31st Colloquium on African Languages and Linguistics, University of Leiden
Hannes Hirzel, University of Zürich, email: hirzel@spw.unizh.ch
Using Shoebox 5.0 for analysing African languages.

How to optimize analysing an African language text corpus by exploiting old and new features of the Shoebox 5.0 interlinearization program:

A demonstration from Akan and Swahili

This presentation focuses on some instrumental aspects of research and teaching methods: It shares experiences we made in Zürich in different projects using the Shoebox computer program.

In particular some information will be given about the new features of version 5, which came out last year.

These notes will be available on the web under www.spw.unizh.ch/tools/shoebox. A more elaborate Akan example will be put there as well.

1) Overview

The Shoebox program supports all aspects of producing the classical trilogy of the descriptive linguist: Text collection, lexicon and grammar.

Fig. 1 shows the setup of data and the most prominent functions Shoebox provides:

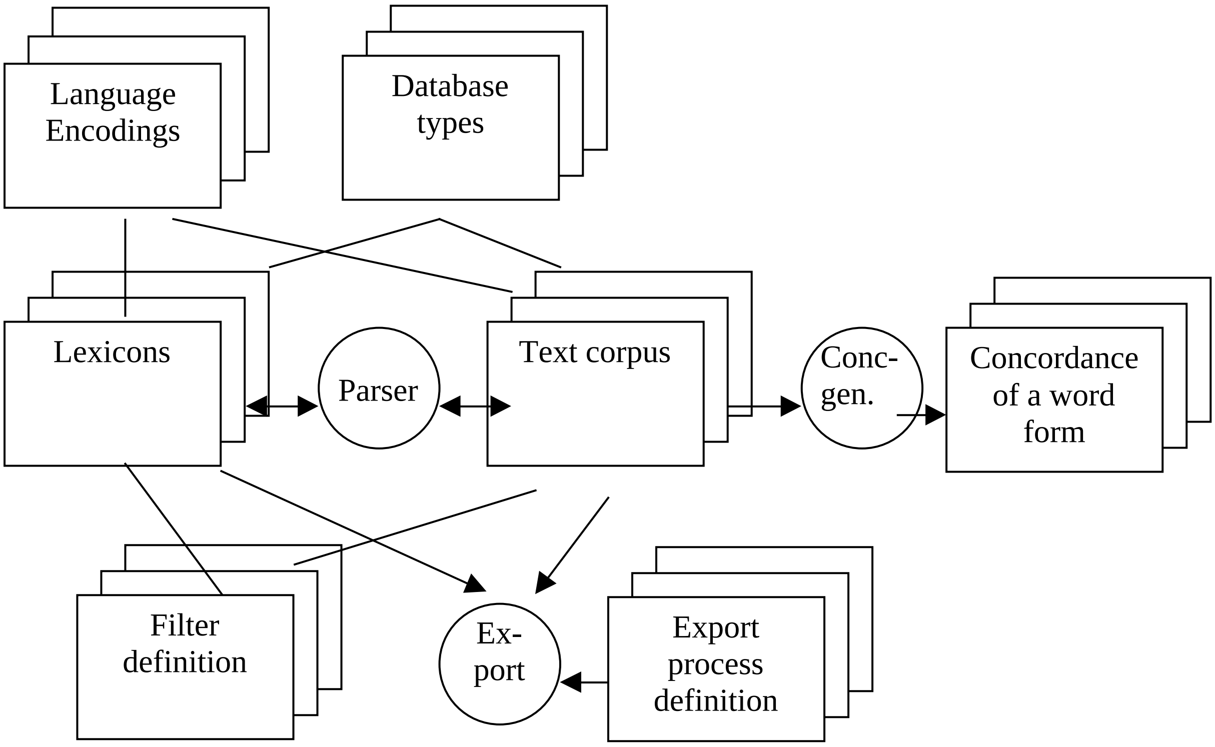
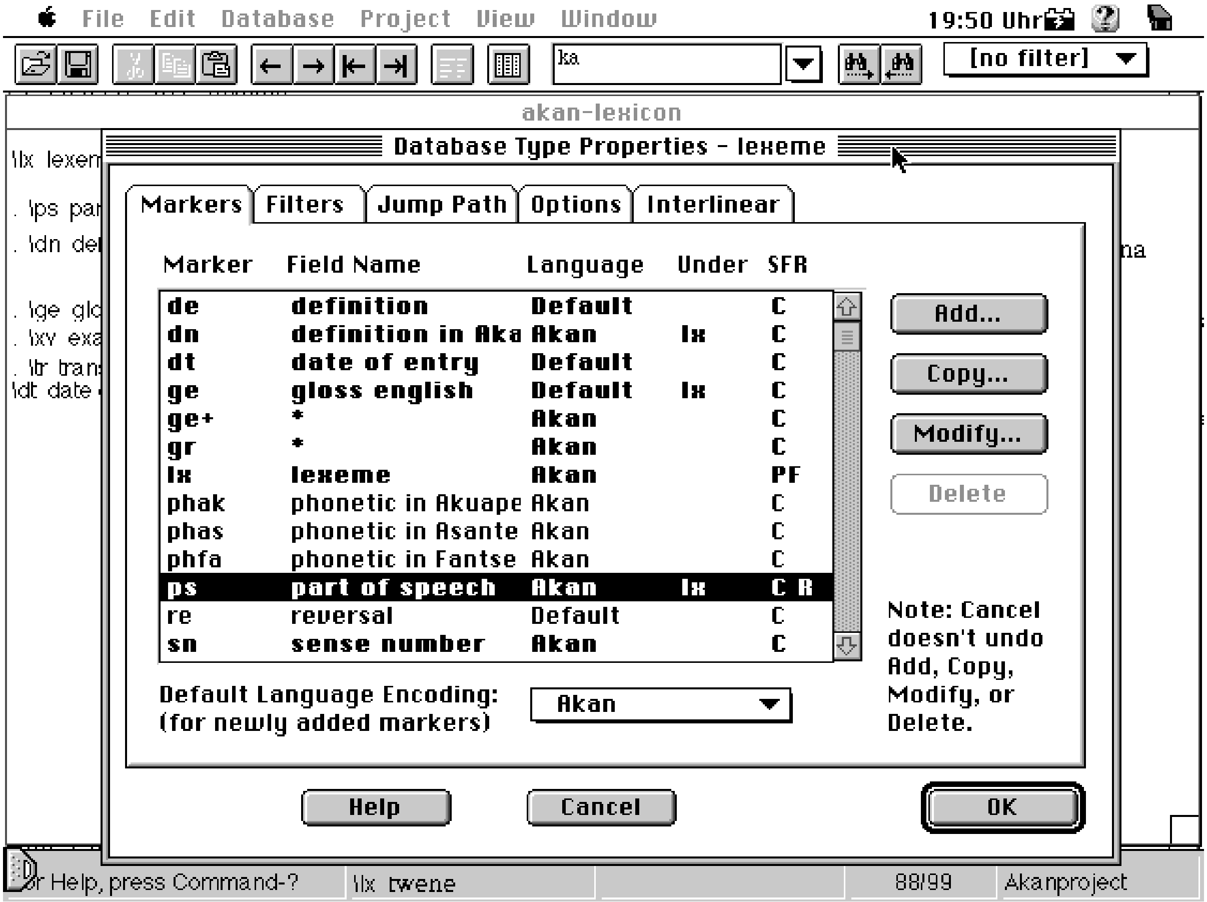


Fig 1: Data setup and main functions: Morphological analysis (parsing), concordance generation and export of data to other programs.

Remarks:

- You need not to be aware of these things fully to use of Shoebox successfully in your research.
- Shoebox is not a database system, nor an information retrieval system, nor a text corpus database in the strict sense. But it combines all these aspects in a particular way, which eases the work of the linguist who does descriptive and analytical work.

Object editing:
To understand the use of Shoebox more easily it is necessary to keep in mind, that for setting up Shoebox a lot of objects have to be defined. They are all accessible through dialog boxes like the one in Fig. 2 (Macintosh screenshot). A list of objects – here database fields - is shown together with the buttons 'add...', 'copy...', 'modify...' and 'delete'.



Marker	Field Name	Language	Under	SFR
de	definition	Default		C
dn	definition in Aka	Akan	lx	C
dt	date of entry	Default		C
ge	gloss english	Default	lx	C
ge+	*	Akan		C
gr	*	Akan		C
lx	lexeme	Akan		PF
phak	phonetic in Akuape	Akan		C
phas	phonetic in Asante	Akan		C
phfa	phonetic in Fantse	Akan		C
ps	part of speech	Akan	lx	C R
re	reversal	Default		C
sn	sense number	Akan		C

Default Language Encoding: (for newly added markers) Akan

Note: Cancel doesn't undo Add, Copy, Modify, or Delete.

Fig 2: Dialog box for editing data base field definitions

Principle for interlinearizing:
Morphological analysis strategies implemented in the Shoebox parsing mechanism:

- Work from the outside of a word form to the inside (root).
- Prefer longest match.
- Consider as well word forms, which the user put in the '\a' (allomorph) field.
- No priority of prefixes or suffixes is respected.
- Present a dialog box for the user to choose if there are ambiguities.

2) Application scenarios

ALI Akan

(ALI = African Linguistics, Akan is spoken in Ghana; Twi is a dialect of it.)

In Zürich a language course for learning Akan with special emphasis on African linguistics was developed. Pilot courses have been held in cooperation with Berlin and Leiden.

The course uses hypertext with sound files and pictures on a CDROM.

For producing the lexicon of the course Shoebox was used: Every entry in the dictionary contains a reference to a sound file. Each entry is marked with the lesson it belongs to. Export functions generate vocabulary pages in HTML and RTF format.

Demonstration: Entry structure and html export.

ALI Akan II in Leiden, May 2001 used Shoebox for doing textual research.

Demonstration: Concordance – uses of sñ

Demonstration : Nyansa text.

- Discontinous constitutents
- Tone

Akan Dictionary Project

Minimal entry structure

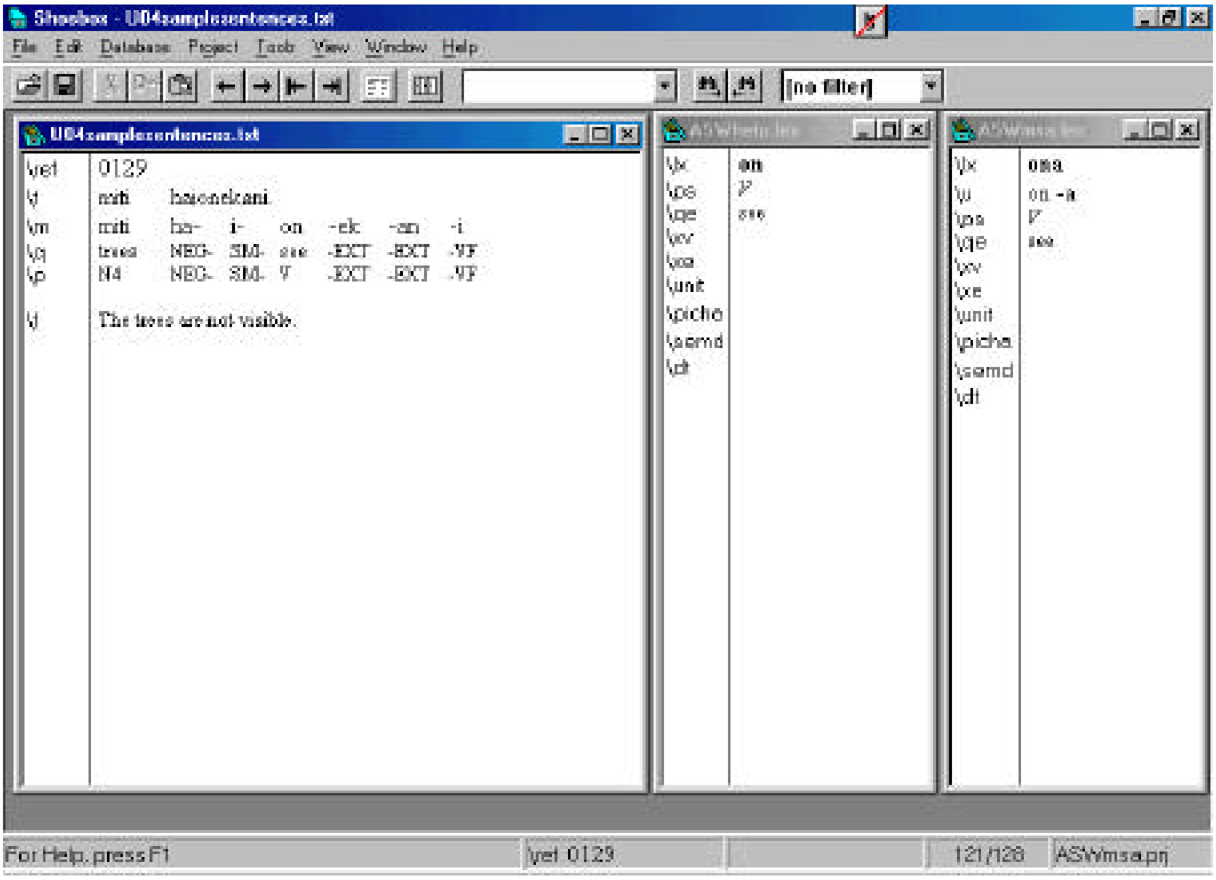
- Lexis with tone markings
- Transcription in 3 dialects (Akuapim, Asante-Twi, Fante)
- Part of speech
- Definition in Ghanaian language
- English definition
- Example in Akan
- English translation
- Semantic domain

First intended use: Generate a domain specific glossary for non-formal education (neologisms).

ALI Swahili

A language learning program for Swahili beeing developed at the University of Zurich.

Shoebox is used for interlinearization:
Most example sentences will be provided with an analysis. The Shoebox interlinear text is exported and postprocessed into html files by a custom made conversion program. Fig. 3 shows the setup. On the left side an interlinear example sentence (in the file 'U04samplesentences.txt') is shown. In the middle a database which aids parsing ('ASWhelp.lex') and on the right hand side the learners dictionary (ASWmsa.lex).



```
\ref   0129
\t     miti     haionekani
\m     miti     ha-    i-    on    -ek    -an    -i
\g     trees    NEG-   SM-   see   -EXT   -EXT   -VF
\p     N4       NEG-   SM-   V     -EXT   -EXT   -VF

\f     The trees are not visible.
```

```
\lx      on
\ps      V
\ge      see
\ev
\ea
\unit
\picho
\semd
\dt
```

```
\lx      ona
\u       on -a
\ps      V
\ge      see
\ev
\xe
\unit
\picha
\semd
\dt
```

Fig. 3 ALI Swahili Shoebox setup.

Interesting features

- Script Definition Facility: Sequence of ASCII character keystrokes is stored in the files but on the screen the script is rendered with a full script. The equivalences are defined in a file. Useful for data exchange over the internet and long term archiving. Can be used for defining right to left scripts as well.

- French program version available, but no french documentation.

- Simple XML export facility.
- Sort order – collation sequence: Fully configurable.
- WORD export for interlinear texts: Works, but not very well.
- Runs on Mac and Windows the same way with the same data.

Cooperation: Data exchange

What is needed for working together with colleagues?

- The same version of Shoebox has to be installed (e.g. 4.0 or 5.0 which came out in 2000)
- Special TrueType font (if necessary in the Windows or Macintosh font folder).
- All the files set up in the same working directory.

Procedure

- Pack the working directory and if necessary the font into a 'zip' file (or another kind of packed file). The zip-file is normally relatively small.
- Send the zip file by email as an attachment to the receiver.
- The receiver has to unpack the zip file, install the font if necessary and to double click on the blue Shoebox project icon and then he has the same setup the sender had.
- Start up Shoebox as well, phone the colleague and you can discuss open issues by walking through the texts and lexicons.
- This approach works between Africa and Europe. Many African cities now have Internet cafes to send and receive emails, Universities often have email too. Phoning is still expensive but possible in some cases.
- A very comprehensive set of Shoebox files fits on a diskette: Air mail is an alternative.
- Make sure you use an recent AntiVirus program!

Evaluation

Taking the risk of stateting the obvious: Shoebox is not a translation tool, but a glossing aid for word by word glossing. Morphological analysis is supported but no syntax analysis;Shoebox helps to do book-keeping and to be consistent. It is a good dictionary making tool.

Experience:

- Shoebox interlinearizing works well with Swahili, a language with a complex morphology without tone.
- With some precautions it's feasible to be used with Akan as well, a language with more isolate constituents and tone.

Parsing mechanism:

- Shoebox presents an automatic matching mechanism which can be put to use. But it sometimes means that you have to make compromises to avoid the mechanism to fail. Some phenomena are not easy to deal with: One has to find workarounds..

Concordance

- An often overlooked tool; relatively easy to set up; very useful for textual research.

Lexicography:

- Aid to produces state of the art files which can be easily processed by a publisher. Ensures consistency when assigning semantic domains and grammatical categories.

Interlinerarization:

- Analytical lexicon versus common lexicon: If you set up a lexicon for analytical use it is often not useful to use it as a standard lexicon. Workaround: Use two lexicons.

Word formulas

- Word formulas are a new feature in Shoebox 5.0. They help to do additional disambiguations. A promising things but we didn't use it yet.

Single user tool

- Shoebox is designed to be used on a single computer. Therefore if several people cooperate in a project the synchronisation has to be done by organizational means. In an African context you probably have one or two persons who key in the data on one computer other researchers wrote on index cards. This approach works fine and has been used to do very comprehensive dictionaries.

Conclusion:
Shoebox is not a general all encompassing tool but a tool which fills a particular need in the process of researching and teaching. Some parts of it are easy to use and you can work for years that way producing valuable results. Other things are still difficult to put into use. But existing set-ups can be copied and used as templates.

AND: Do work through the Computer Based Training which comes with the releases CDROM. It's worth the effort: It takes about 1 to 2 days full time to do the lessons. Besides the user manual there are a lot of application notes on the release CDROM.

Glossary, Abbreviations and additional remarks.

Allomorph In the Shoebox sense an alternate string of characters (i.e. for example a whole word form) to be matched instead of the keyword _field. Conventionally marked with "\a". On may assign an _underlying form.

CCT Consistent Changes Table: A simple computer language used within Shoebox to make consistent changes; used in _import and _export processes.

Concordance Here a dynamically generated list of a keyword or part of it in context. May be saved in a separate file for later reference.

Database A collection of _records.

Database type The description of the structure, the content and the allowed values (for example parts of speech or semantic domain) of a database.

Export Conversion of a Shoebox file to a file format to be processed by another program, for example to _ RTF.

Field A combination of a _tag with a _text element.

Filter Definition of a selection of a part of a _database, for example all nouns of a certain class, all entries having a back vowel after an initial obstruent or all entries belonging to the semantic domain 'cooking'. Applying a filter in Shoebox gives as well the number of occurrences in the result set. Useful for lexicostatistic work.

Font A set of graphemes. Limited to at most 256 in the same font in Shoebox. Combinations of base characters and diacritics is possible thus giving more possibilities.

HTML Hypertext Markup Language: The files displayed in web browsers are mostly in the HTML format. Similar approach as _SFM files but less content oriented.

Import Conversion of a _'text-only' file into _ SFM, i.e. marking the text with tags.

Parser Mechanism in Shoebox which defines together with the lexicon entries (_lexeme, _allomorph and _underlying form fields) how the word forms are analysed into morphemes.

SFM Standard format marker: A file format which has been used by SIL internally in the last twenty years. Text elements have tags in front of them which indicate the nature of the text element. The tags are introduced by a backslash ("\") character and followed by a space. The format can be quite easily converted to _XML. Hierarchies cannot be directly expressed. Shoebox solves this problem by expressing the hierarchy in the _ Database type; but this doesn't show up in the XML export. For many cases this limited XML export is OK.

Record An entry which is displayed on one screen. A record consists of 'fields'. In a text corpus a record may encompass just a sentence or a paragraph. In a dictionary a lexeme and the according explanations.

RTF Rich-Text-Format: A file format developed by Microsoft to render a text with font, size and emphasis attributes. Most text processing system can read RTF.

Sort order Possibility in Shoebox to define a language specific alphabetic sort order (collation sequence).

True Type A file format used for describing _fonts used by Apple and Microsoft.

TTAdopter TrueType Adopter program: A program which allows to convert Windows _True Type fonts to a form usable on the Macintosh.

Tag In linguistic data processing every text element has to be marked as the computer can only relate similar symbols. Tagging has to be done mostly manually.

Text element For example a lexeme, a gloss, a part of speech, an example sentence or a note. (_field).

Text-only file A file stored on the computer in a format which can be opened by all text processing programs. The preferred format to archive language data. The conventions used in representing data has been changing at an incredible rate the last twenty years. Text-only files are still readable and will be in the future. _SFM and _XML files are 'text-only' files. Text-only files which contain tags need a viewer program and an accompanying interpretation description to display. properly. In Shoebox this information is in the _Database type file: Style information (e.g. font, size, emphasis) is rendered by the viewer program.

Underlying form In an underlying form a predefined morphological analysis is given. Shoebox is forced to do a look up of each of its constituents. See also 'allomorph'.

Word formulas Shoebox 5.0 term for describing morphological selection restrictions, for example suffix –*s* in English after a noun means plural, after a verb it means 3SG.

View On can view the same data in different ways at the same time. (Record view, list view, view without and with _filters)

XML Extended markup language. Similar to _ SFM but allows hierarchical data structures. The most recent web browsers can display this format. Shoebox can produce XML files. Use not yet widespread.

References

Akan Dictionary, University of Legon and University of Zurich, www.spw.unizh.ch/afrling/akandic

ALI Akan, Akan language learning program used in the Socrates/Erasmus program – www.spw.unizh.ch/afrling/aliakan

ALI Swahili, Swahili language learning program, University of Zurich – www.spw.unizh.ch/afrling/aliswahili - the project description and the first Unit

Bird, Stephen, Linguistic data consortium, University of Pennsylvania; Web pages: Linguistic annotation and exploration tools – www.ldc.upenn.edu/annotation and www.ldc.upenn.edu/exploration

Coward, David, Making Dictionaries, SIL; PDF-file on Shoebox CD – well written recommendations how to do practical dictionary making work.

Hayashi, Larry ,Discovering and Testing Linguistic Generalisations Using Interactive Concordances, SIL, Dallas – www.ldc.upenn/exploration/LSA/hayashi.html

Severn, John, Maluku, SIL Indonesia - Shoebox - A Phonological Tool? - http://www.sil.org/computing/noc/vol14/142shoeb.htm

Shoebox homepage: www.sil.org/computing/shoebox

Tavultesoft keyboard manager, www.tavultesoft.com - a utility program do define language specific keyboards.

Vydrine, V. (1999). Manding-English Dictionary. Vol. 1. St. Petersburg: Dimitri Bulanin Publishing House, 316 p. (Back cover gives an example of a complex entry structure; the dictionary was made with Shoebox).