An Overview of FieldWorks and Related Programs for Collaborative Lexicography and Publishing Online or as a Mobile App

David Baines
SIL International
E-mail: david_baines@sil.org

Abstract

The FieldWorks ecosystem provides open-source tools for linguists whether working alone or in distributed teams.

FieldWorks is a comprehensive tool for managing linguistic data. It has an extensive selection of fields for each lexical entry and areas for storing grammatical data and interlinear texts. The bulk editing tools can save hours of work by operating on many entries at once. FieldWorks can be used to create mono- or multi-lingual dictionaries and has excellent support for complex scripts. Comprehensive help and resources are available within the tool, which is designed for trained linguists.

Language Forge is an online dictionary creation tool that allows collaborators to browse, comment or contribute to a lexicon. The project manager can control the roles for each team member. Language Forge shares the FieldWorks data allowing users of either tool to modify a shared lexicon. Language Forge can be used with minimal training as it exposes only a small subset of the FieldWorks data.

Webonary is an online platform for publishing dictionaries and their reversal index. The linguist can update the data on Webonary from within FieldWorks as often as desired. Dictionary App Builder facilitates the creation of Android and iOS apps from the FieldWorks data.

Keywords: FieldWorks, Language Forge, collaboration, multilingual, complex scripts, Online publishing, Webonary, mobile publishing, Dictionary App Builder

1 Introduction

In this paper I will briefly discuss the capabilities of FieldWorks (FLEx) with a particular emphasis on possibilities it provides for collaboration, both with other linguists and with other software. There are two broad categories of software that work with FLEx; software that enables collaboration on the same data and software that prepares the data for publication in various media. Key limitations of FLEx will also be considered. The aim of this paper is to enable the reader to know whether (or not) FieldWorks and its ecosystem will be of use to them. Recommendations are given about how to investigate further should this paper be insufficient to determine the suitability of FLEx for a particular project.

2 FieldWorks

2.1 Introduction and a Little History

FieldWorks was once a suite of programs developed by SIL International as a repository for the academic data that a field-based linguist would want to store about a single language and culture.

Language Explorer was the linguistic component and Data Notebook managed anthropological notes. The Data Notebook functions have been incorporated into Language Explorer, and "FieldWorks Language Explorer" (FLEx) is currently the only program under development. Writing about the first release in November 2006, Moe (2008) describes its purpose as follows: "Language Explorer is designed to create and manage a dictionary, create and maintain a text corpus, interlinearise texts, and study morphology". The website for FLEx describes its functions in this way:

Fieldworks Language Explorer (FLEx) enables linguists to be highly productive when building a lexicon and interlinearising texts. Powerful bulk editing tools can save hours of work. Fieldworks allows control of which fields and entries show up in a dictionary publication. Through Pathway, beautiful dictionaries can be exported easily. Send/Receive Project allows users to collaborate with colleagues located anywhere (2018).

2.2 FieldWorks: Data

FieldWorks is built on a complex data model which includes hundreds of possible fields. Dictionary entries can be described and annotated in great detail. There is also the facility to add custom fields to any project which needs to store further information. Custom fields can be added at the level of entries, senses, examples and allomorphs. One great advantage of a system that is built on a database structure is the ability to maintain referential integrity. FLEx will manage the lexical relationships between entries, so, for example, they can only be added if both entries exist in the data. Similarly before the deletion of a related item, FLEx will show a warning to alert the user. Should the entry be deleted then the lexical relationship is automatically deleted from the referent. FLEx maintains a clear distinction between the data and the presentation of the data. This allows the data to be presented in many different forms through many different media. Comprehensive dictionary formatting options are available which use HTML and CSS to format the output ready for draft printing. For even finer control over the format of the output, other programs such as Pathway can be used.

2.3 FieldWorks: Collaboration

Version 8 of FLEx, released in April 2014, added the ability for multiple users to collaborate on a shared dataset. A project repository may be stored online, on a network drive or even on a USB stick. Online hosting of repositories is available at languagedepot.org.

[LanguageDepot] is provided as a service to language communities by the Language Software Group of the Linguistics Institute at Payap University, Thailand. It is for teams collaborating with FLEx, WeSay and Language Forge (2018).

2.4 FieldWorks: Writing System and Complex Script Support

Every piece of textual data must be expressed in one language or another, and knowledge of the language is vital to understand its meaning. You may have seen the T-shirt with the slogan "There are only 10 types of people in the world, those who understand binary and those who don't." The code switching from English to 'Binary' isn't obvious when both languages share the same script and that can cause confusion or a loss of information. A single language can be represented with multiple scripts, for example English can be represented in Latin, IPA, Morse code or Braille. Most programs rely on the operator to know or to recognize both the script and the language of each piece of data. However, a linguist creating an orthography for an unwritten language needs to process vernacular data in closely related writing systems, such as phonetic and phonemic data both expressed in IPA. In some cases the data is identical, and without making the writing system explicit it is impossible to

know exactly what the data means. To address these potential problems FieldWorks explicitly stores the language and script information as metadata on each piece of data entered. FieldWorks supports the vast majority scripts including right to left and complex scripts such as Arabic, Devanagari and Tamil and even the sloping Arabic script: Nastaliq.

2.5 FieldWorks: Rapid Word Collection

Gathering words for a dictionary in a minority language would often take a single field-based linguist many years. In order to accelerate the work the Rapid Word Collection method was developed. This involves people from the language community working together over the course of a two week workshop. During the workshop thousands of words and senses are gathered by the participants. The method was used by speakers of Gusilay in the town of Thionck-Essyl in the south of Senegal. They collected a total of 12,485 words in 11 days, a fairly typical result for a RWC workshop. FLEx includes a tool to facilitate the Rapid Word Collection method.

While the text corpus method can produce similar results it can't be used for a language where a sufficient text corpus does not yet exist. In many cases FLEx has been used to facilitate and inform the creation of a writing system for a minority language community. However, since such people have only recently begun to write their own language, few texts are available. For many minority languages it will be many years before a text corpus exists that is sufficiently large for a text-corpus approach.

2.6 FieldWorks: Interlinear Texts and Discourse Analysis

Two parsers are included in FLEx, one is XAmple and the other is the phonological rule-based parser HermitCrab.NET. Once the lexicon is sufficiently complete and the grammatical rules and information have been supplied, these parsers can be used to facilitate the analysis and interlinearising of texts. FieldWorks is able to analyses a text and provide guesses at the most likely morphemes and translation equivalents. This can greatly accelerate the work of interlinearising. FLEx also contains a tool to facilitate discourse analysis of longer texts that show discourse features.

2.7 FieldWorks: Limitations and Contingencies

2.7.1 Collaboration

The send/receive system within FLEx enables asynchronous collaboration. If two team members edit the same item of data and then send/receive, FLEx will show that the edits are in conflict and allow the users to resolve it. Good inter-team communication reduces the number of conflicts, as does the practice of using send/receive before and after each session of work. If the project is stored on languagedepot then team members need an internet connection to be able to send and receive, although the bulk of the work in FLEx can be done while offline.

2.7.2 OS support

FieldWorks is now released for Windows and Linux, but there are no plans to support any other platform. However, while it won't run natively on a Mac it runs well on Windows in Parallels or VirtualBox.

2.7.3 Training Requirements

Users need a good understanding of linguistics to be able to make full use of the program. However there are comprehensive resources available from the help menu that describe how to use the tools for

lexicography and interlinearization. The use of the parsers is also described in some detail. A series of training videos have been produced that should be of great help to new users of FLEx, and a few universities are now teaching FLEx as part of a Field Methods class in their undergraduate linguistics course.

2.7.4 Designed for Manual Analysis

FieldWorks is designed primarily to assist linguists in their own analysis of a language. It isn't designed for corpus linguistics or automatic analysis of large quantities of text, nor is it particularly helpful for comparing words across multiple languages.

2.7.5 Migrating Data from Other Tools

Data can be imported into FLEx from Standard Format Marker (SFM) files used by Toolbox and Shoebox. The task of importing data may be fairly easy if the data is consistent and simple. Often the task is complicated by the need to make the data consistent and explicit before importing. Some data formats, such as CSV, can fairly easily be transformed into SFM, but others may require significant work or specific skills in order to convert them.

The import process makes a series of XML files, each one created by processing the previous one with an XSLT. Advanced users may use one of those XML formats in order to import their data into FLEx. In theory an XSLT could be produced that would convert an existing XML or TEI file into an XML format that could be imported into FLEx.

Expert help is available for the process of importing data into FLEx from SIL International's Dictionary and Lexicography Services.

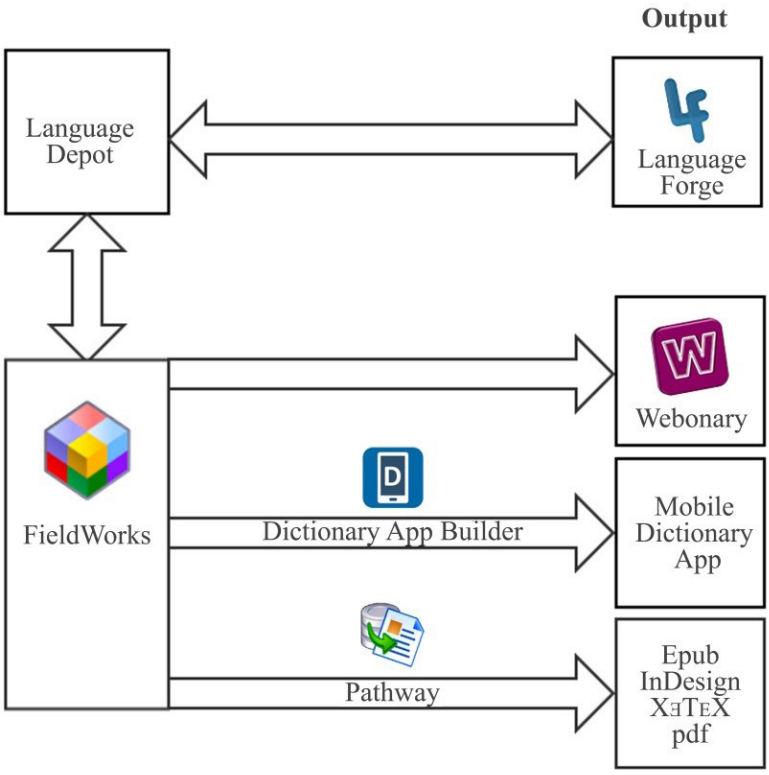

Figure 1: FieldWorks ecosystem overview.

3 Language Forge for online dictionary creation.

A visual representation of the interactions between FieldWorks and related programs is provided in Figure 1. This shows that the FLEx data can be stored in a LanguageDepot repository. Language Forge provides an online interface to the data so that the same set of data can be edited with either FLEx or Language Forge. As it operates within a browser, Language Forge can be used on most operating systems and devices, including tablets and smartphones. Language Forge may be particularly useful when many people from a language community wish to create a dictionary in their own language. One advantage that it has over FLEx is its real-time collaboration feature. "Members of the project will see entries updated as it happens by others in the same project" (Language Forge website 2018). Another advantage is that the interface is much simpler than that of FieldWorks, and it is suitable for use by those without linguistic training. Language Forge presents a only subset of the lexical data that is available in FLEx.

4 Webonary for Online Dictionary Publishing

Webonary offers online dictionary publication, particularly for minority language dictionaries. Webonary currently has dictionaries in 119 different languages, from 38 countries. "Webonary gives language groups the ability to publish bilingual or multilingual dictionaries on the web with a minimum of technical help" (2018). In one sense the task of creating a dictionary is never ending, so a small team with many other demands on their time may never get around to sharing the results of their work-to-date. Webonary lowers the technical barriers to publishing a dictionary online, and thus language communities that wish to create a dictionary for their own use, or are documenting their own language, can easily publish the results of their efforts. For language communities with no dictionary at all, even a draft dictionary is useful. The site therefore encourages editors to publish early and update often, and this is easy to do from within FieldWorks. A graphic indicates the publication-status on a scale from "Rough-draft" through to "Formally published". Sharing the data early in the project provides the possibility of gaining input from the readers of the dictionary in the form of comments on the site. Useful feedback may be gained to improve existing entries and add new ones.

5 Dictionary App Builder for Smartphone Dictionary Apps

Dictionary App Builder (DAB) makes it possible to create a mobile app using data exported from FLEx. No programming is necessary in order to create the apps; however the program does have multiple dependencies making the installation of DAB more complex than it is for most programs. The Mac OS version can create iOS and Android apps, while the Windows and Linux versions can only create Android apps. DAB allows control over many aspects of the apps it produces, including the choice of colors, fonts, and icons to be used. "The apps do not require an internet connection. All content can be packaged together for offline use and distribution, or audio can be made available online for download when needed" (Dictionary App Builder website 2018). Dictionary App Builder also contains an interface for localization, so that it is very easy for the editors to provide apps with an interface in any language. Apps can be published through the Google or Apple stores. Users of the Android apps created with DAB can share the app, complete with its data, with other Android users. This isn't possible on iOS, as Apple prevents side-loading of apps.

6 Pathway for Typeset Output

Pathway provides a way to transform the dictionary data, exported from FieldWorks, into other formats. These include word processing formats such as DOCX and ODF. For higher-end typesetting, InDesign and XƎTEX output formats are available. These options allow greater control over the style and formatting of the dictionary prior to publication.

XƎTEX is an extension of TEX that integrates TEX's typesetting capabilities with (a) the Unicode text encoding standard (supporting most of the world's scripts) and (b) modern font technologies (TrueType and OpenType) and text layout services (XƎTEX website 2018).

7 Conclusion

FieldWorks is useful for teams and individual linguists manually analyzing a language. It supports the vast majority of complex scripts, and maintains detailed metadata about the language and script of each data point. Team members can collaborate by synchronizing the data regularly with a server. Language Forge provides real-time collaboration on a sub-set of the data. Many publishing options are provided: through Dictionary App Builder for mobile apps, Webonary for online, and Pathway for print and Epub publication. There are active user communities where answers to specific questions or advice about specific use-cases can be obtained. Finally, I would encourage experimentation with the programs if there is doubt as to their suitability for a given project.

References

Dictionary App Builder. Accessed at: https://software.sil.org/dictionaryappbuilder/ [07/03/2018]
FieldWorks. Accessed at: https://software.sil./fieldworks/ [07/03/2018].
LanguageDepot. Accessed at: https://public.languagedepot.org/ (login required) [07/03/2018]
Language Forge. Accessed at: https://Language Forge.org/ [07/03/2018]
Moe, R. (2008). FieldWorks Language Explorer 1.0 *SIL Forum for Language Fieldwork 2008-011*, pp.1-4. Dallas, USA.
Webonary. Accessed at: https://www.webonary.org/ [09/03/2018]
Pathway. Accessed at: https://software.sil.org/pathway/ [09/03/2018]
XƎTEX. Accessed at: http://xetex.sourceforge.net/ [25/03/2018]
Rapid Word Collection. Accessed at: http://rapidwords.net/ [14/6/2018]