Online Linguistic Database

Joel Dunham
Context

- The tasks of documenting and analyzing endangered languages are urgent
- These are hard tasks
- ... but there is an intuition that the “grunt” work could be accomplished more quickly
Online Linguistic Database (OLD)

- Software for creating web applications that facilitate language documentation and linguistic analysis

- www.onlinelinguisticdatabase.org
OLD overview

- a program for creating web applications
- collaboratively created language databases
- open source – [https://github.com/jrwdunham/old](https://github.com/jrwdunham/old)
- good documentation
- platform-agnostic (Mac, Linux, Windows)
- Python (Pylons), MySQL (SQLAlchemy), JavaScript (CoffeeScript, Backbone), HTML5
a program for creating web applications ...
Aside on versions

OLD 0.2.7

- in production: Blackfoot, Okanagan, Gitxsan, Nata, etc.
- Python server-side logic
- HTML/JavaScript GUI

OLD 1.0

- under development
- RESTful web service (Python)
- HTML5/CoffeeScript GUI
multi-user, concurrent collaboration

• Imagine a field methods class producing a valuable resource of structured and formatted endangered language data simply as a byproduct of their normal workflow.
multi-user, concurrent collaboration

- A typical field methods class generates at least 1,000 hours of recordings plus transcriptions & analysis over the course of the 5 years of its impact.
structure, presentation, exploration
structure, presentation, exploration

- *le* (the)
- *chien* (dog)
- *s* (plural)

- *les* (the PL)
- *les chien-s* (the dogs)

- *D-Phi N-Phi*
structure, presentation, exploration

'a long time ago there were three people camping'

amo_itao_kaayaa.wav
creat_storyEdited_oct14.wav
play audio I link to audio
structure, presentation, exploration

Noreen Super Doll Story v.2 Oct 22, 08

Sara Johansson

entered: 3 years ago

last modified: 2 years ago

(1) na aakiikoan áyinni ni otsitapiim
   ann-wa aakiikoan áyinni ann-yi ot-atapiim
   DEM-PROX.SG girl IMPF-grasp DEM-OBV.SG 3-doll
   ki na saahkomaaapi itotâ'iroyi
   ki ann-wa saahkomaaapi iit-ot-á-ipoi
   and DEM-PROX.SG boy LOC-3-IMPF-stand
   ‘The girl was holding the doll and the boy was standing (watching).’

(2) na aakiikoan iitsitâ'iroyi anni
   ann-wa aakiikoan LOC-?-IMPF-speak DEM-OBV.SG
   DEM-PROX.SG girl 3-doll
   otsitapiim ot-atapiim
   ‘The girl is talking about her doll’
structure, presentation, exploration

Noreen Super Doll Story v.2 Oct 22, 08

entered by: Sara Johansson
exported by: Joel Dunham

Jan 26, 2011

(1) na aakiikoan áiyinni ni otsitapíím
ann-wa aakiikoan á-yinni ann-yi ot-atapíím
DEM-PROX.SG girl IMPF-grasp DEM-OBV.SG 3-doll
ki na saahkómaapi iitotá’poyi
ki ann-wa saahkómaapi iit-ot-á-ipoyi
and DEM-PROX.SG boy LOC-3-IMPF-stand
‘The girl was holding the doll and the boy was standing (watching).’

(2) na aakiikoan iitsitá’poyi anni
ann-wa aakiikoan iit-it-á’ipoyi ann-yi
DEM-PROX.SG girl LOC-?-IMPF-speak DEM-OBV.SG
otsitapíím
ot-atapíím
3-doll
‘The girl is talking about her doll’
input validation
input validation

Add a Form

PRIMARY DATA

Broad Phonetic Transcription keyboard

Error: the broad phonetic transcription you have entered is not valid. Only graphemes from the broad phonetic inventory and the space character are permitted. See the Application Settings and Orthographies & Inventories pages for more details.

zaβ

b d g t c h k m n n
η r r β s j y w j i
í u ú é ã o ó e é c
ó a á
search

- an unlimited number of filter expressions composed via boolean operators into a tree structure
- regular expressions
there are 3 words in the transcription
the morpheme break field contains the morpheme iksi
None of the translations contains the string Bark or bark
it was entered earlier than Apr 24, 2013 at 4:57 p.m.
it was modified after Apr 24, 2013 at 4:57 p.m.
it was not elicited on the first or the third day of 2012
it was entered by a user with an odd-numbered ID
it has an ID greater than 10,000
it is not tagged as pseudo-data
category search

- search for high-level morpho-syntactic patterns via the system-generated *syntactic category string* value
- E.g., find DPs with one or more adjectival modifiers:
  - D (A )+ N
  - *the big dog*
  - *the big bad cat*
unambiguous morpheme search

• get exactly the morpheme you are looking for via the system-generated *break gloss* category value

• E.g., search for s|PL to get exactly the morpheme /s/ glossed as “PL” and not s|PRES or i|PL
phrase structure search

• supply your forms with syntactic representations via the syntax field:

  (TP
    (DP
      (D le-s)
      (N chien-s))
    (VP
      (V courr-aient))))
phrase structure search

- give me all forms where a TP immediately dominates a DP and a VP:

\[ TP < DP < VP \]
Imagine you could specify phonological mappings in a familiar notation and use those specifications to implement phonological generators and parsers.
phonological models

• Applications:

  • enter morpheme segmentations and have phonological models generate orthographic and phonetic transcripts automatically

  • incorporate phonological models into morphophonological parsers that output morpheme segmentations and glosses when given transcriptions as input

  • specify competing phonologies, test them against large datasets and compare them
phonological models

- Finite State Transducers – one way to do it
- *context-sensitive* phonological rewrite rules (cf. Chomsky & Halle, 1968) actually describe *regular relations* (Johnson 1972) and these can be represented by *finite-state transducers* (FSTs) (cf. Karttunen & Beesley, 2001)
phonological models

context-sensitive

context-free

regular
phonological models

• a phonology represented as ordered context-sensitive rewrite rules can be implemented as an FST
• FSTs are computationally tractable
• FSTs work equally well for parsing and generation
• OLD uses foma (open source, https://code.google.com/p/foma/) to compile phonologies to FSTs so we can put them to good use
phonological models

Blackfoot phonological rules (Frantz 1997)

Appendix B: Phonological Rules

1. GEMINATION \( C_1 \rightarrow C_2 / \_+C_2 \)
   
nitánIt + k + wa \rightarrow nitánIk + k + wa (7,15. \rightarrow nitánikka)
   "He told me."

2. s - INSERTION \( \emptyset \rightarrow s / I_t \)
   
nitánItawa \rightarrow nitápIlstawa (7. \rightarrow nitánistawa)
   "I told him."

3. s - CONNECTION
   a. \( \emptyset \rightarrow s / C_+s \)
      
nit + siksipawa \rightarrow nitssiksipawa
      "I bit him"
      ‘I rested’
   b. \( \emptyset \rightarrow I / V(\_)+_s, \)
      where \( _s \) is not part of a suffix.
phonological models

Frantz’s Blackfoot phonology as a Foma script

```plaintext
104  $  
105  # 1. C1-C2  ->  C2C2$  
106  # Geminations$  
107  define pGem plosives "-" -> p || _ p;$  
108  define tGem plosives "-" -> t || _ t;$  
109  define kGem plosives "-" -> k || _ k;$  
110  define gemination pGem .o. tGem .o. kGem;$  
111  $  
112  # 2. It  ->  Ist$  
113  # s-Insertion (assumes that "breaking I" is a phoneme)$  
114  define sInsertion [...] -> s || I _ t;$  
115  $  
116  # 3.a. C-s  ->  Css$  
117  # s-Connection A$  
118  define sConnectionA "-" -> s || stops _ s;$  
119  $  
120  # 3.b. V('.')-s  ->  V('.')-is$  
121  # s-Connection B$  
122  # condition: where 's' is not part of a suffix$  
123  # present implementation: rule is optional$  
124  define sConnectionB [...] (->) i || vowels ("""") "-" _ s;$  
125  $  
```
Phonology

∅ → s | I _ t
I → ∅
w → ∅ | # _
...

phonological models

generate

/waanlt/

[aanist]
Phonology

∅ -> s | I _ t
I -> Ø
w -> Ø | # _

parse

phonological models

/waanIt/
/waanist/
/aanIt/
...

[aanist]
phonological models

• “But I work in OT, autosegmental representations, etc. – why should I rewrite my phonology as ordered rewrite rules?”

• practical benefits of implementation in the OLD

• probably not a bad idea to try to capture your generalizations in a variety of frameworks
Imagine the database could auto-generate a morphology that accepted sequences of morphemes that constitute valid words and rejected those that do not.
Morphology

D -> [ le, la ]
N -> [ chien, chat ]
Phi -> [s]

word -> [ N | D | N-Phi | D-Phi ]

Morphological models

Words Corpus

les chiens
le-s chien-s
the-PL dog-PL
‘the dogs’
D-Phi N-Phi

le chat
le chat
the cat
‘the cat’
D N

Lexicon Corpus

le
le
the
‘the’
D

la
la
the
‘the’
D

chien
chien
dog
‘dog’
N

chat
chat
cat
‘cat’
N

s
s
PL
‘plural’
Phi
Morphology

D $\rightarrow \{ \text{le}, \text{la} \}$

N $\rightarrow \{ \text{chien}, \text{chat} \}$

Phi $\rightarrow \{ s \}$

word $\rightarrow \{ \$

\text{N} | \text{D} | \text{N-Phi} | \text{D-Phi} \$

\}$

('chien', 'dog'), 's', 'PL')

chien-s'

morphological models
Morphology

D -> [ le, la ]
N -> [ chien, chat ]
Phi -> [s]

word -> [
  N |
  D |
  N-Phi |
  D-Phi
]

Ø ↔ ‘s-chien’
morphological parsers

parse

Frantz’
Blackfoot Phonology as FST

[nitsspiyi]
morphological parsers

recognize

Ø

Blackfoot Morphology FST extracted from database

/-N---M-i---t---si-hsspl-yy-I-/
morphological parsers

Blackfoot Morphology FST extracted from database

("nit", 'I'), '2', ('ihpiyi', 'dance')
morphological parsers

( ('nit', '1'), '-', ('ihpiyi', 'dance') ),
( ('nit', '1'), '-', ('sspi', 'among'), '-', ('yi', '0') ),
( ('nit', '1'), '-', ('sspi', 'among'), '-', ('yi', '3PL') ),
( ('nit', '1'), '-', ('sspi', 'among'), '-', ('yi', '3pl') ),
( ('nit', '1'), '-', ('sspi', 'among'), '-', ('yi', '4PL') ),
( ('nit', '1'), '-', ('sspi', 'among'), '-', ('yi', 'be') ),
( ('nit', '1'), '-', ('sspi', 'among'), '-', ('yi', 'have') ),
( ('n', '1'), '-', ('it', 'LOC'), '-', ('ihpiyi', 'dance') ),
( ('n', '1'), '-', ('it', 'loc'), '-', ('ihpiyi', 'dance') )
morphological parsers

Phonology FST

Morphology FST

Morphophonological FST

Ranker

Language model corpus

((nit', 'I'), '-', ('ihpiyi', 'dance'))

nitsspiyi

Sunday, April 28, 2013
morphological parsers

- user writes a phonology
- user assembles three corpora: lexicon, words and language model

Morpho-phonological parser
morphological parsers
repurposing data

• imagine our collaboratively created repository of language data could be used to advance the goals of those involved in language revitalization and teaching

• imagine a linguistics field methods class being able to immediately give back to the speech community
repurposing data: OLD architecture

Browser App
GUI for fieldwork & research

Mobile App
talking dictionary

Mobile App
learning games

OLD RESTful web service

get data quickly & flexibly

HTTP
JSON

HTTP
JSON

HTTP
JSON
repurposing data:

- File (audio): recording of speaker saying *chien*
- File (image): image of a dog
- Form *chien* (‘dog’)
Conclusions

• The OLD is an interesting and useful application for language documentation, linguistic analysis and (potentially) revitalization efforts
Conclusions

• There is useful work to be done in the intersection of theoretical linguistics, language documentation, computational linguistics, language revitalization and software development
Conclusions

• Use the OLD for your own fieldwork

• Use the OLD in your next field methods class

• Contribute to the OLD
  • contribute to the code of the core web service app
  • write a user-facing app that makes use of an OLD web service!
  • use it, critique it, read the docs, critique them, ...
Works cited

Thank you

• NWLC organizers & presenters
• brilliant & hard working field linguists
• SSHRC ITST grant no. 849-2009-0056
OLD

- www.onlinelinguisticdatabase.org