Citing the spoken word: Adventures in language documentation

Nicholas Thieberger
School of Languages and Linguistics
The University of Melbourne
Department of Linguistics
University of Hawai‘i at Mānoa
Me, selwan ag kupi eḿae, tiawi itraus traus traus traus, natrauswen ga itaos nlag. Itrausi pan kaipa. Me komam uta laap kin uto mau, a?
Malen umat, inom.

But when you are far away the old man can talk and talk and talk, his story is like the wind. He tells it and it is gone. But there aren't many of us left. When we die, it will be finished.

Il n’est pas d’avantage possible à la linguistique de représenter fidèlement le parler des langues vivantes même à l’aide de ces transcriptions arbitraires, tout à fait insuffisantes, imaginées faute de mieux, et qu’aucun accord n’est parvenu à uniformiser. ... tout celà dont la linguistique est incapable, le phonographe dès maintenant en offre la possibilité.”

“Nor is it possible for linguistics to faithfully represent the speech of living languages even with the help of its arbitrary and totally insufficient transcriptions, made up for want of better and which no agreement has ever made uniform. ... all that of which linguistics is incapable, the phonograph from now on makes possible.”

(Léon Azoulay 1900:175)
‘there is the special methodological interest in the frank and full presentation of all available linguistic evidence’ (Malinowski 1935: 4)
‘If I could, by a good phonographic record, counterfeit the living voice of Tokulubakiki: [...] I should certainly be better able to *translate* the text in the sense of imparting to it its full cultural flavour and significance. Again, if by cinematographic picture I could reproduce the facial expression, the bodily attitude, the significant gestures, this would add another contextual dimension.’ (Malinowski 1935: 26)
“I'm not particularly interested in 'smoothed-over' versions of native culture. I like the stuff in the raw, as felt and dictated by the natives. [...] The genuine, difficult, confusing, primary sources. These must be presented, whatever else is done. [...] There are too many glib monographs, most of which time will show to be highly subjective performances. We need to develop in cultural anthropology that *anxious respect for documentary evidence* that is so familiar to the historian, the classical scholar, the Orientalist. [...] where is the raw evidence? I can't tell whether a given statement is common native knowledge or is merely your interpretation of one man's say-so.”

‘There is no reason whatever why, in the future, an exact and physiological study of speech should not use the apparatus of sound films for reproducing fully contextualised utterances.’ (Malinowski 1935: 26)

Malinowski, Bronislaw. 1935. *Coral gardens and their magic*. Vol II
Descriptive grammars written since Sapir and Malinowski may have been built on a corpus of recordings, but that is not evident from the finished product.
Perspectives on grammar writing

Payne: ‘good linguistic descriptions provide the raw data for much research into the organization of the human mind’

In recent years the ‘emphasis has shifted back to the centrality of data.’

Studies in Language, 2006, 30:2 p.236
second person, i.e., the linguist), 17.1.3 (asking address to convey prohibition to other person); negative suggestion 20.19.3 (if the sense is approximately 'let's not go there')

- negation of morally/legally improper behaviour ('mustn't') 94.1.6, 94.4.0.2
- 79.4.4, 67.2.2.3, 68.4.5, 68.5.3, 75.1.3 (three), 75.2.2, 76.1.4, 76.2.2/5, 77.1.3, 77.2.1/2/5, 83.2.2/3 (three), 89.1.5, 85.11.1/4/7, 88.3.1/2/3

- note that most of these exx. are in a sequence of short texts about customary kinship-based behaviour; inability (usually translated as English cannot with present t.)
- tense 20.7.3, 20.6.5, 42.6.4 (all involving sense 'cannot see'); negative intension (1st person) or predication (any person) 17.13.4 ('I won't speak you'), 28.1.4 (L2a), 62.5.2 (L2a), 90.2.7 ('He will not [be forced to] go to jail'; perhaps in moral/legal sense but if so not focusing on subject's responsibility, rather than named coercing agents), 164.2.4 ('we[EXDu] might not return' with particle /ug/ 'maybe', §12.11, 164.3.3 [EXDu], 165.7 (3PL, prediction).

- Past Potential Negative is uncommon in the texts. It occurs in 120.6.1-3, and there is a good ex. from Hughes in TNT 27. Past Potential Negative is used in negating either clause in a counter-factual condition, see discussion §16.7. The ex. in TNT 27 is as follows in my transcription:

\[(15.xli!)\]
\[\text{yag}i-	ext{maga: } \text{bara}='\text{bl}-\text{mi} \quad \text{bara}='\text{guru}:'\]
\[\text{neg- you-hit-[it]for-then you left them}\]
\[\text{You shouldn't have hit them, you should have left them [alone]} \quad \text{(TNT 27)}\]

Both verbs are in the same form (Past Potential), but /yagi/-negates only the first, so that the second verb is positive (but still potential) and represents the antithesis of the prohibition. Postposition /maga/ is common with such contexts; see §12.21.

- There are also some special uses of the "Future Negative" form with /yagi/ which are best dealt with in other sections; for negative purposive clause see 69.2.3 and discussion §16.5, for negated clauses in hypothetical (as opposed to counterfactual) conditionals see ex. in 62.1.2/3 and discussion §16.7, and for gerundial 'while not VERB-ing' in sense 'before VERB' see ex. 47.19.3 and discussion §16.6.

Most textual exx. of /wa=-ri/ have a co-occurring inflected verb, the overall construction being Past (Actual) Negative or Present Negative (Chapter 8). The Past Negatives in NNMT are 3.1.4, 5.1.4/5, 7.6.2/3, 14.1.3, 15.1.8, 17.1.5, 19.2, 21.10.2, 28.10.4, 41.7.3/4, 47.2.5, 47.2.5/6, 47.5.4 (two), 47.6.2/3, 47.7.3/4 (three), 47.9.1, 47.11/2/3, 47.13.2, 47.17.6, 47.22.3/7, 51.1.3, 69.5.7, 69.19.1/3, 70.10.2, 71.6.3, 73.4.1, 73.6.1 (if transcription correct), 100.4.1 (two), 113.3.3, 113.7.2, 117.1.2, 117.6.5, 119.4.1, 120.3.2, 120.5.1/3, 129.15.1, 161.5.2, 161.26.1, 162.17.1, 162.18.4, 162.28.6, 163.2/3, 166.10.2/4, 166.22.2 (two), 167.1.2, 170.1.2.

There is, in addition, an intriguing example of Past (Actual) Negative where one might have expected a Past Potential Negative for semantic reasons, suggesting that in the past tense the Potential form is more marked than its counterpart (Future) is in the nonpast system: 10.2.3 ('he father could not see') contrast 'inability' exx. of /yagi/, top of this page.

First version: 1990

Australian indigenous languages information stacks

© AIATSIS 1995
Fieldwork in Eratap and Erakor villages, Vanuatu 1996–

Grammar based in a corpus

Examples cited to original sources

All material archived
Fieldwork in Eratap and Erakor villages, Vanuatu 1996–

Grammar based in a corpus

Examples cited to original sources

All material archived
How to create a citable set of textual material linked to primary media?
Time-aligned transcripts

Transcriber:  http://sourceforge.net/projects/trans/

Elan:  http://www.mpi.nl/tools/elan.html

Transana, Exmaralda, Anvil ...
Audiamus

Created for interacting with field recordings via their transcripts

Version 1 – HyperCard – 1999,
Version 2 – RunTime Revolution 2002

Text 3 Litrapong a natopu, by Kalsarap Namaf. (Play all)
Story from tape 2000lb told by Kalsarap Namaf. Transcribed and translated into Bislama by Manuel Wayane. The story concerns a natopu or spirit called Litrapong, also known in Bislama as a Lisesep. The telling of this story is also recorded on video.

Akit tumau wi tae esan ipi, go
akit tu= mau wi tae esan i= pi go
1plm 1p.ːRS= all know place 3sgRS= be and

Litrapong ne, amurin na kanrikmus ki,
Litrapong ne = mur ː iː n na kaː nriː mus kiː o
name this 1sgRS= want -TS -3sg0 COMP 1sgRS= tell -2pl0 PREP -3sg0
We all know that place, and this Litrapong, I want to tell you about.

Litrapong ne, ipi naflak ni apu.
Litrapong ne = ipi naflak ni apu
name this 3sgRS= be clan of grandfather

Gar apu rato pusrek ser tete nrak.
gar apu raː to puserek ser tete nrak
3pl.grandfather 2/3d.RS= HAB talk every some time
Litrapong. She is of grandfather’s clan. Those two, grandfather and Litrapong, would talk every now and then.

Ser nrak rato pusrek rato ur naor iskei.
ser nrak raː to puserek raː to ur naor iː skeː
every time 2/3d.RS= HAB talk 2/3d.RS= HAB follow place 3sgRS= one

Atae naliati iskei
aː tae aːliat iː skeː
1sgRS= know day 3sgRS= one
Every time they would talk, they would be at one place. I know one day

1.26 Litrapong ne, ipi naflak ni apu. Litrapong ne = ipi naflak ni apu name
1.33 ne = pi naflak ni apu name this 3sgRS= be clan of grandfather Gar
1.42 be clan of grandfather Gar apu rato pusrek ser tete nrak. gar apu raː =
1.49 pusrek ser tete nrak. gar apu raː to puserek ser tete nrak 3pl.grandfather
3.57 3sgRS= come ilakor mai saof apu = lakoːr mai saof apu 3sgRS= approximate come
3.62 apu = lakoːr mai saof apu 3sgRS= approximate come visit grandfather when we
575 him herbal medicine. Ser nawes naːpu, na a ga preg nalkis ikano nrogtiawi. ser
585 nalkis ikano nrogtiawi. ser nawes naːpu aː ga preg nalkis iː kano nrogtiawi ev...
Audiamus

- Analysis of a linguistic media corpus via text
- Building a citable corpus of media via linked transcripts
- A working tool (now) that gives access to a media corpus via a time-aligned textual index
- Takes the output of Transcriber or Elan as input
- Built in proprietary software, Runtime Revolution (rapid prototyping)
- Win/Mac (&Linux) standalone versions

Audiamus workflow

Media file
.wav, .mp3,
.mov, .mpeg, etc.

Transcribed and linked
(Transcriber, Elan)

Audiamus

Toolbox

Quicktime

Tab-delimited text
Audiamus

Ability to cite primary data

We worked and we got tired, if we wanted to stay healthy, we would (have to) get tired first.
Audiamus

Import
Tab-delimited text
Limsi label (Transcriber)

Export
Tab-delimited text
XML
Toolbox
Quicktime
Outputs from *Audiamus*

**Toolbox**

\_sh v3.0 498 SE Text

\aud 98009b 22.94 28.64
\tx (T) A, kagaag traus nmatu iskei,

\aud 98009b 29.16 37.82
\tx go nanwei iskei, rato. Me ratik ki tesa.

\aud 98009b 37.96 47.06
\tx Me ipiatlak teesa iskei ipaakor faat elau.
I will then go and cut bamboo. I cut bamboo and bring it, I get it ready

A ler mai, ka fo pei m$asel roat, a m$asel roat inom
Benefits of Audiamus

Local control of data
Immediate access (no web connection)
Ability to interact with and improve the transcript
Ability to 'clip' citation forms
Imports and exports several formats
Find and concordance function
Creation of ‘playlists’ organised by theme rather than by media
Freely distributed under GNU General Public License
Problems of Audiamus

Data has to be imported, ideally it would be in XML and read by an Audiamus-like browser.

Audiamus implies a workflow and so version control is a problem.

Proprietary software means development can only be done by buying Runtime Revolution.
Typical workflow resulting in well-formed data

Recording - named

- analogue digitised/digital captured
  - archival digital file
  - descriptive metadata added
  - transcribed and linked (using e.g. Transcriber or Elan)

- concordance of texts, navigation tool

- Media corpus instantiates links to media (e.g. Audiamus)

  - output to e.g. Toolbox for interlinearising

  - Texts, dictionary etc

- archived
Citability

Citation of the spoken word requires a persistent form of the cited form.

Persistence of name and location is provided by a digital archive.
How will WOLP maintain persistent identification of texts?

How will WOLP license the use of texts?
Cataloging texts

DOI? Handle.net?

What repository to use?

Example of PARADISEC
149 collections
6,517 items
43,301 files
4.67 TB
2653 hours of audio data
Data representing 648 languages from 60 countries
Online catalogue: paradisec.org.au/catalog

Collection level information

- **Persistent identifier**: NT1-98002
- **Catalogue link**: http://store.apac.edu.au/paradisec/repository/NT1/98002
- **Item name**: CD09-10
- **Date created**: 1998-09-27
- **Date in ISO format**: 1998-09-27
- **Countries**: VANUATU
- **Languages**: Efate, South
- **Subject language**: Efate, South
- **Content language**: Efate, South
- **Collector**: Thieberger, Nick
- **Operator**: Bone, Colin

Collection digitised in 2001. NICK THIEBERGER COLLECTION DATA. 51 CDs, total of 38 hours of language material. Original medium: audio cassette. Date of recordings: 1996-2000. All burned as AIFF data. Archive copy of each CD in AIFF data.
Online catalogue: paradisec.org.au/catalog
## Metadata

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td>CD09-10</td>
</tr>
<tr>
<td><strong>Access Rights</strong></td>
<td>standard, as per PDSC Access form</td>
</tr>
<tr>
<td><strong>Contributor (depositor)</strong></td>
<td>Thieberger, Nick</td>
</tr>
<tr>
<td><strong>Contributor (recorder)</strong></td>
<td>Thieberger, Nick</td>
</tr>
<tr>
<td><strong>Contributor (speaker)</strong></td>
<td>Kalman, Petro</td>
</tr>
<tr>
<td></td>
<td>Kalopog, Kali</td>
</tr>
<tr>
<td></td>
<td>Tenene, Waia</td>
</tr>
<tr>
<td><strong>Coverage (Box)</strong></td>
<td>northlimit=-17.7235; southlimit=-17.8235; westlimit=168.217; eastlimit=168.317</td>
</tr>
<tr>
<td><strong>Coverage (ISO3166)</strong></td>
<td>VU</td>
</tr>
<tr>
<td><strong>Date (W3CDTF)</strong></td>
<td>1998-09-27</td>
</tr>
<tr>
<td><strong>Date Created (W3CDTF)</strong></td>
<td>1998-09-27</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>There are time-aligned transcripts of this item. Language as given: South Efate, Bislama</td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>Digitised: yes; Media: audiocassette;</td>
</tr>
<tr>
<td><strong>Identifier</strong></td>
<td>NT1-98002</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>Bis</td>
</tr>
<tr>
<td></td>
<td>erk</td>
</tr>
<tr>
<td><strong>Language (ISO639)</strong></td>
<td>Bis</td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>standard, as per PDSC Access form</td>
</tr>
<tr>
<td><strong>Subject</strong></td>
<td>South Efate language</td>
</tr>
<tr>
<td><strong>Subject (ISO639)</strong></td>
<td>erk</td>
</tr>
<tr>
<td><strong>Subject (OLAC)</strong></td>
<td>language_documentation</td>
</tr>
</tbody>
</table>
Open Language Archives Community (OLAC)
http://www.language-archives.org/language/erk

Resources in and about the South Efate language

ISO 639-3: erk

The combined catalog of all OLAC participants contains the following resources that are relevant to this language:

- Primary texts
- Lexical resources
- Language descriptions
- Other resources about the language
- Other resources in the language

Other known names and dialect names: Efate, South, Erakor, Fate, Southern Efate

Primary texts

Data access (paradisec.org.au/repository)

Welcome to the PARADISEC repository
Please log in to identify yourself.

Username:
Password:
Log in

Repository collections
Here's a listing of the collections viewable by your username nxs567. Refer to the repository website glossary for an explanation of terms and abbreviations.

<table>
<thead>
<tr>
<th>Collection</th>
<th>Items</th>
<th>Files</th>
<th>Size</th>
<th>Duration</th>
<th>Catalog Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB1</td>
<td>11</td>
<td>44</td>
<td>29.38GB</td>
<td>14:47:59:60</td>
<td>metadata</td>
</tr>
<tr>
<td>AC1</td>
<td>58</td>
<td>397</td>
<td>64.81GB</td>
<td>33:15:06:00</td>
<td>metadata</td>
</tr>
<tr>
<td>AC2</td>
<td>1.025</td>
<td>12,567</td>
<td>18.06GB</td>
<td>00:00:00:00</td>
<td>metadata</td>
</tr>
<tr>
<td>AM2</td>
<td>14</td>
<td>56</td>
<td>36.65GB</td>
<td>19:08:27:40</td>
<td>metadata</td>
</tr>
<tr>
<td>AM3</td>
<td>13</td>
<td>46</td>
<td>32.41GB</td>
<td>16:19:28:90</td>
<td>metadata</td>
</tr>
<tr>
<td>AM4</td>
<td>1</td>
<td>2</td>
<td>1.50GB</td>
<td>00:45:17:99</td>
<td>metadata</td>
</tr>
<tr>
<td>AP1</td>
<td>4</td>
<td>14</td>
<td>6.00GB</td>
<td>03:01:18:90</td>
<td>metadata</td>
</tr>
<tr>
<td>AR1</td>
<td>13</td>
<td>42</td>
<td>23.00GB</td>
<td>14:04:24:40</td>
<td>metadata</td>
</tr>
<tr>
<td>AR2</td>
<td>3</td>
<td>8</td>
<td>6.09GB</td>
<td>05:58:15:00</td>
<td>metadata</td>
</tr>
<tr>
<td>AS1</td>
<td>6</td>
<td>20</td>
<td>10.81GB</td>
<td>05:26:47:90</td>
<td>metadata</td>
</tr>
<tr>
<td>BE1</td>
<td>1</td>
<td>14</td>
<td>2.15MB</td>
<td>00:00:00:00</td>
<td>metadata</td>
</tr>
<tr>
<td>BH1</td>
<td>21</td>
<td>42</td>
<td>3.75GB</td>
<td>03:40:37:50</td>
<td>metadata</td>
</tr>
<tr>
<td>BM1</td>
<td>2</td>
<td>28</td>
<td>3.02GB</td>
<td>04:40:35:40</td>
<td>metadata</td>
</tr>
<tr>
<td>BP1</td>
<td>2</td>
<td>12</td>
<td>6.01GB</td>
<td>03:01:29:80</td>
<td>metadata</td>
</tr>
<tr>
<td>BP2</td>
<td>9</td>
<td>32</td>
<td>15.73GB</td>
<td>07:55:18:80</td>
<td>metadata</td>
</tr>
<tr>
<td>BP3</td>
<td>3</td>
<td>6</td>
<td>1.65GB</td>
<td>00:40:47:11</td>
<td>metadata</td>
</tr>
<tr>
<td>CR1</td>
<td>5</td>
<td>16</td>
<td>5.68GB</td>
<td>02:51:47:30</td>
<td>metadata</td>
</tr>
<tr>
<td>CH1</td>
<td>42</td>
<td>202</td>
<td>23.28GB</td>
<td>14:54:39:10</td>
<td>metadata</td>
</tr>
</tbody>
</table>
Data access (paradisec.org.au/repository)

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Duration</th>
<th>Class</th>
<th>Status</th>
<th>Metadata</th>
<th>Catalog Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT1-98002-98002A.mp3</td>
<td>37.39MB</td>
<td>00:40:50.19</td>
<td>audio</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-98002A.wav</td>
<td>413.37MB</td>
<td>00:40:55.74</td>
<td>audio</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-98002B.mp3</td>
<td>37.35MB</td>
<td>00:40:47.97</td>
<td>audio</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-98002B.wav</td>
<td>413.00MB</td>
<td>00:40:53.52</td>
<td>audio</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-98002az.xml</td>
<td>39.74KB</td>
<td></td>
<td>xml</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-98002bz.xml</td>
<td>46.42KB</td>
<td></td>
<td>xml</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-A.tab</td>
<td>23.65KB</td>
<td></td>
<td>unknown</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-A.xml</td>
<td>39.73KB</td>
<td></td>
<td>xml</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-A001.jpg</td>
<td>1.60MB</td>
<td></td>
<td>image</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-A002.jpg</td>
<td>1.65MB</td>
<td></td>
<td>image</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-A003.jpg</td>
<td>1.64MB</td>
<td></td>
<td>image</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-A004.jpg</td>
<td>1.68MB</td>
<td></td>
<td>image</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-A005.jpg</td>
<td>1.68MB</td>
<td></td>
<td>image</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98002-A006.jpg</td>
<td>1.66MB</td>
<td></td>
<td>image</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
</tbody>
</table>

Here's a listing of the files of item NT1/98002 viewable by your username nxt563. Refer to the repository website glossary for an explanation of terms and abbreviations used in the repository listings.
Reuse

Recording > Transcript

Preg nafkal skot na̱mēr nig Emlakul.

NT1-98002-A 316.43 320.317
Fight with the people from Malakula.
### English - South Efate

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>i</td>
<td>j</td>
<td>k</td>
</tr>
<tr>
<td>l</td>
<td>m</td>
<td>n</td>
<td>o</td>
<td>p</td>
<td>q</td>
<td>r</td>
<td>s</td>
<td>t</td>
<td>u</td>
<td>v</td>
</tr>
</tbody>
</table>

**A - a**

- abalone
- abort
- above
- absent
- accompany
- act
- active
- adam's apple
- add
- admire
- adopt
- adult
- adze
- aeroplane
- after
- afterbirth?
- afternoon
- again
- against
- age-mate
- agree
- aim, point with finger
- -

**South Efate**

This dictionary is a work in progress. This edition can be cited as: Nicholas Thieberger (2007-01-16) Dictionary of South Efate (http://paradisec.org.au/SELexicon/index-english/main.htm). This work is the result of a collaborative project between speakers of South Efate, mainly from Emotok village, Efate, Vanuatu, and Nicholas Thieberger who has written a grammar of the language. It was produced using Shoebox and LexiquePro.

### Examples

- **nakpe**
  - n: tree sp., young leaves eaten together with *ntali* fruit. Also use the wood to make fish traps by planting sticks vertically in the water to create a fence.

- **nakrik**
  - n: Fig tree. *Ficus virgata*.

- **nakrok**
  - n: vine sp. Strong vine.

- **nakse**
  - n: wild kava, creeping thick vine, leaf eaten, but not the root. Its leaves can stop a devil that is following you.

- **nakte**
  - pron. n: mine (1st person singular possessive).

- **nakur**
  - n: Indian Mulberry, not used as a dye today, but it was used as a yellow dye for mats and hair. Laverdière (1912) says it was used for sickness, and that it was mixed with lime to make a red paint. Also
nakser

wild_kava

wild kava, creeping thick vine, leaf eaten, but not the root. Its leaves can stop a devil that is following you.

info from William Wayane and Sailas Kalfapun 3/10/98
nakser

wild_kava

wild kava, creeping thick vine, leaf eaten, bush info from William Wayane and Sailas Kalfapa

CVC.CVC

n

n

nak.ser

2

19, film 3 11/4/00 JM

plants

3_21.jpg

NT1-98015-B 1896.74 1900.431

nakser.wav

[silence]

[lead in silence]

id 31 October, Endis reading words

sfiu

kat porpres npatin

p$oro

nrig

ntan

[ silence]

nafet

puk

kuom

nakser

p$au$pSal

ptom

tusuer

kat

urur

telekor

kuyaf

nqal

sisi

hair, Layari (1915) says it was used for sickness, and that it was mixed with lime to make a red paint. Also

dye for mats and
Online dictionary with sound and image links, constructed from well-formed underlying data
Waya Tenene, Story of Erakor and Mare

People from Mare came here in the time of the Colonial Government.
The French government sent them as police.
They came to be police on Efate.
On the French side, the French government had a fight in Malakula and it sent them to Malakula.
To fight the people of Malakula.
Those from Mare who came, those from New Caledonia who came because the French government called them.
The police came here.

Teni Emar non kin rumai pak san tetuei mal ni kolonial kafman.
Franis kafman isert kir rumai reki na polis.
Rumai pi polis ni Efai.
Nanne ni Frani, Franis kafman go ipi atiak nasifik nen iso Emalakul go itpekir pak Emalakul.
Preg nasifik skot nantier nig Emalakul.
Teni Emar non kin rumai, teni Caledoni non rumai pa omo naken Franis kafman isoSoor nafer.
Polis rumai pak esa.
Me, selwan ag kupi e.nama, tiawi itraus traus traus traus traus, natrauswen ga itaos nlag. Itrausi pan kaipa. Me komam uta laap kin uto mau, a? Malen umat, inom.

But when you are far away the old man can talk and talk and talk, his story is like the wind. He tells it and it is gone. But there aren't many of us left. When we die, it will be finished.

iTunes installation

Erakor school, Efate, Vanuatu
Example of reuse of interlinear glossed text (IGT)

Ipitlak nmatu inru, rato elag Eãuf.

3sRS= have  woman 3sRS= two 3d.exRS= stay above p.name

There are two women, they lived up at Bufa.

NT1-98004-A 210.34, 212.2
Interlinearising tools

TRANSC(ript) for CPM
IT for DOS and Macintosh in the 1980s
Shoebox and its successor Toolbox
ITE (Jacobson n.d.)
Fieldworks
TypeCraft

For a detailed listing of annotation tools see:
http://annotation.exmaralda.org/index.php/Linguistic_Annotation
Le contenu de l'archive (mode d'emploi / rechercher)

L'archive du LACITO ("Archivage") donne accès à des exemples de parole spontanée, prioritairement dans des langues « rares », enregistrés dans leur contexte social et transcrits en...

L'archive publique contient plus de 200 documents en 45 langues, annotés par une vingtaine de chercheurs (liste non exhaustive car ne comprenant pas les données en cours de numérisation).

- Afrique
- Balkans
- Caucase
- Océanie
- Népal et Asie
- Moyen-Orient et Amérique du Sud

Corpus de langues d'Afrique (présentation)

<table>
<thead>
<tr>
<th>Langue</th>
<th>Famille</th>
<th>Pays</th>
<th>Chercheur(s)</th>
<th>Archive</th>
</tr>
</thead>
<tbody>
<tr>
<td>grand-comorien</td>
<td>Bantu</td>
<td>Comores</td>
<td>Marie-Françoise Rombi</td>
<td></td>
</tr>
<tr>
<td>landu</td>
<td>Bantu</td>
<td>Tanzanie</td>
<td>Margaret Durham</td>
<td></td>
</tr>
<tr>
<td>mandari</td>
<td>Bantu</td>
<td>Tanzanie</td>
<td>Margaret Durham</td>
<td></td>
</tr>
<tr>
<td>mahora</td>
<td>Bantu</td>
<td>Comores</td>
<td>Marie-Françoise Rombi</td>
<td></td>
</tr>
<tr>
<td>mbʉguwe</td>
<td>Bantu</td>
<td>Tanzanie</td>
<td>Margaret Durham</td>
<td></td>
</tr>
<tr>
<td>nyilamba</td>
<td>Bantu</td>
<td>Tanzanie</td>
<td>Margaret Durham</td>
<td></td>
</tr>
<tr>
<td>oulédème</td>
<td>Tchadique</td>
<td>Cameroun</td>
<td>Véronique de Colombel</td>
<td></td>
</tr>
</tbody>
</table>

http://lacito.vjf.cnrs.fr/archivage/index.htm
susu valangi tadoma n iseki tayuwa
Nous les Valangi, nous allions dans la forêt pour ramasser
We Valangi went to the forest to collect

kur iseki dañakèsari yuwa ŋkowa nowalala kuko
ia-bas, dans la forêt DoWakèsari, pour tourner ia-bas
there, in the DoWakèsari forest, to turn there

kotayuwire ma takóda na kayi
Le ramassage fini, nous rentrions à la maison.
When the collecting was finished, we came back home.

Puis nous préparions le dîner.

http://lacito.vjf.cnrs.fr/archivage/tools/show_text.php?id=crdo-LAG_FER_SOUND
AILLA text / media formats

Why not use youtube?

Text / media link not provided for
No commitment to longevity
No path from linguistic fieldwork tools
   (especially interlinear glossed text)

Lack of control
   Ideally linguistic institutions would be curators
   IP issues not in control of the depositor
# Wit Go Kusu

**2006-11-27, Tokelau, erk, VU, ...**

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>date</td>
<td>2006-11-27</td>
</tr>
<tr>
<td>creator</td>
<td>Tokelau</td>
</tr>
<tr>
<td>contributor</td>
<td>recorder: Nick Thieberger</td>
</tr>
<tr>
<td>coverage</td>
<td>VU</td>
</tr>
<tr>
<td>subject</td>
<td>(erk)</td>
</tr>
<tr>
<td>language</td>
<td>erk</td>
</tr>
<tr>
<td>type</td>
<td>linguistic-type: Primary Text, discourse-type: Narrative</td>
</tr>
<tr>
<td>mediafile</td>
<td><a href="http://maenad.lite.uq.edu.au/exist/mediafiles/annodex/21110900.anx">http://maenad.lite.uq.edu.au/exist/mediafiles/annodex/21110900.anx</a></td>
</tr>
</tbody>
</table>

---

**Translation:**

1. **s1** Amurin gag puserek, Nick, kafo gag pusereki kusu go wit.  
   *I want to tell you, Nick, I’ll tell you about the rat and the octopus.*

2. **s2** Naliati iskei, elau imat.  
   *One day, it was low tide.*

3. **s3** Elau imat, itrau mat pe mat pe mat me.  
   *It was low tide, a really low tide.*

4. **s4** Ipiatlan nskau ienot nauru iskei  
   *There was a reef around an island.*

5. **s5** Naur ses iskei ito elau me elau imat  
   *This island was in the sea, but the tide was out.*

6. **s6** Malen elau imat kerkrak,  
   *When the tide was very low,*

7. **s7** mat top, go nskau igar userek  
   *very low, the reef was dry right around the island*

8. **s8** pan'Kal nau ses,  
   *right up to the small island.*

9. **s9** Me kusu, kusu iskei, ito Efat  
   *But the rat, this rat, it was from Efate.*
I want to tell you, Nick, I'll tell you about the rat and the octopus.
Wit Go Kusu

2006-11-27, Tokelau, erk, VU, ...

I want to tell you, Nick, I'll tell you about the rat and the octopus.

One day, it was low tide.

There was a reef around an island.

This island was in the sea, but the tide was out.

When the tide was very low,
amurin

<morphemes>
  <morpheme>
    <text xsi:type="morpheme">a</text>
    <text xsi:type="gloss">1sgRS=</text>
  </morpheme>
  <morpheme>
    <text xsi:type="morpheme">mur</text>
    <text xsi:type="gloss">want</text>
  </morpheme>
  <morpheme>
    <text xsi:type="morpheme">-i</text>
    <text xsi:type="gloss">-TS</text>
  </morpheme>
  <morpheme>
    <text xsi:type="morpheme">-n</text>
    <text xsi:type="gloss">-3sgO</text>
  </morpheme>
</morphemes>
We need a standard schema for IGT

No schema (standard representation) for IGT

So ODIN relies on heuristics to locate IGT on the web (as the basis for making typological claims)

http://www.csufresno.edu/odin/

Inferring the nature of IGT from contextual clues
Eopas sample converted from Toolbox XML

```
<database>
<itmgroup>
<itm>105</itm>
<idgroup>
  <id>001</id>
  <aud>200518.aud 20.507 27.202</aud>
  <txgroup>
    <tx>Amurin</tx>
    <mr>a=</mr>
    <mg>1sgRS=</mg>
    <mr>mur</mr>
    <mg>want</mg>
    <mr>-i</mr>
    <mg>-TS</mg>
    <mr>-n</mr>
    <mg>-3sgO</mg>
  </txgroup>
  <fg>I want to tell you a story.</fg>
</idgroup>
```

```
<Interlinear-text>
<phrases>
  <phrase id="s1" startTime="20.507" endTime="27.202">
    <text xsi:type="translation">I want to tell you a story</text>
    <text xsi:type="orthographic">Amurin na kagag</text>
    <words>
      <word>
        <text xsi:type="orthographic">amurin</text>
        <morphemes>
          <morpheme>
            <text xsi:type="morpheme">a=</text>
            <text xsi:type="gloss">1sgRS=</text>
          </morpheme>
          <morpheme>
            <text xsi:type="morpheme">mur</text>
            <text xsi:type="gloss">want</text>
          </morpheme>
          <morpheme>
            <text xsi:type="morpheme">-i</text>
            <text xsi:type="gloss">-TS</text>
          </morpheme>
          <morpheme>
            <text xsi:type="morpheme">-n</text>
            <text xsi:type="gloss">-3sgO</text>
          </morpheme>
        </morphemes>
      </word>
    </words>
  </phrase>
(...)
```
Uploading...

Browsing...

### EthnoER: online presentation and annotation system

Below is a list of the transcripts currently stored in the database. Click on the table header to sort or on the various values to filter the list. Clicking the title will open the transcript.

<table>
<thead>
<tr>
<th>date</th>
<th>creator</th>
<th>title</th>
<th>lang</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>Nick Thieberger</td>
<td>Koaïsono</td>
<td>erk</td>
<td>VU</td>
</tr>
<tr>
<td>2003</td>
<td>Nick Thieberger</td>
<td>Agoli Ermag</td>
<td>erk</td>
<td>VU</td>
</tr>
<tr>
<td>2005</td>
<td>Tekelau Takau</td>
<td>Witi Go Kusu, the octopus and the rat</td>
<td>erk</td>
<td>VU</td>
</tr>
<tr>
<td>1999</td>
<td>Nick Thieberger</td>
<td>Litap story</td>
<td>erk</td>
<td>VU</td>
</tr>
<tr>
<td>1998</td>
<td>Tekelau Takau</td>
<td>Natopo koro, 'Another spirit'</td>
<td>erk</td>
<td>VU</td>
</tr>
<tr>
<td>1996</td>
<td>Nick Thieberger</td>
<td>English Police</td>
<td>erk</td>
<td>VU</td>
</tr>
<tr>
<td>2005</td>
<td>Tekelau Takau</td>
<td>Witi Go Kusu, the octopus and the rat</td>
<td>erk</td>
<td>VU</td>
</tr>
<tr>
<td>1998</td>
<td>Nick Thieberger</td>
<td>Litapu ray</td>
<td>erk</td>
<td>VU</td>
</tr>
<tr>
<td>1990</td>
<td>erk</td>
<td>VU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001-11-10</td>
<td>John Bowden</td>
<td>Tujuh burung, Seven birds</td>
<td>wrnh</td>
<td>TL</td>
</tr>
<tr>
<td>2006-09-29+10:00</td>
<td>Mailei, Kalfapūn</td>
<td>Flesaur and Flepūg, two natopu</td>
<td>erk</td>
<td>VU</td>
</tr>
</tbody>
</table>

Done
Viewing and concordancing...
login functionality: authenticated access to upload/edit/download
upload functionality of video/audio
transcoding video/audio to Ogg Vorbis and Ogg Theora formats for delivery
storage of video/audio
download video/audio
upload functionality of text (ELAN, Transcriber, IGT from Toolbox)
handling of unicode and other i18n characters
validation of text
transcoding text
storage of text
download text in different formats (ELAN, pdf, rtf)
aligned display of text with video/audio
search
citability on the level of syllables linked to video/audio
editing (but not authoring!) of text
open source publication of system
creation of project in Sourceforge
Amurin gag puserek, Nick, kafo gag pusereki kusu go wit. I want to tell you, Nick, I'll tell you about the rat and the octopus. [...]

[...]
Confluence

Several projects are now aiming to produce textual material in small languages, e.g.,

WOLP & Firebird

PARADISEC

AILLA

Potentially each archive and funding project will need to present texts and media
Central service harvesting text collections for search and display
Conclusion

Need for standards for encoding textual material as interlinear text linked to media

Ideally building these standards into a normal linguistic workflow

Creating multiple outputs from a single research effort

Exciting possibilities for return of material and for representation of small languages
How do we catch the wind?

Me, selwan ag kupi eᵐae, tiawi itraus traus traus traus, natrauswen ga itaos nlag. Itrausi pan kaipa. Me komam uta laap kin uto mau, a? Malen umat, inom.

But when you are far away the old man can talk and talk and talk, his story is like the wind. He tells it and it is gone. But there aren't many of us left. When we die, it will be finished.

†Kalfaⁿun Mailei, 1998, Erakor Village, Efate, Vanuatu