Using Linguist's Assistant for Language Description and Translation

Stephen Beale
University of Maryland, Baltimore County
Baltimore, MD
sbeale@cs.umbc.edu

Abstract

The Linguist's Assistant (LA) is a practical computational paradigm for describing languages. LA seeks to specify in semantic representations a large subset of possible written communication. These semantic representations then become the starting point and organizing principle from which a linguist describes the linguistic surface forms of a language using LA's visual lexicon and grammatical rule development interface. The resulting computational description can then be used in our document authoring and translation applications.

1 Introduction

The Linguist's Assistant (LA) is a practical computational paradigm for describing languages. LA approaches the complex task of language description from two directions. From one side, LA is built on a comprehensive semantic foundation. We combine a conceptual, ontological framework with detailed semantic features that cover (or is a beginning towards the goal of covering) the range of human communication. An elicitation procedure has been built up around this central, semantic core that systematically guides the linguist through the language description process, during which the linguist builds a grammar and lexicon that "describes" how to generate target language text from the semantic representations of the elicitation corpus. The result is a "how to" guide for the language: how does one encode a given semantic representation in the language?

Coming at the problem from the other side, LA also allows the linguist to collect language data in a more conventional manner – from naturally occurring texts and linguistically motivated elicitations (for example, a linguist in Vanuatu might want to explore alienable vs. inalienable possession or serial verb constructions using naturally occurring texts). Such texts are semantically analyzed using a convenient semi-automatic document authoring interface ("authored" in our context means that a semantic representation has been prepared), in effect adding them to the standard elicitation corpus. Existing grammar rules and lexical information can then either be confirmed or adjusted, or new descriptive knowledge added that allows the built-in text generator to produce target text that is substantially equivalent to the elicited examples. The result is a "how did" guide for the language: how did a native speaker encode natural text or linguistically focused elicitations?

We believe that the combination of semantically motivated and linguistically motivated elicitation and description provides an ideal balance. The semantic-based elicitation is general and uniform across languages. It provides an efficient and relatively comprehensive standard for describing the majority of the linguistic phenomena in a language. We have found it to be an invaluable starting point in the description process. It is, however, impossible to produce a general semantic-based elicitation scheme that is not overly burdensome on the user. In addition, linguists typically know the "interesting," atypical or difficult aspects of a language. This is where linguistically based elicitation is invaluable.

A third approach to language description is encouraged in the LA framework: acquiring knowledge (lexicon and grammar) to cover pre-authored texts. The semantically and linguistically motivated elicitations from the first two approaches above provide a solid foundation for lexicon and grammar development, but we have found that adding to that the experience and discipline of acquiring the knowledge necessary to generate actual texts is invaluable. This is usually

Proceedings of the IJCNLP 2011 System Demonstrations, pages 5–8,
Chiang Mai, Thailand, November 9, 2011.

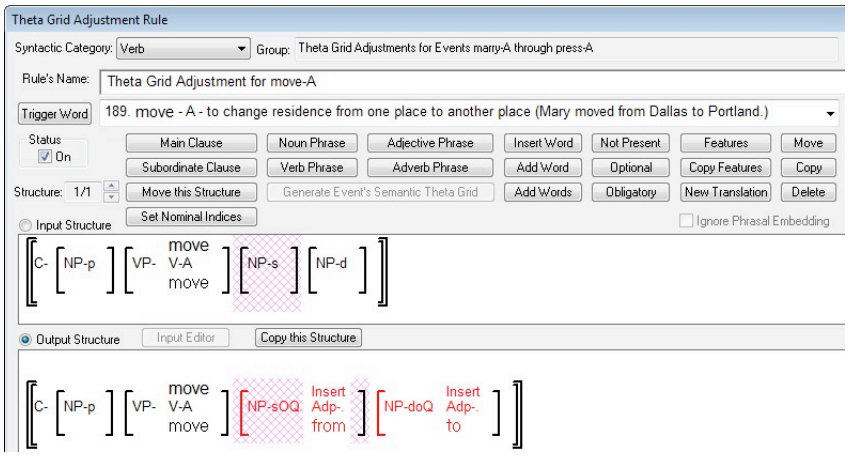

Figure 1. Visual Interface for grammatical rules

the best opportunity for documenting phenomena that are more lexically dependent since the vocabulary in the semantic-based elicitation stage is quite limited. For this reason we include several pre-authored (i.e. semantically analyzed and ready for use in our translation module) community development texts with LA.

Underlying all these approaches to knowledge acquisition in LA is a visual, semi-automatic interface for recording grammatical rules and lexical information. Figure 1 shows an example of one kind of visual interface used for "theta-grid adjustment rules." The figure shows an English rule used to adjust the "theta grid" or "case frame" of an English verb. Grammatical rules typically describe how a given semantic structure is realized in the language. The whole gamut of linguistic phenomena is covered, from morphological alternations (Figure 2) to case frame specifications to phrase structure ordering (Figure 3) to lexical collocations – and many others. These grammatical rules interplay with a rich lexical description interface that allows for assignment of word-level features and the description of lexical forms associated with individual roots (Figure 4). Currently, the linguist is responsible for the creation of rules, albeit with a natural, visual interface that often is able to set up the requisite input semantic structures automatically. We continue work on a module that will allow the semi-automatic generation of rules similar to research in the BOAS (McShane, et al., 2002), LinGO (Bender, at al., 2010), PAWS (Black and Black, 2009) and Avenue (Probst, et al., 2003) projects. Such a module will, we believe, make LA accessible to a larger pool of linguists. We also provide a growing list of rule templates that linguists can use to describe common linguistic phenomena.

Integrated with these elicitation and description tools is a text generator that allows for immediate confirmation of the validity of grammatical rules and lexical information. We also provide an interface for tracking the scope and examples of grammatical rules. This minimizes the possibility of conflicting or duplicate rules while providing the linguist a convenient index into the work already accomplished. And finally, we provide a utility for producing a written description of the language - after all, a computational description of a language is of no practical use (outside of translation applications) unless it can be conveniently referenced. Refer to Beale (submitted) for a comprehensive description of Linguist's Assistant.

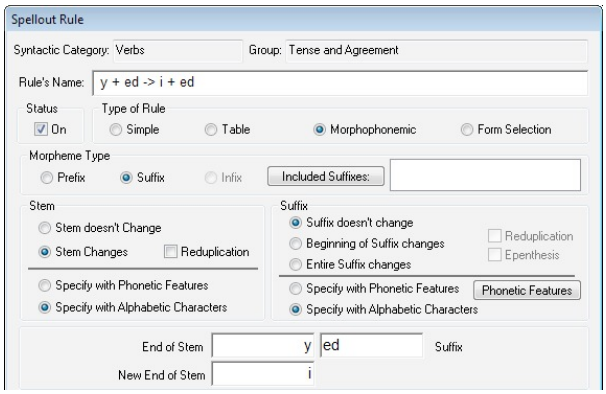

Figure 2. Morphological alternation rule

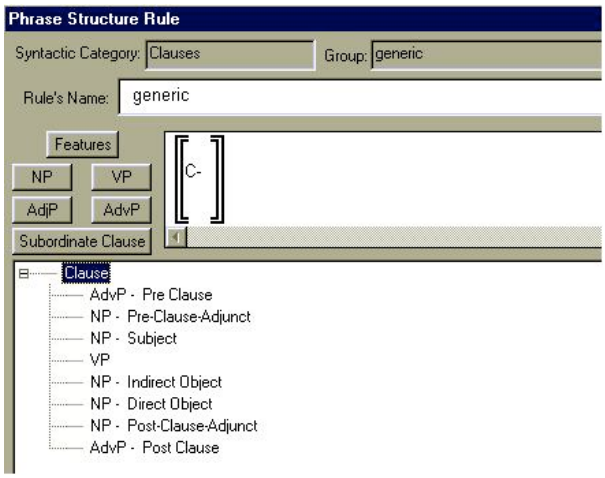

Figure 3. Phrase structure ordering rule

	Stems	Glosses	infinitive	present indic 1st sing
1	aprend	learn	aprender	aprendo
2	habl	speak	hablar	hablo
3	ten	have	tener	tengo
4	viv	live	vivir	vivo

present indic 2nd sing	present indic 3rd sing	present indic 1st pl	present indic 3rd pl
aprendes	aprende	aprendemos	aprenden
hablas	habla	hablamos	hablan
tienes	tiene	tenemos	tienen
vives	vive	vivimos	viven

Figure 4. Lexical forms for Spanish

LA has been used to produce extensive grammars and lexicons for Jula (a Niger-Congo language), Kewa (Papua New Guinea), North Tanna (Vanuatu), Korean and English. Work continues in two languages of Vanuatu, with additional languages planned in the near future. The resulting computational resources have been used in our separate document authoring and translation applications to produce a significant amount of high-quality translations in each of these languages. Figures 5 and 6 present translations of a section of a medical text on AIDS into English Korean. Please reference Beale et al. (2005) and Allman and Beale (2004; 2006) for more information on using LA in translation projects, and for documentation on the evaluations of the translations produced. Note: LA can be used as the language-description module within our larger applications called TA (The Translator's Assistant, for translating health and community development materials, as well as "authoring" new texts) or TBTA (The Bible Translator's Assistant, for those interested in Bible Translation). We argue that the high quality results achieved in translation projects demonstrate the quality and coverage of the underlying language description that LA produces.

Kande's Story 1:1 Title: Kande's mother knows a secret.

Kande's Story 1:2 One day a girl named Kande was sitting near a tree. Kande was reading a book. She had a younger sister named Teshi. Teshi ran to Kande. Teshi was very excited. She said to Kande, "Kande! Kande! I heard certain women talking to each other. Those women said that mother knows a secret! Do you know mother's secret?"

Kande's Story 1:3 Kande said, "I might know mother's secret. We should go to our house and talk to our mother. Mother might tell us about her secret. I'll race you to our house!"

Kande's Story 1:4 Kande and Teshi ran to their house quickly. When Kande and Teshi arrived at the house, they were laughing. They had two younger sisters. One younger sister's name was Falala. And the other younger sister's name was Iniko. Kande and Teshi also had a younger brother named Jumoke. Falala, Iniko, and Jumoke heard Kande and Teshi laughing. So they ran to the door to see Kande and Teshi. Then mother said to all the children, "Be quiet because your father has to sleep." Then she walked from

Figure 5. English translation of a medical text

Korean

Kande's Story 1:1 제목: 칸디의 어머니는 비밀을 알고 있어요.

Kande's Story 1:2 어느 날 칸디라는 소녀가 나무 가까이에 앉아 있었다. 칸디는 책을 읽고 있었다. 칸디는 태쉬라는 여동생이 있었다. 태쉬는 칸디에게 달려갔다. 태쉬는 매우 흥분하였다. 태쉬는 칸디에게 말하였다. "언니! 언니! 나는 어떤 여자들이 서로에게 말하는 것을 들었어요. 이 여자들은 어머니께서 비밀을 알고 계시다고 말하였어요! 언니는 어머니의 비밀을 알고 있어요?"

Figure 6. Korean translation of a medical text

2 Content of the Demonstration

A partial example of the content of the proposed demonstration can be found at http://ilit.umbc.edu/sbeale/LA/ under the "Demo Videos" link. These demonstration videos are part of an online journal article (Beale, submitted) that describes LA in depth. A draft of this journal article can be found at the same website under the "Publications" link.

We will be prepared to demonstrate, as appropriate to the interests of a particular group of participants, the following:

- An overview of LA
- The semantic representation system
- The document authoring system that enables the semi-automatic analysis of new texts or elicitations
- How to create lexicons that are appropriate for different kinds of languages
- How to use the visual rule creation interface to create various kinds of grammatical rules
- Multilingual examples of lexicons
- Multilingual examples of grammatical rules
- Multilingual examples of translation results

We will also prepare 10 minute modules with "hands-on" examples for any interested participants who wish to take a bit more time investigating LA.

3 Previous Experience in Teaching LA

LA is the basis of a semester-long Honor's College class at the University of Maryland, Baltimore County. In that class we present an overview of different types of linguistic phenomena. We then use LA to encode descriptive knowledge of multi-lingual examples of each. The class size is 25 students.

We have also prepared tutorials and online demonstrations (http://ilit.umbc.edu/sbeale/LA/) and informally used LA with a number of field linguists.

4 Required Resources

We require a single projector. Internet service is not necessary.

5 Acknowledgements

The author gratefully acknowledges the partnership of Tod Allman from the University of Texas, Arlington. Dr. Allman is co-developer of LA.

References

Allman, Tod. 2010. The translator‘s assistant: a multilingual natural language generator based on linguistic universals, typologies, and primitives. Arlington, TX: University of Texas dissertation.

Tod Allman and Stephen Beale. 2006. “A natural language generator for minority languages,” in Proceedings of SALTMIL, Genoa, Italy.

Tod Allman and Stephen Beale. 2004. “An environment for quick ramp-up multi-lingual authoring,” International Journal of Translation 16(1).

Stephen Beale. Submitted. “Documenting endangered languages with linguist’s assistant.” Language Documentation and Conservation Journal. Draft available at: http://ilit.umbc.edu/sbeale/LA/papers/DEL-for-LDC-journal.pdf

Stephen Beale, S. Nirenburg, M. McShane, and Tod Allman. 2005. “Document authoring the Bible for minority language translation,” in Proceedings of MT-Summit, Phuket, Thailand.

Emily Bender, S. Drellishak, A. Fokkens, M. Goodman, D. Mills, L. Poulson, and S. Saleem. 2010. “Grammar prototyping and testing with the LinGO grammar matrix customization system,” in Proceedings of the ACL 2010 System Demonstrations.

Sheryl Black and Andrew Black. 2009. “PAWS: parser and writer for syntax: drafting syntactic grammars in the third wave,” http://www.sil.org/silepubs/PUBS/51432/SILForum2009-002.pdf.

Marjorie McShane, Sergei Nirenburg, Jim Cowie, and Ron Zacharski. 2002. “Embedding knowledge elicitation and MT systems within a single architecture,” Machine Translation 17(4), pp.271-305.

Katharina Probst, Lori Levin, Erik Petersen, Alon Lavie and Jaime Carbonell. 2003. “MT for minority languages using elicitation-based learning of syntactic transfer rules,” Machine Translation 17(4), pp.245-270.