Languages of the World

Kevin Duh
June 9, 2014
Goals of this lecture

1. Appreciate the *diversity* of languages

2. Discuss some important linguistic *phenomenon* and *classifications* may help you with your Natural Language Processing research
Outline

1. What is a language?
2. Language Change
3. World Tour
4. Language Universals
What is a language?

• A language is “a product of the collective mind of linguistic groups” -- Ferdinand de Saussure


• “A language is a dialect with an army and navy” – Max Weinreich
  – E.g. Chinese “dialects”, Scandinavian “languages”

  From: http://epyc.yivo.org/content/12_1.php
Definition of language in terms of “Mutual Intelligibility”

• Two caveats:
  – Dialect continuum: A string of dialects may be mutually intelligible, but not transitive
    • E.g. Dutch-German dialect continuum
  – It’s a matter of degree, no clear-cut intelligibility test

• There’s no such thing as “languages”; “Dialects” are all there is.
  – One dialect defined as “standard” language
    • E.g. Tokyo dialect as “Japanese”
Numbers to Know:
How many languages in the world?

- **Conservative estimate**: 6000
  - Peak of diversity: 10,000-15,000 (~15,000 BCE)
- **Skewed distribution**

<table>
<thead>
<tr>
<th>Population range</th>
<th># of Languages</th>
<th>Percentage of world population</th>
</tr>
</thead>
<tbody>
<tr>
<td>100,000,000+</td>
<td>8</td>
<td>40%</td>
</tr>
<tr>
<td>10,000,000-99,999,999</td>
<td>80</td>
<td>39%</td>
</tr>
<tr>
<td>1,000,000-9,999,999</td>
<td>305</td>
<td>14%</td>
</tr>
<tr>
<td>100,000-999,999</td>
<td>93</td>
<td>4%</td>
</tr>
<tr>
<td>10,000-99,999</td>
<td>1,811</td>
<td>0.9%</td>
</tr>
<tr>
<td>1,000 -9,999</td>
<td>1,978</td>
<td>0.1%</td>
</tr>
<tr>
<td>100-999</td>
<td>1,062</td>
<td>0.007%</td>
</tr>
<tr>
<td>1-99</td>
<td>475</td>
<td>0.0002%</td>
</tr>
</tbody>
</table>

Pause and think about this for a bit

What I say here can be expressed equivalently in 6000 other ways, using completely different words and grammar!
Numbers to know:
Largest language by # of speaker

<table>
<thead>
<tr>
<th>Language</th>
<th># of L1 Speakers (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>1,197</td>
</tr>
<tr>
<td>Spanish</td>
<td>414</td>
</tr>
<tr>
<td>English</td>
<td>335</td>
</tr>
<tr>
<td>Hindi</td>
<td>260</td>
</tr>
<tr>
<td>Arabic</td>
<td>237</td>
</tr>
<tr>
<td>Portuguese</td>
<td>203</td>
</tr>
<tr>
<td>Bengali</td>
<td>193</td>
</tr>
<tr>
<td>Russian</td>
<td>167</td>
</tr>
<tr>
<td>Japanese</td>
<td>122</td>
</tr>
<tr>
<td>Javanese</td>
<td>84</td>
</tr>
</tbody>
</table>

Numbers to know:
When did language arise?

200,000 years ago: Anatomically modern humans

50,000 years ago: Behavioral Modernity
  *Language enables cooperation & gossip → larger social groups*

12,000 years ago: Agricultural Revolution

Language arose here?
Or here?

And is there a Language Instinct?

Disclaimer: Dates are inexact. I’m not an expert and there appears to be no definitive answer.
Outline

1. What is a language?
2. Language Change
3. World Tour
4. Language Universals
Change is the cause of diversity

• **Change by Natural Evolution**
  – Slight differences in speaking (usually due to Laziness) leads to large differences after generations
  – E.g. Sound change, re-bracketing, semantic shift

• **Change by Contact (Areal Effect)**
  – Borrowing of phonology, lexicon, and grammar from neighboring languages
    • E.g. Balkan Sprachbund: Albanian, Greek, Romanian, Bulgarian, Macedonian
      → verb-Not-verb, post-article, genitive & dative merger
Sound change

• Principle of least effort, e.g.:
  – “God be with you” → God b’wy → Goodbye
  – Loss of case-endings in Latin → Necessity of word order for grammatical function in English
  – Loss/merger of consonants in Old Chinese → Necessity of Tones

• General change, e.g.:
  – Great Vowel Shift (1350-1700, England)
    • “bite” biːt → balt; “beet”: beːt → biːt
Extension of Grammatical Patterns due to sound change

• Latin had multiple plural rules:
  – sorōrēs “sisters”
  – fēmina → fēminae “women”
  – dominus → domini “master”

• In French, only one plural ending was left due to sound erosion, so -s was extended
Morphological Type Change

Loss of inflection, e.g. Old English → Modern English

Morphemes fuse

Words become grammaticalized as affixes

From: http://languagesoftheworld.info/historical-linguistics/more-on-word-order-morphological-types-and-historical-change.html
Outline

1. What is a language?
2. Language Change
3. World Tour
4. Language Universals
Our Itinerary

• I’ll introduce various language families while we tour the world
  – Note: Don’t confuse geographical and genetic classification; e.g. Languages in Eurasia != Indo-European languages
• For each language family, I’ll point out some interesting phenomena or trivia
  – Warning 1: These phenomena are by no means unique to the language under discussion. May appear elsewhere.
  – Warning 2: Due to time limitation, not all important phenomena will be discussed. Our tour is 走馬看花 style: “viewing the flowers while riding a fast horse”
Indo-European Language Family

Indo-European

**Germanic:** English, German, Swedish, etc.

**Armenian:** Armenian

**Balto-Slavic:** Lithuanian, Russian, Polish, Czech, etc.

**Italic:** Italian, French, Spanish, Romanian, etc.

**Albanian:** Albanian

**Celtic:** Gaelic, Scottish

**Hellenic:** Greek

**Indo-Iranian:** Farsi, Hindi, Bengali, Marathi, etc.
# Discovery of the Indo-European Family

1796: **Sir William Jones** noticed similarity between Sanskrit & Latin

## Comparative Reconstruction:
- Cognates from basic vocabulary (body parts, kinship, nature)
- Identify patterns of sound change & correspondence

<table>
<thead>
<tr>
<th>Language</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irish</td>
<td>aon</td>
<td>do</td>
<td>tri</td>
</tr>
<tr>
<td>Greek</td>
<td>hen</td>
<td>duo</td>
<td>treis</td>
</tr>
<tr>
<td>Latin</td>
<td>unus</td>
<td>duo</td>
<td>tres</td>
</tr>
<tr>
<td>Italian</td>
<td>uno</td>
<td>due</td>
<td>tre</td>
</tr>
<tr>
<td>French</td>
<td>un</td>
<td>deux</td>
<td>trois</td>
</tr>
<tr>
<td>German</td>
<td>einz</td>
<td>zwei</td>
<td>drei</td>
</tr>
<tr>
<td>Swedish</td>
<td>en</td>
<td>tva</td>
<td>tre</td>
</tr>
<tr>
<td>Russian</td>
<td>odin</td>
<td>dva</td>
<td>tri</td>
</tr>
<tr>
<td>Bengali</td>
<td>ek</td>
<td>dvi</td>
<td>tri</td>
</tr>
<tr>
<td>Persian</td>
<td>yak</td>
<td>do</td>
<td>se</td>
</tr>
<tr>
<td>ProtoIE?</td>
<td>Hoi-no?</td>
<td>duwo?</td>
<td>trei?</td>
</tr>
<tr>
<td>Turkish</td>
<td>bir</td>
<td>iki</td>
<td>üc</td>
</tr>
<tr>
<td>Hebrew</td>
<td>‘exad</td>
<td>šnaim</td>
<td>šlosa</td>
</tr>
</tbody>
</table>

Finno-Ugric Family: Finnish, Hungarian, Estonian, etc.

Geographic discontinuity is interesting:
- Urals: probable homeland
- Finnic branch was larger but encroachment by Slavic
- Hungarian branch due to Magyar migration (800CE)
### Finno-Ugric: Agglutinative Morphology

14 cases in Estonian, 15 cases in Finnish, 21 cases in Hungarian:

<table>
<thead>
<tr>
<th>Case</th>
<th>Hungarian Word</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominative</td>
<td>hajó</td>
<td>ship [subject]</td>
</tr>
<tr>
<td>Accusative</td>
<td>hajó-t</td>
<td>ship [object]</td>
</tr>
<tr>
<td>Inessive</td>
<td>hajó-ban</td>
<td>in a ship</td>
</tr>
<tr>
<td>Elative</td>
<td>hajó-ból</td>
<td>out of a ship</td>
</tr>
<tr>
<td>Illative</td>
<td>hajó-ba</td>
<td>into a ship</td>
</tr>
<tr>
<td>Superessive</td>
<td>hajó-n</td>
<td>on a ship</td>
</tr>
<tr>
<td>Delative</td>
<td>hajó-ról</td>
<td>about a ship</td>
</tr>
<tr>
<td>Sublative</td>
<td>hajó-ra</td>
<td>onto a ship</td>
</tr>
<tr>
<td>Adessive</td>
<td>hajó-nál</td>
<td>by a ship</td>
</tr>
<tr>
<td>Ablative</td>
<td>hajó-tól</td>
<td>from a ship</td>
</tr>
</tbody>
</table>

*Note: many of these are encoded by prepositions in Indo-European languages (average 6 cases)*
Basque

• Unrelated to any other language?
• Ergative-absolutive case system
Dravidian Language Family

Distinct from Indo-European in northern India

**Some Characteristics:**
- Rigid SOV word order
- Nouns gender:
  - “rational” (refers to human, deity) vs.
  - “irrational” (refers to children, animal, objects)

Languages of the Caucasus region

Many different language families in this small area!

Trivia: Chechen has 40-60 consonants, ~44 vowels

From: http://en.wikipedia.org/wiki/Languages_of_the_Caucasus
Altaic Language Family (?)

- Macro-family consisting of possibly Turkic, Mongolic, Tungusic
  - Korean & Japanese?
  - Similarities due to genetics or contact?
Vowel Harmony in Turkic

• Turkic: Turkish, Uzbek, Kazakh, Dolgan, etc.

• Vowel Harmony:
  – long-distance assimilation where vowels become similar across intervening consonants in some way
  – E.g. back/front & rounded/unrounded harmonization in Turkish:

  Türkiye’dir  “it is Turkey”
  kapıdir      “it is the dor”
  günþür       “it is the day”
  paltoþur     “it is the coat”
Semitic Language Family: Hebrew, Arabic dialects, Aramaic, Amharic, etc.
Non-concatenative morphology in Semitic (e.g. Arabic)

- Root: 2-4 consonant; Template: vowels in-between
- ktb "write" (as verb)
  - ti-ktib "she writes"
    (prefix ti- means "she", present form is "- - i -")
  - katab-it "she wrote"
    (suffix -it means "she"," past form is "- a - a -")
  - kaatib "writing"
    (present participle "- aa - i -")
  - ma-ktuub "written"
    (past participle "- - uu -")
- ktb "book" (as noun)
  - kitaab: (- i - aa – singular)
  - kutub: (- u - u – plural)
Languages in Sub-Saharan Africa

- Nilo-Saharan
- Niger-Congo
- Khoisan

Characteristics:
- Many are tonal, have large sound inventories and “exotic” sounds, e.g. implosives, clicks
- Large noun classes (Shona: 20)

From: http://en.wikipedia.org/wiki/Languages_of_Africa
Sino-Tibetan Language Family

Tibetan branch: e.g. Tibetan, Burmese

Sinitic branch:

Characteristics:
- Tone
- Isolating morphology
- Noun Classifiers
  - numeral-classifier-noun in Mandarin
  - noun-numeral-classifier in Burmese

Tai-Kadai Family
e.g. Thai – tone (5), isolating, noun classifier

Austro-Asiatic Family
e.g. Vietnamese – tone (6), isolating, noun classifier, 30% vocab via Chinese
e.g. Munda – no tone, agglutinative

Likely areal effects
Austronesian Languages

- **Formosan branch:** ~20 languages in Taiwan (many endangered)
- **Malayo-Polynesian branch:**
  - West: Javanese, Sundanese, Malay, Indonesian, Tagalog, Malagasy, etc.
  - East: Hawaiian, Maori, Fijian, etc.

**Characteristics:**
- Ergative-Absolutive
- Agglutinative morphology
- Small sound inventory: (13 phoneme in Hawaiian)
- Some have VOS, VSO order
- Inclusive/Exclusive 1st person pronoun: “we” includes hearer?
- Reduplication

From: [http://en.wikipedia.org/wiki/Austronesian_languages](http://en.wikipedia.org/wiki/Austronesian_languages)
Reduplication

Sound repetition within a word for semantic or grammatical purpose

e.g. Tagalog:
  - sulat “write” → susulat “will write”
  - hanap “seek” → hahanap “will seek”
  - lakad “walk” → lalakad “will walk”

e.g. Indoneasian:
  - anak “child” → anak anak “all sorts of children”
  - oraN “man” → oraN oraN “all sorts of men”
Languages in Papua New Guinea:
• 800+ languages! (1 language per 200-900km²)
• Diversity due to mountains (natural barriers) and tribal society (cultural barriers)
• Tok Pisin (one of the official languages):
  – Pidgin arose from contact between English & locals
  – Pidgin becomes creole when children learn it as L1
  – Lexicon is mostly from English. Syntax is from where?

Languages in Australia:
• 270 languages, many near extinction
• Trivia - Noun classes in Dyirbal:
  I: masculine & animate; II: feminine, fire, fighting;
  III: all trees with edible fruit; IV: everything else
Languages of America

(there are attempts to group them into macro-families, but controversial)

Some Interesting Phenomena

- **Multiple Argument Agreement in Mohawk:**
  - Verb not only agrees with subject but also object
    - E.g. *shako-* prefix: agreement w/ 3rd person subject and 3rd person object; *ra-*: agreement with just 3rd person subject
  - Noun incorporation: noun root becomes part of the verb, and one less argument to agree with:
    - 3 words: *Wa’-k-hniui-’* (1sg-subj-BUY) *ne* (part) *ka-nakt-a’* (prefix-BED-suffix) → 1 word: *Wa’-ke-nakta-hninu-’*.

- **Three-way case marking in Nez Perce:**
  - Subjects of intransitives, subjects of transitives, objects of transitives → all get different case

- **OVS word order in Carib**

- **Evidential marker in Makah**
Outline

1. What is a language?
2. Language Change
3. World Tour
4. Language Universals
Linguistic Universals and Typology

• Typology: classifies language and aims to describe common properties and diversity
• E.g.: The following Word Orders are common.
  – **SOV**: Japanese, Tamil, Turkish (565 languages in wals.info)
  – **SVO**: Chinese, English, Fula (488 languages in wals.info)
  – **VSO**: Arabic, Tongan, Welsh (95 languages in wals.info)
• Why so few **VOS, OVS, OSV** (total <5%)?
  – Hypothesis: Subjects tend to precede Objects
    • Why? Maybe: Agent before Patient = better info flow
  – Note: some languages have V2 or no dominant order
Typological Generalizations

• SOV tendencies:
  – have postpositions
  – genitive-noun, etc.

• Analytical morphology tendencies:
  – mono-syllable words
  – use of tones
  – use of function words
  – relative fixed word order

• SVO tendencies:
  – have prepositions
  – noun-genitive, etc.

• Synthetic morphology tendencies:
  – poly-syllable words
  – no use of tones
  – fewer function words
  – relative free word order
Check out World Atlas of Language Structures (http://wals.info) for more!

Check out World Atlas of Language Structures (http://wals.info) for more!

Check out World Atlas of Language Structures (http://wals.info) for more!

In: Dryer, Matthew S. & Haspelmath, Martin (eds.)
Leipzig: Max Planck Institute for Evolutionary Anthropology.
(Available online at http://wals.info/chapter/37, Accessed on 2014-06-08.)
Summary

1. What is a language?
2. Language Change
3. World Tour
4. Language Universals
Good References


