NLP

Introduction to NLP

Linguistics

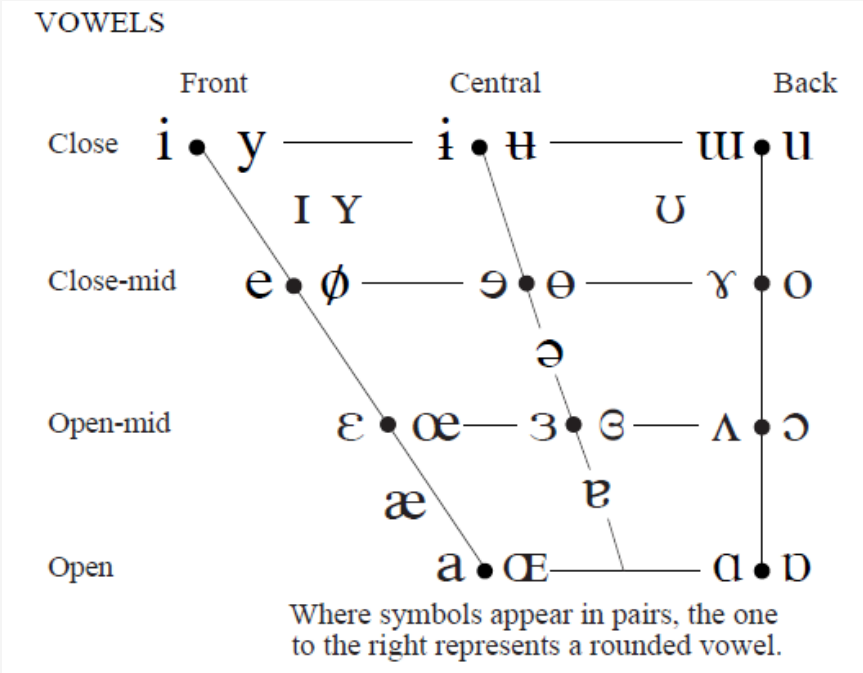
IPA Chart (consonants)

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

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

IPA Chart (vowels)

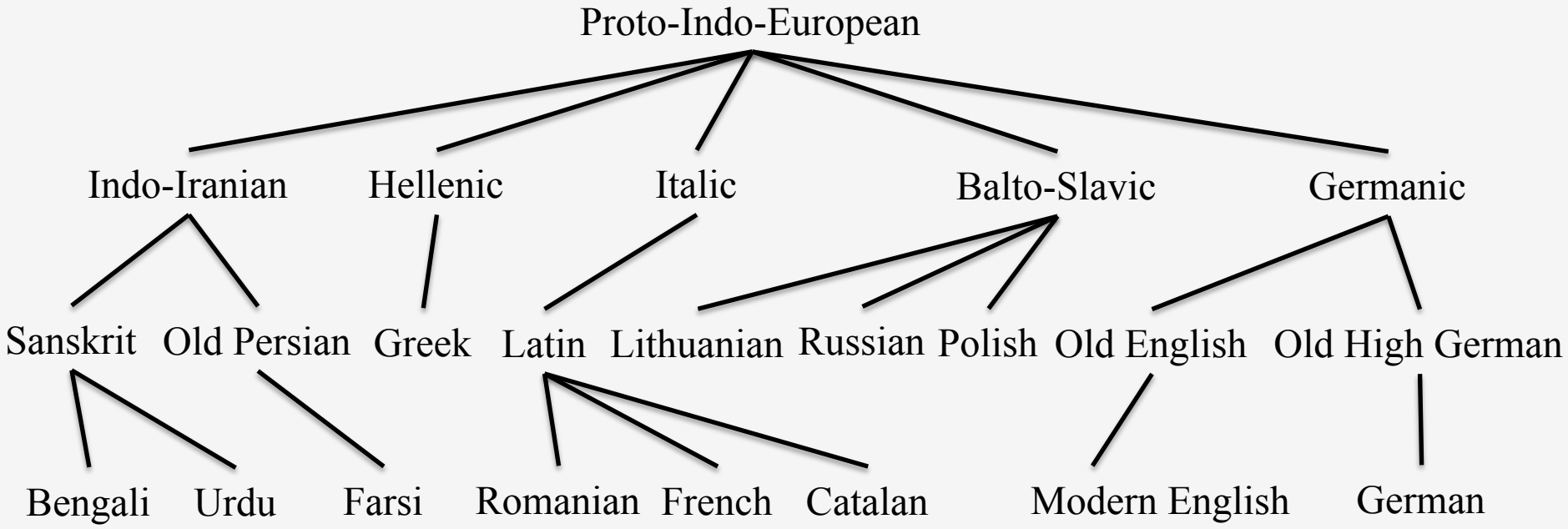


(Many) Languages are Related

- Cognates
  - night (English), nuit (French), Nacht (German), nacht (Dutch), nag (Afrikaans), nicht (Scots), natt (Swedish, Norwegian), nat (Danish), nátt (Faroese), nótt (Icelandic), noc (Czech, Slovak, Polish), ночь, noch (Russian), ноќ, noć (Macedonian), нощ, nosht (Bulgarian), ніч, nich (Ukrainian), ноч, noch/noč (Belarusian), noč (Slovene), noć (Serbo-Croatian), νύξ, nyx (Ancient Greek, νύχτα/nychta in Modern Greek), nox/nocte (Latin), nakt- (Sanskrit), natë (Albanian), noche (Spanish), nos (Welsh), nueche (Asturian), noite (Portuguese and Galician), notte (Italian), nit (Catalan), nuèch/nuèit (Occitan), noapte (Romanian), nakts (Latvian) and naktis (Lithuanian), all meaning "night" and derived from the Proto-Indo-European (PIE) *nók^{w}ts, "night".

From wikipedia

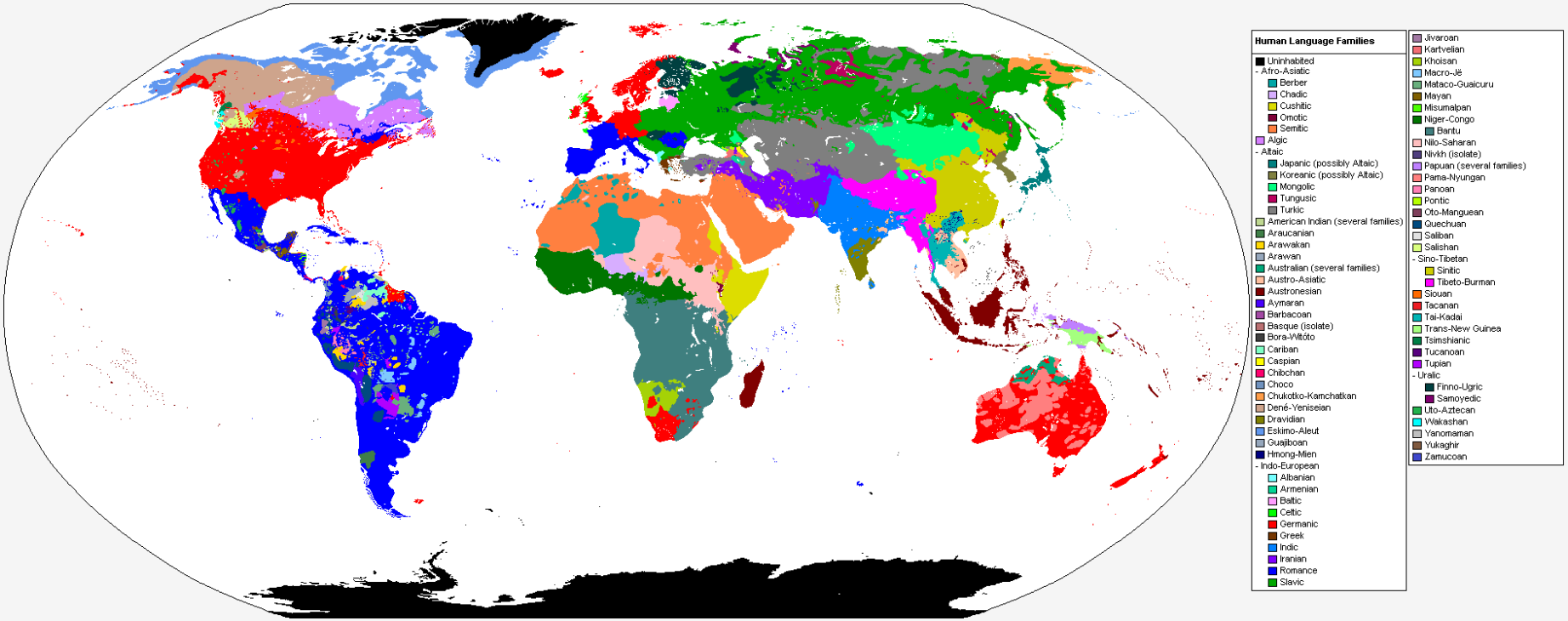
Some Indo-European languages



Some non-Indo-European Languages

- Altaic
  - Turkish
- Uralic (Finno-Ugric)
  - Finnish
  - Hungarian
- Semitic
  - Arabic
  - Hebrew
- Uto-Aztecan

Language Families



By Industrius at English Wikipedia. Later version(s) were uploaded by Mttll at English Wikipedia. (Image:BlankMap-World.png by User:Vardion) [GFDL (www.gnu.org/copyleft/fdl.html)], via Wikimedia Commons

Language Diversity

Afro-Asiatic (374)
Alacalufan (2)
Algic (44)
Altaic (66)
Amto-Musan (2)
Andamanese (13)
Arafundi (3)
Arai-Kwomtari (10)
Arauan (5)
Araucanian (2)
Arawakan (59)
Arutani-Sape (2)
Australian (264)
Austro-Asiatic (169)
Austronesian (1257)
Aymaran (3)
Barbacoan (7)
Basque (1)
Bayono-Awbono (2)
Border (15)
Caddoan (5)
Cahuapanan (2)
Carib (31)
Central Solomons (4)
Chapacura-Wanham (5)
Chibchan (21)
Chimakuan (1)
Choco (12)
Chon (2)
Chukotko-Kamchatkan (5)
Chumash (7)
Coahuiltecan (1)
Constructed language (1)
Creole (82)
Deaf sign language (130)
Dravidian (85)
East Bird's Head-Sentani (8)
East Geelvink Bay (11)
East New Britain (7)
Eastern Trans-Fly (4)
Eskimo-Aleut (11)
Guahiban (5)
Gulf (4)
Harakmbet (2)
Hibito-Cholon (2)
Hmong-Mien (38)
Hokan (23)
Huavean (4)
Indo-European (439)
Iroquoian (9)
Japonic (12)
Jivaroan (4)
Kartvelian (5)
Katukinan (3)
Kaure (4)
Keres (2)
Khoisan (27)
Kiowa Tanoan (6)
Lakes Plain (20)
Language isolate (50)
Left May (2)
Lower Mamberamo (2)
Lule-Vilela (1)
Macro-Ge (32)
Mairasi (3)
Maku (6)
Mascoian (5)
Mataco-Guaicuru (12)
Mayan (69)
Maybrat (2)
Misumalpan (4)
Mixed language (23)
Mixe-Zoque (17)
Mongol-Langam (3)
Mura (1)
Muskogean (6)
Na-Dene (46)
Nambiquaran (7)
Niger-Congo (1532)
Nilo-Saharan (205)
Nimboran (5)
North Bougainville (4)
North Brazil (1)
North Caucasian (34)
Oto-Manguean (177)
Panoan (28)
Pauwasi (5)
Peba-Yaguan (2)
Penutian (33)
Piawi (2)
Pidgin (17)
Quechuan (46)
Ramu-Lower Sepik (32)
Salishan (26)
Salivan (3)
Senagi (2)
Sepik (56)
Sino-Tibetan (449)
Siouan (17)
Sko (7)
Somahai (2)
South Bougainville (9)
South-Central Papuan (22)
Tacanan (6)
Tai-Kadai (92)
Tarascan (2)
Tequistlatecan (2)
Tor-Kwerba (24)
Torricelli (56)
Totonacan (12)
Trans-New Guinea (477)
Tucanoan (25)
Tupi (76)
Unclassified (73)
Uralic (37)
Uru-Chipaya (2)
Uto-Aztecan (61)
Wakashan (5)
West Papuan (23)
Witotoan (6)
Yanomam (4)
Yele-West New Britain (3)
Yeniseian (2)
Yuat (6)
Yukaghir (2)
Yuki (2)
Zamucoan (2)
Zaparoan (7)

Ethnologue (7358 languages)

Language Changes

- Grimm's Law
  - Voiceless stops turn into voiceless fricatives
  - Voiced stops become voiceless stops
  - Voiced aspirated stops change to voiced stops or fricatives
- Example 1
  - Ancient Greek: πούς, Latin: *pēs,* Sanskrit: *pāda*
  - English: *foot*, German: *Fuß*, Swedish: *fot*
- Example 2
  - Ancient Greek: κύων, Latin: *canis*, Welsh: *ci*
  - English: *hound*, Dutch: *hond*, German: *Hund*

NACLO Problem

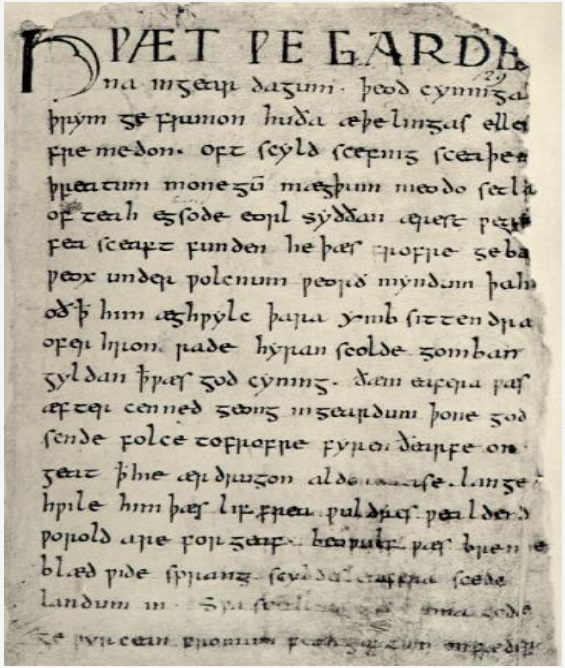
- All in the Family
  - http://www.naclo.cs.cmu.edu/problems2012/N2012-D.pdf

NACLO Solution

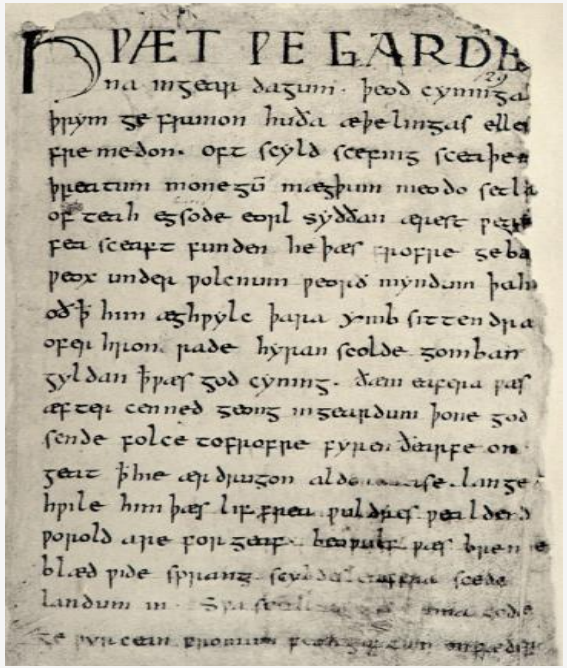
- All in the Family
  - http://www.naclo.cs.cmu.edu/problems2012/N2012-DS.pdf

Question

- Can you guess the source, language, and period of this text?



Answer



- Beowulf
- Epic poem
- 8th–11th Century
- Old English

Beowulf

Hwæt! We Gardena in geardagum,
þeodcyninga, þrym gefrunon,
hu ða æþelingas ellen fremedon.
Oft Scyld Scefing sceaþena þreatum,
monegum mægþum, meodosetla ofteah,
egsode eorlas. Syððan **ærest** wearð
feasceaft funden, he þæs frofre gebad,
weox under wolcnum, weorðmyndum þah,
oðþæt him æghwylc þara ymbsittendra

Lo! the Spear-Danes' glory through splendid achievements
The folk-kings' former fame we have heard of,
How princes displayed then their prowess-in-battle.
Oft Scyld the Scefing from scathers in numbers
From many a people their mead-benches tore.
Since **first** he found him friendless and wretched,
The earl had had terror: comfort he got for it,
Waxed 'neath the welkin, world-honor gained,
Till all his neighbors o'er sea were compelled to …

erst (as in *erstwhile*) = first

http://lit.genius.com/ http://www8.georgetown.edu/departments/medieval/labyrinth/library/oe/texts/a4.1.html
http://www.gutenberg.org/files/16328/16328-h/16328-h.htm
http://www.nvcc.edu/home/vpoulakis/Translation/beowulf1.htm
http://en.wikipedia.org/wiki/File:Beowulf.firstpage.jpeg

ATHELING.—Prince, nobleman.
BAIRN.—Son, child.
BARROW.—Mound, rounded hill, funeral-mound.
BATTLE-SARK.—Armor.
BEAKER.—Cup, drinking-vessel.
BEGEAR.—Prepare.
BIGHT.—Bay, sea.
BILL.—Sword.
BOSS.—Ornamental projection.
BRACTEATE.—A round ornament on a necklace.
BRAND.—Sword.
BURN.—Stream.
BURNIE.—Armor.
CARLE.—Man, hero.
EARL.—Nobleman, any brave man.
EKE.—Also.
EMPRISE.—Enterprise, undertaking.
ERST.—Formerly.
ERST-WORTHY.—Worthy for a long time past.
FAIN.—Glad.
FERRY.—Bear, carry.
FEY.—Fated, doomed.
FLOAT.—Vessel, ship.
FOIN.—To lunge (Shaks.).
GLORY OF KINGS.—God.
GREWSOME.—Cruel, fierce.
HEFT.—Handle, hilt; used by synecdoche for 'sword.'
HELM.—Helmet, protector.
HENCHMAN.—Retainer, vassal.
HIGHT.—Am (was) named.
HOLM.—Ocean, curved surface of the sea.
HIMSEEMED.—(It) seemed to him.
LIEF.—Dear, valued.
MERE.—Sea; in compounds, 'mere-ways,' 'mere-currents,' etc.
MICKLE.—Much.
NATHLESS.—Nevertheless.
NAZE.—Edge (nose).
NESS.—Edge.
NICKER.—Sea-beast.
QUIT, QUITE.—Requite.
RATHE.—Quickly.
REAVE.—Bereave, deprive.
SAIL-ROAD.—Sea.
SETTLE.—Seat, bench.
SKINKER.—One who pours.
SOOTHLY.—Truly.
SWINGE.—Stroke, blow.
TARGE, TARGET.—Shield.
THROUGHLY.—Thoroughly.
TOLD.—Counted.
UNCANNY.—Ill-featured, grizzly.
UNNETHE.—Difficult.
WAR-SPEED.—Success in war.
WEB.—Tapestry (that which is 'woven').
WEEDED.—Clad (cf. widow's weeds).
WEEN.—Suppose, imagine.
WEIRD.—Fate, Providence.
WHILOM.—At times, formerly, often.
WIELDER.—Ruler. Often used of God;
WIGHT.—Creature.
WOLD.—Plane, extended surface.
WOT.—Knows.
YOUNKER.—Youth.

Diversity of Languages

- Articles
- Cases (e.g., in Latin)
  - Puer puellam vexat
- Sound systems
  - Glottal stop (the middle sound in “uh-oh”) – pro
  - Velar fricatives – articulated with the back of the tongue at the soft palate
    - Voiceless /x/ – used e.g., in Russian
    - Voiced /ɣ/ – used e.g., in Modern Greek
- Social status (e.g., in Japanese)
  - otousan, お父さん = someone else‘s father
  - chichi, 父 = one’s own father
- Kinship systems (e.g., in Warlpiri) – see next slide

NACLO Problem

- Warlpiri Kinship – by Alan Chang
  - http://www.naclo.cs.cmu.edu/pdf-split/N2013-O.pdf

NACLO Solution

- Warlpiri Kinship
  - http://www.naclo.cs.cmu.edu/pdf-split/N2013-OS.pdf

Language Universals

- Two types
  - unconditional
  - conditional
- Examples
  - All languages have verbs and nouns
  - All spoken languages have consonants and vowels
  - [Greenberg 1] “In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.“
  - [Greenberg 29] “If a language has inflection, it always has derivation.“

WALS: the World Atlas of Language Structures

- http://wals.info
- Feature 83A: Order of Object and Verb
  - by Matthew S. Dryer
  - OV (713 languages), VO (705), no dominant order (101)
  - http://wals.info/feature/83A#2/18.0/152.9
- Other features:
  - 18A Absence of common consonants (by Ian Maddieson): no bilabials (5 languages), no fricatives (49), no nasals (12)
  - 67A Inflectional future tense (by Östen Dahl, Viveka Velupillai): yes (110), no (112)

Links about World Languages

- Ethnologue
  - http://www.ethnologue.com/
- Number words in many languages
  - http://www.zompist.com/numbers.shtml
- Endangered languages
  - http://www.endangeredlanguages.com/
- Google fights to save 3,054 dying languages
  - http://www.cnn.com/2012/06/21/tech/web/google-fights-save-language-mashable/index.html

NLP