Montage: Leveraging advances in grammar engineering, linguistic ontologies, and mark-up for the documentation of underdescribed languages

Emily Bender*, Dan Flickinger†, Jeff Good‡, Ivan Sag†

*University of Washington
Department of Linguistics, Box 354350, Seattle, WA 98195-4340, USA
ebender@u.washington.edu

†Stanford University
CSLI/Ventura Hall, Stanford, CA 94305-2150, USA
{danf, sag}@csli.stanford.edu

‡University of Pittsburgh
Department of Linguistics, 2816 Cathedral of Learning, Pittsburgh, PA 15260, USA
jcgood@pitt.edu

Abstract

The Montage project aims to develop a suite of software tools which will assist field linguists in organizing and analyzing the data they collect while at the same time producing resources which are easily discoverable and accessible to the community at large. Because we believe that corpus methods, descriptive analysis, and implemented formal grammars can all inform each other, our suite of software tools will provide support for all three activities in an interoperable manner.

1. Introduction

The Montage (Markup for ONTological Annotation and Grammar Engineering) project aims to develop a suite of software whose primary audience is field linguists working on underdocumented languages. The tool suite is designed to have five major components: a manual markup tool to allow for basic grammatical annotation of data, a grammar export tool to allow annotated data to be summarized in a way similar to a traditional grammatical description, a labeled bracketing tool for incorporating information about syntactic relations into the data, a "grammar matrix" to assist with development of a precision formal grammar, and a tool which uses manually annotated data and a formal grammar to partially automate the annotation process.

2. Goal of the paper

The goal of this paper is to given on overview of the structure of the Montage toolkit with an emphasis on how it fits into the traditional conception of field work and language documentation and how the tools to be developed build off of existing tools for formal grammar engineering. Section 3 discusses which aspects of language documentation will be enhanced by the Montage toolkit. Section 4 describes the structure of the toolkit from a technical perspective. Section 5 describes how some of the tools which form the core of Montage will be adapted from existing tools for formal grammar engineering.

3. Language documentation and the Montage toolkit

Traditionally, the process of language documentation has been an extraordinarily labor-intensive and time-consuming task. It involves hundreds of hours of elicitation with native speakers. Based on such elicitation, basic documentary resources like audio and video tapes as well as annotated resources like transcribed texts and word lists can then be produced. After this is done, the data collector can begin to perform grammatical analysis on the language. In the ideal case, this work results in the creation of a descriptive grammar, a dictionary, and a small collection of translated and analyzed texts. Often, however, the barriers to the production of these documents are so high that they are never completed. When this is the case, the data from the language typically remains highly inaccessible and is effectively "lost" to the general community.

In the past few years, a small number of organizations have begun the project of developing digital standards and tools in order to make the task of language documentation easier as well as to ensure that digital resources created by field linguists are accessible to a wide audience and will not be lost as digital technology evolves. Some of these initiatives include the Electronic Metastructure for Endangered Languages project[1] (EMELD), the Dokumentation Bedrohter Sprachen project[2] (DoBeS), the Querying Linguistic Databases project[3], and the Open Language Archives Community[4] (OLAC).

These initiatives are developing tools that address some of the issues faced by field linguists. For example, the creation of dictionaries will be facilitated by EMELD's Field Input Environment for Linguistic Data tool[5] (FIELD), and the Elan tool[6], developed by the DoBeS project, is useful for the basic task of transcribing data. In addition to

[1] http://www.emeld.org
[2] http://www.mpi.nl/DOBES/
[3] https://www.fastlane.nsf.gov/servlet/showaward?award=0317826
[4] http://www.language-archives.org
[5] http://saussure.linguistlist.org/cfdocs/emeld/tools/fieldinput.cfm
[6] http://www.mpi.nl/tools/elan.html

these tools, the EMELD project is also developing an ontology of grammatical terms, called the General Ontology for Linguistic Description (GOLD) (Farrar and Langendoen, 2003). This ontology is designed to improve access to digital resources by creating a uniform means of annotating them for grammatical information, without necessarily imposing any particular theory or terminology on researchers.

Such tools represent an enormous change in the software available to field linguists. However, there remains a notable gap: Nothing yet available or currently under development supports the descriptive grammar component of field linguistic research.[7] The Montage toolkit will assist in such grammatical analysis, from foundational descriptive work to the statement and testing of precise hypotheses about grammatical structure.

Figure 1 illustrates which aspects of language documentation Montage is intended to facilitate. For illustrative purposes, the figure includes how two other tools—Elan and FIELD, discussed above—fit into this model of documentation. As schematized in the figure, Montage will (i) assist in creating annotated texts, specifically texts annotated for grammatical information, (ii) include tools for extracting information from the annotated texts to facilitate production of descriptive grammars, and (iii) allow information in descriptive grammars and electronic lexicons to serve as the foundation for the construction of formal grammars. As will be discussed in the next section, such formal grammars will be used by the system to partially automate annotation and analysis of data.

An important feature of Montage will be that it will allow grammatical annotations to be linked to external ontologies for grammatical terms. The use of ontologies will not be enforced in the toolkit, and the researcher will always have the freedom to use their own terminology. However, should they choose to use the terminology provided by the ontology or use other terminology but link it into the ontology, Montage will make this straightforward. The toolkit will, thus, be able to make important contributions to the creation of interoperable linguistic resources. The particular ontology which will be employed during the development of Montage is the GOLD ontology. However, the tookit's design will not restrict the user to any one particular ontology.

While implemented formal grammars have not traditionally been a part of language documentation, we believe that the current state of the art in computational linguistics is such that field linguists can now benefit from the enhanced hypothesis testing of grammar implementation without needing to become expert in a second subspecialty. Because of this, implemented formal grammars have an important position in Figure 1 with respect to the design of Montage—even if they don't fit into the traditional model.

In addition, we expect that the formal grammars produced by the toolkit will be valuable to software engineers working on tools which require knowledge of a language's grammar. To this point, such tools have generally only been available for majority languages. Montage will facilitate the creation of such tools for minority languages.

[7]The SIL tool, the Linguist's Shoebox, which has been in use for over a decade, can allow a linguist to perform basic text markup and, therefore, assist in grammatical analysis. However, this tool does not provide the support for the development of descriptive and formal grammars that is part of the design of Montage.

4. The design of Montage

The Montage toolkit will comprise five different tools, each of which could be used independently but which, when used together, will be designed to greatly enhance the workflow of the field linguist. The five tools are each discussed in turn.

- **Manual markup tool:** This tool will allow the markup of basic linguistic data for grammatical information. Its design will allow it to interface with an ontology of grammatical terms as well as with electronic lexicons so that morphemes in the data can be associated with their lexical entries.

- **Grammar export tool:** This tool will be a type of "smart" export tool to allow data annotated for grammatical information to be put into a format which facilitates traditional grammatical description. For example, it will export interlinearized example sentences as well as grammatial "notes" made by the linguist for particular linguistic constructions. Support will be included for creating both hyper-text grammars and traditional print grammars.

- **Labeled bracketing tool:** This tool will be similar to the markup tool except it will be specifically designed to annotate sentences for the phrase structure and to give grammatical labels to various levels of phrase structure. This tool will, therefore, facilitate syntactic description as well as the formation of formal implemented grammars.

- **Grammar matrix:** The Grammar Matrix is a language-independent core grammar designed to facilitate the rapid development of implemented precision grammars for diverse languages. (It will be discussed in more detail in section 5)

- **LKB/[incr tsdb()] tools:** These are two existing tools, the Linguistic Knowledge Builder and the [incr tsdb()] Competence and Performance Laboratory which will be used together to allow for semi-automatic parsing of data to find candidate sentences for possible grammatical annotation. (These tools will be discussed in more detail in section 5)

Figure 2 schematizes the workflow of grammatical description using the Montage toolkit. An important aspect of workflow using Montage is the "positive feedback loop" seen in the diagram. After the researcher manually marks up a set of data and creates a partial formal grammar, Montage will examine an entire corpus to find sentences not annotated for a particular grammatical feature but which would be good candidates for such annotation. A partially annotated corpus can, therefore, "jump-start" the process of annotating an entire corpus.

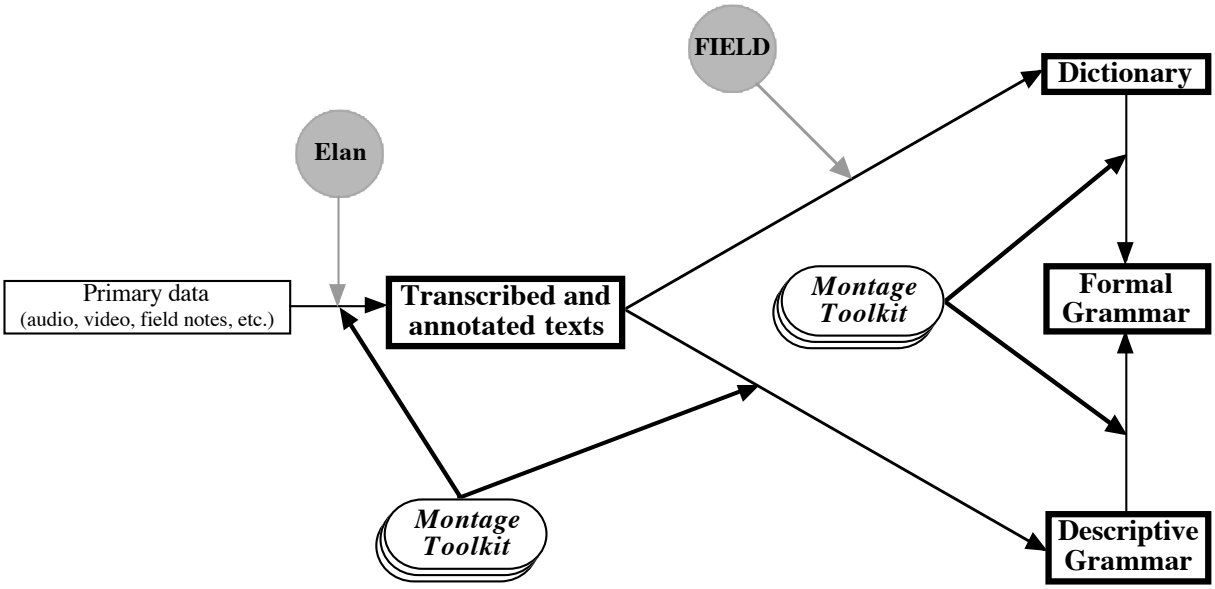

Figure 1: Language documentation and the Montage toolkit

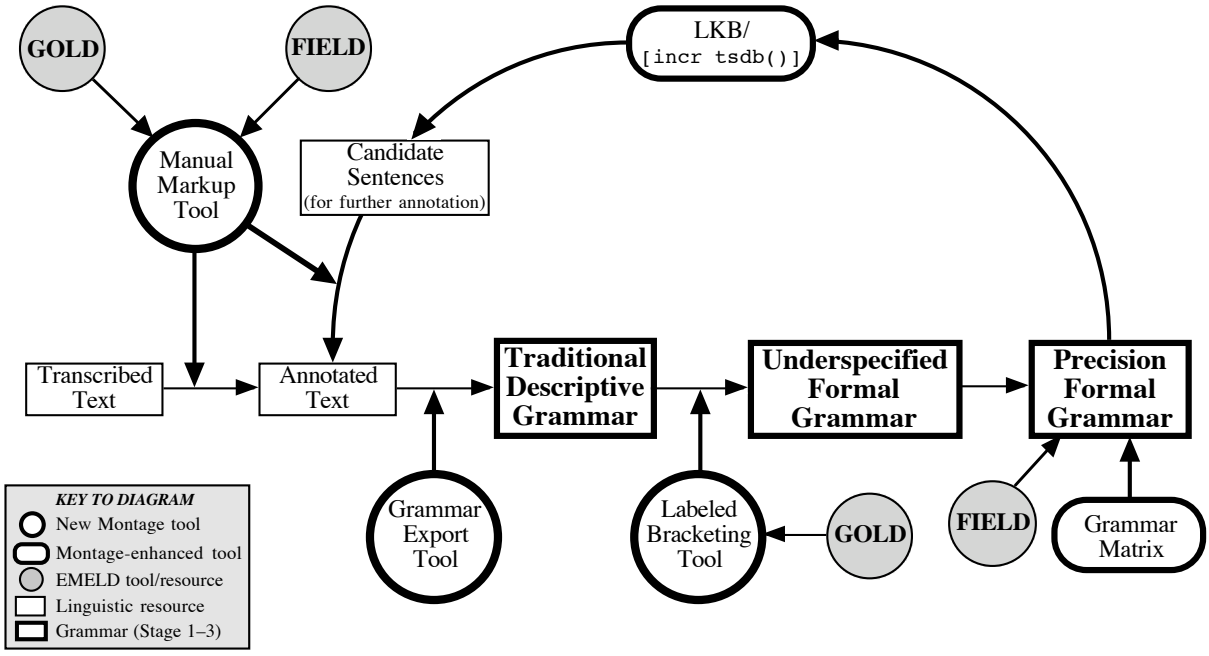

Figure 2: Workflow using the Montage toolkit

As should be clear from Figure 2, the Montage toolkit does not assume that work on the grammar of a language proceeds "serially". Rather, it assumes that work on each resource can assist in work on the other resources. For example, a partial descriptive grammar can assist in the production of a partial formal grammar which, in turn, can assist in the annotation of texts.

This model is designed with two ideas in mind. The first is that traditional field work largely proceeds in this "parallel" fashion—for example, informal work on grammatical description typically accompanies text analysis with no strict division of the work. The second reason for this model of workflow is to ensure that even incomplete grammatical analysis can produce a range of valuable resources. A partially annotated corpus of texts can easily be produced along side of a partial descriptive or formal grammar, for example. This will allow researchers to collaborate more easily on grammatical description and, crucially, will not necessitate that grammatical analysis only be publicly disseminated after it is "complete".

5. Refining existing tools as part of Montage

An important aspect of Montage is that, of its five core tools, two of them will be directly based on existing tools

for grammar engineering; these are the Grammar Matrix and the LKB/[incr tsdb()] tool combination, both developed as part of the LinGO (Linguistic Grammars Online) project.[8] Our use of such tools represents, we believe, an important convergence between the methods of computational linguistics and the methods of descriptive linguistics.

The Grammar Matrix (Bender et al., 2002) is designed to jump-start the process of implementing precision grammars by abstracting knowledge gained in grammar engineering activities done by the LinGO project into a form that can be easily reused by grammar engineers working on new languages. An early prototype of the Grammar Matrix is being used with promising results in the development of grammars for Norwegian (Hellan and Haugereid, 2003), Modern Greek (Kordoni and Neu, 2003), and Italian[9] (all funded by the EU Project 'DeepThought'[10]). Currently, the most elaborated portion of the Grammar Matrix is the syntax-semantics interface. This aspect of the core grammar assists grammar engineers in converging on consistent semantic representations for different languages.

The LKB grammar development environment (Linguistic Knowledge Builder; Copestake (1992, 2002)) , includes a parser and a generator as well as support for developing implemented formal (typed feature structure) grammars. [incr tsdb()] (Oepen, 2001) is a comprehensive environment for profiling and evaluating both grammars and parsing systems which is integrated with the LKB. The system design of Montage will allow the linguist to use the LKB/[incr tsdb()] tools directly, and, in addition, will provide for additional levels of functionality specifically designed to facilitate identification of candidate sentences for grammatical annotation.

While the tools developed by the LinGO project were designed with formal grammars in mind, they assume a model of grammar not dissimilar to that employed by the traditional field linguist, and, thus, can be directly applied to descriptive work. Specifically, both LinGO grammars and traditional descriptive grammars assume a rich category structure is operative in language and that grammatical description consists of generalizations over those categories. The main difference between descriptive grammars and the formal model of grammar employed by the LinGO tools is simply one of precision—in order to be machine readable, a restricted, well-defined set of categories must be rigidly employed for resources using the LinGO tools, while this requirement has not been essential for traditional grammatical description.

However, even though descriptive grammarians have not generally aimed for the level of precision required for computational applications, with the rise of the use of digital resources in all aspects of linguistics, efforts have begun to make descriptive materials precise in a way which would facilitate their being machine-readable.

The EMELD project's work on the GOLD ontology is a good example of research in this vein, since it is an attempt to codify traditional terminology into a well-defined controlled vocabulary of terms which can be used in all kinds of linguistic resources. In order to take full advantage of the accessibility provided by an ontology, we intend to support links to the ontology from both the descriptive and implemented grammars created with Montage. We expect that this work will place new demands on the GOLD ontology. Thus, while Montage has been made possible, to a large extent, by work on ontologies, we expect work developing the toolkit will also be valuable in refining and enhancing the ontologies themselves.

[8] http://lingo.stanford.edu

[9] http://www.celi.it/english/hpsg_itgram.htm

[10] http://www.project-deepthought.net

6. Conclusion

The goal of the Montage project is to make advances in electronic data management and computational linguistics accessible to field linguists working on the documentation of grammars of underdescribed languages. We envision two final products based on the resources of our toolkit. The first is the modern version of the traditional descriptive grammar. Without the inherent limitations of a paper-based format, these electronic grammars will allow easy access to the entire corpus of source examples, enhancing linguistic research. The second is a set of machine-readable resources codifying the grammatical analyses of the language. These resources will be valuable in linguistic hypothesis testing as well as practical applications such as machine translation or computer assisted language learning.

7. References

Bender, Emily M., Dan Flickinger, and Stephan Oepen, 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th Internation Conference on Computational Linguistics*. Taipei, Taiwan.

Copestake, Ann, 1992. The ACQUILEX LKB: Representation issues in the semi-automatic acquisition of large lexicons (ACQUILEX WP No. 36). In Antonio Sanfilippo (ed.), *The (other) Cambridge ACQUILEX papers*. University of Cambridge Computer Laboratory, Technical report No. 253.

Copestake, Ann, 2002. *Implementing Typed Feature Structure Grammars*. Stanford, CA: CSLI Publications.

Farrar, Scott and Terry Langendoen, 2003. A linguistic ontology for the semantic web. *GLOT International*, 7:97–100.

Hellan, Lars and Petter Haugereid, 2003. Norsource – an exercise in the Matrix Grammar building design. In Emily M. Bender, Dan Flickinger, Frederik Fouvry, and Melanie Siegel (eds.), *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI 2003*.

Kordoni, Valia and Julia Neu, 2003. Deep grammar development for Modern Greek. In Emily M. Bender, Dan Flickinger, Frederik Fouvry, and Melanie Siegel (eds.), *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI 2003*.

Oepen, Stephan, 2001. [incr tsdb()] — competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany. In preparation.