LingSync:
Software for language documentation

Joel Dunham, Jessica Coon & Alan Bale

Concordia  McGill  Concordia

4th International Conference on
Language Documentation & Conservation

February 28, 2015
LingSync

- a suite of open source software tools that facilitate collaborative language documentation:
  - language description
  - theoretical linguistic research
  - revitalization efforts
- It enables and entices language documenters with diverse goals to work together to mutual advantage
Goals

• give you an overview of LingSync
• convince you that it’s a good tool for language documentation
• focus on its high-level architecture, specifically its striving for modularity and re-usability
Under development, but production-ready

www.lingsync.org
app.lingsync.org
Under development, but production-ready

www.onlinelinguisticdatabase.org
Under development, but production-ready

LingSync

Morphological Parser Creator

LingSync (+ OLD parser creator)
Pro collaboration & data sharing

- endangered/low-resourced language data are scarce so inefficiencies are especially costly
- field methods classes and research groups produce valuable data, but it’s hard to access
- your work could advance more rapidly if you had access to your peers’ data
Pro collaboration & data sharing

• endangered/low-resourced language data are scarce so inefficiencies are especially costly

• field methods classes and research groups produce valuable data, but it’s hard to access

• your work could advance more rapidly if you had access to your peers’ data
Pro collaboration & data sharing

- endangered/low-resourced language data are scarce so inefficiencies are especially costly
- field methods classes and research groups produce valuable data, but it’s hard to access
- your work could advance more rapidly if you had access to your peers’ data
Requirements

• multi-user access to corpora stored on web servers
• cross-platform & open source
• permissions control
• language/linguistics-specific features
Requirements

• multi-user access to corpora stored on web servers
• cross-platform & open source
• permissions control
• language/linguistics-specific features
Requirements

- multi-user access to corpora stored on web servers
- cross-platform & open source
- permissions control
- language/linguistics-specific features
Requirements

- multi-user access to corpora stored on web servers
- cross-platform & open source
- permissions control
- language/linguistics-specific features
Collaboration, Data-Sharing & Re-use

Google Spreadsheets

LingSync

FLEEx

ELAN

Language Documentation Features

Toolbox
Corpus web service

- Talking Dictionary
- Fieldwork/Documentation GUI
- Language-learning software
Corpus web service

Digital archive

Psycholinguistics experiment builder

Talking Dictionary

Fieldwork/Documentation GUI

Language-learning software
Morphological Parser web service

“LingSync” Corpus web service

“OLD” Corpus web service

“Dative” Fieldwork/Documentation GUI
Features Overview

- APIs
- GUIs
- search
- consistency
- permissions
- modelling & analysis testing
- audio/video
- import/export
- flexible data structure
Features

- APIs
- GUIs
- search
- consistency
- permissions
- modelling & analysis testing
- audio/video
- import/export
- flexible data structure
Features

- APIs
- GUIs
- search
- consistency
- permissions
- modelling & analysis testing
- audio/video
- import/export
- flexible data structure
Features

- APIs
- GUIs
- search
- consistency
- permissions
- modelling & analysis testing
- audio/video
- import/export
- flexible data structure
Features

- APIs
- GUIs
- search
- consistency
- permissions
- modelling & analysis testing
- audio/video
- import/export
- flexible data structure
Features

- APIs
- GUIs
- search
- consistency
- permissions
- modelling & analysis testing
- audio/video
- import/export
- flexible data structure
Features

- APIs
- GUIs
- search
- consistency
- permissions
- modelling & analysis testing
- audio/video
- import/export
- flexible data structure
Features

- APIs
- GUIs
- search
- consistency
- permissions
- modelling & analysis testing
- audio/video
- import/export
- flexible data structure
Features

- APIs
- GUIs
- search
- consistency
- permissions
- modelling & analysis testing
- audio/video
- import/export
- flexible data structure
Features

• APIs
• GUIs
• search
• consistency
• permissions
• modelling & analysis testing
• audio/video
• import/export

• flexible data structure
Web service APIs

- consistent and standardized interfaces for getting and manipulating data stored on a web server
- REST-ful protocol
- JSON exchange format
GET https://corpus.lingsync.org/your-corpus-id

[
  {
    "transcription": "oki",
    "translations": ["hello"],
    ...
  },
  {
    "transcription": "kikatáí'nawaa nitohpóósima",
    "translations": ["Have you seen my cat?"]
    ...
  },
  ...
]
“LingSync” Corpus web service

“OLD” Corpus web service

RESTful HTTP/JSON

RESTful HTTP/JSON
Graphical User Interface

• work with your data without all that coding nonsense
• good user experience
• features and conveniences that help you with fieldwork and documentation-related tasks
Benefits of browser-based applications

- browser-based applications: run in Chrome, Firefox, Safari, Opera, etc.
- cross-platform for free
- technologies have matured rapidly in recent years
- relatively easy to find/train programmers
Benefits of browser-based applications

- browser-based applications: run in Chrome, Firefox, Safari, Opera, etc.
- cross-platform for free
- technologies have matured rapidly in recent years
- relatively easy to find/train programmers
Benefits of browser-based applications

- browser-based applications: run in Chrome, Firefox, Safari, Opera, etc.
- cross-platform for free
- technologies have matured rapidly in recent years
- relatively easy to find/train programmers
Benefits of browser-based applications

• browser-based application: run in Chrome, Firefox, Safari, Opera, etc.
• cross-platform for free
• technologies have matured rapidly in recent years
• relatively easy to find/train programmers
“LingSync” Corpus web service

“OLD” Corpus web service

“Dative” Fieldwork/Documentation GUI
Welcome to Dative, an application for linguistic fieldwork.
Welcome to Dative, an application for linguistic fieldwork.
Search

• Important for linguistic research
• Important part of the API needed by
  • digital dictionaries
  • language-learning software
Search features

- morphological search
- field-sensitive boolean search with regular expressions
- syntactic search (TGrep2)
### Morphological Search

<table>
<thead>
<tr>
<th>morpheme 1</th>
<th>morpheme 2</th>
<th>morpheme 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>shape</td>
<td>shape</td>
<td>shape</td>
</tr>
<tr>
<td>wa</td>
<td>wa</td>
<td>iksi</td>
</tr>
<tr>
<td>gloss</td>
<td>gloss</td>
<td>gloss</td>
</tr>
<tr>
<td>PROX</td>
<td>PRX</td>
<td>PROX</td>
</tr>
<tr>
<td>category</td>
<td>category</td>
<td>category</td>
</tr>
<tr>
<td>NUM</td>
<td>NUM</td>
<td>NUM</td>
</tr>
<tr>
<td>sh-gl-cat</td>
<td>sh-gl-cat</td>
<td>sh-gl-cat</td>
</tr>
<tr>
<td>wa</td>
<td>PROX</td>
<td>NUM</td>
</tr>
</tbody>
</table>
Consistency

- configurable inventory-based input validation
- configurable orthography conversion
- feedback on morpho-lexical consistency
**Orthographic Transcription** *keyboard*

Warning: the orthographic transcription you have entered is not constructable using only the input orthography, punctuation characters and the space character.

<table>
<thead>
<tr>
<th>Grapheme/ Polygraph</th>
<th>Code Point(s)</th>
<th>Input</th>
<th>Input Code Point(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>U+0061</td>
<td>a</td>
<td>U+0061</td>
</tr>
<tr>
<td>á</td>
<td>U+0061, U+0301</td>
<td>á</td>
<td>U+0061, U+0301</td>
</tr>
<tr>
<td>c</td>
<td>U+0063</td>
<td>c</td>
<td>U+0063</td>
</tr>
<tr>
<td>c’</td>
<td>U+0063, U+0315</td>
<td>c’</td>
<td>U+0063, U+0027</td>
</tr>
</tbody>
</table>

**LATIN SMALL LETTER SCHWA,** **COMBINING ACUTE ACCENT; U+0259,** **U+0301**

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>á</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
</tr>
<tr>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
</tr>
<tr>
<td>í</td>
<td>í</td>
<td>í</td>
<td>í</td>
<td>í</td>
<td>í</td>
<td>í</td>
<td>í</td>
<td>í</td>
<td>í</td>
</tr>
<tr>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
<td>é</td>
</tr>
</tbody>
</table>

**LATIN SMALL LETTER Q, COMBINING COMMA ABOVE, MODIFIER LETTER SMALL W; U+0071,** **U+0313,** **U+02B7**

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>w</td>
<td>w</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>ť</td>
<td>ť</td>
<td>ť</td>
<td>ť</td>
<td>ť</td>
<td>ť</td>
<td>ť</td>
<td>ť</td>
<td>ť</td>
<td>ť</td>
</tr>
</tbody>
</table>

Graphical keyboard generated from orthography.
Orthography Conversion

Useful for endangered languages fieldwork where multiple orthographic conventions may co-exist

User chooses an orthography for entering, viewing, and searching data

Inputs are translated to and outputs are translated from representations in the designated storage orthography

Orthography Converter

database

ƛ'ałḵʷ

t'lalhk'w

ƛ'āłḵʷ

ƛ'āłḵʷ

ƛ'āłḵʷ
morphologically complex form

Nitsiitsinoaayaawa  niki  aakiikoaiksi
nit-iit-ino-aa-yi-aawa  ann-iksi  aakkioan-iksi
1-loc-see-dir-3pl-pro  dem-an.pl  girl-an.pl
na  Maanikapi  otaisinoi'isskipaiksi
ann-wa  Maanikapi  ot-á-sinoi'isskip-aa-iksi
dem-an.sg  Maanikapi  3-impf-kiss-dir-an.pl
otsipiissa
ot-ipii-hsi-aawa
3-enter-conj-pro
‘I saw those girls who Maanikapi was kissing.’

lexical form

sonai'isskip
‘kiss’

lexical form

no links indicates no matches

lexical form

aa
‘direct suffix:

blue links indicate exact matches

green links indicate partial matches
Morphological Modeling

- Ongoing work on implementing in Dative a graphical interface to the Morphological Parser Creator (MPC) built into the OLD web service
  - build any number of morphological parsers
    - use them to expedite data creation
    - use them to empirically evaluate the underlying models
MPC Parsers

FST

morphophonology

phonology

morphology

morphotactics

lexicon

[nítsspiyi] →

{ 
/ni-it-ihpiyi/, 
/nit-ihpiyi/, 
/nit-ihp-yi-yi/, 
...
}

ranker

N-gram language model

( 
/nit-ihpiyi/, 
/n-it-ihpiyi/, 
/nit-ihp-yi-yi/, 
...
)
Morphological Parser Creator

- phonological rewrite rules
- lexicon corpus
- morphotactics corpus
- N-gram LM corpus

parser
Morphological Parser
Creator

• tested on Blackfoot (Dunham 2014)
  • highlighted gaps in the standard morpho-phonological analysis of that language
  • parsers built were effective enough to be practical tools (but further improvements are both possible and planned)
Morphological Parser Creator

• tested on Blackfoot (Dunham 2014)
• highlighted gaps in the standard morphophonological analysis of that language
• parsers built were effective enough to be practical tools (but further improvements are both possible and planned)
Morphological Parser Creator

- tested on Blackfoot (Dunham 2014)
- highlighted gaps in the standard morphophonological analysis of that language
- parsers built were effective enough to be practical tools (but further improvements are both possible and planned)
LingSync usage: field methods classes

- Numbers: 12 classes
- Languages: Tibetan, Lezgian, Nepali, Inuktitut, Quechua, Nepali, Teenek, Igikuria, Mi’gmaq, Blackfoot, Nata, Gitksan
- Institutions: McGill University, Concordia University, University of Chicago, University of Western Ontario, Yale University, University of Connecticut, University of Ottawa, Pomona College, University of British Columbia
LingSync usage: corpora

- Numbers:
  - 86 corpora with active investigations
  - 32 corpora with more than 100 documents
- Languages (private information)
- Institutions: 48 different institutions including Harvard, MIT, Cornell, Berkeley, University of Massachusetts, etc.
Field methods experience

- Field methods class on Dharamsala Tibetan (Jessica Coon, McGill):
  - students have access to each other’s data
  - instructor has access to all data elicited by students
  - data not lost once the class ends
  - collaborative in-class elicitation
  - WordPress blog + Simple Interlinear Gloss plugin + LingSync
Field methods experience

• Field methods class on Dharamsala Tibetan (Jessica Coon, McGill):

  • students have access to each other’s data
  • instructor has access to all data elicited by students
  • data not lost once the class ends
  • collaborative in-class elicitation
  • WordPress blog + Simple Interlinear Gloss plugin + LingSync
Field methods experience

- Field methods class on Dharamsala Tibetan (Jessica Coon, McGill):
  - students have access to each other’s data
  - instructor has access to all data elicited by students
  - data not lost once the class ends
  - collaborative in-class elicitation
  - WordPress blog + Simple Interlinear Gloss plugin + LingSync
Field methods experience

- Field methods class on Dharamsala Tibetan (Jessica Coon, McGill):
  - students have access to each other’s data
  - instructor has access to all data elicited by students
  - data not lost once the class ends
  - collaborative in-class elicitation
  - WordPress blog + Simple Interlinear Gloss plugin + LingSync
Field methods experience

• Field methods class on Dharamsala Tibetan (Jessica Coon, McGill):
  • students have access to each other’s data
  • instructor has access to all data elicited by students
  • data not lost once the class ends
  • collaborative in-class elicitation
  • WordPress blog + Simple Interlinear Gloss plugin + LingSync
Field methods experience

- Field methods class on Dharamsala Tibetan (Jessica Coon, McGill):
  - students have access to each other’s data
  - instructor has access to all data elicited by students
  - data not lost once the class ends
  - collaborative in-class elicitation
  - WordPress blog + Simple Interlinear Gloss plugin + LingSync
Conclusions

• LingSync is a useful documentation tool, esp. in field methods classes and research groups

• Bring us your “fieldwork-challenged” linguists (Bettinson 2015) and your programmer linguists!
Conclusions

• LingSync is a useful documentation tool, esp. in field methods classes and research groups

• Bring us your “fieldwork-challenged” linguists (Bettinson 2015) and your programmer linguists!
Conclusions

• How to balance innovative and domain/culture-appropriate tool development while not completely re-inventing the wheel?

• design with modularity (re-usability, pluggability) in mind (web services)
Open Source

• https://github.com/OpenSourceFieldlinguistics
• https://github.com/OpenSourceFieldlinguistics/FieldDB
• https://github.com/jrwdunham/dative
lingsync.org
jrwdunham.com

- pt 2: How to Use LingSync (.pdf, video slides)
- pt 3: Hands-on LingSync (.pdf, video slides)
- pt 4: Future Developments (.pdf, video slides)

“LingSync Tutorial.” Presentation at the “I'd Like to Know More About …” Workshop, organized by the Department of Linguistics, Languages and Cultures and the Language Research Centre at the University of Calgary, November 10, 2014. Video slides.


Acknowledgements

- Alan Bale (Concordia)
- Gina Cook (iLanguage Lab Ltd)
- Jessica Coon (McGill)
- M.E. Cathcart (U Delaware)
- Theresa Deering (Visit Scotland, iLanguage Lab Ltd)
- Josh Horner (Amilia, iLanguage Lab Ltd)
- Yuliya Manyakina (McGill)
- Elise McClay (McGill)
- Gretchen McCulloch (McGill)
- Hisako Noguchi (Concordia)
- Michael Wagner (McGill)
- Jesse Pollak (Pomona College)
- Tobin Skinner (Acquisio, iLanguage Lab Ltd)
- Xianli Sun (Miami University)
- More...