Lexical Databasing

Verónica Grondona
Eastern Michigan University

Outline

- General information about our project
- What we use for lexical databasing
- A data input tool: FIELD
- Conclusions ... or “what we (Excl) have learned from this”

Matacoan and Guaycuruan languages

Chulupí

Chorote

Wichí

Kadiweu

Pilagá

Maká

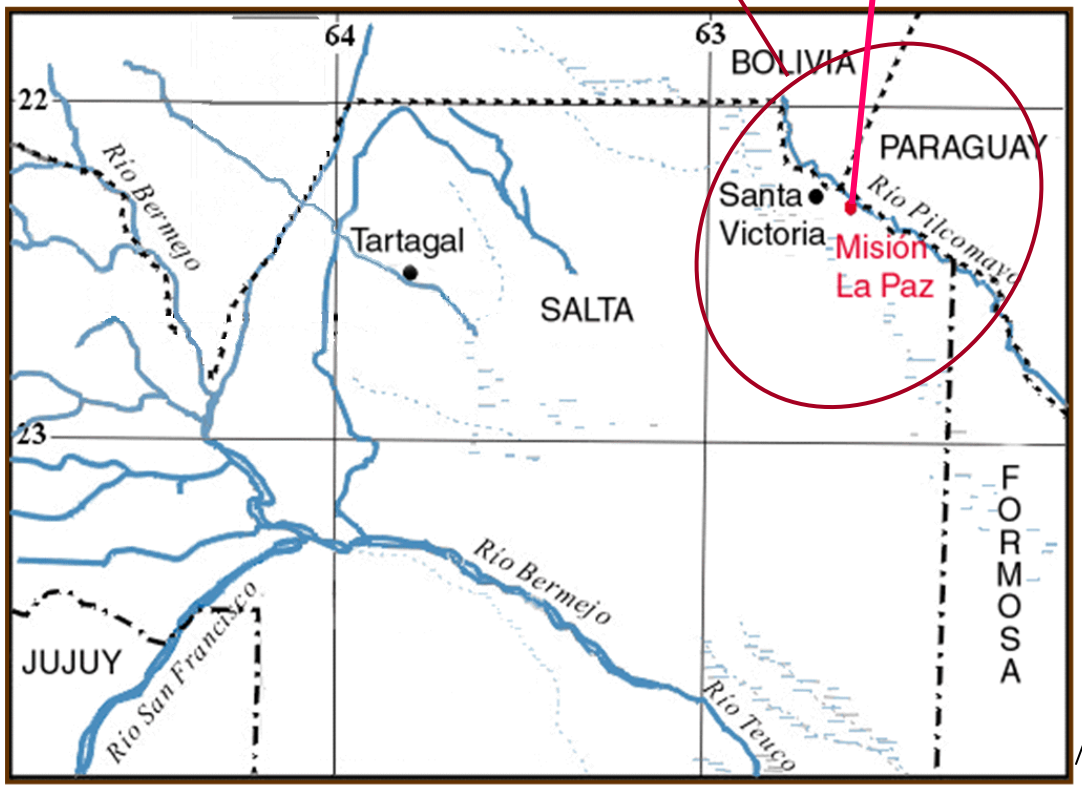
Toba

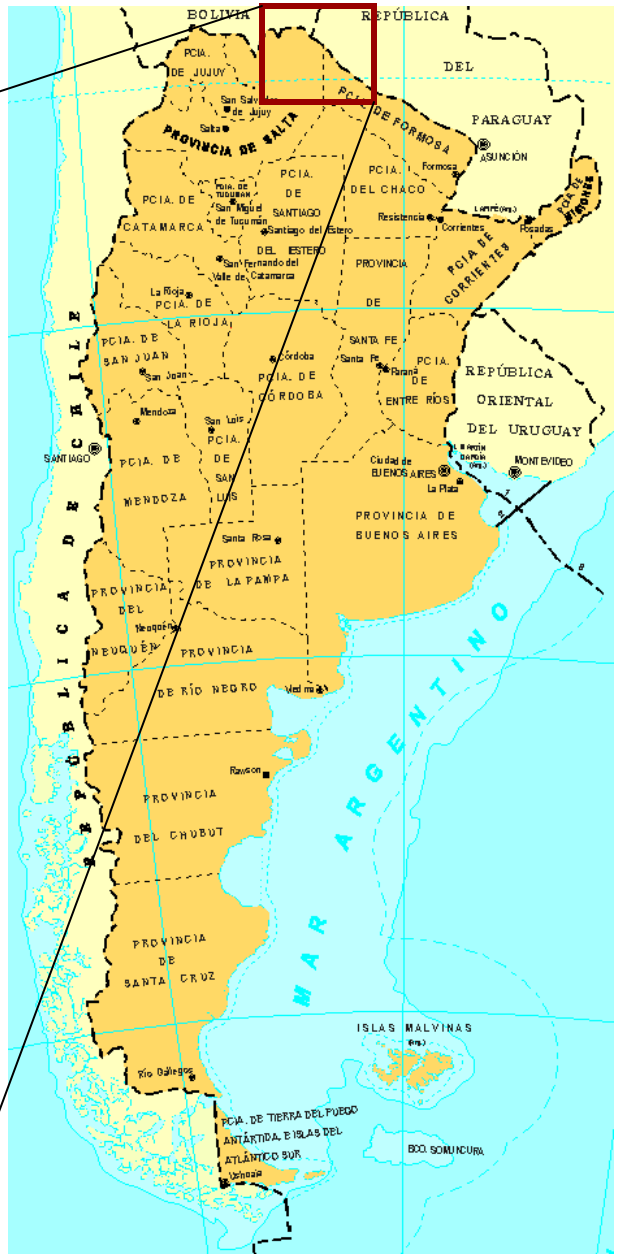
Mocoví

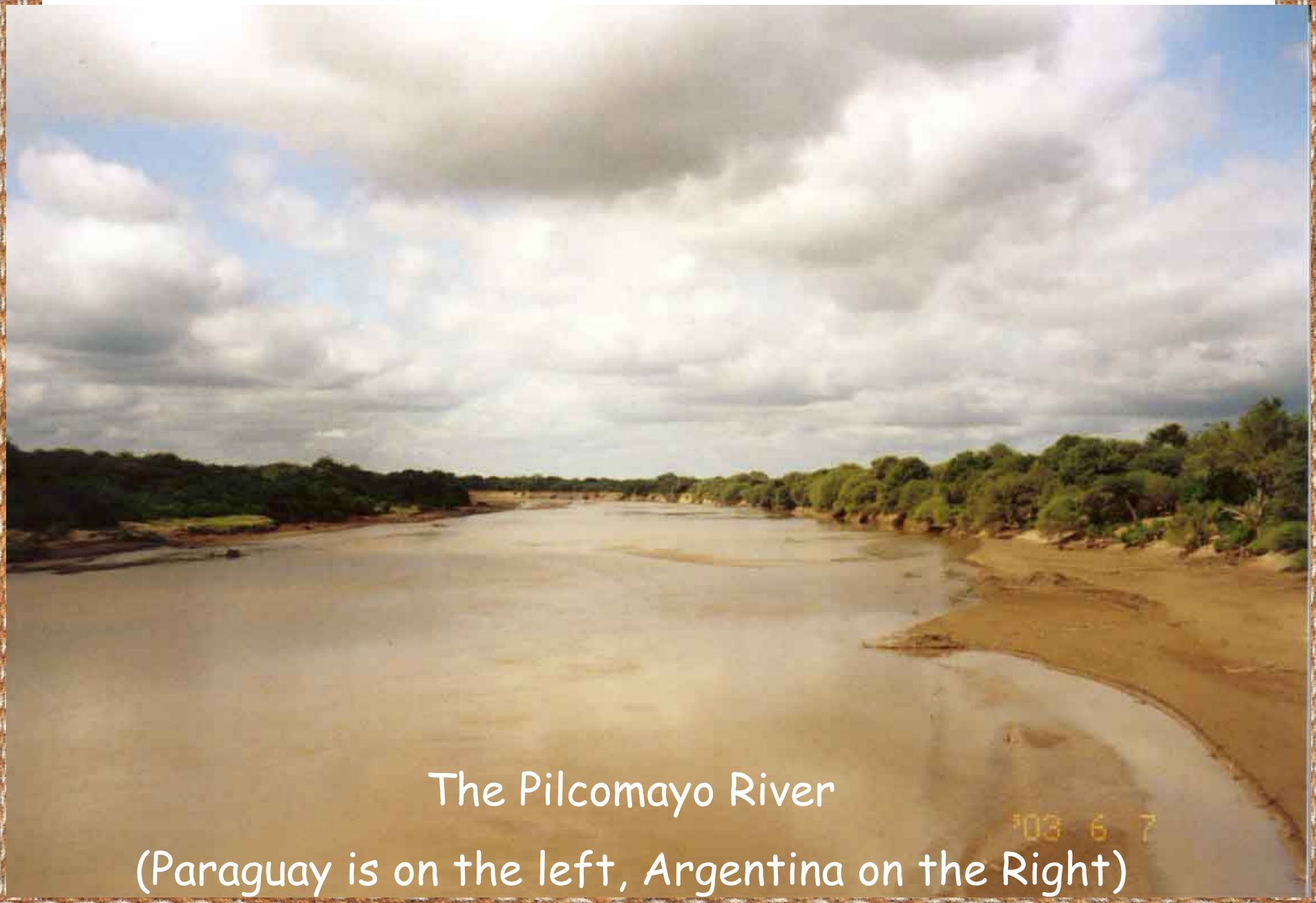
Where we work:

Chorote/Chulupí area

Misión La Paz







The Pilcomayo River

(Paraguay is on the left, Argentina on the Right)

The Project:

- Language Documentation Project (Funded by the HRELDP, SOAS)
- 3 languages: Kadiwéu (Guaycuruan, Brazil), Chorote and Chulupí (Matacoan, Argentina & Paraguay)
- Linguists: Lyle Campbell (U. of Utah), and Filomena Sandalo (UNICAMP), Verónica Grondona (Eastern Michigan U.)
- Objectives:
  - language documentation
  - language description (dictionary and grammar of each language)
  - materials development (Instructional materials, collections of traditional stories, etc.)
  - Comparative and historical studies of languages of this area

Lexical database:

- Excel
  - input fieldnotes
  - organize and analyze data (manually)
  - compare forms from different languages

Advantages of Excel:

- fast input environment
- Unicode
- can export in XML
- highly structured (→ allows for simpler conversion)
- easy to use and understand

Disadvantages of Excel:

- it is not specifically designed for linguistic data
- analysis is done manually
- does not allow for different ways of structuring data for outputs of different sorts (data must be exported to XML and then imported into another program for more ‘automatic’ restructuring)

FIELD:

- is a generalized data input tool
- developed by the EMELD Project (LINGUISTList)
- developed by linguists & experts in language technology with the input of field linguists (using typologically and genetically diverse languages)
- the goal: to produce a flexible tool (for lexical databases) for linguists
- currently works on an Oracle database that exports in XML, is accessible online
- includes CHARwrite

What FIELD does:

- provides a set of terms to analyze the language, also allowing the linguist to use his/her own
- facilitates the analysis of data
- allows the researcher to fit the data into a ‘framework’ (which can be customized & modified all along)
- Unicode compliant
- allows collaborative work
- allows different levels of access to data

More about FIELD:

- it is an initial testing version that is being worked on and improved
- How to try FIELD:
  - www.emeld/org/tools/fieldinput.cfm
  - www.emeld.org

The next step for FIELD:

- user friendly inport of XML files and mapper for XML tags
- user-friendly input worksheet to optimize data-input time for researchers
- some ‘automatization’ of data analysis (e.g. building a sound chart based on the data input in the database)
- mapping of non-standard characters to standard Unicode IPA characters

FIELD:

- at a basic level, it works very well, but it is still beind developed, and a lot more needs to be done
- the EMELD project can only develop it with the help of (field) linguists; we welcome advice and suggestions from the linguistic community

What we (Excl) have learned:

- Excel has worked well for our project for creating lexical databases, in organizing and analyzing data...
- ...but it is not perfect, and might not be the best
- FIELD is basically a very good tool...
- ...but it still needs to be refined and improved
- There are other tools (FileMaker Pro, Shoebox, Toolbox, etc.)
- The ideal tool for linguistic data is yet to be developed
- Of the currently existing tools, it all depends on the researcher's needs and purpose(s)
- When choosing a tool it is advisable to:
  - determine what your needs and purpose(s) are
  - choose one that
    - allows for exporting in XML
    - allows the use of Unicode characters (or allows you to map your own characters to Unicode characters)
    - is highly structured to facilitate mapping and conversion

When choosing a tool:

- It is advisable to determine what your needs and purpose(s) are
- Choose one that
  - allows for exporting in XML
  - allows the use of Unicode characters (or allows you to map your own characters to Unicode characters)
  - is highly structured to facilitate mapping and conversion

A suggestion:

For information and tips on best practice in digital language documentation check out the EMELD School of Best Practice at

http://emeld.org/school/index.html


'04 6 30


NIKE
'04 6 23

Thank you!