Language Documentation

Laura Tomokiyo

What is language documentation?

- Provides “a comprehensive record of the linguistic practices characteristic of a given speech community” (Himmelman 1998)
- Focuses on description and archiving
- Forms the basis for further analysis

Why LD for endangered languages

- Conservation
- Analysis
- Education
- Revitalization
- Reclamation

Why LD for language technologies

- Many language technologists are not experts in linguistics, but *are* experts in the kind of data they need and can collaborate with linguists
- Language technologists need to know about the standards and practices core to linguistics
  - We don't arbitrarily define new technological or mathematical standards and practices, but too often that's exactly what happens for language
- Developing models for documenting the sounds, words, and relationships between words in endangered languages *will help* in creating systems in low-resource and rapid deployment situations

Stone Age resources

- Inscriptions on stones, bones, clay tablets
  - Not produced to provide a linguistic record
  - Yet have been successfully used to explore long-extinct languages
- Hittite reconstruction
  - We know something about government, law, trade, religion
  - What was adolescent conversation like???
  - Is it possible to have the verb in first position in subordinate clauses???

Modern-day resources

- All information needed for further descriptive analysis should be contained in the corpus
- The corpus should conform strictly to established and interoperable standards, practices, and formats
- The corpus should be large enough that important evidence for grammatical structure can be extracted
  - Elicited data?
  - Negative examples?

What does LD look like?

<table>
<tr><th>Primary data</th><th colspan="2">Apparatus</th></tr>
<tr><td rowspan="2">Recordings/records of observable linguistic behavior and metalinguistic knowledge

(possible basic formats: session and lexical database)

Interoperable formats are crucial (plain text, xml)</td><td>Per session</td><td>Overall</td></tr>
<tr><td>Metadata
• Time and location of recording
• Participants
• Recording team
• Recording equipment
• Content descriptors

Annotations
• Transcription
• Translation
• Further linguistic and ethnographic glossing and commentary</td><td>Metadata
• Location of documented community
• Project team(s)
• Participants
• Acknowledgments

General access resources
• Introduction
• Orthographic conventions
• Ethnographic sketch
• Sketch grammar
• Glossing conventions
• Indices
• Links to other resources...</td></tr>
</table>

From Himmelmann 2006

Ethical considerations

- Do no harm
  - Respect cultural norms of privacy, status, compensation
- Reciprocity and equity
  - Plan research collaboratively – the researcher's viewpoint is not the only one
  - The indigenous knowledge system is rich
- Give back
  - What would actually be useful to the community?
- Obtain informed consent
  - Explore oral/communal consent
- Archive and disseminate
  - Shared data is more useful than no data
  - Language is too precious to be proprietary

Formal representations: words and grammar

- Meaning
- Number
- Gender
- Person
- Possessives
- Distance
- Direction
- Voice
- Register
- …

IPA

- International Phonetic Alphabet
- A standard for making distinctions between sounds
- A set of symbols for writing those sounds down
- Corresponding practices for deciding whether related sounds should be
  - Written with the same symbol (allophonic variation / phonemic transcription)
  - Written with modifiers (diacritics esp. for idiolects)
  - Written with two different symbols (phonemic distinction or phonetic transcription)

Formal representations: sounds

- Identify the sounds in a language which, if changed, make a difference in meaning *in that language*
- Characterize the difference between those sounds
- Situate those sounds in the context of
  - The sounds humans can make
  - The sounds of other languages

Orthography ≠ phonetics

- One letter, many sounds
  - equity, equal, beneath
- One sound, many letters
  - cash, character, king, queen

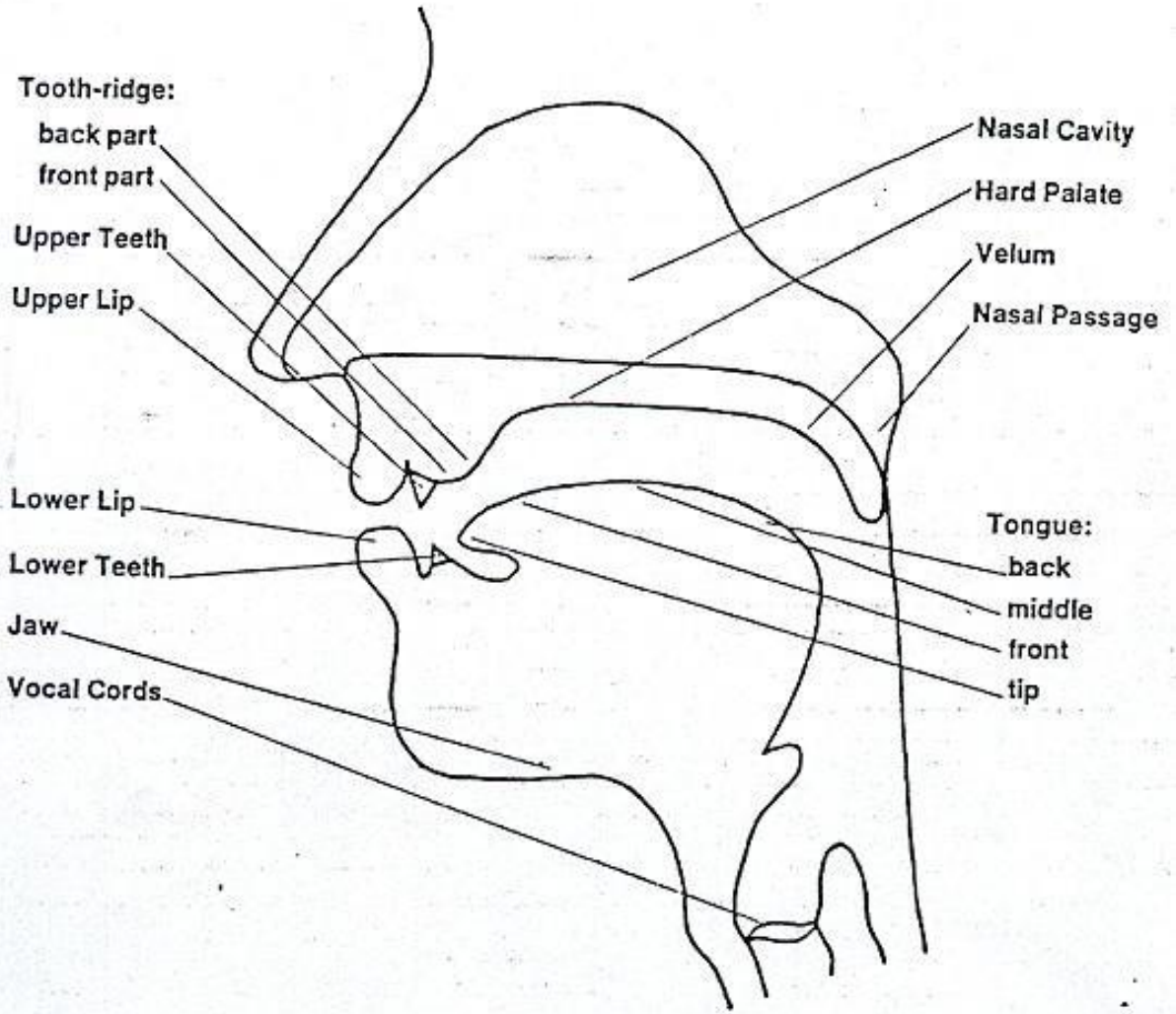
THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

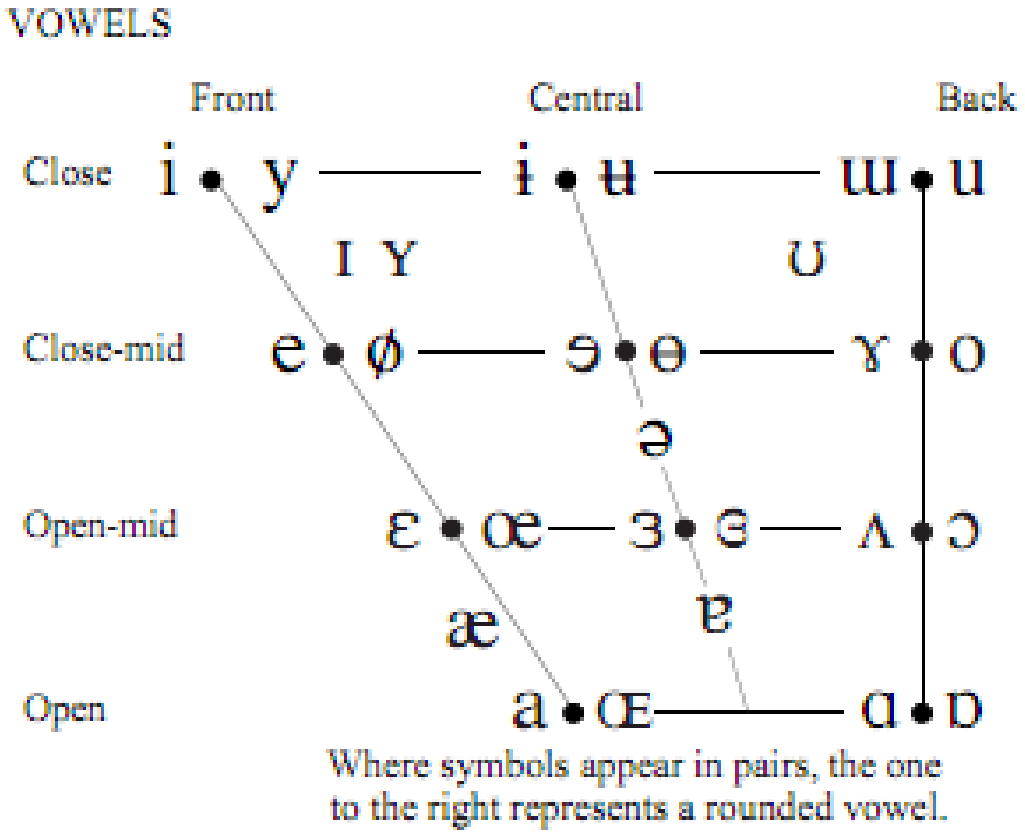
	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

The Organs of Speech



Bert but between



beat

bit

bait

bet

bat

boot /buːt/

put

boat

bought

baht

/biːt/ /bɪt/ /beit/ /bɛt/ /bæt/ /bait/
/bɜt/ /bʌt/ /bətween/
/buːt/ /boːt/ /bɔːt/ /bɑːt/

Disclaimer: American English vowels are not usually the pure sounds

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. n̊ j̊

◌̥	Voiceless	n̥	d̥	◌̤	Breathy voiced	b̤	a̤	◌̪	Dental	t̪	d̪
◌̬	Voiced	s̬	t̬	◌̰	Creaky voiced	b̰	a̰	◌̺	Apical	t̺	d̺
◌ʰ	Aspirated	tʰ	dʰ	◌̼	Linguolabial	t̼	d̼	◌̻	Laminal	t̻	d̻
◌̹	More rounded	ɔ̹		◌ʷ	Labialized	tʷ	dʷ	◌̃	Nasalized		ẽ
◌̜	Less rounded	ɔ̜		◌ʲ	Palatalized	tʲ	dʲ	◌ⁿ	Nasal release		dⁿ
◌̟	Advanced	u̟		◌ˠ	Velarized	tˠ	dˠ	◌ˡ	Lateral release		dˡ
◌̠	Retracted	e̠		◌ˤ	Pharyngealized	tˤ	dˤ	◌̚	No audible release		d̚
◌̈	Centralized	ë		◌̴	Velarized or pharyngealized	ɬ̴					
◌̽	Mid-centralized	e̽		◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)				
◌̩	Syllabic	n̩		◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)				
◌̯	Non-syllabic	e̯		◌̘	Advanced Tongue Root	e̘					
◌˞	Rhoticity	ɚ	a˞	◌̙	Retracted Tongue Root	e̙					

Exercises

- Swadesh list with groups of 3
  - First 10 first
  - Then try the remainder until time is up
  - regroup to discuss differences between speakers, transcriber agreement
- Homework to transcribe
  - http://millercenter.org/president/speeches/speech-332 #paragraph 2

Swadesh List

1. I
2. You
3. We
4. this
5. that
6. who?
7. what?
8. not
9. all (of a number)
10. many
11. one
12. two
13. big
14. long (not 'wide')
15. small
16. woman
17. man (adult male human)
18. person (individual human)
19. fish (noun)
20. bird
21. dog
22. louse
23. tree (not log)
24. seed (noun!)
25. leaf (botanics)
26. root (botanics)
27. bark (of tree)
28. Skin
29. flesh
30. Blood
31. bone

Discussion

- What was hard?
- Where did transcribers differ?
- Where did speakers differ?

ASCII alternatives to the IPA symbols

- Various ASCII representations
  - AA, AH, AX, AY, …
- Biased toward English, and a particular view of English
- Speakers of different languages have different issues
  - Unfamiliar character-pronunciation mapping (j/y)
  - Unfamiliar character set (e.g. Japanese)
  - Inexperience writing the language down (e.g. Iñupiaq)
- Different systems can define own phoneme set, but ultimately need to be multilingual