Implemented Grammars for the Rest of the World: The challenge of Slave

Jeff Good, MPI EVA
(good@eva.mpg.de)
Emily Bender, U. Washington
(ebender@u.washington.edu)
SSILA Winter Meeting
Oakland, California
January 7, 2005

Outline of talk

- Introduce the Montage project (Bender et al. 2004) to an Americanist audience
- Give an overview of the design plan of the entire Montage toolkit
- Focus on the design plan for a system for morphological analysis in Montage, including:
  - Expected benefits for descriptive linguists
  - The ways in which Athabaskan morphology has informed system design
- Companion talk: Today, 3:00, Jewett ABC, Computational linguistics session

Overview: Montage

- The Montage project has the goal of building tools to assist field linguists in doing grammatical analysis
- Specifically, it is developing a toolkit allowing the ordinary working linguist to make use of methods from computational grammar engineering without being grammar engineers themselves
- Implemented grammar: A machine-readable encoding of the grammatical analysis of some linguistic data from a given language

Overview: Montage

- The project is actively cooperating with related initiatives, such as those associated with the E-MELD project
- It has an advisory board of computational linguists and field linguists
- Overarching methodology: Come to a detailed understanding of the nature of the workflow of language documentation in order to pinpoint ways in which computational techniques can best assist the field linguist

Overview: Montage

- Some features of the design of the toolkit
  - At heart, a system for electronically annotating text data for grammatical information
  - Annotation should not be limited to word-level interlinearization but should be flexible to reflect the nature of grammatical discovery
  - System designed to make use of even partial or preliminary analysis

Overview: Montage

- Important aspects of the functionality of the toolkit
  - System for linking morphemes in texts to entries in an electronic lexicon (e.g., a FIELD lexicon)
  - Ability to search for and collate text examples across a range of grammatical parameters

Montage: Morphology

- The focus of this talk is the present design being employed by Montage to build tools for morphological analysis
- The project is specifically using Slave (Rice 1989) as a difficult test case
- The expectation is that a system designed to assist in the morphological analysis of an Athabaskan language should work for a wide range of other languages

Implementation: Why

- Targeted searching and collating
  - Search for morphological annotation, not just phonological string
  - Search for forms where some rule applies
  - Find exceptions to defined rules (both actual and due to mistranscriptions/ typoes)
  - Detection of variant forms to determine if variation is systematic or not

Implementation: Why

- See if rules really do analyze the data as expected and quickly discard analyses which do not account for the data
- Usable by software engineers for machine translation and other computational tools (e.g., spell checkers)
- Facilitate sharing of analyses

Implementation: How

- Default rule formalism: SPE style
- Wizards for interfacing with Grammar Matrix when formalizing common morphological constructions
- Grammar Matrix: A starter kit for the development of implemented grammars with a range of useful predefined grammatical constructions

Slave Morphology

- Position-class system (Rice 1989:437)
- adv=obj= pp#adv#dist#cust#stem# number+DO+deic+theme+asp+conj+mode+subj= cl-stem
- Seventeen total possible positions
- With different phonology across different kinds of boundaries
- "Ideal" position classes are syntactically and semantically arbitrary

Slave Morphology

- Verb classes based on the "classifiers" a verb theme contains (Rice 1989:439–470)
  - Ø-classifier, **Ø**-ʔáh 'eat, chew'
  - h-classifier, **h**-t'ó 'suck'
  - d-classifier, **d**-shin 'sing'
  - l-classifier, ná-**l**-séh 'hunt'
- In some cases, verbs can alternate in their choice of classifier in conservative versus innovative speech (Rice 1989:449–50)

Slave morphology

- An epenthetic "peg element" is inserted before verb stems if they are not otherwise preceded by a syallable (Rice 1989:133)
  - **he**hjį 'I sing' vs. nejį 'you sing'

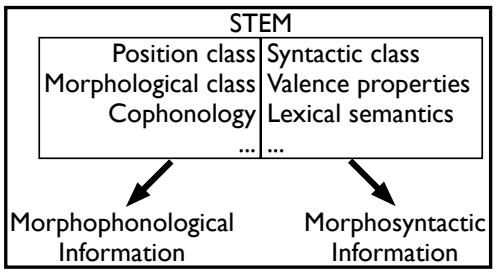
Design conclusion

- A sufficient flexible computational tool for morphological analysis requires the ability for morphophonological generalizations to be made mostly independent from morphosyntactic ones
- Morphophonology: maps surface forms to strings of abstract morphemes
- Morphosyntax: maps strings of abstract morphemes to syntactic/semantic information (feature structures)

Lexical database

- Bipartite lexical database
- Each lexical entry is associated with
  - A "stem", an abstract representation of a morpheme, often an underlying form but in some cases, a morphosyntactic label like "CAUS"
  - Morphophonological information
  - Morphosyntactic information

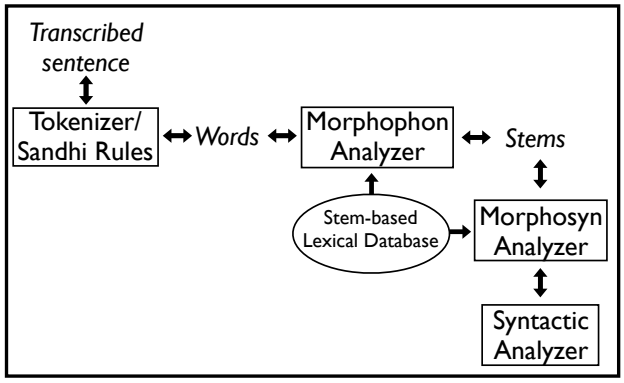
Lexical database



Lexical database

- The nature of the morphophonological and morphosyntactic information would be specified by the linguist
- Some might be entered and accessed via resources produced by a dedicated lexicon-building tool, like E-MELD's FIELD tool
- As part of support for the system, where possible, lexical "templates" might be provided which would work well with particular language groups

Parsing system



Conclusions

- To accommodate the description of morphologically complex languages, a system where morphophonology and morphosyntax are mostly independent is necessary
- The "morphology-syntax" interface in this system consists of the abstract stem linking a morphophonological half of a lexical entry with a morphosyntactic half
- We believe this very "open" system is required for a tool designed for general use by documentary linguists

Acknowledgments

Thanks for helpful discussion to:
Duane Blanchard, Anya Dormer, Scott Drellishak, Ann Gaponoff, David Goss-Grubbs, Jeremy Kahn, Bill McNeill, Matty Noble, Laurie Poulson

References

AGTK: Annotation Graph Toolkit. http://www.ldc.upenn.edu/Projects/AG/
ELAN: EUDICO Linguistic Annotator. http://www.mpi.nl/tools/elan.html
FIELD: Field Input Environment for Linguistic Data. http://emeld.org/tools/fieldinput.cfm
Grammar Matrix: Precision Grammar Starter Kit. http://www.delph-in.net/matrix/
LKB: LKB Grammar Development Environment. http://www.delph-in.net/lkb/
QLDB: Querying Linguistic Databases. http://www.ldc.upenn.edu/Projects/QLDB/
XFST: Xerox Finite State Transducer. http://www.fsmbook.com/

Bender, Emily M., Dan Flickinger, Jeff Good and Ivan A. Sag. 2004. Montage: Leveraging Advances in Grammar Engineering, Linguistic Ontologies, and Mark-up for the Documentation of Underdescribed Languages. Proceedings of the Workshop on First Steps for the Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, LREC 2004, Lisbon, Portugal.

Farrar, Scott and Langendoen, Terence D. 2003. A linguistic ontology for the Semantic Web. *GLOT International* 7:97–100.

Kari, James. 1989. Affix positions and zones in the Athapaskan verb. *International Journal of American Linguistics* 55:424–55.

Rice, Keren. 1989. A grammar of Slave. Berlin: Mouton.

Rice, Keren. 2000. Morpheme and order and semantic scope: Word formation in the Athabaskan verb. Cambridge: Cambridge University.