Notes from previous tasks

- “Your guys” in speech (various corpora, SketchEngine)
- Q/A interviews (NPR ATC, Clojure for scraping & parsing)
- Loanwords in Japanese (KNBC blogs, grep, lookup tools)
- Twain’s 1st person relative clauses (P.Gutenberg, AntConc)
- Loanword in Japanese (WWW, SketchEngine)
- Co-occurring lex/hypernym (Austen corpus, WordNet, NLTK)
- English phrasal verbs (WebCorp, AntConc)
- Hapaxes (Book of Mormon, KJV, NLTK, AntConc)
- Polysemic co-occurrence (NLTK)
- Occurrences of “scandal/fraud/corruption” (Enron, NLTK)

- Search ≠ research
COMPUTATIONAL
PHONETICS, PHONOLOGY

(Computing language sounds)
Phonetics and phonology

- Phonetics: study of the basic sounds (in isolation) that make up a language
  - Phonemic inventory: phones
  - Representation: International Phonetic Alphabet, ArpaBet, and lots of others...
  - Approaches: articulatory, acoustic
- Phonology: study of ways that basic sounds change (in context) in a language
  - Rules, interactions, processes

\[ /h/ \rightarrow \begin{cases} [h] \sim [\theta] & \text{C}_{\text{SON}} \\ [fi] / \_\_ \_ \_ \_ & \text{elsewhere} \end{cases} \]

Figure 0.1: Allophonic variation of /h/
Computational phonology

- Phonological aspects of speech storage and manipulation by computer
- Modeling of speech-specific properties of language(s)
- Systems that assist in capturing, annotating, analyzing, exploiting speech data
- Systems that implement (aspects of) phonological theory
Sound segments

• Consonants (8-98 depending on language)
• Vowels (3-64 depending on language)
• Glides: j, w, etc.
• Diphthongs: aj, uw, etc.
• Binary feature analysis
  • Voicing, labial, interdental, alveopalatal, uvular, glottal, oral, nasal, stop, etc. etc.
  • High, back, front, mid, rounded, nasal, etc.
### The International Phonetic Alphabet (Revised to 1993)

#### Consonants (Pulmonic)

<table>
<thead>
<tr>
<th>Articulation</th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Postalalveolar</th>
<th>Retracted</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>p b</td>
<td>t d</td>
<td>t d</td>
<td>c j</td>
<td>k g</td>
<td>q G</td>
<td>q G</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Nasal</td>
<td>m n</td>
<td>n n</td>
<td>n n</td>
<td>n n</td>
<td>n n</td>
<td>N N</td>
<td>N N</td>
<td>N N</td>
<td>N N</td>
<td>N N</td>
<td>N N</td>
</tr>
<tr>
<td>Trill</td>
<td>b r</td>
<td>r r</td>
<td>t t</td>
<td>t t</td>
<td>t t</td>
<td>t t</td>
<td>t t</td>
<td>t t</td>
<td>t t</td>
<td>t t</td>
<td>t t</td>
</tr>
<tr>
<td>Tap or Flap</td>
<td>f f</td>
<td>s s</td>
<td>s s</td>
<td>s s</td>
<td>s s</td>
<td>s s</td>
<td>s s</td>
<td>s s</td>
<td>s s</td>
<td>s s</td>
<td>s s</td>
</tr>
<tr>
<td>Fricative</td>
<td>f β</td>
<td>θ δ</td>
<td>s z</td>
<td>j c j</td>
<td>x y</td>
<td>x y</td>
<td>x y</td>
<td>x y</td>
<td>x y</td>
<td>x y</td>
<td>x y</td>
</tr>
<tr>
<td>Lateral fricative</td>
<td>l b</td>
<td>l b</td>
<td>l b</td>
<td>l b</td>
<td>l b</td>
<td>l b</td>
<td>l b</td>
<td>l b</td>
<td>l b</td>
<td>l b</td>
<td>l b</td>
</tr>
<tr>
<td>Approximant</td>
<td>v j</td>
<td>l j</td>
<td>l j</td>
<td>j m j</td>
<td>j m</td>
<td>j m</td>
<td>j m</td>
<td>j m</td>
<td>j m</td>
<td>j m</td>
<td>j m</td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
</tr>
</tbody>
</table>

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

#### Consonants (Non-Pulmonic)

**Clicks**
- Bilabial: b, as in:
  - Dental: d, P, dental/alveolar:
    - Palatoalveolar: t, k, Velar:
- Alveolar lateral: s, Alveolar fricative

**Voiced Implosives**
- Bilabial: b

**Ejectives**
- Bilabial: b

**Vowels**
- Front: i, y, i, y
- Central: e, ə, e, ə
- Back: u, u
- Open-mid: e, ə, e, ə
- Open: a, ə, a, ə

Where symbols appear in pairs, the one to the right represents a rounded vowel.

#### Suprasegments

- Primary stress
- Secondary stress
- Long
- Half-long
- Extra-short
- Syllable break
- Minor (foot) group
- Major (intonation) group
- Linking (absence of a break)
- Downstep
- Upstep

#### Tones & Word Accents

- Level
- Contour
- Rising
- Falling
- High
- Mid
- Rising-falling
- Low

#### Diacritics

Diacritics may be placed above a symbol with a descender, e.g. i.

- Voiceless
- Breathy voiced
- Creaky voiced
- Aspirated
- Lingual-labial
- Labialized
- Nasalized
- Palatalized
- Advanced
- Volarized
- Lateral releasable
- Pharyngealized
- Nasal release
- Advanced Tongue Root
- Rhoticity

- Centralized
- Retracted Tongue Root

- Mid-centralized
- Raised
- Lowered
- Non-syllabic

### Other Symbols

- Voiceless labial-velar fricative
- Voiced labial-velar approximant
- Voiceless labial-palatal fricative
- Voiced labial-palatal approximant
- Voiceless epiglottal fricative
- Voiced epiglottal fricative
- Epiglottal plosive
- Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.
- No audible release
Units of phonological analysis

- Phones
- Diphones, triphones, etc.
- Syllable (and subparts)
- Morae
- Prosodic word

- Phoneme: combination of different variants of a basic sound in the language
- Allophone: variant forms of a basic sound segment
  - The basic phoneme /l/ in English has two allophones (voiced and voiceless).
  - The basic phoneme /t/ in English has as many as eight (!) variants.
Syllables

- Widely used unit of phonological structure
- Above sound segment, below word
- Four principal components:
  - Onset: consonants that precede nucleus
  - Rime (rhyme)
    - Nucleus: most sonorous (V, G), obligatory
    - Coda: consonants that follow nucleus
- Basic orthographic component in many languages
Syllable structure

- Basic pattern is the same crosslinguistically
- Sonority is criterion for O/N/C membership
- Determines allophonic variation, stress, many phonological processes
- Wide-ranging implications for other areas of language (e.g. orthography)
Phonological rules

- All languages have several phonological rules
- Starting point is underlying form, result of rule application(s) is surface form
- These rules interact with each other in (often) very complicated ways; dependencies are ordered, interdependent, even cyclic
- Accounting for these rule sequences is a large part of phonological investigation
- Various theories attempt to describe these interactions as succinctly as possible
- Native speakers have little/no explicit knowledge
Phonological rules (Eng)

- CiV-Lengthening
  - Canada/Canadian, felon/felonious, comedy/comedian
- Diphthongization (Caucasus/Caucasian)
- Velar Softening
  - Public/publicity, prodigal/prodigy, medic/medicine
- Palatization
  - Race/racial, diffuse/diffusion, enclose/enclosure
- Yod-deletion (impress → impression)
- Spirantization
  - Explode/explosion, president/presidency
Blackfoot Phonological Rules (Frantz 1997)

Note: + means morpheme boundary, and { } delimit choices.

1. **Gemination:** $C_1 \rightarrow C_2 / _{___} + C_2$
2. **s-Insertion:** $\phi \rightarrow s / I_{-i}$
3. **s-Connection:** $\phi \rightarrow s / C + _{___} s$ and $\phi \rightarrow i / V(') + _{___} s$
4. **o-Replacement:** $o \rightarrow a / _{___} + a$
5. **Coalescence:** $wi(:) \rightarrow o$
6. **Breaking:** $k \rightarrow ks / _{___} i$
7. **Neutralization:** $I \rightarrow i$
8. **Desyllabification:** $\{i \rightarrow y, o \rightarrow w\} / V + _{___} V$ where $i$ and $o$ are unaccented.
9. **Semivowel drop:** $C \rightarrow \phi / _{___} #$
10. **Vowel shortening:** $V_i : \rightarrow V_i / _{___} + V$
11. **i-Loss:** $i \rightarrow \phi / V y _{___} \{a,o\}$
12. **i-Absorption:** $i \rightarrow \phi / s _{___} \{a,o\}$
13. **ih-Loss:** $ih \rightarrow \phi / s _{___} e$
14. **Presibilization:** $ths \rightarrow ss, lths \rightarrow lss$
15. **Semivowel loss:** $C \rightarrow \phi / C _{___}$ where $C \neq \mu$.
16. **y-Reduction:** $yi \rightarrow ii / C _{___} y$
17. **Postasibilation:** $ih \rightarrow s / e _{___}$
18. **t-Affrication:** $t \rightarrow ts / _{___} i$
19. **Glottal metathesis:** $'V \rightarrow V' / V _{___} C$
20. **Glottal loss:** $' \rightarrow \phi / VV: _{___} C$ where $V: is underlyingly long.
21. **Glottal assimilation:** $V_i ' \rightarrow V_i : / _{___} \{e\} C$
22. **Glottal reduction:** $' \rightarrow \phi / _{___} ' $
23. **Vowel epenthesis:** $\phi \rightarrow V_i : / V_i ' _{___} h$
24. **sss-Shortening:** $ass \rightarrow sss / _{___} C$
25. **Accent spread:**

$$V \rightarrow [+\text{accent}] / \left[ \frac{V}{+\text{accent}} \right] _{___} +$$
Phonological processes (Eng)

- Assimilation: makes another sound similar
  - inconsiderate, impossible, illogical, irreverent
- Vowel reduction: neutralize unstressed vowel
  - Canada
- Flapping: alveolar stops between vowels
  - butter
- Epenthesis: adding a sound segment
  - Warm[]th
- Deletion (suppose, parade, police)
- Metathesis: swapping sound segments
  - Ask, spaghetti
Word-level sound rules (Eng)

- Vowel Shortening
  - CiV-Lengthening
  - s-Voicing
    - Velar-Softening
      - Prevocalic Tensing
      - Stressed Tense
      - Vowel Lengthening
    - Palatalization
      - [iu]-Formation
    - Spirantization
      - [j]-Morification
      - [j]-Deletion
Suprasegmentals

- Properties that are found “above the sounds”
- Intonational contours
- Loudness, tempo, rhythm
- Length (casa, cassa (Ith))
- Tone (milh=moth, snare, sleep (Sarcee))
  - Register: level tone, same pitch
  - Contour: rising/falling tone on same segment
  - Diacritics often used in orthography
- Stress
  - Diacritics sometimes used in orthography
Stress

• English
  • Rather irregular, fairly difficult
  • Three levels detectable
  • Secondary stress (raccoon vs. maroon)
  • Stress shifts in nominal compounds
  • Nuclear Stress Rule

• Polish: penultimate
• Turkish, French, Farsi: ultimate
• Czech, Finnish, Georgian: initial
• Cairene Arabic: very complex but regular
Other issues

- Phonotactics (sequences of permissible sounds/letters)
- Dialects, registers
- Speaker variation
- Pitch, accent, tone, contours
- Syllables, phrasal groups, breaths, pauses
- Disfluencies
- Connected speech
Computational frameworks

• Data analysis/annotation toolkits: SP-TK, Praat, and several others
• Phonemicizers (e.g. the CMU dictionary here and LOGIOS here)
• KIMMO (finite-state) etc.
• OT
  • Given constraints, rankings, input
  • Generates outputs, hierarchy of violations
  • Syllable parsers, other applications
  • Toolkits: OTKit, PyPhon, PYCOT
Sound signals

• Spoken language is ultimately a sound wave which must be encoded, decoded
• Physical, acoustical, mathematical properties that apply to sounds in general
• Specific features associated with language(s)
• Optimal ways of storing, manipulating, generating waves (with loss of some information, though!)
• Spectrogram reading tutorials here
Speech transcription

- Visualizing sound waves
- Converting to phonemic or orthographic text
- Annotating with timestamps
- Separating out interesting parts of the speech signal
- Indicating errors, special intonation, isolating noise, linking signal with linguistic properties, waveform editing
- Cataloguing sound inventories in unrecorded languages
Speech understanding

• Speech transcription
• Some kind of task-related processing
  • Fact/document retrieval
  • Robotics
• Speech generation

• Conversation, discourse, dialog
Visualizing speech

- Toolkits: Praat, Audacity, SFS, WaveSurfer, etc.
- Waveforms, spectrograms
- Vocal properties
- File conversion
- Annotation
- Scripting!
Spoken language corpora

- Recordings of speech
  - Microphones, telephones, radio
  - Spontaneous conversations, experimental conditions, scripted and unscripted
  - Differing speakers, various languages
- Annotation: prosody, errors, pause duration, gestures, etc.
- A good overview is [here](#)
Soundex

- Used by US Gov’t to collapse similarly-sounding surnames
  - Retain the first letter
  - Drop a,e,h,i,o,u, w, y elsewhere
  - Encode remaining letters:
    - b, f, p, v --> 1
    - c, g, j, k, q, s, x, z --> 2
    - d, t --> 3
    - l --> 4
    - m, n --> 5
    - r --> 6
  - Collapse same-code sequences.
  - Pad/truncate to obtain LDDD

- L523: Lonsdale/Lomsdalen/Langdon/Lowenstein/Linstromberg, etc.
Two-level phonology

- Maps between surface orthographic and underlying phonological forms
- Finite-state transducer, lexicons
- Define default input/output mappings
- Define rules for regular variations, compiled into table, used in automata
- Modes: generation/production and recognition/parsing
Sample 2L Soundex rule

; Initial PH, PF, and F identical
; P:F <=> # ___ [H: | F:]
; F:0 <=> F: ___
RULE "Initial PH, PF --> F" 5 7
  # P P H F F @
  # F @ @ 0 @ @
1: 2 0 1 1 0 5 1
2: 1 4 3 1 0 5 1
3: 1 0 1 0 0 0 1
4: 0 0 0 1 1 0 0
5: 1 0 1 1 5 0 1 ; FF --> F0
The crossroads

• Many NLP applications treat personal names
  • (CL)IR of text (MUC, TREC, TIPSTER)
  • (CL)IR of spoken documents (TDT)
  • Information extraction (ACE)
  • i18n, l10n
  • OCR/digitization
  • Semantic Web annotation
  • Homeland security and DoD (Aladdin, REFLEX)
  • and, of course,
  • Family history research (PAF, TMG, etc.)
Jobs in speech/comp-phon

- Very employable skill in linguistics
  - Industry, academia, government
- Particularly lucrative with programming background
- Systems development, integration
- Language model development
- System training, testing, documentation
- Speech corpus collection, annotation
Speech corpora

• Samples of spoken data
• Varies by:
  • Language(s)
  • Speaker(s)
    • gender, emotions, age
  • Content
  • Recording/transmission device
  • Ambient conditions
  • Task
  • Annotation type, scheme
TIMIT (phoneme alignment)

0 2200 h#
2200 3320 hh
3320 4044 axr
4044 5037 q
5037 7240 ao
7240 7819 bcl
7819 8120 b
8120 9441 er
9441 10007 n
10007 11947 hv
11947 13869 eh
13869 15373 er
SB SAE (multiparty dialogue)

.trn format structure
2.660  2.805  JOANNE: But,
2.805  4.685  so these slides be real interesting.
6.140  6.325  KEN:  ... Yeah.
6.325  7.710  I think it'll be real interesting

.ca format structure
JOANNE:   But,
%snd:"sbc015.wav"_2660_2805
so these slides be real interesting.
%snd:"sbc015.wav"_2805_4685
KEN:   ... Yeah.
%snd:"sbc015.wav"_6140_6325
I think it'll be real interesting
%snd:"sbc015.wav"_6325_7710
File formats

- Way sound (including speech) files are stored
- Many different formats
  - HTK, Esignal, TIMIT, SDES1 (commercial)
  - NIST (US Gov.)
  - SCRIBE (EEC Esprit)
  - AIFF (Apple)
  - SUNAU8 (Sun)
  - OGI File Format
  - WAV (Microsoft)
WAV files

- 'RIFF': RIFF file identification (4 bytes)
- <length>: length field (4 bytes)
- 'WAVE': WAVE chunk id (4 bytes)
- 'fmt': format sub-chunk id (4 bytes)
- flength: length of format sub-chunk (4 byte integer)
- format: format specifier (2 byte integer)
- chans: # of channels (2 byte integer)
WAV files

- `sampsRate`: sample rate in Hz (4 byte int)
- `bpsec`: bytes per second (4 byte integer)
- `bpsample`: bytes per sample (2 byte integer)
- `bpchan`: bits per channel (2 byte integer)
- `'data'`: data sub-chunk id (4 bytes)
- `dlength`: length of data sub-chunk (4 byte integer)
File contents

- Waveform data
  - Various encodings
- Byte order differs across platforms
- Header
  - Contains information about contents
  - Various formats
  - NIST SPHERE most commonly used
NIST SPHERE header

NIST_1A
  1024
database_id -s5 TIMIT
database_version -s3 1.0
utterance_id -s8 aks0_sa1
channel_count -i 1
sample_count -i 63488
sample_rate -i 16000
sample_min -i -6967
sample_max -i 7710
sample_n_bytes -i 2
sample_byte_format -s2 01
sample_sig_bits -i 16
end_head
Sun audio file header

- magicNumber: magic number 0x2e736e64
- dataLocation: offset to start of data
- dataSize: number of bytes of data
- dataFormat: data format code (1 for 8 bit mu-law)
- sampRate: a sample rate code which is always 8012.821 Hz
- numChan: the number of channels
- info: arbitrary character string min length 4 bytes
Playing various file formats

- First try platform apps
  (iTunes, WMP, etc.)
- Then try standard acoustic tools
  (Praat, SFS, Audacity, WaveSurfer, etc.)
- Then try conversion programs
  (e.g. sox, NIST)
- Ask me
The problem

- Storing and accessing proper nouns crosslinguistically
Issues

• Scope: some 6,000 languages
• Various types of writing systems
• Conventions: culturally/linguistically set
• Crosslinguistic: migrations, minorities
• Diachrony: spelling changes over time
• Innovation: names are continually invented
• Borrowings: names cross barriers
## Romanization issues

<table>
<thead>
<tr>
<th>Arabic</th>
<th>Romanization</th>
</tr>
</thead>
<tbody>
<tr>
<td>شهید بهشتی</td>
<td>Shaheed Baishtee</td>
</tr>
<tr>
<td>فاطمی</td>
<td>Fatimee</td>
</tr>
<tr>
<td>کوی علی</td>
<td>Koy Halawee</td>
</tr>
<tr>
<td>شهر صنعتی</td>
<td>Sharai Sanati</td>
</tr>
<tr>
<td>دانشگاه</td>
<td>Daanishga</td>
</tr>
<tr>
<td>زهرا</td>
<td>Zahra</td>
</tr>
<tr>
<td>ولیا</td>
<td>Wali A…</td>
</tr>
<tr>
<td>اعظم</td>
<td>Ahzam</td>
</tr>
<tr>
<td>آستنده نورد</td>
<td>Aastana Naward</td>
</tr>
<tr>
<td>آستنده</td>
<td>Aastana</td>
</tr>
<tr>
<td>آستنده</td>
<td>Aastarkee</td>
</tr>
<tr>
<td>آستنده</td>
<td>Aastarkee</td>
</tr>
<tr>
<td>آستنده</td>
<td>Aastarkee</td>
</tr>
<tr>
<td>رضا</td>
<td>Reza</td>
</tr>
<tr>
<td>رضا</td>
<td>Reza</td>
</tr>
<tr>
<td>رضا</td>
<td>Reza</td>
</tr>
<tr>
<td>رضا</td>
<td>Reza</td>
</tr>
</tbody>
</table>
Writing systems

- **Alphabetic: (roughly) one symbol / sound**
  - Roman (Bush), Armenian (ԥնֲ), Georgian, etc.

- **Syllabic: (usually) one symbol / syllable**
  - Hiragana, Katakana (ブッシュ), Cherokee, etc.

- **Abugidic (alphasyllabic): CV***
  - Devanagari (बुश), Inuktitut, Lao, Thai, Tibetan, etc.

- **Logographic: (roughly) one symbol / word**
  - Hieroglyphs, Hieratic, Cuneiform, Hanzi (布什), etc.
Special cases

• Hangul
  • underlyingly alphabetic
  • sounds are arranged compositionally into syllabic symbols (부시)

• Abjads
  • alphabetic, but without (some/all) vocalization
  • e.g. Arabic, Hebrew, Persian (بوش)
Normalization

• Direction
  • left-right vs. right-left
  • horizontal vs. vertical
  • boustrophedonic
• Case
  • DeVon vs. Devon
• Vocalization
  • McConnell, St. John
• Diacritics
  • Étienne vs. Etienne
• Punctuation
• Abbreviations
Related computational aspects

• Character sets, fonts, glyphs
• Input/output (keyboard, display)
• Collation (ordering, alphabetization)
A few mapping strategies

• Don’t bother: lexical lookup
• Transcoding
• Transcription
• Transliteration
• Transduction
• Translation
Lexical lookup

- Rote, literal access (e.g. hash tables)
  - Unending, expensive lexicon management task
  - Some automation possible (bitext, text mining)
- Bush → 布殊
- Some large-scale commercial undertakings
  - Hundreds of millions of names and variants, primarily European
  - Similar efforts exist for CJK conversion via lookup
Transcoding

- Rote (mostly) character-by-character symbol conversion (e.g. Unix recode)
- x44 x61 x6e → xee xb3 xdd
- Even codes within a language vary
  - 布什 (Mainland China)
  - 布希 (Taiwan)
  - 布殊 (Hong Kong)
- Osama bin Laden: 10 Hanzi variants
- Unicode helps, but does not solve the problems
Transcription

• Conversion: (spoken) words → script
  • SAMPA (ASCII)
  • International Phonetic Alphabet (linguistics)
    • Bush → buʃ
  • Usually spoken language = transcribed language

• Sometimes as a strategy for crosslinguistic textual conversion

• Variation is a problem: whose dialectal/idiolectal pronunciation should be used?
Transliteration

• Rewrite symbols of source language in target alphabet
  • Bush → Буш
• Source/target sounds don’t always align
  • 32 English spellings for Muammar Gaddafi
  • 6 Arabic spellings for Clinton
• Sensitive to properties of target language
  • e.g. Yuschenko vs. louchtchenko
• Romanization chaos: scores of schemes
Transduction

- Mapping variable correspondences (transcription, transliteration), often (probabilistic) rule-based
- Implemented via algorithmic finite-state automata
  - e.g. Soundex (Russell, American, Daitch-Mokotoff), others
- Bush → बुश

<table>
<thead>
<tr>
<th>Alternate spellings based upon easily confused letters</th>
<th>American soundex alternatives</th>
<th>Daitch-Mokotoff soundex alternatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bcller, Bebler, Beiler, Belber, Belier, Bellcr, Bellen, Bellor, Boller, Bcbler, and 152 others...</td>
<td>Beler, Beller</td>
<td>Aueler, Beler, Fbeler, Feler, Peler, Pfeler, Ppheler, Veler, Weler</td>
</tr>
</tbody>
</table>
Problems with Soundex

- Long names: Sivaramakrishnarao, Sivaramakrishnan, Sivaramarao
- Implausible collapses
- Anglocentric
- Alphabetic-based
- Not very efficient distributionally
Translation

- Most widely used when logographic system is used
  - Names are rendered non-literally, non-phonemically to/from logograph (sequence)
- Great Salt Lake → 大鹽湖
- Creative, most opaque of mapping schemes
Common techniques used

- Machine learning
  - Statistical/stochastic approaches (e.g. n-grams)
  - Rule-based transformational approaches
  - Entropy/noisy channel approaches
- String matching algorithms
  - Levenshtein edit distance (similarity measure)
  - Dynamic programming techniques
- Speech processing (recognition, TTS)
- Bitext mining, alignment metrics, indexing
What’s the best method?

• One of schemes listed previously
  • All approaches are information-losing propositions

• Hybrid approaches combining several of these
  • Pipeline results
  • Poll different engines for optimal results

• How to generalize beyond a handful of languages?