

# Building a Universal Corpus of the World's Languages

Steven Bird

Associate Professor  
Department of Computer Science, University of Melbourne

Senior Research Associate  
Linguistic Data Consortium, University of Pennsylvania

## Three questions

1. How can I find language resources? (OLAC) [10 mins]
2. How can I use language resources? (Universal Corpus) [20 mins]
3. What about for unwritten languages? (MT4LP) [10 mins]
  - *the special challenge of tone languages (PNG Prosody)*
4. How can I access structured language resources?
  - *efficient query for large linguistic databases (tree search)*
  - *high level programming with linguistic data (NLTK)*



## 1. OLAC: Open Language Archives Community

Search  
109k items  
from 45  
archives.

Participating Archives • OLAC • Delivered by the Net Library

OLAC Language Resource Catalog

Search for language resources

Sort by:

- Possible Sorts: all
- Title [4<33>4]
- Id [4<33>4]
- Date [4<33>4]

Browse by:

- Archive browse
- SIL Language and Culture Archives 22377
- IMDI to OAI bridge 15370
- Alaska Native Language Archive 11141
- Graduate Institute of Applied Linguistics Library 8176
- view more...

Online browse

- No 56255
- Yes 49709

This catalog, developed by the Open Language Archives Community (OLAC), provides access to a wealth of information about thousands of languages, including details of text collections, audio recordings, dictionaries, and software, sourced from dozens of digital and traditional archives.

Browse the OLAC records by Geographic region or by Language:

- English (3521)
- Spanish (2925)
- Yucatec (1383)
- Aleut (1125)
- Central Yupik (1116)
- Beaver (1075)
- North Alaskan Inupiatun
- Quech'In (841)
- Bora (748)
- Chiricango (703)
- Ojama (678)
- Altana (618)
- Nepali (603)
- Tawana (588)

[search.language-archives.org](http://search.language-archives.org)



## 1. OLAC: Open Language Archives Community

- language id guesser
- mined 459 institutional repositories
- 90% recall, 70% precision (guessing ISO 639 code)
- Ainu, Alutiq (Yupik), Alutor (Russia), Basque, Faroese, Frisian, Gothic, Hawaiian Creole English, Inuktitut, Itonama (Bolivia), Marathi, Middle High German, Navajo, Occitan, Pitcairn English, Tibetan, Tausug (Philippines), Toba Batak (Indonesia), Yapese (Micronesia)

*metadata ≠ data*



Digitization at the  
Institute for PNG Studies



## Three questions

---

1. How can I find language resources? (OLAC) [10 mins]
2. **How can I use language resources?** (Universal Corpus) [20 mins]
3. What about for unwritten languages? (MT4LP) [10 mins]
  - *the special challenge of tone languages (PNG Prosody)*
4. How can I access structured language resources?
  - *efficient query for large linguistic databases (tree search)*
  - *high level programming with linguistic data (NLTK)*

The Problem of Language Documentation

Half the world's  
languages have fewer  
than 10,000 speakers

37 years of unique indigenous language recordings  
destroyed by accident when govt. building was demolished  
May 2010, Papua New Guinea



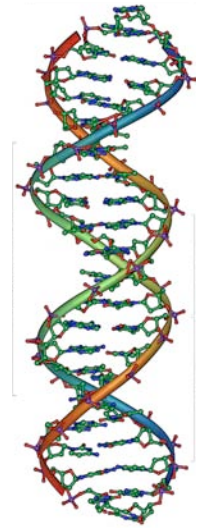
The Grand Aim of Linguistics

0.5%

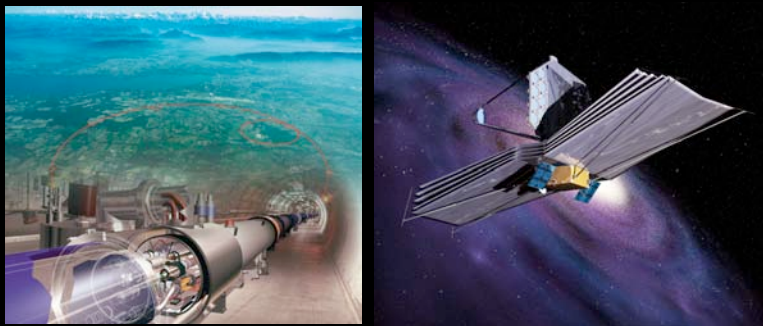
fraction of the world's  
languages addressed by  
"empirical" NLP

## A Universal Corpus

One of the greatest impacts of having the sequence may well be in enabling an entirely new approach to biological research. In the past, researchers studied one or a few genes at a time. With whole-genome sequences ... they can approach questions systematically and on a grand scale. They can study ... how tens of thousands of genes and proteins work together in interconnected networks to orchestrate the chemistry of life.



What grand vision do we have to match the Large Hadron Collider or the James Webb Space Telescope?



## Universal Corpus

1. all languages
2. language-universal processing

## "Documentary Linguistics"

The creation, annotation, preservation and dissemination of transparent records of a language.

-- Tony Woodbury



~180 words per minute  
× 100 hours  
=  
One million words

0.5%

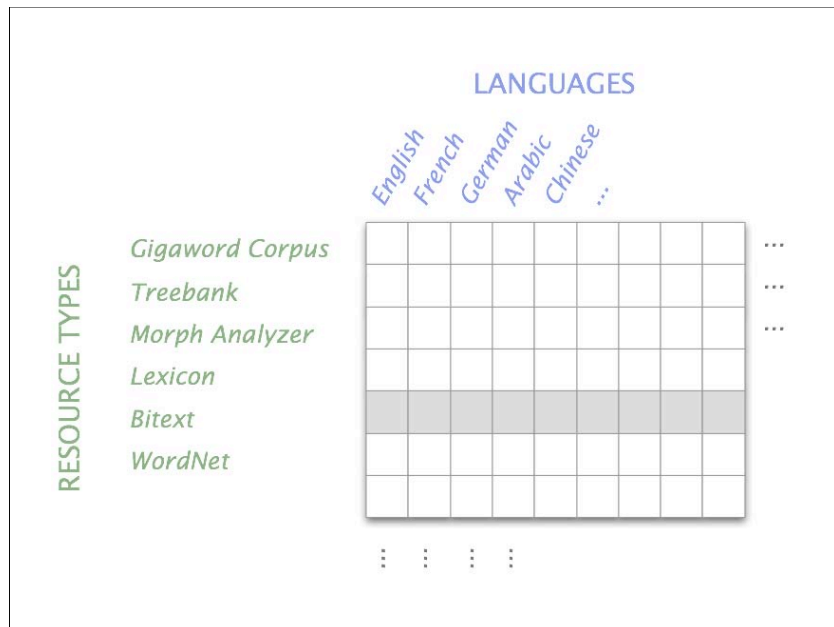
progress towards  
constructing a one million  
word corpus per language

What counts as a  
*Complete Digitization*  
of a language?

We need  
a treebank  
for every  
language!



We have successfully  
captured a language if we  
can translate into and out of  
the language



## Defining the Universal Corpus: *Principles*

---

1. universality
2. machine readability and consistency
3. community effort
4. availability
5. utility
6. centrality of primary data

## Defining the Universal Corpus: *What to include*

---

- metadata
- for written text:  
*primary documents & transcriptions*
- for spoken text:  
*audio recordings & respeaking*
- written transcriptions
- for both written and spoken text:  
*translations, alignments, glosses*

Data Model for a  
Universal Corpus

## Defining the Universal Corpus

### High level view

Aligned Texts					Analyzed Texts													
	deu	spa	fra	eng	deu													
					sent	form	lemma	morph	pos	gloss	head	rel						
$d_1$	sie..	ella..	elle..	she..														
$d_2$					$w_1$	$s_1$	Kühe	Kuh	PL	N	cow	2	SBJ					
:					$w_2$	$s_1$	sind	sein	PL	V	be	0	ROOT					
:																		
$s_1$																		
$s_2$																		
:																		
:																		
$w_1$																		
$w_2$																		
:																		
:																		

## Defining the Universal Corpus:

### Tuple storage (e.g. AWS SimpleDB)

ID: europarl/swedish/ep-00-01-17/18  
 LANGS: swd eng  
 SENT: det gäller en ordningsfråga  
 TRANS: this is a point of order  
 ALIGN: 1-1 2-2 3-3 4-4 4-5 4-6  
 PROVENANCE: pharaoh-v1.2, ...  
 REV: 8947 2010-05-02 10:35:06 leobfld12  
 RIGHTS: Copyright (C) 2010 Uni...; CC-BY

## Defining the Universal Corpus:

### Morphological Annotations

ID: example/001  
 LANGS: eng  
 SENT: the dogs are barking  
 LEX: the dog be bark  
 AFF: - PL PL ING

Gila	abur-u-n	ferma	hamišaluğ	güğüna	amuq'-da-č
now	they-OBL-GEN	farm	forever	behind	stay-FUT-NEG
<i>Now their farm will not stay behind forever.</i>					



## Defining the Universal Corpus: *Lexicons*

---

ID: swedishlex/v3.2/0419  
LANGS: swd eng  
LEX: ordningsfråga  
TRANS: point of order

## Defining the Universal Corpus: *multi-language resources*

---

ID: swadesh/47	ID: swadesh/47
LANGS: fra	LANGS: eng
LEX: chien	LEX: dog

## Defining the Universal Corpus: *A massive store of records*

---

- **bilingual text**

ID: europarl/swedish/ep-00-01-17/18  
LANGS: swd eng  
SENT: det gäller en ordningsfråga  
TRANS: this is a point of order  
ALIGN: 1-1 2-2 3-3 4-4 4-5 4-6  
PROVENANCE: pharaoh-v1.2, ...  
REV: 8947 2010-05-02 10:35:06 leobfld12  
RIGHTS: Copyright (C) 2010 Uni...; CC-BY

- **bilingual lexicons**

ID: swedishlex/v3.2/0419  
LANGS: swd eng  
LEX: ordningsfråga  
TRANS: point of order

- **morphologically analyzed text**

ID: example/001  
LANGS: eng  
SENT: the dogs are barking  
LEX: the dog be bark  
AFF: - PL PL ING

- **comparative wordlists**

ID: swadesh/47	ID: swadesh/47
LANGS: fra	LANGS: eng
LEX: chien	LEX: dog

Building the Corpus



# Language Commons

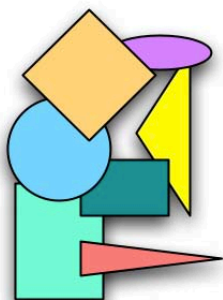
- open access repository
- hosted in the Internet Archive
- lightweight method to contribute



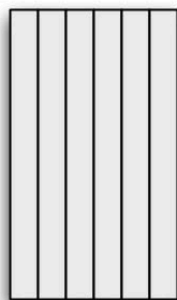
The screenshot shows the Language Commons page on the Internet Archive. The page has a navigation bar with links for Home, Donate, Forums, FAQs, Contributions, Terms, Privacy, & Copyright, Contact, Jobs, and Blog. Below the navigation bar is a search bar and a dropdown menu for 'All Media Types'. The main content area is titled 'language commons' and features several sections: 'Spotlight Item' (Brown Corpus), 'Welcome to language commons' (8 items), 'Most Downloaded Items Last Week' (1. Brown Corpus 18 downloads, 2. Genesis Corpus 4 downloads, 3. Tone in Usanufa: Field Recordings 2 downloads, 4. Chapter 3 of Arabic Phonetics 1 download, 5. Manx Gaelic-Irish lexicon 1 download), 'About the Internet Archive', 'Recently Reviewed Items (more)', and 'This Just In (more)' (Chapter 3 of Arabic Phonetics, Manx Gaelic-Irish lexicon, Tone in Usanufa: Field Recordings).

<http://www.archive.org/details/LanguageCommons>

## Language Commons



## Universal Corpus



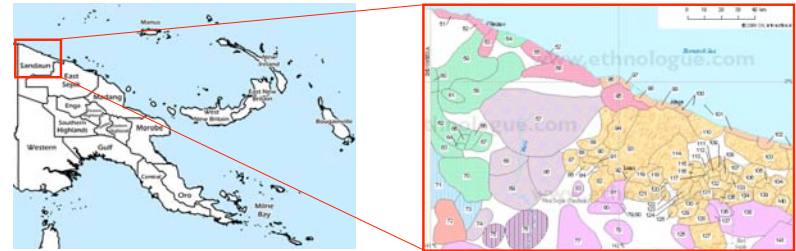
## Three questions

1. How can I find language resources? (OLAC) [10 mins]
2. How can I use language resources? (Universal Corpus) [20 mins]
3. **What about for unwritten languages?** (MT4LP) [15 mins]
  - *the special challenge of tone languages (PNG Prosody)*
4. How can I access structured language resources?
  - *efficient query for large linguistic databases (tree search)*
  - *high level programming with linguistic data (NLTK)*



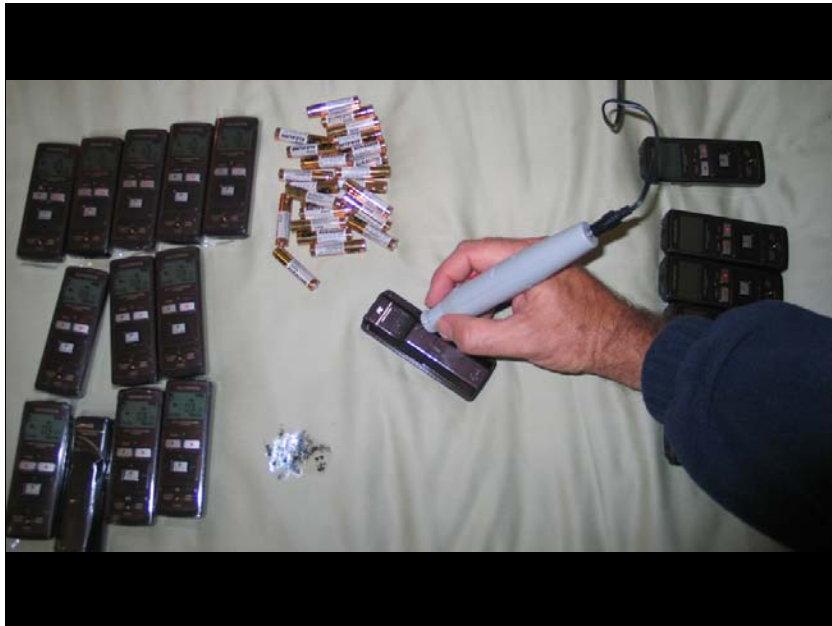
Languages of the World

6,900 distinct languages



Papua New Guinea

Home to >800 languages  
ethnologue.com



## Record natural speech

- genres
- location
- equipment...



## Oral Annotation Protocol

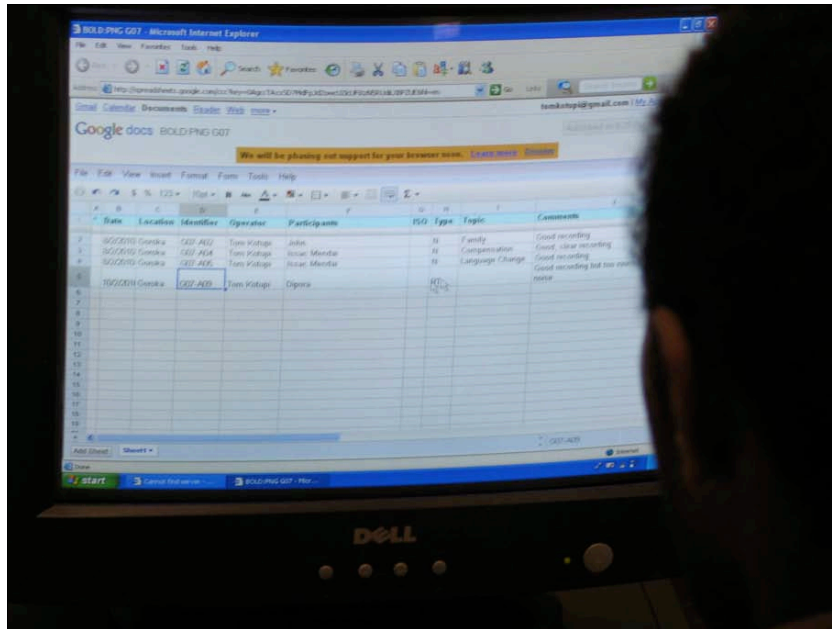


## Logbooks (metadata)

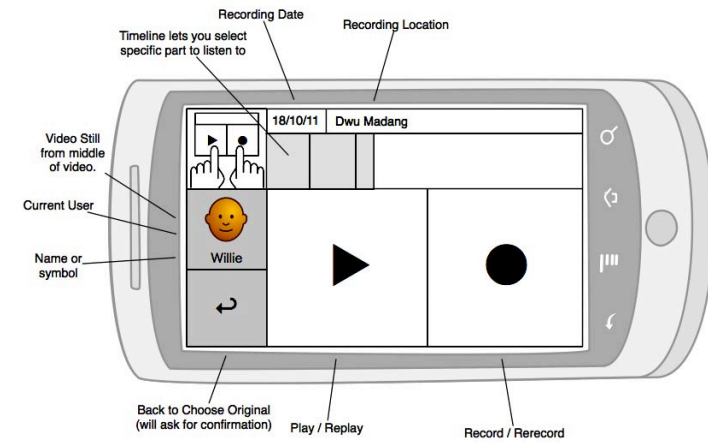


Date	Loc	Identifier	Operator	Participants	Type	Topic	Comments
27/09/10	Poaba, Lae	026-001	Oris IINDIKA	Mr Hendrika Batawäng	N	Legend	- Recording was ok - Permission granted - Oral Translation - done - Transcribing - done.
28/09/10	Poaba, Lae	026-002	✓	Mrs Zanigo Batawäng	P	How to make Blum	- Recording ok - Permission granted - Oral Translation - done - Transcribing - done
29/09/10	Poaba, Lae	026-003	✓	Mr Gann Aze	R	Yam story	- Recording ok - Permission granted - Oral Translation - done - Transcribing - done
30/09/10	Poaba, Lae	026-04	✓	Mrs Zanigo Batawäng	I	Non-verbal Form of communication	- Recording ok - Permission granted - Oral Translation - done - Transcribing - done.





BOLD App for Android Phones  
<http://bold.xpdev-hosted.com/>



## Transcription



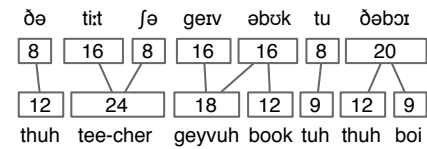


# Writers workshops

## Transcription normalisation

joint work with Adel Foda (University of Melbourne)

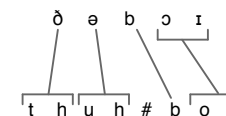
(a) Word alignment



(b) Word pairs

ðə	thuh
ti:t#fə	tee-cher
gɛrv#əbʊk	geyvuh#book
tu	tuh
ðəbɔɪ	thuh#boi

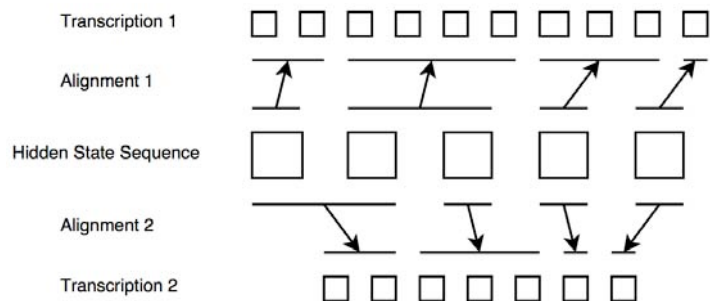
(c) Character alignment



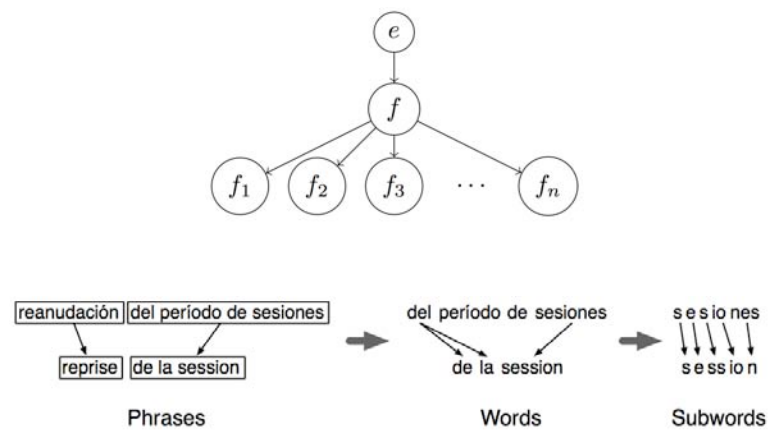
(d) Char mappings

b	b
ð	th
ə	uh
i:	ee
ɔɪ	oi

## Transcription normalisation

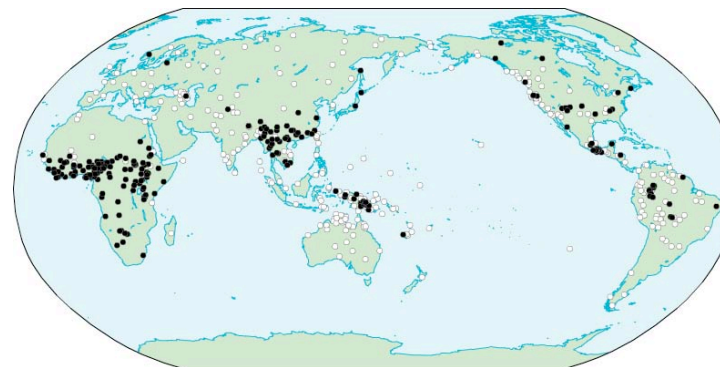


## Cluster-based approach to MT for small languages *joint work with David Chiang (Information Sciences Institute, USC)*



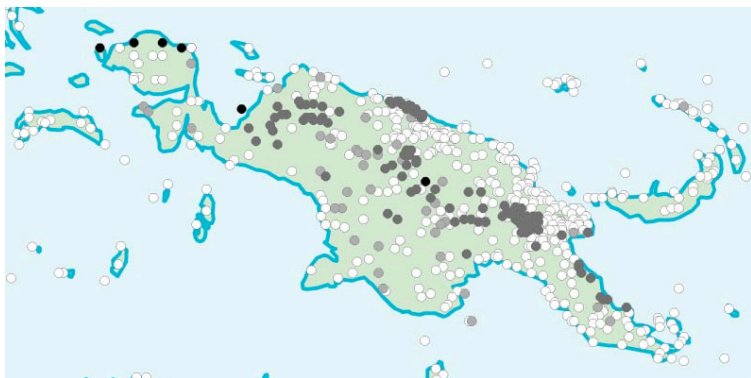
The special problem of tone languages

## Distribution of tone languages



## Tone in New Guinea

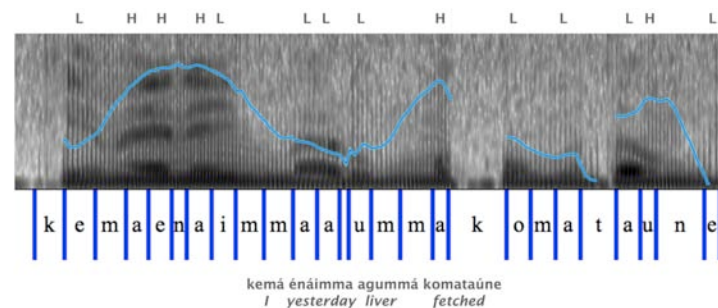
---



white=non-tonal, light gray=pitch accent, dark gray=lexical tone

## Modeling tone

---



## Three questions

---

1. How can I find language resources? (OLAC) [10 mins]
2. How can I use language resources? (Universal Corpus) [20 mins]
3. What about for unwritten languages? (MT4LP) [10 mins]
  - *the special challenge of tone languages (PNG Prosody)*
4. How can I access structured language resources?
  - *efficient query for large linguistic databases (tree search)*
  - *high level programming with linguistic data (NLTK)*

Thank you