Grammar Engineering for Language Documentation

Emily M. Bender
University of Washington

Aarhus University
17 June 2014

Acknowledgments

- Joint work with: Joshua Crowgey, Michael Goodman, Fei Xia, Sumukh Ghodke, Tim Baldwin, Rebecca Dridan, Robert Schikowski, Balthasar Bickel

- This material is based upon work supported by the National Science Foundation under Grants No. 0644097 & No. BCS-1160274. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Introduction/Overview

- What is grammar engineering?
- What can grammar engineering do for language documentation?
- Case studies:
  - Lushootseed ([lut], Coast Salish, Salish, North America) morphophonology
  - Wambaya ([wmb], Mirndi, West Barkly, Australia) morphosyntax, treebanks
  - Chintang ([ctn], Kiranti, Tibeto-Burman, Nepal) automated grammar development from IGT

Computational tools for Linguistics

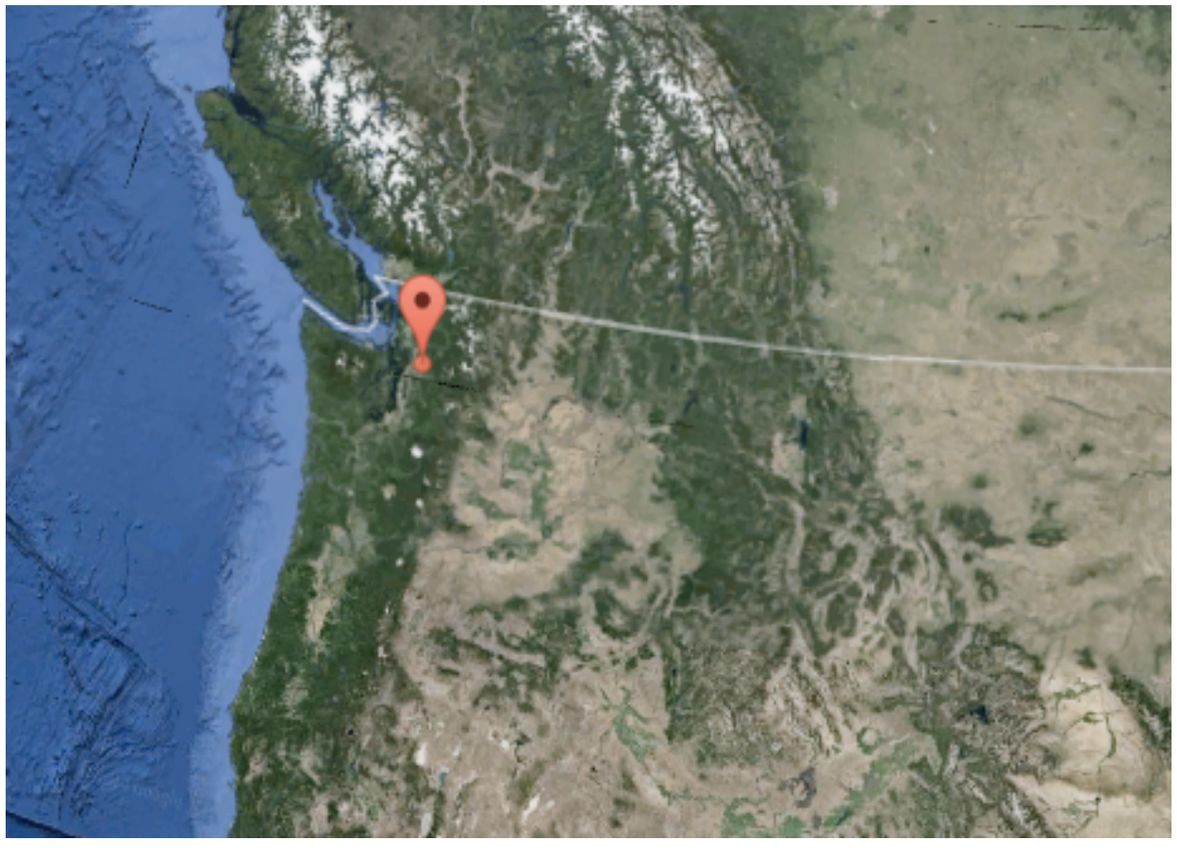
- MS Word
- Text editors
- Excel/spreadsheet software
- Local databases
- Web search
  - For examples
  - For resources (OLAC)
- Concordancer
- Praat
- ELAN
- Shoebox/Toolbox/FieldWorks
- POS tagger
- Morphological analyzer
- Syntactico-semantic parser

Grammar Engineering

- Creating machine-readable implementations of sets of linguistic rules
  - Phonological, morphological, syntactic and/or semantic
- Requires precise definitions of the rules: *formalized*, but not *formalist*
- Requires attention to interactions between the rules
- (Typically) separates declarative grammatical knowledge from procedural algorithms
- Facilitates testing of analyses: against test suites, against naturally occurring corpora
- Contrasts with (and complements): statistical modeling, hand annotation, pen-and-paper formalization

Case study #1: Lushootseed morphology (Work by Joshua Crowgey)

- Lushootseed [lut] is a Coast Salish language spoken in what is now Washington State, USA (Hess 1967)



Lushootseed morphology: Goal

- Create a morphophonological analyzer that relates surface forms (as transcribed) to segmented/regularized forms from Beck and Hess's IGT:

xʷiʔ	gʷəsəsaydubs	ʔə	tiʔəʔ	diʔəʔ
xʷiʔ	gʷə=s=ʔas-hay-dxʷ-b=s	ʔə	tiʔəʔ	diʔəʔ
NEG	SUBJ=NMLZ=STAT-known-CAUS-PASS=POSS.3SG	PREP	PROX	here
It is not known by the children.		Basket Ogress—ʔalatał Martin Sampson		

Lushootseed morhpology: tools

- Foma (Hulden 2009) implementation of xfst and lexc languages (Beesley & Karttunen 2003)

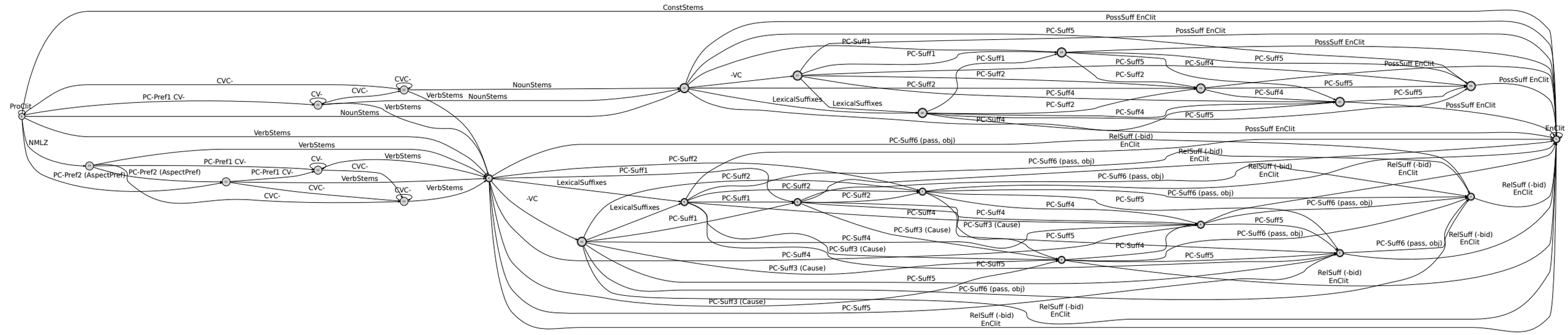
- lexc: morphotactics; define the set of possible underlying forms

- xfst: morphophonology: define the rules that relate underlying forms to surface forms

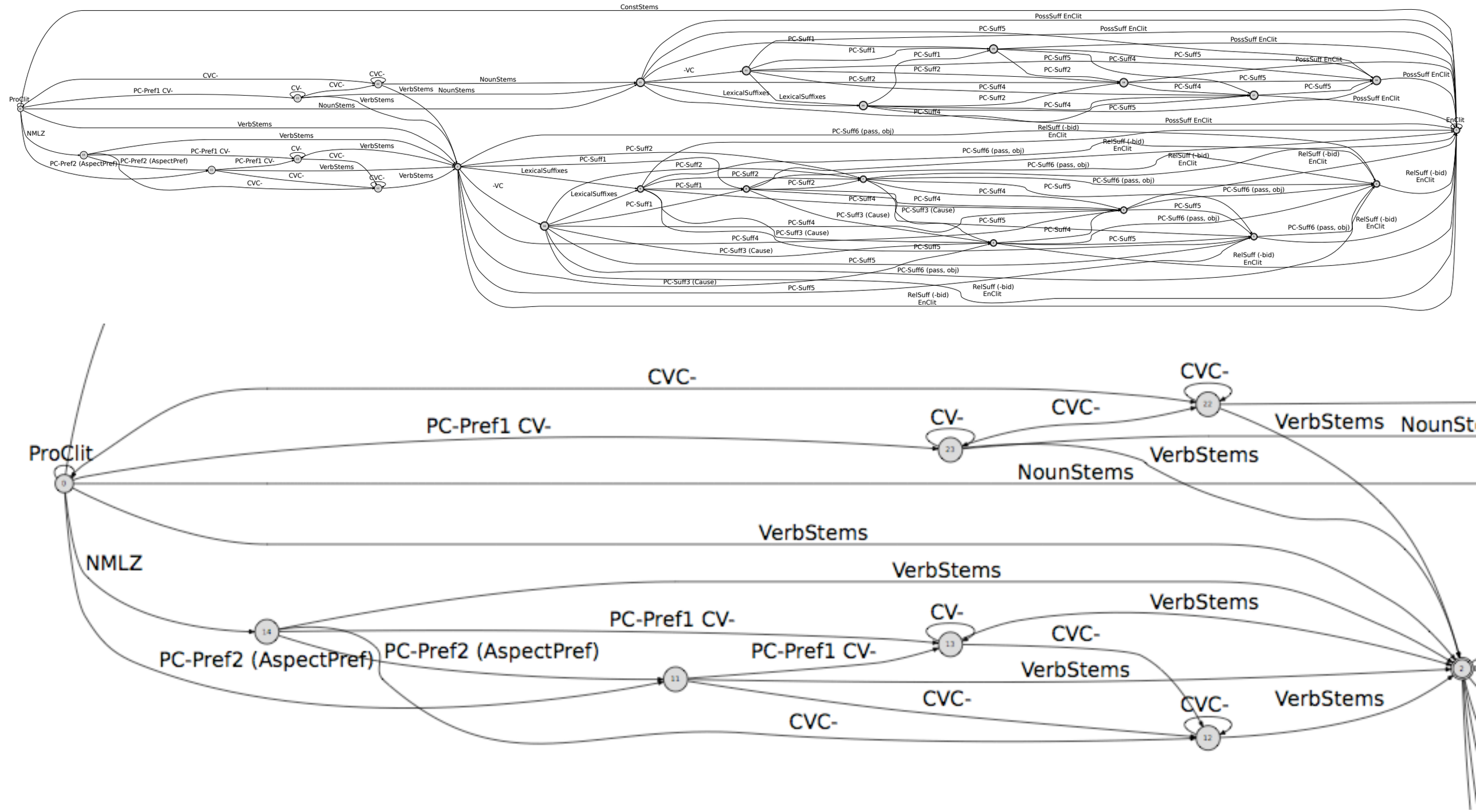
- lexc and xfst both define *finite state transducers* which can be composed into one complex (but efficient) transducer

- Both lexc and xfst use plain text declarations to define grammars, but these can be visualized as graphs

lexc: morphotactics



lexc: morphotactics



lexc: morphotactics

```
define PostRedV RED2* RED1* VerbStems (RED3);
define PostRedN RED2* RED1* NounStems (RED3);
define VPostpref1 (PCpref1) PostRedV (LSuff);
define NPostpref1 (PCpref1) PostRedN (LSuff);
define VPostAspect (AspectPref) VPostpref1;
define NPostsuff1 NPostpref1   (PCsuff1);
define VPostsuff1 VPostAspect (PCsuff1);
define NPostsuff2 NPostsuff1 (PCsuff2);
define VPostsuff2 VPostsuff1 (PCsuff2);
define VPostCausesuff1 VPostsuff2 (Causesuff1);
define VPostsuff4 VPostCausesuff1 (PCsuff4);
define NPostsuff4 NPostsuff2 (PCsuff4);
define VPostsuff5 VPostsuff4 (PCsuff5);
define NPostsuff5 NPostsuff4 (PCsuff5);
define VPostsuff6 VPostsuff5 (PCsuff6) (RelSuff);
define NPostsuff6 NPostsuff5 (poSuff);
define Iword NPostsuff6| (Nmlz) VPostsuff6;
define Word Iword | Const;
regex ProClit* Word EnClit*;
```

xfst: Phonological rules

xfst: Phonological rules

- Linguistic rule:

xfst: Phonological rules

- Linguistic rule: $u{=}? \rightarrow \varnothing \mid \{t, ɬ, ƛ\} \_\_\_ as\text{-}$

xfst: Phonological rules

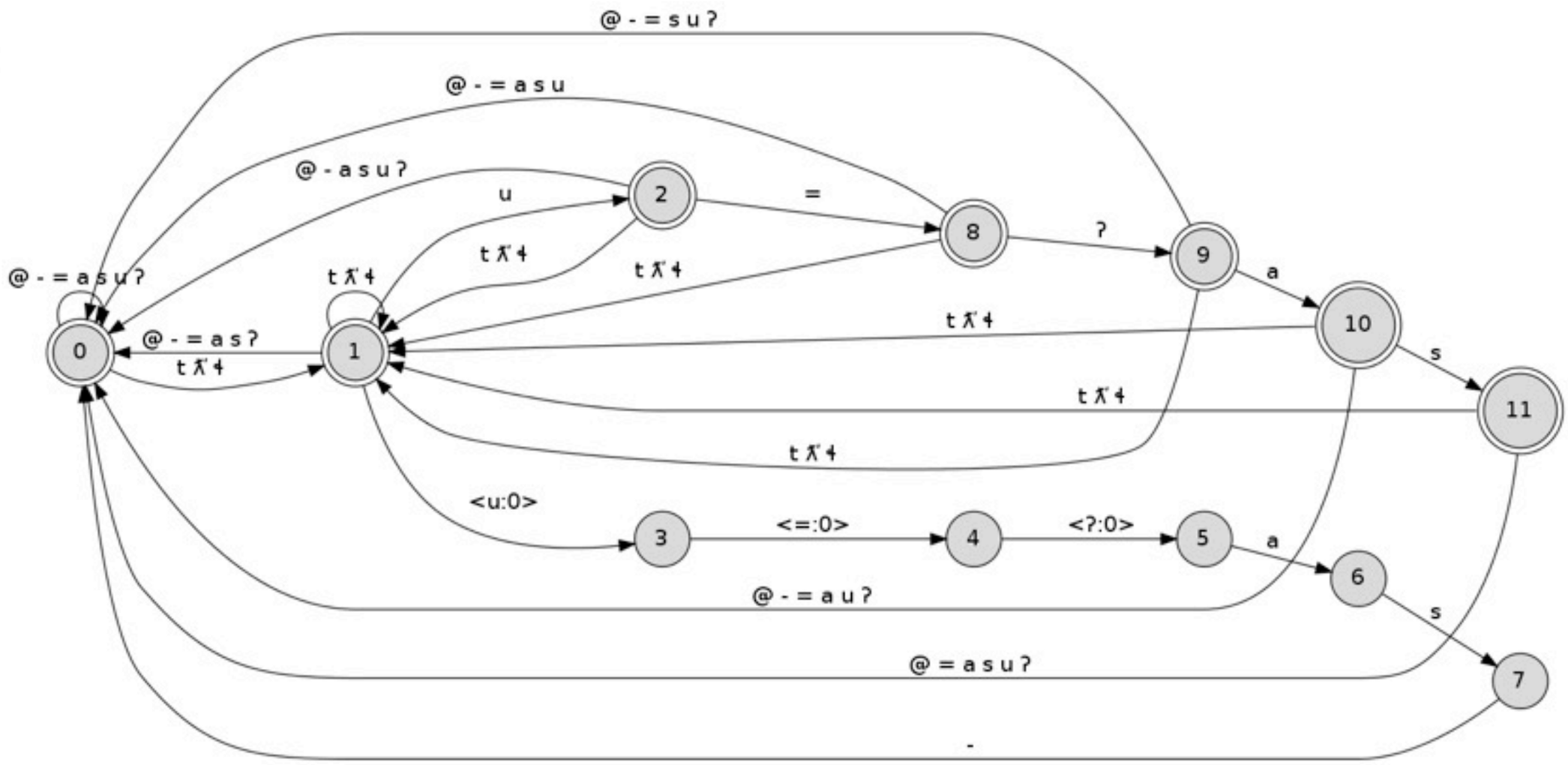
- Linguistic rule: $u{=}? \rightarrow \varnothing \mid \{t, ɬ, ƛ\} \_\_\_ as\text{-}$
- xfst source code: `u = ? -> 0 | [t|ƛ|ɬ] _ a s - ;`

xfst: Phonological rules

- Linguistic rule: $u{=}? \rightarrow \varnothing \mid \{t, ɬ, ƛ\} \_\_\_ as\text{-}$
- xfst source code: `u = ? -> 0 | [t|ƛ|ɬ] _ a s - ;`
- graph:

xfst: Phonological rules

- Linguistic rule: $\mathrm{u{=}?} \rightarrow \varnothing \mid \{\mathrm{t}, \mathrm{ɬ}, \mathrm{ƛ}\} \_\_\_ \mathrm{as\text{-}}$
- xfst source code: `u = ? -> 0 | [t|ƛ|ɬ] _ a s - ;`
- graph:



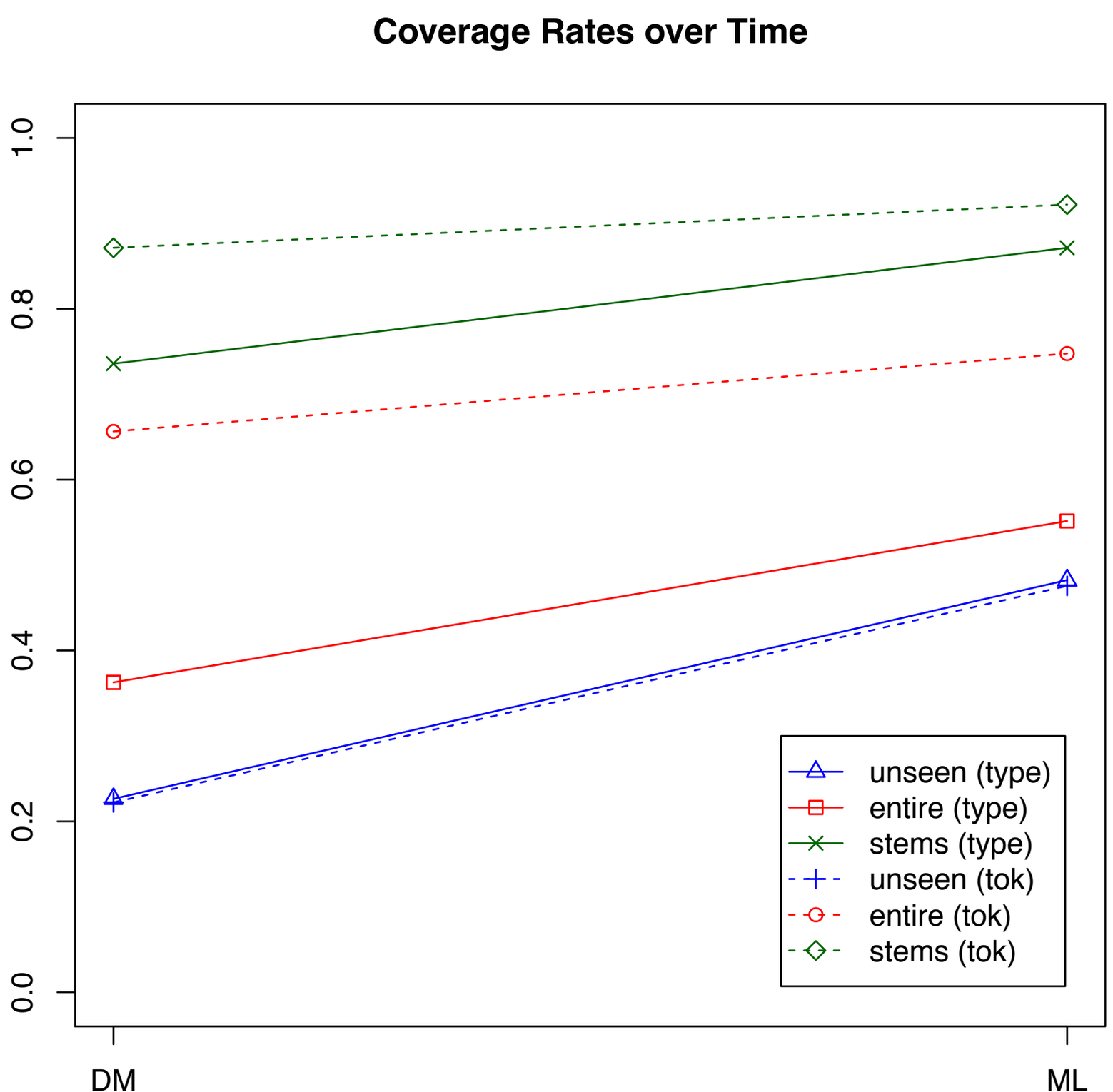
Lushootseed xfst: Development process and evaluation

- Development and evaluation data: Narratives from Beck and Hess (2014) *Tellings from Our Elders: Lushootseed syəyəhub.*

- Work with the narratives one at a time, until the FST has perfect coverage and no overgeneration:
    - for each surface form in the narrative, the FST should produce only legitimate underlying forms, including the one indicated
    - for each underlying form in the narrative, the FST should produce only legitimate surface forms, including the one indicated

- Evaluate that version of the FST with the next narrative, and then repeat

Lushootseed FST: Evaluation so far



How does this help language documentation?

- Verification of analyses
- Identification of errors in glossing
- (Eventually) more rapid glossing of new texts
- More rapid identification of “new” phenomena in new texts

Case study #2: Wambaya morphosyntax

- Wambaya [wmb] is a Mirndi language spoken in the West Barkly Tablelands region (Northern Territory) of Australia (Nordlinger 1998; Green & Nordlinger 2004)

Case study #2: Wambaya morphosyntax

- Wambaya [wmb] is a Mirndi language spoken in the West Barkly Tablelands region (Northern Territory) of Australia (Nordlinger 1998; Green & Nordlinger 2004)

Case study #2: Wambaya morphosyntax

- Wambaya [wmb] is a Mirndi language spoken in the West Barkly Tablelands region (Northern Territory) of Australia (Nordlinger 1998; Green & Nordlinger 2004)

Wambaya morphosyntax: Goal

- Create a grammar that analyses all 794 example sentences from Nordlinger 1998, associated each with an appropriate semantic representation

- Minimize ambiguity found by the grammar; especially avoiding unwarranted structures

- Test the grammar for generalizability on held-out text (narrative from Nordlinger 1998)

- Also: Test the applicability and usefulness of the LinGO Grammar Matrix (Bender et al 2002; Drellishak & Bender 2005)

Wambaya example: IGT

(Nordlinger 1998:223)

Wambaya example: IGT

(1) Ngaragana-nguja ngiy-a gujinganjanga-ni jiyawu
grog-PROP.IV.ACC 3.SG.NM.A-PST mother.II.ERG give
ngabulu.
milk.IV.ACC
‘(His) mother gave (him) milk with grog in it.’ [wmb]

(Nordlinger 1998:223)

Wambaya example: IGT

(Nordlinger 1998:223)

Wambaya example: semantic representation

$$\begin{bmatrix}
\text{LTOP} & \text{h1} \\
\text{INDEX} & \text{e2 (prop-or-ques, past)} \\
\text{RELS} & \left\langle \begin{bmatrix} \textbf{_grog_n_rel} \\ \text{LBL} \quad \text{h3} \\ \text{ARG0} \quad \text{x4 (3, iv)} \end{bmatrix}, \begin{bmatrix} \textbf{proprietive_a_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{e6} \\ \text{ARG1} \quad \text{x7 (3, iv)} \\ \text{ARG2} \quad \text{x4} \end{bmatrix}, \begin{bmatrix} \textbf{_mother_n_rel} \\ \text{LBL} \quad \text{h8} \\ \text{ARG0} \quad \text{x9 (3sg, ii)} \end{bmatrix}, \right. \\
 & \left. \begin{bmatrix} \textbf{_give_v_rel} \\ \text{LBL} \quad \text{h1} \\ \text{ARG0} \quad \text{e2} \\ \text{ARG1} \quad \text{x9} \\ \text{ARG2} \quad \text{x10 (3)} \\ \text{ARG3} \quad \text{x7} \end{bmatrix}, \begin{bmatrix} \textbf{_milk_n_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{x7} \end{bmatrix} \right\rangle \\
\text{HCONS} & \langle\ \rangle
\end{bmatrix}$$

Wambaya example: semantic representation

$$
\begin{bmatrix}
\text{LTOP} & \text{h1} \\
\text{INDEX} & \text{e2 (prop-or-ques, past)} \\
\text{RELS} & \left\langle
\begin{bmatrix} \textbf{\_grog\_n\_rel} \\ \text{LBL} \quad \text{h3} \\ \text{ARG0} \quad \text{x4 (3, iv)} \end{bmatrix},
\begin{bmatrix} \textbf{proprietive\_a\_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{e6} \\ \text{ARG1} \quad \text{x7 (3, iv)} \\ \text{ARG2} \quad \text{x4} \end{bmatrix},
\begin{bmatrix} \textbf{\_mother\_n\_rel} \\ \text{LBL} \quad \text{h8} \\ \text{ARG0} \quad \text{x9 (3sg, ii)} \end{bmatrix},
\begin{bmatrix} \textbf{\_give\_v\_rel} \\ \text{LBL} \quad \text{h1} \\ \text{ARG0} \quad \text{e2} \\ \text{ARG1} \quad \text{x9} \\ \text{ARG2} \quad \text{x10 (3)} \\ \text{ARG3} \quad \text{x7} \end{bmatrix},
\begin{bmatrix} \textbf{\_milk\_n\_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{x7} \end{bmatrix}
\right\rangle \\
\text{HCONS} & \langle\ \rangle
\end{bmatrix}
$$

Wambaya example: semantic representation

$$\begin{bmatrix}
\text{LTOP} & \text{h1} \\
\text{INDEX} & \text{e2 (prop-or-ques, past)} \\
\text{RELS} & \left\langle \begin{bmatrix} \textbf{_grog_n_rel} \\ \text{LBL} \quad \text{h3} \\ \text{ARG0} \quad \text{x4 (3, iv)} \end{bmatrix}, \begin{bmatrix} \textbf{proprietive_a_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{e6} \\ \text{ARG1} \quad \text{x7 (3, iv)} \\ \text{ARG2} \quad \text{x4} \end{bmatrix}, \begin{bmatrix} \textbf{_mother_n_rel} \\ \text{LBL} \quad \text{h8} \\ \text{ARG0} \quad \text{x9 (3sg, ii)} \end{bmatrix}, \right. \\
 & \left. \begin{bmatrix} \textbf{_give_v_rel} \\ \text{LBL} \quad \text{h1} \\ \text{ARG0} \quad \text{e2} \\ \text{ARG1} \quad \text{x9} \\ \text{ARG2} \quad \text{x10 (3)} \\ \text{ARG3} \quad \text{x7} \end{bmatrix}, \begin{bmatrix} \textbf{_milk_n_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{x7} \end{bmatrix} \right\rangle \\
\text{HCONS} & \langle\ \rangle
\end{bmatrix}$$

Wambaya example: semantic representation

$$\begin{bmatrix}
\text{LTOP} & \text{h1} \\
\text{INDEX} & \text{e2 (prop-or-ques, past)} \\
\text{RELS} & \left\langle \begin{bmatrix} \textbf{_grog_n_rel} \\ \text{LBL} \quad \text{h3} \\ \text{ARG0} \quad \text{x4 (3, iv)} \end{bmatrix}, \begin{bmatrix} \textbf{proprietive_a_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{e6} \\ \text{ARG1} \quad \text{x7 (3, iv)} \\ \text{ARG2} \quad \text{x4} \end{bmatrix}, \begin{bmatrix} \textbf{_mother_n_rel} \\ \text{LBL} \quad \text{h8} \\ \text{ARG0} \quad \text{x9 (3sg, ii)} \end{bmatrix}, \right. \\
 & \left. \begin{bmatrix} \textbf{_give_v_rel} \\ \text{LBL} \quad \text{h1} \\ \text{ARG0} \quad \text{e2} \\ \text{ARG1} \quad \text{x9} \\ \text{ARG2} \quad \text{x10 (3)} \\ \text{ARG3} \quad \text{x7} \end{bmatrix}, \begin{bmatrix} \textbf{_milk_n_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{x7} \end{bmatrix} \right\rangle \\
\text{HCONS} & \langle\ \rangle
\end{bmatrix}$$

Wambaya example: semantic representation

$$\begin{bmatrix}
\text{LTOP} & \text{h1} \\
\text{INDEX} & \text{e2 (prop-or-ques, past)} \\
\text{RELS} & \left\langle \begin{bmatrix} \textbf{_grog_n_rel} \\ \text{LBL} \quad \text{h3} \\ \text{ARG0} \quad \text{x4 (3, iv)} \end{bmatrix}, \begin{bmatrix} \textbf{proprietive_a_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{e6} \\ \text{ARG1} \quad \text{x7 (3, iv)} \\ \text{ARG2} \quad \text{x4} \end{bmatrix}, \begin{bmatrix} \textbf{_mother_n_rel} \\ \text{LBL} \quad \text{h8} \\ \text{ARG0} \quad \text{x9 (3sg, ii)} \end{bmatrix}, \right. \\
 & \left. \begin{bmatrix} \textbf{_give_v_rel} \\ \text{LBL} \quad \text{h1} \\ \text{ARG0} \quad \text{e2} \\ \text{ARG1} \quad \text{x9} \\ \text{ARG2} \quad \text{x10 (3)} \\ \text{ARG3} \quad \text{x7} \end{bmatrix}, \begin{bmatrix} \textbf{_milk_n_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{x7} \end{bmatrix} \right\rangle \\
\text{HCONS} & \langle\ \rangle
\end{bmatrix}$$

Wambaya example: semantic representation

$$\begin{bmatrix}
\text{LTOP} & \text{h1} \\
\text{INDEX} & \text{e2 (prop-or-ques, past)} \\
\text{RELS} & \left\langle \begin{bmatrix} \textbf{_grog_n_rel} \\ \text{LBL} \quad \text{h3} \\ \text{ARG0} \quad \text{x4 (3, iv)} \end{bmatrix}, \begin{bmatrix} \textbf{proprietive_a_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{e6} \\ \text{ARG1} \quad \text{x7 (3, iv)} \\ \text{ARG2} \quad \text{x4} \end{bmatrix}, \begin{bmatrix} \textbf{_mother_n_rel} \\ \text{LBL} \quad \text{h8} \\ \text{ARG0} \quad \text{x9 (3sg, ii)} \end{bmatrix}, \right. \\
 & \left. \begin{bmatrix} \textbf{_give_v_rel} \\ \text{LBL} \quad \text{h1} \\ \text{ARG0} \quad \text{e2} \\ \text{ARG1} \quad \text{x9} \\ \text{ARG2} \quad \text{x10 (3)} \\ \text{ARG3} \quad \text{x7} \end{bmatrix}, \begin{bmatrix} \textbf{_milk_n_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{x7} \end{bmatrix} \right\rangle \\
\text{HCONS} & \langle\ \rangle
\end{bmatrix}$$

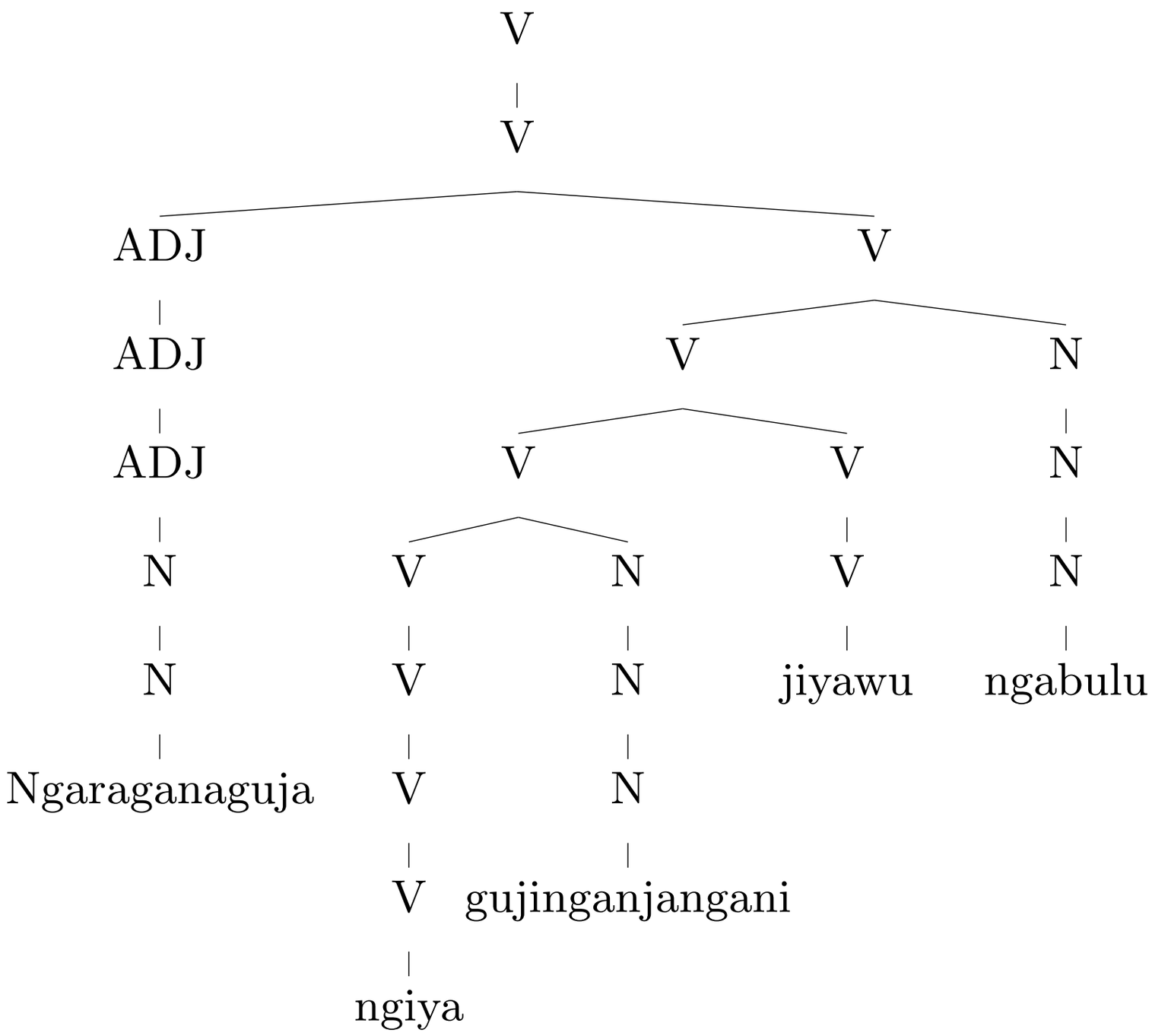
Wambaya example: semantic representation

$$\begin{bmatrix}
\text{LTOP} & \text{h1} \\
\text{INDEX} & \text{e2 (prop-or-ques, past)} \\
\text{RELS} & \left\langle \begin{bmatrix} \textbf{_grog_n_rel} \\ \text{LBL} \quad \text{h3} \\ \text{ARG0} \quad \text{x4 (3, iv)} \end{bmatrix}, \begin{bmatrix} \textbf{proprietive_a_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{e6} \\ \text{ARG1} \quad \text{x7 (3, iv)} \\ \text{ARG2} \quad \text{x4} \end{bmatrix}, \begin{bmatrix} \textbf{_mother_n_rel} \\ \text{LBL} \quad \text{h8} \\ \text{ARG0} \quad \text{x9 (3sg, ii)} \end{bmatrix}, \right. \\
 & \left. \begin{bmatrix} \textbf{_give_v_rel} \\ \text{LBL} \quad \text{h1} \\ \text{ARG0} \quad \text{e2} \\ \text{ARG1} \quad \text{x9} \\ \text{ARG2} \quad \text{x10 (3)} \\ \text{ARG3} \quad \text{x7} \end{bmatrix}, \begin{bmatrix} \textbf{_milk_n_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{x7} \end{bmatrix} \right\rangle \\
\text{HCONS} & \langle\ \rangle
\end{bmatrix}$$

Wambaya example: semantic representation

$$
\begin{bmatrix}
\text{LTOP} & \text{h1} \\
\text{INDEX} & \text{e2 (prop-or-ques, past)} \\
\text{RELS} & \left\langle
\begin{bmatrix} \textbf{\_grog\_n\_rel} \\ \text{LBL} \quad \text{h3} \\ \text{ARG0} \quad \text{x4 (3, iv)} \end{bmatrix},
\begin{bmatrix} \textbf{proprietive\_a\_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{e6} \\ \text{ARG1} \quad \text{x7 (3, iv)} \\ \text{ARG2} \quad \text{x4} \end{bmatrix},
\begin{bmatrix} \textbf{\_mother\_n\_rel} \\ \text{LBL} \quad \text{h8} \\ \text{ARG0} \quad \text{x9 (3sg, ii)} \end{bmatrix},
\begin{bmatrix} \textbf{\_give\_v\_rel} \\ \text{LBL} \quad \text{h1} \\ \text{ARG0} \quad \text{e2} \\ \text{ARG1} \quad \text{x9} \\ \text{ARG2} \quad \text{x10 (3)} \\ \text{ARG3} \quad \text{x7} \end{bmatrix},
\begin{bmatrix} \textbf{\_milk\_n\_rel} \\ \text{LBL} \quad \text{h5} \\ \text{ARG0} \quad \text{x7} \end{bmatrix}
\right\rangle \\
\text{HCONS} & \langle\ \rangle
\end{bmatrix}
$$

Wambaya example: Syntax tree

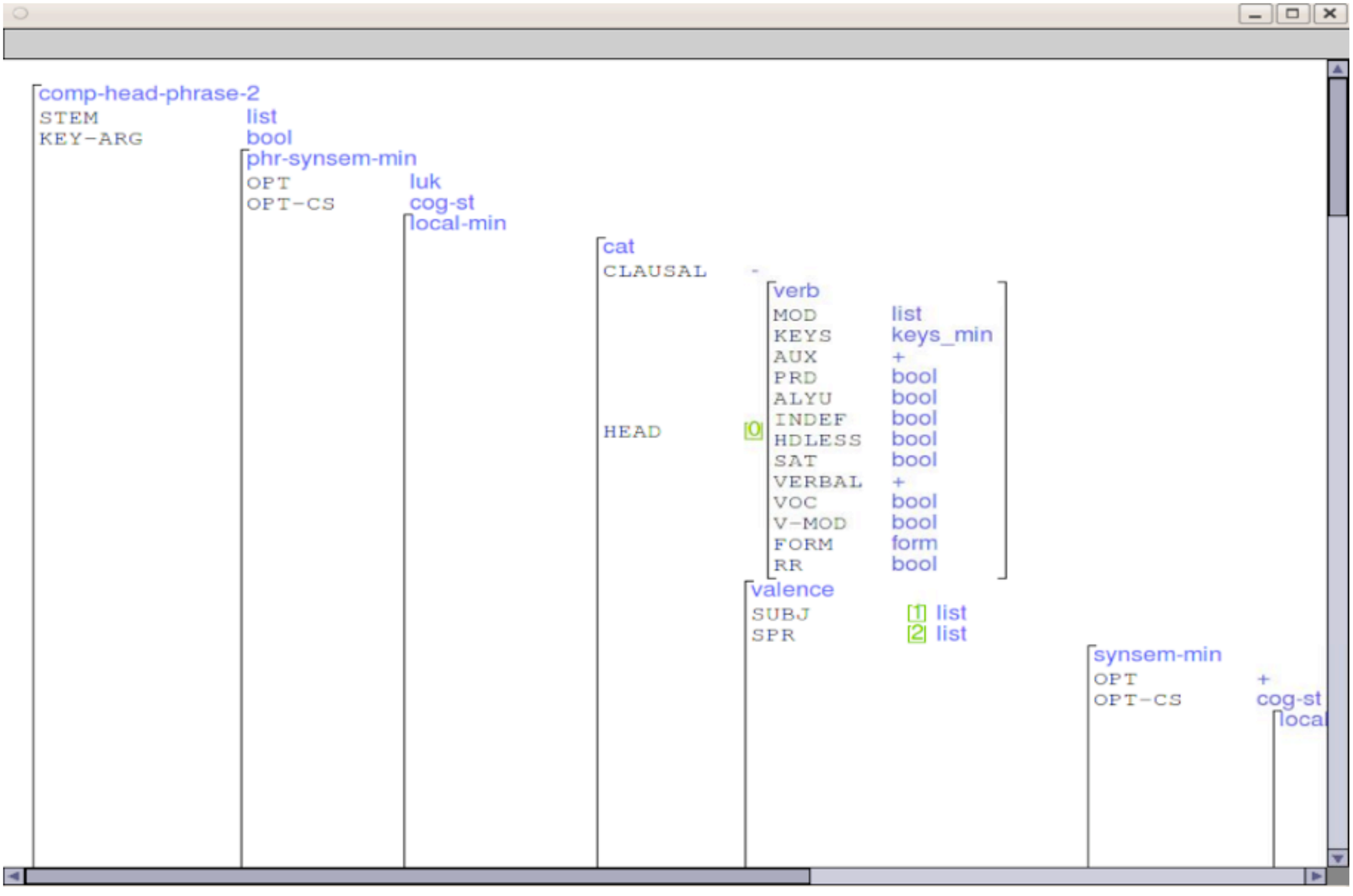


Definition of a grammar rule

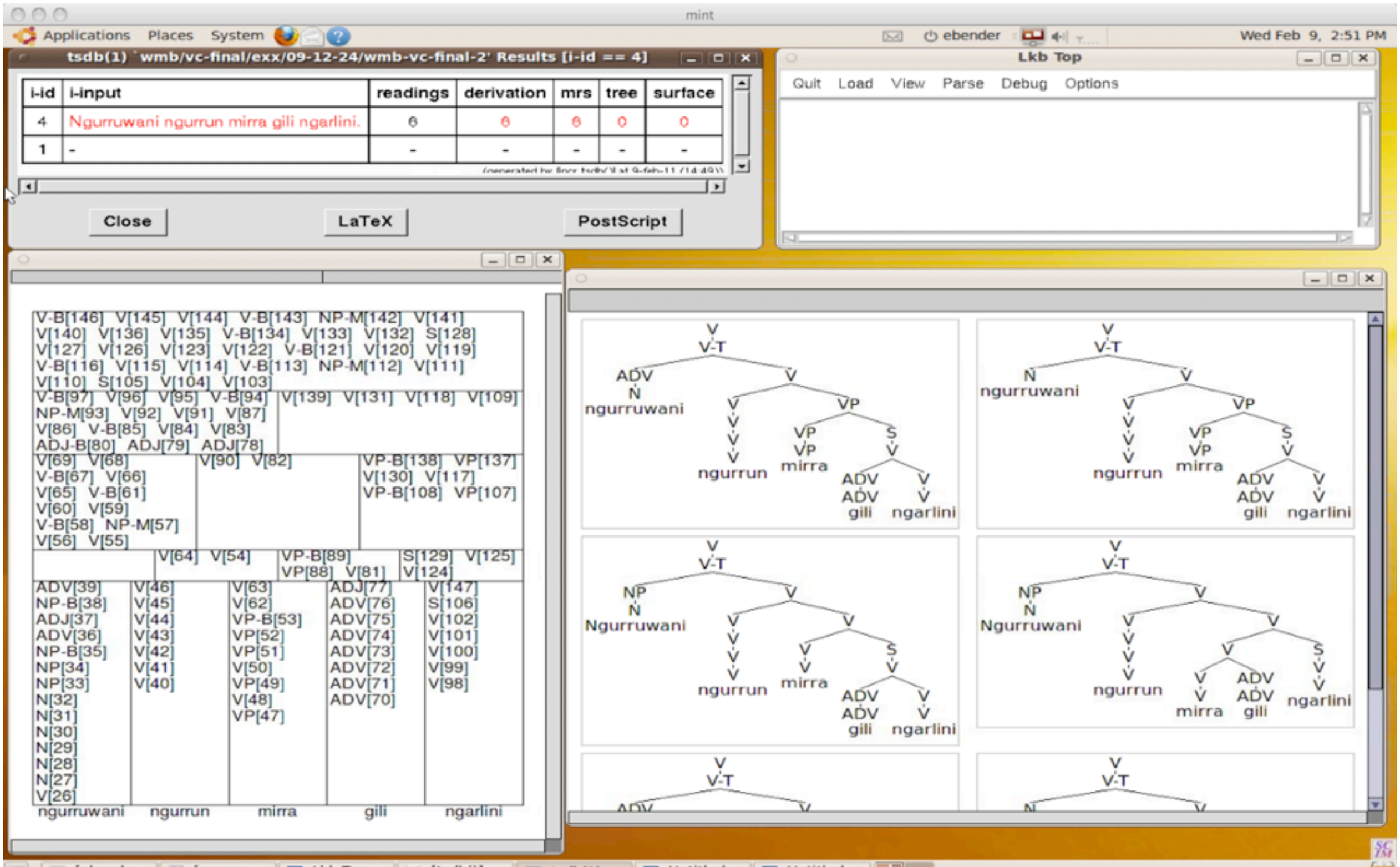
```
wmb-head-2nd-comp-phrase := non-1st-comp-phrase &
  [ SYNSEM.LOCAL.CAT.VAL.COMPS [ FIRST #firstcomp,
                REST [ FIRST [ OPT +,
                               INST +,
                               LOCAL #local,
                               NON-LOCAL #non-local ],
                       REST #othercomps ]],
    HEAD-DTR.SYNSEM.LOCAL.CAT.VAL.COMPS [ FIRST #firstcomp,
                     REST [ FIRST #synsem &
                                  [ INST -,
                                    LOCAL #local,
                                    NON-LOCAL #non-local ],
                            REST #othercomps ]],
    NON-HEAD-DTR.SYNSEM #synsem ].


head-comp-phrase-2 := wmb-head-2nd-comp-phrase & head-arg-phrase.
comp-head-phrase-2 := wmb-head-2nd-comp-phrase & verbal-head-final-
                      head-nexus.
```

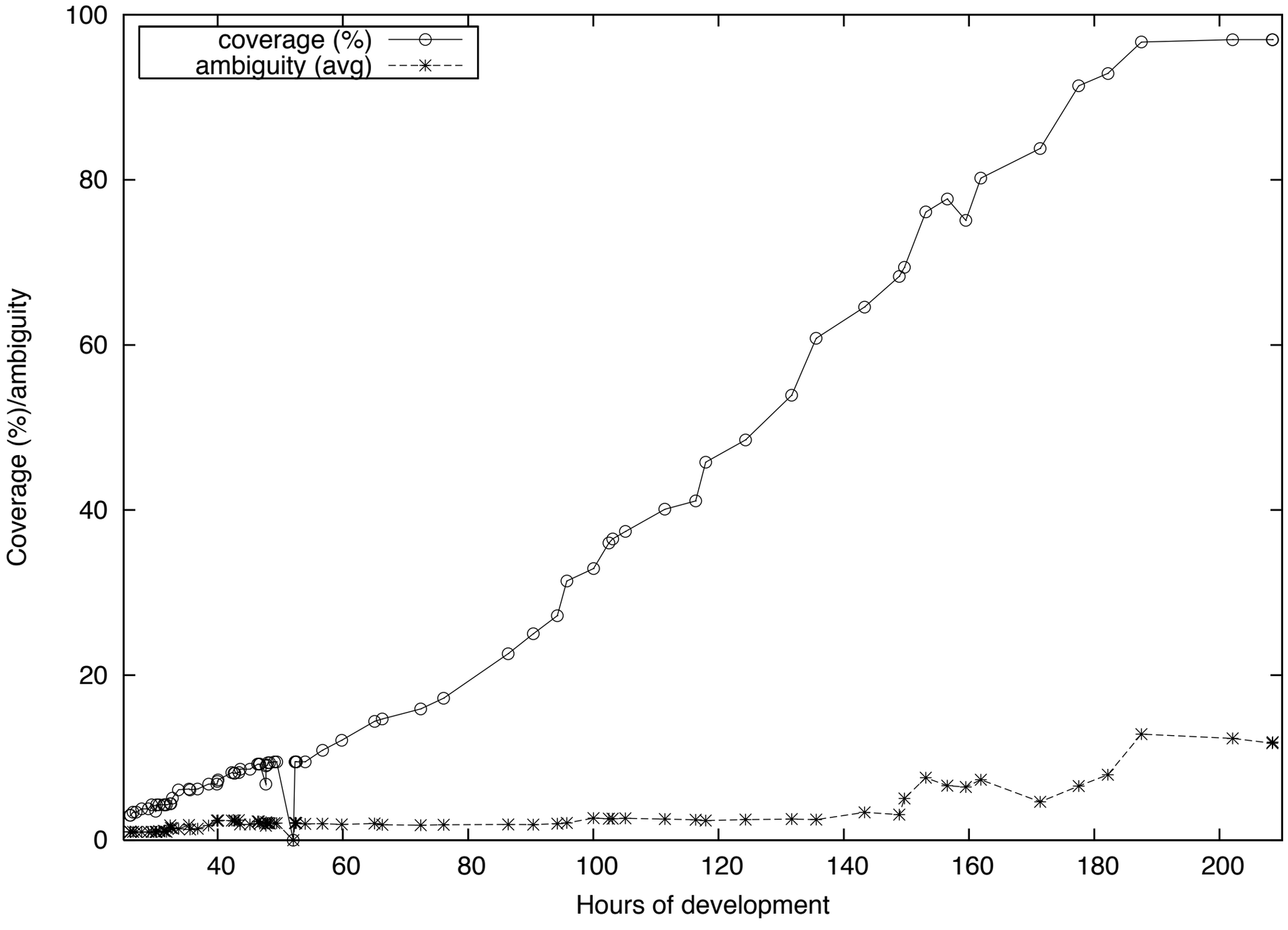
Inspecting a Grammar Rule



A Grammar Rule in Action



Wambaya grammar development



(Bender 2008)

Wambaya grammar evaluation (Bender 2008)

- Held out test data “The two Eaglehawks”
- 72 sentences (orig text: 92, removed 20 seen sentences)
- Run twice: before and after adding lexical entries and adjusting morphophonology only

	correct	parsed incorrect	unparsed	average ambiguity
Existing vocab	50%	8%	42%	10.62
w/added vocab	**76%**	8%	14%	12.56

Wambaya grammar parse selection

- Redwoods parse selection technology (Toutanova et al 2005) with Velldal's (2007) feature set

- Feature selection through 10-fold cross-validation on dev set

- Trained on parsed portion of dev set (732 items; 544 ambiguous)

- Test set results:

	Exact match
Random baseline	18.4%
Trained model	75.0%

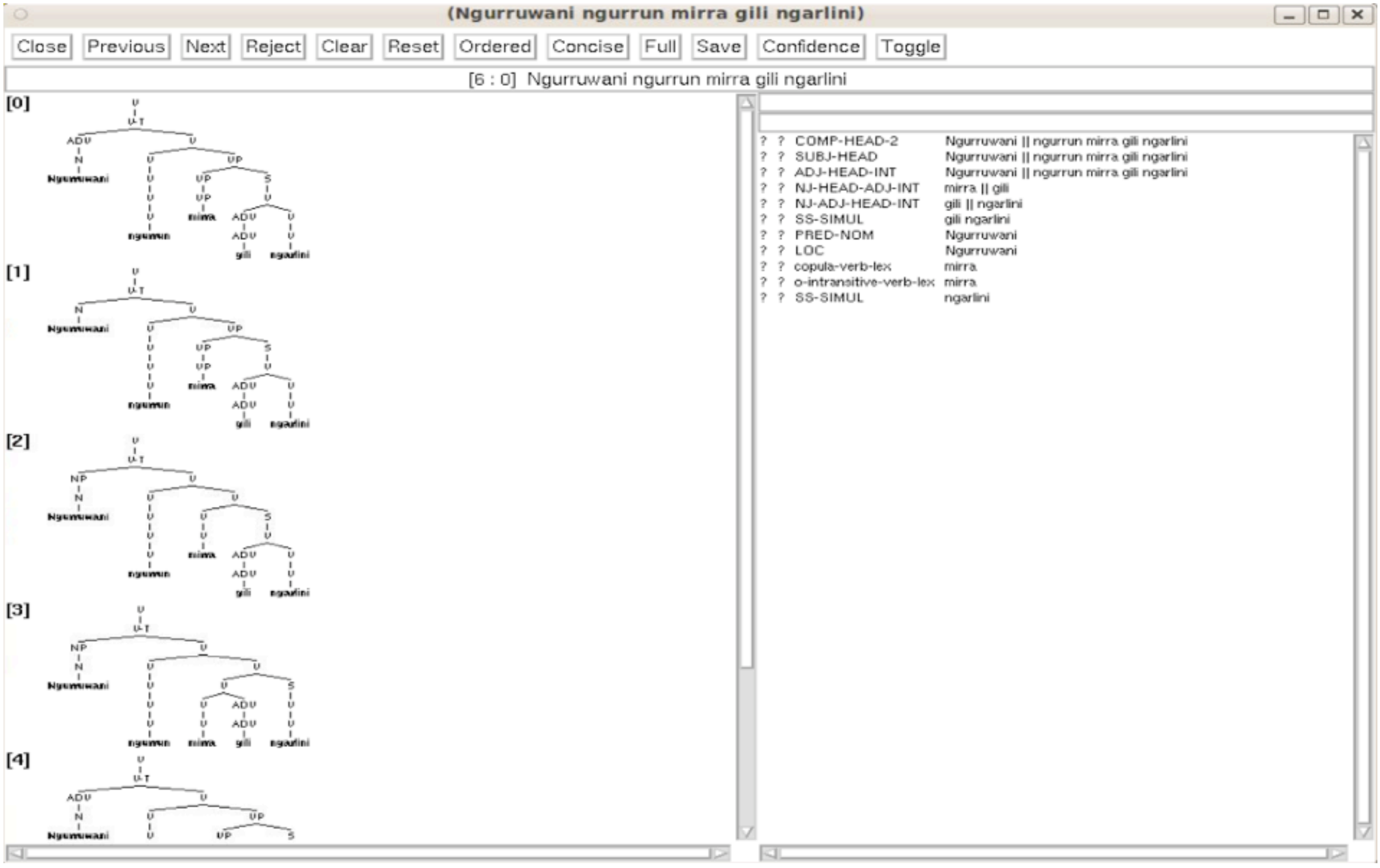
How does this help language documentation?

- Verification of analyses
- Identification of unhandled phenomena
- Processing of new text
- Treebanks

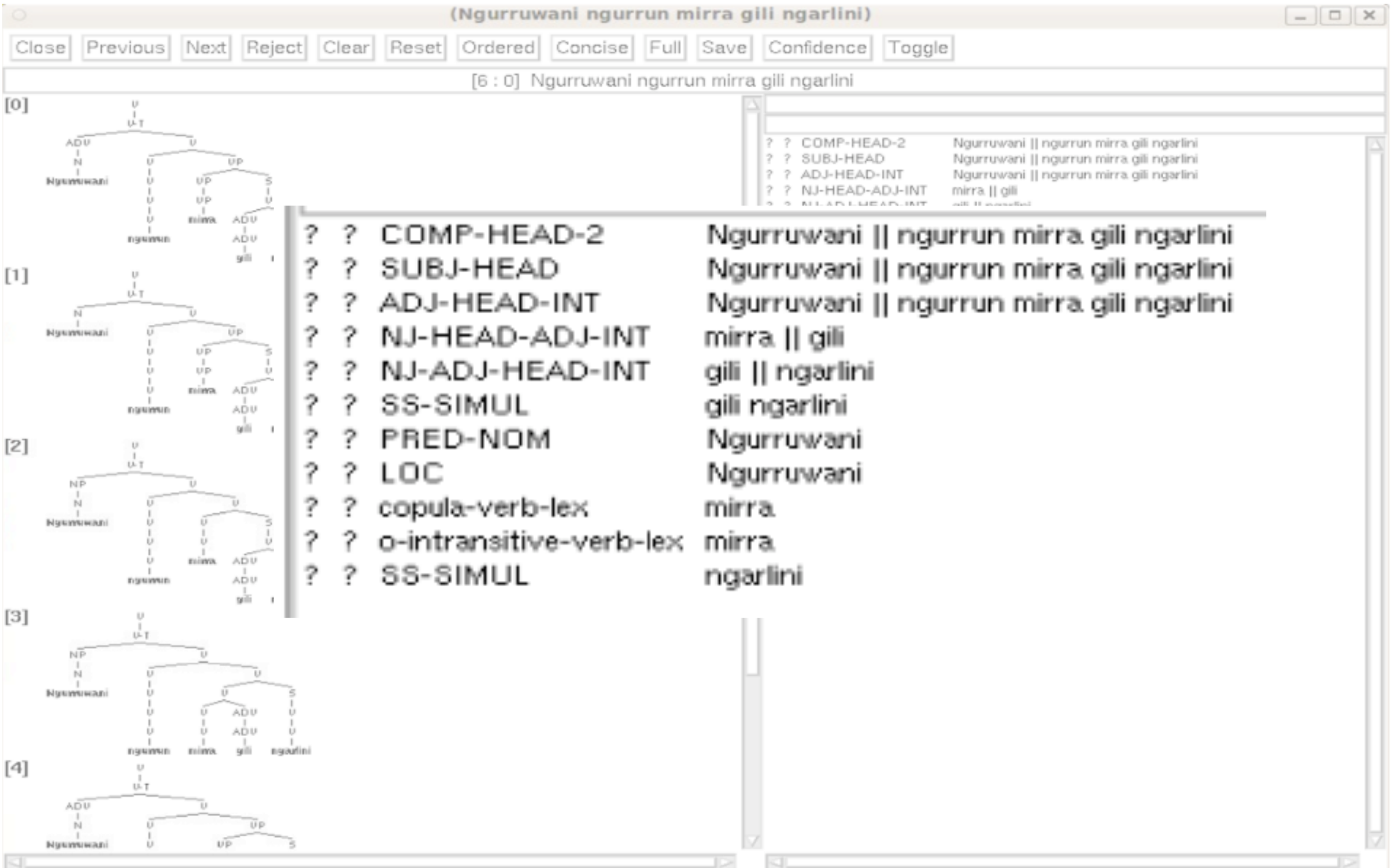
Treebanks

- Old-style (e.g., Penn Treebank, Marcus et al 1993): Develop extensive code book and hand-annotate tree structures for each item.
- New-style (e.g., Redwoods, Oepen et al 2004):
    - Process all items (typically utterances or sentences) with grammar
        - Select intended structure from among those provided by the grammar for each item --- assisted by calculation of discriminants
        - Indicate items with no correct analysis
        - Save decisions to rerun when grammar is updated
- Internally consistent treebanks, which can be updated easily as grammar is improved.

Redwoods Treebanking Tool



Redwoods Treebanking Tool



What Are Treebanks Good For?

- In Computational Linguistics:
    - Training parse-ranking models and other applications of machine learning
- In Language Description:
    - a set of searchable annotations
    - more detailed than IGT
    - more easily kept internally consistent than IGT
    - ... by no means a replacement for IGT!

Treebank Search (Ghodke and Bird 2010)

- Fast queries over large treebanks, including both PTB-style and Redwoods-style

- Sample query over Wambaya data:

  - Find sentences with a complement realized only by a modifier:

```
//DECL[//HEAD-COMP-MOD-2 AND NOT //HEAD-COMP-2
AND NOT //COMP-HEAD-2]
```

  - Find sentences with two overt arguments:

```
//DECL[//J-STRICT-TRANS-VERB-LEX AND
//HEAD-COMP-2 AND //HEAD-SUBJ]
```

Case study #3 Partially automated grammar development for Chintang

- Chintang [ctn] is a Kiranti (Tibeto-Burman) language spoken in Nepal (Bickel et al 2009)

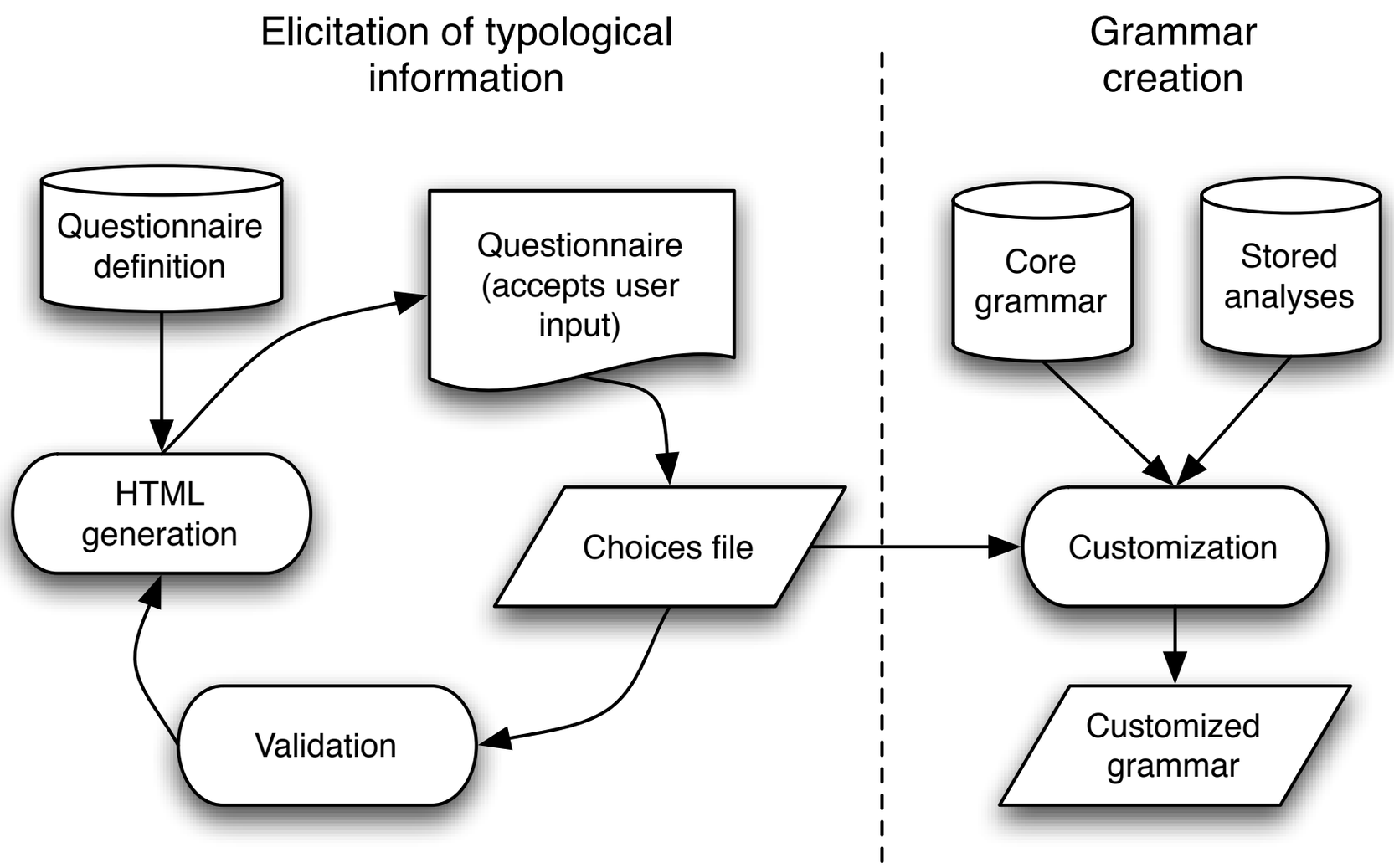
Partially automated grammar engineering: Overview

- LinGO Grammar Matrix, Grammar Matrix customization system
- ToolBox import
- Learning grammar specifications from IGT

The Grammar Matrix

- A repository of implemented analyses, including:
    - A core grammar with analyses of general patterns such as semantic compositionality
    - “Libraries” of analyses of cross-linguistically variable phenomena
    - Accessible via a web-based questionnaire
    - Produces working HPSG (Pollard & Sag 1994) grammars with Minimal Recursion Semantics (Copestake et al 2005), to spec

The Grammar Matrix



The Grammar Matrix:
http://www.delph-in.net/matrix

- * General Information
- * Word Order
- Number
- * Person
- Gender
- * Case
- Direct-inverse
- Tense, Aspect and Mood
- Other Features
- Sentential Negation
- Coordination
- Matrix Yes/No Questions
- Information Structure
- Argument Optionality
- ? Lexicon
- Morphology
- Import Toolbox Lexicon
- Test Sentences
- Test by Generation Options

Archive type: (•) .tar.gz ( ) .zip

Create Grammar | Test by Generation

View Choices File (right-click to download)

Upload Choices File: | Choose File | No file chosen

Implementation of Chintang Morphology (Bender et al 2012)

- Chintang boasts a fairly complex morphological system:

4 prefix slots + STEM + 12 suffix slots

- But then:
  - Prefix slots can freely reorder (Bickel et al 2007)
  - Prefix slots include one for ‘bipartite stems’
  - A single word can contain up to four verb roots, each hosting prefixes and suffixes
  - Complex morphophonology

Implementation of Chintang Morphology

- Focus on morpheme-segmented line of IGT, abstracting away from morphophonology

- 160 verbal lexical rules grouped into 54 position classes
    - Including duplicated prefix/suffix rules to handle the multi-root verbs

- 24 nominal lexical rules grouped into 6 position classes

- Elsewhere in the choices file, define values for case, tense/aspect/mood, person/number/gender

    - => Partial modeling of the syntactico-semantic effects of 131/160 rules

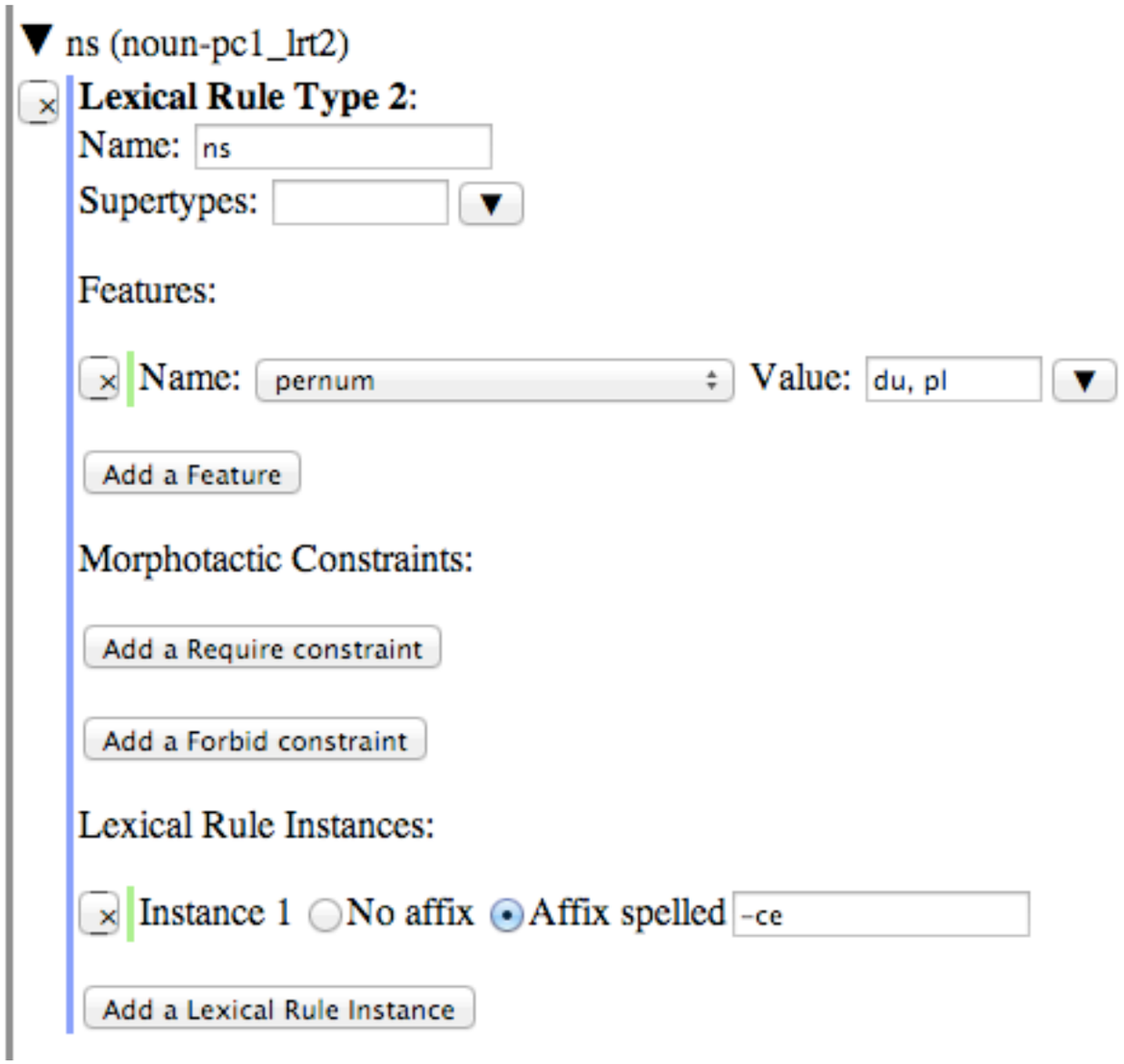
Chintang example

- Example test item (target is second line):

thupro	wassace	uyuwakte	pho
thupro	wassak-ce	u-yuŋ-a-yakt-e	pho
many	bird-NS	3NSS/A-live-PST-IPFV-IND.PST	REP

‘There lived many birds.’ [ctn] story_rabbit.005

Chintang morphosyntax: Sample noun lex rule



Chintang morphosyntax: Verb position classes

Verb Inflection

- neg-prefix (verb-pc1)
- 1st-person-object (verb-pc2)
- non-1st-subj (verb-pc3)
- act-ptcp (verb-pc4)
- suffix1 (verb-pc5)
- suffix2 (verb-pc6)
- suffix3 (verb-pc7)
- suffix4 (verb-pc8)
- suffix5 (verb-pc9)
- suffix6 (verb-pc10)
- suffix7 (verb-pc11)
- suffix8 (verb-pc12)
- suffix9 (verb-pc13)
- suffix10 (verb-pc14)
- suffix11 (verb-pc15)
- suffix12 (verb-pc16)
- suffix13 (verb-pc17)
- suffix14 (verb-pc18)
- suffix15 (verb-pc19)
- ?bps-prefix (verb-pc20)
- v2 (verb-pc21)
- suffix1-final (verb-pc22)
- suffix2-final (verb-pc23)
- suffix3-final (verb-pc24)
- suffix4-final (verb-pc25)
- suffix5-final (verb-pc26)
- suffix6-final (verb-pc27)
- suffix7-final (verb-pc28)
- suffix8-final (verb-pc29)
- suffix9-final (verb-pc30)
- suffix10-final (verb-pc31)
- suffix1-v2 (verb-pc32)
- suffix2-v2 (verb-pc33)
- suffix3-v2 (verb-pc34)
- suffix4-v2 (verb-pc35)
- suffix5-v2 (verb-pc36)
- suffix6-v2 (verb-pc37)
- suffix7-v2 (verb-pc38)
- suffix8-v2 (verb-pc39)
- suffix9-v2 (verb-pc40)
- suffix10-v2 (verb-pc41)
- v3 (verb-pc42)
- ntvz (verb-pc43)
- endoclitics (verb-pc44)
- act-ptcp-v2 (verb-pc45)
- non-1st-subj-v2 (verb-pc46)
- 1st-person-object-v2 (verb-pc47)
- neg-prefix-v2 (verb-pc48)
- endoclitics-v2 (verb-pc49)
- act-ptcp-v3 (verb-pc50)
- non-1st-subj-v3 (verb-pc51)
- 1st-person-object-v3 (verb-pc52)
- neg-prefix-v3 (verb-pc53)
- endoclitics-v3 (verb-pc54)

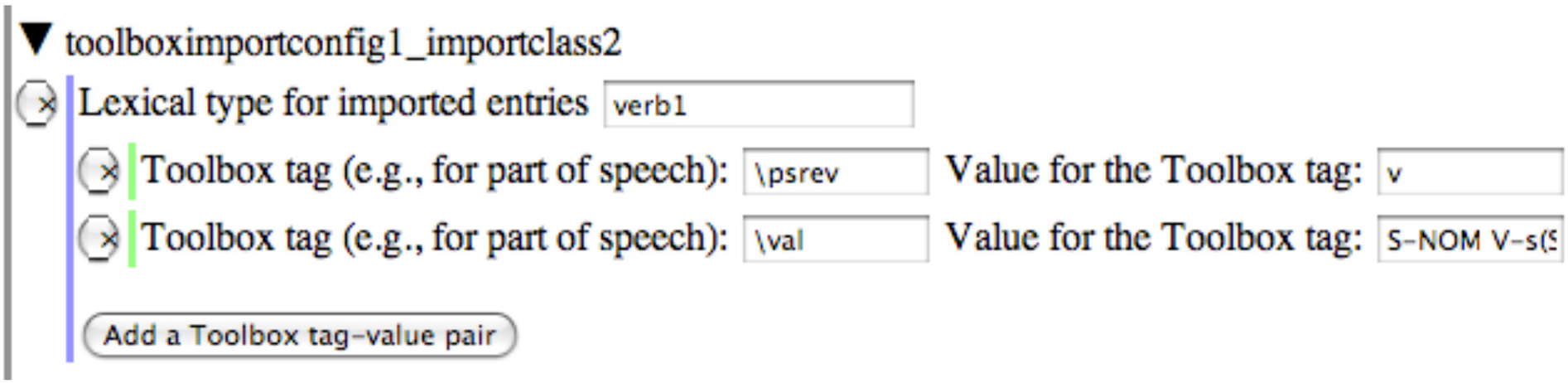
Reusing resources: Importing from Toolbox lexicon

- CPDP has a large Toolbox lexicon for Chintang
- stem forms
- alternate forms
- coarse-grained POS tag (drawn from a set of 30)
- syntactic valence (case for each argument)
- semantic valence (semantic role for each argument)
    - NB: valence information stored as a string

Chintang Toolbox lexicon

```
\lex kond
\id 179
\psrev v
\val A-ERG P-NOM V-a(A).o(P)
\ge search; look.for
\dt 22/Feb/2011
```

Import from Toolbox to Grammar Matrix



- verb1 is target type (defined as intransitive verb with normal case frame)
- \psrev is POS field; value here is ‘v’
- \val is valence field; value here is ‘S-NOM V-s(S)’
- ‘Add a Toolbox tag-value pair’ button makes unlimited tag-value constraints available
- Nothing in this set-up is Chintang-specific

Final import types: entries imported

- Common nouns: 4,741
- Native Chintang intransitive verbs: 282
- Borrowed Nepali intransitive verbs: 142
- Native Chintang transitive verbs: 285
- Borrowed Nepali transitive verbs: 190
- Total: 5,640/9,034 (62%) of all entries; 899/1,440 (62%) of the verbs
- Most frequent unimported POS tags: adverbs (866), adjectives (515), interjections (377), affixes (286)

Grammar Matrix lexical entry: source code

```
khad := trans-verb-lex &
   [ STEM < "kond" >,
     SYNSEM.LKEYS.KEYREL.PRED "_search;look.for_v_rel" ].

trans-verb-lex := erg-abs-transitive-verb-lex.

erg-abs-transitive-verb-lex := transitive-verb-lex &
   [ ARG-ST < [ LOCAL.CAT.HEAD noun &
                               [ CASE erg ] ],
              [ LOCAL.CAT.HEAD noun &
                               [ CASE abs ] ] > ].
```

$$
\begin{bmatrix}
\textit{trans-verb-lex} & & & & & & & & & & \\
\text{STEM} & \langle\ \textit{kond}\ \rangle & & & & & & & & & \\
\text{SYNSEM} & \begin{bmatrix}
\textit{lex-synsem} & & & & & & & & & & \\
\text{CAT} & \begin{bmatrix}
\textit{cat} & & & & & & & & & & \\
\text{HEAD} & \begin{bmatrix} \textit{verb} \\ \text{AUX} & - \\ \text{FORM} & \textit{form} \end{bmatrix} & & & & & & & & & \\
\text{VAL} & \begin{bmatrix}
\text{SUBJ} & \left\langle \boxed{0} \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{noun} \\ \text{CASE} & \textit{erg} \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \text{INDEX} & \boxed{1} \end{bmatrix} \end{bmatrix} \right\rangle \\
\text{COMPS} & \left\langle \boxed{8} \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{noun} \\ \text{CASE} & \textit{nom} \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \text{INDEX} & \boxed{9} \end{bmatrix} \end{bmatrix} \right\rangle
\end{bmatrix} \\
\text{CONT} & \begin{bmatrix}
\text{INDEX} & \boxed{14} \begin{bmatrix} \textit{event} \\ \text{TENSE} & \textit{tense} \\ \text{ASPECT} & \textit{aspect} \\ \text{MOOD} & \textit{mood} \end{bmatrix} \\
\text{RELS} & \left\langle \begin{bmatrix} \textit{arg12-ev-relation} \\ \text{PRED} & \textit{\_search;look.for\_v\_rel} \\ \text{ARG0} & \boxed{14} \\ \text{ARG1} & \boxed{1} \\ \text{ARG2} & \boxed{9} \end{bmatrix} \right\rangle
\end{bmatrix}
\end{bmatrix}
\end{bmatrix} \\
\text{INFLECTED} & \begin{bmatrix} \text{TRANSITIVE-VERB-FLAG} & + \end{bmatrix} & & & & & & & & & &
\end{bmatrix}
$$

Test data narratives

- 1,453 total word tokens
- Varied genres: biographical monologue, Pear Story (Chafe 1980), traditional story, recipe
- Example test item (target is second line):

thupro	wassace	uyuwakte	pho
thupro	wassak-ce	u-yuŋ-a-yakt-e	pho
many	bird-NS	3NSS/A-live-PST-IPFV-IND.PST	REP

'There lived many birds.' [ctn] story_rabbit.005

Lexical coverage results

Narrative	total		# analyzed		% analyzed		avg ambiguity	
	type	token	type	token	type	token	type	token
Durga_Exp	206	489	120	265	58	54	1.24	1.14
choku_yakkheng	152	331	89	184	59	56	1.26	1.20
pear_6-1	206	433	105	203	56	51	1.20	1.62
story_rabbit	85	200	43	69	51	35	1.37	1.23
All	568	1453	324	721	57	50	1.40	1.27

The AGGREGATION Project: Learning grammars from IGT?

- IGT contains a lot of linguistic analysis

jutta khet-a-ŋ-e
shoe buy-PST-1SS/P-IND.PST
'I bought a pair of shoes.' [ctn]

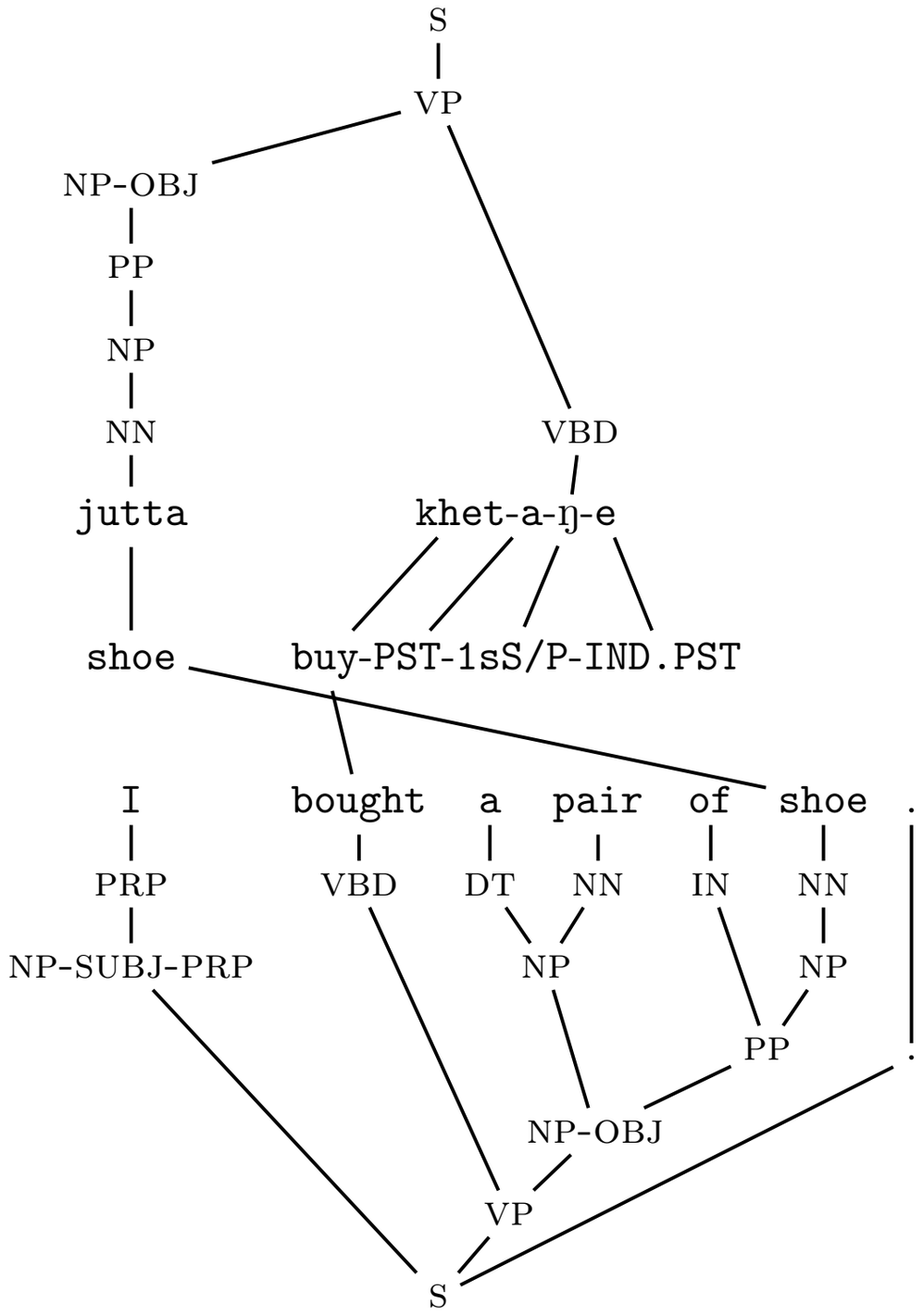
- The Grammar Matrix stores implemented analyses

- Can we write algorithms that extract information from IGT and distill it to a form that the Grammar Matrix can use?

(Bender et al 2013, 2014)

Extracting information from IGT

- Process English translation with parser for English
- Align translation to gloss, and gloss to source
- Project tree structure from English to source language (Xia & Lewis 2007)



Automatic creation of choices files

- Specification of word order and overall case system (Bender et al 2013)
- Specification of case frames for specific verbs (Bender et al 2014)
- Specification of case values for specific nouns (Bender et al 2014)
- Automatic extraction of stems and morphological rules (Wax 2014)

Extracted choices files

Choices file	# verb entries	# noun entries	# det entries
ORACLE	900	4751	0
BASELINE	3005	1719	240
FF-AUTO-NONE	3005	1719	240
FF-DEFAULT-GRAM	739	1724	240
FF-AUTO-GRAM	739	1724	240
MOM-DEFAULT-NONE	1177	1719	240
MOM-AUTO-NONE	1177	1719	240

Choices file	# verb affixes	# noun affixes
ORACLE	160	24
BASELINE	0	0
FF-AUTO-NONE	0	0
FF-DEFAULT-GRAM	0	0
FF-AUTO-GRAM	0	0
MOM-DEFAULT-NONE	262	0
MOM-AUTO-NONE	262	0

Evaluation of customized grammars

choices file	lexical coverage (%)		items parsed (%)		items correct (%)		average readings
	Training Data ($N = 8863$)						
ORACLE	1165	(13)	174	(3.5)	132	(1.5)	2.17
BASELINE	1276	(14)	398	(7.9)	216	(2.4)	8.30
FF-AUTO-NONE	1276	(14)	354	(4.0)	196	(2.2)	7.12
FF-DEFAULT-GRAM	911	(10)	126	(1.4)	84	(0.9)	4.08
FF-AUTO-GRAM	911	(10)	120	(1.4)	82	(0.9)	3.84
MOM-DEFAULT-NONE	1102	(12)	814	(9.2)	52	(0.6)	6.04
MOM-AUTO-NONE	1102	(12)	753	(8.5)	49	(0.6)	4.20
	Test Data ($N = 930$)						
ORACLE	116	(12.5)	20	(2.2)	10	(1.1)	1.35
BASELINE	41	(4.4)	15	(1.6)	8	(0.9)	28.87
FF-AUTO-NONE	41	(4.4)	13	(1.4)	7	(0.8)	13.92
FF-DEFAULT-GRAM	18	(1.9)	4	(0.4)	2	(0.2)	5.00
FF-AUTO-GRAM	18	(1.9)	4	(0.4)	2	(0.2)	5.00
MOM-DEFAULT-NONE	39	(4.2)	16	(1.7)	3	(0.3)	10.81
MOM-AUTO-NONE	39	(4.2)	10	(1.1)	3	(0.3)	9.20

How does this help language description?

- Rapid creation of precision grammars
    - => More quickly get to benefits identified earlier
- Maximize re-use of resources already created

Overview

- What is grammar engineering?
- What can grammar engineering do for language documentation?
- Case studies:
  - Lushootseed ([lut], Coast Salish, Salish, North America) morphophonology
  - Wambaya ([wmb], Mirndi, West Barkly, Australia) morphosyntax, treebanks
  - Chintang ([ctn], Kiranti, Tibeto-Burman, Nepal) automated grammar development from IGT

References

Beck, D., and T. Hess. 2014. *Tellings from Our Elders: Lushootseed syeyeyhub. Volume 1, Snohomish Texts.* Vancouver: UBC Press.

Beck, D., and T. Hess. Forthcoming. *Tellings from Our Elders: Lushootseed syeyeyhub. Volume 2, Tales from the Skagit Valley.* Vancouver: UBC Press.

Beesley, Kenneth R., and Lauri Karttunen. 2003. *Finite State Morphology.* Stanford CA: CSLI Publications.

Bender, Emily M. 2008. Evaluating a crosslinguistic grammar resource: A case study of Wambaya. In *Proceedings of ACL-08: HLT*, 977–985, Columbus, Ohio, June. Association for Computational Linguistics.

Bender, Emily M., Joshua Crowgey, Michael Wayne Goodman, and Fei Xia. 2014. Learning grammar specifications from igt: A case study of chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 43–53, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation* 1–50. 10.1007/s11168-010-9070-1.

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 J. Carroll, N. Oostdijk, and R. Sutcliffe (Eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, 8–14, Taipei, Taiwan.

Bender, Emily M., Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 74–83, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bender, Emily M., Robert Schikowski, and Balthasar Bickel. 2012. Deriving a lexicon for a precision grammar from language documentation resources: A case study of Chintang. In *Proceedings of COLING 2012.*

Bickel, Balthasar, Martin Gaenszle, Novel Kishore Rai, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Netra Paudyal, Judith Pettigrew, Ichchha P. Rai, Manoj Rai, Robert Schikowski, and Sabine Stoll. 2009. Audiovisual corpus of the chintang language, including a longitudinal corpus of language acquisition by six children, plus a trilingual dictionary, paradigm sets, grammar sketches, ethnographic descriptions, and photographs.

Chafe, Wallace. 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production.* Norwood NJ: Ablex Pub. Corp.

Drellishak, Scott, and Emily M. Bender. 2005. A coordination module for a crosslinguistic grammar resource. In S. Müller (Ed.), *The Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar, Department of Informatics, University of Lisbon*, 108–128, Stanford. CSLI Publications.

Ghodke, Sumukh, and Steven Bird. 2010. Fast query for large treebanks. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 267–275, Los Angeles, California, June. Association for Computational Linguistics.

Green, Ian, and Rachel Nordlinger. 2004. Revisiting Proto-Mirndi. In C. Bowern and H. Koch (Eds.), *Australian Languages: Classification and the Comparative Method*, 291–311. Amsterdam: John Benjamins.

Hess, T.M. 1967. *Snohomish Grammatical Structure.* PhD thesis, University of Washington.

Hulden, Mans. 2009. Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, 29–32, Athens, Greece, April. Association for Computational Linguistics.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19:313–330.

Nordlinger, Rachel. 1998. *A Grammar of Wambaya, Northern Australia.* Canberra: Research School of Pacific and Asian Studies, The Australian National University.

Oepen, Stephan, Daniel Flickinger, Kristina Toutanova, and Christopher D. Manning. 2004. LinGO Redwoods. A rich and dynamic treebank for HPSG. *Journal of Research on Language and Computation* 2(4):575–596.

Wax, David. 2014. Automated grammar engineering for verbal morphology. Master's thesis, University of Washington.

Xia, Fei, and William Lewis. 2007. Multilingual structural projection across interlinearized text. In *NAACL-HLT 2007*, Rochester, NY.