In this lecture…

- Overview of the characteristics of different languages/language types
  - Languages vs. dialects
  - Language families
  - Classification/typology, linguistic terminology
  - Writing Systems
  - Spoken vs. Written Languages
How many languages exist in the world?

• Google says:
  – 2,796
  – “almost 5,000”
  – 6,800
  – 4,000-6,000
  – 4,000-10,000
• The truth is: we don’t really know

Language Statistics

• Not sure that all languages have been discovered yet
• E.g. the case of Middle Chulym in 2004: new language or not?
  http://www.languagehat.com/
Language Statistics

• Do we count both living and dead languages?
• “dead” language: no surviving speakers that have acquired language during childhood
• But: language may still be widely used in certain contexts (eg. Latin, Ancient Greek, Coptic,…)

Language Statistics

• **Dialect:** textbook definitions
  – Regional variant of a language
  – Modifications at the lexical and grammatical level
  – Dialects are mutually intelligible
  – Spoken variety without literary tradition
• **Accent:**
  – Modifications at the phonetic/phonological level only
• **Language:**
  – Languages are mutually unintelligible
  – Have a literary tradition
Language Statistics

- Problems with textbook definitions
  - Oral languages
  - Continuum between dialects and languages, distinction is arbitrary
  - Decision may be motivated sociopolitically
- Examples:
  - Dialects of Arabic, Chinese
  - Romanian vs. Moldovan

Language Statistics

- Standard database: Ethnologue
- 15th edition (2005) lists 6,912 known living languages
- Classification based on mutual intelligibility and ethnolinguistic identity
Language Statistics

• Twenty most widely spoken languages

<table>
<thead>
<tr>
<th>#</th>
<th>Language</th>
<th>Speakers (in millions)</th>
<th>#</th>
<th>Language</th>
<th>Speakers (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mandarin Chinese</td>
<td>867.2</td>
<td>11</td>
<td>Wu Chinese</td>
<td>77.2</td>
</tr>
<tr>
<td>2</td>
<td>Spanish</td>
<td>322.3</td>
<td>12</td>
<td>Javanese</td>
<td>75.5</td>
</tr>
<tr>
<td>3</td>
<td>English</td>
<td>309.4</td>
<td>13</td>
<td>Telugu</td>
<td>69.7</td>
</tr>
<tr>
<td>4</td>
<td>Arabic</td>
<td>206.0</td>
<td>14</td>
<td>Marathi</td>
<td>68.0</td>
</tr>
<tr>
<td>5</td>
<td>Hindi</td>
<td>180.8</td>
<td>15</td>
<td>Vietnamese</td>
<td>67.4</td>
</tr>
<tr>
<td>6</td>
<td>Portuguese</td>
<td>177.5</td>
<td>16</td>
<td>Korean</td>
<td>67.0</td>
</tr>
<tr>
<td>7</td>
<td>Bengali</td>
<td>171.1</td>
<td>17</td>
<td>Tamil</td>
<td>66.0</td>
</tr>
<tr>
<td>8</td>
<td>Russian</td>
<td>145.0</td>
<td>18</td>
<td>French</td>
<td>64.8</td>
</tr>
<tr>
<td>9</td>
<td>Japanese</td>
<td>122.4</td>
<td>19</td>
<td>Italian</td>
<td>61.5</td>
</tr>
<tr>
<td>10</td>
<td>German</td>
<td>95.4</td>
<td>20</td>
<td>Urdu</td>
<td>60.5</td>
</tr>
</tbody>
</table>

Language Statistics

• Most widely spoken languages by geographical origin

<table>
<thead>
<tr>
<th>Area</th>
<th>Number of languages</th>
<th>% (of all languages)</th>
<th>% speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>2,092</td>
<td>30.3</td>
<td>11.8</td>
</tr>
<tr>
<td>Americas</td>
<td>1,002</td>
<td>14.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Asia</td>
<td>2,269</td>
<td>32.8</td>
<td>61.0</td>
</tr>
<tr>
<td>Europe</td>
<td>239</td>
<td>3.5</td>
<td>26.3</td>
</tr>
<tr>
<td>Pacific</td>
<td>1,310</td>
<td>18.9</td>
<td>0.1</td>
</tr>
<tr>
<td>Total</td>
<td>6,912</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Linguistic Classification

- Ways of classifying/describing languages
  - Historical relatedness ("language trees", "language families")
  - Linguistic characteristics (typology)

- Are not necessarily correlated

- Knowledge of typological relatedness can help in developing language technology

Language Families

- Katzner (2002): 21 major language families

- 5 largest:
  - Indo-European
  - Afro-Asiatic
  - Niger-Congo
  - Sino-Tibetan
  - Austronesian

- Many smaller families (eg Dravidian, Australian, American Indian)

- Languages with no known affiliation (Basque, Ainu,…)
Language Families

• Indo-European Family

Language Families

• Afro-Asiatic Family
### Language Families

#### Niger-Congo

<table>
<thead>
<tr>
<th>Mande</th>
<th>Kordofanian</th>
<th>Atlantic-Congo</th>
<th>Volta-Congo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malinke</td>
<td>Ko</td>
<td>Fulani</td>
<td>Bantu languages</td>
</tr>
<tr>
<td>Mende</td>
<td>Shwai</td>
<td>Wolof</td>
<td>Swahili</td>
</tr>
<tr>
<td>...</td>
<td>Moro</td>
<td>Serer</td>
<td>Kikuyu</td>
</tr>
<tr>
<td>...</td>
<td>Tira</td>
<td>Temne</td>
<td>Ruanda</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>Yoruba</td>
</tr>
</tbody>
</table>

#### Sino-Tibetan

<table>
<thead>
<tr>
<th>Chinese</th>
<th>Tibeto-Burman</th>
</tr>
</thead>
</table>
| Mandarin Wu Cantonese Min Xiang Hakka Gan | Bodish Himalayan Luish Pyu Kuki Lepcha Burmish ...
|
Language Families

• Austronesian

Borneo-Philippines
Tagalog
Cebuano
Malagasy ...

Malayo-Polynesian

Sunda-Sulawesi
Cham
Malay
Javanese
Balinese
Chamorro ...

Central-Eastern

Central
Timorese
...

Eastern

South

Buli
Biak ...

Oceanic

Tahitian
Tongan
Maori

Language Families

• “artificial” languages
  – Artificially designed languages: Esperanto
  – Pidgins/creoles: developed as a means of communication between different linguistic communities, e.g. Haitian Creole (French + West African languages), Tok Pisin (English + Austronesian languages in Papua New Guinea)

• Lingua franca: non-native language used as a means of communication between native speakers of different languages, can be an existing language
Linguistic Classification

- Phonetics/Phonology
- Prosody
- Morphology
- Syntax/Word Order
- Sociolinguistic aspects
- Writing System

Morphology

- **Morpheme**: smallest meaning-bearing unit in a language, type
- **Morph**: concrete instance of a morpheme, token
- **Allomorph**: variants belonging to the same morpheme, often context-dependent
- **Word**:
  - delimited by whitespace
  - Can be shifted to different positions within sentence
  - Combines to larger syntactic phrases
  - Has meaning, can have internal structure
Morphology

• **Free** morphs: can occur on their own
• **Bound** morphs: need to attach to other morphs
• Word formation processes:
  – **Compounding**: combines two free morphs
  – **Derivation**: combines free + bound morphs to produce word of new class
  – **Inflection**: combines free + bound morphs to produce word of same class

Morphology

• **Stem**: non-minimal morph capable of taking affixes
• **Root**: minimal morph capable of taking affixes
• **Clitics**: cannot exist on their own, must attach to words
Morphology

- Word formation processes:
  - **Affixation**: concatenation of morphs
  - **Clipping**: shortening a word
  - **Reduplication**: reduplication of parts of words
  - **Conversion**: change of part-of-speech
  - **Back-formation**: decomposing a non-compound word
  - **Blending**: combining phonological material of two words
  - **Acronyms, abbreviations**

Morphology

- Languages can be classified according to primary means of word formation
  - **Isolating** languages: form sequences of invariable free morphemes, “no morphology”
  - **Agglutinating** languages: several morphemes per word, linear segmentation of word into morphemes
  - **Fusional** languages: combination of several morphemes per word with drastic change of word form
- Special case: **templatic morphology** (root-and-pattern morphology)
- Most languages have several types of word formation
Syntax

- Languages can be classified according to predominant type of sentence and phrase structure
- Usually according to position of subject (S), verb (V), object (O)
- All possible 6 word orders occur in world’s languages: SOV, SVO, VSO, VOS, OVS, OSV
- SOV and SVO much more frequent than others
- Most languages have several types, with one predominant (more frequent) type

Syntax

- Phrasal structure: classified according to position of head vs. modifiers
  - Modifiers precede head: premodification
  - Modifiers follow head: postmodification
- Primary syntactic categories (word classes, parts-of-speech): N, V, Adj, Adv,…
- Secondary grammatical categories: tense, number, aspect, voice, gender,…
Syntax

- **Tense**: past, present, future
- **Number**: singular, plural
- **Aspect**: present, progressive, habitual,…
- **Voice**: active, passive
- **Gender**: masculine, feminine, neuter,
- **Animacy**: animate, inanimate,…

Sociolinguistic aspects

- **Idiolect**: set of speech patterns specific to one speaker
- **Sociolect**: set of speech patterns specific to a societal group
- **Diglossia**: coexistence of a high-prestige and a low-prestige linguistic variety within one social group
- **Code-switching**: switching between two varieties or languages within same sentence or phrase
Sociolinguistic aspects

- **Honorifics**: special linguistic forms indicating status of speaker/addressee/topic
  - Titles
  - Affixes
  - Inflections
  - Different lexical choices

Writing Systems

- **Writing system**: set of notational symbols + conventions for using them
  - Conveys writer’s message in exact words
- **Pictographic/ideographic systems**: convey message but not in exact words
- **Grapheme**: basic unit in a writing system
- Systems are classified according to types of graphemes they use
Writing Systems

Pictographic writing system
Yukaghir
(Sampson 1985, p. 28)

Writing Systems

writing systems

logographic

phonographic

morphemic polymorphemic syllabic segmental featural
Writing Systems

- **Logographic**: graphemes are semantic units
  - **Morphemic**: (roughly) 1-to-1 correspondence between graphemes and morphemes
  - **Polymorphemic**: grapheme represents several morphemes at once
- **Phonographic**: graphemes represent sound units
  - Further distinction depending on size of sound unit (syllabic, segmental, featural)

Logographic systems: Chinese Hanzi

```
我来自湖南省
```

“I come from Hunan”

- Characters represent words/morphemes
- No whitespace delimiters
- Historically derived from pictograms
- Today: pictographic characters < 3%, most characters used phonetically or in phonographic-logographic combinations
- Word segmentation crucial problem for Chinese NLP
Writing Systems

• Phonographic systems:
  – all Western alphabets (Greek, Roman, Cyrillic,…) –
    segmental subtype, graphemes correspond roughly to
    phone segments (though many ambiguities)
  – Important subtype: consonantal writing systems
    (abjads)
    • only write consonants (and possibly long vowels)
    • short vowels and other phonetic material can be indicated by
      optional diacritics
    • Typically used for Semitic languages but have been adopted to
      a number of other languages as well (Farsi, Hausa, Urdu,…)

Example: Arabic script with/without diacritization:

أَحَبَّ السَّفَرُ إِلَى الْقَاهِرَةُ
أَحَبَّ السَّفَرُ إِلَى الْقَاهِرَةُ

“I like to travel to Cairo”
Writing Systems

• Syllabic systems: graphemes stand for entire syllables
  – For pure syllabic script, need as many graphemes as syllables
  – Historical syllabic scripts: e.g. Linear B (Ancient Greek)
  – Present-day: Japanese Kana
    • Kana = Hiragana / Katakana
    • Kanji = logograms from Chinese
    • Mixture of Kanji and Kana used in practice

Writing Systems

Japanese hiragana

```
 a あ i い u う e ে o お
ka か ki き ku く ke け ko こ
sa さ shi し su す se せ so そ
ta た chi ち tsu つ te て to と
na な ri に nu ぬ ne ね no の
ha は hi ひ fu ふ he へ ho ほ
ma ま mi み mu む me め mo も
ya や yu ゆ yo よ
ra ら ri り ru る re れ ro ろ
Wa わ Wi わ we わ wo わ
```
Writing Systems

• Mixture of syllabic and segmental systems: **alphasyllabaries (abugidas)**
  – Graphemes represent syllables: consonant plus one inherent (default) vowel
  – Other graphemes are derived from basic ones by modifications, e.g. diacritics, rotation, etc.

• Example: Devanagari and derived forms (used for e.g. Khmer, Thai, Gujarati,...)

Writing Systems

Lao abugida

<table>
<thead>
<tr>
<th>重要作用</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
<th>重要</th>
</tr>
</thead>
<tbody>
<tr>
<td>重要作用</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
</tr>
<tr>
<td>重要作用</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
<td>重要</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ng</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
<th>重要作用</th>
</tr>
</thead>
<tbody>
<tr>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
</tr>
<tr>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
<td>重要作用</td>
</tr>
</tbody>
</table>
Writing Systems

Lao abugida

<table>
<thead>
<tr>
<th>ka</th>
<th>ki</th>
<th>ku</th>
<th>ku'</th>
<th>ka:</th>
<th>ki:</th>
<th>ku:</th>
<th>ku:'</th>
<th>ke</th>
<th>kae</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ka]</td>
<td>[ki]</td>
<td>[ku]</td>
<td>[ku]</td>
<td>[ka]</td>
<td>[ki]</td>
<td>[ku]</td>
<td>[ku]</td>
<td>[ke]</td>
<td>[kae]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ko</th>
<th>ke:</th>
<th>kae</th>
<th>koa</th>
<th>ko'</th>
<th>koe</th>
<th>kia</th>
<th>kia</th>
<th>kua</th>
<th>koe:y</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ko]</td>
<td>[ke]</td>
<td>[kae]</td>
<td>[koa]</td>
<td>[ko]</td>
<td>[koe]</td>
<td>[kia]</td>
<td>[kia]</td>
<td>[kua]</td>
<td>[koe:j]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>koe:y</th>
<th>ko'</th>
<th>koe</th>
<th>kua</th>
<th>kaw</th>
<th>kay</th>
<th>kiy</th>
<th>kam</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>[koe:j]</td>
<td>[ko]</td>
<td>[koe]</td>
<td>[kua]</td>
<td>[kaw]</td>
<td>[kay]</td>
<td>[kaj]</td>
<td>[kam]</td>
<td>[k]</td>
</tr>
</tbody>
</table>

Writing Systems

- Featural writing system: graphemes represent elements smaller than the phone: articulatory/phonetic features
  - Rare
  - Most prominent example: Korean Hangul
    - Basic strokes (stylized shapes) for phonetic features of corresponding sound
    - Vowel strokes combine with consonant strokes to form syllables
    - Syllables are surrounded by whitespace
    - In practice, Hangul is used in combination with Chinese Hanzi
Writing Systems

Korean Hangul

<table>
<thead>
<tr>
<th>IPA</th>
<th>p</th>
<th>t</th>
<th>c</th>
<th>k</th>
<th>pʰ</th>
<th>tʰ</th>
<th>cʰ</th>
<th>kʰ</th>
<th>r</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hangul</td>
<td>ṭVERN0</td>
<td>ṭ</td>
<td>ṭN</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭN</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
</tr>
<tr>
<td>IPA</td>
<td>m</td>
<td>n</td>
<td>ŋ</td>
<td>s</td>
<td>b</td>
<td>d</td>
<td>ŋ</td>
<td>g</td>
<td>z</td>
<td></td>
</tr>
<tr>
<td>Hangul</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td></td>
</tr>
<tr>
<td>IPA</td>
<td>i</td>
<td>e</td>
<td>ē</td>
<td>a</td>
<td>o</td>
<td>u</td>
<td>Ø</td>
<td>w</td>
<td>wj</td>
<td>je</td>
</tr>
<tr>
<td>Hangul</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
</tr>
<tr>
<td>IPA</td>
<td>je</td>
<td>ja</td>
<td>jo</td>
<td>ju</td>
<td>jØ</td>
<td>wi</td>
<td>we</td>
<td>we</td>
<td>wa</td>
<td>wØ</td>
</tr>
<tr>
<td>Hangul</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
<td>ṭ</td>
</tr>
</tbody>
</table>

Writing Systems

- **Transcription**: rendering spoken language in writing
- **Transliteration**: conversion from one writing system into another, e.g. from Arabic script into Western alphabet
- **Romanization**: transliteration into Western alphabet, e.g.
  - Chinese: zhuyin, pinyin
  - Japanese: Nishon-shiki
  - Korean: Revised Romanization of Korean
Writing Systems

• Digital text processing: many writing systems have more characters than keys on a standard keyboard
• Before 1990, various encodings, often language-specific; no single encoding could handle all languages; conflicting encodings for same language
• Need for standard

Writing Systems

• ISO standard (International Organization for Standardization)
  – 16 different encodings for major European languages, Arabic, Hebrew (ISO 8859-1-ISO 8859-16)
• Microsoft: internal code pages (CP…)
• Indian Script Code for Information Interchanges (ISCII)
• Since 1991: Unicode
Writing Systems

• Universal character encoding for all living/historical languages, additional symbol sets
• Independent of platforms/operating systems
• Specifies abstract code elements and their order
• Needs to be interpreted by Unicode-compatible software to display actual characters
• Used in most web browsers, text editors, email tools, etc. nowadays

Phonetics/Phonology

• **Phonetics**: study of physical properties of sounds (acoustic, articulatory, perceptual, physical)
• **Phonology**: study of the sound **systems** of languages
• **Phoneme**: smallest unit in a language capable of distinguishing meaning (in a minimal pair)
• **Phone**: concrete realization of a phoneme, sound segment
• **Allophone**: positional variants of phonemes, in complementary distribution
Phonetics

Phonetic description: consonant

THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

<table>
<thead>
<tr>
<th>Plosive</th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Postalveolar</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>t̪̊</td>
<td>d̪̊</td>
<td>c</td>
<td>j̠̊</td>
<td>k̠̊</td>
<td>g̠̊</td>
<td>q̠̊</td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>n</td>
<td>n̩</td>
<td>m̩</td>
<td>n̩</td>
<td>n̩</td>
<td>n̩</td>
<td>n̩</td>
<td>n̩</td>
<td>n̩</td>
</tr>
<tr>
<td>Trill</td>
<td>b</td>
<td>r</td>
<td>r̩</td>
<td>r̩</td>
<td>r̩</td>
<td>r̩</td>
<td>r̩</td>
<td>r̩</td>
<td>r̩</td>
<td>r̩</td>
</tr>
<tr>
<td>Fricative</td>
<td>θ̊</td>
<td>f</td>
<td>v̊</td>
<td>ɵ̊</td>
<td>s̊</td>
<td>z̊</td>
<td>t̊</td>
<td>s̊</td>
<td>z̊</td>
<td>s̊</td>
</tr>
<tr>
<td>Lateral Fricative</td>
<td>l̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
</tr>
<tr>
<td>Approximant</td>
<td>ů</td>
<td>j̊</td>
<td>j̊</td>
<td>j̊</td>
<td>j̊</td>
<td>j̊</td>
<td>j̊</td>
<td>j̊</td>
<td>j̊</td>
<td>j̊</td>
</tr>
<tr>
<td>Lateral Approximant</td>
<td>l̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
<td>ɾ̊</td>
</tr>
</tbody>
</table>

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

Phonetics

Phonetic description: vowels

VOWELS

Front

Close

Close-mid

Open-mid

Open

Central

Back

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

• Language classification often based on phonotactics: constraints on phoneme inventories and combinations
  – complexity of consonant clusters/syllable structure
  – E.g. English, German, etc. allow wide range of syllable structures, Maori only allows CV, V, VV

Prosody

• Phenomena that extend over more than one phonetic segment: pitch, duration, rhythm,…
  • Pitch: perceptual correlate of fundamental frequency (F0) – rate of opening/closing of vocal folds
  • Relevant for tone languages and intonation languages
Prosody

- Tone languages: pitch contour of a word is lexically or grammatically distinctive

<table>
<thead>
<tr>
<th>Word</th>
<th>Tone</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>mā</td>
<td>high level</td>
<td>mother</td>
</tr>
<tr>
<td>má</td>
<td>high rising</td>
<td>hemp</td>
</tr>
<tr>
<td>mǎ</td>
<td>falling-rising</td>
<td>horse</td>
</tr>
<tr>
<td>mà</td>
<td>falling</td>
<td>to scold</td>
</tr>
</tbody>
</table>

Prosody

- Intonation languages: use prosody to signal sentence type, indicate phrase boundaries, and contrastive emphasis
- Often have particular word-stress patterns
  - Fixed – e.g. word-initial (Hungarian) or word-final (French)
  - Free
Problems for NLP

• Multitude of languages – some languages have rich data resources but most don’t
  – Data-sparse approaches for text classification, language modeling, machine translation, annotation (eg POS tagging), etc

• Most languages are oral languages: problems of creating data resources, devising writing systems, etc.

• Plethora of writing systems: how to represent language data in consistent machine-readable form, issues of transliteration, grapheme-to-phoneme conversion

Problems for NLP

• Word definition: many algorithms take word as fundamental modeling unit
  – Languages without whitespace delimiters: problem of finding word segmentation automatically
  – Languages with rich morphology: impractical to take every token delimited by whitespaces as a word: parameter explosion
Problems for NLP

• Variability of language with respect to dialect, topic, speaker/writer, addressee
  – Language modeling
  – Dialogue systems
  – Localization of language software