The Rise of Documentary Linguistics and a New Kind of Corpus

Partners in
Language Development

Gary F. Simons
SIL International

5th National Natural Language Research Symposium
De La Salle University, Manila, 25 Nov 2008

Milestones in corpus linguistics

- 1960s — Brown Corpus of American English
  - 1 million words from a variety of sources
  - With part of speech tagging
- 1970s — Thesaurus Linguae Graecae
  - ~50 million words of Classical Greek literature
- 1980s — COBUILD “Bank of English”
  - Now over 500 million words
- 1990s — Text Encoding Initiative
  - Guidelines for the XML markup of the structure, analysis, and interpretation of text

Spoken corpora

- Digital audio has enabled a new genre of “spoken corpora”; they add recordings to the elements familiar in written corpora, e.g.
  - Paul Thompson, “Spoken Language Corpora.” Chapter 5 in Wynne, M (editor). 2005. *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Available online at http://ahds.ac.uk/linguistic-corpora/
- Stages in developing a spoken corpus
  1. Data collection
  2. Transcription
  3. Markup and annotation
  4. Access

The problem

- Linguists concerned with languages in general (not just major languages) have encountered a problem:
  - Forces of globalization are causing small languages to die out faster than linguists can build conventional corpora to document them.
- What should we be doing in response to this?

Overview of talk

1. Language endangerment as a current issue
2. The rise of documentary linguistics (as distinct from descriptive linguistics) as a response of the linguistics community
3. The emergence of a new kind of corpus

Endangerment hits the radar

- In 1992, *Language* had a special issue on Endangered Languages. Lead article was:
  - Krauss, Michael. 1992. “The world’s languages in crisis,” *Language 68(1):*4-10.
- USA/Canada: 149 of 187 languages were NO longer learned by children (80% were moribund)
  - Australia: 90% were moribund
- Unless we do something:
  - “The coming century will see the death or the doom of 90% of mankind’s languages.”

Endangered languages

- What is an endangered language?
  - *No* — One that is on the verge of extinction
  - *Yes* — Any language for which there is a possibility that parents will no longer be passing it on to their children at the end of this century
- A language can be in common use among children today, but still it is endangered if there are pressures (esp. economic) that could cause language shift within 100 years

An emerging consensus

- Crystal, David. 2000. *Language Death.* Cambridge: Cambridge University Press.
  - Even with a more conservative estimate of 50% loss in 21st century, that is still one language disappearing every 2 weeks.
- A general consensus:
  - 50% of languages are likely to die
  - Another 40% are in danger of dying
  - Only 10% of languages are truly safe

SIL International

Languages by size

Population range	Living languages		Number of speakers	
	Count	*Percent*	*Count*	*Percent*
Over 1,000,000	347	5.0%	5,373,702,347	93.9%
100,000 to 999,999	892	12.9%	283,651,418	4.95%
10,000 to 99,999	1,779	25.7%	58,442,338	1.02%
1,000 to 9,999	1,967	28.5%	7,594,224	0.13%
1 to 999	1,619	23.4%	470,883	0.008%
Unknown	308	4.5%		
Totals	6,912	100.0%	5,723,861,210	100.0%

From *Ethnologue* 15th ed, “Summary by Language Size”

Safe, Endangered, Dying

Population	*Languages of the World*		*Languages of the Philippines*	
> 300,000	688	*10%*	18	*11%*
6,000 to 300,000	2,774	*40%*	105	*64%*
< 6,000	3,450	*50%*	42	*25%*
	6,912		165	

Based on population data from *Ethnologue* 15th ed

Why does it matter?

- The scientific significance
  - Huge loss of data for typology, reconstruction
  - Unique knowledge is lost (e.g. ethnobotanical)
- The social significance
  - When we lose a language and culture, we lose a significant window on human experience
  - As a people's identity and cultural knowledge are eroded by language loss, the fabric of society begins to unravel
  - People in the process of losing their language often have a higher incidence of social problems

Overview

1. *Language endangerment as a current issue*
2. The rise of documentary linguistics (as distinct from descriptive linguistics) as a response of the linguistics community
3. *The emergence of a new kind of corpus*

The community responds

- Documenting endangered languages has become a mainstream issue
- It has become the focus of
  - Conferences, symposia, conference sessions
  - New degree programs and summer institutes
  - New endowed chairs
- Major funding programs:
  - Volkswagen Foundation: DOBES project
  - Hans Rausing: Endangered languages project, SOAS
  - NSF & NEH: Documenting Endangered Languages

SIL International

Traditional practice

- Field linguistics was born in the age of ***descriptive linguistics.***
- The products are a phonology, grammar, lexicon, and corpus of interlinear text.
- These are secondary data based on the analysis of primary data.
- The primary data (*e.g.,* the actual speech events) are not a product; only a means to the end of description.

A new mainstream

- Today ***documentary linguistics*** is on the rise.
- The product is the primary data — a corpus of recorded speech events that document the language in actual use.
- Uses both audio and video recordings.
- Not an alternative to description, but a complement to it.

The seminal work

- Definitive source on documentation *vs* description:
  - Nikolaus Himmelmann, 1998. "Documentary and descriptive linguistics." *Linguistics* 36:165–191.
- Definitions
  - Documentation is "the activity concerned with collecting, transcribing, translating, and commenting on primary data" (190) [+archiving]
  - Aim is "to provide a comprehensive record of the linguistic practices characteristic of a given speech community." Contrasts with description which aims at "the record of a language … as a system of abstract elements, constructions, and rules." (166)

Other key works

- Woodbury, Anthony. 2003. “Defining documentary linguistics.” In Peter Austin (ed.), *Language Documentation and Description 1.* HRELP, SOAS.
- Bird, Steven and Gary Simons. 2003. “Seven dimensions of portability for language documentation and description.” *Language* 79:557-582.
- Recent textbook published by Mouton de Gruyter:
  - Gippert, Jost, Nikolaus P. Himmelmann, and Ulrike Mosel (eds.). 2006. *Essentials of Language Documentation.*

The three basic tasks

"Language Documentation is concerned with compiling, commenting on, and archiving language documents." — Himmelmann

1. ***Compile*** a sample of recordings of a full range of speech event types
2. ***Comment*** on those recordings
   - E.g., transcription, translation, discussion, situational context, informed consent to share
3. ***Archive*** the complete corpus of recordings and commentary with an institution that will provide long-term access

SIL International

Documentation vs. Description

	Documentation	***Description***
What?	Primary data	Secondary data
How?	Observe, Record, Transcribe, Translate	Analyze, Generalize
Who?	Recording specialists, Literate speakers	Professional linguists
Where?	On site	On or off site
When?	Short term	Long term

A call to action

- The situation is urgent:
  - With one language being lost every two weeks, there are not enough linguists coming forward to preserve records of those languages using the traditional descriptive approach.
- Many linguists recognize a new top priority:
  - We must document languages before they are gone forever.
  - We can describe them later using the archived documentation.
- The urgency demands a new approach.

SIL International

Pointing the way

- Woodbury (2003:45) proposes that one could start the documentation process with purely oral techniques:
  - In place of written translation, producing “running UN style translations”
  - In place of written transcription, “starting with hard-to-hear tapes and asking elders to ‘respeak’ them to a second tape slowly so that anyone with training in hearing the language can make the transcription if they wish.”

Overview

1. *Language endangerment as a current issue*
2. *The rise of documentary linguistics (as distinct from descriptive linguistics) as a response of the linguistics community*
3. The emergence of a new kind of corpus

Developing a BOLD approach

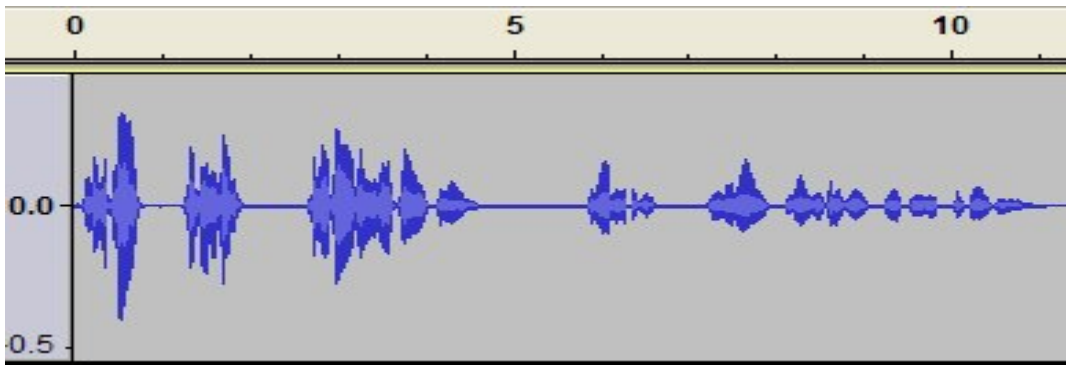
- A team at SIL is working on a method we call:
  - **B**asic **O**ral **L**anguage **D**ocumentation
- In place of the traditional spoken corpus:
  - Compile, Transcribe, Markup/annotate, Archive
- We build an oral documentation corpus:
  - Compile, Comment orally, Archive
- A linguist may be the catalyst, but non-linguists (e.g. community members) can be mobilized to do the work of compiling and commenting

The form of the corpus

- A well-formed BOLD corpus contains:
  - A document introducing the language, people, project, coverage, methods
  - A table of contents listing each item
  - A set of fully commented items
- A fully commented item consists of:
  - Recording
  - Informed consent
  - Situational metadata
  - Oral transcription
  - Oral translation

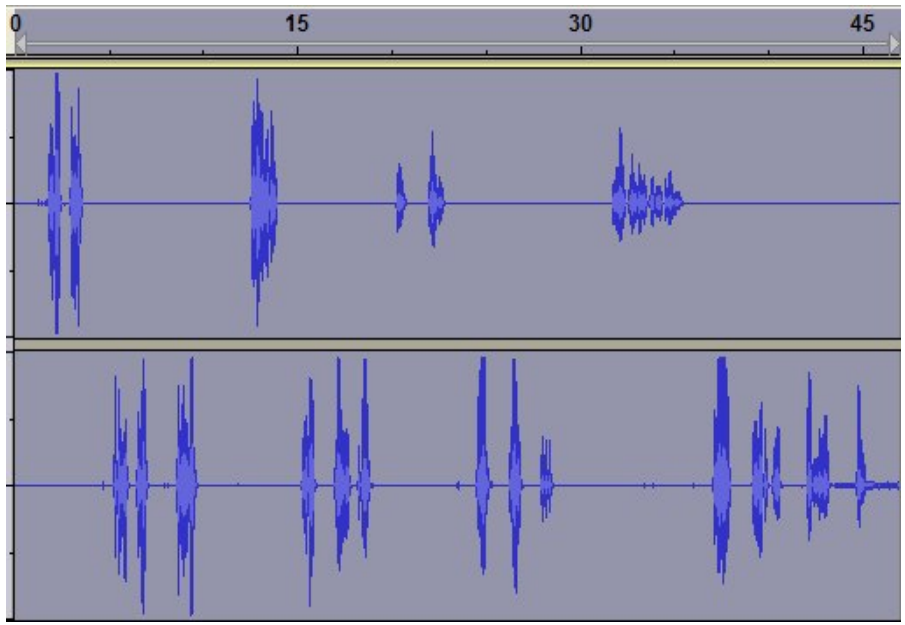
Commenting: An example

- Field testing by Will Reiman (SIL) in Guinea-Bissau
- For instance, a sample from a recorded communicative event in Kasanga [ccj], an endangered language of the Niger-Congo family with only 650 speakers



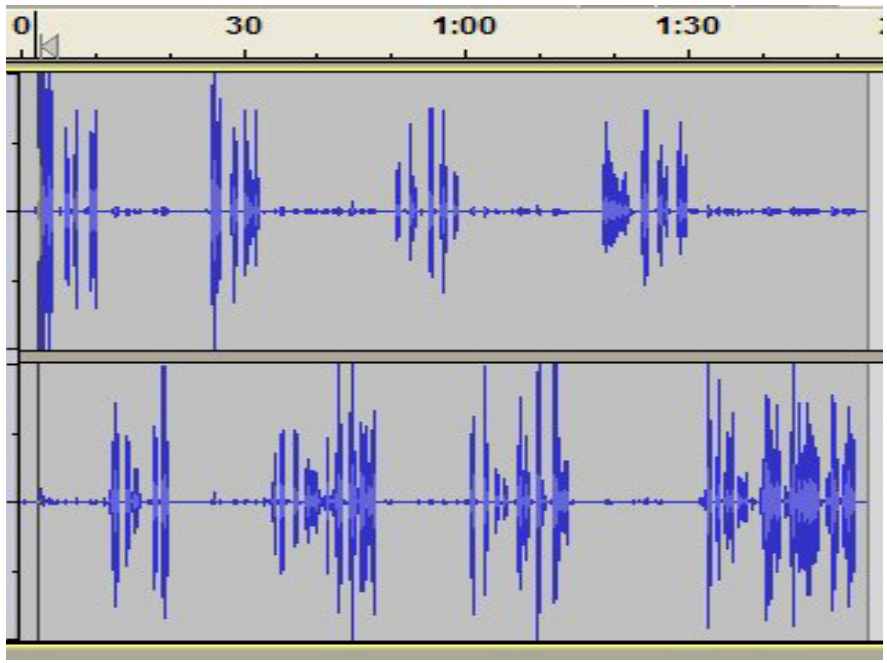
Oral transcription

- In a field “studio”, transcriber hits pause button on original recording at natural breaks, then repeats the segment slowly and carefully
- Original recording fed into left channel
- Oral transcription recorded on right channel



Oral translation

- Follows the same process.
- In this example, two translators participated: first Portuguese Creole, then English
- Original + oral transcription on left channel
- Oral translations recorded on right channel

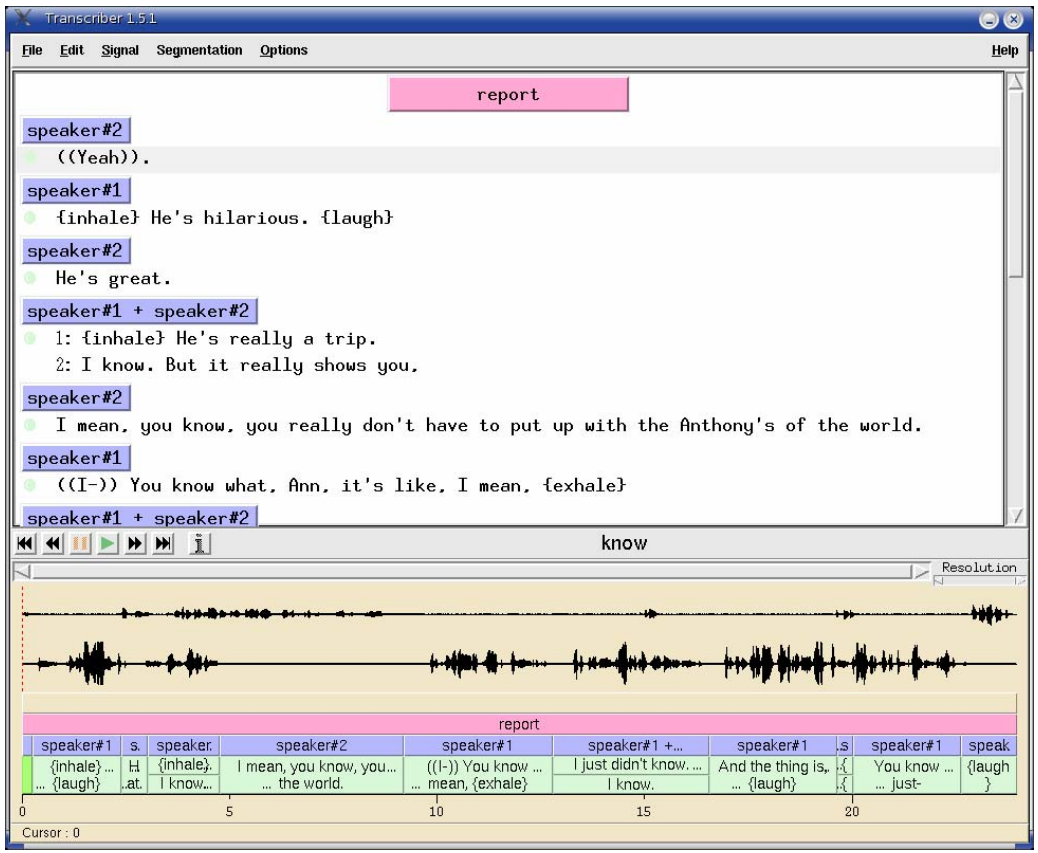


Adding written transcription and translation

- The oral transcription and translation will serve as the source for written transcription and translation
  - Either done immediately and added to the documentary corpus
  - Or done later (even as a different project by different people) as the basis for a new descriptive corpus with links back to the sources in the documentary corpus

Audio transcription

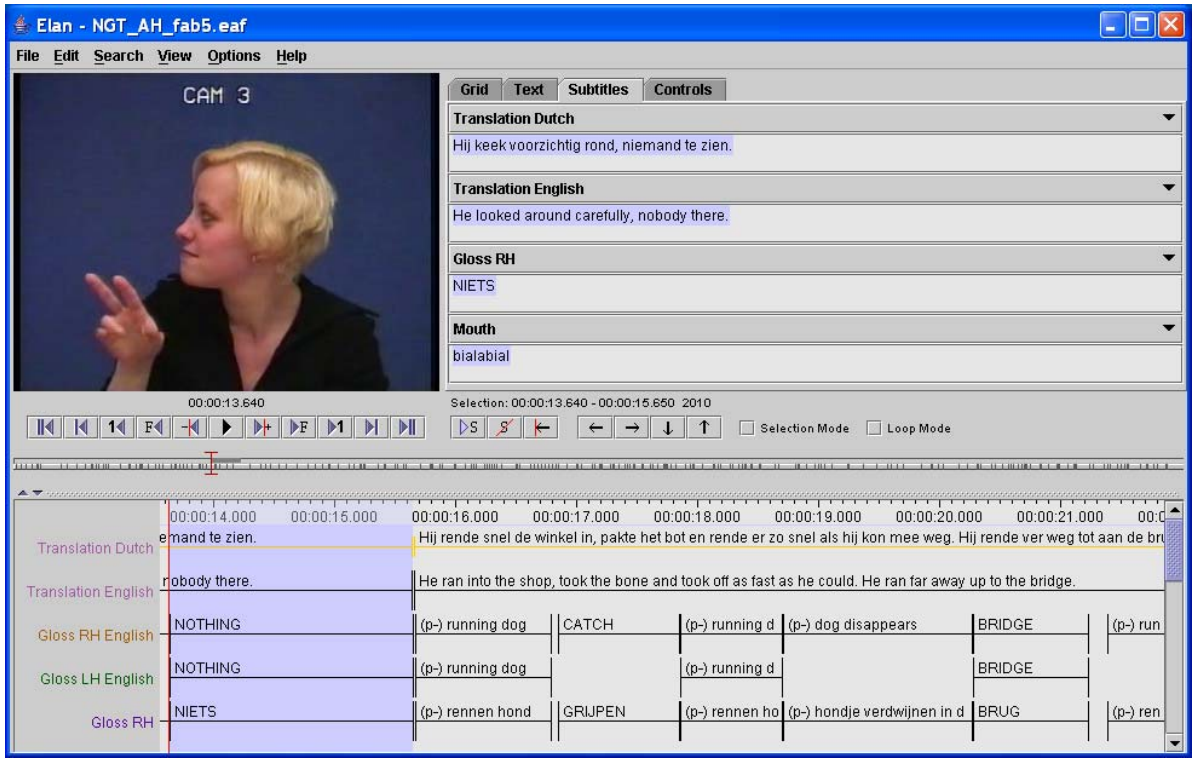
- The most widely used tool is Transcriber
  - Open source: http://trans.sourceforge.net/



- Needed: automate creation of alignment points and audio segments from channel switching in left-right separated input

Video transcription

- The most widely used tool is ELAN
  - Max Planck Institute, Nijmegen
  - http://www.lat-mpi.eu/tools/elan/



Compiling: The breadth of the corpus

Three kinds of items (Himmelmann)

- Communicative events
  - Includes all forms of normal use of the language
  - The corpus should sample a full range
- Elicited lists
  - Standardized word lists
  - Semantic sets like numbers, colors, living things
  - Paradigms of grammatical categories
- Analytical discussions
  - Discussion guided by researcher about the language
  - Conducted in a language of wider communication, so do not require transcription or translation.

Sampling a full range of events

- Begin with a universal grid to sample a full cross-section of event types
  - Events can be classified on a scale of unplanned to planned
  - Exclamations, greetings, small talk, discussion, interview, autobiographical narrative, procedure, speech, folk tale, litany
- Elicit the insider's grid for further sampling
  - Discover the language's own taxonomy for communicative events and get samples of each kind

Other sampling dimensions

- The choice of speakers should involve sampling as well. Universally applicable:
  - Gender
  - Age
- Relevant in some situations:
  - Social stratum
  - Education level
- A large corpus could also sample regional varieties

Compiling: The depth of the corpus

- How big does the corpus need to be?
- Depends on the purpose
  - For historical reconstruction:
    - 100s to 1000s of lexical items
  - For a basic descriptive grammar:
    - At least 100,000 words of running text
  - For good lexicography:
    - Millions of words of running text

Corpus size vs. time

- Speaking speeds: 100 – 200 words per min.
  - 167 w.p.m. = 10,000 words per hour, thus
  - 10 hours = 100,000 word corpus
  - 100 hours = 1,000,000 word corpus
- We currently estimate:
  - It takes 12 hours to process 1 hour of corpus
    - 3 hours to collect
    - 3 hours to transcribe
    - 5 hours to translate
    - 1 hour for corpus management tasks
  - = 100,000 words per person month of 6 hour days

Archiving

- Work not done until corpus is committed to an archive for long term preservation and access
- Open Archival Information System (OAIS) reference model specifies requirements for trustworthy digital archiving (ISO 14721:2003)
- Popular open-source digital library systems:
  - DSpace: **http://www.dspace.org**
  - Fedora: **http://www.fedora.info**
  - EPrints: **http://www.eprints.org**
  - Greenstone: **http://www.greenstone.org**

SIL International

Open Language Archives Community (OLAC)

http://www.language-archives.org

- An open community creating a world-wide virtual library of language resources
- Uses two standards from the digital library world to create an aggregator that supports resource discovery across all institutions:
  - Dublin Core metadata standard
  - Open Archives Initiative (OAI) Protocol for Metadata Harvesting
- Now has 34 participating archives

SIL International

Areas of application (1)

- BOLD corpora will serve the academic community as a basis for:
  - ***Linguistic description*** — providing primary data for the analysis of phonology, grammar, texts, lexicon (even after the language is gone)
  - ***Linguistic training*** — providing data for examples, problems, and theses

Areas of application (2)

- Those involved in education and development for minority language communities will benefit from BOLD corpora since they can support:
  - ***Language learning*** — providing comprehensible input through oral transcription and translation
  - ***Literature development*** — providing source material for new literature and other educational materials

Areas of application (3)

- Minority language communities will benefit from BOLD corpora since they provide a basis for:
  - ***Heritage preservation*** — saving a record of traditional knowledge and of a group's identity as a people
  - ***Language revitalization*** — providing source material to help people learn their language or learn it better

Conclusion

- A language documentation corpus can be developed in a fairly short period of time by using a purely oral approach.
- Archiving such corpora will:
  - Ensure that documentation of endangered languages is preserved before it is too late.
  - Address the need of the scientific community concerning the loss of important information.
  - Address the need of language communities to preserve an aspect of their identify and to support revitalization efforts.