Interoperability for Language Documentation

The Role of Semantic Web Tools

Jeff Good
Department of Linguistics
University at Buffalo
Buffalo, New York, USA
jcgood@buffalo.edu

Tom Myers
N-Topus Software
Hamilton, New York, USA
tommyers@dreamscape.com

Alexander Nakhimovsky
Department of Computer Science
Colgate University
Hamilton, New York, USA
adnakhimovsky@colgate.edu

***Abstract*—The threat of imminent extinction of perhaps half of the world's languages has led to increased efforts aimed at the their documentation. Such documentation necessarily makes use of a suite of tools to handle a diverse array of tasks from time segmentation of recordings, to transcription, to annotation, to publication. No one tool can support all aspects of the workflow for language documentation, but there is, at present, no general solution for interoperation among the tools required. We outline a solution to this problem based on Semantic Web technologies and further suggest that OpenOffice's capabilities for working with RDF make it an ideal choice for tool development for those aspects of workflow related to the production of publications.**

Keywords-Semantic Web, RDF, interoperability, linguistics

I. INTRODUCTION

The last two decades have witnessed increased concern over the threat of extinction of perhaps half of the world's current stock of languages [1]. One consequence of this concern has been increased funding in support of language documentation: collecting data in the field, creating digital archives, and standardizing metadata (among other things). What it means to document a language is an evolving notion [2,3], but some of its core aspects are widely agreed upon. Two kinds of resources, lexicons and glossed texts, are seen as crucial elements of a full set of documentary materials. In the domain of glossed texts, one format, in particular, has gained widespread acceptance as a de-facto standard: the so-called Interlinear Glossed Text (IGT), where transcribed linguistic units (sentences, phrases, words, morphemes) are associated with glosses that are aligned on the page or the computer screen with the units they gloss. We describe it here in detail since it is a data type not well known outside of linguistics.

Structurally, IGT can be modeled as a tree in which each level contains an elaboration (with glosses) of the units of the preceding level; the root of the tree is the entire initial unit together with its transcription and a translation into the language of analysis (English, Russian, etc.) In addition, the use of time-based data (audio and video) has necessitated the addition of a new feature to IGT—time alignment, where the transcriptions and glosses are associated directly with time segments of audio or video recordings.

To illustrate, an example of IGT, from Lezgian [4] is given below in (1). The first line represents a written representation of the language being described, the second line glosses each of the words in the described language, and the last line offers a free translation of the entire sentence. Hyphens in the first line separate words into constituent morphemes and, in the second line, separate components of the gloss into a unit associated with each morpheme. As can be seen in (1), a gloss is not simply a short translation since it is a mix both of translations and technical linguistic abbreviations.[1]

(1) Gila abur-u-n ferma hamišaluğ güğüna amuq'-da-č.
now 3p-OBL-GEN farm forever behind stay-FUT-NEG
"Now their farm will not stay behind forever."

The example in (1) is representable by the tree, described using a table, given in Table I.

TABLE I. TREE REPRESENTATION OF IGT EXAMPLE

Gila aburun ferma hamišaluğ güğüna amuq'dač.									
Now their farm will not stay behind forever.									
Gila	aburun			ferma	hamišaluğ	güğüna	amuq'dač		
Gila	abur	u	n	ferma	hamišaluğ	güğüna	amuq'	da	č
now	3p	OBL	GEN	farm	forever	behind	stay	FUT	NEG

The levels of the tree (and rows of the table) are: the original sentence; English translation; the sentence broken into words; the words broken into morphemes, both lexical and grammatical; and morpheme-by-morpheme gloss, with grammatical morphemes glossed using technical abbreviations (e.g., GEN for Genitive case, FUT for Future tense, etc.).

Once glossed texts and a lexicon of morphemes are created for a language, they can, in principle, be the basis of the creation of any number of additional products, from an edited collection of stories of use to a community creating pedagogical materials to an academic descriptive grammar to a spell-checker.

The specialized needs of language documentation have prompted the creation or wider use of custom software tools. A

[1] The Leipzig Glossing Rules summarize recommend standards in the use of such abbreviations (http://www.eva.mpg.de/lingua/resources/glossing-rules.php).

pioneer and a leader of this effort is SIL International[2] (SIL). SIL's Shoebox, a database utility optimized for the creation of IGT and lexicons, goes back to the 1980s, but one of its key features, the integration of a text database with a lexical database to facilitate parsing, has yet to be effectively replicated in any other widely-used tool (which is due more to social than technological issues). Despite its age and noteworthy limitations—Shoebox and its more recent version Toolbox use an old data format going back to the 1980s that offers no data validation—it remains in wide use, only slowly being replaced by its more powerful successor, FLEx, also from SIL.[3] In the meantime, in Europe, a major center for developing language documentation software was created at the Max Planck Institute for Psycholinguistics in Nijmegen, with stable long-term support from the Volkswagen Foundation. Their program ELAN is the primary tool for time-aligned glossing for language documentation and the most widely-used component of their Language Archiving Technology (LAT) suite of programs.[4]

Toolbox, ELAN and FLEx are the most common specialized tools for language documentation.[5] While all three do some things well, none covers the entire range of language documentation tasks: ELAN has no support for lexicon building while Toolbox and FLEx have no support for time alignment or waveform-based playback that is of great help in transcribing the content of a media file. In addition, none of them provides publishable outputs, a gap that is typically filled by Microsoft Word, even though it offers no ready means of interoperating with custom linguistic software. The custom software itself also offers very limited interoperability, most of it one-way into ELAN that provides import modules for Toolbox and FLEx.[6] (The FLEx module was developed in 2009 in collaboration with Tom Myers and other members of the NSF-funded *Five Languages of Eurasia* project.[7] This was the first known instance of collaboration between MPI and SIL developers.)

Work on the FLEx import module has clearly demonstrated the difficulty of building interoperability on top of a decade of uncoordinated development. Returning to the notion of IGT as a tree, ELAN allows the user to create as many tree levels (called tiers in ELAN) as necessary for analysis; the names of the tiers must be unique but otherwise are up to the user. ELAN also allows, in effect, two or more overlapping trees in a single file. The need for such structures arises when there are two or more speakers whose speeches overlap, for example; or when the analysis tracks both speech and gestures that are not synchronized. FLEx, by contrast, allows only a single tree, and its levels are fixed: paragraph, phrase, word, morpheme. It is thus possible to create structures in ELAN that are not representable in FLEx. On the other hand, every morpheme in FLEx is linked to a lexicon entry so that a revision in the analysis of a text may result in a change in the lexicon and a global replace in the accumulated corpus. ELAN has no such functionality, and no means to represent lexicon links.

It would be an impossible expense to extend ELAN with the functionality of FLEx and vice versa—not to mention the fact that each tool's user interface is optimized for the tasks it was originally intended to perform. The most reasonable way to proceed is to provide both ELAN and FLEx with a unified underlying representation that will represent data from both programs. Some parts of that representation will be ignored by ELAN, and some other parts will be ignored by FLEx, but a lossless round trip between the two programs would be possible (see also [7] for discussion). This approach can also broaden the interoperability between ELAN and Shoebox, and indeed create interoperability between the two SIL programs, Shoebox and FLEx. Finally, the problem of interoperability extends to metadata as well. Digital archives on the web require standard metadata for discovery and retrieval. In the case of language archives, two standards have emerged, one from MPI Nijmegen, the other from the Open Language Archives Community[8] (OLAC), and additional representations are needed to make them compatible.

The obvious choice for a unifying representation is the semantic graph of RDF/OWL. One of its main purposes is specifically to merge heterogeneous representations of overlapping data on the Web. It also has a standard query language, and a rapidly growing arsenal of tools for development. Our goal in this paper is to elaborate how semantic technology can establish interoperability and data integration in the field of language documentation. We examine issues of compatibility among different formats and models for IGT, focusing on four tools: ELAN, Shoebox, FLEx, and OpenOffice.org (OpenOffice)[9]. We discuss its place in the overall model we are developing here in the next section.

II. THE ROLE OF OPENOFFICE

The latest version of OpenOffice has standard interfaces for attaching semantic information to elements and text ranges of an OpenOffice document. The RDF/XML file representing such semantic information can be easily produced, e.g. via a Web application that extracts it from the compressed representation of the document file. OpenOffice has five characteristics that we believe make its use advantageous for language documentation:

- It is an office suite that, for the purposes of language documentation, is essentially equal to MS Office in functionality and usability.
- It is Free and Open Source.

[2] http://sil.org/

[3] Shoebox is, at present, continued by a tool going under the name of Toolbox. See http://sil.org/computing/catalog/show_software.asp?id=79. FLEx (http://sil.org/computing/fieldworks/flex/) stands for FieldWorks Language Explorer tool.

[4] See http://www.lat-mpi.eu/tools/elan/.

[5] There are also even more specialized tool for limited purposes: PRAAT (http://praat.org/) is a powerful tool for phonetic analysis; Lexique Pro (http://lexiquepro.com/), also developed by SIL, is a niche tool for creating and publishing lexicons; Transcriber (http://trans.sourceforge.net/) is excellent for transcribing audio files but its development has slowed over the last several years.

[6] See [5] and [6] for related points.

[7] http://www.philol.msu.ru/~languedoc/eng/

[8] http://language-archives.org/

[9] http://www.openoffice.org/

- It uses standard and open document formats.
- It is readily extensible by software modules written in standard and widely-used languages (Java, XSLT).
- It has built-in support for RDF.

It is possible to extend OpenOffice with modules and interfaces that will support language documentation work. Some such extensions have already been developed within the Pangloss project.[10] It is also possible to maintain the entire RDF graph as a triples table within an OpenOffice document. One can easily construct a workflow that is mostly based in OpenOffice, exporting the documents to ELAN or FLEx or even Toolbox to do specialized work, bringing the results back as an RDF graph, and merging the graphs also within OpenOffice. Thus, we can use Semantic Web technologies not only to enhance interoperation among specialized tools but also to integrate a key kind of general-purpose tool, the word processor, into the linguist's workflow.

OpenOffice can also be used to improve metadata collection. Input–output filters can be written for both the IMDI metadata format[11] from MPI Nijmegen and the OLAC metadata format[12], two commonly used formats within linguistics, as mentioned above. In both cases, metadata will be internally represented by RDF graphs, facilitating interchange between the two formats. Whether or not this will result in significantly improved metadata collection is an open question, but it seems plausible that if users are given a familiar office-document interface within the same program that they use for data creation, their metadata habits will become more reliable. OpenOffice's possibilities for extension would also allow for the development of modules which would mediate the exchange of metadata updates and revisions between the linguist and a language archive, an aspect of workflow not well supported at present.

III. Workflows to support

Language documentation has become the subject matter of a new subfield of linguists called documentary linguistics (see [2] and the collected papers in [8]). Conceptually, a documentary linguist starts with a collection of field recordings. Some of these records are "born digital," others are legacy recordings that have to be digitized. Key products of the linguist's work are a lexicon and a corpus of IGT based on these recordings. The work encompasses a number core steps, shown below.

1. Collect recordings of the target language, typically in a field setting.
2. Create a time-aligned transcription of the recording, with segmentation at approximately the sentence level.
3. Associate the transcription with free translations into an academic language at approximately the sentence level.
4. Associate each word of the transcription with a gloss (i.e., an abbreviated description of the word's semantics and morphosyntax), using a lexical database, if available.
5. Associate items in the text with entries in a lexical database, possibly creating new entries.
6. Produce publishable versions of the lexicon and the text corpus.

The order of steps represents an idealization which will not always be followed in practice.

There will be multiple loops over the same material. Different applications may choose to elaborate some steps as needed to address specific research questions. It is therefore important to support several document flows through the software applications, to accommodate different work styles, and requirements of a specific situation. With lossless round trips between ELAN and FLEx, ELAN or FLEx and OpenOffice, and OpenOffice and Shoebox, the connectivity will be complete. Furthermore, new tools will be able to exploit the common RDF graph as well and, thereby, interoperate other tools whose data can be exposed in the graph at relatively low cost.

IV. The Role of the GOLD Ontology

An agreed-upon ontology of linguistic entities is a necessary condition for merging RDF graphs expressing linguistic data. Substantial work has already been done in this area in the context of the development of the General Ontology for Linguistic Description[13] (GOLD; [9]) on which the efforts described here can be based. GOLD is quite explicitly intended to be on an ontology for linguistic description, not language itself, which makes it appropriate for a project like this one which is intended to facilitate the linguist's work in processing their data rather than, say, machine translation. Obviously, there is a strong relationship between linguist's descriptive categories and categories of relevance to language to itself, though for the purposes of creating interoperable language data it is useful to separate the linguist's conception of their data from any abstract categories underlying language itself.

GOLD, at present, is not well-developed across the range of widely-used linguistic data structures (for example, it defines most of the crucial concepts relating to IGT but not lexicons) and would need to be extended with classes for IGT and lexicon entities to be of maximal use for the work discussed here. For the purposes of development this can be achieved using GOLD's Community of Practice Extension mechanism (see [10]), which allows GOLD to be extended in a normative fashion without needing to alter GOLD itself. The mechanism also provides a framework for developing and testing new concepts before they are migrated into GOLD if this is deemed desirable.

[10] http://code.google.com/p/rosetta-pangloss/

[11] http://www.mpi.nl/imdi/

[12] http://www.language-archives.org/OLAC/metadata.html

[13] http://linguistics-ontology.org/

Despite extensive work on the development of GOLD, it has not yet been widely exploited for the processing of linguistic data, largely due to inadequate tool support. Work is being undertaken at present to make use of GOLD to create an interoperable lexical datanet in the context of the Lexicon Enhancement via the Gold Ontology project[14] (LEGO) and the model described here would promote the widespread use of GOLD in the creation of IGT and links between IGT and lexical data. We therefore see the development of this interoperation model as contributing to the growth of GOLD by making it more straightforward for the linguistic community to create data using GOLD concepts.

V. Major Tasks

There are several key tasks which require detailed work for the model we are elaborating here to be successful. We discuss each of these in turn. Most of them require not only technical work but also standardization and institutional collaboration between the main stakeholders.

A. Development of a uniform system for GUIDS

From a Semantic Web perspective, one of the most important technical desiderata is to establish a way to maintain persistent GUIDs (Globally Unique Identifiers) across any tools which are to be supported by the workflow described here, including ELAN, FLEx, Shoebox, and OpenOffice. For Semantic Web compatibility, these should either take the form of URIs themselves or be readily translatable as URIs. The GOLD community uses PURL[15] (Persistent Uniform Resource Locators), and this approach could be adapted for the purposes of our project, but two hurdles need to be overcome. First, we need a means of coding the relevant GUIDs (or references to them) in the non-RDF formats generated by ELAN, FLEx, and Shoebox. (ELAN and FLEx use XML formats; Shoebox uses a non-XML markup.) Second, we need a general agreement on the kinds of objects to which GUIDs will be assigned in both general terms and in terms of the data models assumed by the target software. While not an ideal long-term solution because of its visibility to the user, in the short-term GUIDs can be stored using custom fields in the various software applications which all allow creating such user-defined fields.

A model for this solution already exists in ELAN's support of import/export of Shoebox files, where ELAN adds fields to a Shoebox database corresponding to timestamps in the media file. These fields have no significance for Shoebox but serve as something akin to GUIDs when imported into ELAN. The issue of what kinds of abstract object should receive GUIDs, and how such objects relate to tool-specific data models is, in principle, a much more difficult one than simply generating and maintaining the GUIDs themselves. However, for one of the datatypes targeted here, IGT, there is widespread agreement on the core features of the data type, which therefore makes the problem much more straightforward in at least this limited domain. For lexicons, relevant work on this problem is being done by the LEGO project discussed above.

B. Extension of the GOLD ontology

In a Semantic Web context, the easiest way to describe and disseminate shared data models to facilitate interoperation is by defining them within an OWL ontology. For the data of interest here, there is already a comparatively well-developed ontology, GOLD, as discussed above. Since GOLD has not yet been applied to a project like this one which is attempting to simultaneously facilitate data and tool interoperability, it inevitably does not contain all of the concepts needed to support all aspects of the required functionality. (For example, while it defines a concept corresponding to a stretch of IGT, it does not have a data structure defining a sequence of stretches IGT, which would be required to describe a whole text as a list of analyzed sentences, which would clearly be needed.)

C. OpenOffice development

Not surprisingly, the developers of tools designed specifically for linguistic analysis like FLEx, ELAN, and Shoebox, already have close ties to the linguistics community and, in general, are open to collaboration with other linguistics projects. However, relatively few linguistics projects have worked with OpenOffice and, therefore, there is a lack of expertise for OpenOffice development within the linguistics community. This problem can be mitigated by dealing with those aspects of the problem not specific to OpenOffice, but rather relevant to handling interchange among linguistic data formats in a more general way, in a separate toolkit from the tools specifically required to interface with OpenOffice. It would even be possible to develop such tools under the aegis of existing projects like the Natural Language Toolkit[16] [11] or the e-Linguistics Toolkit[17] [12], the former of which already has some support for the Shoebox format [13].

The problems specifically associated with interacting with OpenOffice and adapting its user interface in ways that allow linguists to work with their data in ways that are both intuitive and interact well with OpenOffice's other features are obviously not easy ones. However, the work of the Pangloss project, mentioned above, established that OpenOffice's existing capabilities offer solutions to key problems of data import, manipulation, and export. So, the work seams feasible, even if there are still a number of issues to be worked out.

VI. Conclusion

We have argued here that Semantic Web technologies can be used to help solve a major problem of documentary linguistics: Multiple tools are needed to complete critical tasks which may share basic data models at some abstract level but which implement those models in ways which hinder interoperation. In addition, specialist tools tend to focus on specialist problems (e.g., the annotation of texts for linguistic categories) and, therefore, do not support more general aspects of workflow, in particular the production of publications. By translating the outputs of tools presently in use to RDF/OWL with concepts drawn from the GOLD ontology or an appropriately defined extension to the ontology, devising a

[14] http://linguistlist.org/projects/lego.cfm

[15] http://purl.oclc.org/

[16] http://www.nltk.org/

[17] http://uakari.ling.washington.edu/e-linguistics/eltk.html

system of GUIDs usable across tools, and making use of RDF support within OpenOffice, the documentary linguist's workflow can be greatly improved. A welcome side-effect of such an approach is that it will also allow them to more readily produce data that can naturally interoperate with other Semantic Web data and thus facilitate additional kinds of interoperation. Semantic Web technologies, therefore, are valuable not only for their initially formulated purpose of enhancing the World Wide Web but also for streamlining the workflow of individual researchers. While here we have only described, rather than implemented, a solution to tool interoperation in documentary linguistics, key pieces of our proposed solution are already in place, or have been worked on, and we believe our overall solution to be feasible, if not necessarily simple.

REFERENCES

[1] Krauss, Michael. 1992. The world's languages in crisis. *Language* 68:4–10.

[2] Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36:161–195.

[3] Woodbury, Anthony C. 2003. Defining documentary linguistics. In Peter K. Austin (ed.) 2003. Language Documentation and Description, Volume 1. London: Hans Rausing Endangered Languages Project.

[4] Haspelmath, Martin. 1993. A Grammar of Lezgian. Berlin: Mouton.

[5] Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79:557–582.

[6] Trippel, Thorsten. 2006. The missing links in documentary linguistics: An approach to bridging the gap between annotation tools. Proceedings of the E-MELD Workshop on Digital Language Documentation: Tools and Standards: The state of the art. Lansing, Michigan. <http://emeld.org/workshop/2006/papers/trippel.html>

[7] Cochran, Michael, Jeff Good, Dan Loehr, S. A. Miller, Shane Stephens, Briony Williams, and Imelda Udoh. 2007. Report from TILR Working Group 1 : Tools interoperability and input/output formats. Toward the Interoperability of Language Resources Workshop, Stanford, California, July 2007. <http://linguistlist.org/tilr/working-group-reports/Working%20Group%201.pdf>

[8] Gippert, Jost, Nikolaus P. Himmelmann and Ulrike Mosel, (eds.) 2006. Essentials of Language Documentation. Berlin: Mouton de Gruyter.

[9] Farrar, Scott and D. Terence Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International* 7:97-100.

[10] Farrar, Scott. and William D. Lewis. 2007. The GOLD Community of Practice: An infrastructure for linguistic data on the Web. *Language Resources And Evaluation* 41:45–60.

[11] Bird, Steven and Edward Loper. 2004. NLTK: The Natural Language Toolkit. The companion volume to the proceedings of 42nd annual meeting of the Association for Computational Linguistics. Barcelona: Association for Computational Linguistics. 214–217. <http://www.aclweb.org/anthology/P/P04/P04-3031.pdf>

[12] Farrar, Scott and Steven, Moran. 2008. The e-linguistics toolkit. Proceedings of e-Humanities—an emerging discipline: Workshop in the 4th IEEE International Conference on e-Science. IEEE/Clarin, IEEE Press. <http://www.clarin.eu/system/files/private/FarrarMoran08_eling.pdf>.

[13] Robinson, Stuart, Greg Aumann, and Steven Bird. 2007. Managing fieldwork data with Toolbox and the Natural Language Toolkit. *Language Documentation & Conservation* 1:44–57. <http://hdl.handle.net/10125/1725>