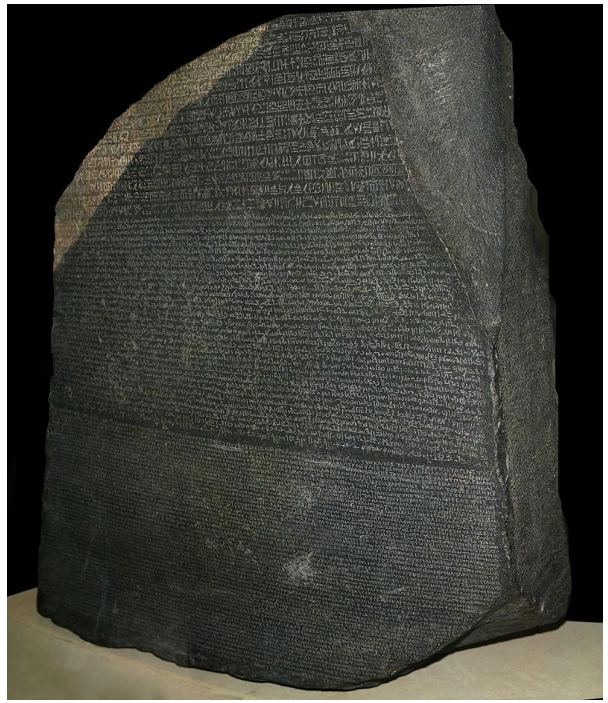
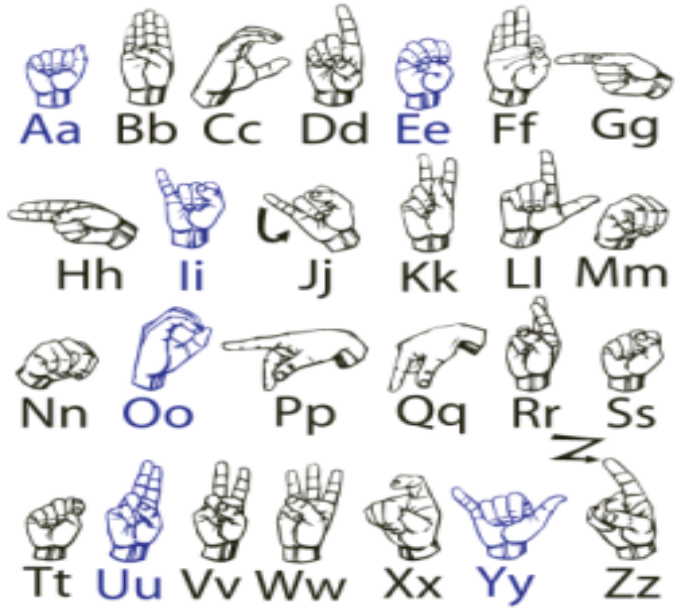
Introduction to NLP

141

A Brief Overview of Languages and Linguistics

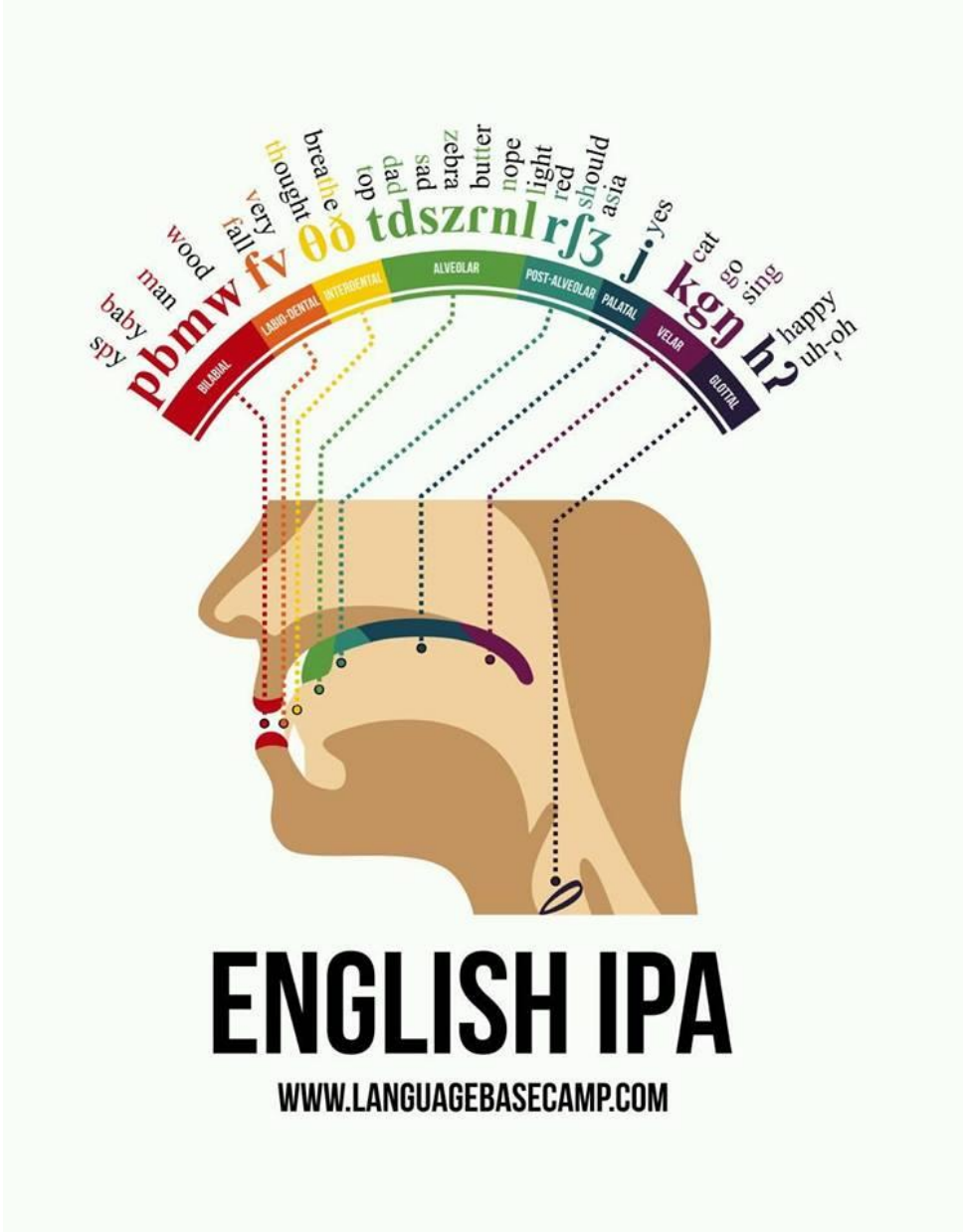


خ	ح	ج	ث	ت	ب	ا
kha	haa	jiim	thaa	taa	baa	alif
ص	ش	س	ز	ر	ذ	د
saad	shiin	siin	zaay	raa	thaal	daal
ق	ف	غ	ع	ظ	ط	ض
qaaf	faa	ghayn	ayn	thaa	taa	daad
ي	و	ه	ن	م	ل	ك
yaa	waaw	ha	nuun	miim	laam	kaaf



Аа	a (as in cat)	Кк	k (as in kick)	Фф	f (as in foot)
Бб	b (as in bus)	Лл	l (as in love)	Хх	h (like "ch" in Bach)
Вв	v (as in very)	Мм	m (as in marry)	Чч	ts (as in puts)
Гг	g (as in good)	Нн	n (as in no)	Цц	ch (as in check)
Дд	d (as in dog)	Оо	o (as in hot)	Шш	sh (as in shut)
Ее	e (as in egg)	Пп	p (as in pot)	Щщ	sht (like "shed" in pushed)
Жж	zh (like's" in leisure)	Рр	r (as in red)	Ъъ	a (like "u" in but)
Зз	z (as in zoo)	Сс	s (as in sit)	Ьь	(consonant softening sound)
Ии	i (as in instant)	Тт	t (as in tree)	Юю	yu (like you)
Йй	y (as in young)	Уу	u (as in yule)	Яя	ya (as in yank)

Consonants in English



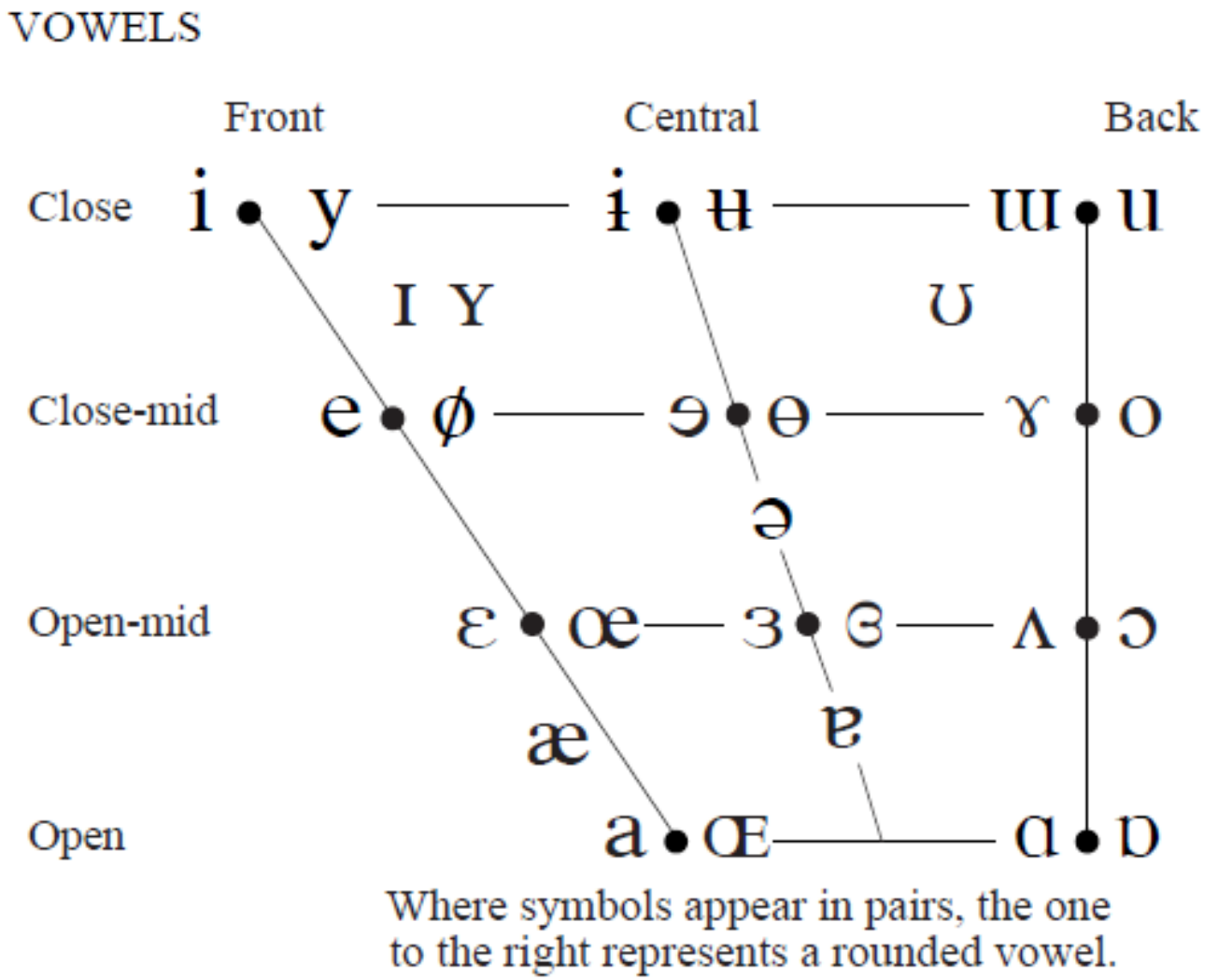
IPA Chart (consonants) 辅音

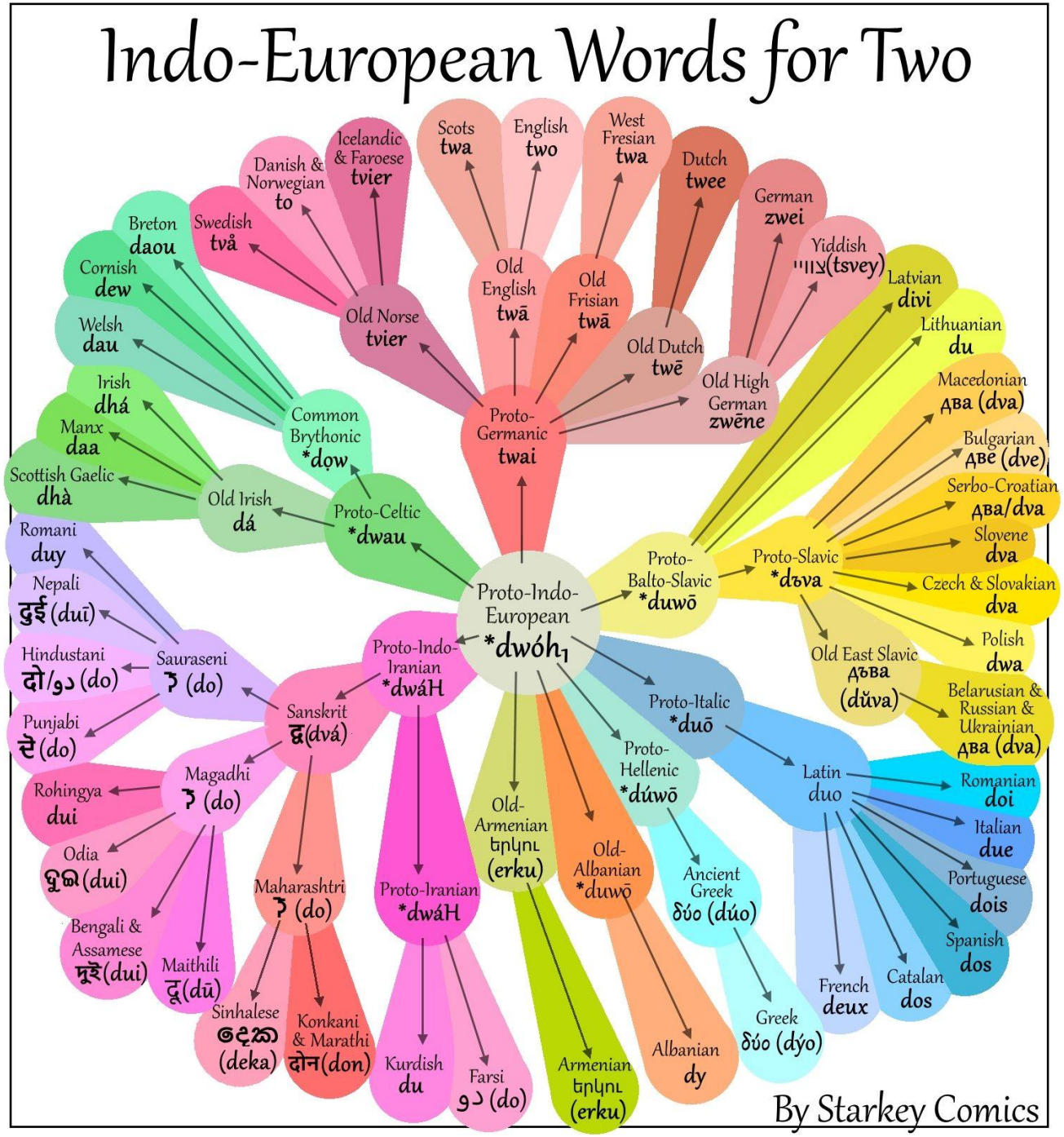
CONSONANTS (PULMONIC) © 2005 IPA

	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	N		
Trill	B			r					R		
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			

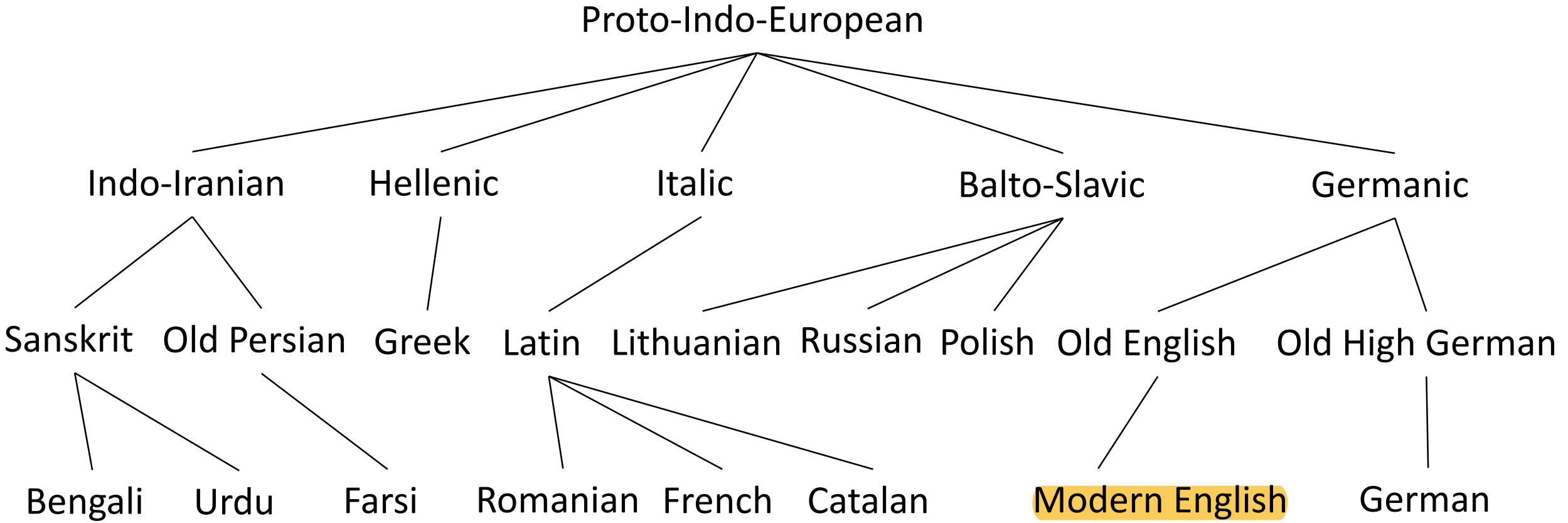
Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

IPA Chart (vowels)




Indo-European Words for Two
Proto-Indo-European *dwóh₁
Proto-Germanic twai
Old Norse tvier
Icelandic & Faroese tvier
Danish & Norwegian to
Swedish två
Scots twa
English two
Old English twā
West Fresian twa
Old Frisian twā
Dutch twee
Old Dutch twē
German zwei
Old High German zwēne
Yiddish צוויי (tsvey)
Common Brythonic *dọw
Breton daou
Cornish dew
Welsh dau
Proto-Celtic *dwau
Old Irish dá
Irish dhá
Manx daa
Scottish Gaelic dhà
Proto-Balto-Slavic *duwō
Latvian divi
Lithuanian du
Proto-Slavic *dъva
Macedonian два (dva)
Bulgarian две (dve)
Serbo-Croatian два/dva
Slovene dva
Czech & Slovakian dva
Polish dwa
Old East Slavic дъва (dŭva)
Belarusian & Russian & Ukrainian два (dva)
Proto-Italic *duō
Latin duo
Romanian doi
Italian due
Portuguese dois
Spanish dos
French deux
Catalan dos
Proto-Hellenic *dúwō
Ancient Greek δύο (dúo)
Greek δύο (dýo)
Old-Albanian *duwō
Albanian dy
Old-Armenian երկու (erku)
Armenian երկու (erku)
Proto-Indo-Iranian *dwáH
Proto-Iranian *dwáH
Kurdish du
Farsi دو (do)
Sanskrit द्व (dvá)
Maharashtri दो (do)
Sinhalese දෙක (deka)
Konkani & Marathi दोन (don)
Magadhi दो (do)
Maithili दू (dū)
Bengali & Assamese দুই (dui)
Odia ଦୁଇ (dui)
Rohingya dui
Sauraseni दो (do)
Hindustani दो/دو (do)
Punjabi ਦੋ (do)
Nepali दुई (duī)
Romani duy
By Starkey Comics

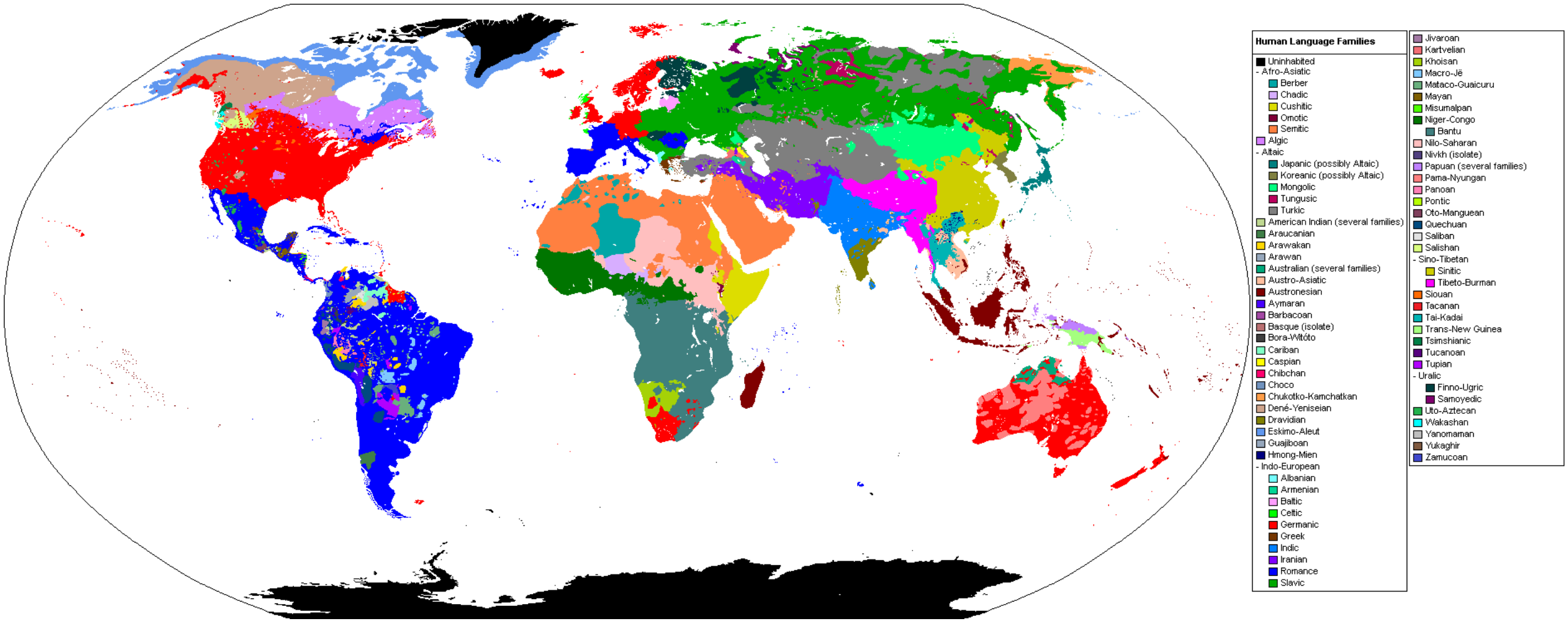
Some Indo-European languages



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)

NACLO Problem

- http://nacloweb.org/resources/problems/2012/N2012-D.pdf
- http://nacloweb.org/resources/problems/2012/N2012-DS.pdf
- Problem by Dragomir Radev

Many languages are related to each other for historical reasons. They may have a common ancestor or they may have borrowed words from each other. Linguists group languages into families and branches, based on their common ancestry.
Here is a list of translations of the first article of the Universal Declaration of Human Rights in 17 languages:

Your task it to identify similarities among these languages and group them into seven clusters (groups) of related languages as sketched in the diagram below:

http://unicode.org/udhr/assemblies/first_article_all.html

Here is a list of translations of the first article of the Universal Declaration of Human Rights in 17 languages:

A. (English) All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

B. (Latin) Omnes homines dignitate et iure liberi et pares nascuntur, rationis et conscientiae participes sunt, quibus inter se concordiae studio est agendum.

C. Vsi ljudje se rodijo svobodni in imajo enako dostojanstvo in enake pravice. Obdarjeni so z razumom in vestjo in bi morali ravnati drug z drugim kakor bratje.

D. Dieub ha par en o dellezegezh hag o gwirioù eo ganet an holl dud. Poell ha skiant zo dezho ha dleout a reont bevañ an eil gant egile en ur spered a genvreudeuriezh.

E. Tuots umans naschan libers ed eguals in dignità e drets. Els sun dotats cun intellet e conscienza e dessan agir tanter per in uin spiert da fraternità.

F. Toate ființele umane se nasc libere și egale în demnitate și în drepturi. Ele sunt înzestrate cu rațiune și conștiință și trebuie să se comporte unii față de altele în spiritul fraternității.

G. Genir pawb yn rhydd ac yn gydradd â'i gilydd mewn urddas a hawliau. Fe'u cynysgaeddir â rheswm a chydwybod, a dylai pawb ymddwyn y naill at y llall mewn ysbryd cymodlon.

H. Visi žmonės gimsta laisvi ir lygūs savo orumu ir teisėmis. Jiems suteiktas protas ir sąžinė ir jie turi elgtis vienas kito atžvilgiu kaip broliai.

I. Totu sos èsseres umanos naschint lìberos e eguales in dinnidade e in deretos. Issos tenent sa resone e sa cussèntzia e depent operare s'unu cun s'àteru cun ispìritu de fraternidade.

J. Gizon-emakume guztiak aske jaiotzen dira, duintasun eta eskubide berberak dituztela; eta ezaguera eta kontzientzia dutenez gero, elkarren artean senide legez jokatu beharra dute.

K. Kai rahvas roittahes vällinny da taza-arvozinnu omas arvos da oigevuksis. Jogahizele heis on annettu mieli da omatundo da heil vältämättäh pidäy olla keskenäh, kui vellil.

L. Všetci ľudia sa rodia slobodní a sebe rovní , čo sa týka ich dostôjnosti a práv. Sú obdarení rozumom a majú navzájom jednať v bratskom duchu.

M. Nascinu tutti l'omi libari è pari di dignità è di diritti. Pussedinu a raghjoni è a cuscenza è li tocca ad agiscia trà elli di modu fraternu.

N. Saoláitear na daoine uile saor agus comhionann ina ndínit agus ina gcearta. Tá bauidh an réasúin agus an choinsiasa acu agus dlíd iad féin d'iompar de mheon bhrthreachais i leith a chéile.

O. Visi cilvēki piedzimst brīvi un vienlīdzīgi savā pašcieņā un tiesībās. Viņi ir apveltīti ar saprātu un sirdsapziņu, un viņiem jāizturas citam pret citu brālības garā.

P. Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki ja omatunto, ja heidän on toimittava toisiaan kohtaan veljeyden hengessä.

Q. Wszyscy ludzie rodzą się wolni i równi pod względem swej godności i swych praw. Są oni obdarzeni rozumem i sumieniem i powinni postępować wobec innych w duchu braterstwa.

Here is a list of translations of the first article of the Universal Declaration of Human Rights in 17 languages:

English

A. (English) All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

Latin

B. (Latin) Omnes homines dignitate et iure liberi et pares nascuntur, rationis et conscientiae participes sunt, quibus inter se concordiae studio est agendum.

Slovenian

C. Vsi ljudje se rodijo svobodni in imajo enako dostojanstvo in enake pravice. Obdarjeni so z razumom in vestjo in bi morali ravnati drug z drugim kakor bratje.

Breton

D. Dieub ha par en o dellezegezh hag o gwirioù eo ganet an holl dud. Poell ha skiant zo dezho ha dleout a reont bevañ an eil gant egile en ur spered a genvreudeuriezh.

Romansch

E. Tuots umans naschan libers ed eguals in dignità e drets. Els sun dotats cun intellet e conscienza e dessan agir tanter per in uin spiert da fraternità.

Romanian

F. Toate ființele umane se nasc libere și egale în demnitate și în drepturi. Ele sunt înzestrate cu rațiune și conștiință și trebuie să se comporte unii față de altele în spiritul fraternității.

Welsh

G. Genir pawb yn rhydd ac yn gydradd â'i gilydd mewn urddas a hawliau. Fe'u cynysgaeddir â rheswm a chydwybod, a dylai pawb ymddwyn y naill at y llall mewn ysbryd cymodlon.

Lithuanian

H. Visi žmonės gimsta laisvi ir lygūs savo orumu ir teisėmis. Jiems suteiktas protas ir sąžinė ir jie turi elgtis vienas kito atžvilgiu kaip broliai.

Sardinian

I. Totu sos èsseres umanos naschint lìberos e eguales in dinnidade e in deretos. Issos tenent sa resone e sa cussèntzia e depent operare s'unu cun s'àteru cun ispìritu de fraternidade.

Basque

J. Gizon-emakume guztiak aske jaiotzen dira, duintasun eta eskubide berberak dituztela; eta ezaguera eta kontzientzia dutenez gero, elkarren artean senide legez jokatu beharra dute.

Karelian

K. Kai rahvas roittahes vällinny da taza-arvozinnu omas arvos da oigevuksis. Jogahizele heis on annettu mieli da omatundo da heil vältämättäh pidäy olla keskenäh, kui vellil.

Slovak	L. Všetci ľudia sa rodia slobodní a sebe rovní , čo sa týka ich dostôjnosti a práv. Sú obdarení rozumom a majú navzájom jednať v bratskom duchu.
Corsican	M. Nascinu tutti l'omi libari è pari di dignità è di diritti. Pussedinu a raghjoni è a cuscenza è li tocca ad agis-cia trà elli di modu fraternu.
Irish	N. Saoláitear na daoine uile saor agus comhionann ina ndínit agus ina gcearta. Tá bauidh an réasúin agus an choinsiasa acu agus dlíd iad féin d'iompar de mheon bhrthreachais i leith a chéile.
Latvian	O. Visi cilvēki piedzimst brīvi un vienlīdzīgi savā pašcieņā un tiesībās. Viņi ir apveltīti ar saprātu un sirdsapziņu, un viņiem jāizturas citam pret citu brālības garā.
Finnish	P. Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki ja omatunto, ja heidän on toimittava toisiaan kohtaan veljeyden hengessä.
Polish	Q. Wszyscy ludzie rodzą się wolni i równi pod względem swej godności i swych praw. Są oni obdarzeni rozumem i sumieniem i powinni postępować wobec innych w duchu braterstwa.

Language Families

1.	CLQ	Slavic
2.	BEFIM	Romance
3.	J	Basque
4.	HO	Baltic
5.	DGN	Celtic
6.	KP	Finno-Ugric
7.	A	English

How English has changed over the last 1000 years: the 23rd Psalm

Modern (1989)

The Lord is my shepherd, I lack nothing.
He lets me lie down in green pastures.
He leads me to still waters.

King James Bible (1611)

The Lord is my shepherd, I shall not want.
He maketh me to lie down in green pastures.
He leadeth me beside the still waters.

Middle English (1100–1500)

Our Lord gouerneth me, and nothyng shal defailen to me.
In the sted of pastur he sett me ther.
He norissed me upon water of fyllyng.

Old English (800–1066)

Drihten me raet, ne byth me nanes godes wan.
And he me geset on swythe good feohland.
And fedde me be waetera stathum.


- How can I say "Togetherness"
on german language?
- "Zusammengehörigkeitsgefühl"

Language Diversity

- Articles
  - English vs. Russian
- 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

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

Introduction to NLP

143

Morphology and the Lexicon

Mental Lexicon

- What is the meaning of cat?
  - Its pronunciation?
  - Part of speech?
- What is the meaning of wug?
- What is the meaning of cluvious?

“Runs”

- Two interpretations
- Affixes
  - prefixes, infixes, suffixes, circumfixes, null morpheme

Morphological Examples

- Reduplication
  - amigo = friend, amimigo = friends (in Pangasinan) [Rubino 2001]
  - savali = he travels, savavali = they travel (in Samoan)
- Templatic morphology (e.g., Semitic languages):
  - lmd (learn), lamad (he studied), limed (he taught), lumad (he was taught)
- Circumfixes
  - spielen – gespielt (in German), light – enlighten (in English)
- Pig Latin
  - appyhay
- Massa-*freakin’*-chusetts
  - where can you insert “freakin” in “education”?

Reduplication: A process of forming new words by repeating a morpheme or a part of a word. In the provided examples, "amimígo" in Pangasinan and "savavali" in Samoan are formed by repeating parts of the root words "amigo" and "savali," respectively.

Templatic morphology: A process in which words are formed using a template or pattern, common in Semitic languages like Hebrew and Arabic. The example shows different forms of the root "lmd" (learn) in Hebrew, with different prefixes and suffixes creating various meanings.

Circumfixes: A process where an affix has two parts, one appearing before and the other after the root word. In German, "spielen" becomes "gespielt" with the circumfix "ge-...-t," and in English, "light" becomes "enlighten" with the circumfix "en-...-en."

Pig Latin: A playful language game where the first consonant or consonant cluster of an English word is moved to the end and followed by "-ay." For example, "happy" becomes "appyhay."

Inserting "freakin'": This is an example of playful infixation for emphasis or stylistic effect. In the word "education," you can insert "freakin'" to make it "edu-freakin'-cation." Note that this is informal and not part of standard English.

Answer

- The “freakin’” infix is inserted
- … to the left of the syllable that bears the main stress

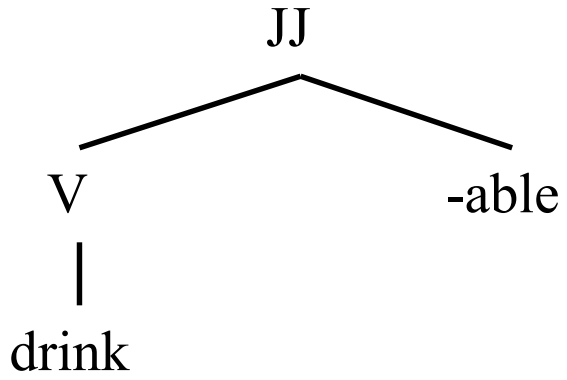
  edu-*freakin’*-cation

  * educa-*freakin’*-tion

  * e-*freakin’*-ducation
- though there can be exceptions

Derivational Morphology

- Example
  - “er” (multiple interpretations)
- What do these morphemes mean?
  - prefix, stem, suffix, ending
  - ness, able, ing, re, un, er (adj)
  - JJ → V + “-able”
- Recursion:
  - unconcernednesses
- Ambiguity
  - uncloggable vs. unbelievable



Answer to the Quiz

- Uncloggable
  - unable to be clogged
  - able to be unclogged
- Unbelievable
  - unable to be believed
  - ? able to be unbelieved

Inflectional Morphology

- Many forms
  - Tense, number, person, mood, aspect
  - Five verb forms in English
  - 40+ forms in French
  - Six cases in Russian:

    http://www.departments.bucknell.edu/russian/language/case.html
  - Up to 40,000 forms in Turkish
    - E.g., you cause X to cause Y to … do Z)

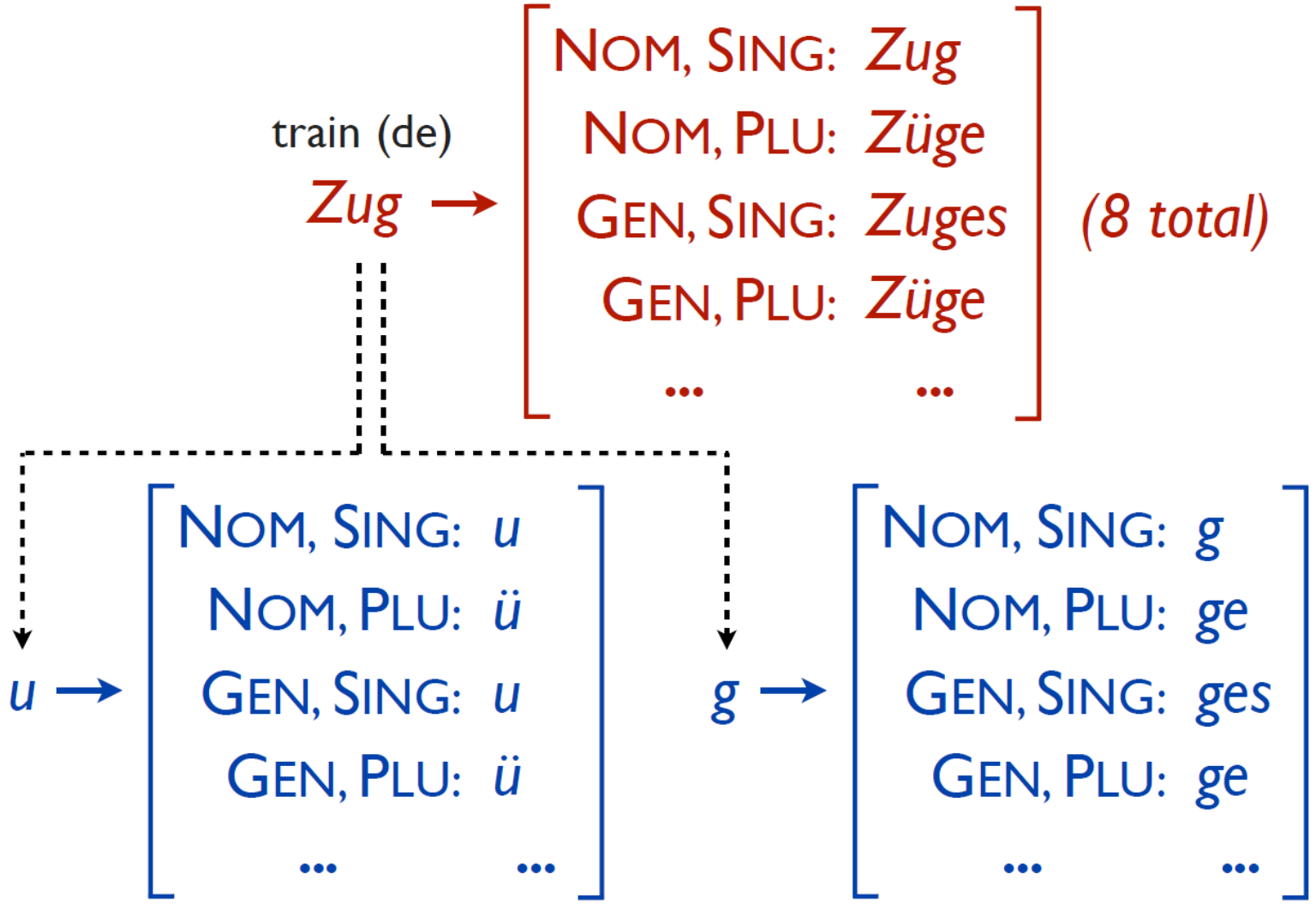
Noun and Pronoun Declension in Latin

	puer, puerī boy m.		*ager, agrī* field m.		*vir, virī* man m.	
	Singular	**Plural**	**Singular**	**Plural**	**Singular**	**Plural**
Nominative	puer	puerī	ager	agrī	vir	virī
Vocative						
Accusative	puerum	puerōs	agrum	agrōs	virum	virōs
Genitive	puerī	puerōrum	agrī	agrōrum	virī	virōrum (virum)
Dative	puerō	puerīs	agrō	agrīs	virō	virīs
Ablative						

	noster, nostra, nostrum our, ours					
	Singular			**Plural**		
	Masculine	**Feminine**	**Neuter**	**Masculine**	**Feminine**	**Neuter**
Nominative	noster	nostra	nostrum	nostrī	nostrae	nostra
Accusative	nostrum	nostram		nostrōs	nostrās	
Genitive	nostrī	nostrae	nostrī	nostrōrum	nostrārum	nostrōrum
Dative	nostrō		nostrō	nostrīs		
Ablative		nostrā				

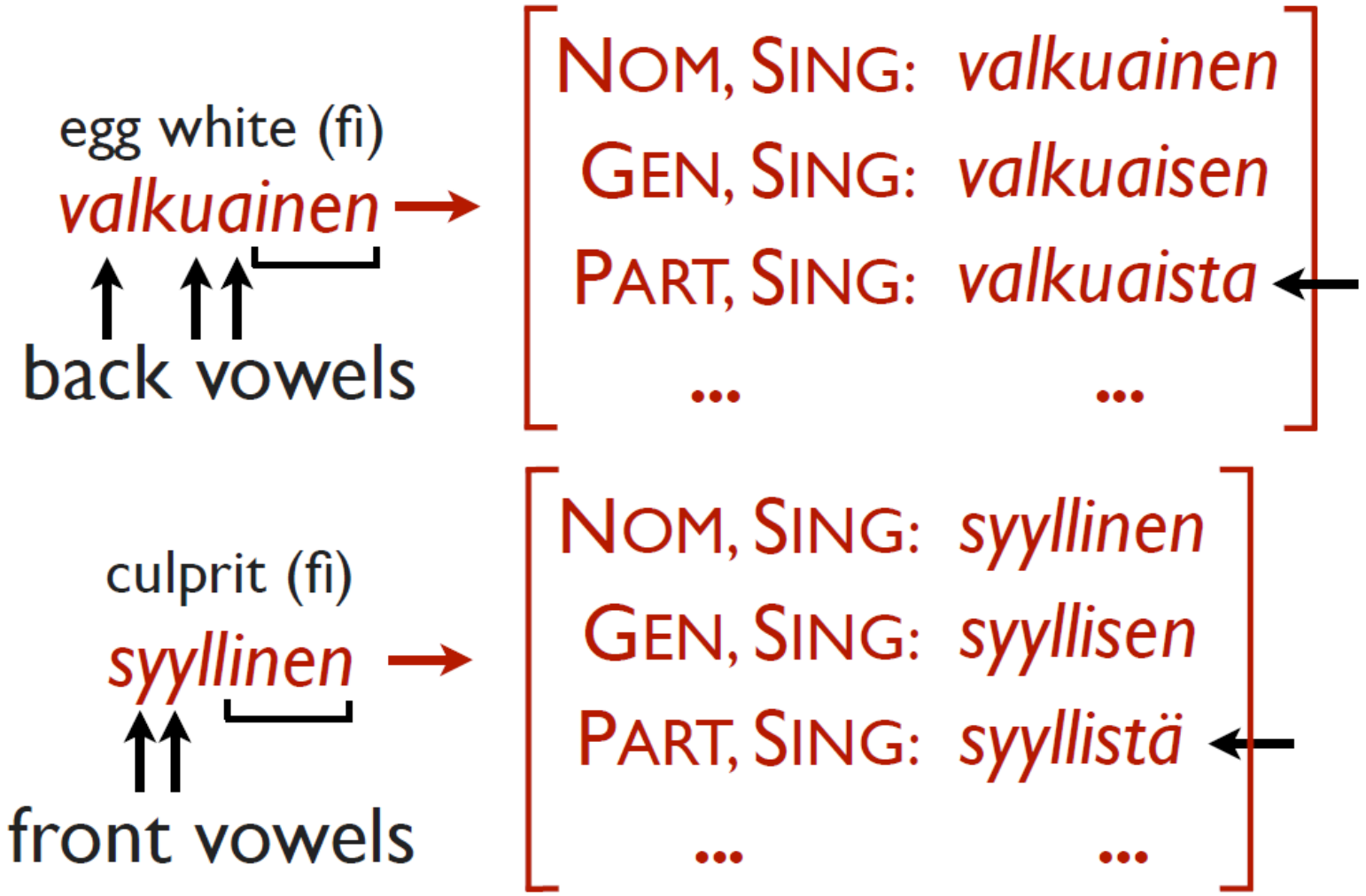
https://en.wikipedia.org/wiki/Latin_declension

Automatic Inflection



Durrett and Denero 2013]

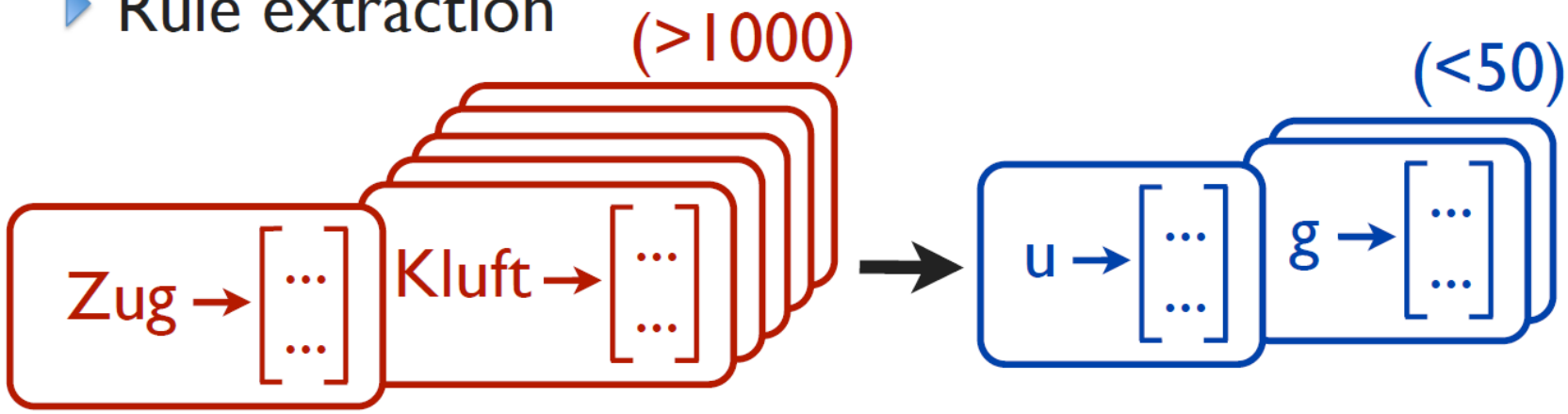
Automatic Inflection



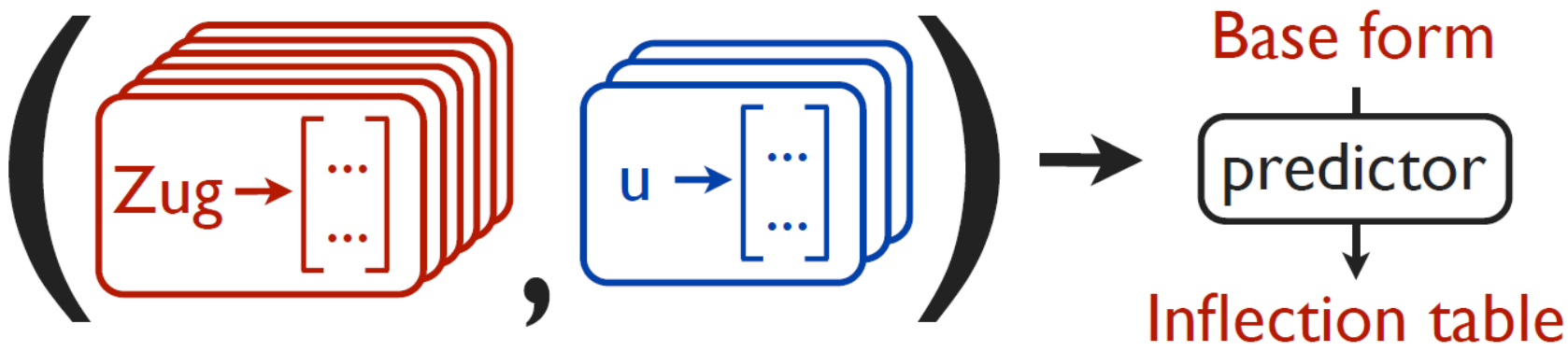
Durrett and Denero 2013]

Automatic Inflection

- Rule extraction



- Paradigm prediction



Durrett and Denero 2013]

Morphological Analysis

- sleeps = sleep + V + 3P + SG
- done = do + V + PP

Nice example from Kemal Oflazer

Dancing in Andalusia

- A poem by the early 20th century Turkish poet Yahya Kemal Beyatlı.

ENDÜLÜSTE RAKS

Zil, şal ve gül, bu bahçede raksın bütün hızı
Şevk akşamında Endülüs, üç defa kırmızı
Aşkın sihirli şarkısı, yüzlerce dildedir
İspanya neşesiyle bu akşam bu zildedir

Yelpaze gibi çevrilir birden dönüşleri
İşveyle devriliş, saçılış, örtünüşleri
Her rengi istemez gözümüz şimdi aldadır
İspanya dalga dalga bu akşam bu şaldadır

Alnında halka halkadır âşufte kâkülü
Göğsünde yosma Gırnata'nın en güzel gülü
Altın kadeh her elde, güneş her gönüldedir
İspanya varlığıyla bu akşam bu güldedir

Raks ortasında bir durup oynar, yürür gibi
Bir baş çevirmesiyle bakar öldürür gibi
Gül tenli, kor dudaklı, kömür gözlü, sürmeli
Şeytan diyor ki sarmalı, yüz kerre öpmeli

Gözler kamaştıran şala, meftûn eden güle
Her kalbi dolduran zile, her sineden ole!

ENDÜLÜSTE RAKS

Zil, şal ve gül, bu bahçede raksın bütün hızı
Şevk akşamında Endülüs, üç defa kırmızı
Aşkın sihirli şarkısı, yüzlerce dildedir
İspanya neşesiyle bu akşam bu zildedir

Yelpaze gibi çevrilir birden dönüşleri
İşveyle devriliş, saçılış, örtünüşleri
Her rengi istemez gözümüz şimdi aldadır
İspanya dalga dalga bu akşam bu şaldadır

Alnında halka halkadır âşufte kâkülü
Göğsünde yosma Gırnata'nın en güzel gülü
Altın kadeh her elde, güneş her gönüldedir
İspanya varlığıyla bu akşam bu güldedir

Raks ortasında bir durup oynar, yürür gibi
Bir baş çevirmesiyle bakar öldürür gibi
Gül tenli, kor dudaklı, kömür gözlü, sürmeli
Şeytan diyor ki sarmalı, yüz kerre öpmeli

Gözler kamaştıran şala, meftûn eden güle
Her kalbi dolduran zile, her sineden ole!

zildedir: a verb derived from the locative case of the noun “zil” (castanet)
“is at the castanet”

dönüşleri: plural infinitive and possessive form of the verb “dön” (rotate)

“their (act of) rotating”

istemez: negative present form of the verb “iste” (want)
“it does not want”

varlığıyla: singular possessive instrumental-case of the noun “varlık” (wealth)
“with its wealth”

kamaştıran: present participle of the verb “kamaş” (blind)
“that which blinds….”

Aligned Verses

Zil, şal ve gül, bu bahçede raksın bütün hızı
Şevk akşamında Endülüs, üç defa kırmızı
Aşkın sihirli şarkısı, yüzlerce dildedir
İspanya neş'esiyle bu akşam bu zildedir

Castañuela, mantilla y rosa. El baile veloz llena el jardín...
En esta noche de jarana, Andalucíá se ve tres veces carmesí...
Cientas de bocas recitan la canción mágica del amor.
La alegría española esta noche, está en las castañuelas.

Castanets, shawl and rose. Here's the fervour of dance,
Andalusia is threefold red in this evening of trance.
Hundreds of tongues utter love's magic refrain,
In these castanets to-night survives the gay Spain,

Zimbel, Schal und Rose- Tanz in diesem Garten loht.
In der Nacht der Lust ist Andalusien dreifach rot!
Und in tausend Zungen Liebeszauberlied erwacht-
Spaniens Frohsinn lebt in diesen Zimbeln heute Nacht!

Agglutinative Languages

- How does English become Turkish?

if	we	will	be	able	to	make	. . .	become	strong
if	we	will	be	able	to	make	. . .	become	strong
. . .	strong	become	to	make	be	able	will	if	we
. . .	sağlam	+laş		+tır		+abil	+ecek	+se	+k

⇓

. . . **sağlamlaştırabileceksek**

Slide from Kemal Oflazer

Aymara

- ch’uñüwinkaskirïyätwa
- ch’uñu +: +wi +na -ka +si -ka -iri +: +ya:t(a) +wa

- I was (one who was) always at the place for making ch’uñu’

ch’uñu	N	‘freeze-dried potatoes’	
	+:	N>V	be/make …
	+wi	V>N	place-of
	+na		in (location)
	-ka	N>V	be-in (location)
	+si		continuative
	-ka		imperfect
	-iri	V>N	one who
	+:	N>V	be
	+ya:ta	1P	recent past
	+wa		affirmative sentencial

Example Courtesy of Ken Beesley

- Finnish Numerals
  - Finnish numerals are written as one word and all components inflect and agree in all aspects
  - **Kahdensienkymmenensienkahdeksansien**

second	tenth	eighth (28th)
kaksi+Ord+Pl+Gen	kymmenen+Ord+Pl+Gen	kahdeksan+Ord+Pl+Gen
kahde ns i en	**kymmene ns i en**	**kahdeksa ns i en**

- Swahili
  - walichotusomea = wa[Subject Pref]+li[Past]+cho[Rel Prefix]+tu[Obj Prefix 1PL]+som[read/Verb]+e[Prep Form]+a[]
    - that (thing) which they read for us
  - tulifika=tu[we]+li[Past]+fik[arrive/Verb]+a[]
    - We arrived
  - ninafika=ni[I]+na[Present]+fik[arrive/Verb]+a[]
    - I am arriving

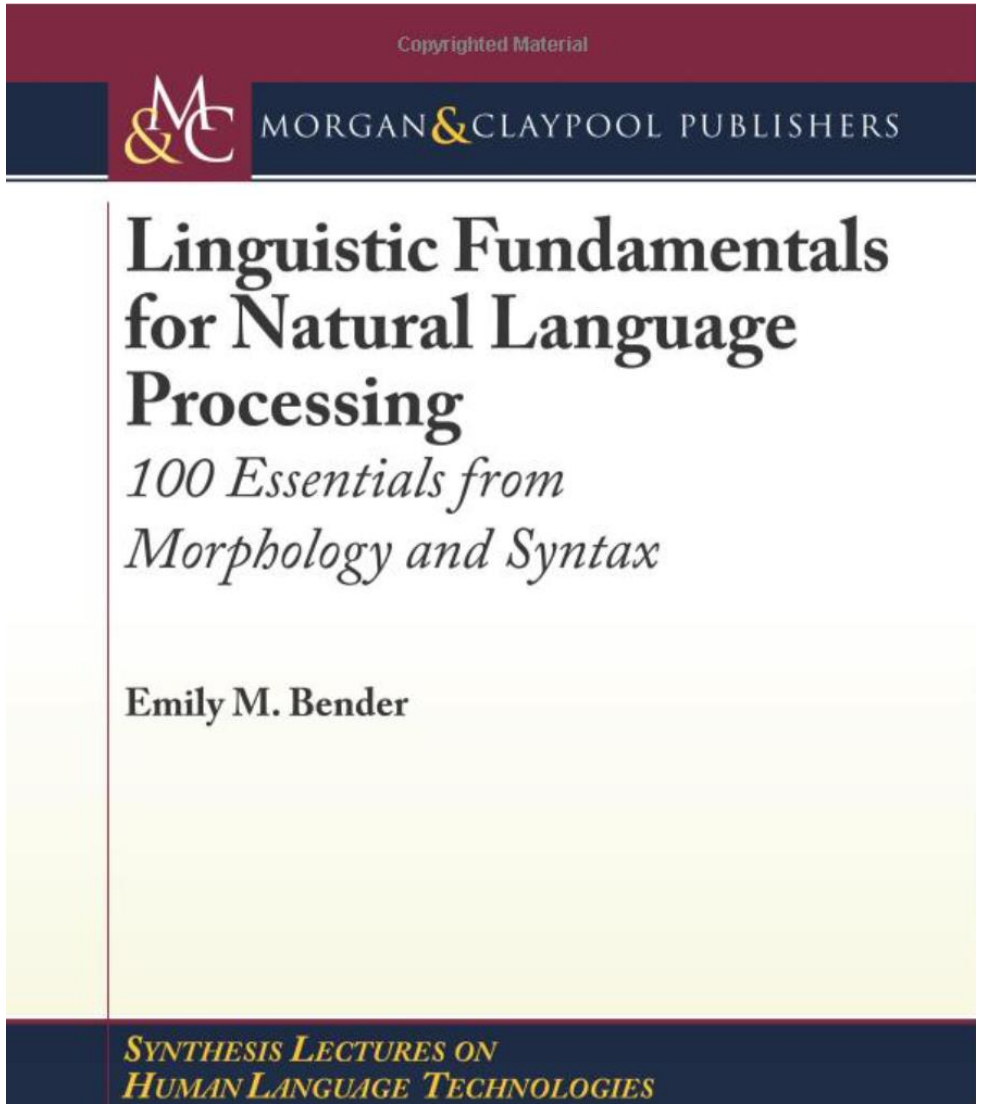
Example from Kemal Oflazer

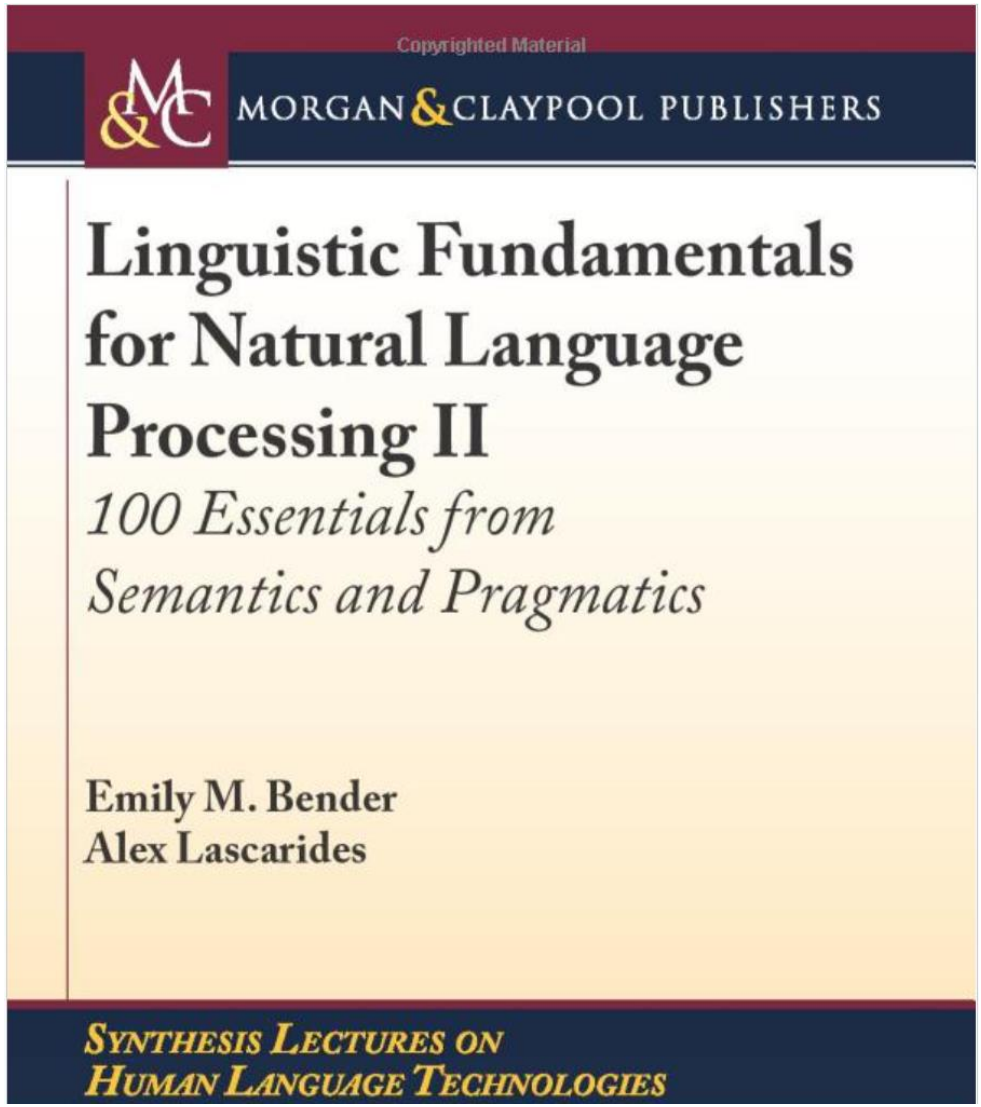
Turkish Vowel Harmony

	Front		Back	
	Unrounded	Rounded	Unrounded	Rounded
High	i	ü	ı	u
Low	e	ö	a	o

- Back vowels
  - in the room → oda**da**
  - at the door → kapı**da**
- Front vowels
  - at home → ev**de**
  - at the lake → göl**de**
  - on the bridge → köprü**de**

Books by Emily Bender





Introduction to NLP

144

Word Distributions

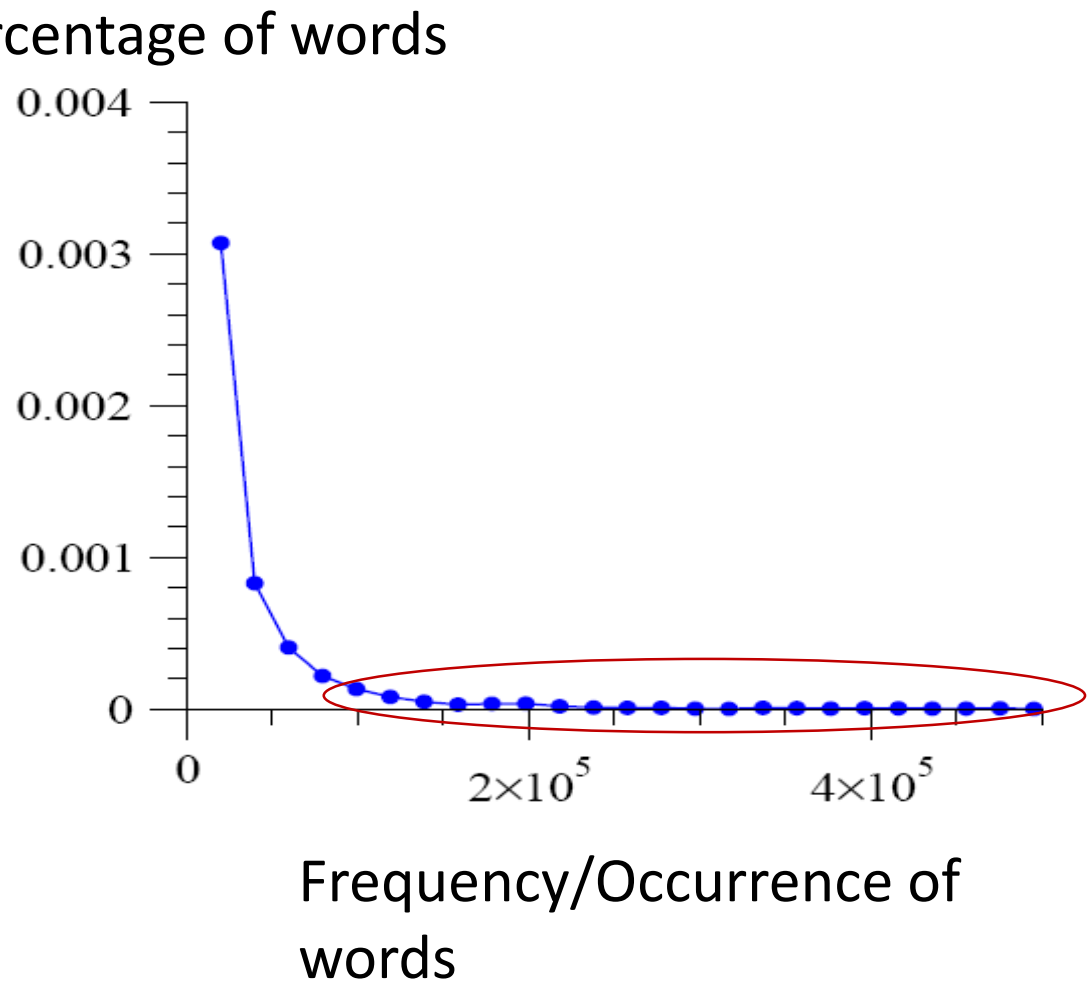
Word Distributions

- Words are not distributed evenly!
  - Same goes for letters of the alphabet (ETAOIN SHRDLU), city sizes, wealth, etc.
- Usually, the 80/20 rule applies
  - 80% of the wealth goes to 20% of the people or it takes 80% of the effort to build the easier 20% of the system
  - more examples coming up…

Stop Words

- Fact:
  - 250-300 most common words in English account for 50% or more of a given text.
- Example:
  - “the” and “of” represent 10% of tokens. “and”, “to”, “a”, and “in” - another 10%. Next 12 words - another 10%.
- Moby Dick Ch.1:
  - 859 unique words (types), 2256 word occurrences (tokens). Top 65 types cover 1132 tokens (> 50%).
- Token/type ratio:
  - 2256/859 = 2.63

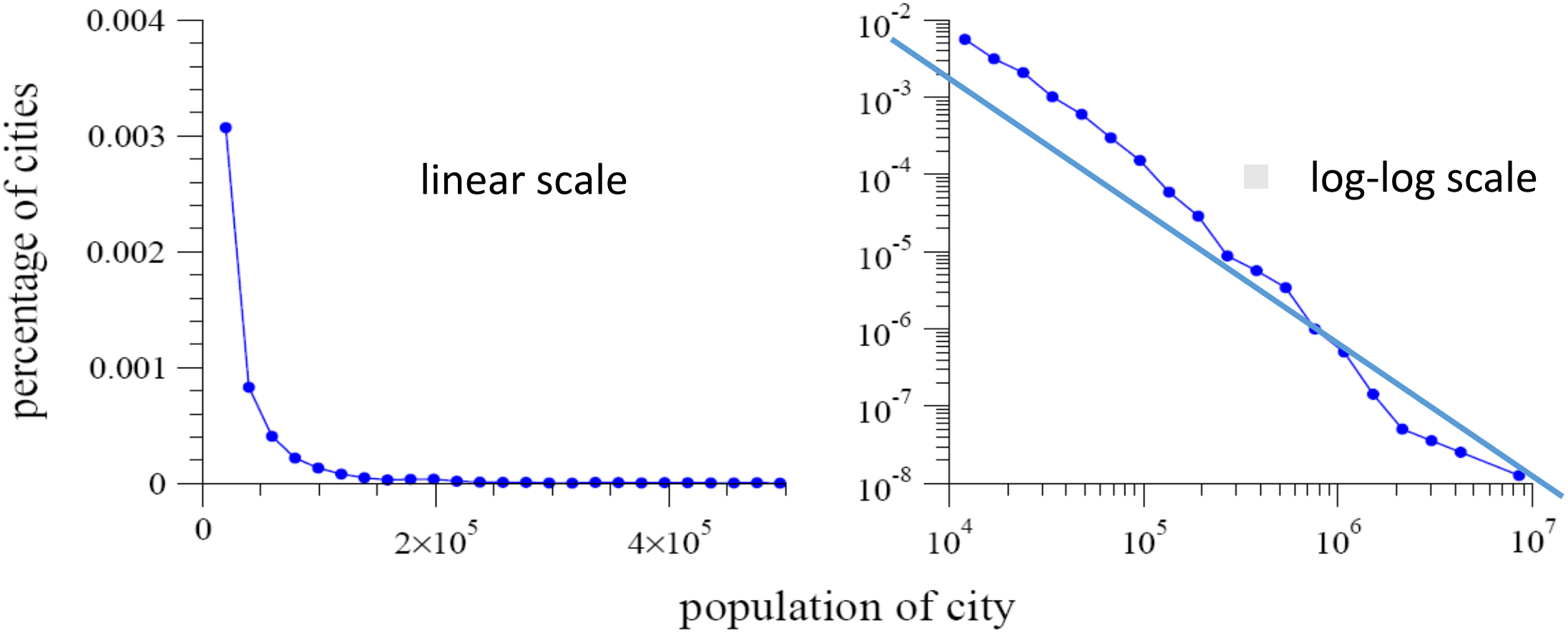
Power-law Distribution



- Power-law
  - Many words with a small frequency of occurrence
  - A few words with a very large frequency
  - High skew (asymmetry)
- Compared to a normal distribution:
  - Many people of a medium height
  - Almost nobody of a very high or very low height
  - Symmetry

Slide from Qiaozhu Mei

Scaling the Axes



- Long-tail on a linear scale - straight line on a log-log plot

Power Law Distribution

- The probability of observing an item of size *x* is given by

$$p(x) = cx^{-\alpha}$$

normalization constant (probabilities over all *x* must sum to 1)

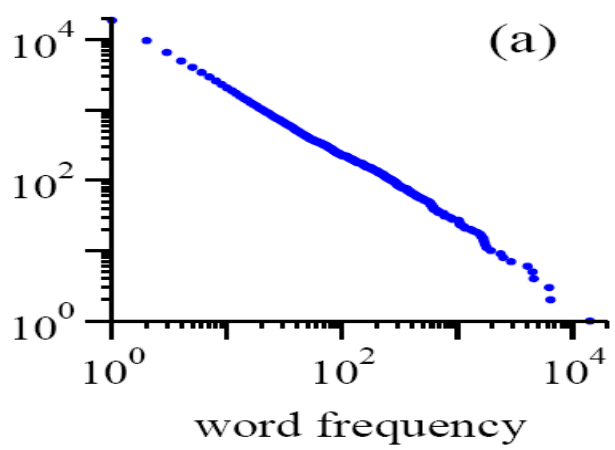
α : scaling exponent, or power law exponent

- Straight line on a log-log plot

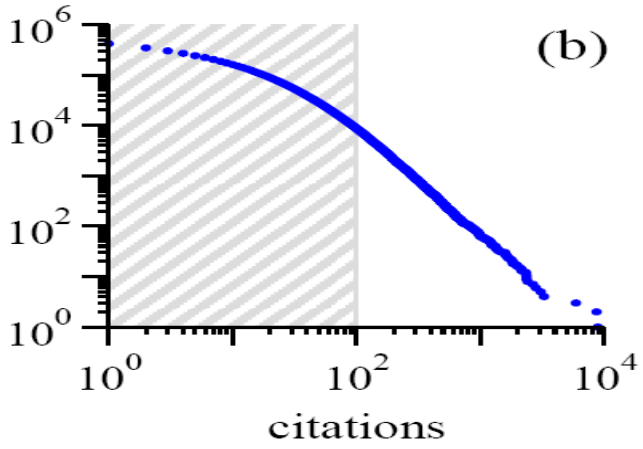
$$\ln\big(p(x)\big) = c - \alpha\, ln(x)$$

Power Laws Are Seemingly Everywhere

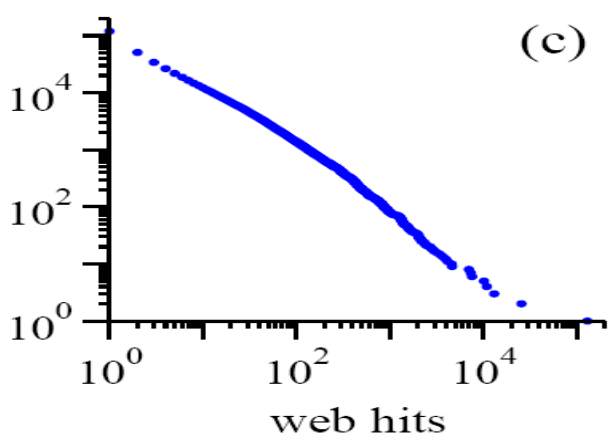
note: these are cumulative distributions



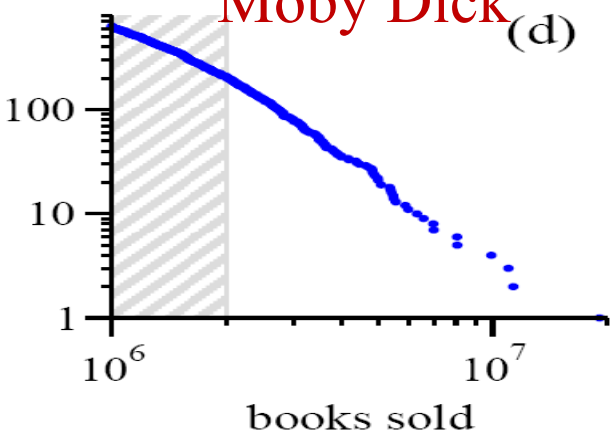
Moby Dick



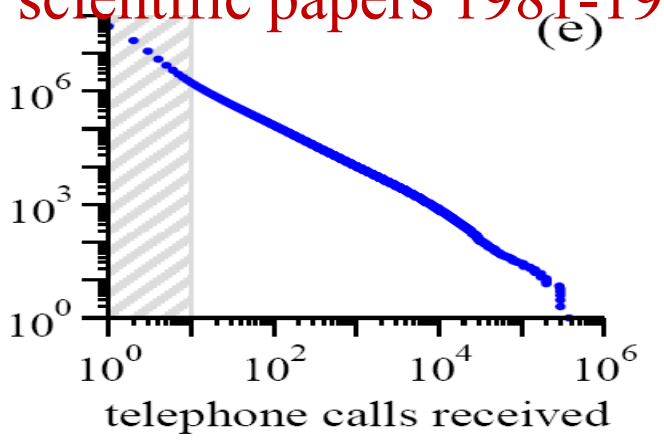
scientific papers 1981-1997



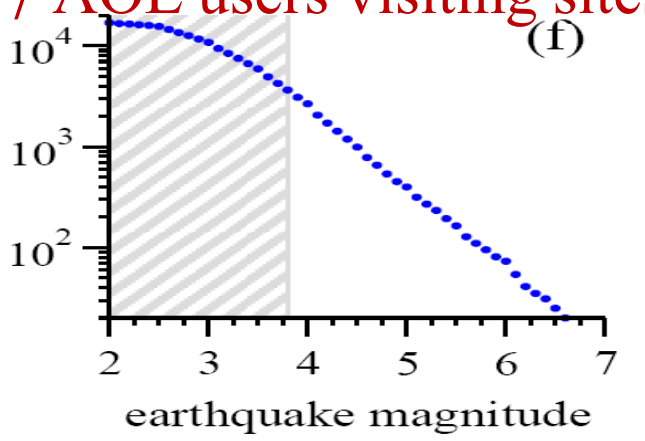
AOL users visiting sites ‘97



bestsellers 1895-1965



AT&T customers on 1 day

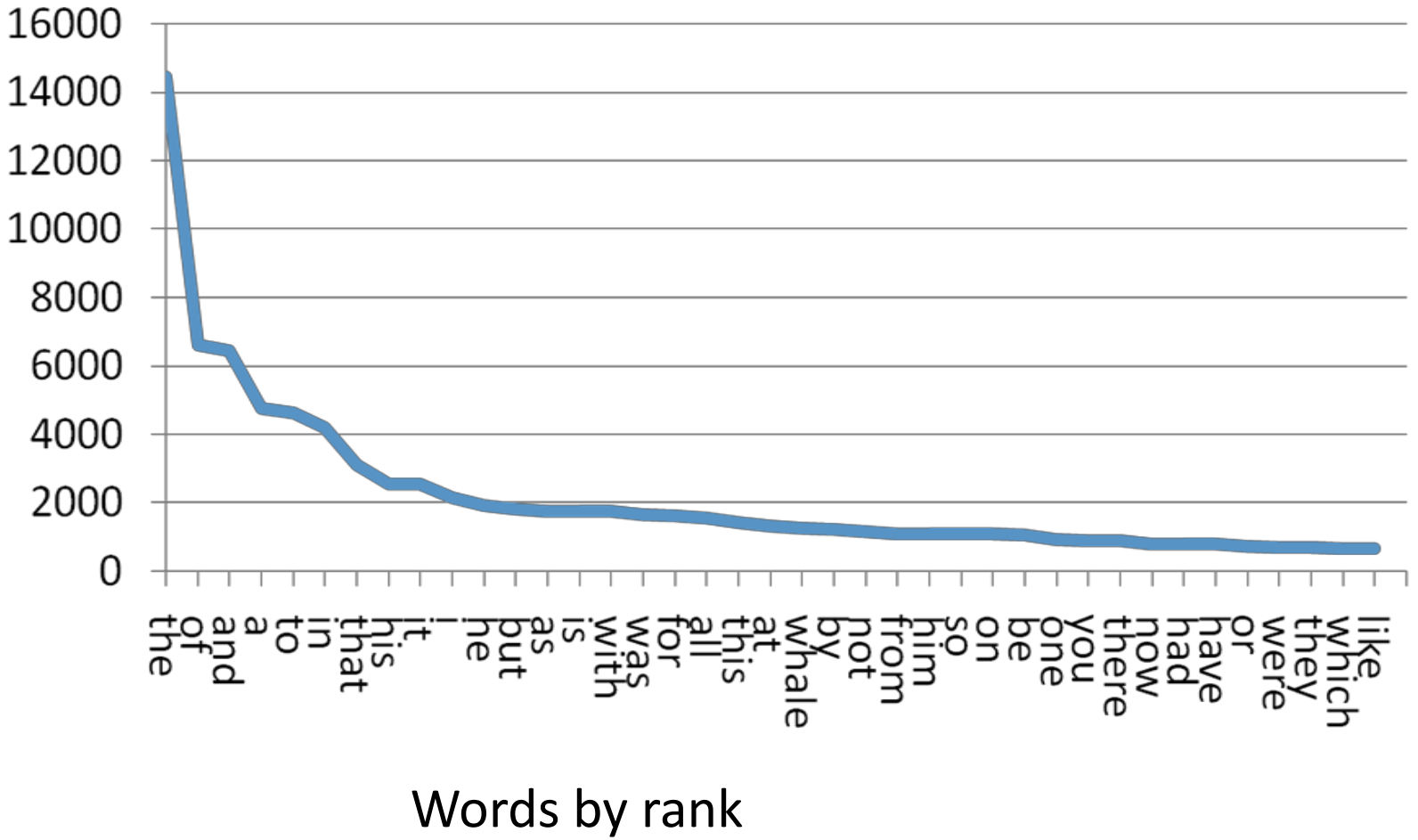


California 1910-1992

Source:MEJ Newman, ’Power laws, Pareto distributions and Zipf’s law’, *Contemporary Physics* **46**, 323–351 (2005)

Zipf's Distribution

Word frequency



$$p(k) \sim k^{-\alpha}$$

Zipf's Law in Natural Language

$$\text{Rank} \times \text{Frequency} \approx \text{Constant}$$

– Constant ≈ 0.1 × Length of collection (in words)
– Not accurate at the tails, but accurate enough for our purposes

Rank	Term	Freq.	Z	Rank	Term	Freq.	Z
1	the	69,9	0.07	6	in	21,3	0.12
2	of	36,4	0.07	7	that	10,5	0.07
3	and	28,8	0.08	8	is	10,0	0.08
4	to	26.1	0.10	9	was	9,81	0.08
5	a	23,2	0.11	10	he	9,54	0.09