2 Whither GOLD?

D. Terence Langendoen

In the Beginning

In the Language Digitization Workshop, the kickoff meeting for the Electronic Metadata for Endangered Language Data (E-MELD) project,[1] in Santa Barbara, California, in June 2001, I made two presentations on linguistic markup (annotation). The first described the general nature of the markup of sound and text files and of databanks that can be derived from them, and the second the work of the Text Encoding Initiative (TEI)[2] on text markup, particularly the chapters on simple analytic mechanisms (SAM), feature structures (FS), and feature system declarations (FSD) of Sperberg-McQueen and Burnard, editors (1994).[3] In these presentations, I made the following observations.

1. For electronically encoded resources to be maximally useful within and across linguistic communities, there must be agreement on transcription and analytic terminology standards within and across languages, with procedures for settling differences among transcription and terminological practices.
2. Linguistic databanks can be developed along the lines of commonly used print data resources, such as comparative wordlists, morphosyntactic paradigms, thesauri, rhyming dictionaries, mono- and multilingual sense dictionaries, and reference grammars, in addition to digitally born types of databanks, such as treebanks and interlinear glossed text (IGT) repositories.
3. Because the TEI recommendations for FS and FSD have not been widely adopted, presumably because of their complexity and the lack of extensive testing on linguistic data,[4] it might be a good idea to try to reach consensus on a simpler form of FS markup using XML that would be adequate to the needs of the linguistics community.[5]

The Birth of GOLD

However, my two newly recruited research assistants, Scott Farrar and Will Lewis, quickly convinced me that a better path would be to take advantage of the infrastructure of the Semantic Web announced in Berners-Lee, Hendler, and Lassila (2001) that was under

development as a reasoning platform for all publicly shared data on the web. Specifically, we would begin to build an ontology for the concepts needed for linguistic analysis as a subcomponent of an upper ontology, such as of SUMO, the Standard Upper Merged Ontology (Pease and Niles 2002; Pease 2007). This ontology, like SUMO and like other domain-specific web ontologies, would be written in one of the markup languages being constructed for the Semantic Web, such as OWL-DL, not in XML. Technically, this did not violate the E-MELD project's endorsement of XML as the markup language of choice for linguistic annotation. Such annotation could still be written in XML but the interpretation of its tags would be determined by the concepts they referred to (i.e., pointed to) in GOLD. FS would be treated as a data type, with its interpretation determined by its connections to GOLD. In Lewis, Langendoen, and Farrar (2001), our first presentation following the kickoff meeting, we pointed out that the real need of the community we were serving would be "to obtain information about endangered languages on the World Wide Web without regard to the tagging schemes that are used in the various websites they consult. Thus [we] cannot impose a markup standard for endangered language websites, even implicitly by developing a data interchange format [such as the TEI]." To make sense of this markup chaos, we proposed the development of a "metatagging" scheme consisting of "a knowledge base and its accompanying tools [that] will act as an interlingua for data comparison," the key to which is an ontology. At the time we submitted the paper for presentation, we had already created an ontology for morphosyntactic concepts with hundreds of nodes drawn from resources provided by the Summer Institute of Linguistics and the Dokumentation Bedrohter Sprachen (DOBES) project, two general linguistics term sets and several dictionaries and grammars of endangered languages, but we had not yet given it a name. At the workshop, we announced our choice: the General Ontology for Linguistic Description (GOLD).

The Development of GOLD within the E-MELD Project

Presentations about GOLD were made at every annual E-MELD workshop from 2002 through the end of the project in 2006, as well as at numerous conferences and workshops around the world, including Langendoen, Farrar, and Lewis (2002), Farrar, Lewis, and Langendoen (2002), Farrar and Langendoen (2004), Simons et al. (2004b), and Lewis (2006). GOLD came to the attention of the linguistics community at large through the publication of Farrar and Langendoen (2003), and the Linguist List began hosting GOLD's website in 2006.[6] Two major accomplishments occurred during this period. First, a proof of concept was achieved for the metatagging scheme proposed in Lewis, Langendoen, and Farrar (2001) to carry out searches over differently encoded datasets of IGT and electronic dictionaries (Simons et al. 2004a, 2004b). Second, the Online Database of Interlinear Text (ODIN) was set up, in which users could select from a list of GOLD morphosyntactic concepts and find instances of IGT harvested from the web in more than 700 languages

that contain morphemes referencing them (Lewis 2006). However, little other progress was made beyond the further refinements of the conceptual structure for morphosyntax, a situation that has continued to this day.

GOLD after E-MELD

At the conclusion of the E-MELD project in 2006, Scott Farrar continued his work for several more years on GOLD's conceptual backbone, particularly on the notion of the linguistic sign itself (Farrar 2007), and on the relative merits of the various versions of OWL for implementing GOLD (Farrar and Langendoen 2010). Will Lewis along with Fei Xia and other collaborators have extended the ODIN's data coverage to nearly 1,300 languages and over 130,000 instances (Lewis and Xia 2010; Xia et al. 2014).[7] Finally, the Lexical Enhancement via the GOLD Ontology (LEGO) project—begun in 2008 under the direction of two of the E-MELD principal investigators, Anthony Aristar and Helen Aristar-Dry, together with Jeff Good—has tagged the entries of 12 lexicons and 11 wordlists with links to GOLD concepts to support cross-linguistic search much in the manner of ODIN.[8] Neither project, however, has extended GOLD's conceptual coverage.

What's Next?

The question of how to sustain the GOLD effort at the end of the E-MELD project was considered by Farrar and Lewis (2007), who proposed that communities of practice take responsibility for constructing GOLD subcomponents for particular languages and language families, and collaborate on determining which cross-linguistic constructs should be incorporated into GOLD itself. However, no effective action has yet been taken on their recommendations. Bender and Langendoen (2010: sec. 4) envisioned a future research environment for linguists called Digital Infrastructure that supports Linguistic Inquiry (DILI) that builds on past and current work and provides the following three capacities, among others:

1. Ready access to large amounts of digital data in text, audio, and audio-video media about many languages, which are relevant to many different areas of research and application both within and outside of linguistics.
2. Facilities for comparing, combining, and analyzing data across media, languages, and subdisciplines, and efforts to enrich DILI with their results.
3. Services to support seamless collaboration across space, time, (sub)disciplines, and theoretical perspectives.

We went on to say, "It is not required that there be a single overarching network for all the annotations in DILI, but it would be desirable if sense could be made of the relations among conceptual networks for different annotation schemes, particularly those that represent

different theoretical perspectives.… This view of the role of conceptual encoding was recently articulated in Farrar and Lewis (200[7]), along with a plan for how to achieve it." Lest this vision be dismissed as pie-in-the-sky fantasy, we pointed out that similarly ambitious research environments already exist for such fields as biochemistry, nanotechnology, and astronomy—so why not linguistics?

Perhaps the lack of such research environments in linguistics is a result of the long history of our field, which sprang up independently in varying language and cultural communities in several parts of the world, or perhaps it's the fractiousness of us linguists, or even the notion that it's harder for ours than for most, if not all, others' fields of inquiry. I think of Scott Farrar, struggling with the problem of characterizing the notion of the linguistic sign for use in GOLD, who finally formulated something that came fairly close to what Louis Hjelmslev (1943 [1962]) proposed. If Farrar is at least in the right ballpark, then the underlying logic will have to be richer than that provided by OWL-DL, which is a decidable version of first-order logic, even putting aside the wondrous complexities of the logical forms needed to represent, for example, reciprocal constructions in the world's languages.[9] The reason is that Farrar's forms have to relate to each other compositionally, both for meaning and for expression.[10] The composition of meanings is governed by whatever conceptual (logical) operation is called for to combine them, such as binding a predicate variable by a quantifier. At the same time, the composition of expressions is governed by a mereological (also logical, but with a different partial ordering) operation such as concatenation, if the expressions are represented as strings, so that at least two distinct logical systems have to be synchronized. The challenge, I think, is well worth undertaking, starting with our taking a fresh look at the proper way to construct conceptual networks for linguistic analysis and annotation.

Notes

1. E-MELD was funded by the US National Science Foundation grant 0094934 to Wayne State University with a subcontract to the University of Arizona.

2. TEI was funded by the US National Endowment for the Humanities, Directorate General XIII of the Commission of the European Communities, Andrew W. Mellon Foundation, and Social Science and Humanities Research Council of Canada.

3. As chair of the TEI Committee on Text Analysis and Interpretation and of the Work Group on Linguistic Description, I had overall responsibility for the preparation of these chapters. The editors and the members of the committee and of the work group were active contributors, particularly Mitch Marcus, who convincingly argued for the importance of FS at the first work group meeting, and Gary Simons, who showed how sets of FS can be validated by FSD, the latter being in effect (partial) grammars of the languages described by those FS sets; see Langendoen and Simons (1995).

4. Mitch Marcus and I gave a tutorial entitled "Tagging Linguistic Information in a Text Corpus" at the June 1990 ACL meeting in Pittsburgh, in which we described the guidelines in preparation for both the Penn Treebank (PTB) and the TEI recommendations for SAM and FS. The PTB, along

with its encoding scheme for English syntactic structure, eventually caught on to become a major resource for computational linguists; the TEI recommendations did not. I still vividly recall Ken Church's making exactly that prediction following our presentation.

5. At its kickoff meeting, the E-MELD project endorsed XML as the markup language it would recommend for linguistic annotation. TEI was originally encoded in SGML but later converted to XML.

6. http://linguistics-ontology.org/.

7. http://odin.linguistlist.org and http://faculty.washington.edu/fxia/odin/. More recently, the ODIN resource has been enriched with the addition of syntactic tiers, and graphical interface tools, but with the links to GOLD removed. For details see Xia et al. (2016).

8. LEGO was supported by the US National Science Foundation award 0753321 to Eastern Michigan University; see http://lego.linguistlist.org.

9. Berners-Lee, Hendler, and Lassila (2001) insisted that the Semantic Web should not deal with the semantics of natural languages. Still, the conceptual networks for linguistic annotation *will* eventually have to deal with them.

10. And for other things as well, but I leave them also aside.

References

Bender, Emily M., and D. Terence Langendoen. 2010. "Computational Linguistics in Support of Linguistic Theory." *Linguistic Issues in Language Technology* 3 (1): 1–31.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284 (5).

Farrar, Scott. 2007. "Using 'Ontolinguistics' for language description." In *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts*. Edited by Andrea Schalley and Dietmar Zaefferer, 175–192. Berlin: Mouton de Gruyter.

Farrar, Scott, and D. Terence Langendoen. 2003. "A Linguistic Ontology for the Semantic Web." *Glot International* 7 (3): 97–100.

Farrar, Scott, and D. Terence Langendoen. 2004. "Comparability of Language Data and Analysis: Using an Ontology for Linguistics." *Symposium on Endangered Data vs. Enduring Practice, 80th Annual Meeting of the Linguistic Society of America*, Boston.

Farrar, Scott, and D. Terence Langendoen. 2010. "An OWL-DL Implementation of GOLD: An Ontology for the Semantic Web." *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*. Edited by Andreas Witt and Dieter Metzing. Dordrecht, Netherlands: Springer.

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 (1): 45–60.

Farrar, Scott, William D. Lewis, and D. Terence Langendoen. 2002. "An Ontology for Linguistic Annotation." *Semantic Web Meets Language Resources: Papers from the AAAI Workshop* (Technical Report WS-02–16), 11–19. Menlo Park, CA: AAAI Press.

Hjelmslev, Louis. 1962. *Prolegomena to a Theory of Language*, 2d rev. Translation by Frances J. Whitfield. Madison: University of Wisconsin Press. Originally published in 1943 as *Omkring Sprogteoriens Grundlæggelse*. Copenhagen: Munksgaard; reprinted 1966. Copenhagen: Akademisk Forlag.

Langendoen, D. Terence, Scott Farrar, and William D. Lewis. 2002. "Bridging the Markup Gap: Smart Search Engines for Language Researchers." *Proceedings of the Workshop on Resources and Tools for Field Linguistics*, *Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands.

Langendoen, D. Terence, and Gary F. Simons. 1995. "A Rationale for the Text Encoding Initiative Recommendations for Feature-Structure Markup." *Computers and the Humanities* 29:191–205. Reprinted in *The Text Encoding Initiative: Background and Context*, edited by Nancy Ide and Jean Veronis, 191–210. Dordrecht, Netherlands: Kluwer.

Lewis, William D. 2006. "ODIN: A Model for Adapting and Enriching Legacy Infrastructure." *Proceedings of the E-Humanities Workshop Held in Cooperation with E-Science 2006: 2nd IEEE International Conference on E-Science and Grid Computing.* Amsterdam.

Lewis, William D., D. Terence Langendoen, and Scott Farrar. 2001. "Building a Knowledge Base of Morphosyntactic Terminology." *Proceedings of the IRCS Workshop on Linguistic Databases,* 150–156. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania.

Lewis, William D., and Fei Xia. 2010. "Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages." *Journal of Literary and Linguistic Computing* 25 (3): 303–319.

Pease, Adam. 2007. "Formal Representation of Concepts: The Suggested Upper Merged Ontology and Its Use in Linguistics." In *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts*, edited by Andrea Schalley and Dietmar Zaefferer, 103–114. Berlin: Mouton de Gruyter.

Pease, Adam, and Ian Niles. 2002. "Towards a Standard Upper Ontology: A Progress Report." *Knowledge Engineering Review* 17 (1): 65–70.

Schalley, Andrea, and Dietmar Zaefferer, eds. 2007. *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts*. Berlin: Mouton de Gruyter.

Simons, Gary F., Brian Fitzsimons, D. Terence Langendoen, William D. Lewis, Scott Farrar, Alexis Lanham, Ruby Basham, et al. 2004a. "A Model for Interoperability: XML Documents as a Distributed Database." *E-MELD Workshop on Databases for Field Linguistics*, Detroit.

Simons, Gary F., William D. Lewis, Scott Farrar, D. Terence Langendoen, Brian Fitzsimons, and Hector Gonzalez. 2004b. "The Semantics of Markup: Mapping Legacy Markup Schemas to a Common Semantics." *Proceedings of the 4th Workshop on NLP and XML* (*NLPXML-2004*), 25–32. Association for Computational Linguistics.

Sperberg-McQueen, Michael, and Lou Burnard, eds. 1994. *Guidelines for Electronic Text Encoding and Interchange* (*TEI P3*). Chicago: The Association for Computers in the Humanities (ACH), the Association for Computational Linguistics (ACL), and the Association for Linguistic and Literary Computing (ALLC).

Xia, Fei, William D. Lewis, Michael W. Goodman, Joshua Crowgey, and Emily M. Bender. 2014. "Enriching ODIN." *Proceedings of the 9th International Conference on Language Resources and Evaluation,* 3151–3157. Reykjavik, Iceland.

Xia, Fei, William D. Lewis, Michael W. Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey, and Emily M. Bender. 2016. "Enriching a Massively Multilingual Database of Interlinear Glossed Text." *Language Resources and Evaluation* 50:321–349.

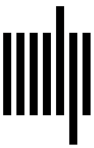
Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences

Edited by: Antonio Pareja-Lora, María Blume, Barbara C. Lust, Christian Chiarcos

Citation:

Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences
Edited by: Antonio Pareja-Lora, María Blume, Barbara C. Lust, Christian Chiarcos
DOI: 10.7551/mitpress/10990.001.0001
ISBN (electronic): 9780262357210
Publisher: The MIT Press
Published: 2020

The open access edition of this book was made possible by generous funding and support from Knowledge Unlatched



The MIT Press

The Open Access edition of this book was published with generous support from the National Science Foundation (grant number BCS-1463196), Pontificia Universidad Católica del Perú, and Knowledge Unlatched.

This book was set in Times New Roman by Westchester Publishing Services. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Pareja-Lora, Antonio, editor. | Blume, María, editor. | Lust, Barbara C., 1941– editor. | Chiarcos, Christian, editor.
Title: Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences / edited by Antonio Pareja-Lora, María Blume, Barbara C. Lust, and Christian Chiarcos.
Description: Cambridge : MIT Press, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2019019588 | ISBN 9780262536257 (paperback)
Subjects: LCSH: Language and languages--Study and teaching. | Language and languages--Research. | Linked data.
Classification: LCC P53 .D398 2019 | DDC 025.06/4--dc23
LC record available at https://lccn.loc.gov/2019019588

10 9 8 7 6 5 4 3 2 1