Sahaptin: A Grammar-Customization Case Study[1]

Scott Drellishak
University of Washington

July 22, 2009

[1] This work was supported by NSF grant BCS-0644097, a gift to the Turing Center by the Utilika Foundation, and the Max Planck Institute for Evolutionary Anthropology

Overview

- Today I'll demonstrate the Grammar Matrix customization system by showing how to create a grammar of Umatilla Sahaptin (Penutian)
- Customization system automatically creates grammars
- User answers a typological questionnaire, gets grammar
- Sahaptin tests all the phenomena I added during my dissertation research:
  - Case on verbal arguments
  - Verb-argument agreement
  - Direct-inverse argument marking
  - Complex feature systems

- Overview
- Background
- Sahaptin
- Libraries in Action
- Conclusion

The LinGO Grammar Matrix (Bender et al., 2002)

- Distill the wisdom of existing broad-coverage grammars
- Provide a typologically-informed foundation for building grammars of natural languages in software
- Syntax-semantics interface consistent with HPSG (Pollard and Sag, 1994) and Minimal Recursion Semantics (MRS) (Copestake et al., 2005), expressed in Type Description Language (TDL), and compatible with the LKB (Copestake, 2002)
- Intended to support all languages—implies support for universal phenomena
- However, there exist phenomena that are widespread but not universal, and we still we want to support them

Matrix Libraries

- Problem: do non-universal phenomena belong in the Matrix?
- Undesirable to burden grammars without the phenomena
- Solution: divide the Matrix into:
  - Universal or "core" Matrix
  - Matrix "libraries" for non-universal phenomena
- Libraries are exposed to the user-linguist via a web-based typological questionnaire
- Based on answers, customize an HPSG grammar of the target language

The Customization System

- Hides details of complex analyses of phenomena
- Produces "starter" grammars, intended to be extended by hand, but increasingly complex "out of the box"
- Flow of grammar development:
  descriptive grammar →
  fill out the questionnaire →
  TDL/LKB software grammar →
  further incremental development (improve/test/debug)

- Overview
- Background
- Sahaptin
- Libraries in Action
- Conclusion

Sahaptin Language

- Pacific NW language, varieties spoken in Washingon and Oregon (Rigsby and Rude, 1996)
- Complex argument marking and agreement: case, agreement, direct-inverse all working together
  - Case marking on verbal arguments
  - Argument marking sensitive to a scale, with proximate and obviative third-person nominals.
  - Two loci of agreement (verbal prefix and second-position enclitic), agreement with both the subject and the object
  - Number: sg/du/pl on nominals, but sg/pl in agreement morphology
  - Inclusive/exclusive distinction in person, but only on the second-position enclitic

Sahaptin Example

- Example:

  (1) *ín=aš á-tux̣nana yáamaš-na*
  I=1SG 3ABS-shot mule.deer-OBJ
  ‘I shot the mule deer.’ (Rigsby and Rude, 1996, 676)

- (Refer to full intransitive and transitive paradigms in (Rigsby and Rude, 1996))

Sahaptin Grammar

- Filled out the questionnaire for a fragment of Sahaptin
- About 80 hours of work (constructing test sentences, analyzing, filling out questionnaire, and debugging)
- System can't handle second-position enclitic—so described Sahaptin as VSO, with prefixes and enclitics as verb morphology (produces legal word order)
- Direct-inverse scale:
  1st > 2nd > 3rd topic > 3rd non-topic
- Vocabulary sufficient to test intrans/trans patterns:
  - One intransitive verb: *wína* 'go'
  - One transitive verb: *q̓ínun* 'see'
  - For NPs, subject and object forms of pronouns

- Overview
- Background
- Sahaptin
- Libraries in Action
- Conclusion

Case

- User can choose among nine patterns for case-marking of verbal arguments
- Customized grammar has appropriate case type hierarchy, types for lexical verbs that specify case on their args
- Though Sahaptin's morphology is complex and the literature refers to some nominal forms as "ergative" and "absolutive", it can be analyzed as nominative-accusative
- [demonstration]

Typology of Agreement

- Agreement is co-variation between two elements in some feature
- Extremely varied phenomenon
- I narrowed the focus to agreement between:
  - Verbs and their arguments
  - Determiners and nouns
- Requires a way to describe features

Features

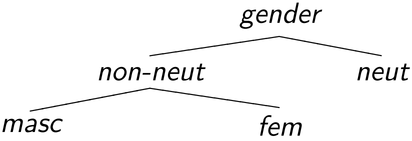
- Three features: person, number, gender
- Questionnaire allows definition of hierarchies
- Directly for number and gender, indirectly for person
- Once defined, features can be used in the lexicon
- [demonstration]

Inflection

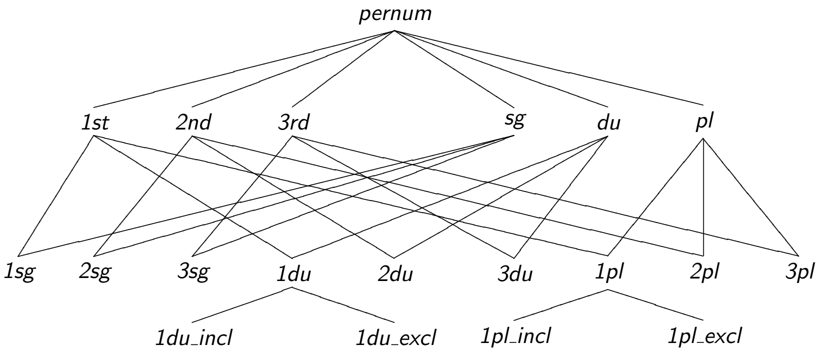
- Model: stems, slots, and morphemes (O'Hara, 2008)
- Inflectional **slots** attach to stems, or to the output of other slots
- One or more **morphemes** appear in each slot
- Morphemes may have one or more features specified
- [demonstration]

Hierarchy Augmentation

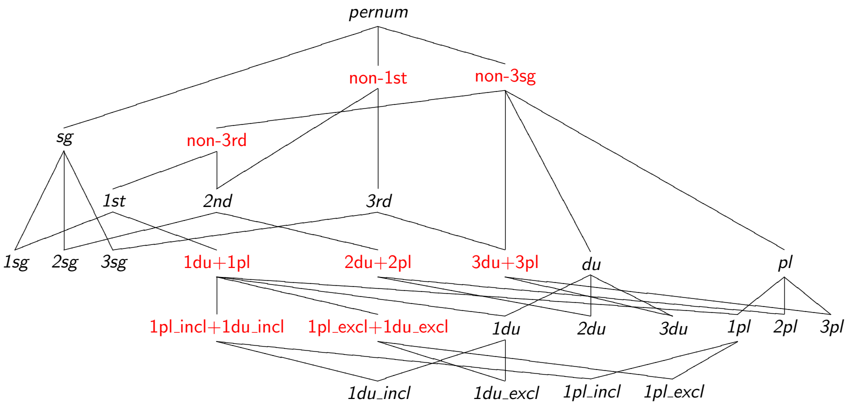
- Way to deal with disjunction
- User can specify multiple feature values using multi-select drop-downs
- System inserts new types into hierarchies
- Example: user says gender (masc, fem, neut), specifies something in the lexicon as gender=masc, fem



Default PERNUM hierarchy



Sahaptin PERNUM hierarchy



Typology of Direct-inverse

- Another variety of verbal argument-marking
- Came across while researching case, but a distinct phenomenon
- NPs are ranked on a scale according to their naturalness as an agent
    - If A outranks O, clause is **direct**
    - If O outranks A, clause is **inverse**
- E.g., 1st person > 2nd person > 3rd person
- Marking can appear on the verb, on the NP arguments, or both

Direct-inverse in Fox

- Example: Fox (Algonquian)
- Scale: 2nd > 1st > 3rd proximate > 3rd obviative

(2) *ne -waapam-aa -wa*
1SG see-DIRECT 3
'I see him.'

(3) *ne -waapam-ek -wa*
1SG see-INVERSE 3
'He sees me.' (Comrie, 1989, 129)

- Some languages also have **proximate** and **obviative** forms to distinguish 3rd person NPs
- Obviative argument is "demoted" on the scale

Direct-inverse in Fore

- Example: Fore (Trans-New Guinea)
- Scale: pron, name, kin > human > anim > inanim

(4) *yaga: wá aegúye*
pig man 3SG.hit.3SG
'The man kills the pig'

(5) *yaga:-wama wá aegúye*
pig-DLN man 3SG.hit.3SG
'The pig kills the man'

(6) *wa yága:-wama aegúye*
man pig-DLN 3SG.hit.3SG
'The pig kills the man' (Scott, 1978, 116)

- Case marking varies, but verbs aren't marked

Implementing Direct-inverse

- User describes the grammatical scale in the questionnaire
- System produces a sheaf of lexical rules to produce direct and inverse variants of transitive verbs
- Features on those rules come from the features specified in the scale
- New feature, DIRECTION, records whether direct or inverse
- Verb inflection or case-marking on arguments can be conditioned on DIRECTION
- [demonstration]

Direct-inverse agreement: SC-ARGS

- Verb-subject and verb-object agreement is common
- Some direct-inverse languages have inflection that agrees with the more highly-ranked argument
- e.g., Sahaptin and Cree
- To model this, add a feature SC-ARGS to signs
- This list contains the arguments in scale-order
- Agree with higher = agree with first element
- Agree with lower = agree with second element
- [demonstration]

- Overview
- Background
- Sahaptin
- Libraries in Action
- Conclusion

Testing the Sahaptin Grammar

- Created sentences from the intransitive and transitive patterns, and ungrammatical sentences by permutation
- 89 grammatical, 6076 ungrammatical
- 8 of the ungrammatical sentences actually parsed—correspond to unfilled cells in the paradigm from R&R
- Example:

(7)	*i-q̓ínun*	*pɨ́n-TOP*	*piinamanáy*
	3SG-see	3SG.NOM-TOP	3DU.OBJ

‘He/she/it sees them.’

- Parses because features specified on *i-* cover 3SG.NOM-TOP subject, 3DU.OBJ object

Conclusion

- Sahaptin exercises the new libraries for case, agreement, direct-inverse languages:
  - Parts of the analyses are novel: entire analysis of direct-inverse, including SC-ARGS
  - Hierarchy augmentation is a powerful technique for turning description into precision grammar
- Also refined our understanding of Sahaptin:
  - Sahaptin is nominative-accusative and direct-inverse, no need for ergative-absolutive pattern
  - Sahaptin shows agreement that's sensitive to the direct-inverse scale

References

Bender, Emily M., Flickinger, Dan and Oepen, Stephan. 2002. The Grammar Matrix. In *Proceedings of COLING 2002 Workshop on Grammar Engineering and Evaluation*, Taipei, Taiwan.

Comrie, Bernard. 1989. *Language Universals & Linguistic Typology, Second Edition*. Chicago: University of Chicago.

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

Copestake, Ann, Flickinger, Dan, Pollard, Carl and Sag, Ivan A. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language & Computation* 3(2–3), 281–332.

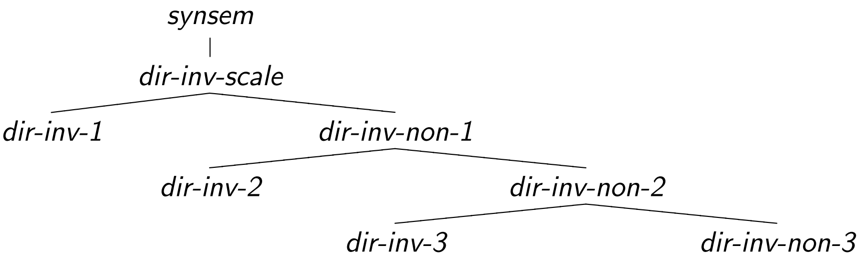
O'Hara, Kelly. 2008. *A Morphotactic Infrastructure for a Grammar Customization System*. Masters Thesis, University of Washington.

Pollard, Carl and Sag, Ivan A. 1994. *Head-Driven Phrase Structure Grammar*. Stanford: CSLI.

Rigsby, Bruce and Rude, Noel. 1996. Sketch of Sahaptin, a Sahaptian Language. In Ives Goddard (ed.), *Languages*, pages 666–92, Washington DC: Smithsonian Institution.

Scott, Graham. 1978. *The Fore Language of Papua New Guinea*. Canberra, Australia: Pacific Linguistics.

Direct-inverse: Scale Hierarchy



- Types on the left specify the features for a scale entry
- Types on the right specify features covering the rest of the scale

Direct-inverse: Algonquian Scale

- So, for Algonquian languages:

$$\begin{bmatrix} \textit{dir-inv-1} & \\ \text{PER} & \textit{2nd} \end{bmatrix} \qquad \begin{bmatrix} \textit{dir-inv-non-1} & \\ \text{PER} & \textit{non-2nd} \end{bmatrix}$$

$$\begin{bmatrix} \textit{dir-inv-2} & \\ \text{PER} & \textit{1st} \end{bmatrix} \qquad \begin{bmatrix} \textit{dir-inv-non-2} & \\ \text{PER} & \textit{3rd} \end{bmatrix}$$

$$\begin{bmatrix} \textit{dir-inv-3} & \\ \text{PER} & \textit{3rd} \\ \text{PROXIMITY} & \textit{proximate} \end{bmatrix} \qquad \begin{bmatrix} \textit{dir-inv-non-3} & \\ \text{PER} & \textit{3rd} \\ \text{PROXIMITY} & \textit{obviative} \end{bmatrix}$$

Direct-inverse: Lexical Rules

- For Algonquian, lexical rules for direct specifying:

  SUBJ ⟨*dir-inv-1*⟩ COMPS ⟨*dir-inv-non-1*⟩

  SUBJ ⟨*dir-inv-2*⟩ COMPS ⟨*dir-inv-non-2*⟩

  SUBJ ⟨*dir-inv-3*⟩ COMPS ⟨*dir-inv-non-3*⟩

- ...and three rules for inverse:

  SUBJ ⟨*dir-inv-non-1*⟩ COMPS ⟨*dir-inv-1*⟩

  SUBJ ⟨*dir-inv-non-2*⟩ COMPS ⟨*dir-inv-2*⟩

  SUBJ ⟨*dir-inv-non-3*⟩ COMPS ⟨*dir-inv-3*⟩