

## Creating an online dictionary as a tool for language maintenance



Fataluku  
Language  
Project

Ruben Stoel  
Leiden University  
Netherlands



[www.fataluku.com](http://www.fataluku.com)

## Online dictionaries for endangered languages

- endangered languages
- Fataluku project
- the data
- the dictionary file
- the database
- searching

[www.fataluku.com](http://www.fataluku.com)

## National language vs. local languages

- many countries have more than one language:
  - a national language, used in education
  - several local languages, used at home
- language shift:
  - speakers start using the national language when talking with their children
  - local languages become endangered

[www.fataluku.com](http://www.fataluku.com)

## Problems of endangered languages

- low status
- few speakers
- not written / no standard spelling
- often several dialects with no standard dialect

[www.fataluku.com](http://www.fataluku.com)

## Language documentation

- recording, annotation and translation of texts (stories, conversations, songs, etc.)
- producing a lexical database
- archiving
- providing access for speakers / linguists

[www.fataluku.com](http://www.fataluku.com)

## Why a dictionary?

- introduces a standard spelling, so speakers may start writing their language
- enhances the prestige and visibility of the language for speakers (and also linguists)
- may serve as a tool for language maintenance or revitalization

[www.fataluku.com](http://www.fataluku.com)

### Why an online dictionary?

- can be made available immediately, even when the data is still limited
- can be updated continuously
- can reach speakers living abroad
- is cheaper to produce than a printed dictionary

[www.fataluku.com](http://www.fataluku.com)

### Fataluku Project

- Fataluku: Papuan language spoken in East-Timor
- Official language is Portuguese, dominant language is Tetun (Austronesian)
- Fataluku speakers are shifting to Tetun
- Indonesian (Malay) and English are also important

[www.fataluku.com](http://www.fataluku.com)

### Fataluku Project

- Website with recordings, transcription, and translation of traditional stories, songs etc.
- Online dictionary: multilingual (Fataluku ↔ English / Tetun / Malay / Portuguese)
- Users are mainly Fataluku speakers living abroad and UN personnel

[www.fataluku.com](http://www.fataluku.com)

### Data for online dictionary

- data from the literature
  - published / unpublished wordlists
  - other studies about the language
- data gathered during fieldwork
  - words from recorded texts
  - translate wordlists
  - describe pictures or films

[www.fataluku.com](http://www.fataluku.com)

### During fieldwork

- collect words and translations
- check existing wordlists
- create a standard spelling
- collect grammatical paradigms
- collect examples (phrases and/or full sentences)

[www.fataluku.com](http://www.fataluku.com)

### Data: words

- standard spelling and spelling variants
- grammatical variants (inflection)
- phonological information (accent, tones)
- morphological information (plurals etc.)
- grammatical category (part of speech)
- semantic category
- translations
- source

[www.fataluku.com](http://www.fataluku.com)

### Data: examples

- standard spelling
- word forms used in the example that are different from citation forms
- translations
- source

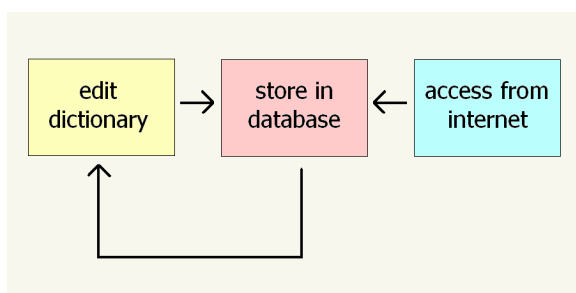
www.fataluku.com

### The dictionary file

- Eventually the data will be stored in a database that can be accessed from the internet
- During fieldwork the data is stored in a text file, which is easy to edit without special software and offers maximal flexibility (contrary to XML)

www.fataluku.com

### Work flow



www.fataluku.com

### Contents of dictionary file

- Each line contains either:
  - the lexeme separator '###'
  - a blank line (to improve readability)
  - a label code followed by one or more arguments. The label codes are project-specific but codes can be added at any time if necessary.

www.fataluku.com

### Fataluku label codes

- **fat**: Fataluku word
  - standard spelling
- **acc**: accent
  - irregular accent only, not in standard spelling
  - **nucece** vs. **núcece**
- **plu**: plural
  - irregular plurals only

www.fataluku.com

### Fataluku label codes

- **mut**: consonant mutation
  - **tahine** vs. **cahine**
- **var**: variant spellings
  - /i/ phoneme: **taia** vs. **taja**, **taya**, **taza**
  - long vowels: **le** vs. **lee**
  - final vowel: **iparu** vs. **ipar**
  - glottal stop: **fa'i** vs. **fai**
  - compounds: **pura fa'i** vs. **purafa'i**

www.fataluku.com

### Fataluku label codes

- **mor**: morphemes (in compounds)
  - **lafai pa'i** contains morphemes **lafai** and **fa'i**
- **cat**: category (part of speech)
  - noun, verb, function word, etc.
- **sem**: semantic category
  - animal, plant, kinship, etc.

www.fataluku.com

### Fataluku label codes

- **src**: source
  - existing wordlists, our own project, or both
- **eng**: English translation(s)
- **mal**: Malay translation(s)
- **tet**: Tetun translation(s)
- **por**: Portuguese translation(s)

www.fataluku.com

###

fat: tapule  
 mut: capule  
 var: tapul, capul  
 cat: verb  
 src: CNF  
 eng: buy  
 tet: sosa  
 mal: membeli  
 por: comprar

###

www.fataluku.com

fat: calu  
 acc: cálu  
 var: cal  
 plu: calafu, calafuru  
 cat: noun  
 sem: kinship  
 src: CNF  
 eng: grandparent, grandfather,  
 grandmother  
 tet: bei-kalu, aman-tuak, inan-bei  
 mal: moyang, kakek, nenek  
 por: avô

www.fataluku.com

### The database

- The dictionary is stored in a relational database to make it accessible from the internet
- A separate table is needed for each one-to-many relationship:
  - one Fataluku word may have several spelling variants
  - one Fataluku word may have several English translations (etc.)

www.fataluku.com

### Tables

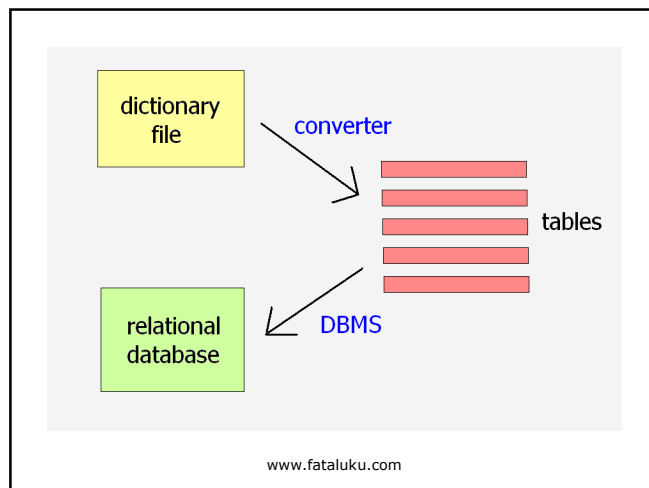
- Tables have a primary key to identify each row; this key must be unique, so it cannot be a Fataluku word because of homonyms
  - **vari** 'nest' vs. **vari** 'to hear'
- All tables except **fataluku** have a foreign key, which refers to a row in the **fataluku** table

www.fataluku.com

### Tables

- fataluku id, word, accent, category, semantic, source
- mutation id, form, fataluku
- plural id, form, fataluku
- variant id, form, fataluku
- morpheme id, form, fataluku
- english id, translation, fataluku
- tetun id, translation, fataluku
- malay id, translation, fataluku
- portuguese id, translation, fataluku

www.fataluku.com



### Converter

- Java program that splits the dictionary file into several files, each containing a table
- It assigns primary keys and foreign keys automatically
- The tables can then be loaded into the database by a standard procedure

www.fataluku.com

<pre>### fat: nana cat: noun sem: animal src: CNF eng: python, snake tet: labak mal: ular sawa por: cobra, serpente ###</pre>	fataluku
	312, nana, , noun, animal, CNF
	english
	341, python, 312
	342, snake, 312
	tetun
	275, labak, 312
malay	
277, ular sawa, 312	
portuguese	
229, cobra, 312	
230, serpente, 312	

www.fataluku.com

### The user interface

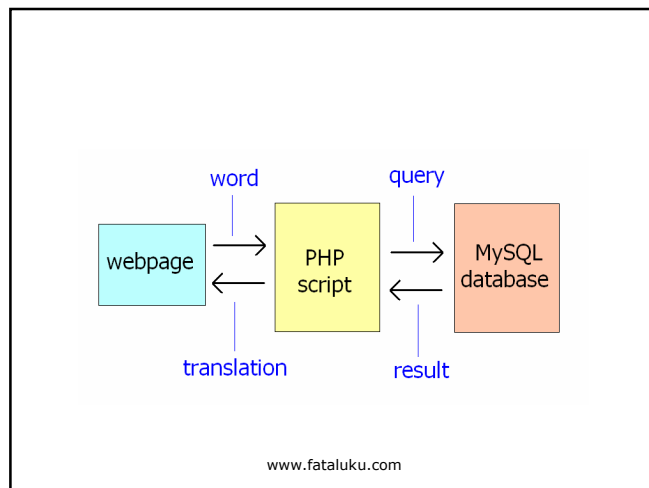
- A user accesses the dictionary through a web page

Fataluku online dictionary

Translate from/into:  Fataluku,  English,  Tetun,  Malay,  Portuguese

Include word groups containing this word     Show examples     Show sources

www.fataluku.com



### PHP script

- Translation from Fataluku: find the word in the **fataluku**, **mutation**, **plural**, or **variant** tables
- Translation into Fataluku: find the word in the **english**, **tetun**, **malay**, and / or **portuguese** tables
- Save the corresponding primary keys (**fataluku**) or foreign keys (other tables)

www.fataluku.com

### PHP script

- For each key in the list, print the Fataluku word and its translations
- Print also the word category, and, if present, mutation forms, plurals and / or accentuation patterns
- Do not print spelling variants

www.fataluku.com

### Example

- The user enters: **child**. This word is looked up in all tables. It is found twice in the **english** table, but not in other tables:
  - 360, **child**, 327
  - 564, **child**, 487
- The foreign keys are saved in a list. Each number in this list is then compared to the **id** field of the **fataluku** table and the **fataluku** field of other tables.

www.fataluku.com

### Example

- The number **327** is found in these tables:
  - fataluku: 327, **kinamoko**, , **noun**, , **CNF**
  - plural: 3, **moko-mokoru**, 327
  - english: 359, **boy**, 327  
360, **child**, 327
  - tetun: 289, **labarik**, 327  
290, **klosan**, 327
  - malay: 292, **kanak-kanak**, 327  
293, **pemuda**, 327

www.fataluku.com

### Examples

- The result is then printed as follows:
  - the Fataluku word and its category
  - accent / mutation / plural forms (if present)
  - translations in each language
- The same is done for the other numbers in the list

www.fataluku.com

#### Fataluku online dictionary

Translate from/into:  Fataluku,  English,  Tetun,  Malay,  Portuguese

Include word groups containing this word  Show examples  Show sources

(Fat.) **kinamoko** (noun) (plural: **moko-mokoru**)  
 (Eng.) boy, child  
 (Tet.) labarik, klosan  
 (Mal.) kanak-kanak, pemuda  
 (Por.) rapaz, menino, maroto

(Fat.) **moco** (noun)  
 (Eng.) child, son, daughter  
 (Tet.) oan  
 (Mal.) anak  
 (Por.) filho, enteado

www.fataluku.com

### Option: Compounds

- Standard search → exact match:  
**aia** 'rain' does not find:  
**aia pari** (rain wind) 'storm'
- Include compounds → also finds words containing this morpheme:  
**aia** also finds: **aia pari**  
but does not find: **taia** 'sleep'

[www.fataluku.com](http://www.fataluku.com)

### Other options

- Show examples: prints all examples containing this word
- Show sources: especially important for seldomly used words or in case of conflicting information

[www.fataluku.com](http://www.fataluku.com)

### Future extensions

- Autocomplete option for words
- Search in specified semantic categories only, so user can discover new words
- User can submit new words
- Advances search options for linguists (e.g. find irregular plurals)
- Include other endangered languages of East Timor

[www.fataluku.com](http://www.fataluku.com)