Crawling, Archiving, and Corpus Creation

Michael Cysouw & Hans-Jörg Bibiko

Max Planck Institute for Evolutionary Anthropology

Goal: typological sample of language corpora

- include ‘small’ languages with **limited data** available
- use **all data** (as much as possible) available
- **linguistic specialists** have to be involved

Specialist’s worries should also be our worries

- **details** become very important
  - transcription
  - context
  - author
- **trackability** of corpus creation
- original context is **recoverable**
- **corrections** should be possible

Proposed phases of corpus construction

- collection
- extraction
- synchronization
- versioning

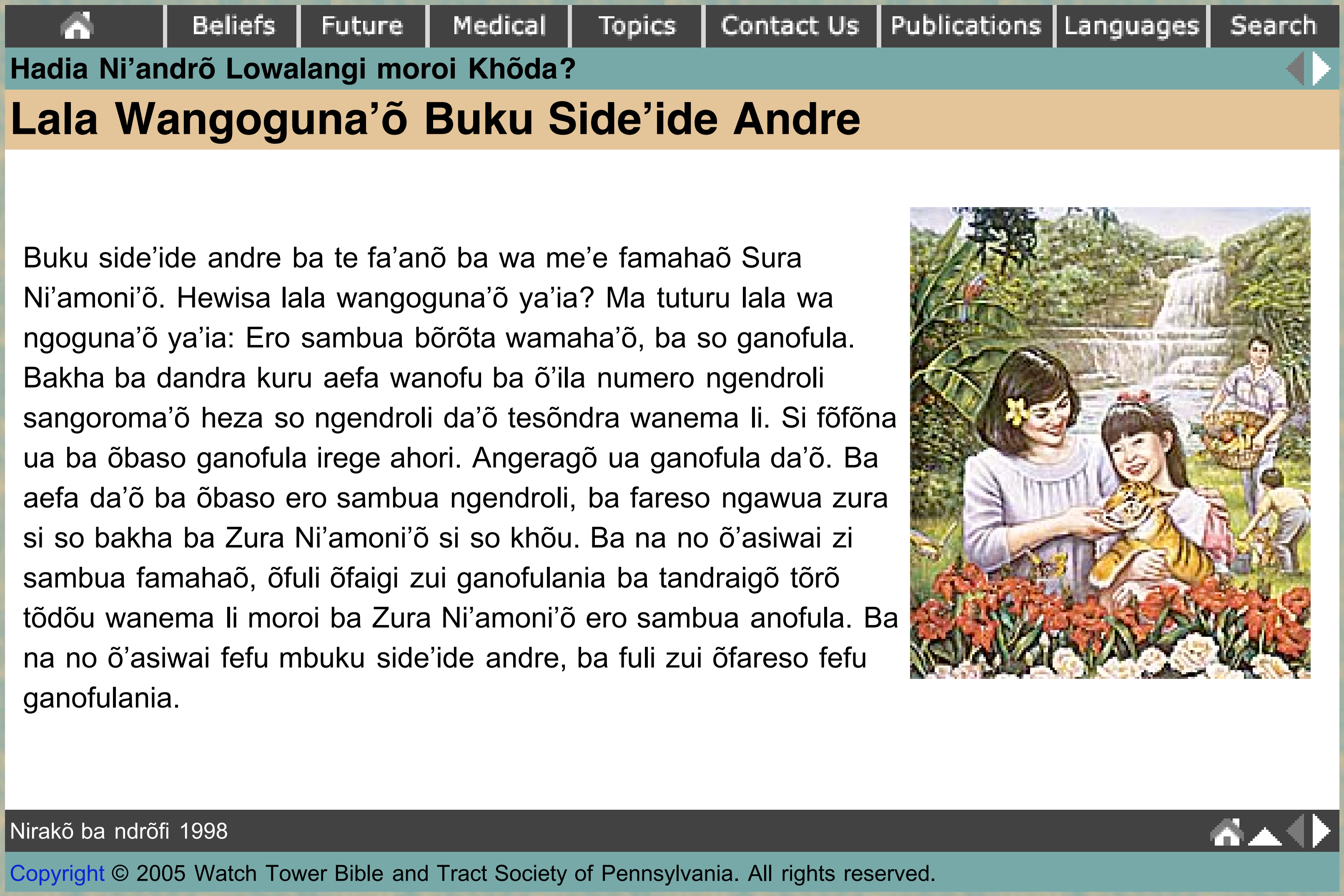
Every phase should be recoverable!

Collection

- keep **original text** in archive (possibly with strong access restrictions)

- annotate files with **metadata**
  - SIL language code
  - file format
  - text encoding used (store font is needed!)
  - date, genre, ...

Web-crawling



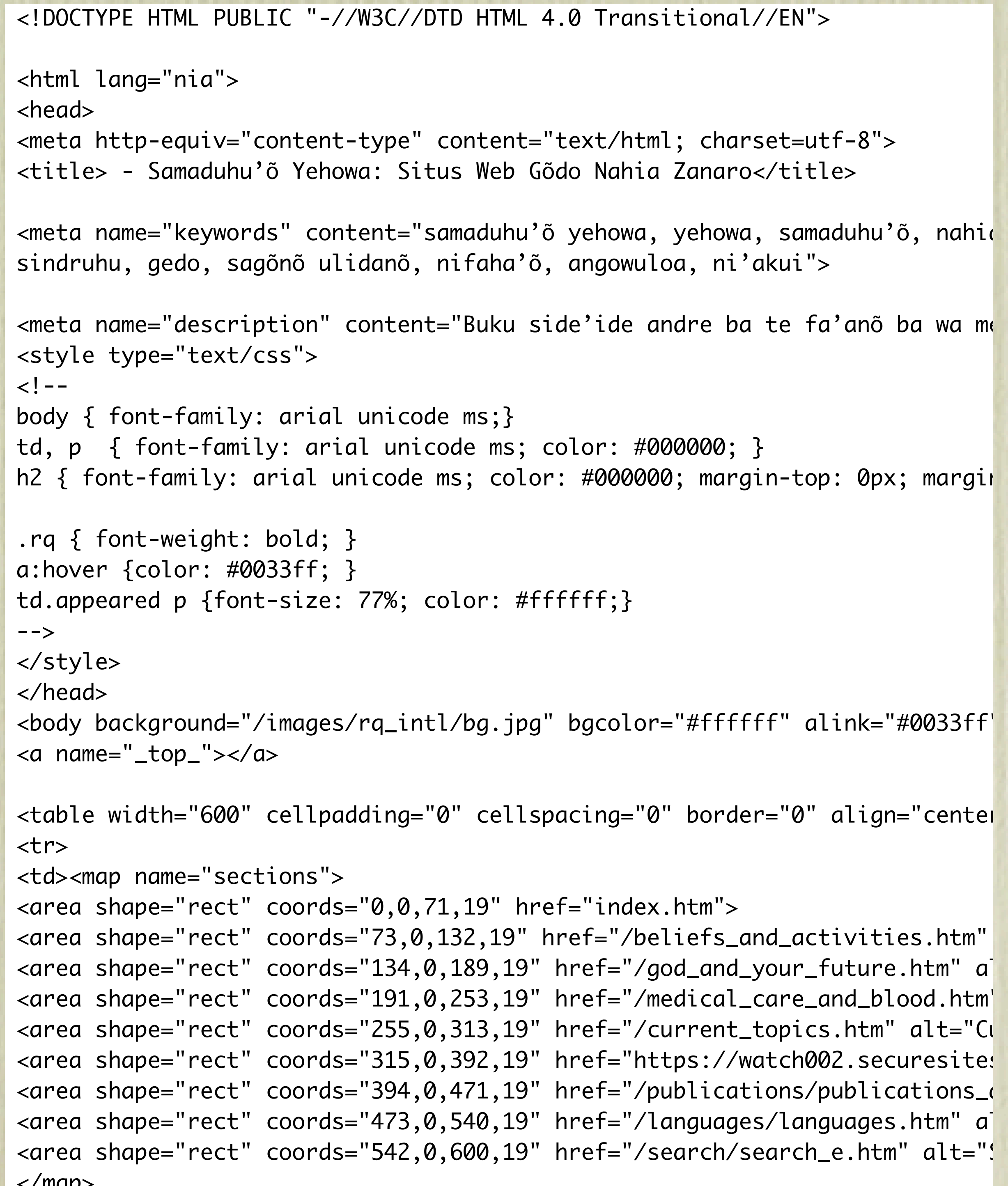
Hadia Ni'andrõ Lowalangi moroi Khõda?

Lala Wangoguna'õ Buku Side'ide Andre

Buku side'ide andre ba te fa'anõ ba wa me'e famahaõ Sura Ni'amoni'õ. Hewisa lala wangoguna'õ ya'ia? Ma tuturu lala wa ngoguna'õ ya'ia: Ero sambua bõrõta wamaha'õ, ba so ganofula. Bakha ba dandra kuru aefa wanofu ba õ'ila numero ngendroli sangoroma'õ heza so ngendroli da'õ tesõndra wanema li. Si fõfõna ua ba õbaso ganofula irege ahori. Angeragõ ua ganofula da'õ. Ba aefa da'õ ba õbaso ero sambua ngendroli, ba fareso ngawua zura si so bakha ba Zura Ni'amoni'õ si so khõu. Ba na no õ'asiwai zi sambua famahaõ, õfuli õfaigi zui ganofulania ba tandraigõ tõrõ tõdõu wanema li moroi ba Zura Ni'amoni'õ ero sambua anofula. Ba na no õ'asiwai fefu mbuku side'ide andre, ba fuli zui õfareso fefu ganofulania.

Nirakõ ba ndrõfi 1998

Keep source



```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">

<html lang="nia">
<head>
<meta http-equiv="content-type" content="text/html; charset=utf-8">
<title> - Samaduhu’õ Yehowa: Situs Web Gõdo Nahia Zanaro</title>

<meta name="keywords" content="samaduhu’õ yehowa, yehowa, samaduhu’õ, nahi
sindruhu, gedo, sagõnõ ulidanõ, nifaha’õ, angowuloa, ni’akui">

<meta name="description" content="Buku side’ide andre ba te fa’anõ ba wa me
<style type="text/css">
<!--
body { font-family: arial unicode ms;}
td, p  { font-family: arial unicode ms; color: #000000; }
h2 { font-family: arial unicode ms; color: #000000; margin-top: 0px; margi

.rq { font-weight: bold; }
a:hover {color: #0033ff; }
td.appeared p {font-size: 77%; color: #ffffff;}
-->
</style>
</head>
<body background="/images/rq_intl/bg.jpg" bgcolor="#ffffff" alink="#0033ff"
<a name="_top_"></a>

<table width="600" cellpadding="0" cellspacing="0" border="0" align="cente
<tr>
<td><map name="sections">
<area shape="rect" coords="0,0,71,19" href="index.htm">
<area shape="rect" coords="73,0,132,19" href="/beliefs_and_activities.htm"
<area shape="rect" coords="134,0,189,19" href="/god_and_your_future.htm" a
<area shape="rect" coords="191,0,253,19" href="/medical_care_and_blood.htm"
<area shape="rect" coords="255,0,313,19" href="/current_topics.htm" alt="Cu
<area shape="rect" coords="315,0,392,19" href="https://watch002.securesite
<area shape="rect" coords="394,0,471,19" href="/publications/publications_
<area shape="rect" coords="473,0,540,19" href="/languages/languages.htm" a
<area shape="rect" coords="542,0,600,19" href="/search/search_e.htm" alt="
</map>
```

Extraction

- extract **relevant pieces** of text
- transfer to **unicode**
- build **transcription profile**
  - characters used
  - include multigraphs (gh, sch, ò, m´¸, ...)
- this phase can probably be **automatised**

Extraction

```
Hadia Ni'andrõ Lowalangi moroi Khõda?
Lala Wangoguna'õ Buku Side'ide Andre

Buku side'ide andre ba te fa'anõ ba wa me'e famahaõ Sura Ni'amoni'õ. Hewisa lala
wangoguna'õ ya'ia? Ma tuturu lala wa ngoguna'õ ya'ia: Ero sambua bõrõta wamaha'õ,
ba so ganofula. Bakha ba dandra kuru aefa wanofu ba õ'ila numero ngendroli
sangoroma'õ heza so ngendroli da'õ tesõndra wanema li. Si fõfõna ua ba õbaso
ganofula irege ahori. Angeragõ ua ganofula da'õ. Ba aefa da'õ ba õbaso ero sambua
ngendroli, ba fareso ngawua zura si so bakha ba Zura Ni'amoni'õ si so khõu. Ba na
no õ'asiwai zi sambua famahaõ, õfuli õfaigi zui ganofulania ba tandraigõ tõrõ
tõdõu wanema li moroi ba Zura Ni'amoni'õ ero sambua anofula. Ba na no õ'asiwai
fefu mbuku side'ide andre, ba fuli zui õfareso fefu ganofulania.
```

```
  '  ,  .  :  ?  a  A  B  b  d  e  E  f  g  H  h  i  K
k  L  l  m  M  N  n  o  õ  r  S  s  t  u  w  W  y  z  Z
```

Synchronization

- **interlink** transcription profiles (probably through rough IPA matching)
- for this, a **language specialist** is needed
- **combine texts** into corpus

Versioning

- it should be made possible to **correct errors**
- a system of **versioning** is necessary
- the **organization** of this is still unclear

Uses

- typological **surveys**
- language **classification**
- **T9 dictionaries** for minorities

Thank you