

Lexical Databases and FIELD

Verónica Grondona, *Eastern Michigan University*
grondona@linguistlist.org

1. Our Project

- documentation and description of 3 endangered languages of the Chaco region of South America (Chorote, Chulupí and Kadiwéu)
- researchers: Lyle Campbell (U. of Utah), Filomena Sandalo (UNICAMP, Brazil), Verónica Grondona (Eastern Michigan U.)
- Funded by the Hans Rausing Endangered Languages Programme, SOAS

The Goals:

- language documentation
- language description (dictionaries and grammars of each language)
- materials development (Instructional materials, collections of traditional stories, etc.)
- comparative and historical studies of languages of this area

2. Lexical databases in the Chaco project

- Excel spreadsheets to
 - input fieldnotes
 - organize and analyze data (manually)
 - compare forms from different languages
- Advantages:
 - fast input environment
 - uses Unicode
 - can export XML
 - highly structured (→ allows for simpler conversion)
 - easy to use and understand
- Disadvantages:
 - it is not specifically designed for linguistic data
 - analysis is done manually
 - does not allow for different ways of structuring the data for outputs of different sorts (for this the data needs to be exported to XML and then imported into some program that allows for such automatic restructuring)

3. FIELD

- is a generalized data input tool
- developed by the EMELD project (LINGUIST List, Eastern Michigan University, Wayne State University, University of Arizona)
- developed by linguists and experts in language technology with the input of field linguists, and using typologically different languages as pilot languages
- The goal: to produce a flexible tool which is customizable to accommodate language data of many different language families and typological configurations
- currently works on an Oracle database that exports in XML, and is accessible online, but a stand-alone non-Oracle database version is being developed
- includes CHARwrite©, a web-based IPA keyboard for the input of phonetic transcription
- What it does:
 - provides a set of terms (a “Term Set”) to analyze the language, but it also allows the researcher to use his/her own terms
 - facilitates the analysis of data (especially morphological & semantic analysis) allowing the researcher to find patterns in the data
 - allows the researcher to fit data into a “framework” (which can be customized and modified all along)
 - is Unicode compliant (If the researcher does not use strict IPA standard symbols they need to be mapped to Unicode characters)
 - allows collaborative work among different researchers; it keeps records of what each person does, when, etc.)
 - allows the researcher to assign different levels of access to data
 - allows a “keyboard set-up”, i.e. to set up a series of keystrokes in one’s keyboard to type certain non-standard characters (they must be keystrokes that are not already used by Windows)
 - will build a paradigm from the data that is input
 - using XSL stylesheets → allows the researcher to build different types of dictionaries depending on what is needed (e.g. one dictionary for native speakers of the language, a dictionary with a (slightly?) different format for linguists, maybe a different format for pedagogical purposes, etc.)
 - allows comparative studies among languages in the database
 - allows the upload of XML files (with appropriate mapping of XML tags)

- it is an initial testing version that is being improved
 - The beta version allows you to try the input tool and browse forms that already have been entered
- Under development also is a tool (Onto-Gloss, being developed at Wayne State University) that allows the mapping of a Term Set onto a text that is saved as XML and then input into FIELD
- How to try FIELD:
 - <http://www.emeld.org/tools/fieldinput.cfm>, or <http://www.emeld.org> and follow the link to the FIELD Data Input Tool
- an enhanced version of FIELD with enhanced capabilities is under way, with the following features
 - user-friendly import of XML files and mapper for XML tags
 - user-friendly input worksheet to optimize data-input time for the researcher
 - build a sound chart based on data that input in DB
 - mapping non-standard characters to standard Unicode IPA characters
- The EMELD project can only develop with your help; we welcome the advice and suggestions of the linguistic community.
- at a basic level FIELD works very well, but it is still being developed, and a lot more needs to be done

4. Conclusions

- Excel has worked well for our project for creating lexical databases, but it is not necessarily the best
- FIELD is basically a very good tool, but it still needs to be refined and improved
- The ideal tool for linguistic data is yet to be developed. Of the existing tools, it all depends on the researcher's needs and purpose(s). But when choosing a tool, it is advisable to choose one that allows for exporting in XML, that allows the use of Unicode characters, and that is highly structured to facilitate mapping and conversion.

For suggestions and tips on best practice in digital language documentation check out the EMELD School of Best Practice at <http://emeld.org/school/index.html>

Appendix A

List of Lexicon Management Tools (from the EMELD School of Best Practice)

The Linguist's Shoebox
 Shoebox
 WordCorr
 Toolbox
 Toolbox
 Kura 2.0-1-2.1.2
 LinguaLinks Workshops
 Wordcorr 2
 A Simple Concordance Program
 Ellogon
 FIELD
 Lexique Pro
 LT XML
 PamSurv
 PhoneBox
 Rook
 SIL FieldWorks
 System Quirk

You can find suggestions that may help you decide what software program is best for you at the EMELD School of Best Practice, on the page that provides advice on "Choosing Software" (<http://emeld.org/school/classroom/software/index.html>)

