

Introduction to Shoebox/Toolbox

Stuart Robinson, Max Planck Institute
stuart.robinson@mpi.nl or stuart@zapata.org

October 11, 2005

Contents

1	Shoebox/Toolbox	2
1.1	Pros and Cons	2
1.1.1	Pros	2
1.1.2	Cons	2
1.2	Dictionaries	2
1.2.1	Organizing Lexical Entries in Shoebox	2
1.2.2	Sample Dictionary	3
2	On-line Resources	5
3	Automatically Processing Shoebox Data	5

1 Shoebox/Toolbox

According to the official Shoebox web site:

“Shoebox is a computer program that helps field linguists and anthropologists integrate various kinds of text data: lexical, cultural, grammatical, etc. It has flexible options for sorting, selecting, and displaying data. It is especially useful for helping researchers build a dictionary as they use it to analyze and interlinearize text. The name Shoebox recalls the use of shoe boxes to hold note cards on which definitions of words were written in the days before researchers could use computers in the field.”

1.1 Pros and Cons

1.1.1 Pros

1. Good integration of lexicon and text interlinearizing
2. User-friendly GUI
3. Widely used and fairly feature-rich
4. Good documentation

1.1.2 Cons

Basically, the main problem is that the data is neither XML nor a relational database—i.e.,

1. No recursion
2. Limited analytical capabilities (no standardized query language)
3. Few mechanisms for constraint enforcement (exception: range sets)
4. Proprietary

1.2 Dictionaries

1.2.1 Organizing Lexical Entries in Shoebox

The following components represent the basic core of a lexical entry:

Lexeme the head word or lexeme for the entry

Part of Speech the grammatical class of the lexeme

Gloss one-word translation for use in interlinear texts

Definition translation or definition as it would appear in a printed dictionary

Notes any notes concerning the entry

Root derivational source of the entry

Timestamp when the entry was last edited (automatically updated by Shoebox)

Sample entry with all of these fields:

```
(1) \lx aaova
    \rt aao
    \ps N.F
    \ge granddaughter
    \gp tumbuna
    \nt SoDa
    \sf KIN
    \dt 25/Aug/2004
    \ex Aaova eira iria uriopaoi.
    \xp Bubu meri em i kam.
    \xe Granddaughter came.
```

1.2.2 Sample Dictionary

A Shoebox dictionary of Rotokas (Papuan, Bougainville) is currently under development by the author. Here's a sample entry:

```
(2) \lx tasiasi
    \ps V.B
    \ge stomp
    \eng stomp on
    \eng step on repeatedly
    \gp krungutim
    \vx 2
    \arg OBL
    \cm -ia
    \dt 25/Aug/2005
    \ex Guruvara-ia tasiasiparevoi.
    \xp Em i wok long krungutim ol lip.
    \xe He is stomping on the leaves.
    \ex Kuvukuvu tou-ia tasiasipai ovusia gurukopai.
    \xp Ol i wok long krungutim giraun na giraun i pairap.
    \xe They are stomping on the ground while it shakes.
```

Derived lexemes are related to their sources by the field `\rt` field—e.g., *kasipu* ‘to be angry’ \Rightarrow *kasipupie* (*kasipu* + *pie*) ‘to make angry’:

```

(3) \lx kasipupie
    \rt kasipu
    \ps V.B
    \ge anger
    \eng anger
    \eng enrage
    \gp mekim kros
    \vx 2
    \arg 0
    \dt 25/Aug/2005
    \ex Teapi rera kasipupieive orareoreopaoro.
    \xp ???
    \xe Don't make him angry talking among yourselves.
    \ex Ae, visii ragai kasipupietavoi.
    \xp Hei, yupela i mekim mi kros.
    \xe Hey, you guys are making me angry.

```

This creates derivational chains—e.g., *tarai* ‘understand’ \Rightarrow *taraipie* ‘teach’ \Rightarrow *orataraipie* ‘learn (literally: self-teach)’.

Lexemes are not unique, as can be seen from the following two entries for *keke*, distinguished by the part-of-speech field (`\ps`):

```

(4) \lx keke
    \ps V.B
    \ge look.at
    \eng look at
    \gp lukim
    \vx 2
    \arg 0
    \cmt Example needed
    \sc PERCEPTION
    \dcp TRUE
    \dt 02/Aug/2005

    \lx keke
    \ps V.A
    \ge look
    \gp luk
    \dcs True
    \cm -ia
    \vx ???
    \dt 17/Jul/2005
    \ex Vearo kekepau ragai osireiaro-ia vii oupa ovusia ragai taviri.
    \xp Yu luk gut tru long ai bilong mi bai mi maritim yu sapos yu tok orait.
    \xe You look good in my eyes I will marry you if you tell me.

```

No mechanism exists in Shoebox for enforcing uniqueness constraints, as far as I am aware.

2 On-line Resources

To learn more:

- www.sil.org/computing/shoebox/
- www.sil.org/computing/toolbox/

A list of Shoebox-related links is maintained by yours truly here:
www.zapata.org/stuart/shoebox

3 Automatically Processing Shoebox Data

It is also possible to process Shoebox data automatically. For those who know the programming language Python, there are tools under development for the Natural Language Toolkit (Loper and Bird, 2002, 2004).

- www.python.org
- nltk.sourceforge.net

References

Steven Bird and Gary Simons. Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582, 2003.

Edward Loper and Steven Bird. NLTK: The natural language toolkit. 2002. URL <http://arxiv.org/abs/cs/0205028>.

Edward Loper and Steven Bird. NLTK: The natural language toolkit. 2004. URL <http://www ldc.upenn.edu/sb/home/papers/nltk.pdf>.