WELT: Using Graphics Generation in Linguistic Fieldwork

Morgan Ulinski*
mulinski@cs.columbia.edu

Anusha Balakrishnan*
ab3596@columbia.edu

Daniel Bauer*
bauer@cs.columbia.edu

Bob Coyne*
coyne@cs.columbia.edu

Julia Hirschberg*
julia@cs.columbia.edu

Owen Rambow†
rambow@ccls.columbia.edu

*Department of Computer Science †CCLS
Columbia University
New York, NY, USA

Abstract

We describe the WordsEye Linguistics tool (WELT), a novel tool for the documentation and preservation of endangered languages. WELT is based on WordsEye (Coyne and Sproat, 2001), a text-to-scene tool that automatically generates 3D scenes from written input. WELT has two modes of operation. In the first mode, English input automatically generates a picture which can be used to elicit a description in the target language. In the second mode, the linguist formally documents the grammar of an endangered language, thereby creating a system that takes input in the endangered language and generates a picture according to the grammar; the picture can then be used to verify the grammar with native speakers. We will demonstrate WELT's use on scenarios involving Arrernte and Nahuatl.

1 Introduction

Although languages have appeared and disappeared throughout history, today languages are facing extinction at an unprecedented pace. Over 40% of the estimated 7,000 languages in the world are at risk of disappearing. When languages die out, we lose access to an invaluable resource for studying the culture, history, and experience of peoples around the world (Alliance for Linguistic Diversity, 2013). Efforts to document languages and develop tools in support of collecting data on them become even more important with the increasing rate of extinction. Bird (2009) emphasizes a particular need to make use of computational linguistics during fieldwork.

To address this issue, we are developing the WordsEye Linguistics Tool, or WELT. In the first mode of operation, we provide a field linguist with tools for running custom elicitation sessions based on a collection of 3D scenes. In the second, input in an endangered language generates a picture representing the input's meaning according to a formal grammar.

WELT provides important advantages for elicitation over the pre-fabricated sets of static pictures commonly used by field linguists today. The field worker is not limited to a fixed set of pictures but can, instead, create and modify scenes in real time, based on the informants' answers. This allows them to create additional follow-up scenes and questions on the fly. In addition, since the pictures are 3D scenes, the viewpoint can easily be changed, allowing exploration of linguistic descriptions based on different frames of reference. This will be particularly useful in eliciting spatial descriptions. Finally, since scenes and objects can easily be added in the field, the linguist can customize the images used for elicitation to be maximally relevant to the current informants.

WELT also provides a means to document the semantics of a language in a formal way. Linguists can customize their studies to be as deep or shallow as they wish; however, we believe that a major advantage of documenting a language with WELT is that it enables such studies to be much more precise. The fully functioning text-to-scene system created as a result of this documentation will let linguists easily test the theories they develop with native speakers, making changes to grammars and semantics in real time. The resulting text-to-scene system can be an important tool for language preservation, spreading interest in the language among younger generations of the community and recruiting new speakers.

We will demonstrate the features of WELT for use in fieldwork, including designing elicitation sessions, building scenes, recording audio, and adding descriptions and glosses to a scene. We will use examples from sessions we

Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 49–54,
Baltimore, Maryland USA, June 23-24, 2014. ©2014 Association for Computational Linguistics

have conducted with a native speaker of Nahuatl, an endangered language spoken in Mexico. We will demonstrate how to document semantics with WELT, using examples from Arrernte, an Australian aboriginal language spoken in Alice Springs. We will also demonstrate a basic Arrernte text-to-scene system created in WELT.

In the following sections, we will mention related work (Section 2), discuss the WordsEye system that WELT is based on (Section 3), describe WELT in more detail, highlighting the functionality that will appear in our demonstration (Section 4), and briefly mention our future plans for WELT (Section 5).

2 Related Work

One of the most widely-used computer toolkits for field linguistics is SIL Fieldworks. FieldWorks is a collection of software tools; the most relevant for our research is FLEx, Fieldworks Language Explorer. FLEx includes tools for eliciting and recording lexical information, dictionary development, interlinearization of texts, analysis of discourse features, and morphological analysis. An important part of FLEx is its "linguist-friendly" morphological parser (Black and Simons, 2006), which uses an underlying model of morphology familiar to linguists, is fully integrated into lexicon development and interlinear text analysis, and produces a human-readable grammar sketch as well as a machine-interpretable parser.

Several computational tools aim to simplify the formal documentation of syntax by eliminating the need to master particular grammar formalisms. First is the PAWS starter kit (Black and Black, 2012), a system that prompts linguists with a series of guided questions about the target language and uses their answers to produce a PC-PATR grammar (McConnel, 1995). The LinGO Grammar Matrix (Bender et al., 2002) is a similar tool developed for HPSG that uses a type hierarchy to represent cross-linguistic generalizations.

The most commonly used resource for formally documenting semantics across languages is FrameNet (Filmore et al., 2003). FrameNets have been developed for many languages, including Spanish, Japanese, and Portuguese. Most start with English FrameNet and adapt it for the new language; a large portion of the frames end up being substantially the same across languages (Baker, 2008). ParSem (Butt et al., 2002) is a collaboration to develop parallel semantic representations across languages, by developing semantic structures based on LFG. Neither of these resources, however, are targeted at helping non-computational linguists formally document a language, as compared to the morphological parser in FLEx or the syntactic documentation in PAWS.

3 WordsEye Text-to-Scene System

WordsEye (Coyne and Sproat, 2001) is a system for automatically converting natural language text into 3D scenes representing the meaning of that text. WordsEye supports language-based control of spatial relations, textures and colors, collections, facial expressions, and poses; it handles simple anaphora and coreference resolution, allowing for a variety of ways of referring to objects. The system assembles scenes from a library of 2,500 3D objects and 10,000 images tied to an English lexicon of about 15,000 nouns.

The system includes a user interface where the user can type simple sentences that are processed to produce a 3D scene. The user can then modify the text to refine the scene. In addition, individual objects and their parts can be selected and highlighted with a bounding box to focus attention.

Several thousand real-world people have used WordsEye online (http://www.wordseye.com). It has also been used as a tool in education, to enhance literacy (Coyne et al., 2011b). In this paper, we describe how we are using WordsEye to create a comprehensive tool for field linguistics.

Vignette Semantics and VigNet To interpret input text, WordsEye uses a lexical resource called VigNet (Coyne et al., 2011a). VigNet is inspired by and based on FrameNet (Baker et al., 1998), a resource for lexical semantics. In FrameNet, lexical items are grouped together in frames according to their shared semantic structure. Every frame contains a number of frame elements (semantic roles) which are participants in this structure. The English FrameNet defines the mapping between syntax and semantics for a lexical item by providing lists of valence patterns that map syntactic functions to frame elements.

VigNet extends FrameNet in two ways in order to capture "graphical semantics'," the knowledge needed to generate graphical scenes from language. First, graphical semantics are added to the frames by adding primitive graphical (typically, spatial) relations between the frame element fillers. Second, VigNet distinguishes between meanings of words that are distinguished graphically. For example, the specific objects and spatial relations in the graphical semantics for *cook* depend on the object being cooked and on the culture in which it is being cooked (cooking turkey in Baltimore vs. cooking an egg in Alice Springs), even though at an abstract level *cook an egg in Alice Springs* and *cook a turkey in Baltimore* are perfectly compositional semantically. Frames augmented with graphical semantics are called *vignettes*.

4 WordsEye Linguistics Tool (WELT)

In this section, we describe the two modes of WELT, focusing on the aspects of our system that will appear in our demonstration.

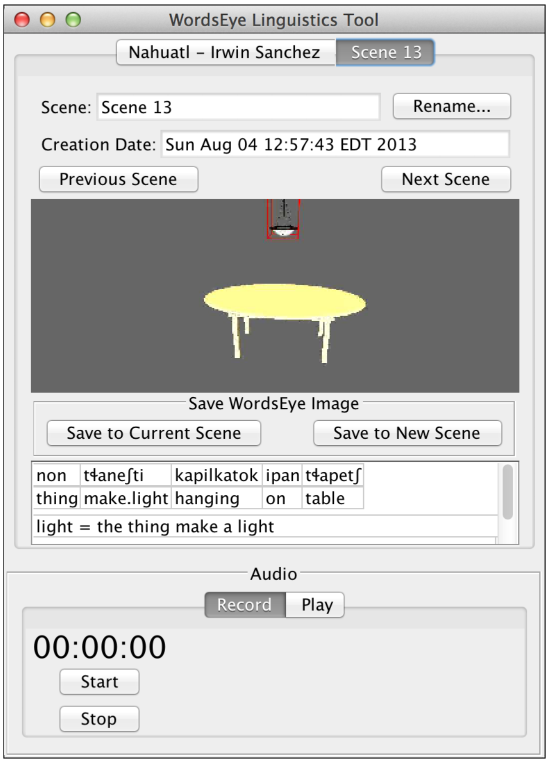
4.1 Tools for Linguistic Fieldwork

WELT includes tools that allow linguists to elicit language with WordsEye. Each elicitation session is organized around a set of WordsEye scenes. We will demonstrate how a linguist would use WELT in fieldwork, including (1) creating an elicitation session, either starting from scratch, or by importing scenes from a previous session; (2) building scenes in WordsEye, saving them to a WELT session, and modifying scenes previously added to the session, either overwriting the original scene or saving the changes as a new scene; (3) adding textual descriptions, glosses, and notes to a scene; and (4) recording audio, which is automatically synced to open scenes, and playingit back tto review any given scene. A screen shot of the scene annotation window is included in Figure 1.

To test the fieldwork capabilities of WELT, we created a set of scenes based on the Max Planck topological relations picture series (Bowerman and Pederson, 1992). We used these scenes to elicit descriptions from a native Nahuatl speaker; some examples of scenes and descriptions are included in Figure 2.

4.2 Formal Documentation of a Language

WELT also provides the means to formally document the semantics of a language and create a text-to-scene system for that language. The formal documentation allows precise description of the lexical semantics of a language. We will demonstrate both the user interface for documenting semantics, as well as a text-to-scene system for Arrernte created with WELT.



Scene: Scene 13

Creation Date: Sun Aug 04 12:57:43 EDT 2013

non	tɬaneʃti	kapilkatok	ipan	tɬapetʃ
thing	make.light	hanging	on	table

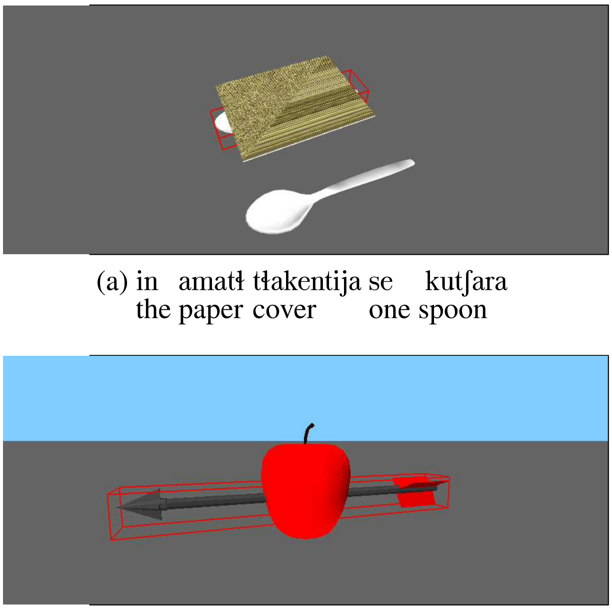
light = the thing make a light

Figure 1: WELT interface for annotating a scene

When a sentence is processed by WordsEye, it goes through three main stages: (1) morphological analysis and syntactic parsing, (2) semantic analysis, and (3) graphical realization. We will walk through these modules in the context of WELT, discussing (a) the formal documentation required for that component, (b) the processing of an example sentence through that component, and (c) the parts of that component that will feature in our demonstration. We will use the Arrernte sentence shown in (1) as a running example.

(1) artwe le goal arrerneme
man ERG goal put.nonpast
The man kicks a goal.

Morphology and Syntax WELT first parses a sentence into its morphology and syntax. Since the focus of WELT is documentation of semantics, the exact mechanisms for parsing the morphology and syntax may vary. To document Arrernte, we are using XFST (Karttunen et al., 1997) to model the morphology and XLE (Crouch et al., 2006) to model the syntax in the LFG formalism (Kaplan and Bresnan, 1982). These are mature systems that we believe are sufficient for the formal documentation of morphology and syntax. In future, we will provide interfaces to the third-party tools so that common information, like the lexicon, can



(a) in amatɬ tɬakentija se kutʃara
the paper cover one spoon

(b) in kwawitɬ tɬapanawi tɬakoja se mansana
the stick pass.thru in.middle one apple

Figure 2: Nahuatl examples elicited with WELT

be shared.

Running each word of the sentence through the morphological analyzer in XFST transforms the verb *arrerneme* into 'arrerne+NONPAST.' The other tokens in the sentence remain unchanged. Parsing the sentence with XLE gives the c-structure shown in Figure 3(a) and the f-structure shown in Figure 3(b). The f-structure will be passed on to the semantics module.

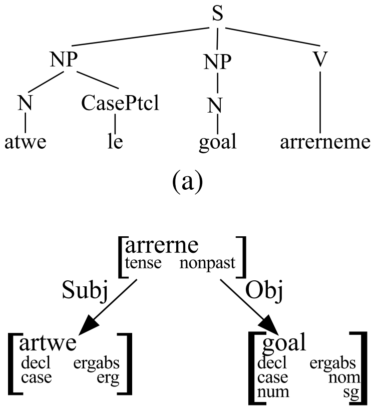


Figure 3: C-structure (a) and f-structure (b) for *artwe le goal arrerneme*.

We have added one additional feature to the morphology and syntax module of WELT's text-to-scene system: an interface for selecting an f-structure from multiple options produced by XLE, in case the grammar is ambiguous. This way, a linguist can use the WELT text-to-scene system to verify their semantic documentation even if the syntactic documentation is fairly rough. We will demonstrate this feature when demonstrating the Arrernte text-to-scene system.

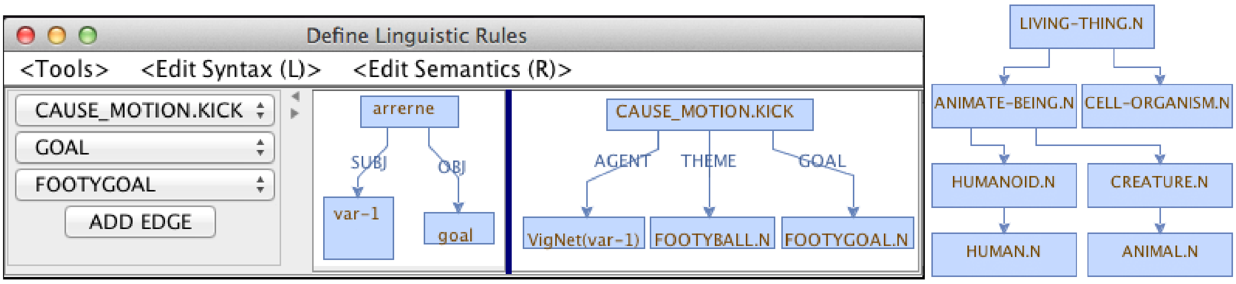
Semantics The WELT semantics is represented using VigNet, which has been developed for WordsEye based on English. We will assume that large parts of VigNet are language-independent (for instance, the set of low-level graphical relations used to express the graphical semantics is based on physics and human anatomy and does not depend on language). Therefore, it should not be necessary to create a completely new VigNet for every language that will be used in WELT. In future, we will develop tools for modifying VigNet to handle linguistic and cultural differences as they occur.

In order to use VigNet with other languages, we need to map between the formal syntax of the language being studied and the (English) lexical semantics required currently by VigNet. One instance showing why this is necessary occurs in our example Arrrente sentence. When discussing football in English, one would say that someone *kicks a goal* or *makes a goal*. In Arrente, one would say *goal arrerneme*, which translates literally to "put a goal." Although the semantics of both sentences are the same, the entry for "put" in the English VigNet does not include this meaning, but the Arrernte text-to-scene system needs to account for it.

To address such instances, we have created an interface for a linguist to specify a set of rules that map from syntax to semantics. The rules take syntactic f-structures as input and output a high-level semantic representation compatible with VigNet. The left-hand side of a rule consists of a set of conditions on the f-structure elements and the right-hand side consists of the semantic structure that should be returned. Figure 4(a) is an example of a rule mapping Arrernte syntax to semantics, created in WELT.

In addition to these rules, the linguist creates a simple table mapping lexical items into VigNet semantic concepts, so that nouns can be converted to graphical objects. We have created a mapping for the lexical items in the Arrernte grammar; a partial mapping is shown in Table 1.

We now describe the semantic processing of our example Arrernte sentence, assuming a set of rules consisting solely of the one in Figure 4(a) and the noun mapping in Table 1. The f-structure in Fig-



(a) (b)

Figure 4: Syntax-semantics rule (a) and semantic category browser (b) from WELT

Lexical Item	*artwe*	*panikane*	*angepe*	*akngwelye*	*apwerte*	*tipwele*
VigNet Concept	PERSON.N	CUP.N	CROW.N	DOG.N	ROCK-ITEM.N	TABLE.N

Table 1: A mapping from nouns (lexical items) to VigNet semantic concepts

ure 3(b) has main predicate *arrerne* with two arguments; the object is *goal*. Therefore, it matches the left-hand-side of our rule. The output of the rule specifies predicate CAUSE_MOTION.KICK with three arguments. The latter two are straightforward; the Theme is the VigNet object FOOTYBALL.N, and the Goal is FOOTYGOAL.N. To determine the Agent, we need to find the VigNet concept corresponding to `var-1`, which occupies the subject position in the f-structure. The subject in our f-structure is *artwe*, and according to Table 1, it maps to the VigNet concept PERSON.N. The resulting semantic representation is augmented with its graphical semantics, taken from the vignette for CAUSE_MOTION.KICK (vignette definition not shown for lack of space). The final representation is shown in Figure 5, with lexical semantics at the top and graphical semantics below. The WordsEye system then builds the scene from these constraints and renders it in 3D.

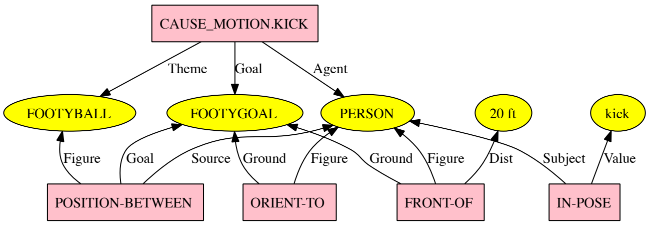


Figure 5: The semantics (lexical and graphical) for sentence (1)

WELT provides an interface for creating rules by defining the tree structures for the left-hand-side and right-hand-side of the rule. Every node on the left-hand-side can optionally contain boolean logic, if for example we want to allow the subject to be [(*artwe* 'man' `OR` *arhele* 'woman') `AND` `NOT` *ampe* 'child']; so rules can be as simple or complex as desired. Rules need not specify lexical items directly; it is also possible to refer to more general semantic categories. For example, a rule could select for all verbs of motion, or specify a particular constraint on the subject or object. In figure 4(a), for instance, we may want to only allow animate subjects.

Semantic categories are chosen through a browser that allows the user to search through all the semantic categories defined in VigNet. For example, if we want to find the semantic category to use as a constraint on our example subject, we might start by searching for *human*. This takes us to a portion of a tree of semantic concepts centered around HUMAN.N. The semantic categories are displayed one level at a time, so we initially see only the concepts directly above and directly below the word we searched for. From there, it's easy to select the concepts we are interested in, and go up or down the tree until we find the one we want. Below HUMAN.N are HUMAN-FEMALE.N and HUMAN-MALE.N, but we are more interested in the more general categories above the node. A screen shot showing the result of this search is shown in Figure 4(b). Above HUMAN.N is HUMANOID.N; above that, ANIMATE-BEING.N. Doing a quick check of further parents and children, we can see that for the subject of 'put goal,' we would probably want to choose ANIMATE-BEING.N over LIVING-THING.N.

The table mapping lexical items to VigNet concepts is built in a similar way; the lexicon is automatically extracted from the LFG grammar, and the user can search and browse semantic concepts to find the appropriate node for each lexical item.

We will demonstrate the WELT user interface which supports the creation of syntax-to-semantics rules, creates the mapping between nouns in the lexicon and VigNet concepts, and verifies the rules using the WELT text-to-scene system. We will show examples from our documentation of Arrernte and demonstrate entering text into the Arrernte text-to-scene system to generate pictures.

5 Summary and Future Work

We have described a novel tool for linguists working with endangered languages. It provides a new way to elicit data from informants, an interface for formally documenting the lexical semantics of a language, and allows the creation of a text-to-scene system for any language.

This project is in its early stages, so we are planning many additional features and improvements. For both modes of WELT, we want to generate pictures appropriate for the target culture. To handle this, we will add the ability to include custom objects and modify VigNet with new vignettes or new graphical semantics for existing vignettes. We also plan to build tools to import and export the work done in WELT in order to facilitate collaboration among linguists working on similar languages or cultures. Sharing sets of scenes will allow linguists to reuse work and avoid duplicated effort. Importing different versions of VigNet will make it easier to start out with WELT on a new language if it is similar to one that has already been studied. We might expect, for instance, that other Australian aboriginal languages will require the same kinds of cultural modifications to VigNet that we make for Arrernte, or that two languages in the same family might also have similar syntax to semantics rules.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1160700.

References

Alliance for Linguistic Diversity. 2013. The Endangered Languages Project. http://www.endangeredlanguages.com/.

C. Baker, J. Fillmore, and J. Lowe. 1998. The Berkeley FrameNet project. In *36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 86–90, Montréal.

C. Baker. 2008. FrameNet, present and future. In *The First International Conference on Global Interoperability for Language Resources*, pages 12–17.

E. Bender, D. Flickinger, and S. Oepen. 2002. The Grammar Matrix. In J. Carroll, N. Oostdijk, and R. Sutcliffe, editors, *Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

S. Bird. 2009. Natural language processing and linguistic fieldwork. *Computational Linguistics*, 35(3):469–474.

C. Black and H.A. Black. 2012. Grammars for the people, by the people, made easier using PAWS and XLingPaper. In Sebastian Nordoff, editor, *Electronic Grammaticography*, pages 103–128. University of Hawaii Press, Honolulu.

H.A. Black and G.F. Simons. 2006. The SIL FieldWorks Language Explorer approach to morphological parsing. In *Computational Linguistics for Less-studied Languages: Texas Linguistics Society 10*, Austin, TX, November.

M. Bowerman and E. Pederson. 1992. Topological relations picture series. In S. Levinson, editor, *Space stimuli kit 1.2*, page 51, Nijmegen. Max Planck Institute for Psycholinguistics.

M. Butt, H. Dyvik, T.H. King, H. Masuichi, and C. Rohrer. 2002. The parallel grammar project. In *2002 Workshop on Grammar Engineering and Evaluation - Volume 15*, COLING-GEE '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

B. Coyne and R. Sproat. 2001. WordsEye: An automatic text-to-scene conversion system. In *SIGGRAPH*.

B. Coyne, D. Bauer, and O. Rambow. 2011a. Vignet: Grounding language in graphics using frame semantics. In *ACL Workshop on Relational Models of Semantics (RELMS)*, Portland, OR.

B. Coyne, C. Schudel, M. Bitz, and J. Hirschberg. 2011b. Evaluating a text-to-scene generation system as an aid to literacy. In *SlaTE (Speech and Language Technology in Education) Workshop at Interspeech*, Venice.

D. Crouch, M. Dalrymple, R. Kaplan, T. King, J. Maxwell, and P. Newman, 2006. *XLE Documentation*. http://www2.parc.com/isl/groups/nltt/xle/doc/xle.

C. Filmore, C. Johnson, and M. Petruck. 2003. Background to FrameNet. In *International Journal of Lexicography*, pages 235–250.

R.M. Kaplan and J.W. Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In J.W. Bresnan, editor, *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, Mass., December.

L. Karttunen, T. Gaál, and A. Kempe. 1997. Xerox finite-state tool. Technical report, Xerox Research Centre Europe, Grenoble.

S. McConnel, 1995. *PC-PATR Reference Manual*. Summer Institute for Linguistics.