Deconstructing descriptive grammars

Jeff Good
University at Buffalo

Much work within digital linguistics has focused on the problem of developing concrete methods and general principles for encoding data structures designed for non-digital media into digital formats. This work has been successful enough that the field is now in a position to move past "retrofitting" digital solutions onto analog structures and to consider how new technologies should actually change linguistic practice. The domain of grammaticography is looked at from this perspective, and a traditional descriptive grammar is reconceptualized as a database of linked data, in principle curated from distinct sources. Among the consequences of such a reconceptualization is the potential loss of two valued features of traditional descriptive grammars, here termed *coverage* and *coherence*. The nature of these features is examined in order to determine how they can be integrated into a linked data model of digital descriptive grammars, thereby allowing us to benefit from new technology without losing important features intrinsic to the structure of the traditional version of the resource.

1. FROM RECODING TO RECONCEPTUALIZING. The field of linguistics is now well aware of the need to use new digital technologies to encode linguistic data with care in order to ensure its portability across user communities, computational environments, and even time (Bird & Simons 2003).[1] This has resulted in a range of work examining the best means through which traditional linguistic resources can be re-encoded in digital form. Proposals have been made, for instance, that offer conceptual models and accompanying digital implementations for lexical resources (Bell & Bird 2000; Poornima & Good 2010), interlinear glossed text (Bow et al. 2003; Schroeter & Thieberger 2006; Palmer & Erk 2007), grammatical paradigms (Penton et al. 2004), and descriptive grammars (Good 2004). Other work has gone beyond this to codify general principles for conducting this kind of research, as seen, for instance, in Nordhoff's (2008) examination of possible ideal requirements for electronic grammars and in the "meta-model" approach to lexical encoding embodied by Lexical Markup Framework, developed in the context of work on natural language processing (Francopoulo et al. 2009) (see also Wittenburg et al. (2002) and Trippel (2006; 2009)).

This work has produced important results, and, in particular, has made clear that, even if important kinds of linguistic data may still await proper study, the

[1] I would like to thank audience members at the Colloquium on Electronic Grammaticography, held at the Second International Conference on Language Documentation and Conservation at the University of Hawai'i at Mānoa, February 11–13, 2011, for their comments on the work leading to this paper. Many of the ideas developed here have been influenced by informal discussions with a number of individuals during the last several years, in particular Michael Cysouw and Sebastian Nordhoff.

challenges of encoding them linguistically are presumably solvable using existing technologies, in particular generalized markup systems like XML (see Good (2011: 225–227) for discussion of XML in the context of language documentation). Furthermore, it is even possible to extract from this body of research an informal general procedure for devising new encoding schemes: (i) survey existing practice for presenting data in a given domain, (ii) devise a conceptual model of the data that can be understood as providing an underlying form for the surveyed presentation (or "surface") forms, (iii) relate the various components of that model to linguistic practice, focusing, in particular, on how they can be derived from more general principles regarding what constitutes appropriate methodology for linguistic analysis, and (iv) propose a concrete way of encoding that model using archival markup formats that is as consistent with those general methodological principles as possible. While I am not aware of any one publication that incorporates the totality of these procedures, the combination of Good (2004) and Nordhoff (2008) can be understood as an illustration of this approach in the domain of descriptive grammars.

At the same time, the sensible reliance of such work on *existing* practice makes it ill-suited for considering the ways in which new technologies should prompt more fundamental reconceptualizations of the kinds of products that the field of linguistics should produce as a result of technological changes. This is illustrated, for instance, by Palmer & Erk's (2007) revision to Bow et al.'s (2003) proposals for encoding interlinear glossed text. The latter is sufficient for dealing with interlinear glossed text's traditional function of providing a succinct analysis of the lexical and grammatical content of a given stretch of phrasal data. However, it is not well-suited for an additional function that is clearly desirable (though non-traditional): automated (or semi-automated) annotation.

The main goal of this paper is to consider how new models for data encoding—developed independently from the field of linguistics—might prompt us to consider revisions to our models for producing traditional grammatical descriptions. In particular, detailed consideration will be given regarding how the work required to describe a language's grammar on the basis of documentary products might be done in a highly distributed fashion by making use of emerging Web technologies (sections 2 and 3). However, as we will see, there is a danger when considering such a possibility: The introduction of a new conceptualization of a linguistic resource, with clear positive features, may inadvertently lead to the loss of valued features embedded within traditional models. This requires the development of means to reintegrate what has been "lost" into the new conceptualization, which can be done by augmenting received technologies with solutions more specific to linguistics (sections 4 and 5). However, the solutions discussed here only allow us to retain part of what would be lost, underscoring that the transition from traditional products to digital ones must be led by linguists' needs rather than the non-linguistic agendas that drive the development of most new technologies (section 6).

This paper is primarily conceptual in nature, rather than reporting on the results of a specific technological implementation. Accordingly, at times, it will be somewhat speculative, though an attempt will be made, whenever possible, to ground any speculation in relevant existing technological efforts, often from within linguistics itself. The intended audience for this paper are so-called ordinary working linguists, rather than those more directly engaged in applying emerging technologies to linguistic work. At the same time, I hope that some of its key ideas will be of interest to both groups. Those familiar with work on the technologies to be discussed will be aware of the fact that some of the points made here are relatively well-known outside of linguistics. Therefore, in some places, the aim of the discussion will not be to outline the significance of these new technologies generally but, rather, to present them in a way that makes their utility clearer to a documentary and descriptive linguistic audience—that is, to linguists who are now, or may in the future, be creating new grammatical descriptions.

2. THE TECHNOLOGICAL CONTEXT. While some data encoding technologies (e.g., XML) have become so ubiquitous in work on language documentation and description as to require little introduction, the technological context which inspires the present work—that of the so-called Semantic Web—has not yet been particularly widely employed within linguistics. The leading idea behind the development of the Semantic Web is to extend the document-centered World Wide Web to all kinds of data, adding, in effect, an explicit layer of meaning to a network architecture that was originally designed to simply link together pages intended to be interpreted by humans.[2]

Rather than consisting of a monolithic piece of technology, the Semantic Web is better understood as resulting from the interactions of a set of logically-independent technological pieces. Some of these are relatively simple in and of themselves but, nevertheless, provide crucial elements of the infrastructure needed to augment the World Wide Web with semantic information. Moreover, since the Semantic Web is intended to build on the World Wide Web, many of its core technologies are the same as those of the World Wide Web itself, as seen, for example, in its use of Unicode for character encoding (see, e.g., Anderson (2003) and Gippert (2006: 345–351) for discussion of Unicode in a linguistic context). At the same time, in order to extend the World Wide Web, there are also technologies that underlie the Semantic Web that are not part of the World Wide Web. The most prominent of these within work on language documentation and description is almost certainly Web Ontology Language (OWL), which has been used to express the linguistic information encoded in the General Ontology for Linguistic Description (GOLD) (Farrar

[2] The Semantic Web Frequently Asked Questions page (http://www.w3.org/2001/sw/SW-FAQ) produced by the World Wide Web Consortium serves as useful introduction to the Semantic Web.

& Langendoen 2003; Farrar & Lewis 2007; Farrar & Langendoen 2009).[3] GOLD will be returned to shortly below.

The Semantic Web technologies that will play a significant role in the discussion here are given in table 1. (The first of these, the URI, is also a World Wide Web technology and likely to be familiar to many readers.) A brief summary of the relevant function that each takes on is included in the table, though this description should not be understood to be exhaustive in a general Semantic Web context. Both in the table, and elsewhere, the presentation of the Semantic Web and relevant component technologies is simplified somewhat for the purposes of exposition.

ACRONYM	TECHNOLOGY	FUNCTION
URI	Uniform Resource Identifier	Unique identification of entities
RDF	Resource Description Framework	Description of entities
OWL	Web Ontology Language	Encoding of general facts

Table 1: Some Semantic Web Technologies

Taken together the three technologies listed in table 1 allow for (i) the unique reference to anything in the real or mental world (e.g., a language, an utterance, or a phoneme) in the form of URIs, (ii) the specification of "facts" about those entities (e.g., *English is a language* or *this sentence makes use of a passive construction*) in the form of RDF expressions, and (iii) the specification of the overall properties of the conceptual model that the entities and facts are embedded within (e.g., *sentences have grammatical properties* or *lexical items are comprised of a combination of sound, meaning, and grammatical properties*) in the form of OWL statements. While the ability to specify such information falls short of the rich content of a traditional descriptive grammar, it should be clear that even this relatively minimal apparatus could allow for the production of a significant amount of description of a given language's grammar. In particular, it would permit many of the "low-level" facts that are part of any complete grammatical description to be encoded in a machine-readable form on the Web.

It would be inappropriate to cover the full technical details of URIs, RDF, or OWL here. However, it will be useful if something can be said about what these technologies "look like" in concrete terms, especially since none of them is a prototypical instance of a "technology". In Semantic Web applications, URIs look more or less like the familiar Uniform Resource Locators (URLs) associated with web pages, taking on a form like http://example.org/English. URLs can, in fact, be understood as a type of URI that both identifies a given entity (e.g., a web page or a file) and also specifies how the entity can be found on the World Wide Web. While Semantic Web URIs look like URLs, and may even behave like URLs in resolv-

[3] The latest version of GOLD can be viewed at http://linguistics-ontology.org/.

ing to a web page when accessed on a browser, strictly speaking they need not be URLs. In concrete terms, this would mean that if a URI, which was not also a URL, were entered into a browser, it would not resolve to any web page (and produce, for example, an error page).

The idea that a URI may look like URL, but not act like one, is potentially counterintuitive to those accustomed to working with the World Wide Web rather than the Semantic Web. However, it is a reasonable consequence of attempting to build an online repository of information on top of the existing Web rather than in a completely new environment. In this case, the standard Web mechanism for uniquely referring to a given web page is simply extended for uniquely referring to *anything*, whether or not it happens to be associated with a web page. Of course, alternatives are possible. For instance, the Handle System provides another mechanism for creating unique identifiers (see Broeder et al. (2006) for discussion of the use of the Handle System in developing linguistic infrastructure).[4]

An important feature of URIs is that they are not merely unique within some local system, as might be the case, for instance, for identification numbers of the sort commonly associated with records in a database. Rather, they are universally unique in the context of the Web, at least when used as intended. That is, in Semantic Web terms, anything that needs to be referred to gets a completely unique identifier. The vast proliferation of identifiers that this entails may, at first, sound problematic. However, one must bear in mind that the World Wide Web has grown vastly since its first inception without breaking down, illustrating the durability of URIs as a mechanism for providing globally unique references.

Turning now to the second technology listed in table 1, Resource Description Framework (RDF) is a means for making machine-readable three-part statements, or *triples*, of the form SUBJECT PREDICATE OBJECT. The subject is a URI for some entity, the predicate is a URI for a possible relationship between a subject and an object, and the object consists of either of a URI or a limited class other objects, such as a text string. Figure 1 gives an example of a representation of two RDF triples. The first (Triple 1) is intended to relate a subject URI referring to the English language to an object URI referring to the phoneme *p* via a predicate that states that that phoneme is found in English (i.e., "English has the phoneme *p*"). The second (Triple 2) relates the phoneme *p*, now serving as a subject of a triple, to its standard transcription, the text string "p", serving as the object of the triple (i.e, "The phoneme *p* has the transcription 'p"').

RDF has not yet been widely deployed for describing traditional linguistic data,

[4] The Handle System is used by this journal as a means of persistent identification of published papers. For example, the handle for Newman (2007) is 10125/1724. This handle can be resolved, using an appropriate online service, to a web page where a copy of the paper can be found. One way to do this is to simply append the handle to http://hdl.handle.net/, producing the URL http://hdl.handle.net/10125/1724.

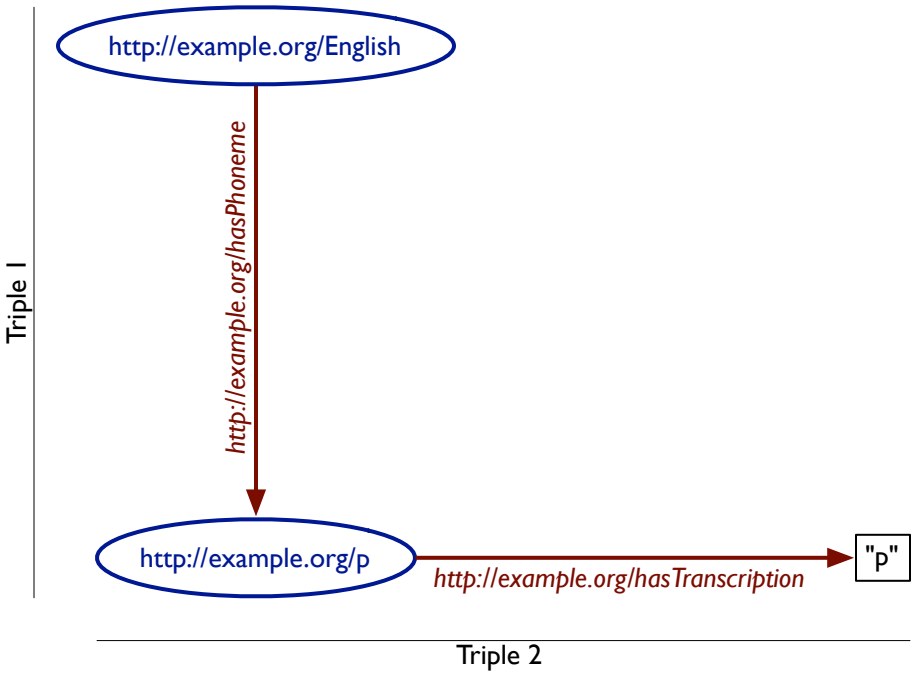

Figure 1: Two RDF triples

though there have been some attempts, both as exemplary cases (Simons 2005) and in working systems (Good & Hendryx-Parker 2006) (see also Cysouw (2007: 63–65)).[5] It has also seen attention in research in computational linguistics (see, e.g., Ide et al. (2003)).

The information encoding model of RDF is, in principle, expressible in a variety of formats, including a standardized XML format, which facilitates exchange of data described in RDF. More generally, RDF can be understood as a means for encoding data that can usefully be modeled in the form of a graph consisting of nodes connected by labeled arcs, as is seen in figure 1.[6] Of course, RDF is not the only way of describing data in graph form, though it is the focus here because of its role as a key means of expressing "atomic" statements about entities in the context of the Semantic Web.[7]

[5] The websites for the World Atlas of Language Structures (WALS) (Dryer & Haspelmath 2011) and the World Loanword Database (Haspelmath & Tadmor 2009) also allow some of their data to be exported in RDF format, though in the case of WALS, this is limited to bibliographical data.

[6] The fact that the connections between nodes in RDF representations can be labeled or "typed" makes them richer than the connections found in typical hypertext which merely links documents (or parts of documents) without specifying the semantic nature of those links. Therefore, while "hypertext grammars" (Evans & Dench 2006: 29) would clearly represent an advance over traditional grammars, they would fall short of the possibilities afforded by the Semantic Web.

[7] Another common way of expressing graphs with labeled arcs in linguistic work is through the use

An important feature of graphs (whether or not they are expressed in RDF) is that, as discussed by Ide & Suderman (2007), they facilitate the merging of information from distinct sources. As long as two related sets of data expressed in graph form use the same identifiers for nodes referring to the same entities, the two graphs can simply be joined wherever nodes are shared. If we consider figure 1, for instance, one data source could state that English makes use of the phoneme *p*, while another could associate *p* with a transcription, with the two being joined by their common node, http://example.org/p. While this is a relatively trivial example, graph merger of this kind can become quite powerful when the merged graphs each contain rich and largely complementary information (see section 3.1).

URIs and RDF, when brought together, are key pieces to the idea of creating significant amounts of *Linked Data*, which is seen as a crucial step towards the broader vision of the Semantic Web (Bizer et al. 2009: 15) and provides a useful metaphor for understanding the goals of the Semantic Web more generally.[8]

The final technology listed in table 1 is Web Ontology Language (OWL), which allows for the expression of *ontologies*. In this context an ontology can be understood as a means for expressing general knowledge about a given domain in a form that can be understood by machines. This might include statements like, *a past tense is a kind of tense* or *a phoneme inventory is comprised of phonemes*. Basic statements like these could, in fact, be stated using RDF which is flexible enough to encode both very specific statements as well as general ones. OWL, however, provides a means for expressing certain kinds of generalizations that are not standardly expressible in RDF. In fact, OWL can itself be viewed as an augmentation of RDF in much the same way as the Semantic Web augments the World Wide Web. To pick one example, OWL provides a standard way of stating that one property is the "inverse" of another property. This would allow, for instance, a machine to infer that, if *the phoneme* p *has the transcription 'p'*, then *the symbol 'p' is a possible transcription of the phoneme* p. As discussed above, there has already been significant work on an OWL-based ontology in the context of descriptive linguistics in the form of the GOLD project (Farrar & Langendoen 2003; Farrar & Lewis 2007; Farrar & Langendoen 2009), and there has also been work using OWL to support the mobilization of descriptive language materials (Beck et al. 2007).

Before moving on, it is important to bear in mind that URIs, RDF, and OWL are merely specific technical solutions to more general problems. Their significance in the present context is the way that they are integrated into the larger vision of the Semantic Web. Of course, the Semantic Web, too, is a specific technical solution to the broad problem of how information can be shared and exchanged efficiently.

---

of feature structures as found, for example, in Head-driven Phrase Structure Grammar (Sag et al. 2003: 50–51).

[8] A workshop on Linked Data in Linguistics was held on March 7-9, 2012, in Frankfurt, Germany (see http://ldl2012.lod2.eu/).

Its relevance for the field of linguistics is twofold. First, it offers a model for a new way of managing research results in an increasingly internet-driven data management world. Second, it is a specific instantiation of such a model with considerable support outside of linguistics, allowing linguists to take advantage of technological infrastructure that has already been developed elsewhere.

The next section of this paper will consider how an initiative like the Semantic Web might prompt us to reconsider what it means to create a descriptive grammar of a language.

3. MULTIPLE FACETS OF GRAMMARS.

3.1. DISENTANGLING PUBLICATION. In this section, I will largely abstract away from the technical details delineated in section 2 and, instead, focus on how the model embodied by the conjunction of the technologies in table 1 could be exploited to create new methods for producing grammatical descriptions. In principle, a number of the ideas developed here could have been put forth decades ago. After all, many of the key concepts embodied by the Semantic Web (e.g., unique identification) are hardly new. However, in practice, before the rise of the World Wide Web, work along such lines would have been largely impossible to apply concretely to documentary and descriptive research. We have now reached a point, by contrast, where the development of models that, at one point, would have been merely speculative or "futuristic" can actually be implemented in specific tools, at least in prototype form. This makes it important for the field to begin to consider the relevant issues proactively in order to avoid accidentally adopting technological solutions that might, at first, appear to be appropriate but which may actually be built on assumptions that will prove problematic in the long run. (See Boynton et al. (2010: 134–138) for relevant discussion in a language documentation context.)

A striking feature of the model that the Semantic Web offers is the extent to which it leads to a view of scholarly work in general, and grammaticography more specifically, wherein a number of elements that were previously intertwined due to the restrictions of paper publication can now be decoupled from one another. Based on ideas found in Neylon (2010), we can break down the functions of traditional "monolithic" scholarly publications along the lines of what is described in (1).

(1) a. **Registering:** Scholarly publishing allows an individual or set of individuals to officially establish that they should be associated with a given set of ideas or research results.
b. **Filtering:** Scholarly publishing provides mechanisms through which a given user can locate the information they are interested in both by providing a means through which the quality of a given piece of research can be quickly assessed and by providing tools for discovery (e.g., through the use of keywords).

c. **Curation:** Scholarly publishing works with carefully aggregated sets of data that are brought together to tell a specific "story".
d. **Archiving:** Scholarly publishing produces resources designed to be usable in the long-term.
e. **Reusing:** Scholarly publishing is associated with standardized mechanisms through which research results can be reused in a manner deemed acceptable by the research community (e.g., by providing a stable citation for a given resource).

Of the five functions of publishing given in (1), the one of greatest interest here—and the one which would would seem to be most profoundly impacted by the technologies discussed in section 2—is almost certainly (1c), the curatorial function. The traditional model of a descriptive grammar as a kind of monograph encourages us to see the thousands of tiny observations that form a complete description as part of a single research "outcome". The graph-based model of the Semantic Web, by contrast, explicitly makes each of these observations visible as distinctive connections among discrete objects. This was already schematized in figure 1, with a relatively simplistic example. Figure 2 offers something comparable with a more complex example that abstracts away from some of the more technical aspects of RDF.

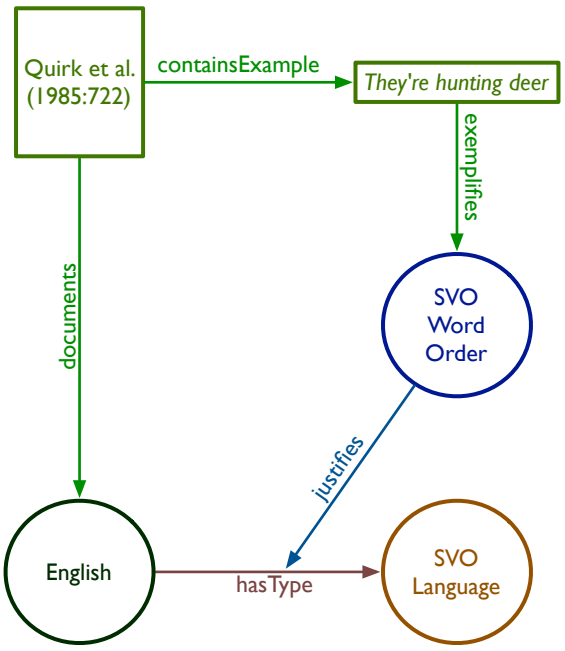

Figure 2: A fragment of a descriptive grammar in graph form

Figure 2 represents a set of low-level statements which, when combined, allow one to make a claim like *English is an SVO language*. At the top of the figure, an

example is indicated as being extracted from a source that documents the English language. This example is observed to show SVO word order, which, in turn, justifies the general classification of English as an SVO language. In RDF terms, this would mean breaking down the classification of English as SVO into five distinct statements (or triples). There are obvious elements of simplification involved in the figure, though it should be sufficient for purposes of exemplification.

By way of further illustration, figure 3 represents further information about one of the nodes in figure 2, the one representing the English language. This figure gives reference information for English, specifically a language code and its genealogical parent. It can be understood here as representing information about the same entity as described in figure 2, but coming from a different source.

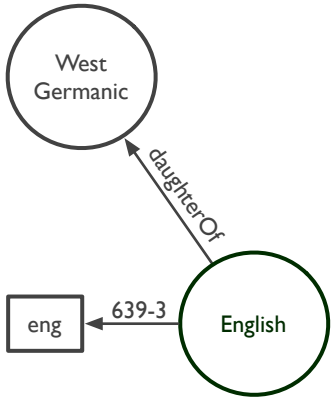

Figure 3: Classificatory information for English in graph form

Figure 4 illustrates one of the positive features of graph-based representations of information (see also section 2): The fact that they allow information from different sources to be straightforwardly merged as long as common node identifiers are employed. In figure 4, the content of figures 2 and 3 are brought together into a single graph.

Of course, there is nothing particularly innovative about combining related pieces of information from distinct sources. The power of graph-based representations like the one seen in figure 4 is the way in which they allow this process to be, at least partly, automated and the way in which they make visible the nature of the connections between data sources in a more precise way than is possible with standard academic citations.

This latter point is of interest here when we consider the inherently "distributed" task of writing a descriptive grammar—even if it is only written by a single author. They are typically the distillation of a number of years, or even decades, of work on a language and untold numbers of small observations, preliminary analyses, reanalyses, etc. (see Weber (2006: 417–418) for relevant discussion). Moreover, the

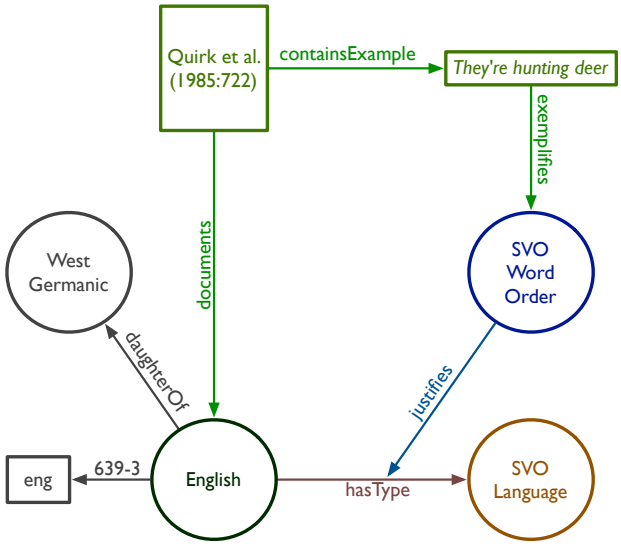

Figure 4: Merging two graphs

distributed nature of the work involved in grammar writing has become significantly more pronounced in recent years with the rise of the documentary paradigm. This has stressed methodological and theoretical separation between research outputs that can be classified as "documentary" from those that are "descriptive" (Himmelmann 1998; Woodbury 2011: 168–169). The next section will consider then what a "graph-based" approach to data might mean for grammar writing, especially in the context of the newly placed emphasis on documentation. A more concrete way of looking at this issue would be to ask: How would grammar creation be different if crafting each of the component statements of a classification like the one represented in figure 2 was the responsibility of different linguists?

Before moving on, it seems worth emphasizing that many of the issues to be addressed below—for example, ensuring that a grammar has adequate coverage or that its analyses are coherent—existed long before the development of digital approaches to research. What has changed is that, now that research can, in principle, be done in a much more distributed fashion, the utility of general solutions to problems like these has become more apparent. In the creation of single-authored grammars, the main control on coherence, for example, has been the authors themselves. But clearly this concern must be approached differently if one wants to make use of the time and the skills of ten, or even a hundred, contributors working on the description of a single language. Moreover, as we will see in the next section, there has been impetus for grammatical descriptions to be developed in more distributed fashion that has arisen completely independently from the growth of the Semantic Web itself.

3.2. TOWARDS DISTRIBUTED GRAMMAR AUTHORING. The extent to which the activities comprising documentation and description should be viewed as easily severable has been questioned (see, e.g, Evans (2008: 346–348)). Nevertheless, it seems uncontroversial that the possibilities that new technologies offer for creating documentary products should have a significant impact on the creation of grammatical descriptions (Evans & Dench 2006: 24–25) (see also Good (2010: 120–122)). Moreover, the documentary paradigm has emphasized the need for a more collaborative approach to the collection and analysis of language data, integrating members of speaker communities, as well as experts from allied disciplines, more directly into the linguist's research activities (Himmelmann 2006: 15–16; Dwyer 2006: 54–55; Grenoble 2010: 293–399; Woodbury 2011: 176–177).

While not always emanating from the same fundamental concerns, both of these ideas share a comparable impact when it comes to research which has as one of its goals the production of a descriptive grammar. On the one hand, new data collection and annotation technologies have led to an emphasis on ensuring that the provenance of a descriptive claim can be straightforwardly verified by associating it with the relevant supporting documentary materials. Ideally, these should be in the form of fully transcribed texts as well as audio and video records (see, e.g., Bird & Simons (2003: 571); Nordhoff (2008: 299); Thieberger (2009)). This requires tools and methods for making the "documentary chain" from recording to analysis explicit, which amounts to creating a new set of intermediate linguistic resources comprising each of the relevant links in that chain. A prominent example of this is time-aligned annotation, which connects a transcription directly to the recording containing what is being transcribed. This has resulted in the widespread use of a relatively new kind of linguistic resource which encodes documentary and descriptive annotations (see Schultze-Berndt (2006)) directly with an indication of start and end times within a media file that those annotations can be associated with. This is found, for example, in resources produced by the ELAN annotation tool (see Berez (2007) for a review). What we see in this case is that the task of annotation, which formerly was disseminated primarily as embedded within finished products, can now be associated with an "intermediate" resource reflecting an important aspect of the underlying work.[9]

On the other hand, collaborative models for collecting and analyzing linguistic data cause us to shift perspective from an approach where a single individual is responsible for all stages of grammatical analysis to one where the various stages of the documentary and descriptive workflow might be the primary responsibility of different contributors (see, e.g., Thieberger (2004) and Bowern (2011: 461–462) for discussions of workflow). This adds an additional element of "decomposition" to the traditional way of working. In large part, new data management and com-

[9] Of course, "intermediate" objects like this could be created without the aid of new digital technologies and some of them can be found in archives of field notes. However, before the rise of the internet, they were not typically made widely available or as carefully curated.

munication technologies are a prerequisite to the practical application of such collaborative models. However, their ultimate motivation is largely social in nature and emanates from changes in the conception of what constitutes ethical and appropriate research practices (see, e.g., Rice (2006a: 124–134) and Dobrin & Berson (2011: 201–206)). They can also be understood as a response to language endangerment, insofar as the impending loss of a language is understood as a loss not only to linguistics, but to speaker communities and other disciplines as well. This has, thereby, caused linguists to seriously consider the need for approaches incorporating a diverse array of stakeholders in the collection and analysis of language data. Here, then, we see how a set of changes in practice, driven by social considerations, can at least be partly supported by new technologies—in this case, technologies which facilitate work being done in distributed fashion.

Taken together, these two trends place increasing emphasis on the individual "pieces" of work involved in the creation of descriptive grammars, as opposed to treating them as a monolithic whole—the view encouraged by the traditional publication model. Moreover, independent of developments within work on language documentation itself, a more distributed approach to work forming the basis of descriptive grammars, in principle, has an additional potential advantage: It can facilitate more efficient use of research resources. An individual who is skilled at transcription may not be adept at morphological analysis, and a specialist in semantics may not be the ideal person to work on a language's phonemic system—and this is not to mention the problems that may arise when a linguist is asked to be not merely a grammarian but also an archivist, ethnographer, a lexicographer, or even a "linguistic social worker" (Newman 1998: 14–17) (see also Evans (2008: 342–343)).

We arrive then, at a potential future where we move away from the publication-centered view of a descriptive grammar as a single-authored monograph to one where it consists of the compilation of set of "facts" about a language, each associated with a distinct provenance. This view of a "grammar" not only has clear connections to the current documentary paradigm, but it is also consonant with the reconceptualization of data embodied by the Semantic Web: Research results are "atomized" as it were, the barriers to registering a result (in the sense of (1a)) are significantly reduced, and the connections between discrete results can be made more explicit. Of course, it is a long road from the representation of a relatively simple observation like that seen in figure 2 to the construction of a graph representing a "complete" descriptive grammar. Nevertheless, we now have a core conceptual model of such a structure, technology that can implement that model, and even a field-internal motivation to use such a model. This means the creation of such an object is no longer simply something only to be imagined.

Within this broad vision, making the results of research public no longer needs to be delayed until it reaches the threshold of a "publishable unit" (see Broad (1981)) but can be done as soon as a useful observation is made, even if it consti-

tutes something as simple as the discovery of a new minimal pair or a single unusual pattern of agreement. Of course, such "micro-discoveries" would not be associated with the same level of prestige as a curated publication.[10] What is important is that the Semantic Web, in principle, can allow them to be associated with the elements of publishing appropriate to them, e.g., registration, archiving, and reuse (see (1)), even if they fall short of the whole traditional publication "package".

There are clear potential advantages to reconceptualizing descriptive grammars as distributed, multi-authored resources. However, it is immediately apparent that valued features of the traditional descriptive grammar would be lost under such an approach unless additional measures are taken. In particular, the curatorial aspect of publishing (see (1c)) imposes various important characteristics on the assembled "facts" which constitute descriptive grammars, two of which I will focus on here, *coverage* (section 4) and *coherence* (section 5). Of course, these are only parts of what constitutes a "good" grammar (see, e.g., Noonan (2006); Rice (2006b)), an issue which will be briefly returned to in section 6.2.

4. COVERAGE.

4.1. COEXTENSIVITY AND COMPLETENESS. An important aspect of good descriptive grammars is the extent to which their discussion (i) adequately addresses phenomena represented in the available documentation on a language and (ii) presents a reasonable overview of a language's entire grammatical system.[11] I refer to these properties under the umbrella term *coverage* here, using the term *coextensivity* for the relationship between a description and available documentation and *completeness*, to refer to the extent to which a given description addresses those issues that are taken, at the time of publication, to be sufficiently central to basic linguistic theory (see Dryer (2006)) that they would be deemed necessary in a "complete" description of a language.

Completeness has seen more explicit attention than coextensivity, perhaps most famously in the form of Comrie & Smith's (1977) descriptive questionnaire. This can not only be used as a set of guidelines for ensuring that a grammar has covered a wide range of grammatical topics deemed descriptively significant (Noonan 2006: 360) but has even formed the basis of full grammatical descriptions (such as

[10] The idea of publishing a "micro-discovery" can be clearly connected to the notion of micro-blogging, most prominently associated at present with the online service Twitter (http://twitter.com), which has been the subject of work considering how the content of micro-blogs can be made available not just on the World Wide Web but within the Semantic Web as well (Passant et al. 2010). See also Cysouw (2007: 64) for relevant discussion on the notion of a "micro-publication" within work on language typology.

[11] By *available documentation*, I refer to the extent of the documentary resources (e.g., recordings, transcribed texts, lexical material, etc.) that those working on the description of a given language have access to when doing their work.

Huttar & Huttar (1994)). An important aspect of completeness is that determining just what constitutes something like a "complete" description is within the purview of the general community of linguists rather than those working on a particular language.

The notion of coextensivity is instead connected to the actual documentation available in the production of a descriptive grammar. Therefore, a grammatical description of a language for which relatively little documentation exists may be considered to have satisfactory coextensivity even if its level of completeness is unambiguously inadequate. This would be the case, for instance, if the only material on a language that was available was a vocabulary list which might allow for the production of a phonological sketch, but little else. As such, it is important to distinguish between what is referred as coextensivity here and what one might call *documentary coverage*. This latter notion might be used to characterize the extent to which a documentary corpus actually includes the information needed to create a complete grammar of a language (see Berge (2010) for some discussion), regardless as to whether or not a descriptive grammar has actually been created on the basis of that record. A key distinction between coextensivity and completeness is that what constitutes completeness in a grammatical description can be laid out in general terms. Adequate coextensivity, by contrast, is linked to the particularities of the available documentation.

Coextensivity has seen not seem as much attention as completeness presumably because, before the development of the current documentary paradigm, it was difficult to evaluate due to the inaccessibility of the documentary materials on which descriptions were based. Even if a given descriptive grammar made clear, for instance, what percentage of available recordings or texts had been used in creating it, an inability to examine those texts would have made it essentially impossible for a reader to gauge the adequacy of its coextensivity. However, to the extent that the documentary bases of descriptions are expected to become more widely disseminated, explicit attention to adequacy in coextensivity would seem warranted. As pointed out by Evans & Dench (2006: 25), new technologies are unlikely to result in significantly more analyzed materials than was previously possible. After all, the time it takes the linguist to conduct careful analysis will not change in proportion to the amount of material that can be made available. This means that it is likely to be the case, at least for the foreseeable future, that grammars will only be based on a sample of collected materials. Therefore, the extent to which the studied sample may be representative of the language as a whole will be a significant concern.

An immediate problem with adopting a distributed approach to the production of electronic grammars is that the model in and of itself does not allow us to gauge the extent to which a given set of statements about a language's grammar is adequate with respect to coextensivity and completeness. Properly addressing such dimensions of coverage has normally been the responsibility of the single author of a traditional descriptive grammar. Dealing with them in a distributed context re-

quires us to consider how we can augment Semantic Web (or similar) technologies with data processing and analysis methods more specific to the domain of language data. This is the topic of the next two sections, which, in turn, discuss possible digital approaches to completeness (section 4.2) and coextensivity (section 4.3).

4.2. MODELING COMPLETENESS. In understanding how to ensure adequate completeness in a descriptive grammar, it is first important to keep clear the fact that just what constitutes a "good" description in terms of completeness is not a technological problem but a scientific one, and it is driven by what is taken to constitute basic linguistic theory (Dryer 2006) at a given point in time. The key issue here is not, therefore, deciding what phenomena need to be included in a "complete" description, but, rather, how to digitally represent completeness in a way that facilitates creating grammars with a high degree of completeness and also allows us to automatically (or semi-automatically) evaluate the level of completeness attained by some digital descriptive grammar.

Of the problems to be discussed here, dealing with completeness is probably the easiest since there are already reasonable models to work from, in particular in work interested in typologically-oriented language comparison. Zaefferer (2006), for example, describes a system for creating a cross-linguistic reference grammar database which attempts to balance the need to ensure that each language is adequately described in its own terms against allowing comparable features across languages to be compared. In Semantic Web terms, this approach could be generalized first by formulating a pre-determined list of elements for cross-linguistic comparison expressed in the form of an accessible digital object. Then, the completeness of a digital descriptive grammar could be (at least partly) gauged by the extent to which each member of that list of elements is or is not associated with a specific RDF statement relating them to the other observations comprising a digital descriptive grammar.

Some work has suggested that the relevant elements of comparison should preferentially be drawn from the onomasiological domain rather than the semasiological one (Zaefferer 2006: 122, Cristofaro 2006: 140–142). However, it seems likely that a full consensus specification of completeness will need to include specification of of both functional and formal features. Comrie & Smith (1977: 28), for example, contains a question regarding, in general terms, what kinds of morphological elements are used to encode the syntactic or semantic functions of nouns, regardless of the specific functions of those elements. Similarly categories like *head-marking* and *dependent-marking*, though having a functional component, primarily target formal variation, but have nevertheless been the subject of significant typological investigation (see, e.g., Nichols (1992)).

In any event, while an up-to-date specification of ideal completeness is lacking, this does not appear to be the result of particular technological impediments but, rather, a lack of social effort. If there were sufficient interest, an initial proposal

could probably be developed with relatively little work by examining and selectively merging the topics of a work like Comrie & Smith (1977) with more up-to-date surveys of specific typological phenomena, where available. In particular, the collected grammatical domains found in Dryer & Haspelmath (2011) would serve as a good recent "snapshot" of the typological state-of-the-art already available in electronic form (see also Levin et al. (2007: 262–266)).[12]

The specific digital form of an object expressing the components of completeness could straightforwardly be based on the familiar notion of a questionnaire, where each question would be associated with a unique identifier. These questions would be "answered" via an RDF triple linking the topic of the question to supporting descriptive materials, where relevant via an intermediary node that would specify a categorial response to the issue raised by the question. The resulting series of "links" between questions and answers could then function very much like an index in a traditional descriptive grammar, though with the additional expressive power and utility afforded by digital technology (see also Zaefferer (2006: 115) and Cristofaro (2006: 162)). Though not dealing with the creation of full-fledged descriptive grammars, relevant work has been done in the domain of typological database construction. In particular, models have been proposed which seek to clearly separate the problem of isolating language-specific data illustrating the presence or absence of a feature from classifying a language (or construction) on the basis of that data into one of a fixed number of "types" (Bickel & Nichols 2002; Cysouw 2007).

Figure 5 augments figure 4 to schematize the integration of an element associated with completeness with the analysis of a particular language. The notion of *Basic clausal word order* is treated as an element that is part of completeness and is introduced into the overall descriptive graph of English by being associated with the earlier treatment of English as an SVO language.

Other elements of completeness could be associated with comparable nodes to what is seen in figure 5, though the relationships between the element of completeness and the descriptive analysis will, of course, not always be as simple. For instance, if a language showed split word order, this would require a more elaborate specification in the graph, just as it requires a more elaborate description in a traditional descriptive grammar. Similarly, if an element of completeness was associated with a phenomenon not attested in the language being described, conventions could be adopted to explicitly indicate its absence, comparable to what is found in the index of Haspelmath (1993) (see Good (2004: §2.1)). Furthermore, just as in a work like Comrie & Smith (1977), where grammatical questions are arranged in a hierarchy, a full scheme for completeness need not be composed simply of a "bag" of questions, but could be specified with additional structure and information—either

[12] There is at least one instance of a widely disseminated language description tool, Fieldworks Language Explorer, which incorporates a kind of grammar model in its design that can facilitate achieving completeness (see Butler & van Volkinburg (2007); Rogers (2010) for reviews).

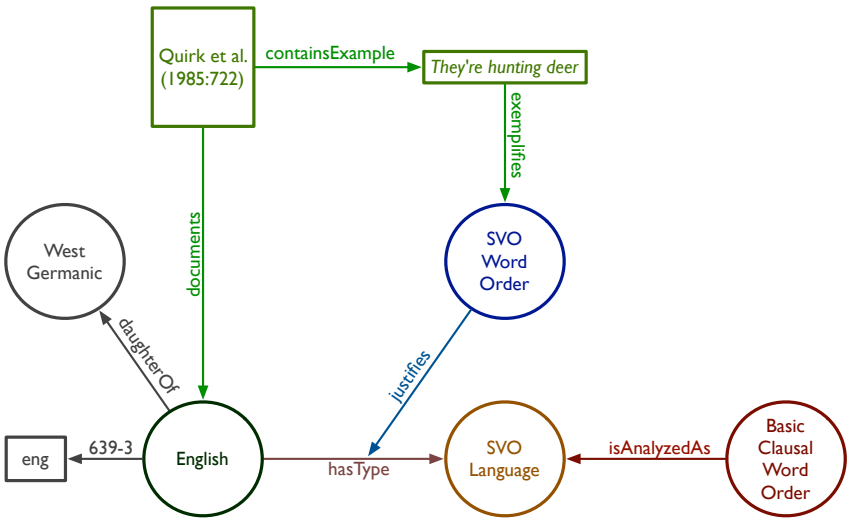

Figure 5: Adding completeness to a graph-based description

using RDF or some other means.

Ultimately, the problem of understanding completeness can be understood as a kind of data modeling, and better completeness would amount to "filling out" more of the elements of an accepted model. For this particular aspect of descriptive grammars, consensus on the shape of the model itself—whether digitally encoded or not—appears to be the most difficult issue, with the technical problems being relatively attenuated.

4.3. MODELING COEXTENSIVITY. Unlike completeness, which must clearly be connected to some community-wide consensus of what a description should include, coextensivity is particularized to the documentary record of a language. If the available documentary information is quite limited, then it might be expected that a description will be based (more or less) on all of the available data. However, for languages subject to even a moderate level of investigation, this will often simply not be possible. Rather, the general goal of the description is not that it be based on a detailed examination of all of the data. Instead, it should be based on a sample of the data that results in a description that is representative of all of the collected data.[13]

As mentioned above, there does not appear to have been significant work on how to measure adequacy of coextensivity. Nichols (2005), however, considers how much material is needed to produce adequate descriptions characterized in terms of numbers of words, clauses, and hours, and this would serve as a good starting point for work on this topic. Ultimately, these recommendations are probably best consid-

[13] As will be discussed in section 5, there are clear connections between coextensivity and aspects of coherence in descriptive grammars.

ered to be connected to adequacy in documentation rather than coverage. However, they might serve as proxies for coextensivity: For instance, her calculation that about 100,000 running words will allow for "basic documentation" suggests that a descriptive grammar based on a 100,000-word sampling of a much larger corpus is likely to be sufficient to allow for a reasonable level of analysis of the remaining part of the corpus by a future investigator, though there will probably still be significant gaps.

Bird (2010) (see also Reiman (2010)) describes a documentation model aimed at under-resourced languages involving the collection of oral texts, accompanied by oral annotation of a fraction of those texts, and a further written annotation for a fraction of the orally-annotated texts. While the clear intention of this process is to provide sufficient annotation (in oral and written form) of a language to allow the unanalyzed portions to be analyzed in the future (Bird 2010: 7), without further research, it seems impossible to know whether any description creatable on the basis of such a degree of annotated documentation would actually be sufficient for analyzing the entirety of the collected materials without the aid of a native speaker. In any event, the question of what kind of sample of documentary materials can be considered representative enough to form the basis of a description that would also cover the remaining materials appears to be an interesting one, and work in this area would be quite useful for developing general methods for assessing adequacy in coextensivity.

For transcribed documentary materials, there is at least the possibility of using a more direct means to assess coextensivity of a description. If a description can be expressed in a machine-readable form, automated methods could be employed to apply that description to the entire available dataset (see also section 5.2.3 for related discussion).[14] The coextensivity of the machine-readable description could then be considered to be adequate if it can provide appropriate parses for unanalyzed material. Of course, while some analytical domains, such as morphological parsing, have seen significant work in terms of relating traditional documentation to machine-readable parsers (see, e.g., Black & Simons (2008)), most domains of grammar are not yet well covered. Moreover, while there is at least some relevant work in the domain of syntax (see section 5.2.4), there seems to be little to no work along these lines in the domain of phonology (e.g., allowing a user to define a phoneme inventory and phonotactic constraints and checking to see if transcriptions are consistent with them).[15]

While completeness appears to be representable via data modeling, as discussed

[14] In principle, these methods could be applied to untranscribed materials as well if they could somehow be associated with a reasonable, automatically-derived transcription using techniques from work on speech recognition. But, of course, that adds a significant, additional element of complexity.

[15] There are, however, tools that allow one to discover things like phonotactic constraints on the basis of existing transcriptions, for example Phonology Assistant (see Dingemanse (2008) for a review).

in section 4.2, this general solution does not appear to apply to coextensivity. The goal of creating a grammatical description on the basis of documentation is, in some sense, to discover the data model of a language in the first place. Rather, a more appropriate model for coextensivity would appear to be one based on notions from natural language processing related to the ability of a computational system, based on an examination of a subset of available data, to assign analyses to data that the system has not been "trained" on (see Resnik & Lin (2010) for overview discussion).[16] Implementing this kind of approach generally for digital descriptive grammars is likely to be quite difficult, however. Creating the relevant kinds of computational systems, even for well-described languages, is, at this point, still quite time-consuming and requires resources well-beyond those usually available to those engaged in language description (though, see sections 5.2.3 and 5.2.4 for discussion of different lines of research attempting to address this issue). Moreover, existing methods are often heavily reliant on the availability of large quantities of textual materials, which are simply unavailable for most languages (see also Bird (2011)).

5. COHERENCE.

5.1. THE COMPONENTS OF COHERENCE. A second clear problem with the possibility of taking a distributed approach to the development of descriptive grammars is that, without specific attention, they run the risk of becoming incoherent due to distinct conventions and analyses adopted by different researchers working on pieces of the documentation or description. Maintaining coherence is, of course, a problem for all kinds of research, and my goal here will not be to develop the notion generally, but, rather, to try to specifically model the components of coherence in descriptive grammars. While the components to be discussed, listed in (2), may not exhaust all of what is required for coherence in descriptive grammars, they appear to represent at least four prominent ones.

(2) a. **Consistency in terminology for language-specific categories**: The use of terms for language-specific categories is ideally consistently applied throughout.
b. **Clarity in terminology for the general audience:** The relationships between language-specific categories and comparative concepts (in the sense of Haspelmath (2010)) are ideally made explicit.
c. **Consonance of analyses with documentation:** The descriptive analyses should ideally be in agreement with what is found in the entire documentary record.

[16] Abney & Bird (2010); Bird (2011) offer parallel proposals in suggesting that one metric for determining whether available documentation is adequate for capturing the properties of a language is that it is sufficient for training a machine translation system.

d. **Compatibility of analyses with each other:** The analyses of specific grammatical patterns are ideally compatible with each other throughout the description.

As indicated, the components in (2) represent ideals, and even single-authored grammars will fail to adhere to them fully. Nevertheless, they suggest points to pay special attention to if grammatical analysis is to be distributed since coherence is likely to be an especially problematic area in this regard.

There are clear connections between certain aspects of coverage and certain aspects of coherence, at least when these two concepts are understood informally. This is most clearly seen in relation to coextensivity (see section 4.1) and the components of coherence given in (2c) and (2d). Whether or not the coextensivity of a description would be considered adequate clearly hinges on the extent to which it is consonant with the documentation and the extent to which all of its analyses can be brought together into a non-contradictory whole. At the same time, it seems reasonable to separate these notions. Coextensivity is intended to reflect the link between available documentation and what level of description is possible given that documentation, while the components of coherence are more general than this. Nevertheless, practically speaking, as will be indicated in sections 5.2.3 and 5.2.4, an important consequence of the connection between coextensivity and those two components of coherence is that they may require overlapping technological support.

In the next section, I will discuss how the components of coherence in (2) could be modeled digitally and discuss existing technologies that would be relevant for the implementation of those models, thereby complementing the discussion in section 4 and suggesting additional ways in which the Semantic Web vision might be augmented to facilitate the creation of digital descriptive grammars in a distributed fashion. Many of the points to be discussed below should resonate even for those working on traditional grammatical monographs, but, again, the idea that the work might be done by many individuals, rather than just one, brings the relevant issues to the fore. Under such an approach, it will no longer be obvious who is in charge of the "quality control" necessary to achieve coherence, which necessarily prompts us to consider how we might develop means of ensuring that it is maintained that do not depend on the presence of a central "author".

5.2. MODELING COHERENCE.

5.2.1. CONSISTENCY. In discussing *consistency* for the terminology used in a grammatical description, I refer only to consistency for the terms used to describe the categories found in the language in question. This dimension of coherence, therefore, is not intended to apply to the use of terms for general linguistic concepts, of the sort contained within the GOLD ontology (see section 5.2.2). The distinction between language-specific categories and general linguistic ones is not always

well-maintained within descriptive grammars, though it is found, for instance, in Haspelmath (1993: 11) (on the basis of a practice employed in Comrie (1976: 13)). In that grammar, capitalized terms are used for language-specific categories and lower-case terms for general linguistic notions (see Good (2004: §2.1)).

Ultimately, the issue of maintaining consistency in a description can, in large part, be understood as a problem of terminology management (or *terminography*), which is a distinctive area of research in its own right (see, e.g., Wright & Budin (1997: 1–3) and Cabré (1999: 115–159)). Terminology management has some overlap with lexicography. However, it is primarily oriented with relating concepts to forms, rather than forms to concepts, as is typical of lexicography (Cabré 1999: 7–8) (thus making it more comparable to the onomasiological rather than the semasiological approach to descriptive grammars).

Descriptive linguistics generally already involves a fair amount of informal terminology management (as evidenced, for instance, by ubiquitous glossing abbreviation lists and efforts like Bickel et al. (2008)). Therefore, even if we did not adopt a distributed approach to the writing of descriptive grammars, the field could clearly benefit from more robust (and ideally partially automated) techniques for managing the terms used to describe a given language. Furthermore, once one considers the possibilities for more distributed authorship, such techniques would seem to become a necessity in order to facilitate harmonization of terms across content contributors, whether these are in Semantic Web form from the start or partial Semantic Web annotation is attempted for legacy resources. Work of the latter sort could specifically build on existing research in the area of (semi-)automated term extraction (see Ahmad & Rogers (2001)).

On the whole, technological support for the consistent use of terminology within a grammatical description appears to be largely underdeveloped. However, it seems like a potentially profitable area in which to focus efforts in the near term. This is due to the possibility to make use of existing work on terminology management in general, as well as on terminological support for the component of descriptive coherence termed clarity here (see (2b)). This latter area of research will be discussed in the next section, and it can likely serve as a useful model for the development of tools facilitating consistency in the use of language-specific terminology, as will be briefly discussed below.

5.2.2. CLARITY. A well-known problem of linguistic description is the use of the same word to refer to different grammatical concepts or the use of different words to refer to the same concept across descriptions (see, e.g., Cysouw et al. (2005: §1) and Zaefferer (2006: 114)). In either case the potential for confusion is clear, and the issues are especially acute for work attempting to automate, or partly automate, language comparison on the basis of digital materials.

This has been one of the key motivations behind the development of the GOLD ontology (Farrar & Langendoen 2003; Farrar & Lewis 2007; Farrar & Langendoen

2009), already discussed in section 2, which represents the longest-running effort for the exploitation of Semantic Web technologies for use in descriptive linguistics. GOLD provides a set of standardized concepts relevant to grammatical description that can be used to formally define a term used in a language description in general linguistic terms. For instance, it would allow for the specification that the English Past Tense verb form expresses a meaning that can be reasonably related to the general notion of *past tense* specified in GOLD. Of course, in this case, the terminology is not particular problematic. However, when dealing with a form like the Latin Perfect, which would be generally characterizable as a combination of *perfective* and *past*, rather than a *perfect* (Comrie 1976: 13), being able to relate a language-specific term to more general categories with readily-accessible definitions, as GOLD allows for, is clearly valuable.

GOLD both provides a standardized termset (with associated URIs for each term) and structures the members of the termset into an ontology (consisting of a taxonomy plus some additional information) in order to facilitate automated processing of linguistic data. Such an ontology is not a strict requirement to achieve clarity in use of terminology, and a somewhat simpler model is provided by ISOcat (Kemps-Snijders et al. 2008a;b).[17] ISOcat provides an open registry for data categories relevant to linguistic resources, allowing an individual linguist or groups of linguists to publicly register the terms they use and associate them with basic descriptions of the meaning of the terms. It also provides a unique identifier for each term which can be used in a Semantic Web context.[18]

ISOcat has a somewhat more "open" model than GOLD, insofar as it allows different groups to register their own categories. By contrast, the addition of new categories to GOLD is more centrally managed. At the same time, GOLD's community model explicitly allows for different subcommunities of linguists to extend the ontology to suit their specific needs (see Farrar & Lewis (2007: 53–55)). Taken together, an open registration system like ISOcat in conjunction with tools making it straightforward to associate an ISOcat category with the appropriate GOLD concepts could provide a significant degree of support for some of the issues relating to consistency in the use of terminology discussed in section 5.2.1.

There does not yet appear to be significant use of resources like GOLD or ISOcat to enhance traditional descriptive work. However, due to the efforts that have been expended on their development, terminological clarity can probably be considered, at present, the best supported component of coherence discussed here.

[17] http://www.isocat.org/

[18] While ISOcat's structure does not allow for the specification of an ontology for its categories, there has been work attempting to develop a degree of ontological structure around the ISOcat categories in a parallel resource (Wright et al. 2010).

5.2.3. CONSONANCE. The development of the documentary paradigm has altered expectations regarding the extent to which the data on which a description is based should be accessible. This brings to the fore an issue with respect to coherence which was always present but was often of little practical significance: To what extent is a given description, which will often be based on a detailed examination of only a subset of documentary materials, in agreement with the entire body of available documentation for a language? Section 4.3 discussed related issues from the point of view of ensuring adequate coextensivity.

Because the documentary expectations that have made this problem a practical concern are relatively new, this issue does not appear to have received significant attention. The focus, up to the present, has instead been on developing methods through which a given descriptive claim can be verified on the basis of supporting documentation in a relatively richly annotated corpus (see, e.g., Thieberger (2009)). But, in the long run, it would also be ideal if it were possible for sparsely annotated documentation to be automatically processed on the basis of existing connections between documentation and description in order to locate possible cases of discord between the documentation and the description. This processing could involve such things as, for example, the detection of phonological processes that fail to apply as expected, the discovery of unaccounted for members of a morphological paradigm, or flagging uses of a discourse marker that do not match its description.

These are, of course, difficult tasks to the extent that they require the development of sophisticated parsers based on machine-readable grammars—the descriptive grammatical equivalent of debugging software (see also section 4.3). One potentially promising relevant line of work in this regard involves automated processing techniques that make use of manual annotation of a fragment of a corpus combined with *active learning* in which the user provides feedback to an automated system to help improve its performance. This has been applied to the domain of interlinear glossed text (Baldridge & Palmer 2009; Palmer et al. 2009; Palmer 2009; Palmer et al. 2010) and could, in principle, be applied to other domains as well, reducing the effort required to create useful parsing tools for a given language. However, the path from experimenting with these methods to providing robust tools for checking the complicated relationships involved in consonance between documentary products (e.g., transcribed texts) and descriptive ones (e.g., a complete grammar based on those texts) is likely to be a relatively long one. (A different line of research involving machine-readable grammars, more relevant to the notion of compatibility, but with potential applications to consonance as well, will be discussed below in section 5.2.4.)

An open question in research along these lines is where the most effective "balancing point" might be for manual versus automatic annotation and whether or not even relatively simple types of annotation might facilitate the use of automated methods in ways that make more effective use of an expert's time. For instance, it will often be the case that documentary recordings will contain stretches of different

languages, most typically a language of wider communication and a language being documented. Annotating which stretch is in which language may be able to be done by someone without special linguistic expertise on a subset of the recordings to train a machine to do the work across the whole documentary corpus. This would give an expert linguist a head-start on more complex kinds of annotation by making it easier for them to locate the most important stretches of recorded data. The maintenance of consonance in the relationship between documentary products and derivative descriptive ones would, thereby, be facilitated. Such a scenario suggests the possibility that taking full advantage of a more distributed approach to grammatical description may require us to consider crafting non-prototypical documentary products—in this case a very sparse kind of linguistic annotation.[19]

5.2.4. COMPATIBILITY. The final dimension of coherence to be discussed here, compatibility, has already seen some attention in the computational linguistics literature in the area of grammar engineering, which seeks to create machine-readable versions of formal grammars (see, e.g., Bender (2008b)). Among other things, these grammars allow linguists to automatically test the extent to which their analysis of a given phenomenon interacts with other analyses as expected, something which is more or less impossible to do by hand.[20] Moreover, there have been efforts to ensure that grammar engineering work done for better-resourced languages can be used to facilitate the development of machine-readable grammars for lesser-resourced languages (Bender et al. 2010), addressing an issue raised at the end of section 4.3 in the discussion of coextensivity. Such tools can even potentially play a role in helping linguists choose among competing descriptive hypotheses (Bender 2010).

Bender (2008a) reports the results of work which made use of a general grammar engineering system to create a machine-readable grammar for a language typologically quite distinct from those languages that have informed most computational research. Strikingly, it was possible to create a reasonable new machine-readable grammar for this language in a timeframe of about six weeks. This is, of course, only a small fraction of the time it takes to write a traditional descriptive grammar. It suggests that work in grammar engineering has reached a point where relatively limited collaborations between linguists specializing in this area and those working on underdescribed languages may yield worthwhile results for language description with respect to ensuring compatibility among analyses. More gener-

[19] This scenario also recalls recent work, such as Snow et al. (2008), which suggests that internet marketplaces where workers can be recruited to perform relatively simple annotation tasks at much lower rates than individuals with linguistic training may be useful for language research. However, see Fort et al. (2011) for discussion of ethical complications potentially associated with using such marketplaces, related to the limited rights and wages typically granted to workers.

[20] Grammar engineering can also play a role in maintaining consonance (see section 5.2.3), insofar as it can help locate instances of data that cannot be analyzed at all by a given formal grammar, suggesting gaps in a description.

ally, this work provides a model for how to move forward in the development of computational techniques to facilitate the creation of digital descriptive grammars: Lessons learned from the development of computational tools for better-resourced languages, often at great cost, can ultimately be applied to lesser-resourced languages at much lower cost.

The work described above is primarily focused on syntactic phenomena. Comparable tools exist for some aspects of morphophonological analysis (see, e.g., Black & Simons (2008)), though there has not yet been much work on integrating tools from each of these domains (though see Bender & Good (2008) for some discussion of possibilities). Furthermore, other domains of grammar do not appear to be well-supported yet at all, with phonology standing out as an area where the lack of obvious tools does not seem to present major technical problems but, rather, results from a lack of dedicated effort. For example, a tool allowing a linguist to specify phonotactic constraints and ensure their overall compatibility would appear to be much more straightforward to develop than the already existing tools supporting syntactic analysis mentioned above. Nevertheless, to the best of my knowledge no such tool exists. (See also section 4.3.) In other domains, like semantics and pragmatics, the tool gap is less surprising given the difficulties of conducting even informal descriptive work in these areas.

While the discussion here has been primarily in terms of issues regarding parsing language data, some computational systems are designed to be bidirectional, both parsing and generating language data (see Bender & Langendoen (2010: 6) for discussion relevant in the present context). While I am not aware of specific proposals in this regard, the ability of such systems to generate data has potential applications for testing compatibility (as well as consonance) of grammatical analyses adopting onomasiological approaches. This would require a machine-readable means to describe the relevant "functions" which would serve as an input for the generation of predicted forms in a given language. The generated forms could be compared against attested forms to gauge the extent to which they match each other.

6. CONCLUSION.

6.1. BUILDING ON EXISTING INFRASTRUCTURE. Overall, we have seen above that, if we take the distributed model of research implied by Semantic Web technologies seriously, we are presented with the problem of losing desirable characteristics of traditional resources not embedded within the design of the Semantic Web itself. However, with appropriate conceptual models of key components of a traditional resource, we can devise explicit characterizations of what we have lost which allow us to see how we can augment the Semantic Web (or any comparable endeavor) in ways that reincorporate the "missing" features into our new kinds of resources.

In focusing on coverage and coherence, the overall vision that emanates from

this discussion is one where Semantic Web technologies would form a basic infrastructure to express statements in ways that make them straightforwardly registrable and reusable, but where the set of statements comprising a grammatical description would be subject to additional data processing and validation. This would involve techniques ranging from the development of formal data models (to help verify completeness), to the deployment of tools for terminology management (to help ensure consistency), to the use of machine-readable formal grammars (to help test for compatibility).

At the same time, significant areas have been identified where tool support or appropriate models of practice are lacking, though a path to developing those tools or models can be identified. This was seen, for instance, in the domain of coextensivity, where there has been relatively little research on determining what an appropriate documentary "sample" might look like. It was also seen in the domain of consonance where tools for verifying that nowhere in the documentation is a description contradicted have not seen serious attention on either the conceptual or implementation side. In both cases, methods from natural language processing were put forth as presenting possible solutions to these problems.

Nevertheless, a significant result, I believe, of this survey is the extent to which many existing technologies provide models (if not necessarily "off-the-shelf" tools) for how we might go about developing a tool "ecology" (see Good (2007)) for distributed grammatical description with less effort than might appear to be needed at first. Importantly, even if a given tool type requires re-implementation to be usable in a documentary and descriptive context, modeling a new tool on an existing one is likely to save considerable time and resources, when set against developing it completely anew.

6.2. DECONSTRUCTING *OUR* PROBLEMS. To conclude, this has paper has been intended to be an exercise in what one might call "theoretical" electronic grammaticography (though grounding the discussion in specific relevant technologies). While its starting point (see section 2) was a technical discussion of the Semantic Web, ultimately its specific technical details were of less importance than the model that it provided for a more distributed approach to data dissemination and curation, which has clear potential applications for many areas of scholarship, but especially descriptive grammars. This is due to their complexity in terms of the relationship of descriptive claims to documentation, the breadth of their subject matter, and the interconnectedness of the elements of description.

While the discussion here has been framed as one where the distribution of the work of describing a language is dispersed across multiple contributors, the long-term nature of most descriptive work also means that it is distributed across time, even if there is only a single main creator. Because of this, many of the models and techniques described here would certainly also be of value in cases where effort is expended primarily by one person, but over a long enough period that they may find

it difficult to keep track of their own success in terms of coverage and coherence.

Two particular issues were the focus here, coverage and coherence. These are undoubtedly important aspects of traditional descriptive grammars. However, it should be emphasized that they are far from the whole story. For instance, one of the criteria listed by Rice (2006b: 396) as an aspect to writing an effective grammar is "richness of illustration", a clearly important concern not considered here at all.

Moreover, the framework introduced here does not allow for the expression, in any straightforward way, of the idea that languages have a basic "plan" or structural "genius", to borrow from the famous formulation of (Sapir 1921: 127). This idea has been reflected in the intuitions of both formal and descriptive linguists (see, e.g., Baker (1996: 6–9) and Evans & Dench (2006: 3–4)). However, it is difficult to imagine how it could be expressed in the deliberately reductionist framework of the Semantic Web, making it clear that we should not understand a reconceptualization process like the one offered here as a means of replacing our traditional understanding of what makes a descriptive grammar "good". Rather, it should be seen as an exercise in understanding how we can use technology to enhance what we already know to be good (see also Dobrin et al. (2009: 42–43)).

Ultimately, the goal of a study like this one is not to set our agenda on the basis of what a given technology offers but, rather, to clarify which existing technologies can fulfill our needs and to map out plans for the creation of new technologies. The Semantic Web may have prompted consideration of many of the ideas discussed here, but it cannot serve as a substitute for crafting a vision for the future of linguistic resources with our own values serving as its foundation.

REFERENCES

Abney, Steven & Steven Bird. 2010. The human language project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 88–97. Stroudsburg, Penn.: Association for Computational Linguistics.

Ahmad, Khurshid & Margaret Rogers. 2001. Corpus linguistics and terminology extraction. In Sue Ellen Wright & Gerhardt Budin (eds.), *Handbook of terminology management, volume 2: Application-oriented terminology management*, 725–760. Amsterdam: Benjamins.

Anderson, Deborah. 2003. Using the Unicode standard for linguistic data: Preliminary guidelines. In *Proceedings of E-MELD 2003: Digitizing and annotating texts and field recordings*. East Lansing, Michigan, July 11–13. http://www.e-meld.org/workshop/2003/anderson-paper.pdf.

Baker, Mark C. 1996. *The Polysynthesis Parameter*. Oxford: Oxford University.

Baldridge, Jason & Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, 296–305. Stroudsburg, Penn.: Association for Computational Linguistics. http://www.aclweb.org/anthology-new/D/D09/D09-1031.

Beck, Howard, Sue Legg, Elizabeth Lowe & M. J. Hardman. 2007. Aymara on the internet: A step toward interoperability and user access. In Peter K. Austin, Oliver Bond & David Nathan (eds.), *Proceedings of Conference on Language Documentation and Linguistic Theory*, 29–38. London: SOAS.

Bell, John & Steven Bird. 2000. A preliminary study of the structure of lexicon entries. In *Proceedings from the Workshop on Web-Based Language Documentation and Description*, Philadelphia, December 12–15, 2000. http://www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html.

Bender, Emily M. 2008a. Evaluating a crosslinguistic grammar resource: A case study of Wambaya. In *Proceedings of ACL-08: HLT*, 977–985. Stroudsburg, Penn.: Association for Computational Linguistics. http://www.aclweb.org/anthology/P/P08/P08-1111.

Bender, Emily M. 2008b. Grammar engineering for linguistic hypothesis testing. In Nicholas Gaylord, Stephen Hilderbrand, Heeyoung Lyu, Alexis Palmer & Elias Ponvert (eds.), *Proceedings of the Texas Linguistics Society X Conference: Computational linguistics for less-studied languages*, 16–36. Stanford: CSLI.

Bender, Emily M. 2010. Reweaving a grammar for Wambaya. *Linguistic Issues in Language Technology* 3. http://elanguage.net/journals/index.php/lilt/article/view/662.

Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson & Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation* 8. 23–72.

Bender, Emily M. & Jeff Good. 2008. Implementation for discovery: A bipartite lexicon to support morphological and syntactic analysis. In Rodney L. Edwards, Patrick J. Midtlyng, Colin L. Sprague & Kjersti K. Stensrud (eds.), *Proceedings of the Chicago Linguistic Society 41: The panels*, 1–15. Chicago: Chicago Linguistic Society.

Bender, Emily M. & D. Terence Langendoen. 2010. Computational linguistics in support of linguistic theory. *Linguistic Issues in Language Technology* 3. http://www.elanguage.net/journals/index.php/lilt/article/view/661.

Berez, Andrea. 2007. Technology review: EUDICO Linguistic Annotator (ELAN). *Language Documentation & Conservation* 1. 283–289. http://hdl.handle.net/10125/1718.

Berge, Anna. 2010. Adequacy in documentation. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language documentation: Practice and values*, 51–66. Amsterdam: Benjamins.

Bickel, Balthasar, Bernard Comrie & Martin Haspelmath. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme by morpheme glosses. http://www.eva.mpg.de/lingua/pdf/LGR08.02.05.pdf.

Bickel, Balthasar & Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the International Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26–27 May 2002*, Nijmegen: ISLE and DOBES. http://www.mpi.nl/lrec/2002/papers/lrec-pap-20-BickelNichols.pdf.

Bird, Steven. 2010. A scalable method for preserving oral literature from small languages. In Gobinda Chowdhury, Chris Khoo & Jane Hunter (eds.), *The role of digital libraries in a time of global change: 12th International Conference on Asia-Pacific Digital Libraries (ICADL 2010)*, 5–14. Berlin: Springer.

Bird, Steven. 2011. Bootstrapping the language archive. *Linguistic Issues in Language Technology* 6. http://elanguage.net/journals/index.php/lilt/article/view/2580.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documention and description. *Language* 79. 557–582.

Bizer, Christian, Tom Heath & Tim Berners-Lee. 2009. Linked data—The story so far. *International Journal on Semantic Web and Information Systems* 5. 1–22.

Black, Andrew H. & Gary F. Simons. 2008. The SIL FieldWorks Language Explorer approach to morphological parsing. In Nicholas Gaylord, Stephen Hilderbrand, Heeyoung Lyu, Alexis Palmer & Elias Ponvert (eds.), *Proceedings of the Texas Linguistics Society X Conference: Computational linguistics for less-studied languages*, 37–55. Stanford: CSLI.

Bow, Catherine, Baden Hughes & Steven Bird. 2003. Towards a general model for interlinear text. In *Proceedings of E-MELD 2003: Digitizing and annotating texts and field recordings*. East Lansing, Michigan, July 11–13. http://e-meld.org/workshop/2003/bowbadenbird-paper.html.

Bowern, Claire. 2011. Planning a language documentation project. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 459–482. Cambridge: Cambridge University.

Boynton, Jessica, Steven Moran, Helen Aristar-Dry & Anthony Aristar. 2010. Using the E-MELD School of Best Practices to create lasting digital documentation. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language documentation: Practice and values*, 133–146. Amsterdam: Benjamins.

Broad, William J. 1981. The publishing game: Getting more for less. *Science* 211. 1137–1139.

Broeder, Daan, Remco van Veenendaal, David Nathan & Sven Strömqvist. 2006. A grid of language resource repositories. In *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)*. Los Alamitos, California: IEEE Computer Society. doi:10.1109/E-SCIENCE.2006.261065.

Butler, Lynnika & Heather van Volkinburg. 2007. Review of Fieldworks Language Explorer (FLEx). *Language Documentation & Conservation* 1. 100–106. http://hdl.handle.net/10125/1730.

Cabré, M. Teresa. 1999. *Terminology: Theory, methods, and applications*. Amsterdam: Benjamins.

Comrie, Bernard. 1976. *Aspect*. Cambridge: Cambridge University.

Comrie, Bernard & Norval Smith. 1977. Lingua descriptive studies: Questionnaire. *Lingua* 42. 1–72.

Cristofaro, Sonia. 2006. The organization of reference grammars: A typologist user's point of view. In Felix Ameka, Alan Dench & Nicholas Evans (eds.), *Catching language: The standing challenge of grammar writing*, 137–170. Berlin: Mouton de Gruyter.

Cysouw, Michael. 2007. A social layer for typological databases. In Andrea Sansò (ed.), *Language resources and linguistic theory*, 59–66. Milano: FrancoAngeli.

Cysouw, Michael, Jeff Good, Mihai Albu & Hans-Jörg Bibiko. 2005. Can GOLD "cope" with WALS? Retrofitting an ontology onto the World Atlas of Language Structures. In *Proceedings of E-MELD 2005: Linguistic ontologies and data categories for language resources*. Cambridge, Mass., July 1–3. http://emeld.org/workshop/2005/papers/good-paper.pdf.

Dingemanse, Mark. 2008. Review of Phonology Assistant 3.0.1. *Language Documentation & Conservation* 2. 325–331. http://hdl.handle.net/10125/4350.

Dobrin, Lise M., Peter K. Austin & David Nathan. 2009. Dying to be counted: The commodification of endangered languages in documentary linguistics. In Peter K. Austin (ed.), *Language documentation and description, volume 6*, 37–52. London: Hans Rausing Endangered Languages Project.

Dobrin, Lise M. & Josh Berson. 2011. Speakers and language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 188–211. Cambridge: Cambridge University.

Dryer, Matthew S. 2006. Descriptive theories, explanatory theories, and basic linguistic theory. In Felix Ameka, Alan Dench & Nicholas Evans (eds.), *Catching language: The standing challenge of grammar writing*, 207–234. Berlin: Mouton de Gruyter.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2011. *The World Atlas of Language Structures online*. Munich: Max Planck Digital Library. http://wals.info/.

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 31–66. Berlin: Mouton de Gruyter.

Evans, Nicholas. 2008. Review of *Essentials of language documentation* ed. by Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel. *Language Documentation & Conservation* 2. 340–350.

Evans, Nicholas & Alan Dench. 2006. Introduction: Catching language. In Felix Ameka, Alan Dench & Nicholas Evans (eds.), *Catching language: The standing challenge of grammar writing*, 1–39. Berlin: Mouton de Gruyter.

Farrar, Scott & D. Terence Langendoen. 2003. A linguistic ontology for the semantic web. *Glot International* 7. 97–100.

Farrar, Scott & D. Terence Langendoen. 2009. An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In Andreas Witt & Dieter Metzing (eds.), *Linguistic modeling of information and markup languages: Contributions to language technology*, 45–66. Berlin: Springer.

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

Fort, Karën, Gilles Adda & K. Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics* 37. 413–420.

Francopoulo, Gil, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet & Claudia Soria. 2009. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation* 43. 57–70.

Gippert, Jost. 2006. Linguistic documentation and the encoding of textual materials. In Jost Gippert, Nikolaus Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 337–361. Berlin: Mouton de Gruyter.

Good, Jeff. 2004. The descriptive grammar as a (meta)database. In *Proceedings of E-MELD 2004: Linguistic databases and best practice*. Detroit, Michigan. July 15–18. Http://www.e-meld.org/workshop/2004/jcgood-paper.html.

Good, Jeff. 2007. The ecology of documentary and descriptive linguistics. In Peter K. Austin (ed.), *Language documentation and description, volume 4*, 38–57. London: Hans Rausing Endangered Languages Project.

Good, Jeff. 2010. Valuing technology: Finding the linguist's place in a new technological universe. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Valuing technology: Finding the linguist's place in a new technological universe*, 111–131. Amsterdam: Benjamins.

Good, Jeff. 2011. Data and language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 212–234. Cambridge: Cambridge University.

Good, Jeff & Calvin Hendryx-Parker. 2006. Modeling contested categorization in linguistic databases. In *Proceedings of E-MELD Workshop 2006: Tools and standards: The state of the art*, Lansing, Michigan. June 20–22. http://e-meld.org/workshop/2006/papers/GoodHendryxParker-Modelling.pdf.

Grenoble, Lenore A. 2010. Language documentation and field linguistics: The state of the field. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language documentation and field linguistics: The state of the field*, 289–309. Amsterdam: Benjamins.

Haspelmath, Martin. 1993. *A grammar of Lezgian*. Berlin: Mouton.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86. 663–687.

Haspelmath, Martin & Uri Tadmor (eds.). 2009. *World loanword database*. Munich: Max Planck Digital Library. http://wold.livingsources.org/.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin: Mouton de Gruyter.

Huttar, George L. & Mary L. Huttar. 1994. *Ndyuka*. London: Routledge.

Ide, Nancy, Alessandro Lenci & Nicoletta Calzolari. 2003. RDF instantiation of ISLE/MILE lexical entries. In *Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the model right (LingAnnot '03)*, 30–37. Stroudsburg, Penn.: Association for Computational Linguistics. doi:10.3115/1119296.1119301.

Ide, Nancy & Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the linguistic annotation workshop (LAW '07)*, 1–8. Stroudsburg, Penn.: Association for Computational Linguistics. http://portal.acm.org/citation.cfm?id=1642059.1642060.

Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg & Sue Ellen Wright. 2008a. ISOcat: Corralling data categories in the wild. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Paris: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2008/pdf/222_paper.pdf.

Kemps-Snijders, Marc, Menzo Windhouwer & Sue Ellen Wright. 2008b. Putting data categories in their semantic context. In *Proceedings of e-Humanities—an emerging discipline: Workshop in the 4th IEEE International Conference on e-Science*, Indianapolis, Indiana, December 10. http://www.clarin.eu/system/files/e-Humanities-ISOcat-final.pdf.

Levin, Lori, Jeff Good, Alison Alvarez & Robert Frederking. 2007. Automatic learning of grammatical encoding. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling & Chris Manning (eds.), *Architectures, rules and preferences: A festschrift for Joan Bresnan*, 253–275. Stanford: CSLI.

Newman, Paul. 1998. We has seen the enemy and it is us: The endangered languages issue as a hopeless cause. *Studies in the Linguistic Sciences* 28. 11–20.

Newman, Paul. 2007. Copyright essentials for linguists. *Language Documentation & Conservation* 1. 28–43.

Neylon, Cameron. 2010. What would scholarly communications look like if we invented it today? http://cameronneylon.net/blog/what-would-scholarly-communications-look-like-if-we-invented-it-today/.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago.

Nichols, Johanna. 2005. E-MELD School of Best Practice: How many words do you need? http://e-meld.org/school/classroom/text/lexicon-size.html.

Noonan, Michael. 2006. Grammar writing for a grammar-reading audience. *Studies in Language* 30. 351–365.

Nordhoff, Sebastian. 2008. Electronic reference grammars for typology: Challenges and solutions. *Language Documentation & Conservation* 2. 296–324.

Palmer, Alexis. 2009. *Semi-automated annotation and active learning for language documentation*. Austin, Texas: University of Texas Ph.D. dissertation.

Palmer, Alexis & Katrin Erk. 2007. IGT-XML: An XML format for interlinearized glossed text. In *Proceedings of the linguistic annotation workshop (LAW '07)*, 176–183. Stroudsburg, Penn.: Association for Computational Linguistics. http://portal.acm.org/citation.cfm?id=1642059.1642087.

Palmer, Alexis, Taesun Moon & Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, 36–44. Stroudsburg, Penn.: Association for Computational Linguistics. http://aclweb.org/anthology-new/W/W09/W09-1905.pdf.

Palmer, Alexis, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell & Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology* 3. http://elanguage.net/journals/index.php/lilt/article/view/663.

Passant, Alexandre, John G. Breslin & Stefan Decker. 2010. Rethinking microblogging: Open, distributed, semantic. In Florian Daniel & Federico Michele Facca (eds.), *Proceedings of the 10th International Conference on Web Engineering (ICWE 2010)*, 263–277. Berlin: Springer.

Penton, David, Catherine Bow, Steven Bird & Baden Hughes. 2004. Towards a general model for linguistic paradigms. In *Proceedings of E-MELD 2004: Linguistic databases and best practice*. Detroit, Michigan, July 15–18. http://e-meld.org/workshop/2004/bird-paper.pdf.

Poornima, Shakthi & Jeff Good. 2010. Modeling and encoding traditional wordlists for machine applications. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the common ground (NLPLING '10)*, 1–9. Stroudsburg, Penn.: Association for Computational Linguistics.

Quirk, Randolph, Sydney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Reiman, D. Will. 2010. Basic oral language documentation. *Language Documentation & Conservation* 4. 254–268. http://hdl.handle.net/10125/4479.

Resnik, Philip & Jimmy Lin. 2010. Evaluation of NLP systems. In Alexander Clark, Chris Fox & Shalom Lappin (eds.), *The handbook of computational linguistics and natural language processing*, 271–295. Oxford: Wiley-Blackwell.

Rice, Keren. 2006a. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics* 4. 123–155.

Rice, Keren. 2006b. A typology of good grammars. *Studies in Language* 30. 385–415.

Rogers, Chris. 2010. Review of Fieldworks Language Explorer (FLEx) 3.0. *Language Documentation & Conservation* 4. 78–84. http://hdl.handle.net/10125/4471.

Sag, Ivan A., Thomas Wasow & Emily Bender. 2003. *Syntactic theory: A formal introduction (second edition)*. Stanford: CSLI.

Sapir, Edward. 1921. *Language: An introduction to the study of speech*. New York: Harcourt, Brace, and Company.

Schroeter, Ronald & Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In Linda Barwick & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 99–124. Sydney: The University of Sydney.

Schultze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 213–251. Berlin: Mouton de Gruyter.

Simons, Gary F. 2005. Beyond the brink: Realizing interoperation through an RDF database. In *Proceedings of E-MELD 2005: Linguistic ontologies and data categories for language resources*. Cambridge, Mass., July 1–3. http://e-meld.org/workshop/2005/papers/simons-paper.pdf.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky & Andrew Y. Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, 254–263. Stroudsburg, Penn.: Association for Computational Linguistics. http://portal.acm.org/citation.cfm?id=1613715.1613751.

Thieberger, Nicholas. 2004. Documentation in practice: Developing a linked media corpus of South Efate. In Peter K. Austin (ed.), *Language documentation and description, volume 2*, 169–178. London: Hans Rausing Endangered Languages Project.

Thieberger, Nicholas. 2009. Steps toward a grammar embedded in data. In Patience Epps & Alexandre Arkhipov (eds.), *New challenges in typology: Transcending the borders and refining the distinctions*, 389–407. Berlin: Mouton de Gruyter.

Trippel, Thorsten. 2006. *The Lexicon Graph Model: A generic model for multimodal lexicon development*. Saarbrücken: AQ-Verlag.

Trippel, Thorsten. 2009. Representation formats and models for lexicons. In Andreas Witt & Dieter Metzing (eds.), *Representation formats and models for lexicons*, 165–184. Berlin: Springer.

Weber, David J. 2006. Thoughts on growing a grammar. *Studies in Language* 30. 417–444.

Wittenburg, Peter, Wim Peters & Sebastian Drude. 2002. Analysis of lexical structures from field linguistics and language engineering. In Manuel González Rodríguez & Carmen Paz Suarez Araujo (eds.), *Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002)*, 682–686. Paris: European Language Resources Association.

Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–186. Cambridge: Cambridge University.

Wright, Sue Ellen & Gerhardt Budin (eds.). 1997. *Handbook of terminology management, volume 1: Basic aspects of terminology management*. Amsterdam, Benjamins.

Wright, Sue Ellen, Marc Kemps-Snijders & Menzo Windhouwer. 2010. The OWL and the ISOcat: Modeling relations in and around the DCR. In *Proceedings of the Language Resource and Language Technology Standards Workshop (LREC10-W4)—State of the art, emerging needs, and future developments*. http://www.windhouwer.nl/menzo/professional/papers/Wright_OWL_DCR.pdf.

Zaefferer, Dietmar. 2006. Realizing Humboldt's dream: Cross-linguistic grammatography as data-base creation. In Felix Ameka, Alan Dench & Nicholas Evans (eds.), *Catching language: The standing challenge of grammar writing*, 113–135. Berlin: Mouton de Gruyter.

Jeff Good
jcgood@buffalo.edu