

Implementing Comparative Reconstruction with Finite-State Methods

James Kilbury
HHU Düsseldorf
01.07.2011

my goals for today

- present a new finite-state example for comparative reconstruction and multilingual lexica ***including inflectional morphology***
- show how our recent work fits together
- show how this example fits into my own work in computational linguistics

a personal retrospective

- learned (some) Russian in school – 1960-64
- studied Old Church Slavonic – 1966
- studied mathematical logic, then TG -- 1966
- wrote about Old Icelandic morphology – 1969
- focus on inflection with complex morphophonemics
- graduate studies – 1970-74
 - historical-comparative Slavic and Germanic
 - theoretical basis of reconstruction with Hockett
 - dissertation on morphophonemics
- computer science – 1980-84

DATR as a formal framework

- representation language developed by Gerald Gazdar and Roger Evans 1989 –
- for declarative definition of **non-monotonic multiple inheritance** networks
- designed especially to define lexica and capture **linguistic generalizations**
- concentration on **inflectional morphology**
- Network Morphology and PF Morphology

serbocro.dtr (Gazdar, 1989)

- typical early DATR study (Serbo-Croatian)
- uses **inheritance** to capture relations
 - within paradigms (**syncretism**)
 - between paradigms (**inflection classes**)

problem: morphophonemics not dealt with

>> file

sbcnx.dtr (Kilbury, 199?-2007)

- implements **finite-state transducers** in DATR to capture **morphophonemic alternations**
- achieves more compact lexical entries
- realization, syncretism, and classes as before

problems:

- no good solution to cascades of rules
- lack of homogeneity
- no way to parse

>> file

new idea for reconstruction

If we can implement finite-state transducers as DATR nodes, then we can implement non-monotonic inheritance hierarchies of transducers in DATR.

This is just what we need to implement historical-comparative reconstruction in DATR in order to reduce the variance of different languages in a family to the invariance of their genetic relationship.

hcr*.dtr (Kilbury, 199?-2004)

- uses DATR inheritance to capture **genetic relations** (i.e. **regular sound correspondences**) within a language family
- **inheritance network of transducers**

problems:

- cascades of replacement rules
 - no way to parse (i.e. reconstruct proto-forms from lexical cognate sets)
- >> file**

xfst as a new formal framework

- work by Johnson, Kay, Kaplan as basis
- implementation and language for regular languages and relations by Lauri Karttunen
- provides **rewrite rules** as reg expressions
- realizational morphology easily captured
- allows generation **and parsing**

problem: how to capture (non-monotonic)

inheritance relations (i.e. advantage of DATR)

sbc*.txt (Kilbury, 2004 -)

- extends the standard rewriting techniques of *xfst* and realizational morphology to ***tag-rewriting for non-monotonic relations between inflectional classes***
- ***combines the advantages of DATR and xfst for inflectional morphology***

>> file

hcr*.txt (Kilbury, 2008)

- a **formal model** of comparative reconstruction based on **multitape automata** of Kay
- introduces **language tags** for multiple tapes
- allows **generation** (daughter forms derived from proto-forms) as well as **parsing** (proto-forms from lexical cognate sets)
- hcr4.txt as ***tree model*** of relationship
- hcr5.txt as ***wave model***

problem: no inflectional morphology dealt with

>> file

oscn.txt (Kilbury, 2011)

- sbc20.txt (Serbo-Croatian)
including noun inflection adapted to
Old Church Slavonic
- OCS serves as an approximation of the
proto-language for a **multilingual** Slavic
example (same lexicon as upper language)

>> file

orsn.txt (Kilbury, 2011)

- **monotonic extension** of ocsn.txt
- extends previous example to cover
 - **O**ld Church Slavonic
 - **R**ussian and
 - **S**erbo-Croatian **n**oun morphology

>> file

Fazit

- incorporates and summarizes all of our techniques and previous work
- provides a framework for future work
 - historical reconstruction of Germanic, ...
 - synchronic multilingual lexica

Time to stop!