Sparse Transcription
Rethinking Oral Language Processing

Steven Bird, Charles Darwin University
steven.bird@cdu.edu.au
(paper for this talk available on request)
Why we transcribe

It is easy to capture a large amount of audio

We want to get text into the linguistics or NLP pipelines

But: transcription bottleneck
The tools

Elan: Glossing with no lexicon!

FLEX: Transcription with no speech
Computational approaches

NO TWO THE SAME
What is transcription?
Kunwinjku: polysynthetic language, 2000 speakers

kabirrimalaydjarrkunjdjangkan

birri- malay- djarrk- kunj- djangka- n
we.incl morning together kangaroo hunt non.past

let's all hunt kangaroo together in the morning
Source

1. Record

2. Respeak
   = oral
   transcription
   (now)

3. Transcribe
   (sometime in the future)
Oral Transcription (interactive)

ku-bolk-nahnah-ni

birri-bolk-nahnah-ni

birri-bolk-nahna-ni

they-country-look.after-imperf

Do enough so I can transcribe later

birri-h-ni birri-bolk-nah-na-ni ku-mekke kun-red they-BE they-look-after that country

BANGARDI MARALNGURRA
GUNBALANYA
OCTOBER 2019
Mahn ngandibe’kkal Nagay Balang Nabirdbird
wajh ngayawmalewan bi’i nahne bi’i, kahbi’i
Kubodne ngareduh nahne nquiri’siin dolobbo
kure nakhobang ngareduh Konkumo marneydyol meg
Buladjang. Nahne walem. Dja Koroko bi’rihi
Marahmarah. Ngabenjime bi’rihye’aghuwar dja
birriwemwarre dja mak ya’sow marahmarah ngabenjime
Yoh ngabo’iud x ngabere Kure nda’i bi’ihidi nasyungki
nawu marahmarah mak marah kabi’rye’ungki bol
marahmarah Koroko bi’rihi bi’rihye’iinwani.
Mah ngadibekkan ngaye balang nabirdbird wanjh nagayawmulewan nahne bim kahbimdi nauw kubodme ngardduk nahne ngurrihnan dolobbo kure nahkohbanj ngardduk kornkumo nganmarneyolyolmeng Buladjang nahne walem dja korroko birrihni mawahmawah ngabenymie birriyahyahwurd dja birriwernwarre dja mak yayaw mawahmawah ngabenymie yoh ngabbard kornkumo ngadberre kumekke birrihdi nayungki nawu mawahmawah mak nauw mawah kabirriyungki ?? mawahmawah korroko birrihni birribolknahnani kore kunred kore Buladjang kahdi wanjh ngurrrina kahbimdi nahne ngurrrina kahbimdi kumekke kore birrihwam nayungki birrire mayh birriyawani kore bindidahmeng bindimarneyolyolmeng mak yuwn mak... second birriwam with no story, they had only little bit story... second birriwam yiman yereyere same thing happened yimeng bolkki birridjalwam anyway birridjalwam same birriwam...
Interactive Transcription

Play a recording from the computer

Pause

Speak what we are hearing

KAMARRANG GUYMALA
KABULWARNAMYO
OCTOBER 2018
What's going on?

- We recognise repeated forms, in the midst of unrecognised material
- There is always unrecognised material
- We can only skip it (wastes time to try to transcribe it)
  - noise in the signal (ambient, human)
  - disfluency, speech impediments
  - unknown vocabulary (incl loanwords)
- More examples of why we want to transcribe words, not phones...
Transcribing words

1. kayadirri ~ ka-birri-yaw-dirri
2. berre ~ bedberre
3. mahne ngalengman ngan-bedde
4. ka-bourk-mang ~ ka-borurrk-mang ~ ka-bo-durrk-mang
5. kadiakodjuke ~ konhda ka-bandi-yaw-kodj-djuhke
English example: d'ya d'ya see?

DO WE WRITE WORDS OR PHONES?

1. dou jou si – enables further analysis of the text
2. ɣəɣəsi – enables accurate phone recognition

`field linguists [should modify] their [transcription] practice so as to assist the task of machine learning' – Seifart et al 2018

`linguists should aim for exhaustive transcriptions that are faithful to the audio ... mismatches result in high error rates down the line' – Michaud et al 2018
Transcription as observation

We hear a form repeatedly, and add it to our list, with a canonical representation.

Many of these speech tokens will be significantly reduced.

Transcription = pairing a locus of speech with an entry in a proto-lexicon.
This is not new

"I listened to stories and wrote down words... my glossary is really growing"

– Franz Boas (quoted in Sanjek 1990)
Transcription as observation

"No matter how careful I think I am being with my transcriptions, from the very first text to the very last, for every language that I have ever studied in the field, I have had to re-transcribe my earliest texts in the light of new analyses that have come to light by the time I got to my later texts... You can probably expect to be transcribing and re-transcribing your texts until you get to the final stages of your linguistic analysis and write-up." (Crowley 2007)
Transcription as observation

"... a transcription, whatever the type, is always the result of an analysis or classification of speech material. Far from being the reality itself, transcription is an abstraction from it. In practice this point is often overlooked, with the result that transcriptions are taken to be the actual phonetic ‘data’." (Cucchiarini 1993)
Transcription as observation

Those who deal with the spoken word... seem to regard phonography as little more than a device for moving the scene of alphabetic notation from the field interview to the solitude of an office... The real analysis begins only after a document of altogether pre-phonographic characteristics has been produced... The alphabet continues to be seen as an utterly neutral, passive, and contentless vehicle – Tedlock 1983
In spite of this...
Documentary workflow: transcriptions = data

The importance of the transcript resides in the fact that for most analytical procedures it is the transcript and not the original recording which serves as the basis for further analyses – Himmelmann 2006

For the scientific documentation of a language it would suffice to render all recordings utterance by utterance in a phonetic transcription with a translation – Mosel 2006
Transcription? phonetic vs IGT

Workflow: phonetic transcription → glossed text
NLP pipeline: transcriptions = data

1. automatic phone transcription
2. automatic word segmentation
3. the rest of the pipeline...
Word segmentation?
BAKED INTO THE DATA!

\[
\begin{align*}
t\text{emp}m \sim t\text{en} \ p\text{m} & \quad \text{‘ten pin’ (homorganic nasal assimilation)} \\
\text{hæd}ji \sim \text{hæd} \ ji & \quad \text{‘had your’ (palatalisation)} \\
t\text{en}t\text{SENTS} \sim t\text{en} \ s\text{ENTS} & \quad \text{‘ten cents’ (coarticulation of nasal and fricative)} \\
\text{lor\text{a}nd} \sim \text{lɔ} \ a\text{end} & \quad \text{‘law and (order)’ (intrusive ‘r’)}
\end{align*}
\]
Who can do this?
ANSWER: ONLY (COMPUTATIONAL) LINGUISTS

phonetic transcription

IGT

automatic phone transcription

automatic word segmentation
1. A long time ago, a woman and her son both stayed in a village. They lived there.

2. Menipa goliha venala zegipa getaminako hilih. Father got sick, wife, baby, and baby died.


4. Mota litaoko napadoake izg nama peletokanga. Very quickly grew pig, birds killed me.

5. Izelahina gizopa otoko wine. His mother looked after went.
Transcription as observation

mevet chat

just another way to identify a lexical entry
no segmentation
we can always discover elided material between words
Collaborative workflow
PhD project of Éric Le Ferrand

1. for each phrase:
2. verify forms automatically recognised in previous iteration
3. tag with new lexical items that we can confidently identify
   a. speak a form that was not identified
   b. automatically locate it in the phrase
   c. elicit translation, add to lexicon
4. retrain word models
### Scaling up

(a) Low Resource Scenario, with enough data to train a basic speech recogniser

(b) Zero Resource Scenario: unsupervised word spotting by co-indexing and labelling recurring forms

(c) Tapered Corpus: an open-ended speech collection, some translations, fewer transcriptions, cf Fig. 2
From protolexicon to lexicon

New PhD project!

Protolexicon contains repeated forms. We need to analyse it, splitting and merging...

Only then will our transcriptions be a `nice polite sequence of morphemes' (Bender 2019)

Construct the actual lexicon

Evaluation: "Lexeme error rate"
Dismantling the transcription bottleneck

If transcriptions = data

1. want a surrogate for the signal: phonetic transcription
2. must transcribe everything
3. must transcribe as the first step

If transcriptions = observation

1. want any useful annotation of the signal: includes "lexical" transcription
2. we transcribe what we can observe (always provisional, always more audio)
3. we can translate first
This is a return to orthodoxy: The Tapered Corpus

(1) **Core Corpus**: a central body of data, the basis of a complete linguistic analysis;
(2) **Indexed Corpus**: equipped with a complete lexicon, an indexed list of morphemes with glosses and morphological classifications;
(3) **Transcribed Corpus**: transcriptions prepared as soon as possible after the recording to reduce the frustrations of cold notes;
(4) **Translated Corpus**: not transcribed but translated into some familiar language, with indications of the social contexts;
(5) **Raw Corpus**: unprocessed recordings.

- The quantities of data at each level follow a power law, based on the amount of curation they require (after Twaddell 1954, Samarin 1967)
- Translation precedes transcription (capturing meaning is more urgent than re-representing our data)
Three approaches to design

How does an outsider encourage people to keep their language strong?

1. capture (is it effective?)
2. address a cause of language shift: low prestige, relevance
3. address another cause: functionality
LanguageParty.org
Venez à notre Language Party!

Berbère, Mongol, Kanak, Arménien, ...
Contes dans les langues d’origine avec traduction
mardi 26 nov à 18h30 à la bibliothèque
Oral Language Learning

APPROPRIATING SIMPLE APPS

(a) Learning terms of address by capturing selfies and recording a short bilingual dialogue

(b) Learning key phrases by capturing segments where a term is spoken and translated

(c) Learning linked to places, where multiple people describe the same set of places
Conclusion 1: The Status Quo

In NLP, we extend our text-based methods to handle speech input by adding a speech-to-text component.

Therefore, for unwritten languages: convert speech data to text data and continue as before.

The existence of automatic speech recognition provides proof of concept.

But, these are not "unwritten languages", but ...
Conclusion 2: Problems with status quo

1. **over-values transcriptions**, treating them as data when they are nothing more than contingent observations (‘Premature systematisation keeps nature's surprises hidden’ – Lenat and Feigenbaum 1987)

2. **under-values words**, treating them as the byproduct of boundary detection when they are meaningful units which often overlap in the speech stream

3. **trivialises the role of linguists** to phone transcription when they can conduct complex workflows involving iteration and interaction

4. **excludes the speech community**, the workforce, and the main beneficiary (NB for these people, FAIR != fair)
Conclusion 3: Transcription as Observation

**transcription = data:** transcribe phones / fully / first

**transcription = observation:**

- map locations in the speech stream to an inventory of meaningful units
- generalises over dense and sparse transcription
- transcription = discover repeated meaningful units
  (regardless of whether they are canonical)
Conclusion 4: Have we addressed the bottleneck?

Don't waste scarce resources on unimportant tasks!

- phone transcription is extremely time consuming for spontaneous speech in the presence of coarticulation and disfluency
- much of the phonetic detail is not necessary as long as we are identifying words

Instead, allocate resources to the central acts of transcription:

- identify meaningful units without deciding on their formal status (as morphemes, words, or multiword expressions)
- identify meaningful units without baking-in boundaries (then train word-spotters)
- allocate effort to the units of interest (ie words with meanings), to improve our ability to identify oral texts for closer attention (including dense transcription)
Conclusion 5: New promises of scalability

1. it is open to participation by local speakers, given their superior ability to identify meaningful units in the speech stream, even in the presence of noise

2. it faces the Zipfian distribution of words; word-spotting enables us to allocate effort according to decreasing frequency, and to the topics / texts of interest

3. it treats each additional resource as auxiliary information, i.e. further supervision to help annotate the signal; this necessary flexibility in the face of diverse language situations, different constellations of data and skills