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

From: http://en.wikipedia.org/wiki/File:Ferdinand_de_Saussure_by_Jullien.png

- "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 (~15,000BCE)
- Skewed distribution

Population range	# of Languages	Percentage of world population
100,000,000+	8	40%
10,000,000-99,999,999	80	39%
1,000,000-9,999,999	305	14%
100,000-999,999	93	4%
10,000-99,999	1,811	0.9%
1,000 -9,999	1,978	0.1%
100-999	1,062	0.007%
1-99	475	0.0002%

Source: Ethnologue - http://www.ethnologue.com/statistics/status

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

Language	# of L1 Speakers (in millions)
Chinese	1,197
Spanish	414
English	335
Hindi	260
Arabic	237
Portuguese	203
Bengali	193
Russian	167
Japanese	122
Javanese	84

Source: Ethnologue - http://www.ethnologue.com/statistics/status

Numbers to know:
When did language arise?

200,000 years ago: Anatomically modern humans

Language arose here?
Or here?

And is there a Language Instinct?

50,000 years ago: Behaviorial Modernity

Language enables cooperation & gossip → larger social groups

12,000 years ago: Agricultural Revolution

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

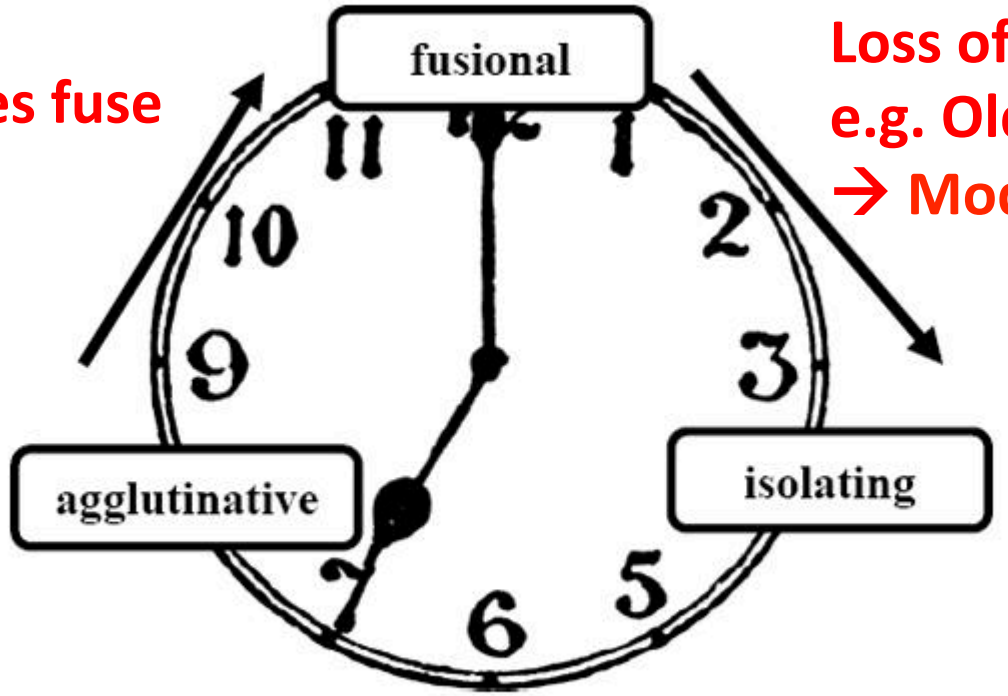
Sound change

- 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" bi:tə → balt; "beet": be:t → bi:t

Extension of Grammatical Patterns due to sound change

- Latin had multiple plural rules:
  - sorōr**ēs** “sisters”
  - fēmina → fēmin**ae** “women”
  - dominus → domin**i** “master”
- In French, only one plural ending was left due to sound erosion, so **-s** was extended

Morphological Type Change



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

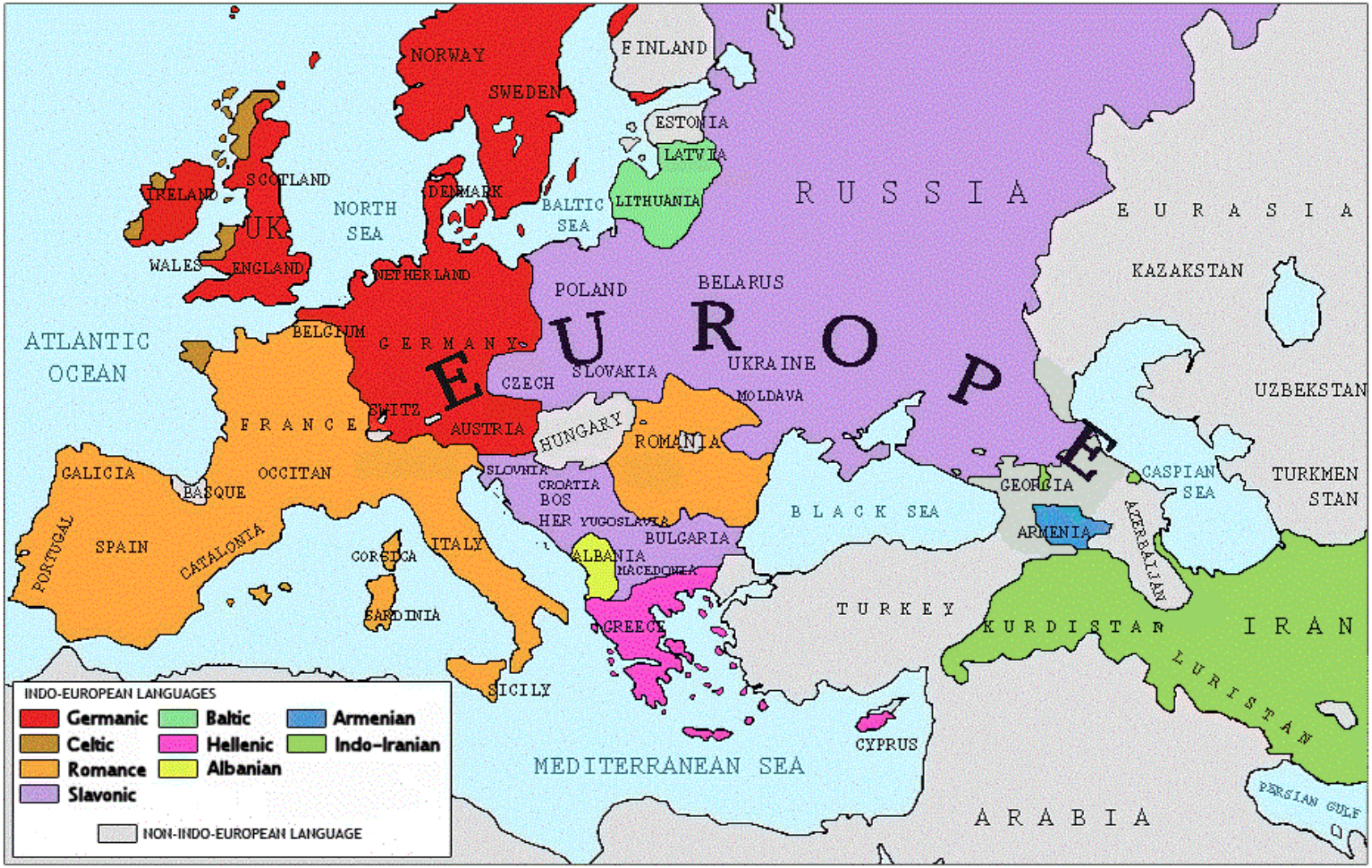
Outline

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



From: http://en.wikipedia.org/wiki/Indo-European_languages

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.

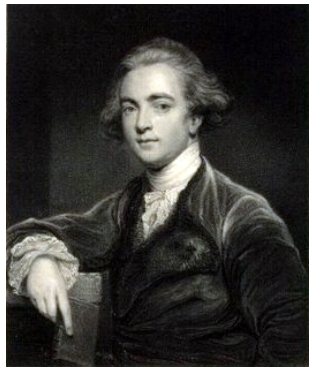
Discovery of the Indo-European Family

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

IE

Not IE

1796: Sir William Jones noticed similarity between Sanskrit & Latin



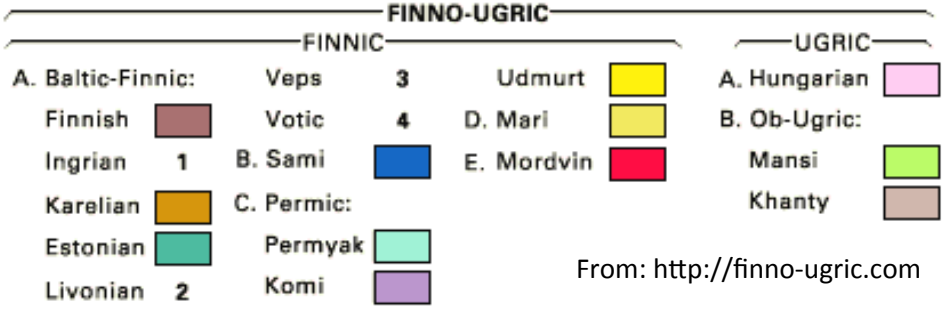
From: http://en.wikipedia.org/wiki/William_Jones_(philologist)

Comparative Reconstruction:

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

Finno-Ugric Family:

Finnish, Hungarian, Estonian, etc.



From: http://finno-ugric.com

<u>Geographic discontinuity</u>
is interesting:
- Urals: probable homeland
- Finnic branch was larger but encroachment by Slavic
- Hungarian branch due to Magyar migration (800CE)

Finno-Ugric: Agglutinative Morphology

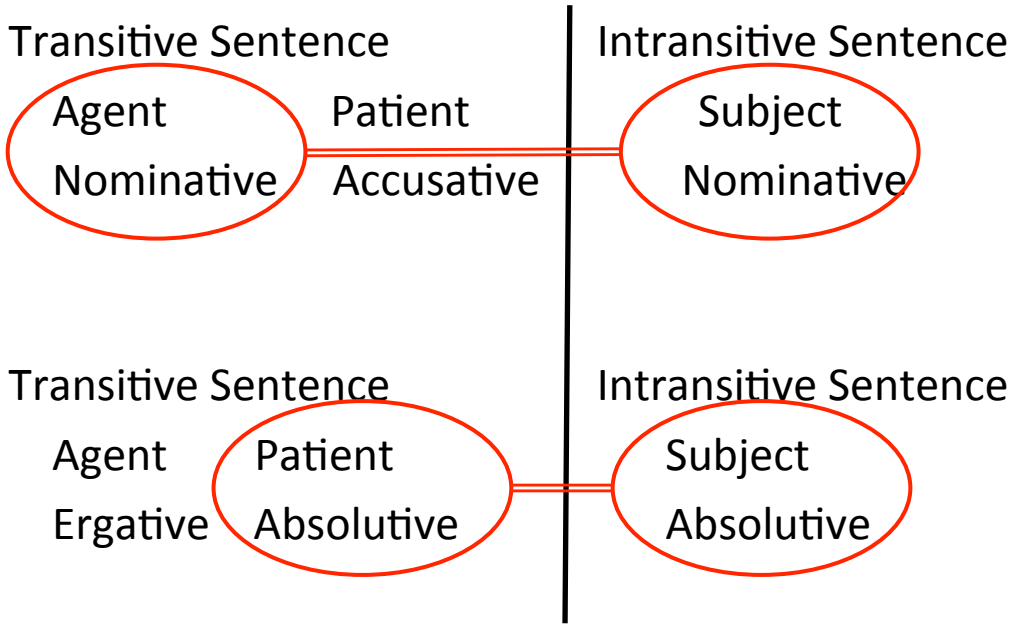
14 cases in Estonian, 15 cases in Finnish, 21 cases in Hungarian:

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

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

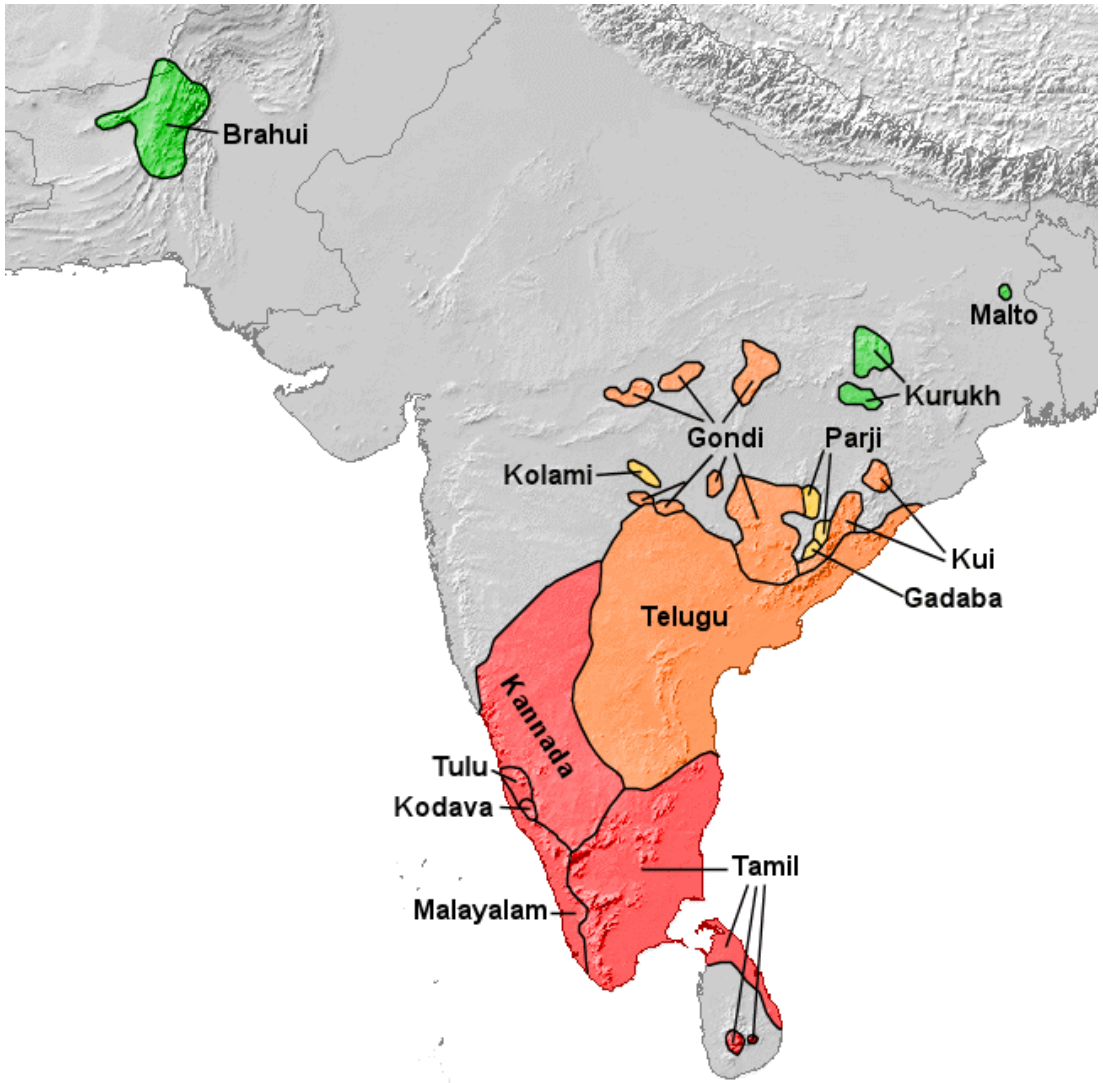
Basque

- Unrelated to any other language?
- Ergative-absolutive case system



From: http://en.wikipedia.org/wiki/Basque_language

Dravidian Language Family



From: http://en.wikipedia.org/wiki/Dravidian_languages

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 Caucuses 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, Tungustic
  - Korean & Japanese?
  - Similarities due to genetics or contact?

From: http://en.wikipedia.org/wiki/Altaic_languages

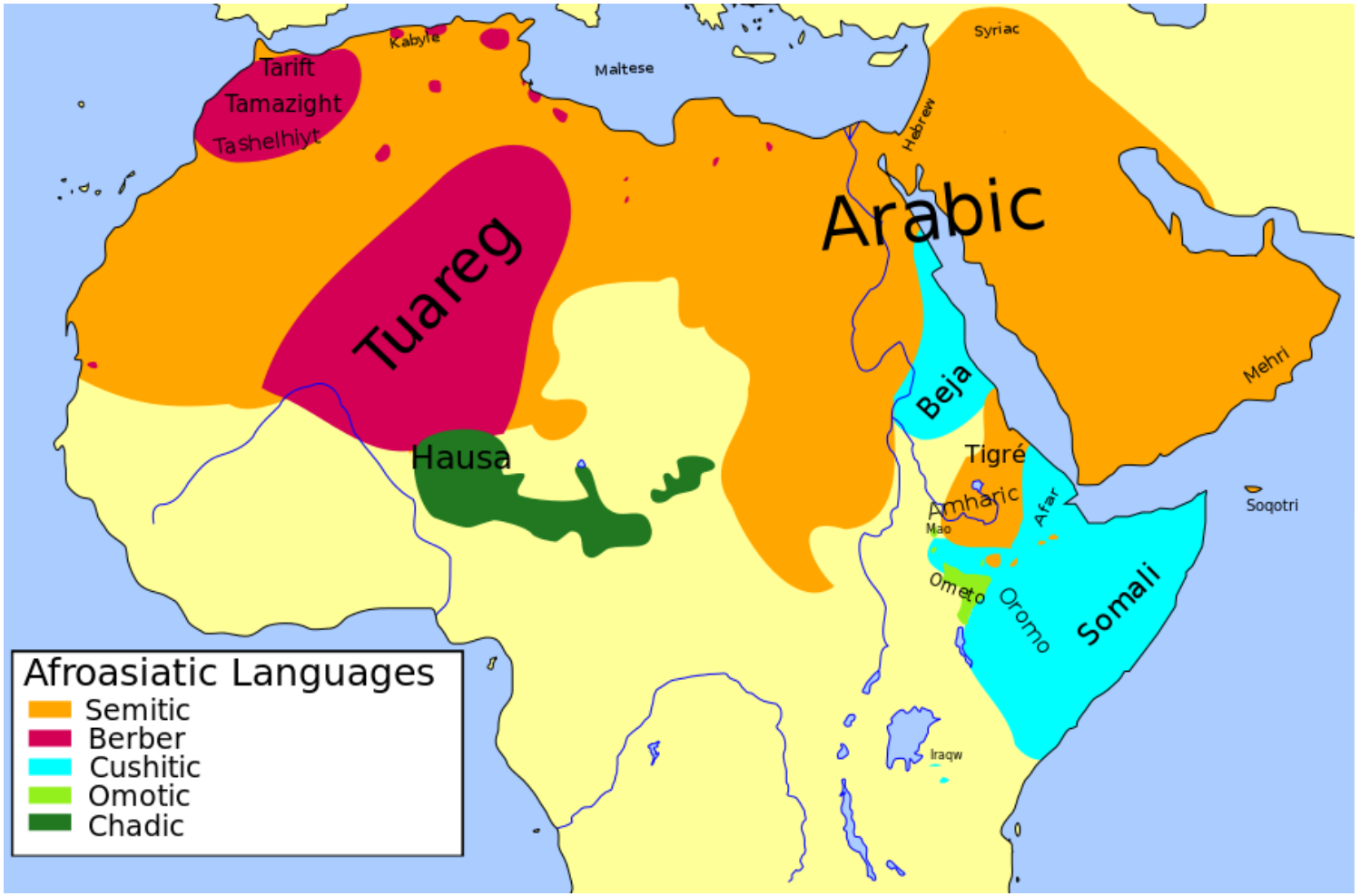
Vowel Harmony in Turkic

- 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"
kapı**dır**	"it is the dor"
gün**dür**	"it is the day"
palto**dur**	"it is the coat"

Semitic Language Family:

Hebrew, Arabic dialects, Aramaic, Amharic, etc.



From: http://en.wikipedia.org/wiki/Semitic_languages

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:

From: http://en.wikipedia.org/wiki/Chinese_language

Characteristics:

- Tone
- Isolating morphology
- Noun Classifiers

  numeral-classifier-noun in Mandarin

  noun-numeral-classifier in Burmese

Tai-Kadai Family

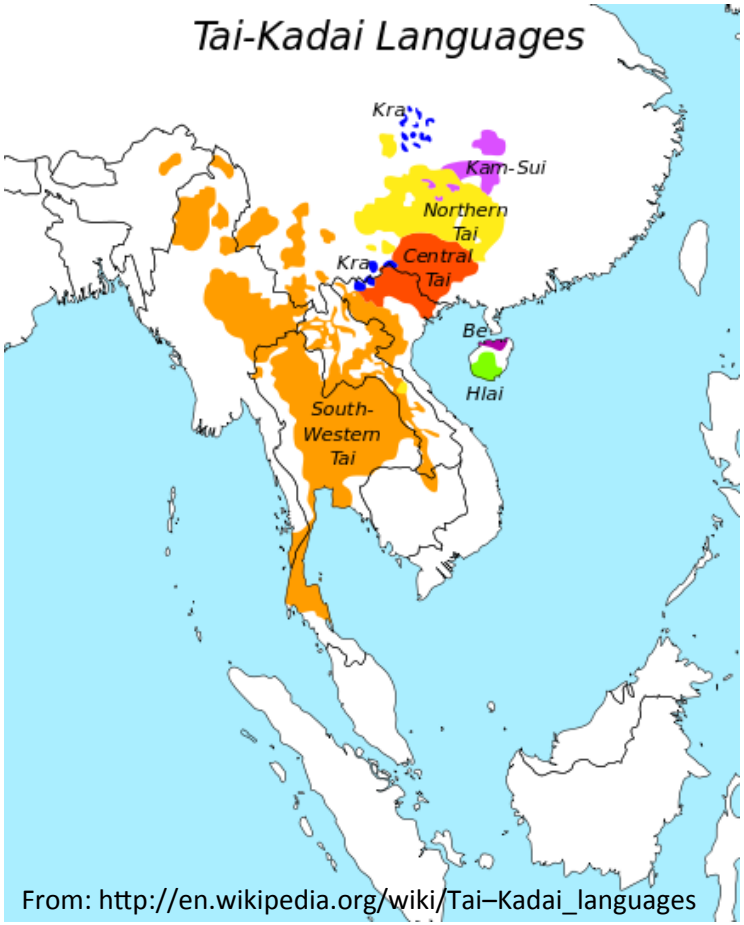
e.g. Thai – tone (5), isolating, noun classifier

Austro-Asiatic Family

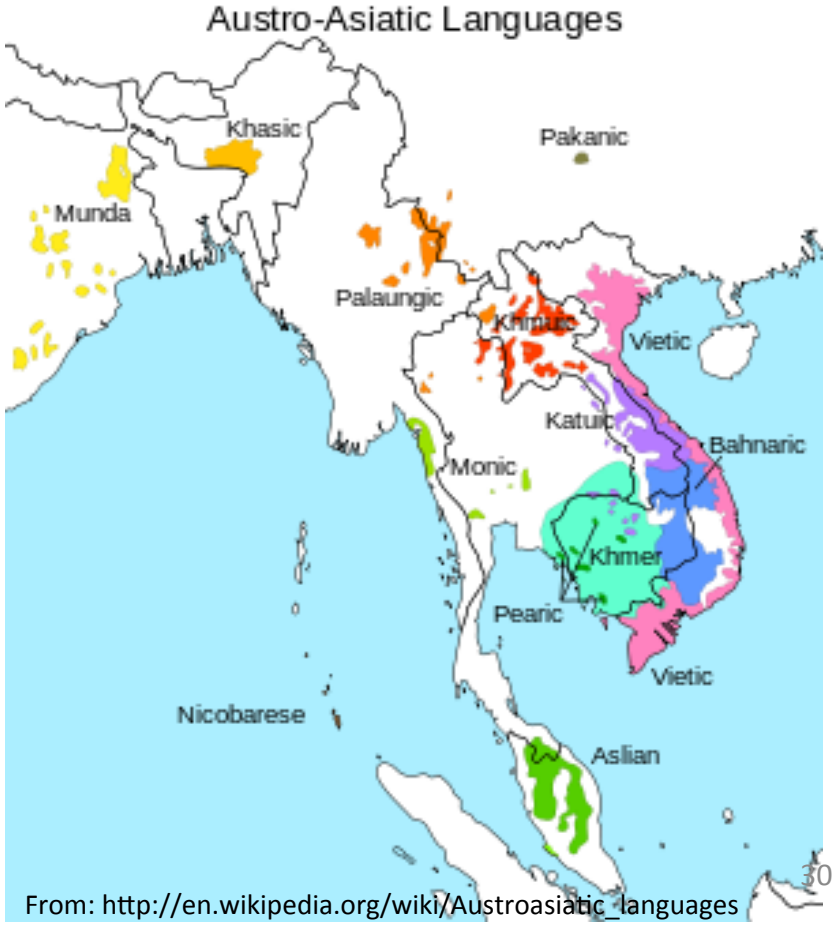
e.g. Vietnamese – tone (6), isolating, noun classifier, 30% vocab via Chinese

e.g. Munda – no tone, agglutinative

Likely areal effects



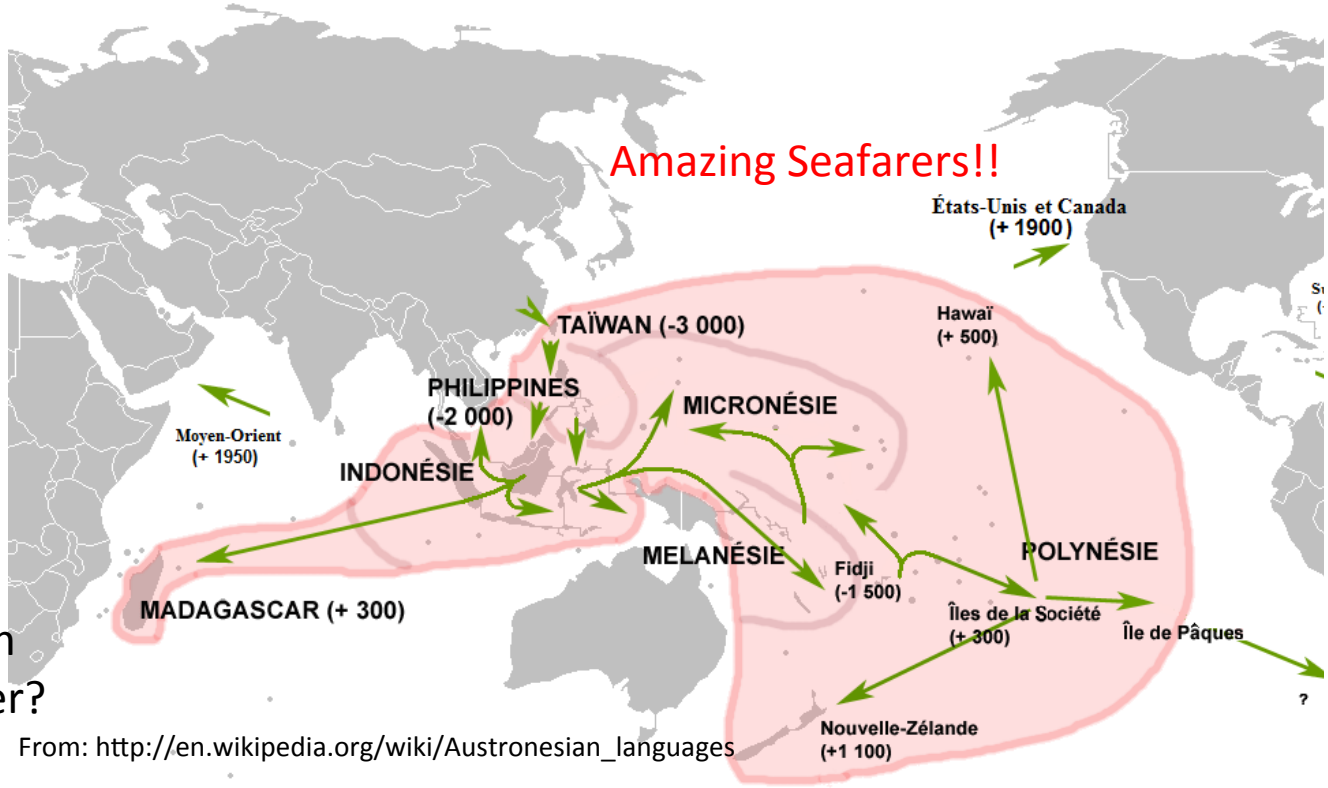
From: http://en.wikipedia.org/wiki/Tai–Kadai_languages



From: http://en.wikipedia.org/wiki/Austroasiatic_languages

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.



From: http://en.wikipedia.org/wiki/Austronesian_languages

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

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"

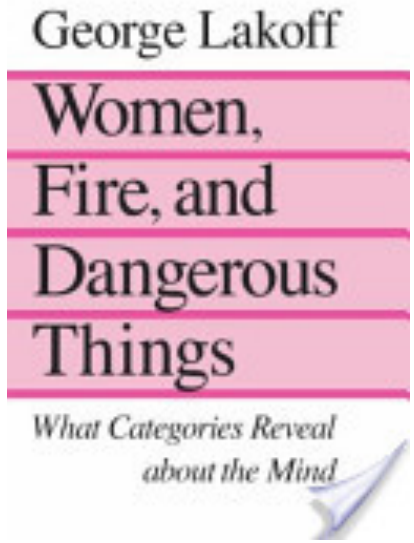
Languages in Papua New Guinea:

- 800+ languages! (1 language per 200-900km^2)
- 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:

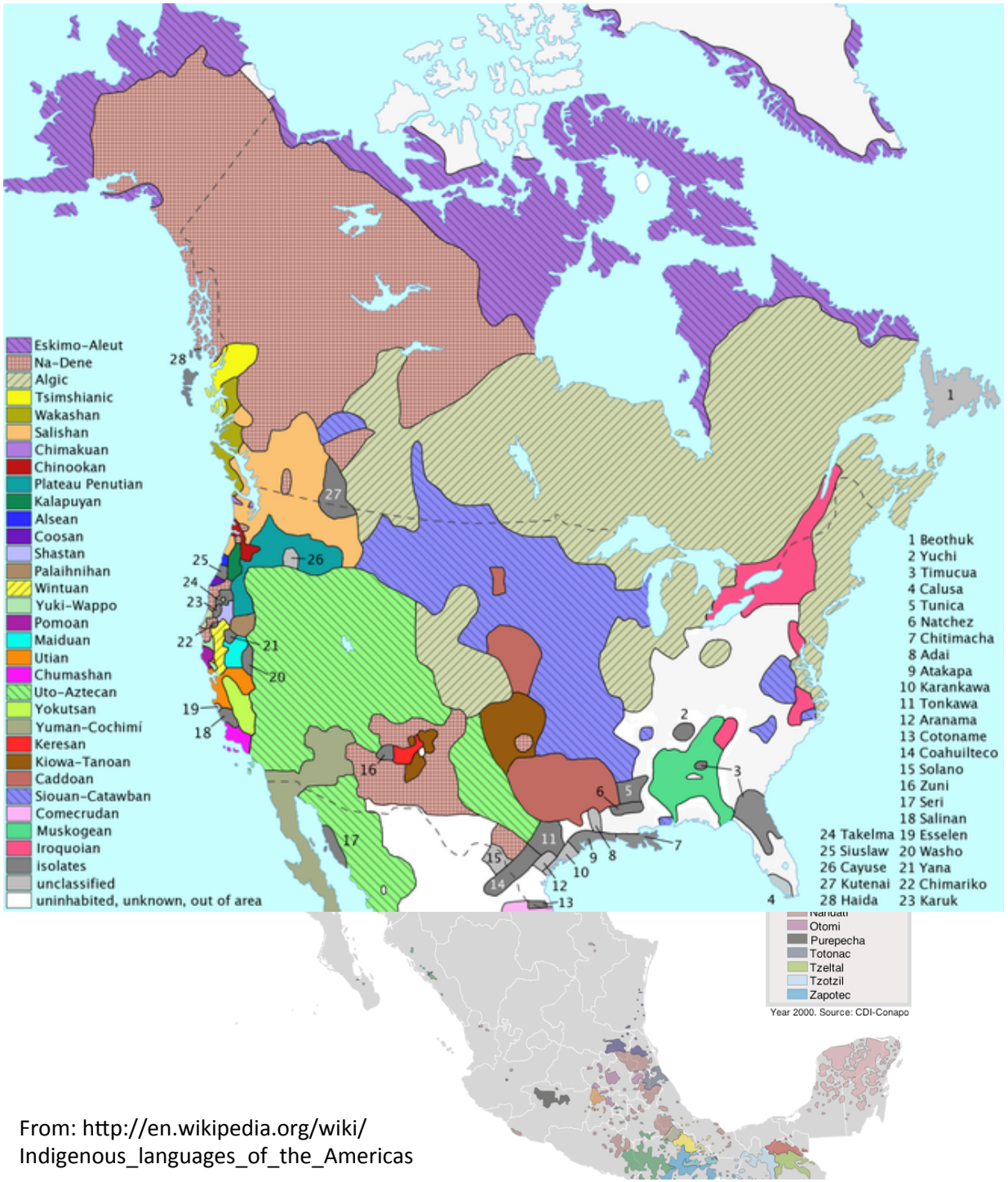
- 270 languages, many near extinction
- Trivia - Noun classes in Dyirbal:

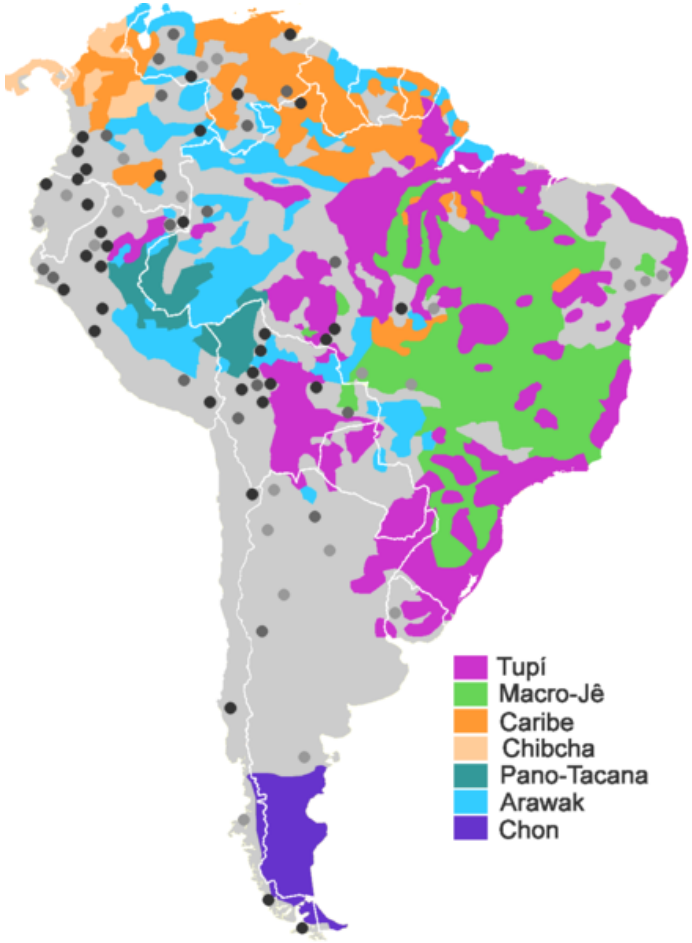
  I: masculine & animate; II: feminine, fire, fighting;
  III: all trees with edible fruit; IV: everything else



Languages of America

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





From: http://en.wikipedia.org/wiki/Indigenous_languages_of_the_Americas

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

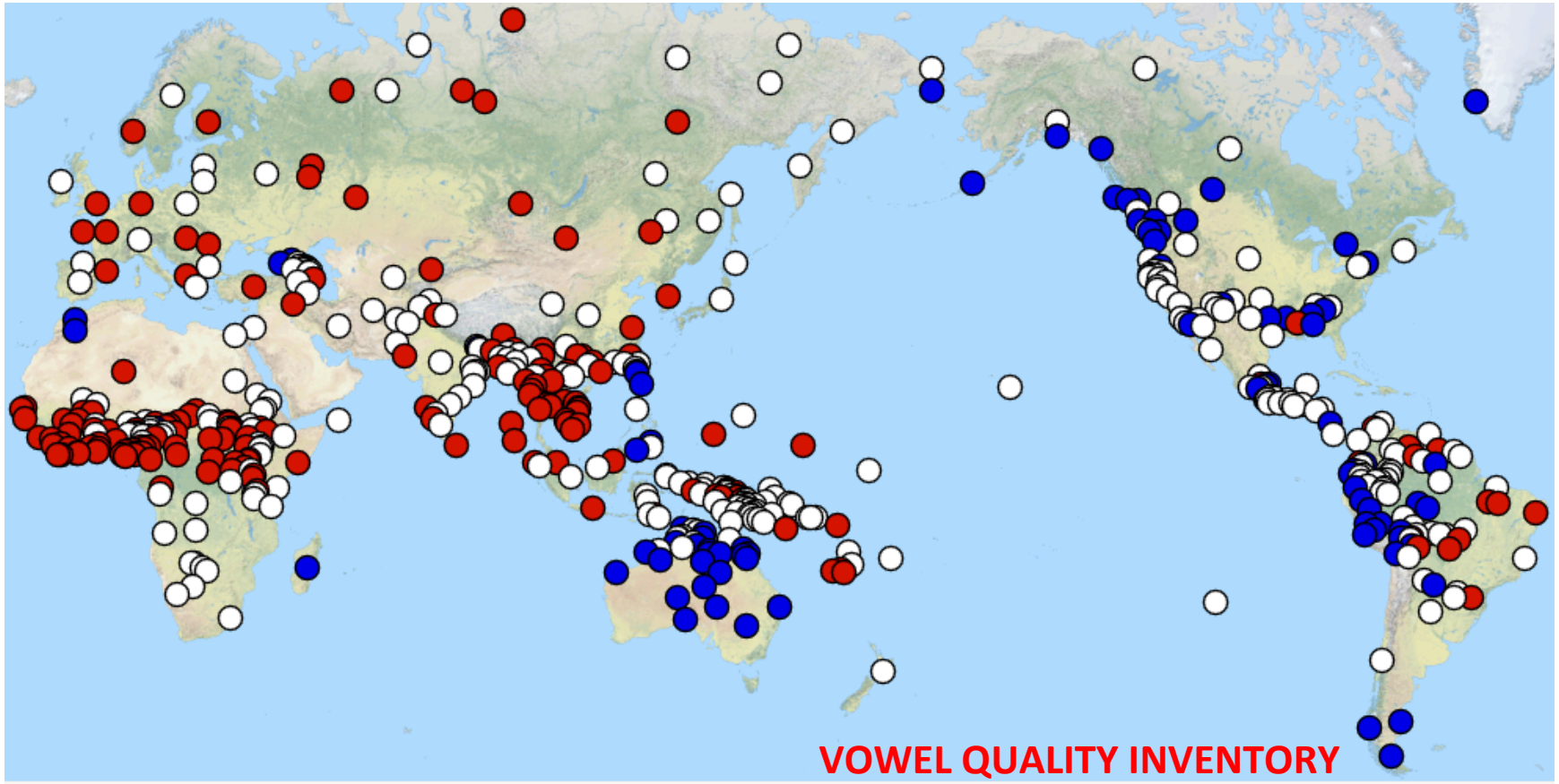
Linguistic Universals and Typology

- 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

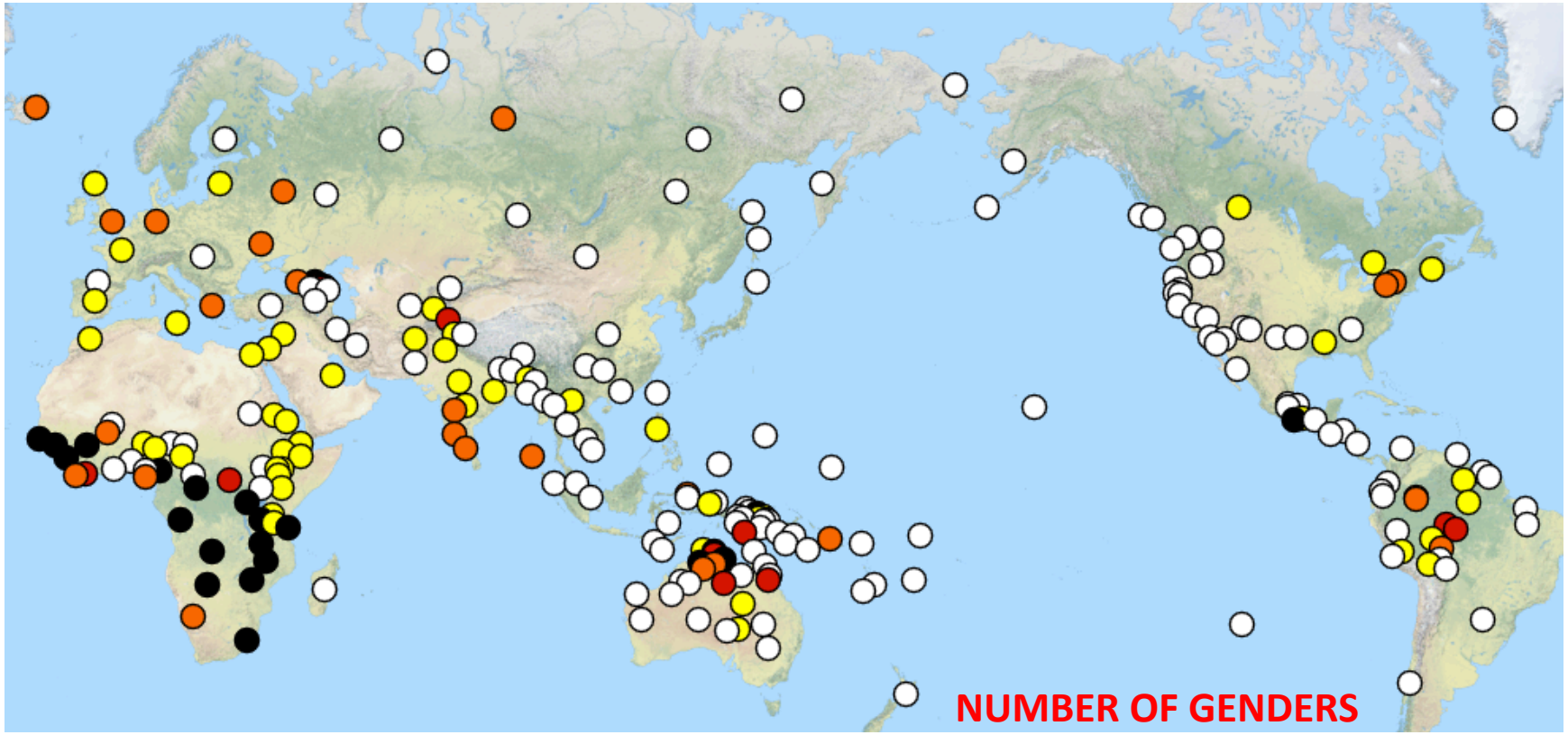
- SOV tendencies:
  - have postpositions
  - genitive-noun, etc.
- Analytical morphology tendencies:
  - mono-syllable words
  - use of tones
  - use of function words
  - relative fixed word order
- SVO tendencies:
  - have prepositions
  - noun-genitive, etc.
- Synthetic morphology tendencies:
  - 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!



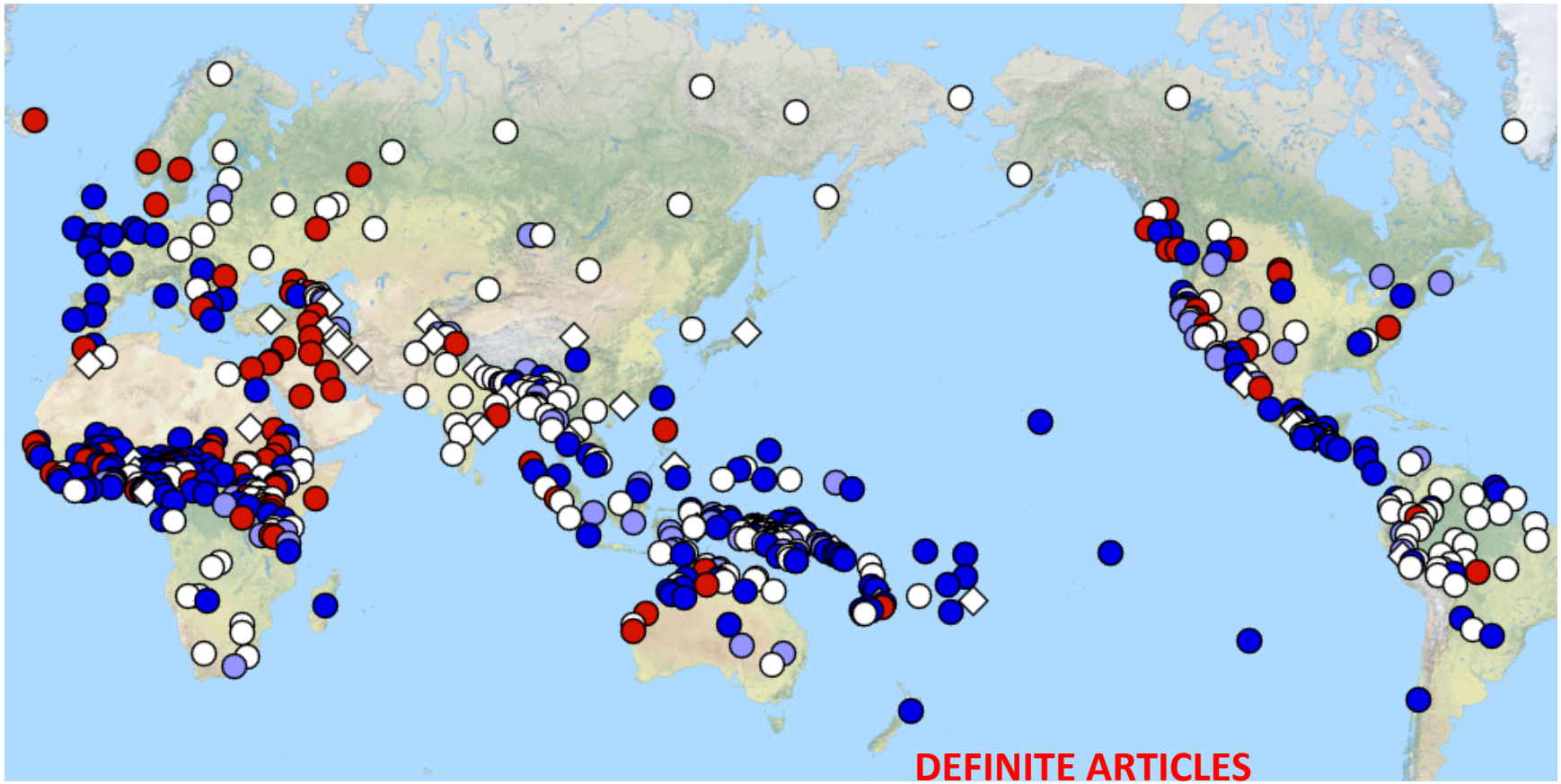
Ian Maddieson. 2013. Vowel Quality Inventories.
In: Dryer, Matthew S. & Haspelmath, Martin (eds.)
The World Atlas of Language Structures Online.
Leipzig: Max Planck Institute for Evolutionary Anthropology.
(Available online at http://wals.info/chapter/2, Accessed on 2014-06-08.)

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



Greville G. Corbett. 2013. Number of Genders.
In: Dryer, Matthew S. & Haspelmath, Martin (eds.)
The World Atlas of Language Structures Online.
Leipzig: Max Planck Institute for Evolutionary Anthropology.
(Available online at http://wals.info/chapter/30, Accessed on 2014-06-08.)

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



Matthew S. Dryer. 2013. Definite Articles.
In: Dryer, Matthew S. & Haspelmath, Martin (eds.)
The World Atlas of Language Structures Online.
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

- Bernard Comrie (Ed.) (2009) ***The World's Major Languages***, 2nd ed. New York, NY: Routledge
- Asya Pereltsvaig (2012) ***Language of the World: An Introduction***. Cambridge Univ. Press
- John McWhorter (2001) ***The Power of Babel***. HarperCollins Press
- Matthew Dryer & Martin Haspelmath (Eds.) (2013) ***The World Atlas of Language Structures Online***. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online: http://wals.in)
- Bernard Comrie, Stephen Matthews, Maria Polinsky (Eds.) (1998) ***The Atlas of Languages***. Bloomsbury Publishing