home page (www.cs.sfu.ca/∼anoop).

3.2 The UW team

The UW team includes Fei Xia (the PI), William Lewis (a contractor), and a research assistant.

One of Fei Xia's major research activities, supported by two NSF grants (BCS-0720670 and BCS-0748919), is to create a framework that allows the rapid development of resources for resource-poor languages and then use the automatically created resources to perform cross-lingual study on a large number of languages to discover linguistic knowledge. As part of the research, Xia designed new algorithms for IGT detection (extracting IGT from linguistic documents) and language ID (identifying language code for extracted IGT). The algorithms have been incorporated in the new ODIN system and consequently increased the size of ODIN by more than 360%. She has also developed a method to enrich IGT data and then extract syntactic information (e.g., context-free rules) to bootstrap NLP tools such as POS taggers and parsers. The enrichment algorithm first parses the English translation with an English parser, then aligns the language line and the English translation via the gloss line, and finally projects syntactic information (e.g., POS tags and phrase structures) from English to the language line. The algorithm was tested on 538 IGTs from seven languages and the word alignment accuracy was 94.1% and projection accuracy (i.e., the percentage of correct links in the projected dependency structures) was 81.5%.

William Lewis was the PI of the Data-Driven Linguistics Ontology (DDLO) project, an NSF-funded SGER (BCS-0411348) that supported the creation of the original ODIN system. He has already consulted with a number of syntacticians and other linguists about the types of queries that would be of most interest to linguists. The ODIN database and the projection algorithm from Xia and Lewis's prior work will be used in our proposed work. The query list compiled by Dr. Lewis can serve as a starting point for defining a query language.

3.3 Teamwork

The PI's of the two teams have a history of working together. They were engaged in common projects while at the University of Pennsylvania. The projects have been diverse, involving the building of linguistic resources such as Treebanks, and building various NLP tools such as parsers and morphological analyzers. They were involved in the XTAG project at the University of Pennsylvania which was building a wide-coverage English computational grammar using linguistic expertise (Chung-hye Han), computational meta-grammar techniques (Fei Xia) and corpus evaluation (Anoop Sarkar) within the same project.