LingSync & the Online Linguistic Database

New models for the collection and management of data for language communities, linguists and language learners

Joel Dunham [1] Gina Cook [2] Josh Horner [3]

[1] University of British Columbia

[2] iLanguage Lab

[3] Amilia

ComputEL Workshop, June 26 2014
52nd Annual Meeting of the Association for Computational Linguistics (ACL)

LingSync & OLD

LingSync & the Online Linguistic Database
New models for the collection and management of data for language communities, linguists and language learners

Joel Dunham [1] Gina Cook [2] Josh Horner [3]

[1]University of British Columbia

[2]iLanguage Lab

[3]Amilia

ComputEL Workshop, June 26 2014
52nd Annual Meeting of the Association for Computational Linguistics (ACL)

Hello everyone. My name is Joel and this is Gina.

1. We are here to talk to you today about LingSync and the Online Linguistic Database.
2. These are free non-proprietary open source tools that facilitate collaborative fieldwork on endangered languages.
3. This presentation will provide a very high overview of the systems for a mixed audience.
4. However, we are at a transition point since we are beginning to merge these two tools under the LingSync name. We're very excited to be here to be part of the discussion about how mobile or web technologies and computational linguistic advances can benefit fieldwork on endangered languages.

Context

- Context of rapid language loss (Harrison 2007, Krauss 1992)
- Language diversity and vitality are valuable scientifically, culturally, and socially (Hallett et al. 2007)
- Computational approaches and formalisms can benefit fieldwork (Bird 2009)

- Context of rapid language loss (Harrison 2007, Krauss 1992)
- Language diversity and vitality are valuable scientifically, culturally, and socially (Hallett et al. 2007)
- Computational approaches and formalisms can benefit fieldwork (Bird 2009)

1. So, it hardly needs to be stated that we are in a situation where a significant portion of the world's languages are predicted to be extinct or "sleeping" in the very near future.
2. For those who might be skeptical of a strong link between social well-being and language vitality Hallett et al. (2007) describes a study which finds a correlation between higher levels of Aboriginal language use in British Columbia First Nations bands and lower youth suicide rates.
3. And clearly we're all here because we think that computational linguistics can contribute to the documentation, study, and/or revitalization of endangered languages if such a collaboration can be facilitated.

Debates

- How to balance our limited time resources?

LingSync & OLD

Background

Endangered languages fieldwork

Debates

- How to balance our limited time resources?

1. There are a number of ongoing debates.
2. First, how we should balance our time given the urgent problem of language endangerment.

Debates

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)

1. Some argue that we could use the majority of our time for recording speech, re-speakings, and translations, and for generating artifacts for the future.

Debates

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)

1. Others argue that we can use advances in computational linguistics to expedite language description and analysis.

Debates

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)
  - In-depth theoretical analysis can uncover previously unknown generalizations which can make the community proud to speak their unique language (Murray 2014)

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)
  - In-depth theoretical analysis can uncover previously unknown generalizations which can make the community proud to speak their unique language (Murray 2014)

1. There is a related debate between typologists and theoretical linguists about how targeted data is elicited, including using translation from the metalanguage and the gathering of grammaticality judgments--- Murray argues, convincingly in my estimation, that hypothesis-driven fieldwork involving the elicitation of negative data can lead to the discovery of significant generalizations that, under a purely framework independant descriptive approach, would remain hidden.

Debates

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)
  - In-depth theoretical analysis can uncover previously unknown generalizations which can make the community proud to speak their unique language (Murray 2014)
- A duty to collaborate?

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)
  - In-depth theoretical analysis can uncover previously unknown generalizations which can make the community proud to speak their unique language (Murray 2014)
- A duty to collaborate?

1. A related debate asks whether the various types of fieldworkers have a duty to collaborate.

Debates

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)
  - In-depth theoretical analysis can uncover previously unknown generalizations which can make the community proud to speak their unique language (Murray 2014)
- A duty to collaborate?
  - Language revitalization is paramount; collaboration is imperative (Gerdts 2010)

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)
  - In-depth theoretical analysis can uncover previously unknown generalizations which can make the community proud to speak their unique language (Murray 2014)
- A duty to collaborate?
  - Language revitalization is paramount; collaboration is imperative (Gerdts 2010)

1. At one extreme lies the view that the time overhead which collaboration entails is necessary despite the loss of academically publishable results of fieldwork.

Debates

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)
  - In-depth theoretical analysis can uncover previously unknown generalizations which can make the community proud to speak their unique language (Murray 2014)
- A duty to collaborate?
  - Language revitalization is paramount; collaboration is imperative (Gerdts 2010)
  - Collaboration is not always desired (Crippen & Robinson 2013)

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)
  - In-depth theoretical analysis can uncover previously unknown generalizations which can make the community proud to speak their unique language (Murray 2014)
- A duty to collaborate?
  - Language revitalization is paramount; collaboration is imperative (Gerdts 2010)
  - Collaboration is not always desired (Crippen & Robinson 2013)

1. Others point out that, for a variety of reasons, collaboration is not always possible. The contexts of language endangerment are themselves diverse. Not all communities are commited to documentation or revitalization. Not all communities need or desire the help of academic fieldworkers. In some cases, the political landscape of an endangered language community is just too complex for a field linguist to navigate. In these types of situations, the best course of action for all parties involved may be for linguist fieldworkers to respectfully and ethically pursue their own research program.

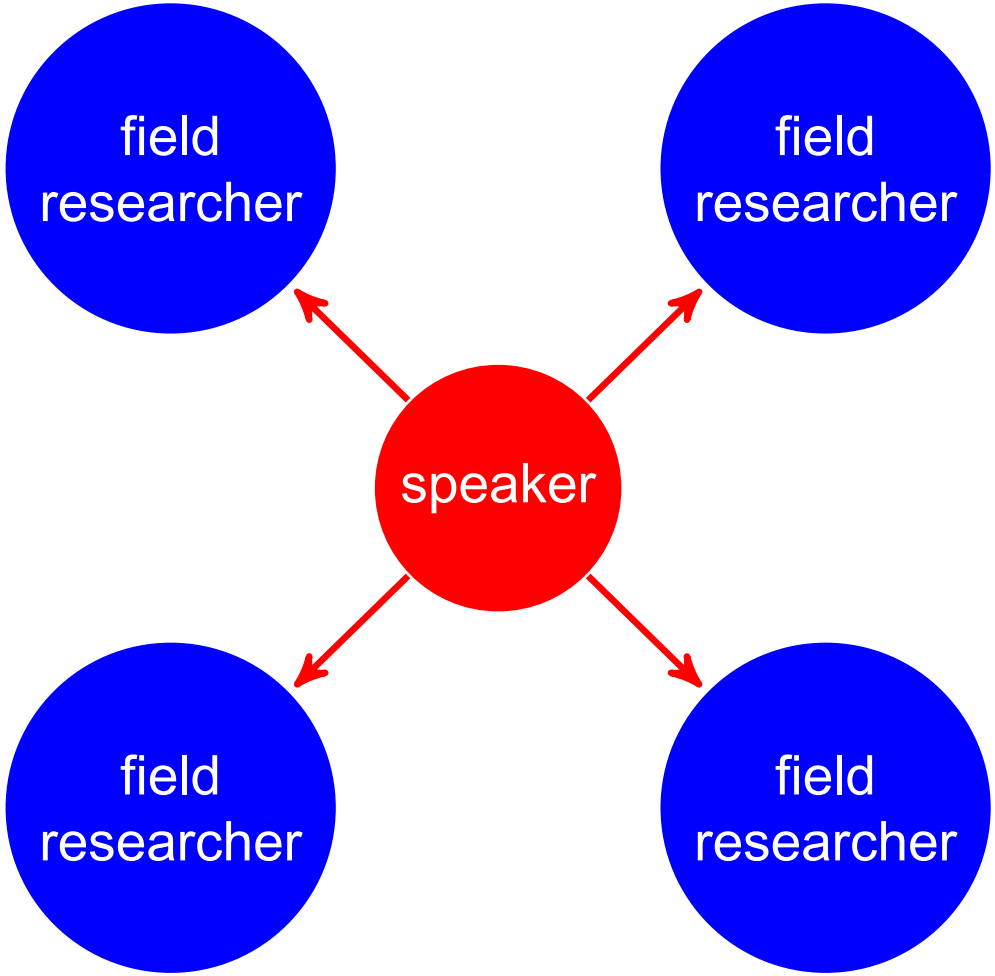
Debates

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)
  - In-depth theoretical analysis can uncover previously unknown generalizations which can make the community proud to speak their unique language (Murray 2014)
- A duty to collaborate?
  - Language revitalization is paramount; collaboration is imperative (Gerdts 2010)
  - Collaboration is not always desired (Crippen & Robinson 2013)
- Whatever we decide, let's maximize openness, transparency, access, sharing and reuse of both data

- How to balance our limited time resources?
  - Less descriptive artifacts, more aligned speech recording, more grass-roots community involvement can revitalize the language (Woodbury 2003, Bird et al. 2014)
  - Better descriptive artifacts can prevent “data graveyards” (Gippert et al. 2006, cf. Beale 2014)
  - In-depth theoretical analysis can uncover previously unknown generalizations which can make the community proud to speak their unique language (Murray 2014)
- A duty to collaborate?
  - Language revitalization is paramount; collaboration is imperative (Gerdts 2010)
  - Collaboration is not always desired (Crippen & Robinson 2013)
- Whatever we decide, let's maximize openness, transparency, access, sharing and reuse of both data

1. These are complex and relevant issues. However, to a certain extent the LingSync/OLD approach can sidestep them. The attitude we advocate is something along the lines of "can't we all just get along?". We want to help contribute towards an infrastructure which facilitates, but does not require, collaboration between stakeholders, a data structure which is flexible and yet structured enough to be useful to different types of linguist, translator, language teacher and fieldworker. We want to help fieldworkers to make their data more useful to their peers without incurring significant loss of their time for their own research program.

Collaboration

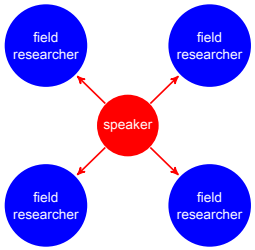


LingSync & OLD

Background

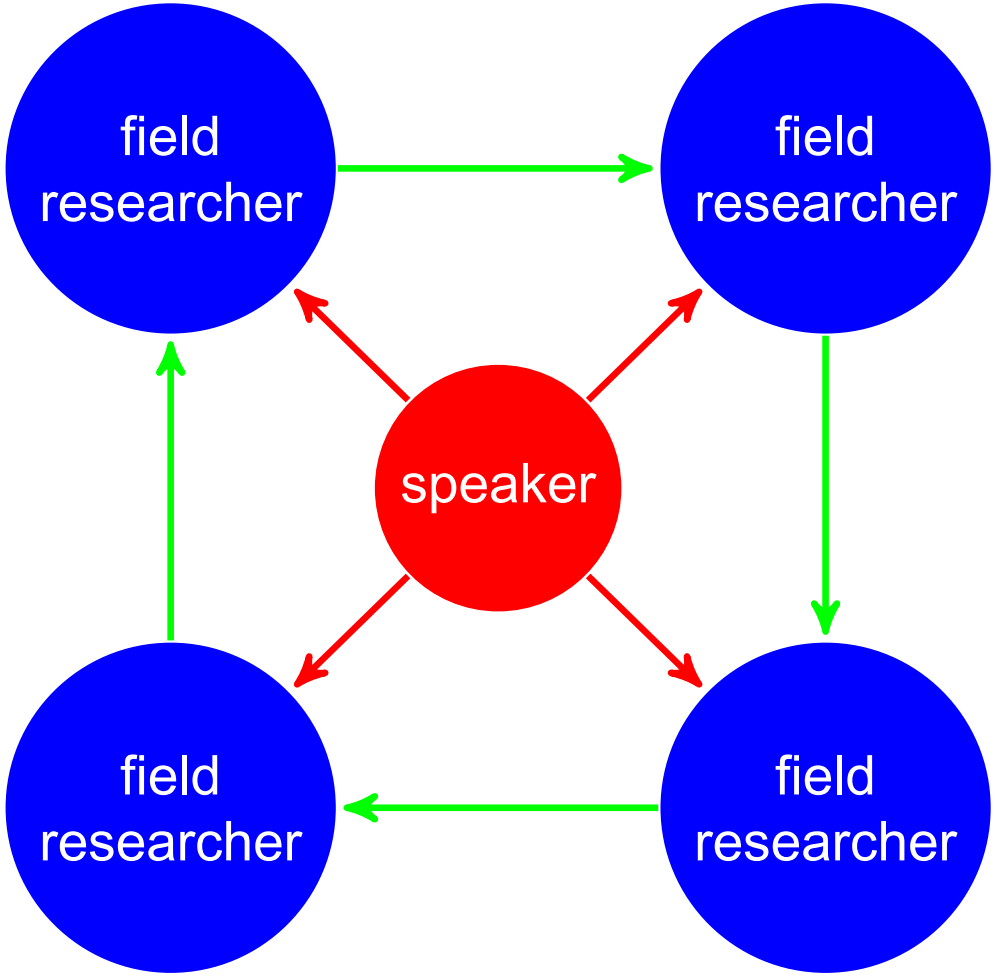
Endangered languages fieldwork

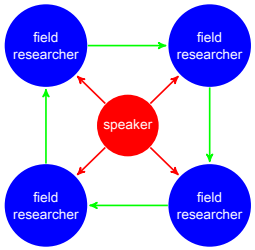
Collaboration



blank note.

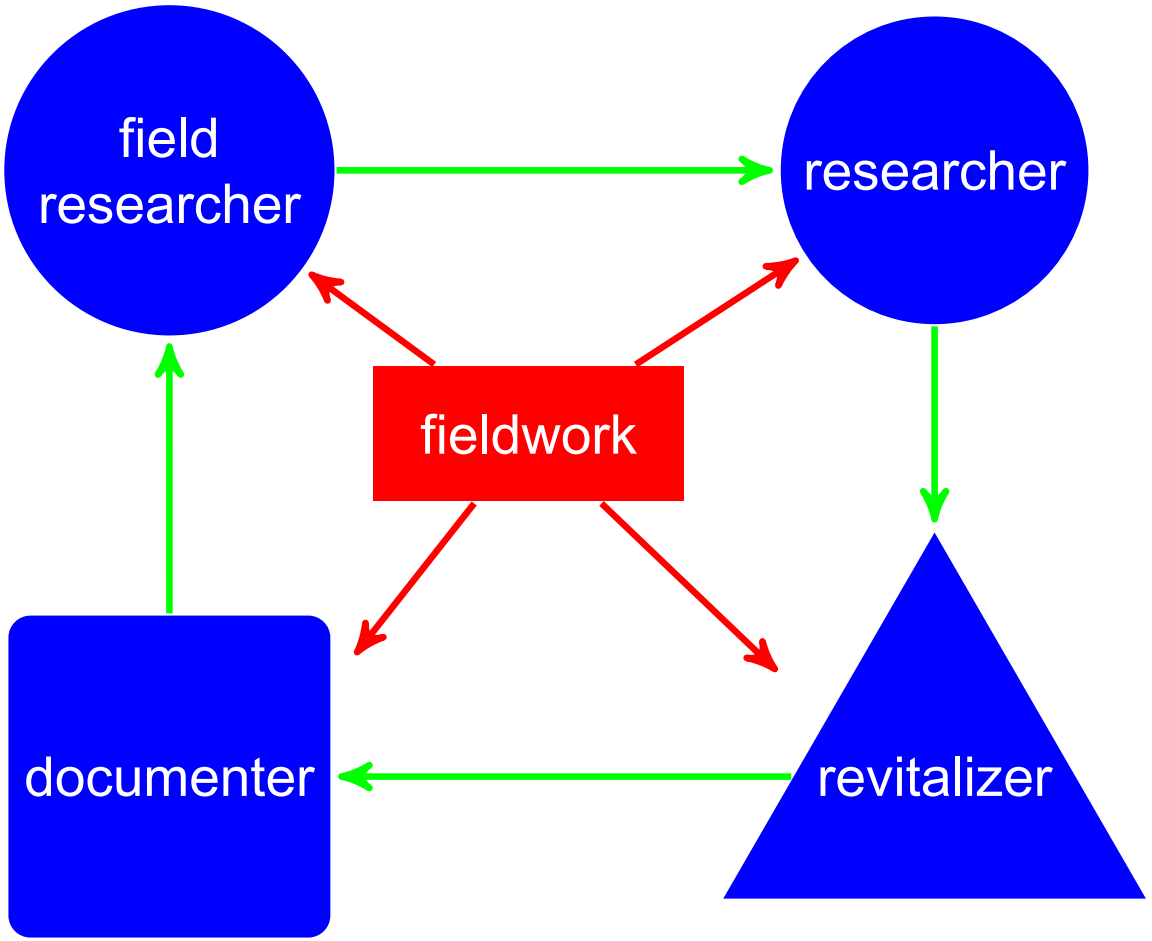
Collaboration



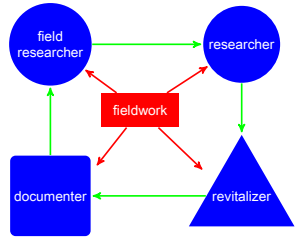

field researcher
field researcher
speaker
field researcher
field researcher

blank note.

Collaboration

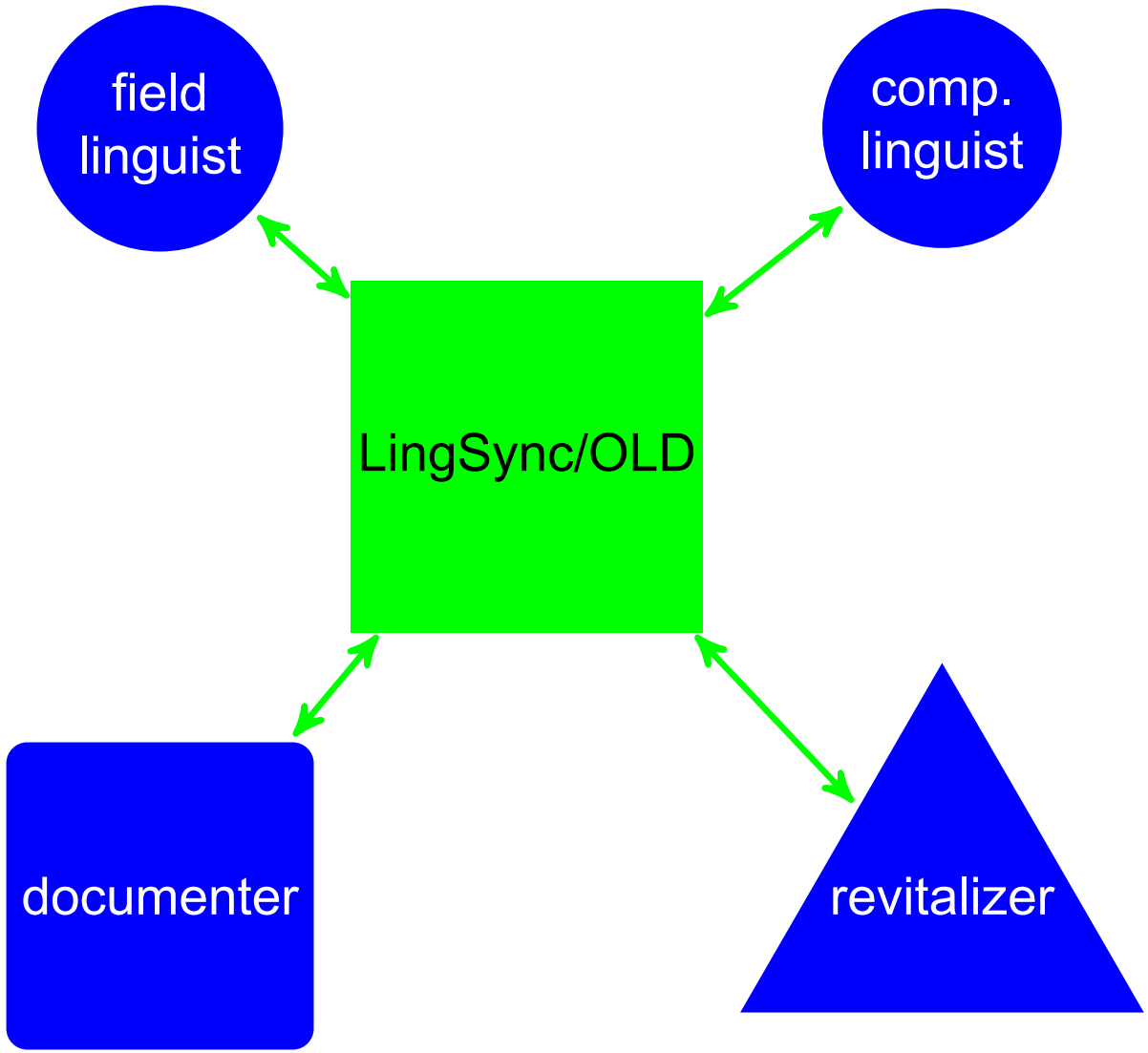


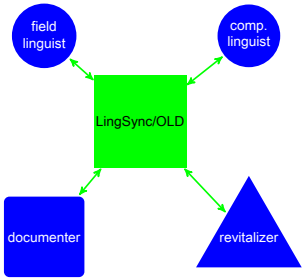
LingSync & OLD
Background
Endangered languages fieldwork
Collaboration



blank note.

Collaboration





1. LingSync helps language workers of various types and with different goals to share data and collaborate, if they want to.
2. As will be discussed on YouTube, the systems offer features and conveniences which respond to the requirements of different types of fieldworker and fieldwork situation and which may make the system worthwhile beyond the primary collaboration- and data-sharing functionality.

Requirements

1. So here we will briefly review the requirements that guided the development of LingSync and the OLD.

Requirements

Requirement 1 . Integration of primary data

1. Rather uncontroversially, the software must be able to handle primary data as a first-class citizen of the system. In particular, we need to help fieldworkers share audio and video recordings (including experimental stimuli). The system should allow for the alignment of audio/video with transcriptions and other textual data; the audio/video and textual data should be displayed simultaneously for easy cross-reference. It should be possible to record audio right into the application and, if possible, the text/audio alignment process should be automated or partially automated and audio/video should be searchable.

Requirements

Requirement 1 . Integration of primary data
Requirement 2 . Curation of data

1. The system should facilitate the curation of data, that is, its iterative and collaborative refinement over time. For example, the initial output of elicitation could be simply an audio recording, metadata about the source of the recording and a transcription of salient forms. Then, subsequent waves of data curation can involve transcription at various levels and/or the creation of morphological analyses and annotations of various types, for example, tagging and categorizing. Various automations of the data curation process are should be easy to script for power users.

Requirements

Requirement 1 . Integration of primary data
Requirement 2 . Curation of data
Requirement 3 . Inclusion of stakeholders

1. The software should allow for the inclusion of the various stakeholders in the endangered languages fieldwork enterprise. That is, the software should be useful for language community members, fieldworkers engaged in community-based documentation, education, and revitalization projects as well as linguistic research teams with members of various types of expertise and primary focus (e.g., theoretical, typological, historical, computational)

Requirements

Requirement 1 . Integration of primary data
Requirement 2 . Curation of data
Requirement 3 . Inclusion of stakeholders
Requirement 4 . Openable data

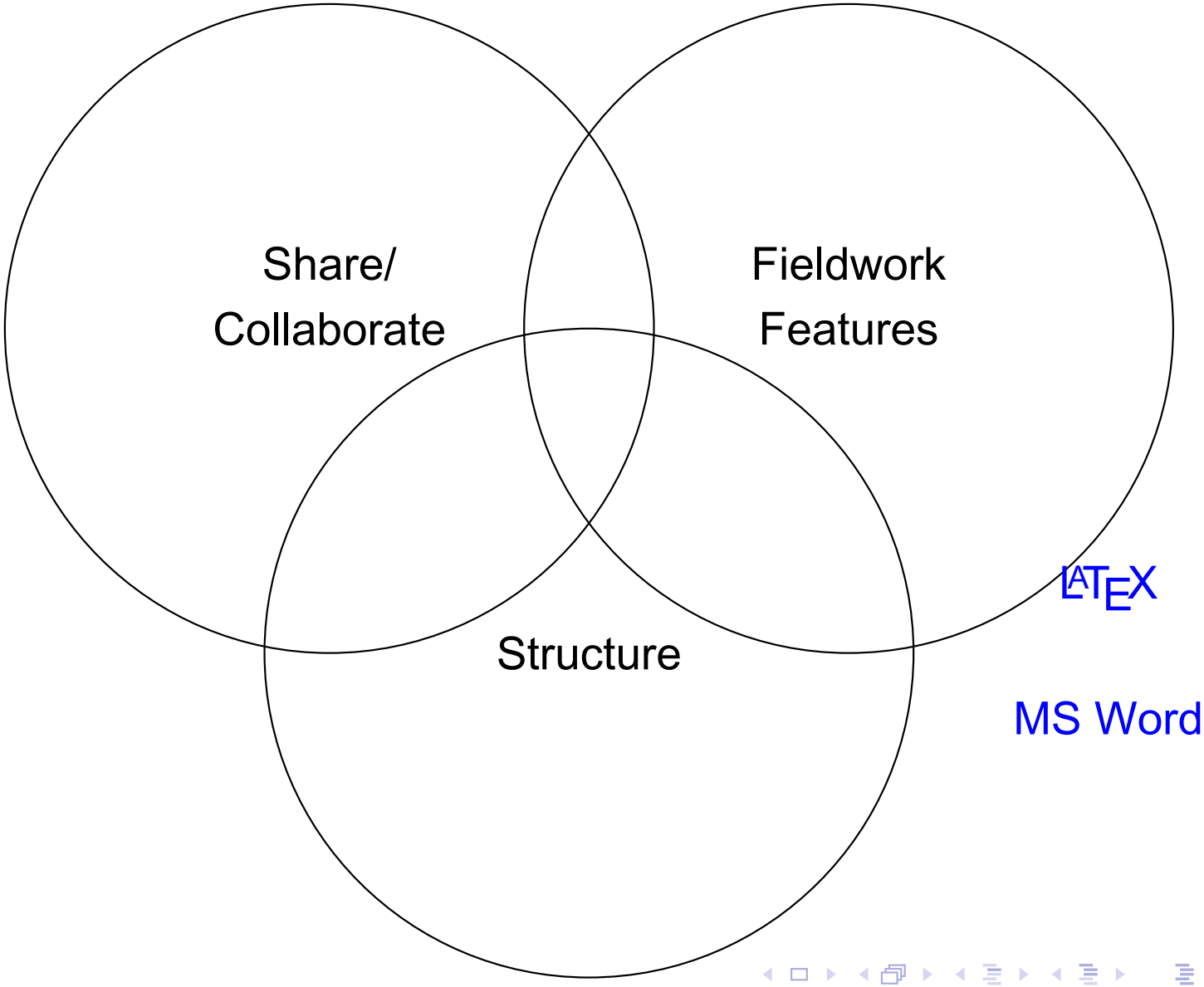
1. The system should allow fieldworkers to produce data which is relatively easy to Open. That is, data are available for reuse via various GUIs and APIs while the field work is underway a corpus should also be configurable such that access to portions of the data can be restricted to respect licensing and informed consent forms which speakers and communities have requested, if necessary.

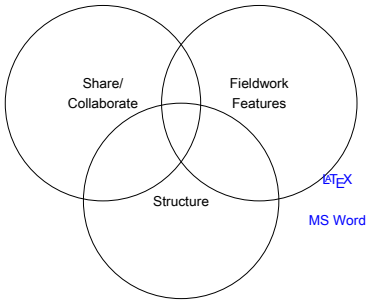
Requirements

Requirement 1 . Integration of primary data
Requirement 2 . Curation of data
Requirement 3 . Inclusion of stakeholders
Requirement 4 . Openable data
Requirement 5 . User productivity

1. The system should enable user productivity. It should include well-designed and cute user interfaces as well as conveniences which speed up repetitive tasks. It should also help with tasks that are particular to linguistic fieldwork without making the user interface complex or clunky.

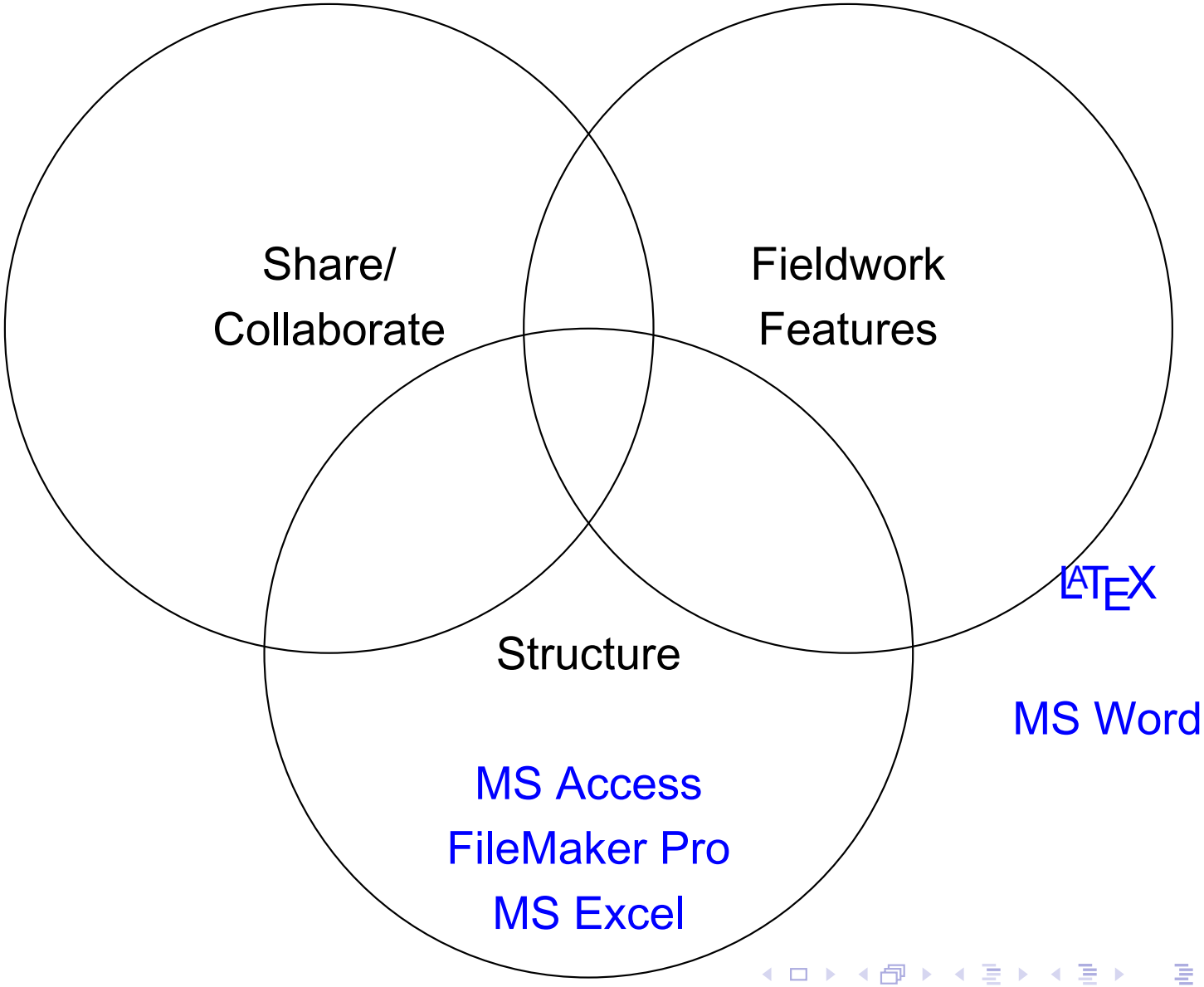
Existing Software

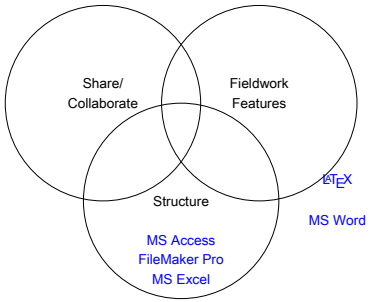




1. Most field workers use Microsoft Word to type up their data. The user interface is one they already know so there is no training to be done for them to immediately begin entering data. While efficient for the needs of field workers the data is difficult to search and reuse by collaborators.

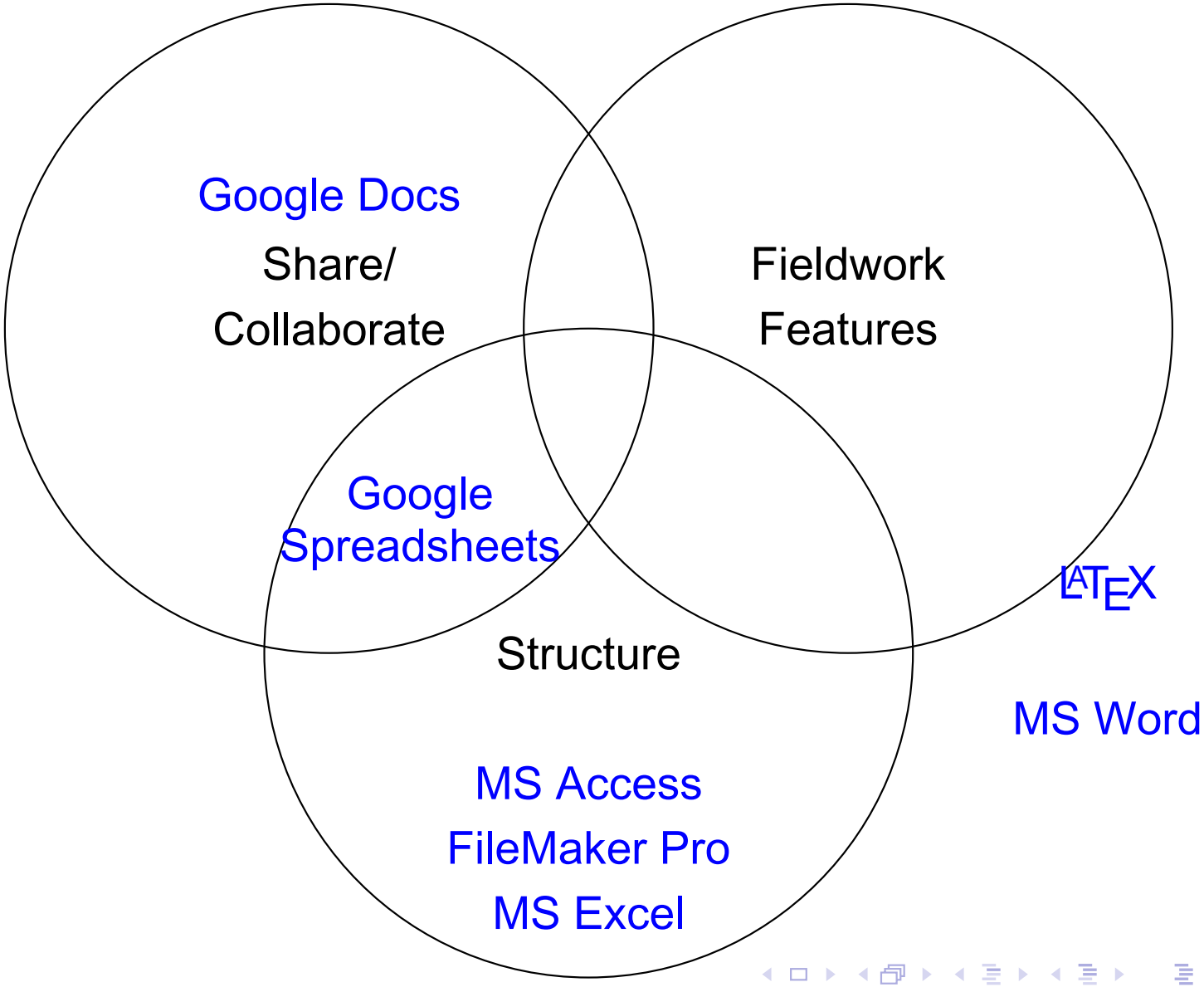
Existing Software

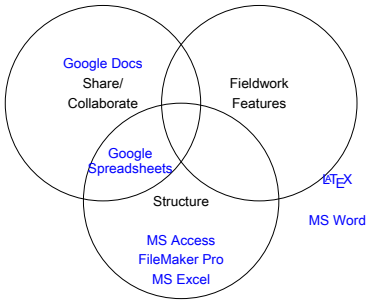




1. Microsoft Excel, Microsoft access and FileMaker Pro are more structured and produce data which is easier for computational linguist collaborators to use.

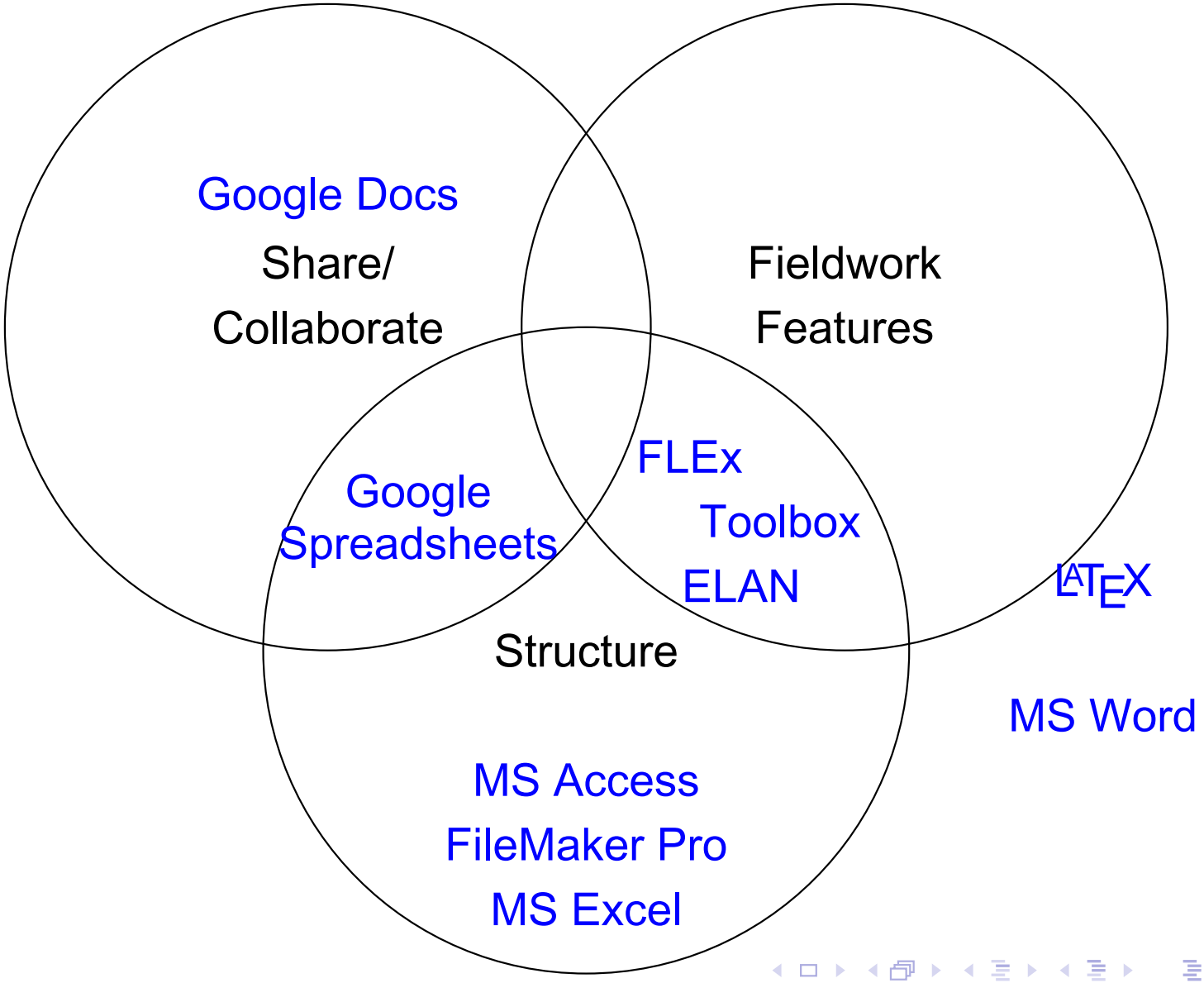
Existing Software

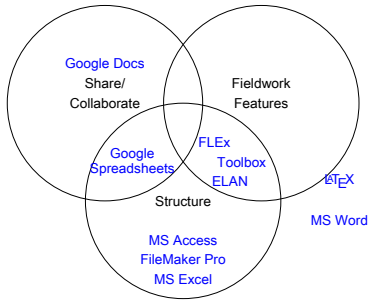




1. Google Docs are better than Microsoft Word in that multiple users can view and edit the data at the same time.
2. Google Spreadsheet is even better Google Docs in that the data is structured and can be accessed using a programming interface API.

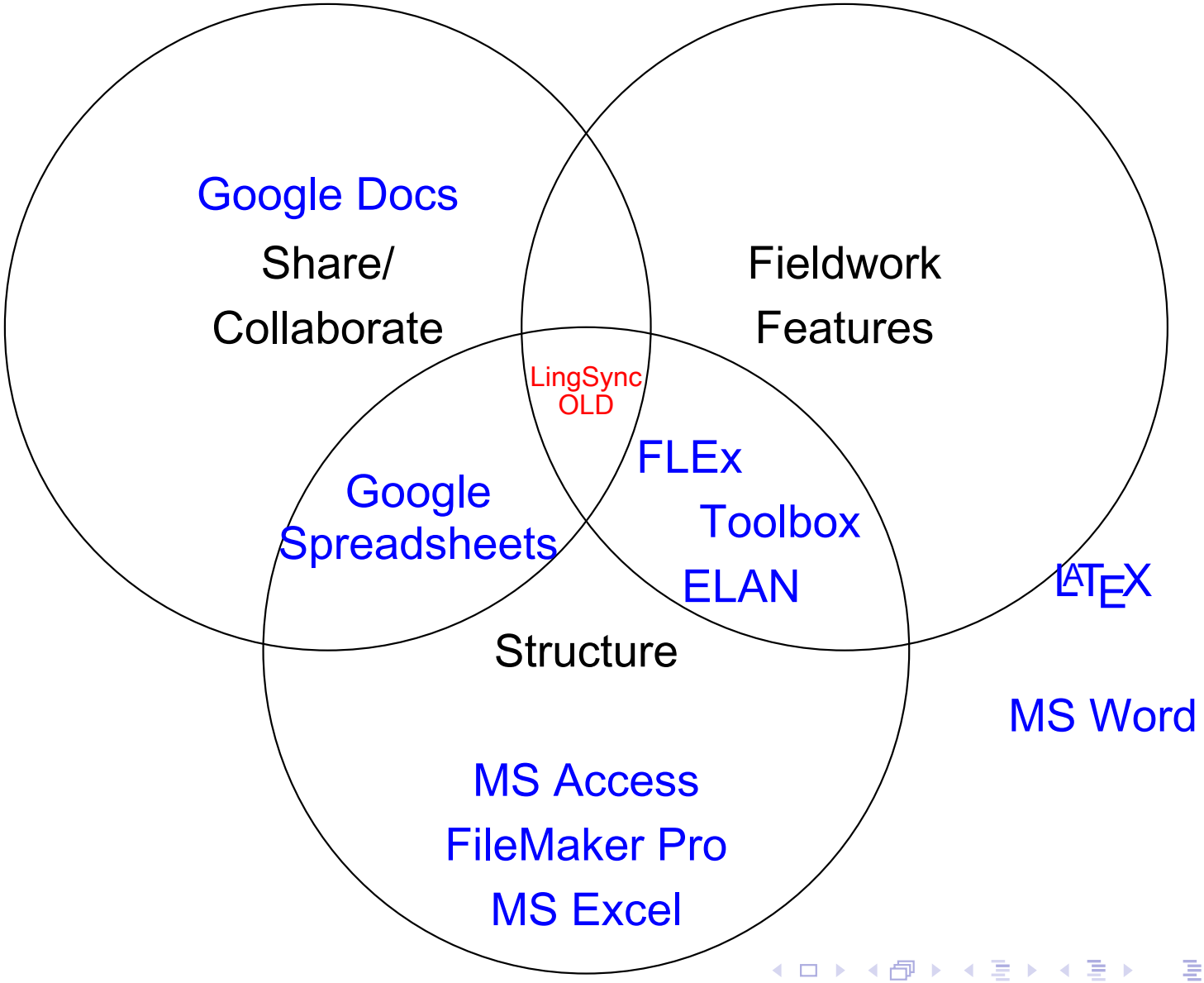
Existing Software

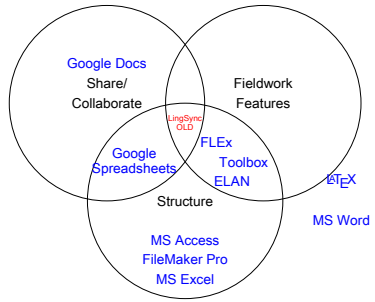




1. On the other hand we have FLEx, Toolbox, and ELAN which provide features specifically designed to facilitate fieldwork tasks: such as presentation of data in IGT format, grammar modelling and automated morphological parsing, and export to formats commonly used by linguists.

Existing Software





1. Like Google Spreadsheets, LingSync and the OLD allow multiple contributors to share data but they also support field work features and integrate well with existing tools for field work.

Ad hoc Solutions

	Offline	Doesn't require programming	Runs on non-windows computers	Text	NoSQL	Cloud	Works on mobile devices/context	Export	Version control	Scriptable	Collaboration	Doesn't require training	API	Searchable	Import	Open Source	Scaleable	Non-Proprietary Storage	Audio	Video	IGT assistance	Discoverable
LingSync/OLD	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓+	✓	✓	✓	✓	✓	✓	✓	✓
Google Spreadsheet	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓									
Google Docs	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓									✓	
EMELD & OLAC		✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓		✓	✓	✓	?		✓
Video Archives/ YouTube		✓	✓		✓	✓	✓				✓	✓	✓							✓	✓	✓
LDC & Git			✓	✓	✓	✓	✓	✓	✓	✓	✓-		✓	✓	✓		✓	✓	✓	✓		
Gate & Git	✓		✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓			✓	
Protege & Git	✓		✓		✓	✓		✓	✓	✓	✓-		✓	✓+	✓	✓	✓	✓				
Flex	✓	✓	✓	✓	✓			✓			✓-			✓+	✓	✓			✓-			
ToolBox	✓	✓		✓	✓			✓						✓	✓							
LaTeX & Dropbox	✓		✓	✓	✓	✓	✓		✓	✓	✓-	✓				✓		✓				
Word & Dropbox	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓-	✓										
FileMaker	✓	✓	✓	✓				✓		✓	?	✓		✓	✓							
Elan	✓	✓	✓	✓	✓			✓		✓					✓	✓		✓	✓	✓	?	
SFS & Git	✓		✓	✓	✓	✓		✓	✓	✓	✓-			✓	✓			✓	✓		✓	
Praat & Git	✓	✓	✓	✓	✓			✓	✓	✓	✓-					✓		✓	✓			
Excel	✓	✓	✓	✓				✓		✓		✓										
Paper	✓	✓	✓	✓	✓		✓					✓										

Figure: Many ad hoc software combinations are used by teams.

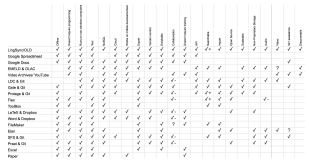
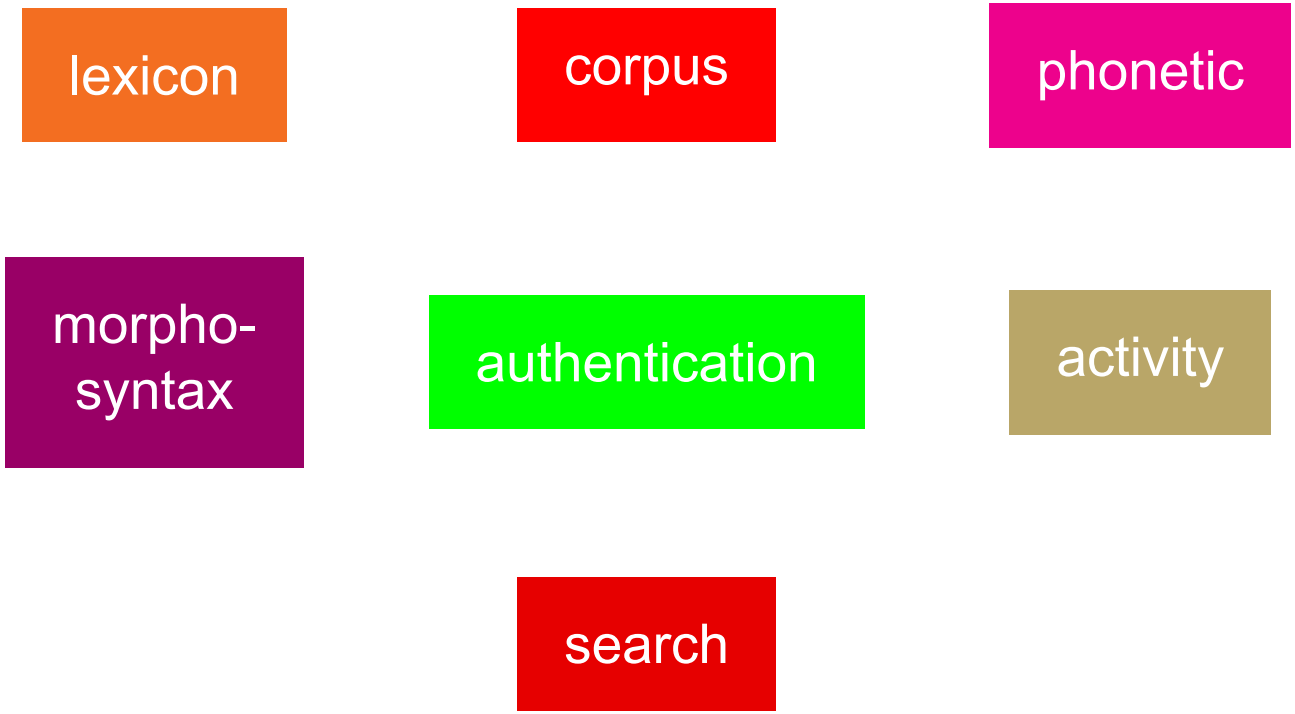


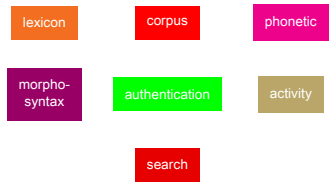
Figure: Many ad hoc software combinations are used by teams.

1. There are more than just three levels of comparison. In this table you can see the multitude of other ad hoc combinations which fieldworkers use to meet their requirements.
2. There is no one solution which can facilitate collaborative inclusive curation of data, while fieldwork is underway.
3. This is why we began glueing together existing open source modules and providing cute user interfaces to create what has come to be LingSync.

LingSync Architecture

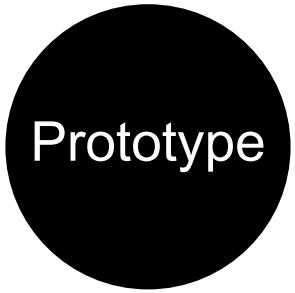


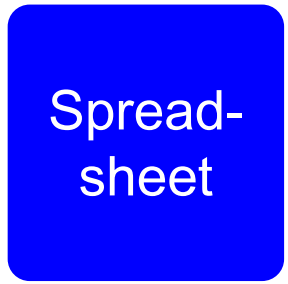
LingSync & OLD
New models for data collection and management
Architecture
LingSync Architecture

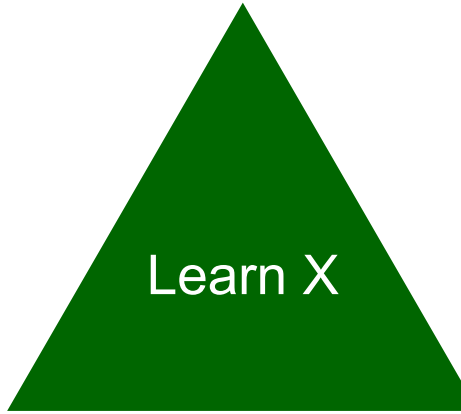


LingSync has many web services

LingSync Architecture





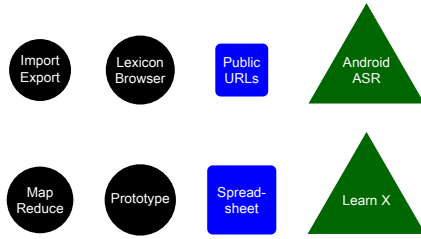


LingSync & OLD

New models for data collection and management

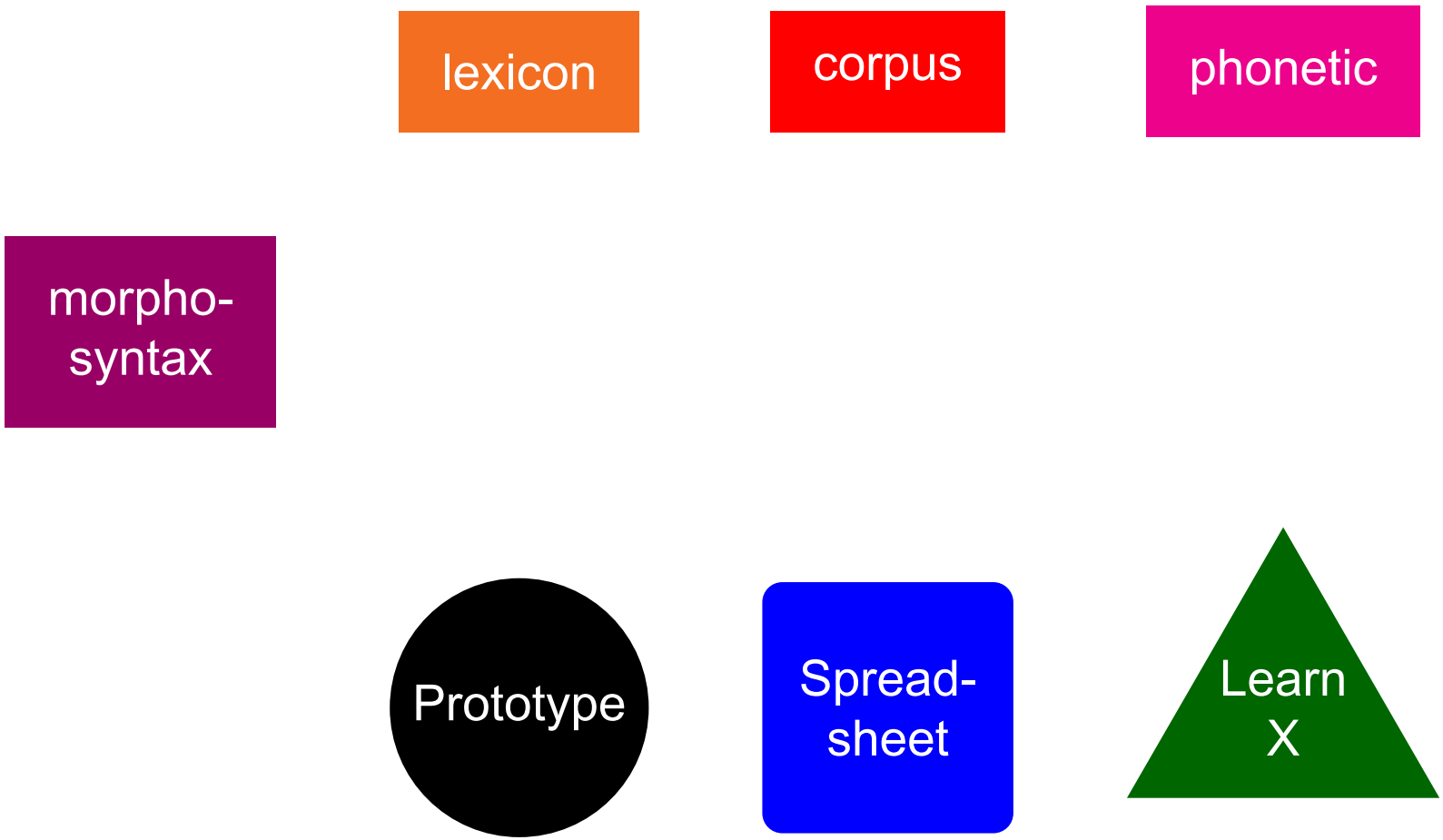
Architecture

LingSync Architecture



And many user interfaces for different stakeholders, in both mobile and desktop contexts

LingSync Architecture

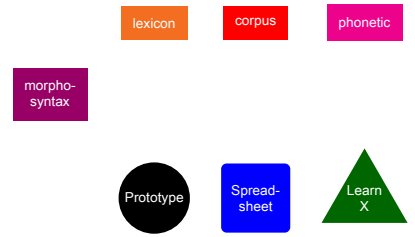


LingSync & OLD

New models for data collection and management

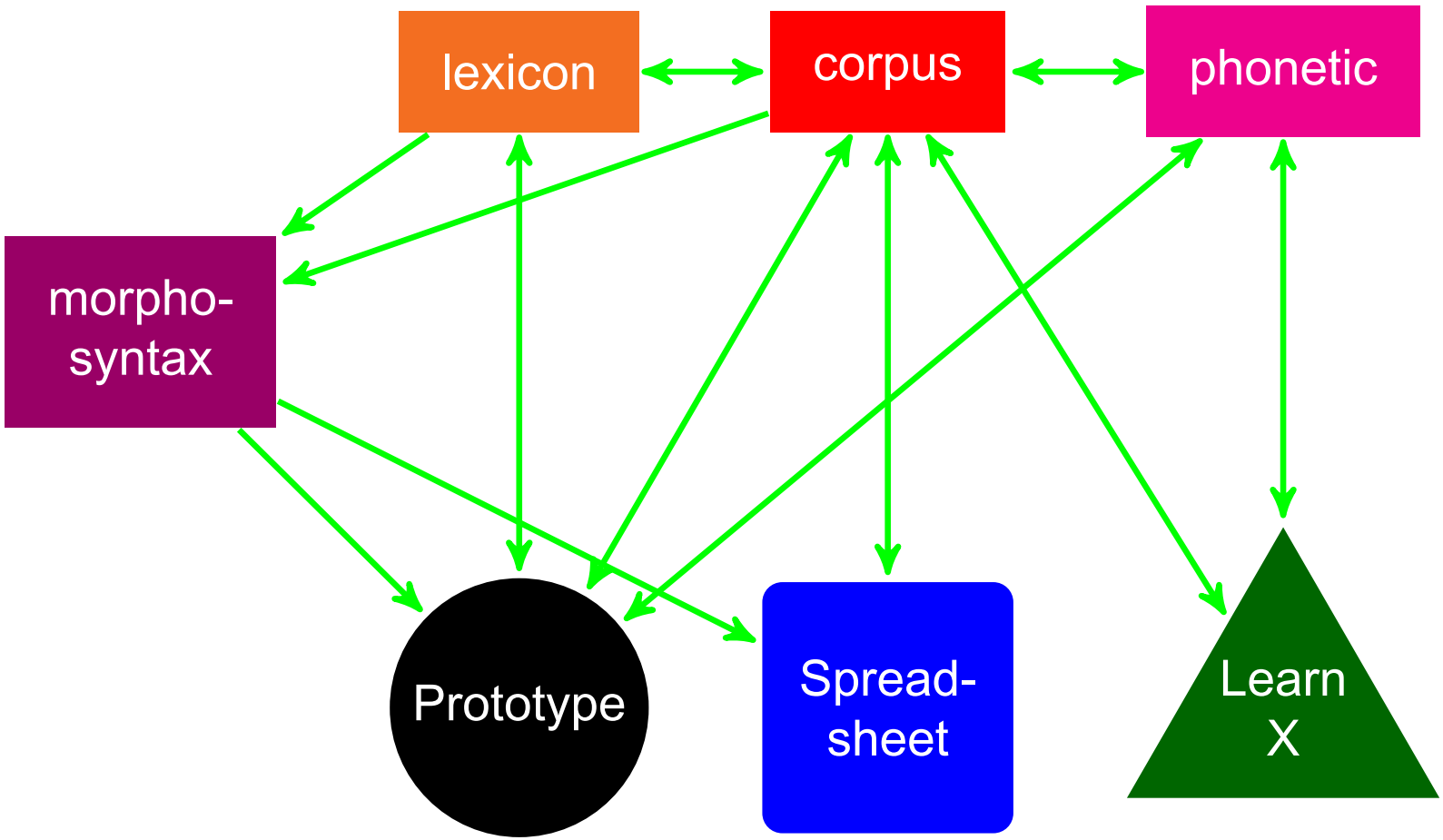
Architecture

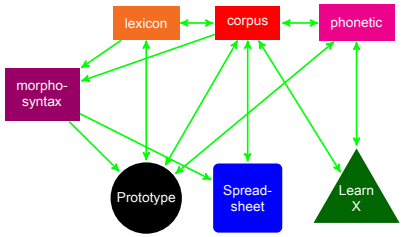
LingSync Architecture



Lets look how the core web services and user interfaces

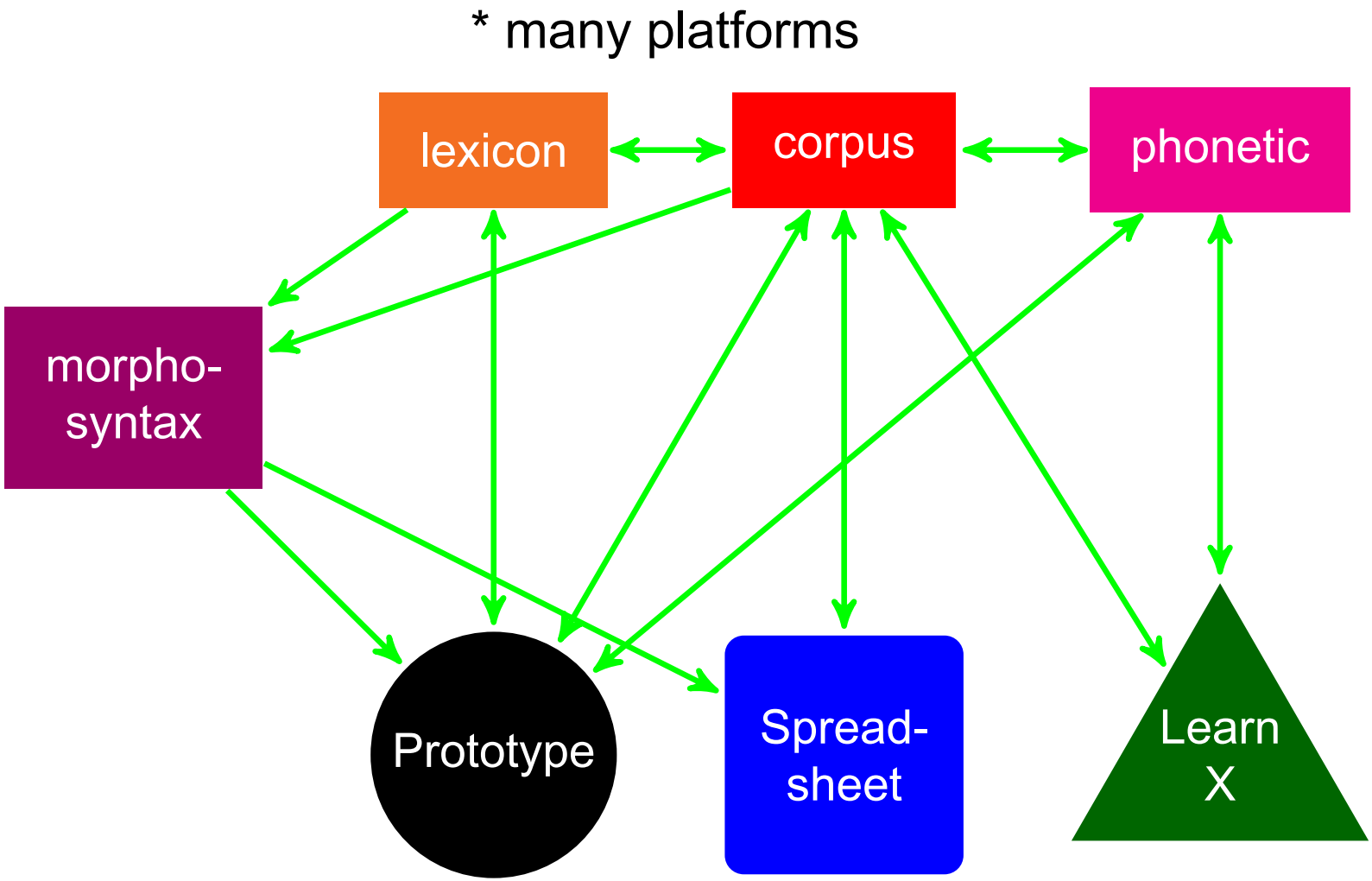
LingSync Architecture





are connected

LingSync Architecture

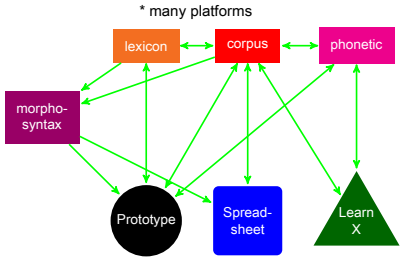


LingSync & OLD

New models for data collection and management

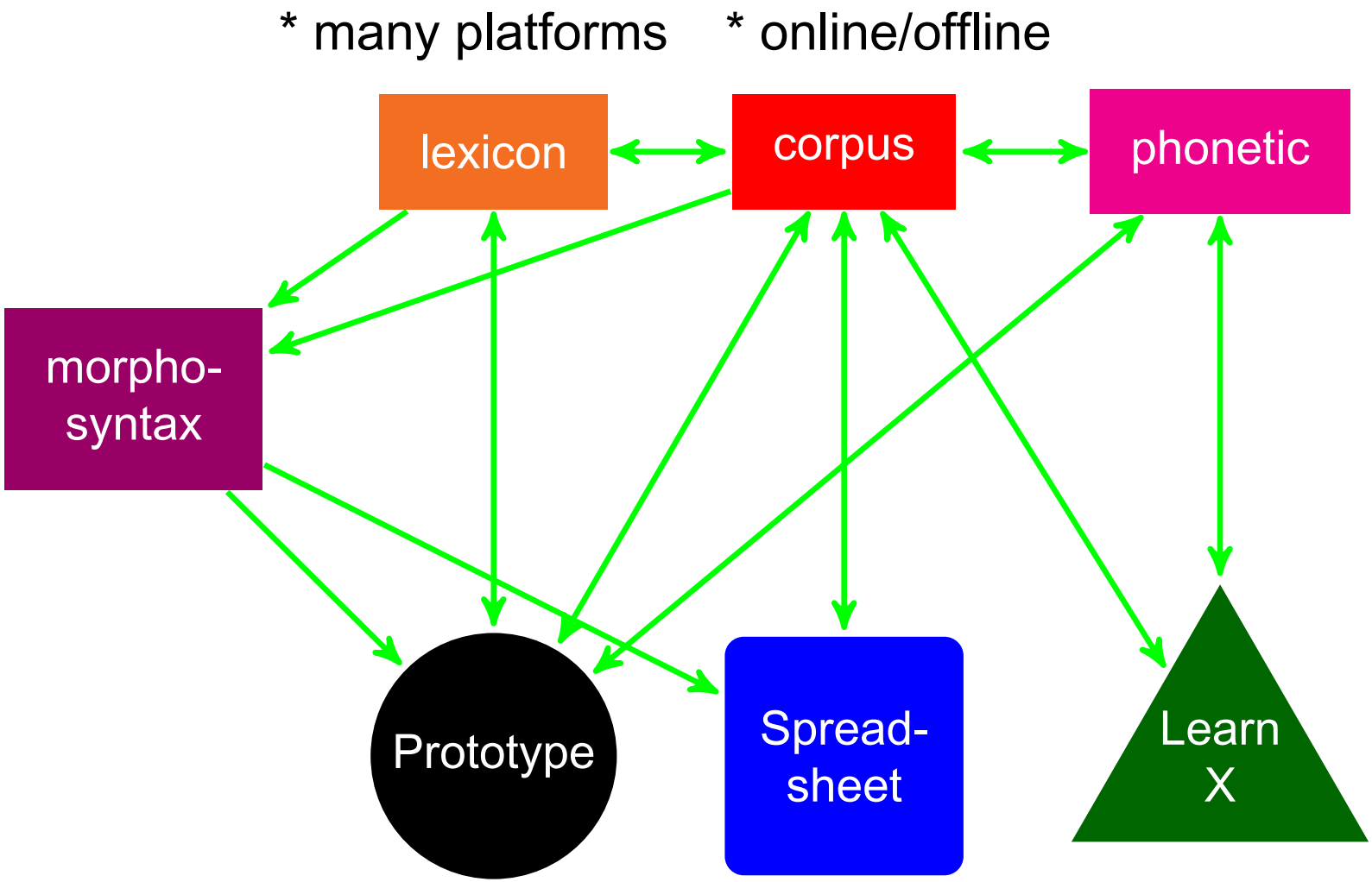
Architecture

LingSync Architecture



to serve many platforms

LingSync Architecture

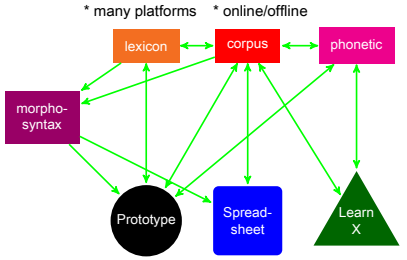


LingSync & OLD

New models for data collection and management

Architecture

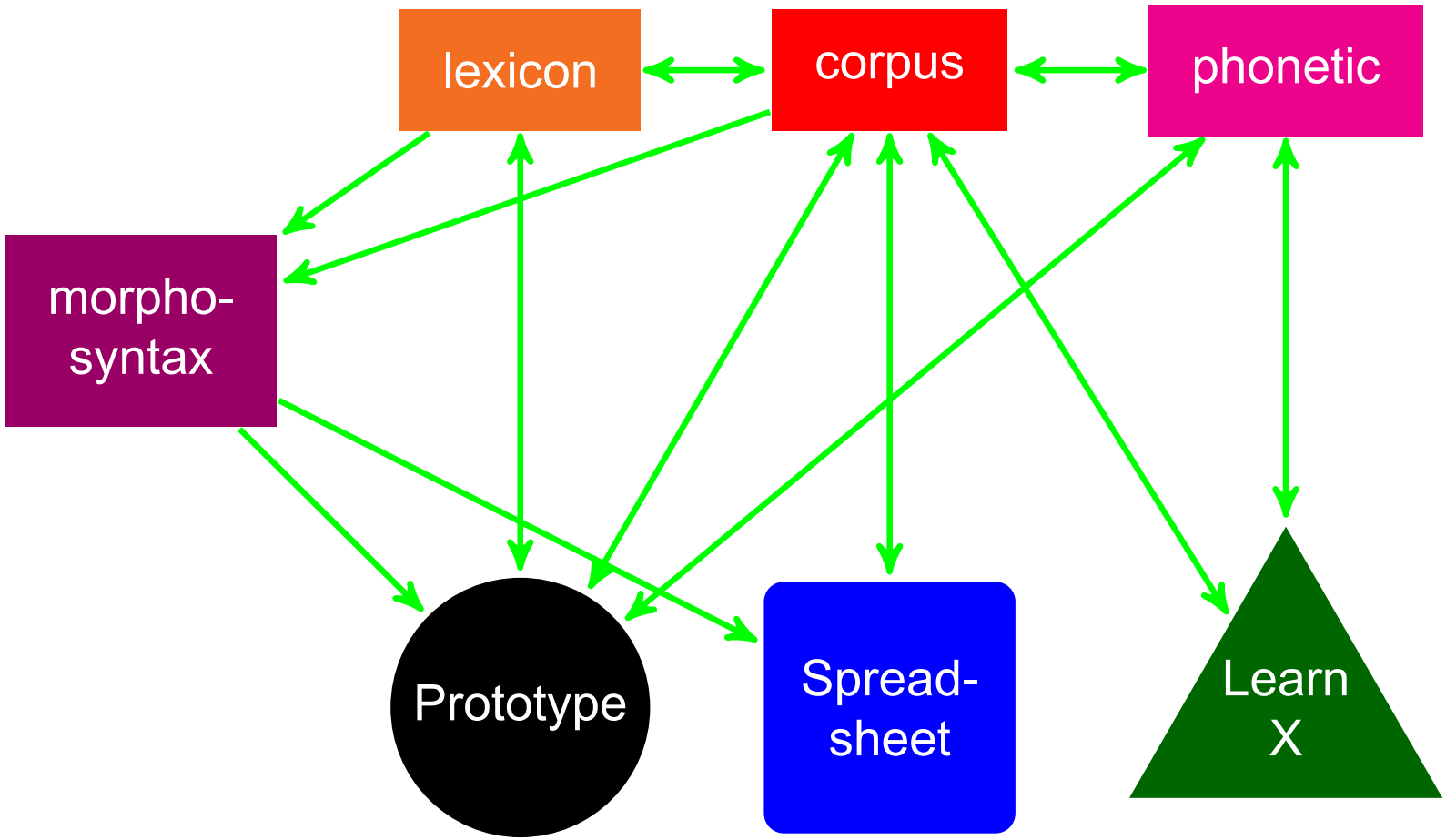
LingSync Architecture

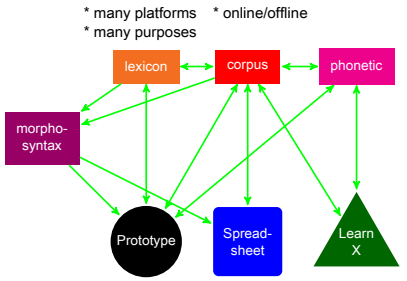


in online, offline and low bandwidth situations

LingSync Architecture

* many platforms
* many purposes
* online/offline

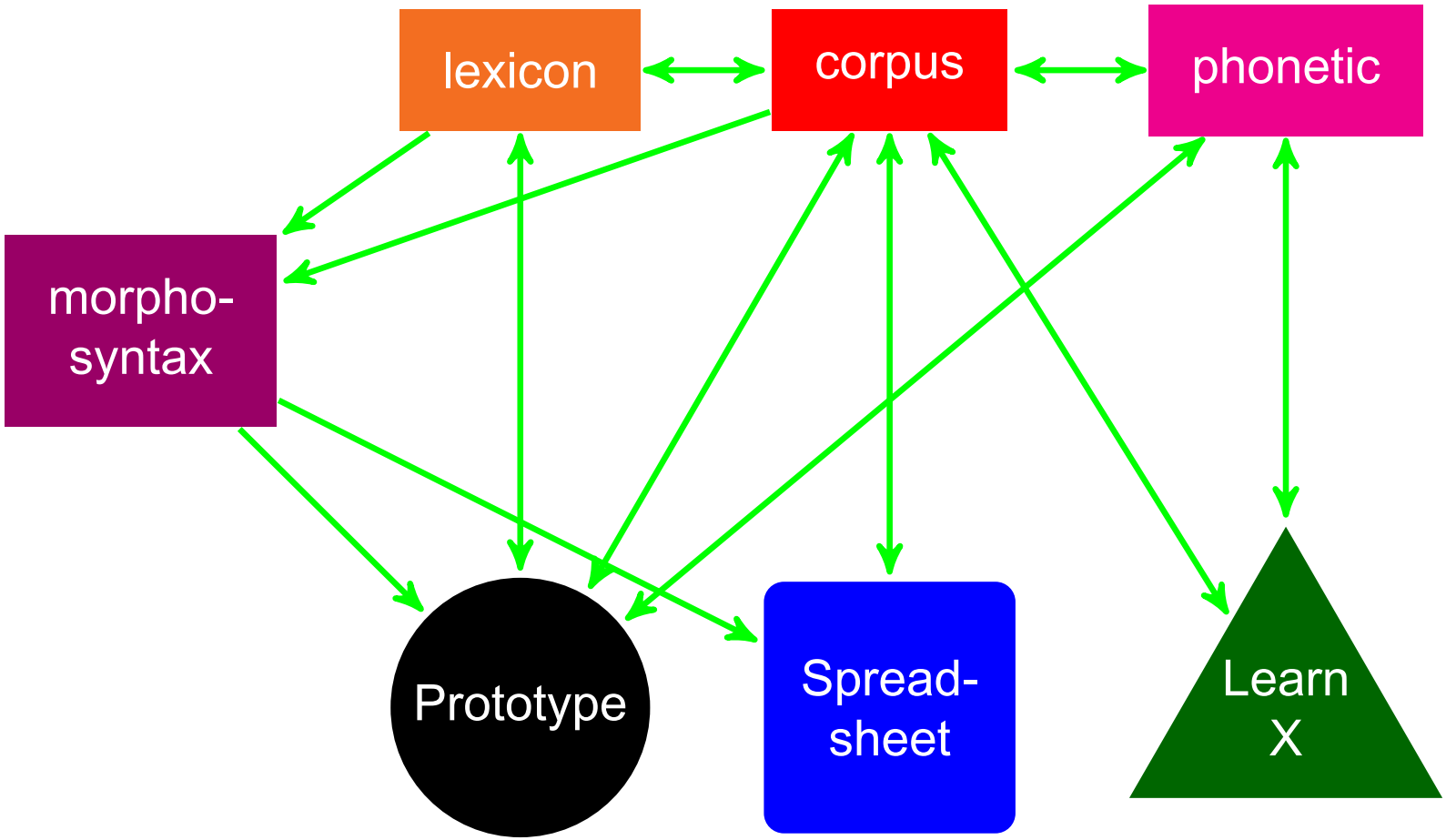




letting the data grow for teams with many purposes

LingSync Architecture

* many platforms
* many purposes
* online/offline
* extendable

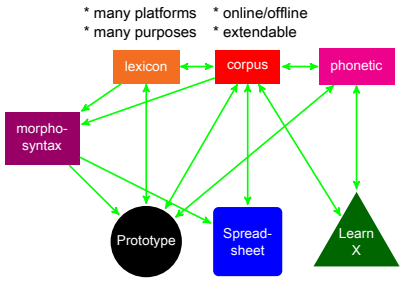


LingSync & OLD

New models for data collection and management

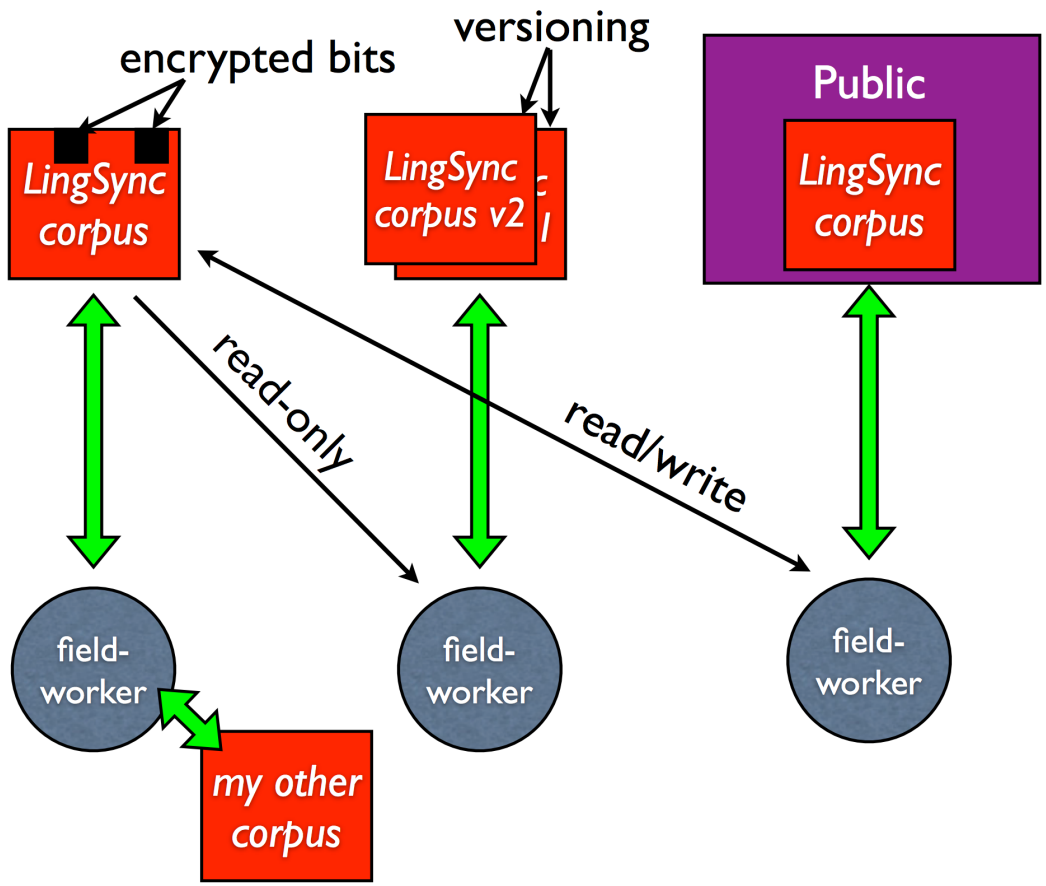
Architecture

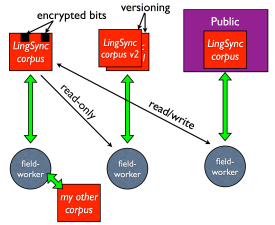
LingSync Architecture



the architecture is modular and extendable

Corpora





1. A user can create any number of corpora. They may grant access at various levels to any one of his/her corpora. A corpus may also be made public so that it can be accessed without password-based authentication.
2. Portions of a corpus can be encrypted at a fine-grained level if that kind of control over access is required.
3. Finally, all data has version numbers which means that changes can be undone and are traceable to cleaning scripts or humans who made the changes. Clearly, this is an important feature in the context of collaborative data creation.

Generality in data structure

- FLEx, Toolbox, etc.
  - lexical entry is primary
  - Boasian trilogy: texts, grammar, and dictionary
- LingSync and OLD
  - datum/form is primary
  - elicitations, corpora, texts, grammar, dictionary, handouts, language lessons

- FLEx, Toolbox, etc.
  - lexical entry is primary
  - Boasian trilogy: texts, grammar, and dictionary
- LingSync and OLD
  - datum/form is primary
  - elicitations, corpora, texts, grammar, dictionary, handouts, language lessons

1. The data structures assumed by LingSync and the OLD are arguably more general than those of similar applications like FLEx and Toolbox.
2. This greater generality allows these tools to be useful to a wider range of fieldworkers.
3. The fundamental unit of data is something quite unconstrained which in LingSync is called a "datum" and in the OLD is called a "form".
4. This abstract data unit may be used to represent sentences, phrases, words, or morphemes.
5. Texts, corpora, and records of elicitation sessions can then be constructed as (possibly ordered) sets of these data points.
6. Similarly, grammars can be created as texts which embed via reference these data points. And dictionaries could be constructed from these units as well.

Generality in data structure

- FLEx, Toolbox, etc.
  - lexical entry is primary
  - Boasian trilogy: texts, grammar, and dictionary
- LingSync and OLD
  - datum/form is primary
  - elicitations, corpora, texts, grammar, dictionary, handouts, language lessons

- FLEx, Toolbox, etc.
  - lexical entry is primary
  - Boasian trilogy: texts, grammar, and dictionary
- LingSync and OLD
  - datum/form is primary
  - elicitations, corpora, texts, grammar, dictionary, handouts, language lessons

1. This contrasts with the data structures that underpin FLEx and Toolbox; these tools assume that a grammar and a dictionary with supporting texts are the ultimate goals of the fieldworkers who use them. However, this is not always the case.

User Adoption

	Active	Investigating	In-active	Total
Public Corpora	2	1	2	5
Private Corpora	15	37	321	373
Users	38	43	220	301
Documents	13,408	2,763	4,541	23,487
Disk Size	1GB	.9GB	5.3GB	7.2GB

Table: Data in LingSync corpora (Feb 14, 2014). Active corpora: $>$300 activities; Investigating corpora: 300-10 activities; Active users: $>$100 activities; Investigating users: 100-10 activities.

LingSync & OLD

New models for data collection and management

User adoption

User Adoption

	Active	Investigating	In-active	Total
Public Corpora	2	1	2	5
Private Corpora	15	37	321	373
Users	38	43	220	301
Documents	13,408	2,763	4,541	23,487
Disk Size	1GB	.9GB	5.3GB	7.2GB

Table: Data in LingSync corpora (Feb 14, 2014). Active corpora: >300 activities; Investigating corpora: 300-10 activities; Active users: >100 activities; Investigating users: 100-10 activities.

1. There are, in total, some 300 users, 400 corpora and 24,000 documents, i.e., the general-purpose data points mentioned above.

language	forms	texts	audio	GB	speakers
Blackfoot (bla)	8,847	171	2,057	3.8	3,350
Nata (ntk)	3,219	32	0	0	36,000
Gitksan (git)	2,174	6	36	3.5	930
Okanagan (oka)	1,798	39	87	0.3	770
Tlingit (tli)	1,521	32	107	12	630
Plains Cree (crk)	686	10	0	0	260
Ktunaxa (kut)	467	33	112	0.2	106
Coeur d'Alene (crd)	377	0	199	0.0	2
Kwak'wala (kwk)	98	1	1	0.0	585
TOTAL	19,187	324	2,599	19.8	

Table: Data in OLD applications (Feb 14, 2014)

language	forms	texts	audio	GB	speakers
Blackfoot (bla)	8,847	171	2,057	3.8	3,350
Nata (ntk)	3,219	32	0	0	36,000
Gitksan (git)	2,174	6	36	3.5	930
Okanagan (oka)	1,798	39	87	0.3	770
Tlingit (tli)	1,521	32	107	12	630
Plains Cree (crk)	686	10	0	0	260
Ktunaxa (kut)	467	33	112	0.2	106
Coeur d'Alene (crd)	377	0	199	0.0	2
Kwak'wala (kwk)	98	1	1	0.0	585
TOTAL	19,187	324	2,599	19.8	

Table: Data in OLD applications (Feb 14, 2014)

1. Here we can see that fieldwork is being performed on nine endangered and/or under-documented languages using the OLD software.
2. The languages that have seen the most use are Blackfoot, Nata, Gitksan, Okanagan, and Tlingit.
3. The speaker population figures in the rightmost column are from Ethnologue and are probably optimistic or out-of-date. In any cases, most of these languages are endangered to highly endangered.
4. We claim that these usage statistics indicate that field workers are actively seeking new tools like LingSync and the OLD.

Plugins

- Audio
- Lexicon

LingSync & OLD
Plugins & Reusing existing tools and libraries
Audio
Plugins

- Audio
- Lexicon

In this section I will show some screenshots of a couple of our more interesting plugins for

audio processing and

lexicon visualization.

Kartuli Speech Recognizer

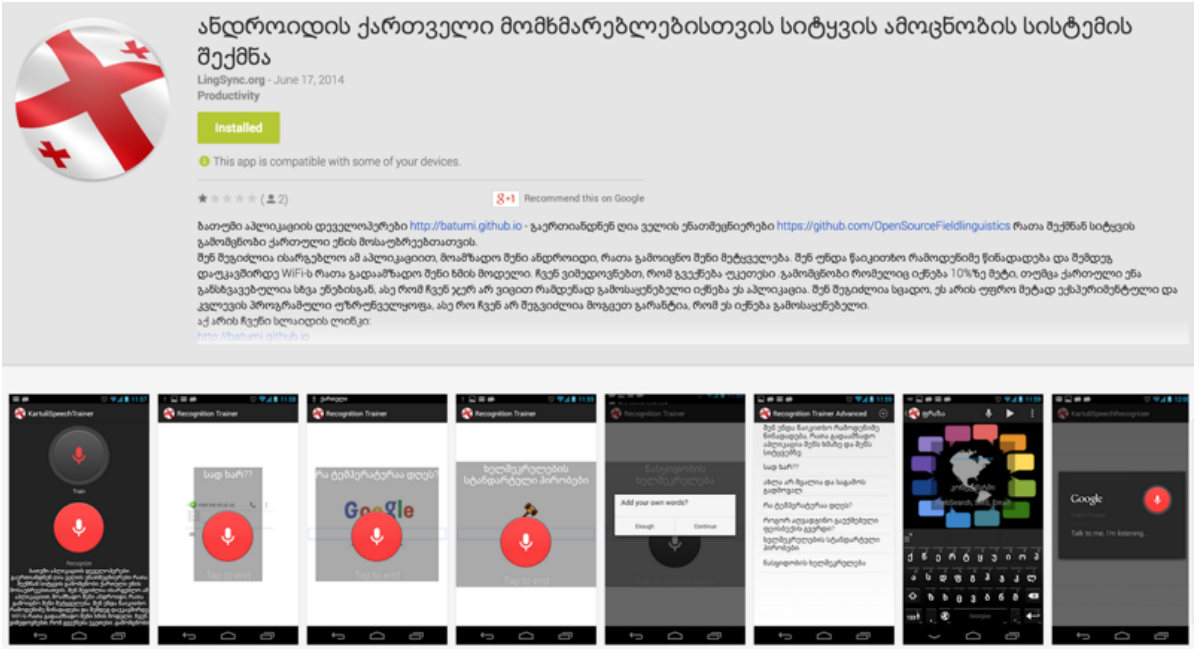


Figure: Screenshot of the Speech Recognition trainer

LingSync & OLD
Plugins & Reusing existing tools and libraries
Audio
Kartuli Speech Recognizer

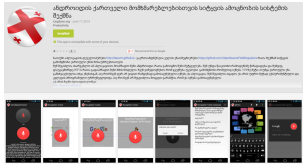
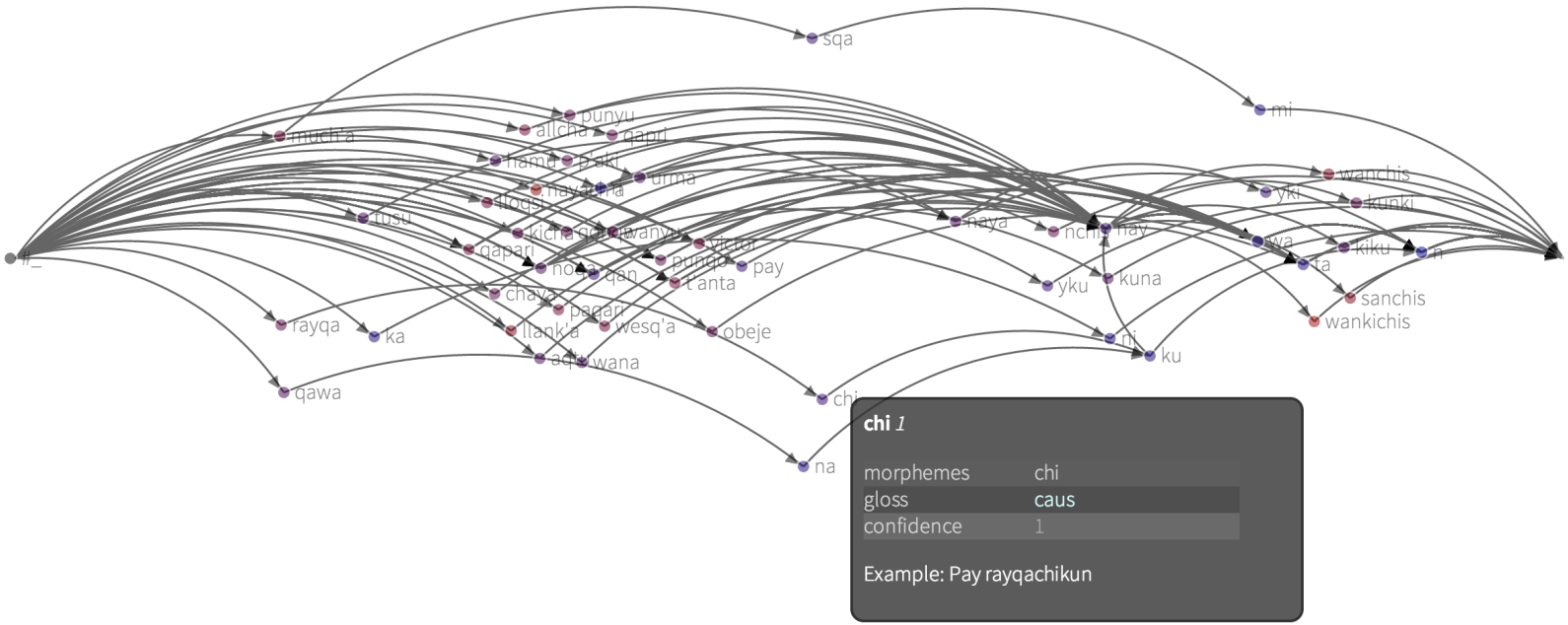


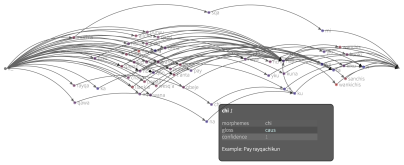
Figure: Screenshot of the Speech Recognition trainer

1. The Android speech recognition app is an app which is available on Google play for Kartuli speakers.
2. It was built last semester while I was in the field in Batumi.
3. The app uses the Learn X interface to permit users to train it to to their voice and vocabulary.
4. The sentences the users say become datum in their private corpus which is in turn used to re-train their own personal language model.
5. If you would like to see it in action you can download it from the Play store, (search for Kartuli speech recognizer) or come see us at the demos later.
6. We dont expect recognition rates better than 10% but we are hoping by letting users import their SMS or other text on their Android it will be come personalized enough to recognize their own speech in limited contexts such as SMS messages.

Force directed graph of morphemes in context

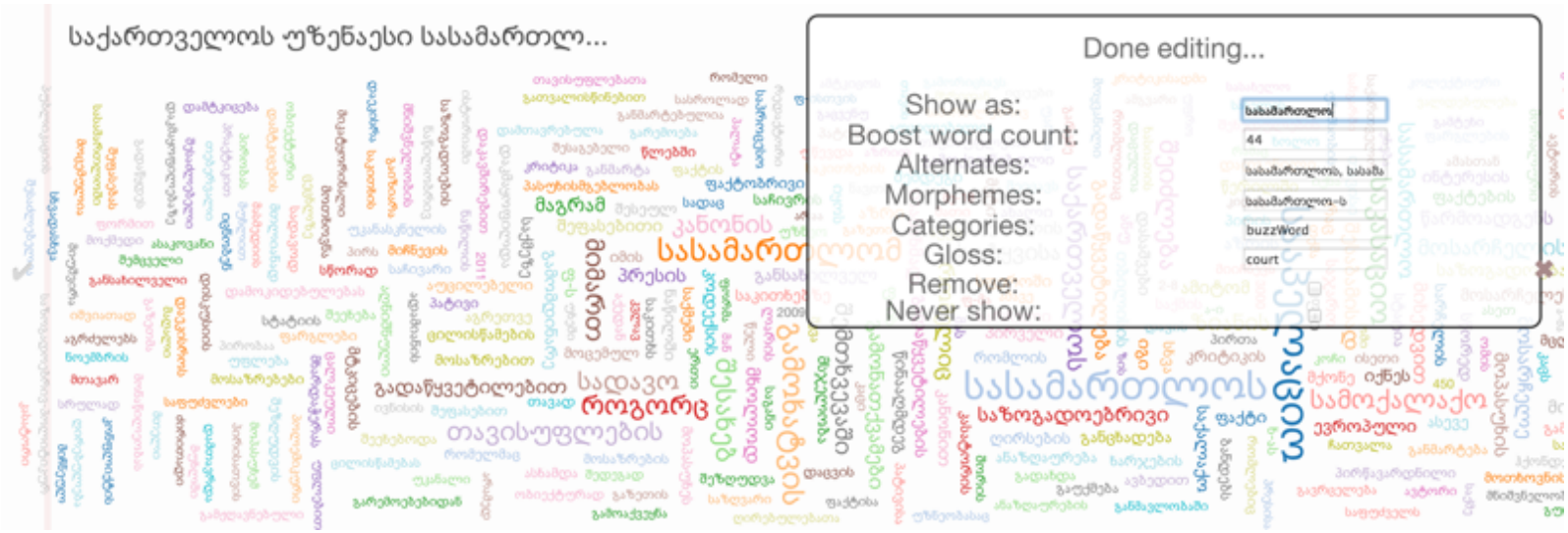


Force directed graph of morphemes in context

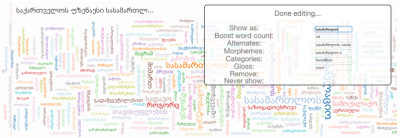


1. This visualization lets you see precendence order of morphemes in your corpus in a connected graph
2. The node on the left is the beginning of the word and the node on the right is the end of the word.
3. You can also choose not to plot the end nodes, and then you can see if the corpus has a focus on one morpheme. This data here is from M.E. dissertation about the interactions of the morpheme's -naya and -ta in Cusco Quechua, and this can be seen as they are the focal points of the graph.

WordCloud showing words by frequency



WordCloud showing words by frequency



1. The second visualization is a word cloud visualization which Josh built using Jason Davies' D3 word cloud layout engine.
2. Unlike Wordle, it runs in Javascript (not a Java applet) and so it works on iPads, Androids and all browsers.
3. While it's not as beautiful as Wordle, it supports the full unicode character set.
4. We added some logic for language independent automatic detection of function words (stop words) and tokenization.
5. My language consultants use this to clean head words, add segmentation, gloss or other lexical information. They get visual feedback in that the content-ful words begin to pop out as they clean, and the cloud becomes more representative of the meaning of the document.
6. The app is on Google Play and on the Chrome Store, search for iLanguage Cloud or come see us at the demos.

OLD morphological parsers (Blackfoot)

Goals

- Present and motivate the OLD's approach to creating morphological parsers
- Demonstrate the use of this functionality via two Blackfoot parsers

Goals

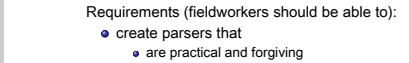
- Present and motivate the OLD's approach to creating morphological parsers
- Demonstrate the use of this functionality via two Blackfoot parsers

1. Now I'd like to discuss the OLD's morphological parser functionality. I will present and motivate the OLD's approach to creating morphological parsers with reference to two parsers created for Blackfoot, an endangered language which is from the Algonquian language family and which is spoken in Alberta, Canada and Montana, USA.
2. This portion of our presentation should segue nicely into the next talk which discusses a similar but interestingly different approach to modelling the morphology and phonology of another endangered Algonquian language: Plains Cree.

OLD Morphological Parsers

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving



1. Here are the requirements that guided the design of the OLD's functionality for creating morphological parsers
2. Fieldworkers should be able to create morphological parsers that are practical and forgiving. By practical I mean that they should be able to suggest correct, or largely correct, word analyses during data entry. By forgiving I mean that I want fieldworkers to be able to create parsers without first having a full and perfect analysis of the morphophonology of their language of study.

OLD Morphological Parsers

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving
  - make use of existing fieldworker skills

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving
  - make use of existing fieldworker skills

1. Since fieldworkers are being encouraged to create their own parsers, they should be able to do so by making use of the skills that they already have. That is, they should not be required to first learn wholly unfamiliar formalisms.

OLD Morphological Parsers

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving
  - make use of existing fieldworker skills
  - make use of data in the system

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving
  - make use of existing fieldworker skills
  - make use of data in the system

1. Fieldworkers should be able to build parsers using the expertly analyzed data that already exist in their databases; for example, morphologically analyzed words in IGT format and categorized and phonemically transcribed lexical entries.

OLD Morphological Parsers

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving
  - make use of existing fieldworker skills
  - make use of data in the system
  - are tailored to a specific purpose

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving
  - make use of existing fieldworker skills
  - make use of data in the system
  - are tailored to a specific purpose

1. Fieldworkers should be able to create different parsers for different purposes. Examples include orthographic parsers which parse orthographic transcriptions and phonetic parsers which parse phonetic transcriptions. It should also be possible to tailor phonetic parsers to the grammars of individual speakers or dialects. Fieldworkers should also be able to create analysis-specific variants of all of these.

OLD Morphological Parsers

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving
  - make use of existing fieldworker skills
  - make use of data in the system
  - are tailored to a specific purpose
  - facilitate automated analysis testing

LingSync & OLD
Plugins & Reusing existing tools and libraries
Morphology
OLD Morphological Parsers

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving
  - make use of existing fieldworker skills
  - make use of data in the system
  - are tailored to a specific purpose
  - facilitate automated analysis testing

1. Since the parsers are comprised of computational implementations of analyses and models of the lexicon, morphology, and phonology, they should facilitate the automated testing of these analyses and models against specified data sets in the system.

OLD Morphological Parsers

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving
  - make use of existing fieldworker skills
  - make use of data in the system
  - are tailored to a specific purpose
  - facilitate automated analysis testing
  - are reusable: spell-checkers, pronunciation dictionaries, …

LingSync & OLD
Plugins & Reusing existing tools and libraries
Morphology
OLD Morphological Parsers

Requirements (fieldworkers should be able to):

- create parsers that
  - are practical and forgiving
  - make use of existing fieldworker skills
  - make use of data in the system
  - are tailored to a specific purpose
  - facilitate automated analysis testing
  - are reusable: spell-checkers, pronunciation dictionaries, …

1. The components of the parsers should be reusable for, say, the creation of spell-checkers or the generation of pronunciation dictionaries that can be used in audio-transcription aligners.

OLD Morphological Parsers

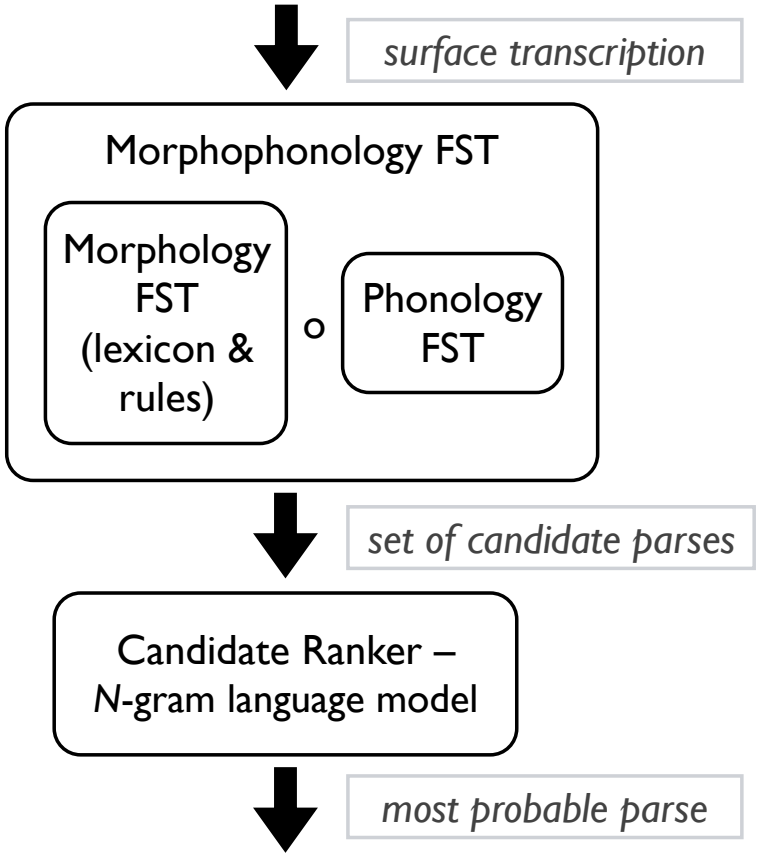


Figure: Architecture of an OLD morphological parser.

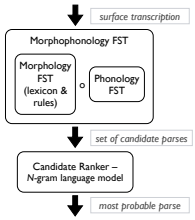


Figure: Architecture of an OLD morphological parser.

1. This diagram provides a high-level overview of the components of an OLD morphological parser.
2. The morphophonology is a finite-state transducer that takes a surface transcription as input and returns a set of candidate parses as output; it is the composition of a phonology transducer and a morphology transducer.
3. The candidate ranker assigns a probability to each candidate parse returned by the morphophonology; it is built upon an N-gram language model.

OLD Phonologies

- Phonology: a transducer wholly defined by the fieldworker as ordered context-sensitive rewrite rules

- Phonology: a transducer wholly defined by the fieldworker as ordered context-sensitive rewrite rules

1. An OLD phonology is a transducer that is defined explicitly and in its entirety by the fieldworker as an ordered list of context-sensitive rewrite rules.

OLD Phonologies

- Phonology: a transducer wholly defined by the fieldworker as ordered context-sensitive rewrite rules

(1) /nit-ihpiyi/ $\rightarrow$ <nitsspiyi>

- Phonology: a transducer wholly defined by the fieldworker as ordered context-sensitive rewrite rules

(1) /nit-ihpiyi/ → <nitsspiyi>

1. Example (1) shows a Blackfoot word segmented into phonemically transcribed morphemes and its phonetico-orthographic surface representation.

OLD Phonologies

- Phonology: a transducer wholly defined by the fieldworker as ordered context-sensitive rewrite rules

(1) /nit-ihpiyi/ $\rightarrow$ <nitsspiyi>

(2)
```
define phonology [
    [ "-" -> s || t _ i ] .o.
    [ i h -> s || s _ ] ] ;
#test nit-ihpiyi -> nitsspiyi
```

- Phonology: a transducer wholly defined by the fieldworker as ordered context-sensitive rewrite rules

(1) /nit-ihpiyi/ → <nitsspiyi>

(2)
```
define phonology [
    [ "-" -> s || t _ i ] .o.
    [ i h -> s || s _ ] ] ;
#test nit-ihpiyi -> nitsspiyi
```

1. (2) shows the definition of a simple phonology FST that implements the mapping in (1).
2. The phonology is written in the regular expression rewrite rule language that is accepted by FST toolkits like foma and Xerox's XFST.

OLD Phonologies

- Phonology: a transducer wholly defined by the fieldworker as ordered context-sensitive rewrite rules

(1) /nit-ihpiyi/ $\rightarrow$ <nitsspiyi>

(2)
```
define phonology [
    [ "-" -> s || t _ i ] .o.
    [ i h -> s || s _ ] ] ;
#test nit-ihpiyi -> nitsspiyi
```

- make use of existing fieldworker skills

- Phonology: a transducer wholly defined by the fieldworker as ordered context-sensitive rewrite rules

(1) /nit-ihpiyi/ → <nitsspiyi>

(2)
```
define phonology [
    [ "-" -> s || t _ i ] .o.
    [ i h -> s || s _ ] ] ;
#test nit-ihpiyi -> nitsspiyi
```

- make use of existing fieldworker skills

1. Since these rewrite rules are simply notational variants of the SPE-style rewrite rules that most linguistic fieldworkers are familiar with, the system makes use of existing fieldworker skills.
2. The second line in (2), for example, is a rule which transforms the hyphen morpheme delimiter to an "s" when it occurs after a "t" and before an "i". A fieldworker can immediately begin using this formalism after learning a few simple notational oddities, such as the use of double vertical lines in place of the forward slash that is customary in rule-based phonology texts.

OLD Phonologies

- Phonology: a transducer wholly defined by the fieldworker as ordered context-sensitive rewrite rules

(1) /nit-ihpiyi/ $\rightarrow$ <nitsspiyi>

(2)
```
define phonology [
    [ "-" -> s || t _ i ] .o.
    [ i h -> s || s _ ] ] ;
#test nit-ihpiyi -> nitsspiyi
```

- make use of existing fieldworker skills
- facilitate automated analysis testing

- Phonology: a transducer wholly defined by the fieldworker as ordered context-sensitive rewrite rules

(1) /nit-ihpiyi/ → <nitsspiyi>

(2)
```
define phonology [
      [ "-" -> s || t _ i ] .o.
      [ i h -> s || s _ ] ] ;
#test nit-ihpiyi -> nitsspiyi
```

- make use of existing fieldworker skills
- facilitate automated analysis testing

1. Also, since the fieldworker has full control over the definition of phonological transformations (or spelling rules), they can use the functionality to test their analyses of this component of the grammar. For example, the parser can be used to automate the discovery of words in the database that its morphophonology cannot analyze; This may lead to modifications of the underlying analyses and models.
2. One restricted but useful example of automated analysis testing is provided by the phonology test comment beneath the phonology definition. The left side of the arrow is the underlying representation and the right side is the surface representation. The OLD recognizes these comments in FST rewrite rule scripts and reports back on what percentage of the tests are passed by the phonology. This facilitates test-driven phonology development.

OLD Morphologies and Morpheme LMs

(3) Nitsspiyi
nit-ihpiyi
1-dance
agra-vai

'I danced.'

(3) Nitsspiyi
nit-ihpiyi
1-dance
agra-vai
'I danced.'

1. OLD applications are full of morphologically analyzed words like the Blackfoot one in (3).

OLD Morphologies and Morpheme LMs

(3) Nitsspiyi
nit-ihpiyi
1-dance
agra-vai

'I danced.'

(4)

w	$\rightarrow$	agr - vai
agr	$\rightarrow$	nit
vai	$\rightarrow$	ihpiyi

(3) Nitsspiyi
nit-ihpiyi
1-dance
agra-vai

'I danced.'

(4)
w	→	agr - vai
agr	→	nit
vai	→	ihpiyi

1. (4) is a rewrite rule representation of the morphological generalizations that can be extracted from a corpus of forms like (3) and implemented as an FST, in either the Lexicon Compiler (lexc) or Regular Expression Rewrite Rule (regex) language.
2. Note that the OLD does not currently allow users fine-grained control over morphology specification. That is, it is not currently possible to use flag diacritics to implement long-distance dependencies as illustrated by Snoek et al. in the precedings of this workshop. Of course, the OLD interface could easily be modified to allow users complete control over morphologies, perhaps with the option of using the system-generated lexical and morphotactic generalizations as a starting point.

OLD Morphologies and Morpheme LMs

(3) Nitsspiyi
nit-ihpiyi
1-dance
agra-vai

'I danced.'

(4)

w	→	agr - vai
agr	→	nit
vai	→	ihpiyi

(5)

<s>	nit	ihpiyi
nit	ihpiyi	</s>

(3) Nitsspiyi
nit-ihpiyi
1-dance
agra-vai

'I danced.'

(4)
w	→	agr - vai
agr	→	nit
vai	→	ihpiyi

(5)
<s>	nit	ihpiyi	
nit	ihpiyi	</s>	

1. (5) shows the morpheme trigrams that can be extracted from a corpus of IGT forms like (3).

OLD Morphologies and Morpheme LMs

(3) Nitsspiyi
nit-ihpiyi
1-dance
agra-vai

'I danced.'

(4)

w	$\rightarrow$	agr - vai
agr	$\rightarrow$	nit
vai	$\rightarrow$	ihpiyi

(5)

<s>	nit	ihpiyi
nit	ihpiyi	</s>

- make use of data in the system

(3) Nitsspiyi
nit-ihpiyi
1-dance
agra-vai

'I danced.'

(4)
w	→	agr - vai
agr	→	nit
vai	→	ihpiyi

(5)
<s>	nit	ihpiyi
nit	ihpiyi	</s>

- make use of data in the system

1. Since the morphology FST and the morpheme LM are generated by the system based on data in the database, a significant portion of the parser creation work is automated using the endangered language data present in the OLD application.

Blackfoot Parsers

- 2 parsers, both with identical morphologies and LMs:

LingSync & OLD
Plugins & Reusing existing tools and libraries
Morphology
Blackfoot Parsers

- 2 parsers, both with identical morphologies and LMs:

1. Now I'd like to quickly discuss two Blackfoot parsers created using the OLD morphological parser functionality.
2. They have identical morphologies and morpheme language models.

Blackfoot Parsers

- 2 parsers, both with identical morphologies and LMs:
  - lexical items from standard dictionary (Frantz 1995)

LingSync & OLD

Plugins & Reusing existing tools and libraries

Morphology

Blackfoot Parsers

- 2 parsers, both with identical morphologies and LMs:
  - lexical items from standard dictionary (Frantz 1995)

1. The lexica of these parsers are the morphemes present in the standard dictionary of the language, Frantz and Russell's Blackfoot Dictionary of Stems, Roots, and Affixes.

Blackfoot Parsers

- 2 parsers, both with identical morphologies and LMs:
  - lexical items from standard dictionary (Frantz 1995)
  - morphotactic rules extracted from 3,414 well analyzed words types (BLA OLD)

- 2 parsers, both with identical morphologies and LMs:
  - lexical items from standard dictionary (Frantz 1995)
  - morphotactic rules extracted from 3,414 well analyzed words types (BLA OLD)

1. The morphotactic rules, that is, the set of category and delimiter sequences that correspond to valid words of the language, were extracted from the 3,414 well analyzed word types in the Blackfoot OLD. Well analyzed words are those with fieldworker-created morphological analyses such that all of the morphemes used have lexical entries in the database.

Blackfoot Parsers

- 2 parsers, both with identical morphologies and LMs:
  - lexical items from standard dictionary (Frantz 1995)
  - morphotactic rules extracted from 3,414 well analyzed words types (BLA OLD)
  - morpheme LM estimated from 3,425 gold standard well analyzed words (BLA OLD)

- 2 parsers, both with identical morphologies and LMs:
  - lexical items from standard dictionary (Frantz 1995)
  - morphotactic rules extracted from 3,414 well analyzed words types (BLA OLD)
  - morpheme LM estimated from 3,425 gold standard well analyzed words (BLA OLD)

1. The morpheme language models were estimated based on counts from the set of well analyzed words in the database such that a single best analysis can be identified. This is the gold standard used for both LM creation and evaluation. Each parser was created five times with a different randomly sampled 90% training set and overall parser performance was tested against the corresponding 10% test set.

Blackfoot Parsers

- 2 parsers, both with identical morphologies and LMs:
  - lexical items from standard dictionary (Frantz 1995)
  - morphotactic rules extracted from 3,414 well analyzed words types (BLA OLD)
  - morpheme LM estimated from 3,425 gold standard well analyzed words (BLA OLD)
- Parser 1: phonological rules and allomorphic alternations from grammar (Frantz 1997)

- 2 parsers, both with identical morphologies and LMs:
  - lexical items from standard dictionary (Frantz 1995)
  - morphotactic rules extracted from 3,414 well analyzed words types (BLA OLD)
  - morpheme LM estimated from 3,425 gold standard well analyzed words (BLA OLD)
- Parser 1: phonological rules and allomorphic alternations from grammar (Frantz 1997)

1. The two parsers differ in their phonologies. Parser 1 has a phonology which implements, as faithfully as I could manage, the general phonological rules of the grammar and the lexically conditioned allomorphic alternations described in that same text.

Blackfoot Parsers

- 2 parsers, both with identical morphologies and LMs:
  - lexical items from standard dictionary (Frantz 1995)
  - morphotactic rules extracted from 3,414 well analyzed words types (BLA OLD)
  - morpheme LM estimated from 3,425 gold standard well analyzed words (BLA OLD)
- Parser 1: phonological rules and allomorphic alternations from grammar (Frantz 1997)
- Parser 2: Parser 1's phonology with length and prominence contrasts removed during generation; massively over-analyzes

- 2 parsers, both with identical morphologies and LMs:
  - lexical items from standard dictionary (Frantz 1995)
  - morphotactic rules extracted from 3,414 well analyzed words types (BLA OLD)
  - morpheme LM estimated from 3,425 gold standard well analyzed words (BLA OLD)
- Parser 1: phonological rules and allomorphic alternations from grammar (Frantz 1997)
- Parser 2: Parser 1's phonology with length and prominence contrasts removed during generation; massively over-analyzes

1. Parser 2 is an attempt to improve on Parser 1 by modifying the phonology so that it recognizes many more surface forms. During generation, this phonology maps long segments to their short counterparts and de-accents accented characters. Orthographically transcribed words must be de-accented and de-lengthened prior to parsing attempts.

Blackfoot Parser – Results

parser	succ.	F-score	prec.	rec.	phon.	morphon.	LM
1	0.14	0.32	0.53	0.23	0.21	0.20	0.72
2	0.17	0.40	0.40	0.39	0.60	0.60	0.28

Table: Blackfoot OLD morphological parser results.

LingSync & OLD
Plugins & Reusing existing tools and libraries
Morphology
Blackfoot Parser – Results

parser	succ.	F-score	prec.	rec.	phon.	morphon.	LM
1	0.14	0.32	0.53	0.23	0.21	0.20	0.72
2	0.17	0.40	0.40	0.39	0.60	0.60	0.28

Table: Blackfoot OLD morphological parser results.

1. This table summarizes the results of evaluating these parsers.
2. While neither has a very impressive overall success rate---14% and 17%, respectively---their F-scores show that they will be practical in expediting the creation of IGT representations. With a GUI that presents a ranked list of parses during data entry, fieldworker-contributors will be able to edit auto-generated parses that contain, roughly and on average, half of the correct morphemes.
3. Also, the morphophonology transducer of Parser 2 is 60% accurate, meaning that 60% of the time the correct analysis is in the candidate set that it returns. Therefore users could scroll through the ranked list of candidates and choose the correct one; which would be faster than typing it out manually.

Blackfoot Parser – Results

parser	succ.	F-score	prec.	rec.	phon.	morphon.	LM
1	0.14	0.32	0.53	0.23	0.21	0.20	0.72
2	0.17	0.40	0.40	0.39	0.60	0.60	0.28

Table: Blackfoot OLD morphological parser results.

parser	succ.	F-score	prec.	rec.	phon.	morphon.	LM
1	0.14	0.32	0.53	0.23	0.21	0.20	0.72
2	0.17	0.40	0.40	0.39	0.60	0.60	0.28

Table: Blackfoot OLD morphological parser results.

1. The phonology of Parser 1 has a 21% success rate while that of Parser 2 has a 60% success rate. There are two primary explanations for the poor performance of Parser 1's phonology, i.e., that which is faithful to the grammar. First, the grammar's phonology does not account for non-lexical prominence, i.e., accenting. Second, there is a lot of inconsistency in terms of orthographic transcription in this multi-contributor data set.

Blackfoot Parser – Results

parser	succ.	F-score	prec.	rec.	phon.	morphon.	LM
1	0.14	0.32	0.53	0.23	0.21	0.20	0.72
2	0.17	0.40	0.40	0.39	0.60	0.60	0.28

Table: Blackfoot OLD morphological parser results.

LingSync & OLD
Plugins & Reusing existing tools and libraries
Morphology
Blackfoot Parser – Results

parser	succ.	F-score	prec.	rec.	phon.	morphon.	LM
1	0.14	0.32	0.53	0.23	0.21	0.20	0.72
2	0.17	0.40	0.40	0.39	0.60	0.60	0.28

Table: Blackfoot OLD morphological parser results.

1. It is interesting that the phonology of Parser 2 performs so much better yet the gains in overall parse rate and F-score are not as significant. Parser 2 essentially puts a far greater burden on the language model, a burden which is too much for an LM based on counts from an extremely small corpus of 3,400 words. Parser 2's morphophonology transducer produces approximately 1,800 analysis candidates per input word. As more well analyzed words are added to the database with the aid of the parsers, the LM will improve and, as a result, so too will the parser as a whole.

Blackfoot Parser – Results

parser	succ.	F-score	prec.	rec.	phon.	morphon.	LM
1	0.14	0.32	0.53	0.23	0.21	0.20	0.72
2	0.17	0.40	0.40	0.39	0.60	0.60	0.28

Table: Blackfoot OLD morphological parser results.

LingSync & OLD
Plugins & Reusing existing tools and libraries
Morphology
Blackfoot Parser – Results

parser	succ.	F-score	prec.	rec.	phon.	morphon.	LM
1	0.14	0.32	0.53	0.23	0.21	0.20	0.72
2	0.17	0.40	0.40	0.39	0.60	0.60	0.28

Table: Blackfoot OLD morphological parser results.

1. Of course, another means of improving overall performance would be to improve the performance of the phonology without causing it to massively overgenerate during analysis and put such a burden on the LM, that is, by using fieldworker knowledge to write a phonology that can predict the location of accented vowels. This is in the works.

OLD Morphological Parsers – Discussion

OLD parsers exemplify a particular set of principles for guiding computationally-assisted endangered languages fieldwork:

OLD parsers exemplify a particular set of principles for guiding computationally-assisted endangered languages fieldwork:

1. The OLD approach to creating morphological parsers illustrates some general strategies that I would like to advocate in terms of computationally-assisted fieldwork on endangered languages.

OLD Morphological Parsers – Discussion

OLD parsers exemplify a particular set of principles for guiding computationally-assisted endangered languages fieldwork:

- exploit both fieldworker expertise and automated methods of grammar induction

OLD parsers exemplify a particular set of principles for guiding computationally-assisted endangered languages fieldwork:

- exploit both fieldworker expertise and automated methods of grammar induction

1. We need systems that can both make use of the expertise of the fieldworkers who are creating these data sets, while also helping them to bootstrap the process with statistical and machine learning techniques. This is interestingly different from NLP where the overarching requirement would seem to be the production of results with as little expert guidance as one can get away with.
2. By allowing fieldworkers real control over the specification of models, we not only make use of that skill set, we also make it easier for them to automate the testing of their analyses, and thereby to improve those analyses.
3. At the same time, we should be forgiving, i.e., we should not REQUIRE a complete mastery of components of the grammar prior to the creation of useful features like automatic morphological analyzers.

OLD Morphological Parsers – Discussion

OLD parsers exemplify a particular set of principles for guiding computationally-assisted endangered languages fieldwork:

- exploit both fieldworker expertise and automated methods of grammar induction
- use familiar representations:
  - yes: (nit-ihpiyi, 1-dance, agra-vai)
  - no: ihpiyi+1+SG+PST

OLD parsers exemplify a particular set of principles for guiding computationally-assisted endangered languages fieldwork:

- exploit both fieldworker expertise and automated methods of grammar induction
- use familiar representations:
  - yes: (nit-ihpiyi, 1-dance, agra-vai)
  - no: ihpiyi+1+SG+PST

1. We should also create systems that make use of representations that are familiar to fieldworkers. This has already been mentioned with respect to SPE-style context-sensitive rewrite rules in the specification of phonologies. Another example of this, however, can be seen in the fact that OLD parsers return morphological analyses in IGT format, i.e., a morpheme segmentation line, a gloss line, and a category line.

OLD Morphological Parsers – Discussion

OLD parsers exemplify a particular set of principles for guiding computationally-assisted endangered languages fieldwork:

- exploit both fieldworker expertise and automated methods of grammar induction
- use familiar representations:
  - yes: (nit-ihpiyi, 1-dance, agra-vai)
  - no: ihpiyi+1+SG+PST
- allow for multiple models

OLD parsers exemplify a particular set of principles for guiding computationally-assisted endangered languages fieldwork:

- exploit both fieldworker expertise and automated methods of grammar induction
- use familiar representations:
  - yes: (nit-ihpiyi, 1-dance, agra-vai)
  - no: ihpiyi+1+SG+PST
- allow for multiple models

1. We need to also enable fieldworkers to implement different models for different purposes. In some situations, prescriptive models are needed such as orthographic parsers that can parse orthographic transcriptions and which can be used, for example, in the creation of spell-checkers. In other situations, descriptive models of various types are needed. For example, speaker-, dialect-, and/or analysis-specific models that can analyze and generate phonetic representations and which can be used to test analyses and, for example, to create pronunciation dictionaries for audio-to-text alignment tools.

Take Homes

- Finding effective tools for fieldwork on endangered languages is non-trivial
- All stakeholders, All devices
- Modular Web Services and Plugin Architecture
- Open Source, Open Development, Open Data
- Providing glue for fieldworkers and computational linguists to collaborate

- Finding effective tools for fieldwork on endangered languages is non-trivial
- All stakeholders, All devices
- Modular Web Services and Plugin Architecture
- Open Source, Open Development, Open Data
- Providing glue for fieldworkers and computational linguists to collaborate

1. We hoped to demonstrate that finding tools for fieldwork for endangered languages is not trivial task.
2. We wanted to include all stakeholders and all devices
3. The way we achieved this was by using modular web services and a plugin architecture
4. Everything is open source, our development is open which means you can see all the features and requests as well as completed milestones and future milestones. we enable open data but do not enforce it
5. We provide the glue for field workers and computational linguists to collaborate without becoming frustrated or feeling like they are loosing precious time which could be used to further their own research programs.

Acknowledgements

Tobin Skinner, Elise McClay, Louisa Bielig, MaryEllen Cathcart, Theresa Deering, Yuliya Manyakina, Gretchen McCulloch, Hisako Noguchi, Brian Doherty, Gay Hazan, Oriana Kilbourn, Kim Dan Nguyen, Rakshit Majithiya, Mietta Lennes, Nivja de Jong, Ton Wempe, Kyle Gorman, Curtis Mesher, Beso Beridze, Tornike Lasuridze, Zviadi Beradze, Rezo Turmanidze, Jason Smith, Martin Gausby, Pablo Duboue, Xianli Sun, James Crippen, Michael McAuliffe, Patrick Littell, Faryal Abbasi, Farah Abbasi, Tamila Paghava, Esma Chkhikvadze, Nina Gatenadze, and Mari Mgeladze, Jessica Coon, Alan Bale, Michael Wagner, Henry Davis, Lisa Matthewson, Alexandre Bouchard-Côté

SSHRC Connection Grant (#611-2012-0001)
SSHRC Standard Research Grant (#410-2011-2401)
SSHRC Image, Text, Sound & Technology (#849-2009-0056)

LingSync & OLD
(Our Team)
Acknowledgements

Tobin Skinner, Elise McClay, Louisa Bielig, MaryEllen Cathcart, Theresa Deering, Yuliya Manyakina, Gretchen McCulloch, Hisako Noguchi, Brian Doherty, Gay Hazan, Oriana Kilbourn, Kim Dan Nguyen, Rakshit Majithiya, Mietta Lennes, Nivja de Jong, Ton Wempe, Kyle Gorman, Curtis Mesher, Beso Beridze, Tornike Lasuridze, Zviadi Beradze, Rezo Turmanidze, Jason Smith, Martin Gausby, Pablo Duboue, Xianli Sun, James Crippen, Michael McAuliffe, Patrick Littell, Faryal Abbasi, Farah Abbasi, Tamila Paghava, Esma Chkhikvadze, Nina Gatenadze, and Mari Mgeladze, Jessica Coon, Alan Bale, Michael Wagner, Henry Davis, Lisa Matthewson, Alexandre Bouchard-Côté

SSHRC Connection Grant (#611-2012-0001)
SSHRC Standard Research Grant (#410-2011-2401)
SSHRC Image, Text, Sound & Technology (#849-2009-0056)

We would like to thank our linguistics student interns, computer science student interns and countless other open source software developers who directly or indirectly helped build LingSync/OLD to what it is today and will be in the future.

We would like to thank the ComputEL workshop reviewers, our users and would be users for providing feedback, suggestions, asking tough questions, and sending bug reports.

We would like to thank our language consultants for their friendship, patience and for sharing their language with us.

We would also like to thank the SSHRC council for explicitly supporting "open source approaches to knowledge mobilzation."

References - ბიბლიოგრაფია

- MaryEllen Cathcart, Gina Cook, Theresa Deering, Yuliya Manyakina, Gretchen McCulloch, and Hisako Noguchi. 2012. LingSync: A free tool for creating and maintaining a shared database for communities, linguists and language learners. In Robert Henderson and Pablo Pablo, editors, Proceedings of FAMLi II: workshop on Corpus Approaches to Mayan Linguistics 2012, pages 247-250.
- Jonathon E. Cihlar. 2008. Database development for language documentation: A case study in the Washo language. Master's thesis, University of Chicago.
- N.H. De Jong and T Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. Behavior research methods, 41(2):385-390.
- Joel Dunham. 2014. The Online Linguistic Database: Software for linguistic fieldwork. PhD dissertation, UBC. (To appear.)
- Benoit Farley. 2012. The Uqailaut project. http://www.inuktitutcomputing.ca, January.
- Jeff Good. 2012b. Valuing technology: Finding the linguists place in a new technological universe. In Louanna Furbee and Lenore Grenoble, editors, Language documentation: Practice and values, pages 111131. Benjamins, Amsterdam.
- D. Hallett, M. J. Chandler, and C. E. Lalonde. 2007. Aboriginal language knowledge and youth suicide. Cognitive Development, 22(3):392–399.
- Stuart Robinson, Greg Aumann, and Steven Bird. 2007. Managing fieldwork data with ToolBox and the Natural Language Toolkit. Language Documentation & Conservation, 1(1):4457.
- R. Schroeter and N. Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In Sebastian Nordoff, editor, Sustainable Data from Digital Fieldwork. University of Sydney, Sydney.
- Nick Thieberger. 2012. Using language documentation data in a broader context. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek, editors, Potentials of Language Documentation: Methods, Analyses, and Utilization. University of Hawaii Press, Honolulu.
- Doug Troy and Andrew J. Strack. 2014. Metimankwiki kimehsoominaanaki we follow our ancestors trail: Sharing historical Myaamia language documents across myaamionki. In Proceedings of the 2014 Myaamiaki Conference.

LingSync & OLD
(Our Team)
References - ბიბლიოგრაფია

- MaryEllen Cathcart, Gina Cook, Theresa Deering, Yuliya Manyakina, Gretchen McCulloch, and Hisako Noguchi. 2012. LingSync: A free tool for creating and maintaining a shared database for communities, linguists and language learners. In Robert Henderson and Pablo Pablo, editors, Proceedings of FAMLi II: workshop on Corpus Approaches to Mayan Linguistics 2012, pages 247-250.
- Jonathon E. Cihlar. 2008. Database development for language documentation: A case study in the Washo language. Master's thesis, University of Chicago.
- N.H. De Jong and T Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. Behavior research methods, 41(2):385-390.
- Joel Dunham. 2014. The Online Linguistic Database: Software for linguistic fieldwork. PhD dissertation, UBC. (To appear.)
- Benoit Farley. 2012. The Uqailaut project. http://www.inuktitutcomputing.ca, January.
- Jeff Good. 2012b. Valuing technology: Finding the linguists place in a new technological universe. In Louanna Furbee and Lenore Grenoble, editors, Language documentation: Practice and values, pages 111131. Benjamins, Amsterdam.
- D. Hallett, M. J. Chandler, and C. E. Lalonde. 2007. Aboriginal language knowledge and youth suicide. Cognitive Development, 22(3):392–399.
- Stuart Robinson, Greg Aumann, and Steven Bird. 2007. Managing fieldwork data with ToolBox and the Natural Language Toolkit. Language Documentation & Conservation, 1(1):4457.
- R. Schroeter and N. Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In Sebastian Nordoff, editor, Sustainable Data from Digital Fieldwork. University of Sydney, Sydney.
- Nick Thieberger. 2012. Using language documentation data in a broader context. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek, editors, Potentials of Language Documentation: Methods, Analyses, and Utilization. University of Hawai Press, Honolulu.
- Doug Troy and Andrew J. Strack. 2014. Metimankwiki kimehsoominaanaki we follow our ancestors trail: Sharing historical Myaamia language documents across myaamionki. In Proceedings of the 2014 Myaamiaki Conference.

These are just a few of our references please refer to the paper in the preceedings for more.