SIL's Data Collection.

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

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
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

SIL International

David Baines

david_baines@sil.org
SIL's Data Collection

1934
1959
1976
1987
2000
2017

History

SIL International

David Baines  david_baines@sil.org
SIL's Data Collection.

Pen and paper. Shoeboxes.

Hixkaryana - Des Derbyshire
First portable computer for field linguistics
Shoebox and Toolbox
Lingualinks and FieldWorks
Language Forge and Dictionary App Builder

1934
1959
1976
1987
2000
2017

History  Data  Access  Uses  Q&A
Pen and paper. Shoeboxes.

- Hixkaryana
- Des Derbyshire

First portable computer for field linguistics

Shoebox and Toolbox

Lingualinks and FieldWorks

Language Forge and Dictionary App Builder

1934

1959

1976

1987

2000

2017

A a

a₂ Assoc of mwana wa Fátima child of Fatima

a₂-1) V : obj, sbj pfx he 2) Poss : possnel pfx 6 3) Assoc : assoencl pfx 6 4) Adj : adjncl pfx 6

a₂-2) V : sbj + tam1 pfx 1) he + PST 2) they + PST


a₂-4) V : tam1 pfx PAST

a₂-5) V : sbj pfx 3S+1

a₂-6) Poss : assocex pfx associative prefix

-a₁) V : vf sfx IND

-a₂) Nomsv 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

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1934</td>
<td></td>
</tr>
<tr>
<td>1959</td>
<td></td>
</tr>
<tr>
<td>1976</td>
<td></td>
</tr>
<tr>
<td>1987</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
</tr>
</tbody>
</table>

Pen and paper.  Shoeboxes.

- Hixkaryana
- Des Derbyshire
- First portable computer for field linguistics
- Shoebox and Toolbox
- Lingualinks and FieldWorks
- Language Forge and Dictionary App Builder

David Baines
david_baines@sil.org
SIL's Data Collection.

French-English Scuba
My Projects / French-English Scuba / Edit

Words in dictionary
109 entries

<table>
<thead>
<tr>
<th>Entry</th>
<th>Word</th>
<th>Pronunciation</th>
<th>PictureAttribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Empty] [Empty]</td>
<td>Frn</td>
<td>Frn_IP</td>
<td>en_y</td>
</tr>
<tr>
<td>air</td>
<td>the mix of gases that constitute the atmosphere.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>apnée</td>
<td>the suspension of breathing.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atmosphérique</td>
<td>relating to the atmosphere.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>autonome</td>
<td>having freedom to govern or control oneself.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>azote</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Data Collection: Four sources

<table>
<thead>
<tr>
<th>Data Store</th>
<th>Type of data</th>
<th>Item(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnologue</td>
<td>Language Identification. Language and Country profile data.</td>
<td>7111</td>
</tr>
<tr>
<td>Digital Bible Library</td>
<td>Translations</td>
<td>828</td>
</tr>
<tr>
<td>Webonary</td>
<td>Dictionaries</td>
<td>180</td>
</tr>
<tr>
<td>REAP</td>
<td>Books, Academic papers Dictionaries, Vocabularies. &quot;Everything&quot;.</td>
<td>45000</td>
</tr>
</tbody>
</table>

SIL International

David Baines  david_baines@sil.org
The ISO 639-3 standard identifies every language with a three letter code. **Ethnologue.com** has summary information about every language and country.
Data Collection: Aligned Translations

Aligned parallel corpora are rare.

EU: 24 official languages
Large parallel corpus with 26 languages.
Data Collection: Aligned Translations

Translations for 828 languages. Mostly New Testaments

Aligned at verse level. ~ Sentence
Most have 100k to 400k words.

Digital format - XML
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
Data Collection: Dictionaries - Webonary

**webonary.org**

Dictionaries in 181 languages. 150 have at least 1000 entries. Largest has 46000 entries. Various stages of completion.
Data Collection: Archive - REAP

Published at: [www.sil.org/resources](http://www.sil.org/resources)
19000 for download
45000 citations

Request: [www.sil.org/contact](http://www.sil.org/contact)
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

SIL International
David Baines  david_baines@sil.org
SIL's Data Collection: Translations

Sometimes it’s difficult.
Translations

Sometimes it’s easy.
Dictionaries

Multiple agencies

YouVersion App
bible.is
scriptureearth.org
scripture.api.bible

David Baines
david_baines@sil.org

SIL International
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
SIL's Data Collection: Archive

www.sil.org/resources

Request: www.sil.org/contact
SIL's Data Collection.

1. Individual language study.

Dictionaries and Translations (xml):
1. Comparative language studies.
SIL's Data Collection.
SIL's Data Collection.

Q&A

History  Data  Access  Uses

SIL International  

David Baines  david_baines@sil.org