iCAMPO:
UN APLICACIÓN DE COLECCIÓN DE DATOS LINGÜÍSTICOS*

MARYELLEN CATHCART[a,b]
GINA COOK[b]
THERESA DEERING[b,c]
YULIYA MANYAKINA[d,e]
GRETCHEN MCCULLOCH[d]
HISAKO NOGUCHI[e]
[a]*University of Delaware* [b]*iLanguage Lab* [c]*Visit Scotland* [d]*McGill University* [e]*Concordia University*

Spanish Abstract - ME

English Abstract

LingSync is an OpenSource database system that allows language researchers to securely enter, store, organize, annotate, and share linguistic data. The application is accessible on any device; as it runs in a HTML5 browser, it runs on laptops (Mac 10.5 and above, Linux, Windows, ChromeBooks) as well as on mobile devices (Android and iPhone/iPad). It is suitable for both online and offline use. Furthermore, the application is created with collaborative goals in mind; data is syncable and sharable with other researchers. Researchers can form teams that contribute to a single corpus, where team members use the application to modify and discuss the data. The system also has a simple and friendly user interface, allowing users to drag and drop data (audio, video, text), or record audio/video directly into the database. In addition, the application has

*English title: LingSync: A Free Tool for Creating and Maintaining a Shared Database For Communities, Linguists and Language Learners

We would like to thank the CAML participants for their comments and questions, as well as members of Android Montréal and Javascript Montréal for their comments and feedback on aspects of LingSync presented at meet-ups, as well as Montreal's vibrant start-up community who helped us understand the industry best practices which underpin LingSync's success as a user driven app. We would also like to express our deep thanks to Tobin Skinner, Josh Horner, Elise McClay, Louisa Bielig, Joel Dunham, Gay Hazan, Oriana Kilbourn, Rakshit Majithiya, Kim Dan Nguyen, Pablo Duboue and as well as countless other open source software developers whose code directly or indirectly helped build LingSync to what it is today. We would like to thank Mathieu Legault for his substantial financial contributions which made LingSync advance far quicker than expected, as well as SSHRC Grant ***? Lastly we would like to thank our significant others for their support and patience: without them our passion for better data experiences would have remained just that, a passion and not a reality.

import and export capabilities for multiple file types. LingSync is designed from the ground up to conform to E-MELD and DataOne data management best practices, an important requirement for any database which will house data funded by granting agencies. Most importantly, the application is designed intuitively and theory free, so it is not necessary to be a field linguist or programmer to figure out how it works. LingSync is hosted on cloud servers so that users can use it without knowing how to set up its servers, but also has an installation guide for linguistics department server administrators so that they can set up unlimited data usage on their own department servers.

1 Introduction

1.1 Why LingSync was created

1.1.1 The need - GRETCHEN

LingSync was conceived out of the needs of language researchers doing fieldwork or other large scale data collection. Linguistic fieldwork often requires researchers to travel to places where a stable connection to the internet is not guaranteed. Also, it often involves a group of researchers contributing to building a single database. An ideal linguistic database should therefore work both online and offline as well as making it easy to share and integrate data.

1.1.2 Other programs: pros and cons - GRETCHEN

There are several existing programs used for linguistic fieldwork; however, none of them fully satisfies the needs of field linguists for robust, collaborative, multi-platform data annotation and organization, both online and offline. For example, there are web-based databases which allow collaboration, such as the Online Linguistic Database (OLD), Karuk Dictionary and Texts, and The Washo Project but they only work online, making them unusable for researchers looking to enter new data or search the database while in the field with limited or no internet access. There are also non-web-based software programs such as Toolbox and FLEx/FieldWorks, which are excellent for annotating data and organizing data into various formats (corpus, grammar or lexicon).

However, with these offline tools, researchers each enter data on a single computer, making them more vulnerable to technical difficulties, and meaning that they must use a single device for all work on the language and cannot easily combine their data with others who work on related projects. Moreover, these tools run only on a single platform (either PC or Linux, but not both and not Mac or mobile devices). Another offline tool is general purpose database software such as FileMaker Pro, which can be customized for the purpose of language research. However, this incurs the same problems as other offline tools, while additionally often requiring that a programmer be hired to customize the software for the purpose of linguistic research.

All of the linguistic database programs surveyed did not provide a good user experience. The number of clicks required and the delay between actions did not meet current software engineering best practices. In addition to core functionalities, a good user experience is necessary to ensure quality data management. LingSync grows out of discussion with a number of fieldworkers dissatisfied with currently available options.

1.1.3 Technological background: why we can make this now - GINA

1.2 Design principles - HISAKO

The principal goal of LingSync is to help language researchers collect and organize linguistic data and to facilitate collaborative research work. The main objectives are to provide:

The usefulness and effectiveness of LingSync will be evaluated in two respects: Data Documentation and Data Management. Data documentation will be evaluated following E-MELD Best Practices in Digital Language Documentation,[1] and data management following DataONE Primer on Data Management.[2]

- A self-explanatory, easy-to-use user interface so that researchers can understand and start using the application within seconds of the time the installation is completed.
- Both online and offline functionality so that the fieldwork is not constrained by the internet accessibility.
- Customizable data entry fields to accommodate particular requirements of a research.
- Data sharing, protection and integration functions to facilitate collaboration among researchers and between researchers and language consultants.

Although it is designed primarily for linguists, the application will equally be useful for researchers documenting endangered languages and/or creating dictionaries/grammar books for minority languages, as well as language teachers creating educational materials.

1.2.1 Open (open source, open access)

OpenSource. Being OpenSource allows departments to install and customize the database application to tailor their specific needs without worry that the company behind the software will disappear or stop maintaining the software. In addition, OpenSourcing the software on GitHub will allow linguists with scripting or programming experience to contribute back to the software to make it more customized to their needs, language typologies, or linguistics research areas. It allows the software to continue to grow and improve without any company which seeks to profit from the software.

OpenAccess. Corpora often contain sensitive information, consultant stories and other information which must be kept confidential. Having confidential data in plain text in a corpus forces the entire corpus to be kept confidential. Instead, the system encrypts confidential data and stores the data in the corpus encrypted. To access the plain text the user has to log in and use a password to decrypt the data. This design has important ramifications for exporting data, and for editing the data outside the application. The system allows the user to export data encrypted or decrypted. If the corpus contains sensitive confidential information the system will warn the users if they choose to export the information in a decrypted fashion.

[1] http://emeld.org/school/what.html

[2] http://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

1.2.2 Standards-compliant (Unicode, EMELD, Leipzig, XML)

Unicode. Encoding problems and losing data should be behind us in the days of Unicode. However, many existing fieldlinguistics databases were built in programming languages that did not support Unicode, or where unicode support was added as an afterthought, so the Unicode support is dangerously fragile. Javascript and HTML5 (the technologies used in the system) are 100% Unicode.

EMELD. The application will allow researchers to share their data with anyone interested in their work. The application will be able to import data from ELAN XML, CSV and text file formats, and to export to XML, Latex and WIki formats. These import/export functions will make it easier to exchange data with those who are not using the same application. The system will have users and teams, and permissions for corpora. Permissions will ensure that data can be safely shared and edited by multiple users. Moreover, the corpus will be versioned so that users can track changes and revert mistakes.

Leipzig. The application will be designed for search as this is one of the most fundamental tasks a language researcher must be able to do. The search will go far beyond traditional string matches and database indexes. The application will guess what users do most often, and automate the process for them. Most importantly, the system will have semi-automated glossing based on morpheme segmentation.

XML/JSON. Data will be stored on a CouchDB server, and will be accessible and sharable by multiple users. The installable Chrome extension version of the application allows individual researchers to store a portion of a corpus, or an entire corpus, on their own devices, enabling offline work and quick search.

1. Content: Data is annotated and described using consistent terminology.
   The application allows users to design their annotation terminology per corpus and attaches these categories to each datum of the corpus, ensuring that all datum are annotated in a consistent fashion. The creation of datum also automatically inserts the gloss units into an ontology which is used for search indexes and can be human curated in Module **??**. The Annotations include three types, free text, enumerated grammatical tags, and enumerated datum status (checked with a consultant, to be checked, elicitation error, deleted etc). While all datum are annotated with a consistent terminology, the app is designed to adapt to a particular linguistic framework or theoretical construct, and allows researchers to choose categories for annotation.
2. Format: Data are intelligible regardless of the types of operating system
   The application is entirely in Unicode and exports data in a number of plain text human readable non-proprietary formats including .json, .tex, .txt, .csv, and .xml files. The application runs on PC, Linux and Mac OS, as well as on newer platforms such as Android and ChromeBook.
3. Discovery: Data are searchable and discoverable.
   Within the application, data are discoverable via keyword and string match search. The search module is capable of performing intersective or union search of the data using any of

the annotation fields which are used in the corpus. External to the application the corpora which are public will show up in the OLAC linguistic search engine. [3]
The system is designed not only for data entry, but also for data retrieval. Data stored in the application, if tagged as public by the researcher, is indexable by search engines and open to public view in principle. However authorized researchers (authors of data) have control over who can see, edit and/or export their data. Confidential data are stored encrypted in the database. The application allows two sorts of export. Export of encrypted output maintains encryption on confidential data, allowing researchers to make their public data public, without concern that confidential data or consultant information will become public. Export of decrypted output keeps the data non-proprietary and viewable outside of the app and must be protected as any other confidential data.

4. Access: Data are accessible.
   As the system is online and hosted by reliable servers, the data is considered accessible, unlike tape-data and noted cards.
5. Citation: Data provide citation information.
   Each corpus has a unique URL which can be used for citation. In addition, corpus administrators can add an archive URL and automate their corpus archiving by creating a "Bot" which will archive their data to one of the existing, reputable language archives of the user's choosing.[4] Data are further citable to their primary sources in that all data points are tied to a Session which contains details of data source such as consultant, publication, web page and the time the data are elicited, as well as other metadata controlled by the researcher to ensure that data quality can be traced to its source.
6. Preservation: Data are archived in a way that withstands long-term preservation. [5]
   Data are stored in a host server as JSON files. JSON files are human-readable text files, which have replaced XML data stores as a lighter weight, similarly featured markup language. Therefore the information content would not be lost even if the data format becomes obsolete. In addition to a host server, the data is stored in multiple locations. Researchers have a local storage on their own devices if the application is installed as Chrome extension, which could host a significant portion of (or an entire) corpus.
7. Rights: Rights of authors of data and of language consultants are respected.
   Researchers and language consultants (authors of data) have control over who can see, edit and/or export their data. Data which are confidential, as well as consultant information which is confidential are encrypted prior to storage in the database using the US Federal approved AES encryption standard. This ensures that confidential data cannot be leaked if a corpus is shared or leaked, without the consent of the corpus' author. The terms of use of data will be well documented, and are enforced using AES encryption. As endangered language data may often have a condition of being kept private until its parties have passed away, should the administration member of the corpus die, a policy of obeying his/her heir's decisions with regard to the data shall be put in place and discussed in the terms and conditions.

[3] http://linguistlist.org/olac/index.html
http://www.language-archives.org/documents/implement.html#conventional
[4] http://emeld.org/school/classroom/archives/index.html
[5] http://emeld.org/school/classroom/archives/finding-archives.html

1. Plan: Plan for data management prior to data collection and revise it as necessary during the project.
   What data will be generated: users can use LingSync to manage their data collection process. They can begin by adding a description to their corpus, as well as adding language consultants and collaborators. When they begin collecting data they are first prompted to enter a Session and encouraged to write the session goals. Users are even able to prepare hypotheses in the form of Data lists containing Datum "To be Checked" to prepare for their data elicitation session with their language consultant. In this way LingSync helps a researcher in their collection rationale, collection and analysis methods. When the users add fields to datum they can enter a help text which will pop-up over the datum field as users are entering if they need to know what are the conventions for that field, i.e. IPA transcription could be either phonological or phonetic, the corpus administrator can indicate the conventions using the help text.
   Repository: data can be stored on cloud servers or on department's own server, or researcher's own device. In addition "Bots" have been included as a feature so that users can automatically archive their data in existing, reputable language data archives as described by E-MELD requirements above. Multiple storage locations means less risk of losing data.
   Data organization: Data are stored in a versioned document centered storage solution referred to as "NoSQL." Data is stored in JSON format (exportable to .csv, .xml, .txt, .tex). Unlike SQL databases, NoSQL databases are designed to expand limitlessly and focuses on the ability to provide powerful search of data in context.
   Data management: LingSync allows a number of permissions and data administration. All data is versioned and so mistakes can be recovered easily and discussed via the datum comment feeds.
   Data description: A datum is produced according to leipzig conventions as linguist examples (transcription, morphemes, gloss, translation). Additional metadata description are configured on a corpus basis by the corpus administrator. LingSync offers a list of commonly used annotation fields to researchers which is generated by the popularity of fields among app users.
   Data sharing: Data may be shared with members outside of LingSync via embedding live widgets on their department or lab webpages, or WordPress blogs. Data may also be converted as LaTeX source code to be given directly to members of their data collection team who are not using LingSync. Users can schedule "Bots" to automatically release/publish their data to external web services or language archives.
   Data preservation: Data are stored locally on the users' machines if they use an Android or a Chrome App. The data is also stored on a central server of their choosing, either a server hosted in the cloud, or a department server.
   Budget: LingSync is free and OpenSouce, it can be installed for free on department servers, or institutional data centres, which may require hiring professionals to maintain the database and ensure that the data is properly backed up. LingSync can also run on the cloud and departments can choose to host their data on a cloud hosting provider. The costs usually range between $1 and $10 per month for 1-100 gigabytes of data/data transfer.

2. Collect: Data are collected in such a way to ensure future usability.[6]
   The application provides a template for data entry which is flexible and customizable. The four core data fields (transcription, morpheme segmentation, gloss, translation) are usually required for language data and are set as defaults on the template. Researchers can add extra fields (e.g. IPA transcription, semantic context of utterance) relevant to their research. Contents of data fields (default and additional) can be used in a fine grained way to search, and researchers can organize data using the information contained in the fields. The application is non-proprietary and data collected are exportable in various formats (.csv, .xml, .txt, and .tex) to ensure that data are sharable and usable outside the application.
   When the user exports their corpus a readme.txt is generated which describes the datum annotation fields (using the conventions help text which the users can customize). In this way, when they share their data they can attach the readme.txt to allow others to know what their conventions were when building their corpus.
3. Assure: The quality of data is assured through checks and inspections.
   Datum state (Checked, To be checked, Elicitation Error) and datum comment feeds enables researchers to check, discuss and inspect the data quality. Each datum is also time stamped when entered and modified so that the data history is trackable. Users may also make "Bots" which crawl their data and ensure that it is consistent.
4. Describe: Data are accurately and thoroughly described using the appropriate metadata standard.
   Data will be described at three levels in the application:

   - Corpus: A corpus is a dataset created for a single language/dialect and for a particular purpose. A corpus has a title and a description that will include general information about the corpus such as what the dataset is about, who contributes to the corpus and the purpose of creating the corpus.
   - Session: Each datum in a corpus is tied to a Session which includes metadata such as language, dialect, researcher's name, consultant's code and the goal of the session (e.g. eliciting scope ambiguity).
   - Datum: Data fields and data tags serve for parameters/categories to describe data. The application is not restricted to a particular theoretical construct so that researchers can choose and describe categories appropriate to their research.

5. Preserve: Data are submitted to an appropriate long-term archive.
   Data and associated metadata are stored in a host server for long-term archival purposes. Confidential data and information are stored encrypted using the US Federal approved AES encryption standard to ensure that confidential data cannot be leaked if a corpus is shared or leaked, without the consent of the corpus' author. (See also Data Documentation, Items 1, 5, 6 & 7.)
6. Discover: Data are located and obtained.

[6]While E-MELD recommends Unicode, DataOne recommends plain text ascii characters for variable names, file names, and data. As ascii is inappropriate for linguistic data we will be following E-MELDs recommendation to ensure that IPA, diacritics, semantic calculations and other are maintained in our users data.

The application is accessible through internet search, and data tagged for public view will be discoverable within the application via keyword search. Data permitted for export will be exportable to CSV, text or LaTex formats. (See also 6.1 Data Documentation, Items 3 & 4.)

7. Integrate: Data from disparate sources are combined into one consistent data set.
   Sync function helps integrate data collected by multiple researchers into one corpus. Import/export functions enables integration of data from Filemaker Pro (.csv) and ELAN (.xml), making data integrated with programs researchers commonly use to store their data. Data can be made consistent by the creation of "Bots" who crawl the corpus and automate changes.
8. Analyze: Data are analyzed.
   The four core data fields include morpheme-segmentation and glossing lines, hence the data will contain primary linguistic analysis at the time of the entry to the database. Customizable data entry fields and data tags, as well as the data list functionality allows researchers to organize data ready for further analysis.

1.2.3 User-friendly

Simple. The system will be designed to replace Word Documents or LaTeX documents which is a very common way field linguists store data because it requires no training, doesn't require a complicated set-up for data categories, and takes no time to add new categories. The application will not include categories or linguistic frameworks or theoretical constructs that must be tied to the data. The application will allow data fields and categories to develop organically as data collection proceeds, as opposed to imposing a particular construct upon entry. Researchers will be able to add and change their fields and categories for the data at any point.

Attractive. The system will have a modern design like many of the popular websites such as Google and Twitter. It's layout and background image will be customizable so that the user can change the look and feel of the application to make their eyes comfortable in bright/dark light, or adapt the layout of the widgets to their style of data entry.

Cross-Platform. The application will be available for any device that has an HTML5 compatible browser. Specifically, the application will run *offline/online* in Chrome on Mac, Linux, and Windows computers,[7] as well as *online/online* on Android tablets and phones. The application will run *online only* in Safari and Firefox, and *online only* on iPads and iPhones.

Touch tablets are one of the easiest tools to carry and use in the field; they have a long battery life; they can play videos or show images for the consultant to elicit complicated contexts; and they permit recording audio and video without microphones or cameras which distract consultants. Mobile devices also have apps for push button publishing to YouTube or other audio/video hosting solutions which allow for private data like Google Plus. Furthermore,

[7]The app will also run on ChromeBooks. ChromeBooks are affordable laptops ($299) which use the Chrome operating system created by Google. ChromeBooks are currently available in the UK and online at www.google.com/chromebook/. ChromeBooks have very long battery life and automatically backup data, which makes them good laptops for fieldwork.

Android tablets are particularly easy to program and integrate the microphone/camera directly into the database (Cook, Marquis and Achim 2011).

2 What is iCampo/LingSync?

This application integrates the best functions from existing fieldwork database programs, while avoiding many of the shortcomings discussed above. The dashboard is composed of several widgets. The Data Entry widget is the primary focus, containing four core fields customary for a gloss format (utterance, morpheme-segmentation, gloss, translation). In addition to these fields, researchers can add customized fields, such as phonetic transcription or context for an utterance. Researchers can even upload audio files and link them to the appropriate data. Each data entry is tagged with session info such as the researcher, date of elicitation, language, dialect and consultant's code. The core features are summarized in the following subsections.

2.1 Current functions -GINA

LingSync's functionality can be divided into two groups: functionality for linguistic field databases, and functionality for user friendly community driven software. In this section we will place more emphasis on the linguistic field database functionality and only briefly gloss over some of the more over-arching concerns which make LingSync a user-friendly project with a high proportion of returning users.

2.1.1 Data entry and import

Data entry in LingSync goes beyond just typing or transcribing data. While simply typing in data is the most common use case, LingSync also provides the ability to add comments to any data in the system. This makes it possible to collaboratively enter and discuss data, without modifying or destroying information in the data itself. This means that multiple team members can suggest new segmentation, new gloss information, or qualms about translation or context, and the team can reach a consensus together without blocking team members from accessing or improving data. Comments are also editable and deletable, and can be formatted using Wiki markup which essentially permits the ability to add unlimited documentation to a data record without needing to put the documentation into the record's utterance or translation lines, for example. When working as teams composed of linguists or community members who may speak different dialects, and thus have differing judgements, we believe that comments are a key way teams can provide a maximum amount of access and curation, without worrying about different team members over-writing each other's judgements.

Since all documents in the system are versioned, mis-guided edits by team members can be undone and detected via the team activity feeds §2.1.4. If a team discovers one of its members is not following their team's data curation conventions, the permissions system allows the team to set the individual's permissions to read and comment only.

Data can be categorized by tags as well as by the status of each individual data entry. A record could have a status as simple as "Checked," or even "CheckedWithSeberina" or "ToBeCheckedWithConsultant" if the team is working with multiple consultants and/or dialects. The ability to group data into its validation status further aids organization and permits the team

to gather data for future elicitation sessions, or to send data to consultants to be checked either by exporting the data and sending it by email, or by adding the consultant as a team member.

LingSync is "skinnable", which means that each user can have a different visual representation of the data. Teams can even create non-technical views of the data so that the consultants enjoy being part of the team and feel more connected to the collaborative nature of the language documentation effort.

Data entry and data curation is also fully scriptable. In our user studies we estimate that well over 50% of research time is spent cleaning and curating data, most often to revise old data to update it to what the team's evolving analysis of the data has shown. As such, a fundamental part of field work is exploring and re-analyzing data, so LingSync allows users to create bots which partly automate these tasks. Bots can even be scheduled to run periodically on the corpus, reducing the manual data entry process if, for example, the team decides all data should use the convention "ACC" should be glossed as "CAUS" in the context of "ASP." Bots are able to go through a corpus, and leave comments on data which should be cleaned manually, or even execute the changes after the team has reviewed the changes and approves. There are existing bots for transliteration (conversion of Inuktitut syllabics to romanization) and for duplicating morphemes fields to allomorphs fields, among others. Bots help reduce the redundant tasks, freeing team members to focus on data entry and data analysis. In most teams with long standing databases, it is often the custom to enforce conventions by providing users with drop-downs where they must select only from appropriate options, or go to another screen to add the new option before selecting it. Bots reduce the time dedicated to high quality validation as fields can be populated with autocomplete lists which display options, but still permit new options to be added without visiting an additional screen, and permits the validation to happen after data entry by identifying context to be combined or separated and executing the validation automatically.

Data entry is expected to be grouped by elicitation session. In fact, one expected method of data entry is not data entry at all, but rather recording of an elicitation session followed by typing up the session at a later date. Longer audio files can also optionally be uploaded to the audio web-service §2.2.3 to be automatically split into utterances, reducing data entry and record creation if a team wishes to record elicitation sessions and enter the data later. We strongly recommend this approach to data entry as it permits the team to dedicate 100% of their attention to the speaker while eliciting data, rather than dividing their attention between the speaker and the process of data entry.

One of LingSync's founding principles is that you should only need to enter data once. Whether you enter it in an Excel Spreadsheet, in a handout, in Elan or in FLEx, you should be able to import it into LingSync without needing to re-enter the data. Each record in a LingSync database can have an unlimited number of fields, with unlimited complexity, making it possible to import other formats, and be able to re-export them without losing any information (for example, timed alignments in Elan or Praat).

2.1.2 Auto-glosser

The semi-automatic glosser requires no configuration or set up to be useful. It "learns" from the data in your corpus to guess where morphemes might be segmented, or how morphemes should be glossed. The glosser is also a separate module, meaning that if you have an existing glosser you can plug it in to LingSync. Glossers can also be shared. For example, if you have two Quechua corpora, you can set the glosser url to use either corpus.

The glosser is designed to make the app "smarter" and to reduce the amount of time spent entering predictable information such as glosses. The glosser can use any existing morphological analysis tool to break down the utterance/orthography line into a probable morphological segmentation using known morphemes in the lexicon, and enters a probable gloss for the morphemes in the glossing line. However, the glosser module is designed to reduce redundant data entry, not to provide full, accurate glosses. It is of course crucial that predicted morpheme segmentation and glosses be corrected by users, particularly in languages that have many short or ambiguous morphemes, which will result in more possibility for error in automatic morpheme segmentation.

The glosser uses the auto-generated lexicon to evaluate morphemes both by precedence relation and by gloss. Each corpus has its own lexicon, which is loosely modelled after a mental lexicon, as a network of morphemes, allomorphs, orthographie(s), glosses and translations. It is not a dictionary but rather a connected graph similar to theoretical models of mental lexicons (for a dictionary see the Dictionary Module in § **??**). As a connected graph it is the most useful structure to index datum and search for datum real time while data entry is happening. Currently, there is no user interface to view/edit a corpus' lexicon but if there is enough demand it will prioritized.

2.1.3 Search

LingSync was designed for powerful search. You can search your corpus, or across your corpora. Similar to ELAN (Wittenburg et.al. 2006) The results of your searches can even be saved as a Data Lists which can be sorted, saved for later exporting or curating data for a handout or language learning lesson for heritage speakers.

(1) Search for 'yell' in an entire corpus, (to find examples in gloss and translation and comments etc)

Search can be as simple as a key word search, which will search the entire record, or search with in only one field, or for example one key word in one field or another key word in another field. For those users who like to think in Set Theory LingSync provides you with the ability to look at the Intersection of search results, or in the Union.

(2) Search for 'nay' in the morphemes line, or 'des' in the gloss

For phonologists LingSync lets you search using regular expressions to find segments in context.

(3) Search for 'nay[tk]' in the morphemes line to find context of allomorphy

If there is demand we can add the ability to search for minimal pairs or to search for phonological features in context using a phonology ontology (a general purpose feature geometry/articulatory feature ontology, or a customized ontology created by the users for their language of interest) where feature geometry searches could be used.

Phonological search lets the user search for potential minimal pairs or phonological features in context to verify with consultants, and/or to prepare psycholinguistic experiments.

(4) Search for 'nay-voice' in the morphemes line to find context of surface vowels

Search has a The phonological search module shown in Table **??** is used to search for phonological features in context.

2.1.4 Sharing corpora, activity feed

Corpora in LingSync can be shared as a team, with administrators, who cannot see the data, but can add new team members (e.g. a project coordinator); writers, who cannot read the data but can enter new data (e.g. language consultants, or psycho-linguistic experiment participants); readers, who can see the data but cannot edit it (e.g. external collaborators) and commenters, who cannot edit the data but can provide feedback and offer additional information or corrections (e.g. consultants and/or collaborators). Of course, most teams will choose to give all roles to all users, but these roles permit a wider inclusive data collection team than previously available in other data management tools where the permissions are simply full access or no access.

As a team, you might also want to catch up on recent activity in the corpus. If your corpus is small (only a 100 records) you could simply read each record to see what is new, but for larger corpora or where there is more activity LingSync provides team activity feeds. In the activity feed widget you can see who has modified, commented, created data, as well as recording audio, and putting records in the trash to be deleted later. There are also user activity feeds which are only presented to the user: although less interesting than the actions of one's teammates, they can help users remember what they were working on last time, particularly if it has been months since they last opened the corpus. Our user studies and previous experience as field workers indicate that most users visit their corpus very frequently , when building it, and the usually more sporadically as they need to return only to consult their data. For example, the user activity feed could help you remember that you hadn't finished typing up that elicitation session three months ago before you had to go to class.

2.1.5 Custom settings

LingSync is highly customizable. It comes with the ability to choose from 5 popular dashboards, and even to create your own. Users can decide how many records to show on a page of data, which order to show the records in and many more options. For users who have limited eyesight or who are using screen readers, LingSync provides a high-contrast option as well as a dark option to reduce eyestrain after hours of entering data. LingSync was designed partly by research assistants who had entered data for 40 hours a week, and so it also provides the ability for rich and interesting background pictures to keep entering data visually stimulating.

Each corpus is also fully customizable: team members can add new fields to the corpus, as well as edit its terms of use and other information which can become important if the team decides to share the result of their work with the outside world. We have added many other options to corpora which conform to EMELD recommendations discussed in §1.2

2.1.6 Export

As users of many other data management software, we felt it was crucial that LingSync be nonproprietary and open. One important aspect of this is the ability for teams to export their entire database in any format they choose, in its entirety or only data which are relevant to a certain export goal.

Teams can even save lists of data for dedicated export purposes, such as data for a handout, or data which they are curating to be published as stories for the language community they are working with.

LingSync is also able to export word lists, which can be used either as language learning exercises for heritage speakers or as materials for field methods courses.

It is possible to export an entire corpus either as a zip, as XML or JSON, as plain text, as LaTeX and as CSV.

Beyond export, LingSync databases are fully replicable between servers, which means that team members can have entire copies of their database locally on their laptops, yet still remain in full sync with other team members when they go online. It also means that departments can back up their data to their own department servers without worrying that data may become stale or out of date.

2.2 Plugging into LingSync -GINA

One of the strengths of LingSync is that is built using well-understood web technologies which permit the creation and integration of nearly any existing software as web services, and if the software provides a javascript or HTML5 library or widget, even complex user interfaces can be combined and integrated with LingSync via the NPM and Bower web module management system. In this section we will discuss some of the current web services.

2.2.1 Custom glosser

We have wrapped Benoit Farley's morphological analyzer for inuktitut into a web service using Node.js ***link?

2.2.2 Language learning module for android

A prototype for language learning was build which enables researchers and language teachers to create language learning aids from the data in existing corpora and from the data newly collected for the purpose of language learning. The learning aids aim to help heritage language learners improve their listening and speaking skills. The orthographic lines (i.e. utterance and morpheme lines) and the attached audio or video recordings of a datum are taken as materials to create a lesson.

2.2.3 Integration with ProsodyLab aligner

The phonetic aligner web service it possible to upload audio recordings and the orthographic/utterance lines of datum to create a dictionary unique to the language of the corpus, and to run the ProsodyLab Aligner, a machine learning algorithm which uses Hidden Markov Models to predict boundaries between phones and creates a Praat TextGrid with estimated phone boundaries, saving hours of boundary tagging.

2.2.4 WebSpider

The Web Spider allows teams with limited access to consultants to gather data using blogs or forums or online translations of the bible. The web spider also provides an additional source

of context to assist consultants in providing grammaticality judgements, as well as additional contexts where morphemes appear. For example, "ke" is largely considered a postposition by Urdu-consultants with explicit knowledge, however it is often produced as other functional morphemes in everyday spoken contexts. Blog/forum data can be used to discover these additional contexts. citation needed ***

2.3 Collaborators - GRETCHEN

LingSync is a product of collaboration between field linguists, programmer linguists, and programming fieldlinguists, involving many linguistics students as interns who have been trained to code and subsequently contributed to the app.

–> perhaps Gina you could talk more about this? Maybe move to the next section with other types of people?

3 How is iCampo/LingSync used so far?

3.1 McGill-Listuguj partnership

LingSync is currently in use by the Mi'gmaq Research Partnership, a collaborative project by McGill linguists and Mi'gmaq speakers and learners in Listuguj, Quebec. Several (number?) members of the team working with the language have been entering language data into a shared corpus since Fall 2013.

3.2 Field methods classes

LingSync was used by two field methods classes in Winter 2013, at the University of Ottawa (Keewa?) and Palona (sp? language?) and is being used by an additional three field methods classes in Winter 2014, at McGill (Inuktitut), Yale (Quechua), and the University of Connecticut (language?). The languages of these classes are typologically diverse and the teaching styles are varied. We appreciate all of the contributions of the students and instructors of these classes in reporting issues that they have had so that we can further improve LingSync.

3.3 Other users

Our current user registration statistics show that over 300 people have created user accounts with LingSync, and have created over 1000 corpora. Although many of these accounts appear to be 'testing' accounts to try out the app, ?? number of people have been using LingSync actively. (Server logs show the developers logging in and saving activity but not the content of the data.)

4 Conclusion - YULIYA

References