E16 & DEtool
typesetting language data using ConTEXt

Abstract
This article describes two recent projects in which ConTEXt was used to typeset language data. The goal of project E16 was to typeset the 16th edition of the Ethnologue, an encyclopaedia of the languages of the world. The complexity of the data and the size of the project made this an interesting test case for the use of TEX and ConTEXt. The Dictionary Express tool (DEtool) is developed to typeset linguistic data in a dictionary layout. DEtool (which is part of a suite of linguistic software) uses ConTEXt for the actual typesetting.

Introduction

Some background: SIL is an NGO dedicated to serve the world's minority language communities in a variety of language-related ways. Collecting all sorts of language data is the basis of much of the work. This could be things like the number of speakers of a particular language, relations between different languages, literacy rates and bi- and multilingualism. Much of this data ends up in a huge database, which in turn is used as the source for publications like the Ethnologue.[1] which is an encyclopaedia of languages. It consists of four parts, starting with an introductory chapter explaining the scope of the publication and 25 pages of 'Statistical summaries'. Part 1 has 600 pages with language descriptions, describing all the 6909 languages of the world. Part 2 consists of 200 pages with language maps and Part 3 has of 400 pages of indexes, for Language names, Language Codes and Country names.

Typesetting the Ethnologue

Data flow and directory structure: All the data is stored in an Oracle database running on a secure web server. The XML output is manipulated using XSLT to serve different 'views'. One output path leads to html (for the website http://www.ethnologue.com) and another output path gives TEX-output of with the codes are defined in ConTEXt. Once the data is downloaded from the server, it is stored locally in the 'data' directory of the typesetting system. There is also a 'content' directory containing small files that \input the data files (and do some tricky things with catcodes.) All the content-files are loaded using a 'project' file in the root directory. This (slightly complicated) process allows for easy updating of the data and convenient testing of all the different parts, both separately and together. The macro definitions are all stored in a module.

Module

In good ConTEXt style all the code for this project is placed in a module. A ConTEXt module starts with a header like this:

```
%D \module
%D   [       file=p-ethnologue,
%D        version=2009.01.14
%D          title=\CONTEXT\ User Module,
%D       subtitle=Typesetting Ethnologue 16,
%D         author=Jelle Huisman, SIL International,
%D           date=\currentdate,
%D      copyright=SIL International]
%C Copyright SIL International
```

```
\writestatus{loading}{Context User Module  Typesetting Ethnologue 16}
\unprotect
\startmodule[ethnologue]

All the macro definitions go here... and the module is closed with:

\stopmodule
\protect \endinput
```

With the command `texexec --modu p-ethnologue.tex` it is easy to make a pdf with the module code, comments and even an index.

E16 code examples

A couple of code examples are presented here to give an impression of the project. This is part of the standard page setup for the paper size and the setup of two basic layouts.

```
\definepapersize [ethnologue][width=179mm, height=255mm]

\startmode[book] % basic page layout for the book
\setuppapersize [ethnologue][letter]% paper size for book mode
\setuplayout[backspace=18mm, width=148mm, topspace=7mm, top=0mm,
             header=6mm, footer=7mm, height=232mm]
\stopmode

\startmode[proofreading] % special layout for proofreading mode
\setuppapersize [letter][letter]% paper size for proofreading mode
\setuplayout[backspace=18mm, width=160mm, topspace=7mm, top=0mm,
             header=16mm, footer=6mm, height=250mm]
\stopmode
```

Use of modes: proofreading vs. final output

To facilitate the proofreading a special proofreading 'mode' was defined with wider margins, as shown in the code example in the previous section and with a single column layout (not in this code example). The 'modes' mechanism is used to switch between different setups. This code:

```
%\enablemode[book]
\enablemode[proofreading]
```

is used in a 'project setup' file to switch between the proofreading mode (single column, bigger type) and the book mode showing the layout of the final publication. One other application of modes is the possible publication of separate extracts with e.g. the language descriptions of only one country. This could be published using a Printing on Demand process.

Language description

The biggest part of the publication is the section with the language descriptions. Each language description consists of: a page reference (not printed), the language name, the language code, a short language description and a couple of special 'items' like: language class, dialects, use and writing system. This is an example of the raw data for Belarusian:

```
\startLaDes{ % start of Language Description
\pagereference[bel-BY] % used for index
\startLN{Belarusan }\stopLN % LN: Language name
[bel] % ISO 639-3 code for this language
(Belarusian, Belorussian, Bielorussian, Byelorussian, White Russian,
White Ruthenian). 6,720,000 in Belarus (Johnstone and Mandryk 2001).
Population total all countries: 8,620,000.  Ethnic population:
9,051,080. Also in Azerbaijan, Canada, Estonia, Kazakhstan,
Kyrgyzstan, Latvia, Lithuania, Moldova, Poland, Russian Federation
```

Sine, Dyegueme (Gyegem), Niominka. The Niominka and Serere-Sine dialects mutually inherently intelligible. *Lg Use:* Official language. National language. *Lg Dev:* Literacy rate in L1: Below 1%. Bible: 2008. *Writing:* Arabic script. Latin script. *Other:* 'Sereer' is their name for themselves. Traditional religion, Muslim, Christian. *Map:* 725:28.

Soninke [snk] (Marka, Maraka, Sarahole, Sarakole, Sarangkolle, Sarawule, Serahule, Serahuli, Silabe, Toubakai, Walpre). 250,000 in Senegal (2007 LeClerc). North and south of Bakel along Senegal River. Bakel, Ouaoundé, Moudéri, and Yaféra are principal towns. *Dialects:* Azer (Adjer, Aser), Gadyaga. *Lg Use:* Official language. National language. Also use French, Bambara [bam], or Fula [fub]. *Lg Dev:* Literacy rate in L1: Below 1%. *Other:* The Soninke trace their origins back to the Eastern dialect area of Mali (Kinbakka), whereas the northeastern group in Senegal is part of the Western group of Mali (Xenqenna). Thus, significant differences exist between the dialects of the 2 geographical groups of Soninke in Senegal. Muslim. See main entry under Mali. *Map:* 725:29.

Wamey [cou] (Conhague, Coniagui, Koniagui, Konyagi, Wamei). 18,400 in Senegal (2007), decreasing. Population total all countries: 23,670. Southeast and central along Guinea border, pockets, usually beside Pulaar [fuc]. Also in Guinea. *Class:* Niger-Congo, Atlantic-Congo, Atlantic, Northern, Eastern Senegal-Guinea, Tenda. *Lg Use:* Neutral attitude. Also use Pulaar [fuc]. *Lg Dev:* Literacy rate in L1: Below 1%. *Writing:* Latin script. *Other:* Konyagi is the ethnic name. Agriculturalists; making wine, beer; weaving bamboo mats. Traditional religion, Christian. *Map:* 725:30.

Wolof [wol] (Ouolof, Volof, Walaf, Waro-Waro, Yallof). 3,930,000 in Senegal (2006). Population total all countries: 3,976,500. West and central, Senegal River left bank to Cape Vert. Also in France, Gambia, Guinea-Bissau, Mali, Mauritania. *Class:* Niger-Congo, Atlantic-Congo, Atlantic, Northern, Senegambian, Fula-Wolof, Wolof. *Dialects:* Baol, Cayor, Dyolof (Djolof, Jolof), Lebou (Lebu), Jander. Different from Wolof of Gambia [wof]. *Lg Use:* Official language. National language. Language of wider communication. Main African language of Senegal. Predominantly urban. Also use French or Arabic. *Lg Dev:* Literacy rate in L1: 10%. Literacy rate in L2: 30%. Radio programs. Dictionary. Grammar. NT: 1988. *Writing:* Arabic script, Ajami style. Latin script. *Other:* 'Wolof' is their name for themselves. Muslim. *Map:* 725:32.

Xasonga [kao] (Kasonke, Kasso, Kasson, Kassonke, Khasonke, Xaasonga, Xaasongaxango, Xasonke). 9,010 in Senegal (2006). *Lg Dev:* Literacy rate in L1: Below 1%. *Other:* Muslim. See main entry under Mali (Xaasongaxango).

Seychelles

Republic of Seychelles. 86,000. National or official languages: English, French, Seselwa Creole French. Includes Aldabra, Farquhar, Des Roches; 92 islands. Literacy rate: 62%–80%. Information mainly from D. Bickerton 1988; J. Holm 1989. Blind population: 150 (1982 WCE). The number of individual languages listed for Seychelles is 3. Of those, all are living languages.

English [eng]. 1,600 in Seychelles (1971 census). *Lg Use:* Official language. *Other:* Principal language of the schools. See main entry under United Kingdom.

French [fra]. 980 in Seychelles (1971 census). *Lg Use:* Official language. *Other:* Spoken by French settler families, 'grands blancs'. See main entry under France.

Seselwa Creole French [crs] (Creole, Ilois, Kreol, Seychelles Creole French, Seychellois Creole). 72,700 (1998). Ethnic population: 72,700. *Class:* Creole, French based. *Dialects:* Seychelles dialect reportedly used on Chagos Islands. Structural differences with Morisyen [mfe] are relatively minor. Low intelligibility with Réunion Creole [rcf]. *Lg Use:* Official language since 1977. All domains. Positive attitude. *Lg Dev:* Taught in primary schools. Radio programs. Dictionary. Grammar. NT: 2000. *Writing:* Latin script. *Other:* Fishermen. Christian.

Sierra Leone

Republic of Sierra Leone. 5,586,000. National or official language: English. Literacy rate: 15%. Immigrant languages: Greek (700), Yoruba (3,800). Also includes languages of Lebanon, India, Pakistan, Liberia. Information mainly from D. Dalby 1962; TISSL 1995. Blind population: 28,000 (1982 WCE). Deaf institutions: 5. The number of individual languages listed for Sierra Leone is 25. Of those, 24 are living languages and 1 is a second language without mother-tongue speakers. See map on page 726.

Bassa [bsq]. 5,730 in Sierra Leone (2006). Freetown. *Other:* Traditional religion. See main entry under Liberia.

Bom [bmf] (Bome, Bomo, Bum). 5,580 (2006), decreasing. Along Bome River. *Class:* Niger-Congo, Atlantic-Congo, Atlantic, Southern, Mel, Bullom-Kissi, Bullom, Northern. *Dialects:* Lexical similarity: 66%–69% with Sherbro [bun] dialects, 34% with Krim [krm]. *Lg Use:* Shifting to Mende [men]. *Other:* Traditional religion.

Bullom So [buy] (Bolom, Bulem, Bullin, Bullun, Mandenyi, Mandingi, Mmani, Northern Bullom). 8,350 in Sierra Leone (2006). Coast from Guinea border to Sierra Leone River. Also in Guinea. *Class:* Niger-Congo, Atlantic-Congo, Atlantic, Southern, Mel, Bullom-Kissi, Bullom, Northern. *Dialects:* Mmani, Kafu. Bom is closely related. Little intelligibility with Sherbro, none with Krim. *Lg Use:* Shifting to Themne [tem]. *Lg Dev:* Bible portions: 1816. *Writing:* Latin script. *Other:* The people are intermarried with the Temne and the Susu. Traditional religion. *Map:* 726:1.

English [eng]. *Lg Use:* Official language. Used in administration, law, education, commerce. See main entry under United Kingdom.

Gola [gol] (Gula). 8,000 in Sierra Leone (1989 TISLL). Along the border and inland. *Dialects:* De (Deng), Managobla (Gobla), Kongbaa, Kpo, Senje (Sene), Tee (Tege), Toldil (Toodii). *Lg Use:* Shifting to Mende [men]. *Other:* Different from Gola [mzm] of Nigeria (dialect of Mumuye) or Gola [pbp] (Badyara) of Guinea-Bissau and Guinea. Muslim, Christian. See main entry under Liberia. *Map:* 726:4.

Kisi, Southern [kss] (Gissi, Kisi, Kissien). 85,000 in Sierra Leone (1995). *Lg Dev:* Literacy rate in L2: 3%. *Other:* Different from Northern Kissi [kqs]. Traditional religion, Muslim, Christian. See main entry under Liberia. *Map:* 726:13.

Kissi, Northern [kqs] (Gizi, Kisi, Kisie, Kissien). 40,000 in Sierra Leone (1991 LBT). *Dialects:* Liaro, Kama, Teng, Tung. *Lg Use:* Also use Krio [kri] or Mende [men]. *Other:* Traditional religion. See main entry under Guinea. *Map:* 726:11.

Klao [klu] (Klaoh, Klau, Kroo, Kru). 9,620 in Sierra Leone (2006). Freetown. Originally from Liberia. *Other:* Traditional religion. See main entry under Liberia.

Kono [kno] (Konnoh). 205,000 (2006). Northeast. *Class:* Niger-Congo, Mande, Western, Central-Southwestern, Central, Manding-Jogo, Manding-Vai, Vai-Kono. *Dialects:* Northern Kono (Sando), Central Kono (Fiama, Gbane, Gbane Kando, Gbense, Gorama Kono, Kamara, Lei, Mafindo, Nimi Koro, Nimi Yama, Penguia, Soa, Tankoro,

Figure 1. Example of page with language descriptions

```
(Europe), Tajikistan, Turkmenistan, Ukraine, United States, Uzbekistan.
\startLDitem{Class: }\stopLDitem % LDitem: Language description item
Indo-European, Slavic, East.
\startLDitem{Dialects: }\stopLDitem Northeast Belarusan (Polots,
Viteb-Mogilev), Southwest Belarusan (Grodnen-Baranovich,
Slutsko-Mozyr, Slutska-Mazyrski), Central Belarusan. Linguistically
between Russian and Ukrainian [ukr], with transitional dialects to both.
\startLDitem{Lg Use: }\stopLDitem National language.
\startLDitem{Lg Dev: }\stopLDitem Fully developed. Bible: 1973.
\startLDitem{Writing: }\stopLDitem Cyrillic script.
\startLDitem{Other: }\stopLDitem Christian, Muslim (Tatar).  }
\stopLaDes % end of Language Description
```

The styles for the different elements are defined using start-stop setups. One example is the style for the LDitem (Language Definition item) which was initially coded in this way:

```
\definestartstop % Language Description Item Part 1 % deprecated code!
  [LDitem]
  [before={\switchtobodyfont[GentiumBookIt,\LDitemfontsize]},
   after={\switchtobodyfont[Gentium,\bodyfontpartone]}]
```

Eventually bodyfont switches were replaced by proper ConTEXt-style typescripts, but the idea remains the same: `\definestartstop[something][code here]` makes it possible to use the pair `\startsomething` and `\stopsomething`.

Dynamic running header

As the example of the page with language descriptions (figure 1) shows the Country name is inserted in the header of the page, using the first country on a left page and the last country on the right page. The code used to do this is based on an example in `page-set.tex` in the ConTEXt distribution.

```
\definemarking[headercountryname]
\setupheadertexts[\setups{show-headercountryname-marks}]
 \startsetups show-headercountryname-first
 \getmarking[headercountryname][1][first] % get first marking
 \stopsetups
 \startsetups show-headercountryname-last
 \getmarking[headercountryname][2][last] % get last marking
 \stopsetups
\setupheadertexts[]
\setupheadertexts
  [\setups{text a}][]
  [][\setups{text b}] % setup header text (left and right pages)
\startsetups[text a] % setup contents page a
 \rlap{Ethnologue}
 \hfill
 {\pagenumber}
 \hfill
 \llap{\setups{show-headercountryname-last}}
\stopsetups
\startsetups[text b] % setup contents page b
  \rlap{\setups{show-headercountryname-first}}
  \hfill
  \pagenumber
  \hfill
  \llap{Ethnologue}
\stopsetups
```

Language Name Index

This index lists every name that appears in Part I as a primary or alternate name of a language or dialect. The following abbreviations are used in the index entries: *alt.* 'alternate name for', *alt. dial.* 'alternate dialect name for', *dial.* 'primary dialect name for', *pej. alt.* 'pejorative alternate name for', and *pej. alt. dial.* 'pejorative alternate dialect name for'. The index entry gives the primary name for the language with which the given name is associated, followed by the unique three-letter language code in square brackets. The numbers identify the pages on which the language entries using the indexed name may be found. If the list of page references includes the entry in the primary country, it is listed first. The entry for a primary name also lists page numbers for the maps on which the language occurs.

Figure 2. Start of the language name index

Index
Since all the data for this publication comes from a database it was easy to compile a list of index items from that data. Page numbers were resolved using ConTEXt's internal referencing system. The data contains references using three letter ISO code for language and a two letter country code like this:

```
\pagereference[bel-BY] % ISO code - country code
```

In the file with the index data this reference is linked to an index item:

```
Belarusan [bel], \at[bel-BY]
```

The code `[bel-BY]` is automatically replaced by the right page number(s) producing the correct entry in the index:

Belarusan [bel], 32, 224

Since the language name index (the biggest index) contains more than 100.000 references it can be imagined that typesetting this publication in one run was pushing the limits of TEX. This is the first time that ConTEXt is used to typesetting this publication. The previous version was produced using Ventura but when that program was replaced by InDesign there were some questions about the way in which InDesign works with the automatically generated data. TEX seemed to be the right tool to use for this project and it sparked renewed interest in the use of TEX for other data-intensive publications like dictionaries.

Exploring language

Counting languages is not the only way to collect language data: many linguists move into a language group and take a closer look at the different parts of the actual languages. Some linguists focus on the sounds of a language, others analyse the sentence structure or the way in which language is used in specific communication processes. The collected data is stored in a special database program called FieldWorks. FieldWorks runs on Windows only (though a Linux port is work in progress) and it is a free download from the SIL website[2]. FieldWorks is actually a suite of programs consisting of Data Notebook, Language Explorer and WorldPad. FieldWorks Data Notebook is used for anthropological observations. FieldWorks WorldPad is a 'world ready' text editor with some special script support (including Graphite[3]). FieldWorks Language Explorer (FLEx) is used to store all sorts of language related data. It is basically a complex database program with a couple of linguistics related tools. FLEx contains a lexicon for storing data related to words, meaning(s), grammatical information about words and translations in other languages. Another part of FLEx is the interlinear tool which makes it possible to take a text in one language and to give a 'word for word translation' in another language, for example as a way to discover grammatical structures. FLEx comes with a grammar tool to facilitate the analysis and description of the grammar of a language. Since all language data is stored in the same database there are some interesting possibilities to integrate the language data and analysis tools.

Dictionaries

Once a field linguist has collected a certain amount of data he can start to think about the production of a word list or a real dictionary. To facilitate this a team of programmers has made tool called 'Dictionary Express'. This tool allows for the easy production of dictionaries based on data available in the FLEx database. The user of FLEx gets a menu option 'Print dictionary' and is presented with small window to enter some layout options. Behind the scenes one of two output paths is used: one is based on the use of an OpenOffice document template and another one uses XƎTEX and ConTEXt to typeset the dictionary. XƎTEX was chosen because of the requirement to facilitate the

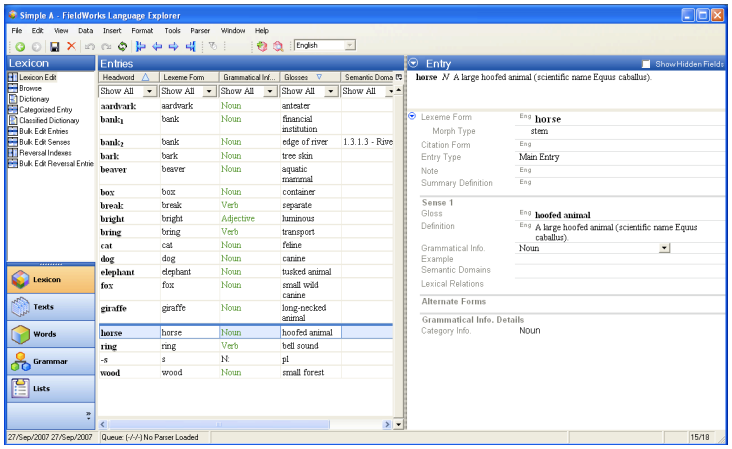

Figure 3. FieldWorks Language Explorer main window

use of the Graphite smart font technology used for the correct rendering of complex non-roman script fonts in use in some parts of the world (see footnote 2). The use of XƎTEX does of course mean that we use ConTEXt MkII.

All data is available in an XML format and converted (using a purpose built converter) to a simple TEX-tagged format. A typical dictionary entry looks like this:

```
\Bentry
\Bhw{abel}\Ehw
\marking[guidewords]{abel}
\Bpr{a.ᵐbɛl}\Epr
\Bps{noun(al)}\Eps
\Blt{Eng}\Elt
\Bde{line, row}\Ede
\Blt{Pdg}\Elt
\Bde{lain}\Ede
\Bps{noun(al)}\Eps
\Blt{Eng}\Elt
\Bde{pole, the lowest of the three horizontal poles to which a fence is
tied and which form the main horizontal framework for the fence. This
is the biggest of the three}\Ede
\Eentry
```

The tags used in this data file include:

- headword (hw): this is the word that this particular entry is about,
- pronunciation (pr): the proper pronunciation of the word written using the International Phonetic Alphabet (IPA),
- part of speech (ps): the grammatical function of the word,
- language tag (lt): the language of the definition or example,
- definition (de): meaning of the headword,
- example (ex): example of the word used in a sentence.
- `\marking[guidewords]{}`: is used to put the correct guideword at the top of each page. (The code used here is inspired by the code used to put country name in the headers in the Ethnologue project.)

Currently most of the required features are implemented. This includes: font selection (including the use of Graphite fonts), basic dictionary layout and picture support. Some of these features are strait-forward and easy to implement. Other features such as picture support required more work e.g. page wide pictures keep floating to the next

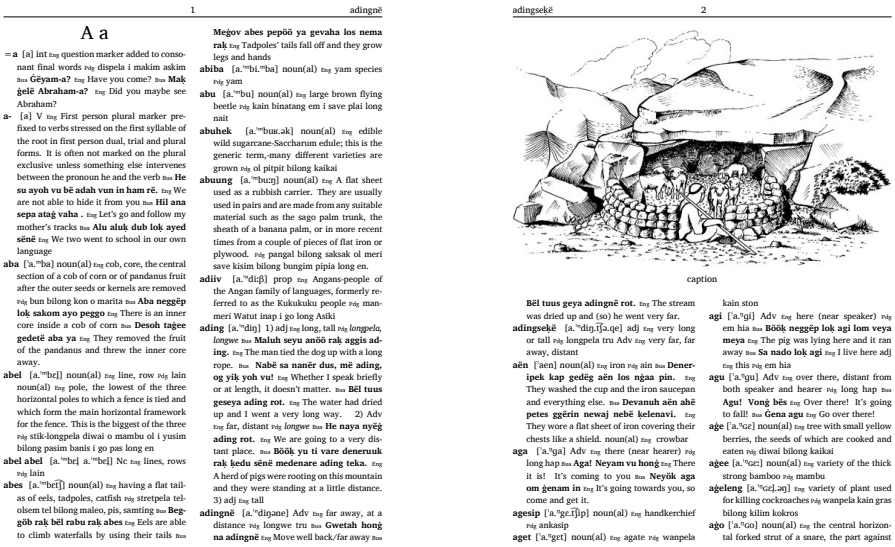

Figure 4. Sample double column dictionary layout

page. Since it is usually a good idea to separate form and content most of the layout related settings are not stored in the data file itself but in a separate settings file which is loaded at the start of the typesetting process. Examples of settings in this file include the fonts and the use of a double column layout. Default settings are used unless the user has specified different settings using the small layout options window at the start of the process.

Currently the test version of this ConTEXt-based system works with a stand alone ConTEXt-installation, using the 'minimals' distribution. One of the remaining challenges is to make a light weight, easy to install version of ConTEXt which can be included with the FieldWorks software. Since the main script used by ConTEXt Mark II is a Ruby script this requires dealing with (removing) the Ruby dependency. It is hoped that stripping the TEX-tree of all unused fonts and code will help too to reduce the space used by this tool. This is currently work in progress.

Footnotes

1. Lewis, M. Paul (ed.), Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International (2009)
2. http://www.sil.org/computing/fieldworks/
3. http://scripts.sil.org/RenderingGraphite

Jelle Huisman
SIL International
Horsleys Green
High Wycombe
United Kingdom
HP14 3XL
jelle_huisman (at) sil (dot) org