SIL's Data Collection.

History of data collection
Data in the collection
Access to the data
Uses
Questions and answers

History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

SIL's Data Collection.

1934	Pen and paper. Shoeboxes.
1959	Hixkaryana - Des Derbyshire
1976	First portable computer for field linguistics
1987	Shoebox and Toolbox
2000	Lingualinks and FieldWorks
2017	Language Forge and Dictionary App Builder

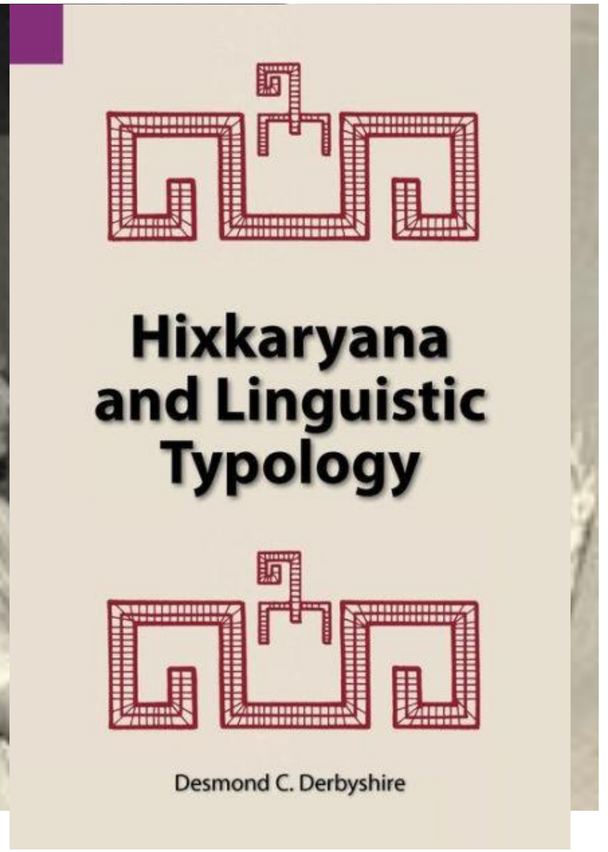
History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

SIL's Data Collection.

1934	Pen and paper. Shoeboxe
1959	Hixkaryana - Des Derbys
1976	First portable computer f
1987	Shoebox and Toolbox
2000	Lingualinks and FieldWo
2017	Language Forge and Dict



History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

SIL's D

1934

1959

1976

1987

2000

2017

History

SIL International

David Baines david_baines@sil.org

SIL's Data Collection.

1934

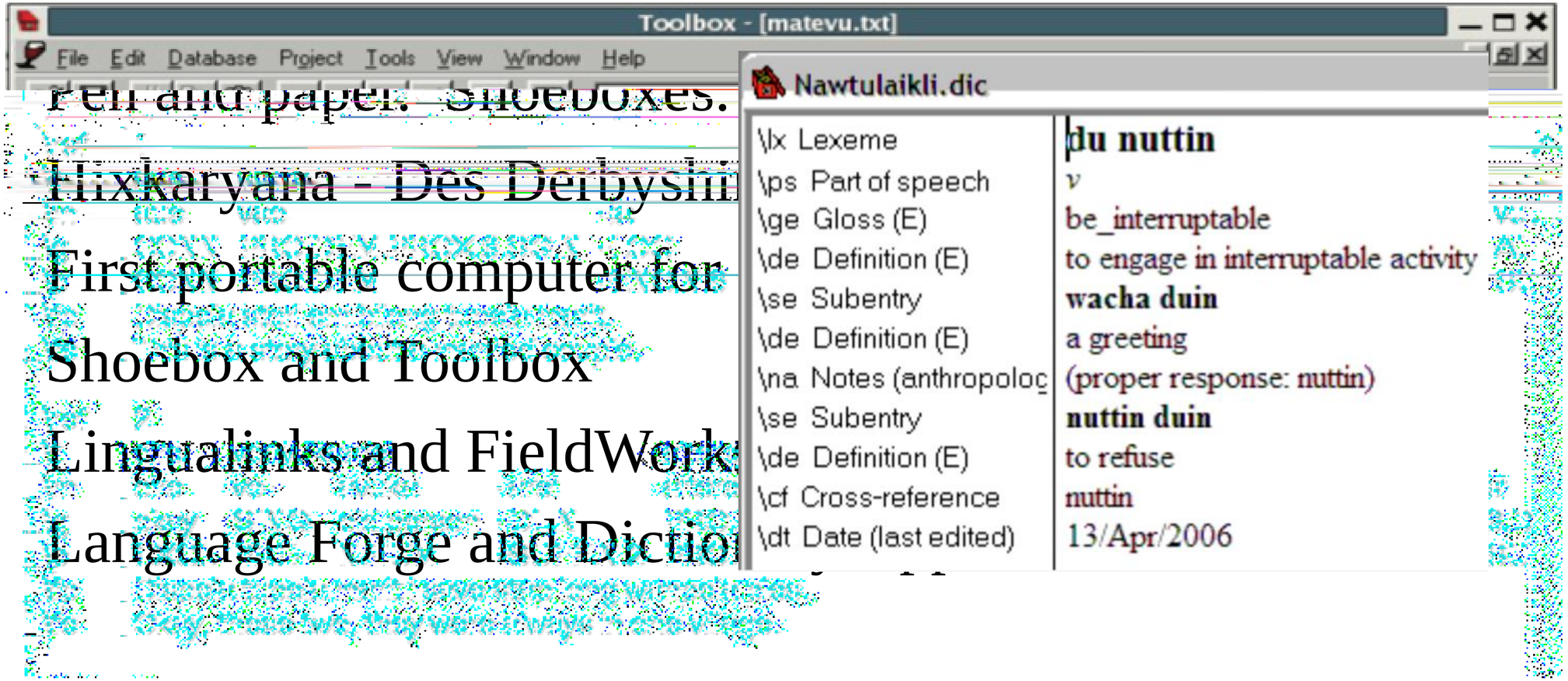
1959

1976

1987

2000

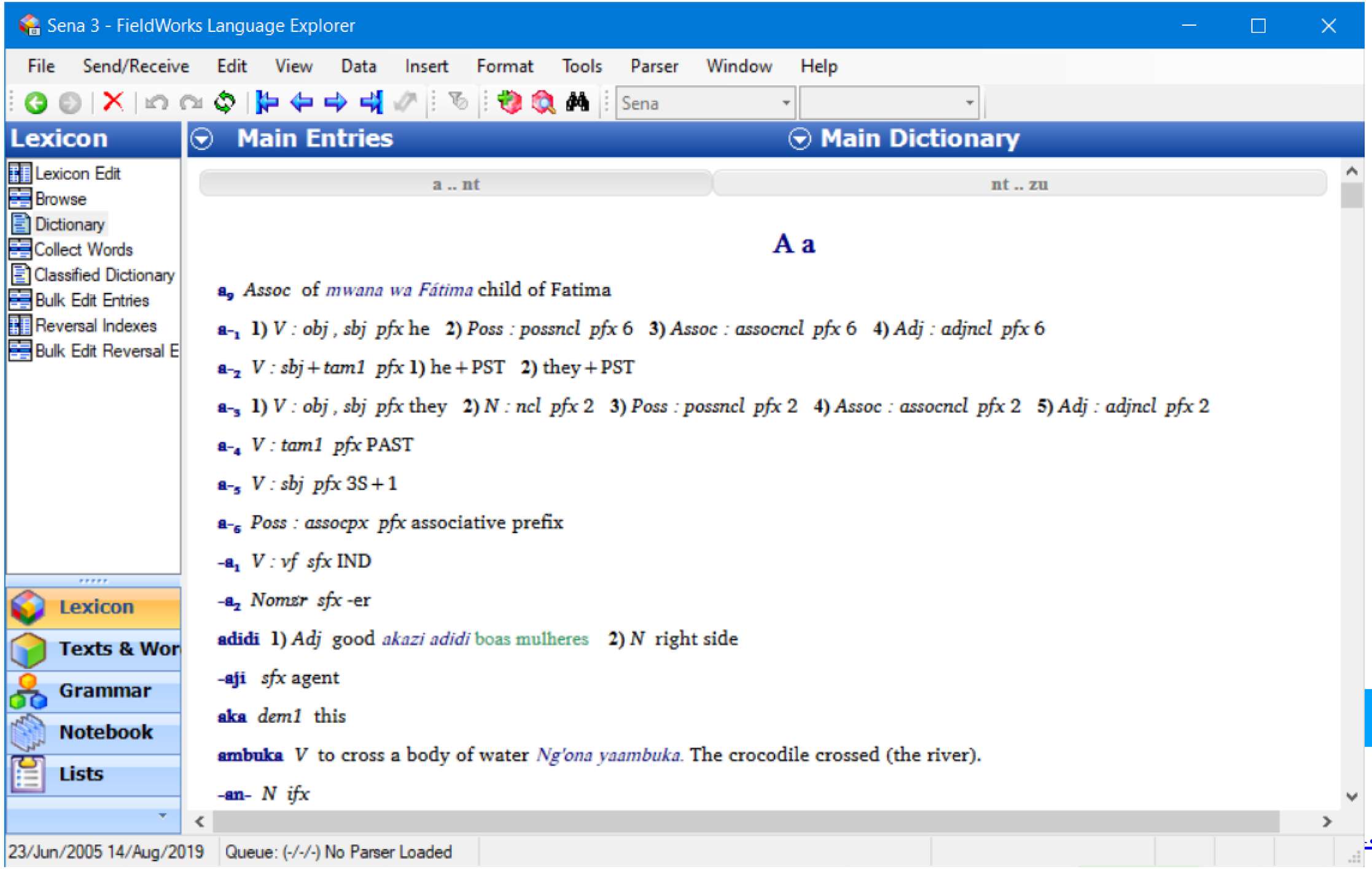
2017



History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org


Sena 3 - FieldWorks Language Explorer
File Send/Receive Edit View Data Insert Format Tools Parser Window Help
Sena
Lexicon
Lexicon Edit
Browse
Dictionary
Collect Words
Classified Dictionary
Bulk Edit Entries
Reversal Indexes
Bulk Edit Reversal E
Lexicon
Texts & Wor
Grammar
Notebook
Lists
Main Entries
Main Dictionary
a .. nt
nt .. zu
A a
a9 Assoc of mwana wa Fátima child of Fatima
a-1 1) V : obj , sbj pfx he 2) Poss : possncl pfx 6 3) Assoc : assocncl pfx 6 4) Adj : adjncl pfx 6
a-2 V : sbj + tam1 pfx 1) he + PST 2) they + PST
a-3 1) V : obj , sbj pfx they 2) N : ncl pfx 2 3) Poss : possncl pfx 2 4) Assoc : assocncl pfx 2 5) Adj : adjncl pfx 2
a-4 V : tam1 pfx PAST
a-5 V : sbj pfx 3S + 1
a-6 Poss : assocpx pfx associative prefix
-a1 V : vf sfx IND
-a2 Nomzr sfx -er
adidi 1) Adj good akazi adidi boas mulheres 2) N right side
-aji sfx agent
aka dem1 this
ambuka V to cross a body of water Ng'ona yaambuka. The crocodile crossed (the river).
-an- N ifx
23/Jun/2005 14/Aug/2019
Queue: (-/-/-) No Parser Loaded

.org

French-English Scuba

Words in dictionary

109 entries

Search Entries | Options

- [Empty] — [Empty]
- air — the mix of gases that constitute the atmospher...
- apnée — the suspension of breathing.
- atmosphérique — relating to the atmosphere.
- autonome — having freedom to govern or control oneself.
- azote

Entry Preview

Entry

Word	Frn	
Pronunciation	Frn_IPA	
	en_v	
PictureAttribution	en	
	Frn	

Data Collection: Four sources

Data Store	Type of data	Items
Ethnologue	Language Identification. Language and Country profile data.	7111
Digital Bible Library	Translations	828
Webonary	Dictionaries	180
REAP	Books, Academic papers Dictionaries,Vocabularies. "Everything".	45000

SIL International

David Baines david_baines@sil.org

Data Collection: Language Identification

- Data Store
- Ethnologue
- Digital Bible Library
- Webonary
- REAP

The ISO 639-3 standard identifies every language with a three letter code.

Ethnologue.com has summary information about every language and country.

History | Data | Access | Uses | Q&A

David Baines david_baines@sil.org

Data Collection: Aligned Translations

Data Store

Ethnologue

Digital Bible Library

Webonary

REAP

The Bible consists of the Old and New Testaments.
English ~728000 words
Old Testament twice as long as the New Testament

History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

Data Collection: Aligned Translations

Data Store

Ethnologue

Digital Bible Library

Webonary

REAP

Aligned parallel corpora are rare.

EU: 24 official languages
Large parallel corpus with 26 languages.

History | Data | Access | Uses | Q&A

David Baines david_baines@sil.org

Data Collection: Aligned Translations

- Data Store
- Ethnologue
- Digital Bible Library
- Webonary
- REAP

Translations for 828 languages.
Mostly New Testaments

Aligned at verse level. ~ Sentence
Most have 100k to 400k words.
Digital format - XML

History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

Data Collection: Aligned Translations

Total 2686 languages with some translation:

Bibles in 695 languages.
New Testaments in 1550 languages.
Portions in 1136 languages.

Data from progress.Bible Sept 2019 snapshot

History | Data | Access | Uses | Q&A

David Baines david_baines@sil.org

Data Collection: Dictionaries - Webonary

Data Store

Ethnologue

Digital Bible Library

Webonary

REAP

webonary.org

Dictionaries in 181 languages.

150 have at least 1000 entries

Largest has 46000 entries.

Various stages of completion.

History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

Data Collection: Archive - REAP

Data Store

Ethnologue

Digital Bible Library

Webonary

REAP

Published at:

www.sil.org/resources

19000 for download

45000 citations

Request : www.sil.org/contact

History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

SIL's Data Collection: Ethnologue

It's easy, sometimes.
The ISO 639-3 data can be downloaded from : www.ethnologue.com/codes

ethnologue.com by subscription
From $5 per month

History | Data | Access | Uses | Q&A

SIL International

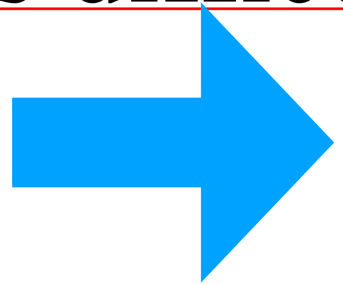
David Baines david_baines@sil.org

SIL's Data Collection: Translations

Sometimes it’s difficult.

Translations

Dictionaries



Multiple agencies

Sometimes it’s easy.

YouVersion App

bible.is

scriptureearth.org

scripture.api.bible

History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

SIL's Data Collection: Dictionaries

Dictionaries can be found on Webonary
Not easily machine readable - xhtml
Few available for download in pdf

More available from the Archives

History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

SIL's Data Collection: Archive

www.sil.org/resources

Request : www.sil.org/contact

History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

SIL's Data Collection.

1. Individual language study.

Dictionaries and Translations (xml):

1. Comparative language studies.
2. Multilingual parallel corpus.

History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

SIL's Data Collection.

History | Data | Access | Uses | Q&A

SIL International

David Baines david_baines@sil.org

SIL's Data Collection.

History

Data

Access

Uses

SIL International

David Baines david_baines@sil.org