Languages without Writing

- *Most languages are unwritten*
  - *How writing systems evolve*
  - *How they are often not ideal*
  - *What should they write*
  - *Writing standards*
  - *Speech Technology for unwritten languages*
  - *Discovering a writing systems*

Writing Systems

- *"Computational Theory of Writing Systems" by Richard Sproat*
- *Actually limited number of writing systems*
  - *Letter, mora, syllable*
  - *Covering all written phenonema*
    - → *Stress, tone, etymology, semantics*

Selecting Writing System

- *Often inappropriate*
  - *Borrowed from colonial power/missionaries*
  - *Not capable of distinguishing issues in language*
  - *Writing can cause changes in the language*
- *Legacy conventions*
  - *'ph' for /f/ for Greek derived words 'gh' for /x/*
  - *Confusing use of new letters 'c' as /k/ /ch/ /s/*
- *Lost letters*
  - *Menzies /m ih ng ih s/ due to lost (yogh) written as z*
  - *"Ye Olde Shoppe" 'ye' is really Þe (thorn)*
- *Writing normalizes pronunciations*
  - *'forehead' was 'for'head' not 'fore head'*

What to write

- *Text and Speech are different languages*
  - *In some cultures, very different*
- *Writing is for taxes*
  - *Not to record speech, but to record facts*
- *Spoken words vs "proper" words*
  - *Words: "I'm gonna hava …"*
  - *Grammar: "went home early" (pro drop)*
  - *Filler words: "Like, ehm, you know, …"*
- *Big distinctions*
  - *Arabic (MSA vs dialects)*
  - *Chinese (Putonghua vs dialects)*
  - *English (Written vs dialects)*

How to write

- *Writing standards*
  - *Consistent spelling*
    - *(even within the same document)*
  - *Should it reflect dialect/sociolect*
    - *Colour/color, coordinate vs coördinate*
    - *I'm vs I am*
    - *Mis-adoption from other languages*
    - *Seoul, Beijing, Gibraltar*
- *Most consistent spelling requires:*
  - *Country-wide education system*
  - *Government definitions*
  - *MS Word support ☺*

Nonwritten Languages

- *Mostly spoken*
  - *Typically not taught in schools*
- *Speakers may be literate in other languages*
  - *Putonghua, Hindi, Spanish, English etc*
- *No regularly written examples*
- *No standard*

Language Technologies for Unwritten Languages

- *Devise your own writing system*
  - *DARPA Transtac for Iraqi Arabic*
  - *Need to train people to read/write it*
  - *Need good tools*
    - *(spell checkers, morphanalysis)*
    - *cannot vs can not vs can't vs can'*
  - *For Iraqi: used Arabic script, but writers confused between MSA and Iraqi.*

Just write in …

- *Just use IPA/Sampa/MyUniversalSystem*
  - *Not enough to define a standard*
  - *Too much variation between writers/speakers*
  - *Not clear what allophonic level to go to*
  - *Its hard to type, requires too much expertise*

Automatic Discovery

- From a set of recorded data
  - "Nice" sentences
  - Use ASR (phoneme) to find initial system
    - → Use other "close" languages to start from
    - → Be rich enough to capture distinctions
  - Iterate (retrain acoustic models)
    - → relabel.
- Build a TTS engine from this labeling
- But how do you write this?

Using ASR to find "writing"

- *(But this is all for English so its not fair)*
- *English female speaker (45 mins speech)*
  - *Original text:*
  - *In a study of 72 ballot issues in Massachusetts and three other states, Boston University political science professor, Betty Zisk, found that 88 percent of the battles were won by the side that spent the most money.*
  - *ASR text:*
  - *In estonian 72 validations in massachusetts in 3 other states boston university political buddy says about 80 percent of the battles won by the side could spend the most money*

  

Using ASR to find writing

- *(Hindi)*
- *Recognize with English Phones*
- *Build TTS voice with result*
- *Iterate*
  - *Build new acoustic models with "english"*
  - *Re-recognize Hindi data with new model*
  - *Repeat (5 times)*
- *"Text": AH N IY AE K S AH T OW T EY Z AE N EY M AH N EY S AH M AA N AH B AA K IY S AH M P"*

Cross Lingual Writing

- *Write in one language and it comes out in the other*
- *Building a Konkani dialog system*
  - *Write the prompts in Hindi*
  - *Use SMT technology to convert to new written form*
  - *Speech comes out in Konkani*

SMT

- *Discover (phonetic) writing system*
- *Learn SMT conversion to known language*
- *What about:*
  - *Syllables, words, can they be discovered*
  - *How much data do you need (40,000 utts?)*
  - *Efficient collection:*
    - → *Speak these Hindi sentences in Konkani*