NLP

Introduction to NLP

151

NLP Tasks

Part of Speech Tagging

The swimmer is getting ready to run in the final race.

Part of Speech Tagging

The swimmer is getting ready to **run** in the final race.

- Run – verb or noun?
- Final – noun or adjective?
- Race – verb or noun?

Part of Speech Tagging

The candidate is preparing for his **run** for the presidency.
The swimmer is getting ready to **run** in the final race.

Parsing

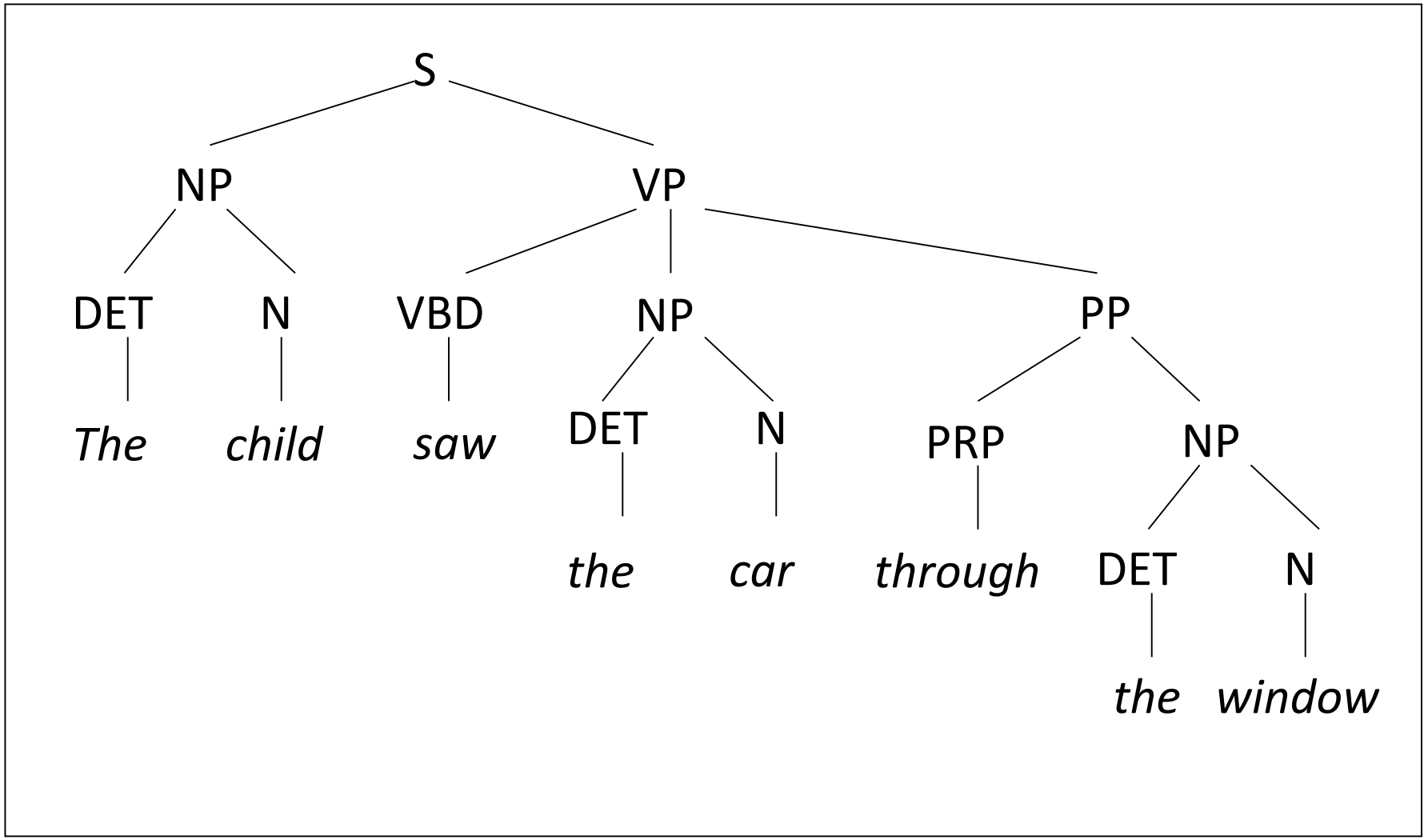
- Myriam slept.
- Myriam wrote a novel.
- Myriam gave Sally flowers.
- Myriam ate salad with a fork.

Phrase-Structure Grammar

S	→	NP	VP
NP	→	DET	N
NP	→	NP	PP
VP	→	VBD	
VP	→	VBD	NP
VP	→	VBD	NP NP
VP	→	VP	PP
PP	→	PRP	NP

DET	→	*the*
DET	→	*that*
DET	→	*a*
N	→	*child*
N	→	*window*
N	→	*car*
VBD	→	*found*
VBD	→	*ate*
VBD	→	*saw*
PRP	→	*in*
PRP	→	*of*
PRP	→	*through*

Parse Trees


S
NP
VP
DET
N
VBD
NP
PP
The
child
saw
DET
N
PRP
NP
the
car
through
DET
N
the
window

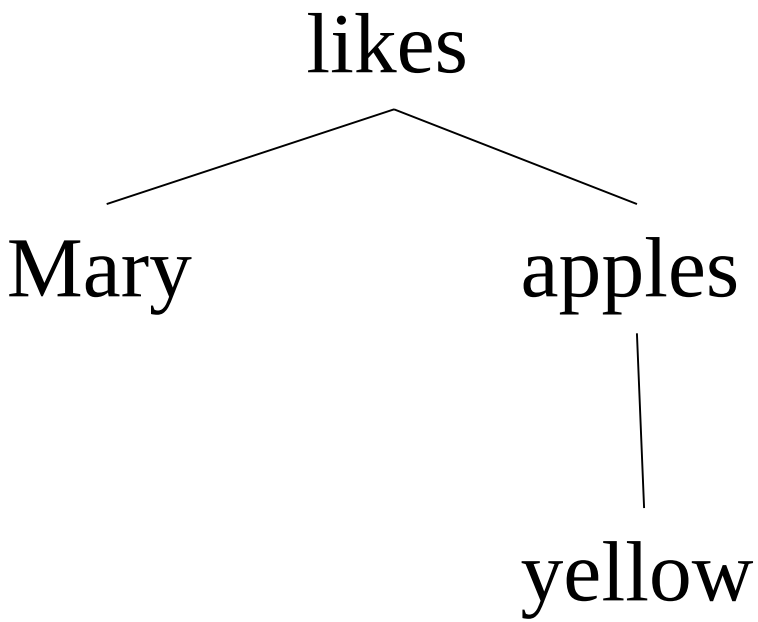
Stanford Parser

Parser Output

```
(ROOT
  (S
    (S
      (NP
        (NP (NN Housing) (NNS starts))
        (, ,)
        (NP
          (NP (DT the) (NN number))
          (PP (IN of)
            (NP
              (NP (JJ new) (NNS homes))
              (VP (VBG being)
                (VP (VBN built))))))
        (, ,))
```

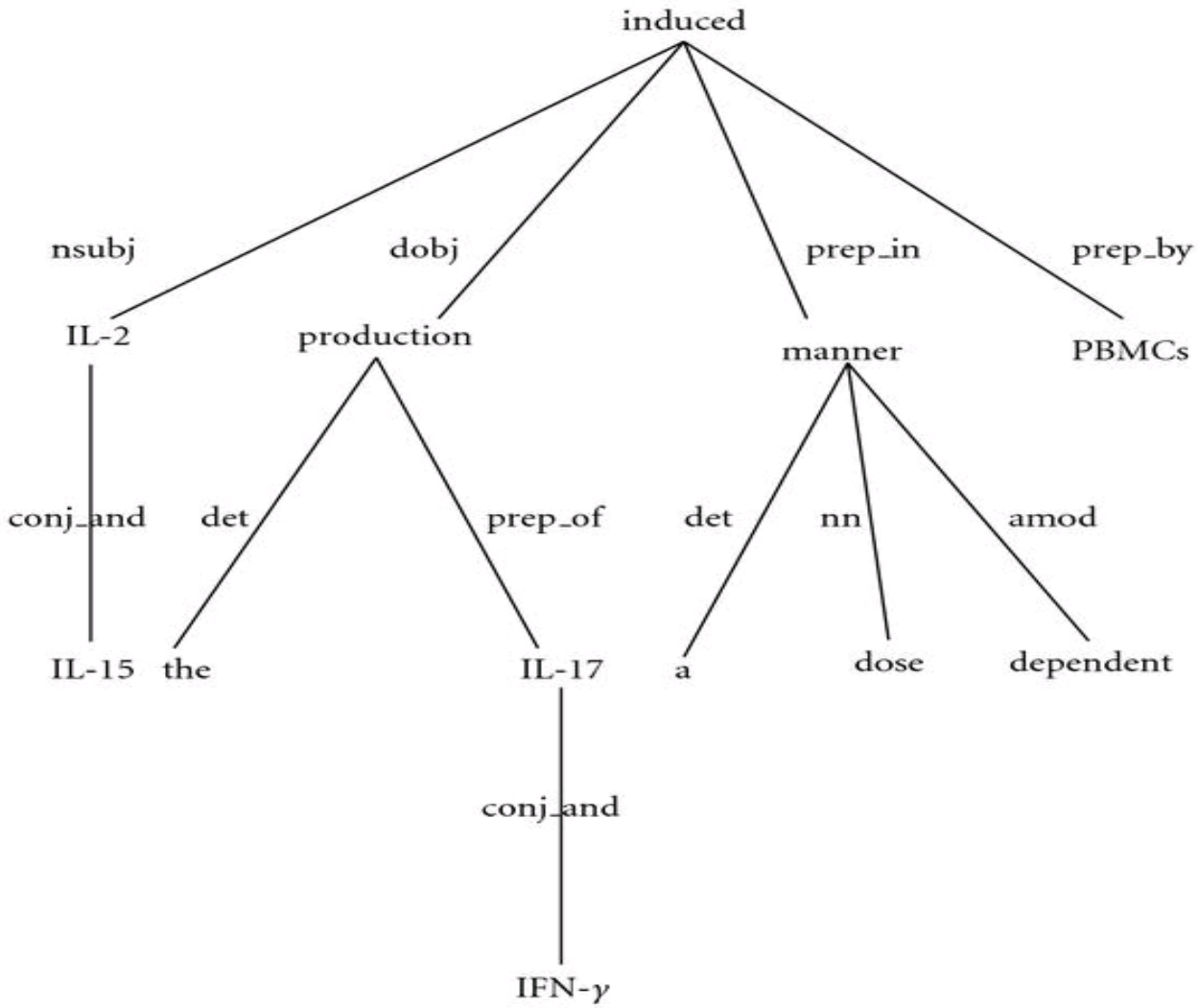
```
(VP (VBD rose)
        (NP (CD 7.2) (NN %))
        (PP (IN in)
          (NP (NNP March)))
        (PP (TO to)
          (NP
            (NP (DT an) (JJ annual) (NN rate))
            (PP (IN of)
              (NP (CD 549,000) (NNS units)))))
        (, ,)
        (ADVP (RB up)
          (PP (IN from)
            (NP
              (NP (DT a) (VBN revised) (CD 512,000))
              (PP (IN in)
                (NP (NNP February)))))))
    (, ,)
    (NP (DT the) (NNP Commerce) (NNP Department))
    (VP (VBD said))
    (. .)))
```

Dependency Parsing



Dependency Parsing

IL-2 and IL-15 induced the production of IL-17 and IFN-γ by PBMCs in a dose dependent manner.



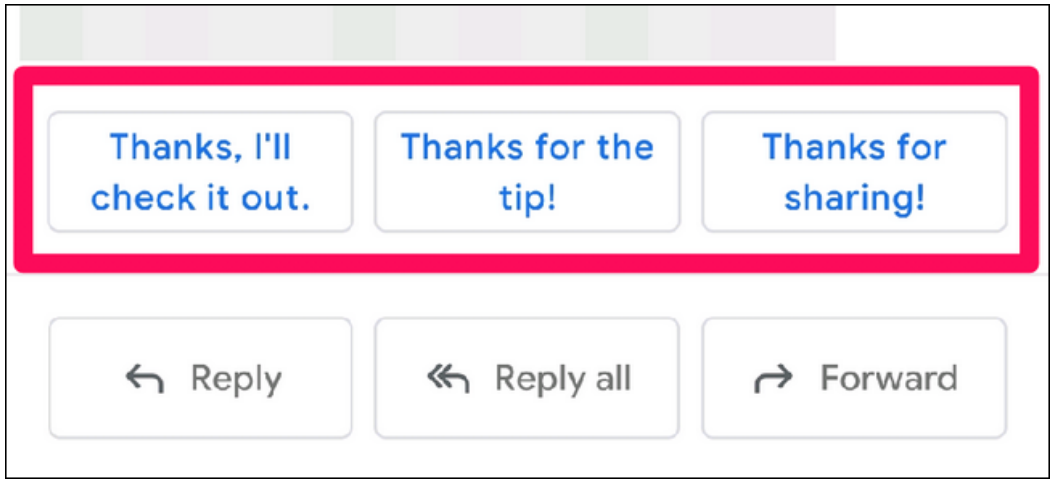
Information Extraction

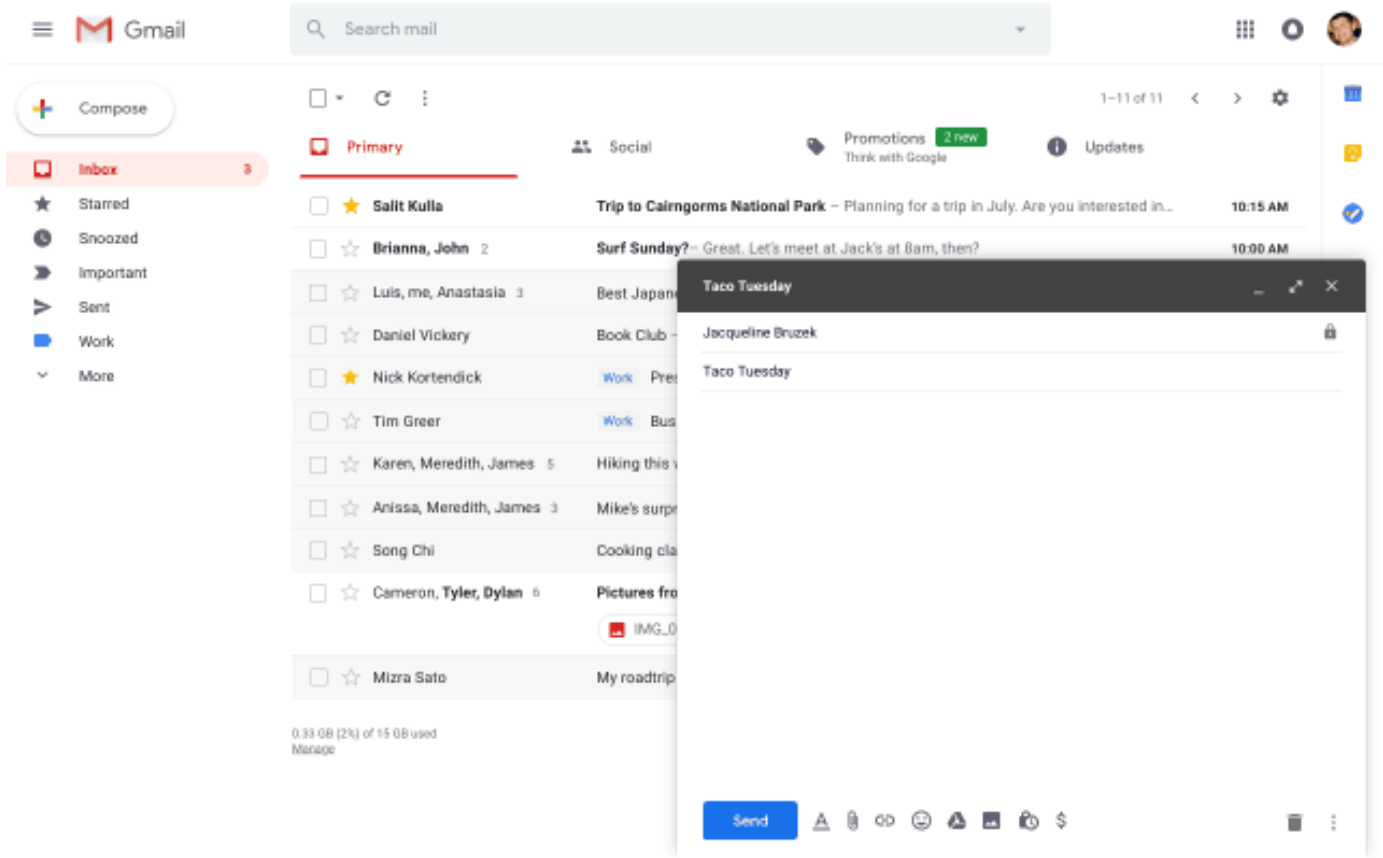
- RESEARCH ALERT-Wells Fargo cuts PPD Inc to market perform
- China Southern Air Upgraded To Overweight From Neutral-HSBC
- CITIGROUP RAISES INGERSOLL RAND <IR.N> TO HOLD FROM SELL
- TCF Financial Corp Raised To Overweight From Neutral By JPMorgan
- BAIRD CUTS KIOR INC <KIOR.O> TO UNDERPERFORM RATING
- BRIEF-RESEARCH ALERT-Global Equities Research cuts LinkedIn to equal weight

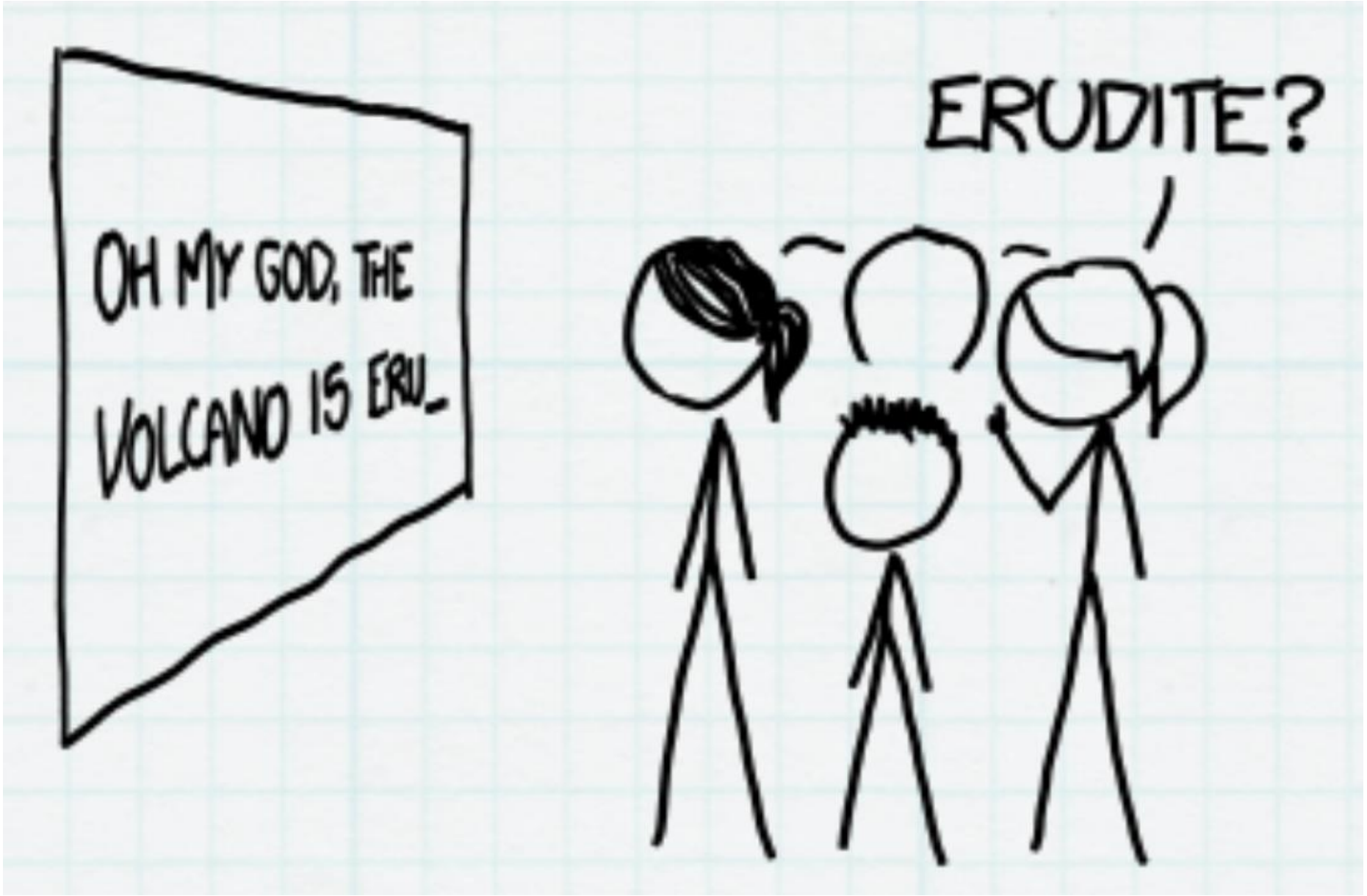
Information Extraction

DATE/TIME	TICKER	COMPANY	SOURCE	OLD	NEW	CHANGE
		PPD Inc	Wells Fargo		market perform	↓
		China Southern Air	HSBC	Neutral	Overweight	↑
	IR.N	INGERSOLL RAND	CITIGROUP	SELL	HOLD	↑
		TCF Financial Corp	JPMorgan	Neutral	Overweight	↑
	KIOR.O	KIOR INC	BAIRD		UNDERPERFORM	↓
		LinkedIn	Global Equities Research		equal weight	↓

Text Completion







https://what-if.xkcd.com/34/

Semantics

- First order logic
- Inference/deduction
- Semantic analysis

$$\forall x,y: \textit{Mother}\,(x,y) \Rightarrow \textit{Parent}\,(x,y)$$

Reading Comprehension

Mars Polar Lander - Where Are You?

(January 18, 2000) After more than a month of searching for a signal from NASA's Mars Polar Lander, mission controllers have lost hope of finding it. The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life ever existed on Mars. Polar Lander was to have touched down December 3 for a 90-day mission. It was to land near Mars' south pole. The lander was last heard from minutes before beginning its descent. *The last effort to communicate with the three-legged lander ended with frustration at 8 a.m Monday.* "We didn't see anything," said Richard Cook, the spacecraft's project manager at NASA's Jet Propulsion Laboratory. The failed mission to the Red Planet cost the American government more than $200 million dollars. Now, space agency scientists and engineers will try to find out what could have gone wrong. They do not want to make the same mistakes in the next mission.
(sources: CBC "For Kids" web page, Associated Press, CBC News Online, CBC Radio news, NASA)

1. When did the mission controllers lose hope of communicating with the lander?
   Answer: *8AM, Monday Jan. 17, 2000*
2. Who is the Polar Lander's project manager?
3. Where on Mars was the spacecraft supposed to touch down?
4. What did the Mars Global Surveyor do?
5. What was the mission of the Mars Polar Lander?

Pranav Anand, Eric Breck, Brianne Brown, Marc Light, Gideon Mann, Ellen Riloff, Mats Rooth, Michael Thelen. 2000. Fun with Reading Comprehension

Word Sense Disambiguation

- “The thieves took off with 100 gold **bars**”.
  - Did they steal 100 drinking establishments?
  - Or 100 measures of a song?

WSD is Important for Translation

- Paul plays soccer
  - Paul joue au football
- Paul plays the guitar
  - Paul joue de la guitare
- “wall” in German
  - die Chinesische Mauer (The Great Wall of China)
  - (otherwise Wand)
- “wall” in Spanish
  - pared, muro, muralla

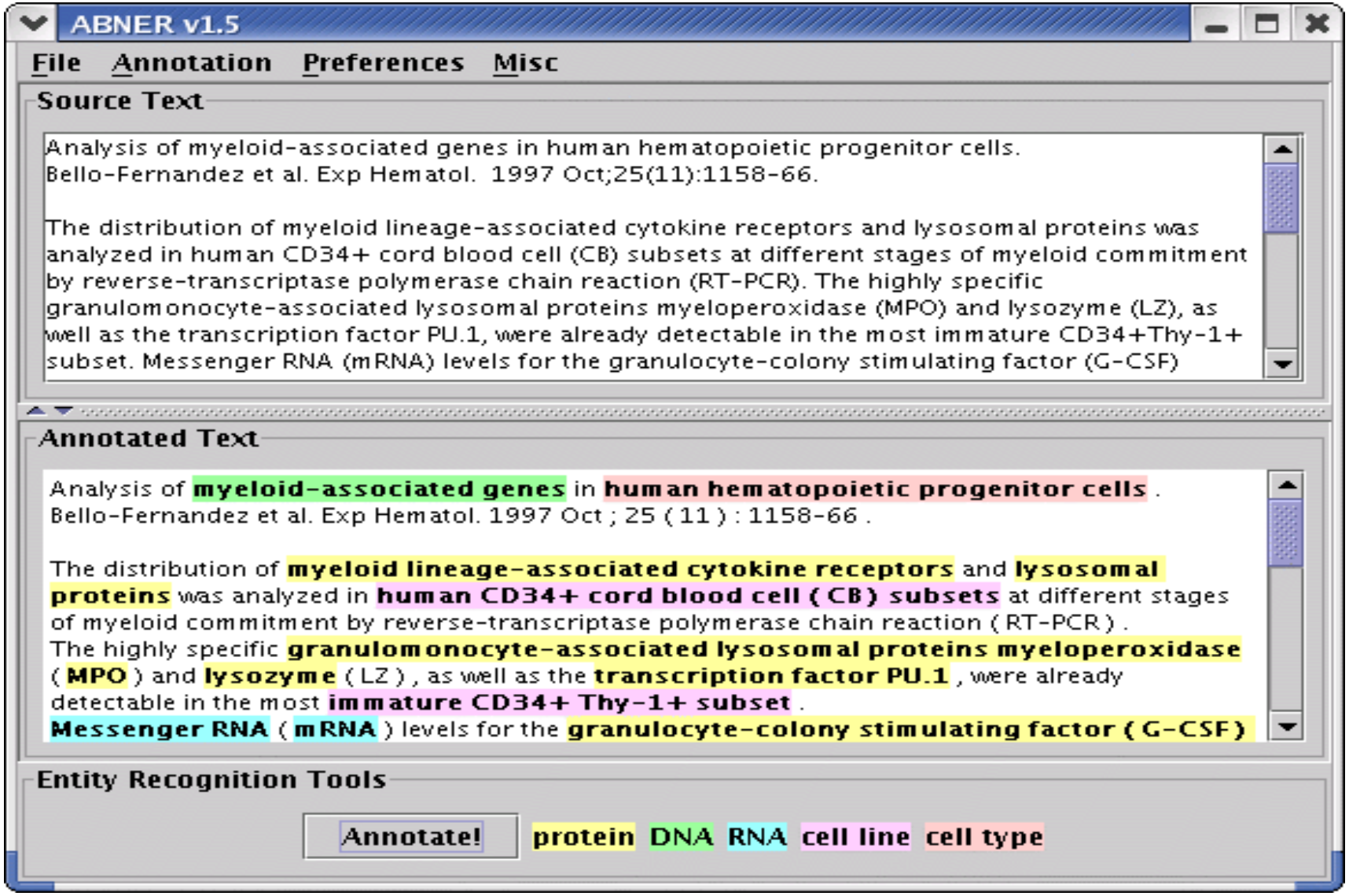
Named Entity Recognition

Wolff, currently a journalist in Argentina, played with Del Bosque in the final years of the seventies in Real Madrid.

```
     Wolff B-PER
         , O
 currently O
         a O
journalist O
        in O
 Argentina B-LOC
         , O
    played O
      with O
       Del B-PER
    Bosque I-PER
        in O
       the O
     final O
     years O
        of O
       the O
 seventies O
        in O
      Real B-ORG
    Madrid I-ORG
         . O
```

- http://cogcomp.cs.illinois.edu/page/demo_view/NER
- http://nlp.stanford.edu:8080/ner/

Named Entity Recognition



http://pages.cs.wisc.edu/~bsettles/**abner**

Coreference Resolution

- Barack Obama visited China. The US president met with his Chinese counterpart.
- Cynthia went to see her aunt at the hospital. She was scheduled for surgery on Monday.
- Because he was sick, Michael stayed home on Friday.

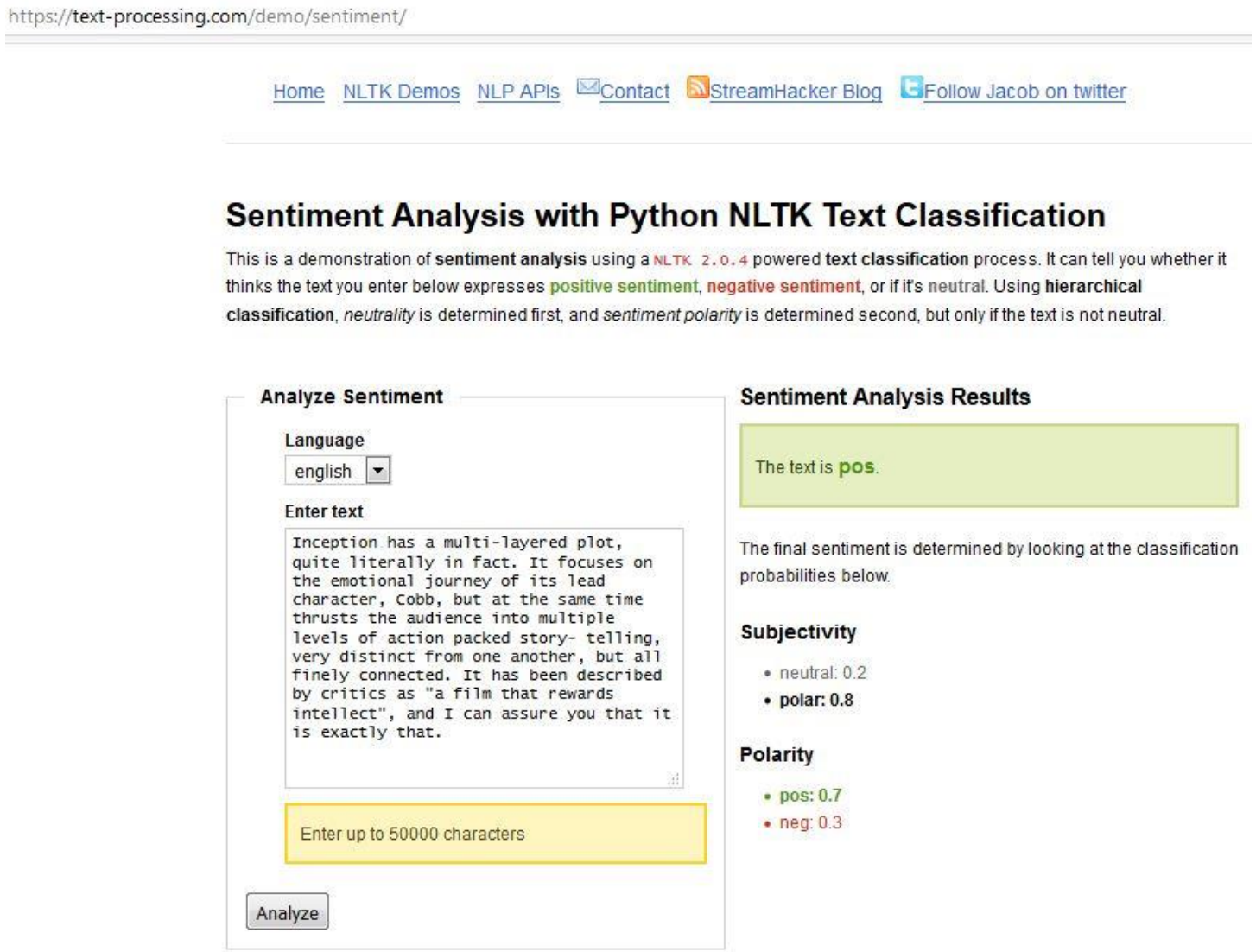
Question Answering

Sentiment Analysis

“I like the camera because I can edit images so easily, exactly as I do my iPad. I have found that its difficult to frame a picture when there isn't a zoom function as with the iPad. With this camera I can adjust my images by cropping as I did with my iPad but better yet, this camera has a built in zoom. A stretch or pinch of the fingers bring in the subject closer or back out again. With this iPhone I can also, as I dido with my iPad, enhance, crop, rotate, red eye reduce, and set a range of tints. I am also quite impressed with the quality of the images. Pretty darn good especially better than I expected for low light situations where I can use the built-in flash! Quite frankly I was quite surprised with these built in features. I also hope too experiment with and learn what HDR photography is. It's built into this iPhone and can be activated by a the touch of an icon. ”

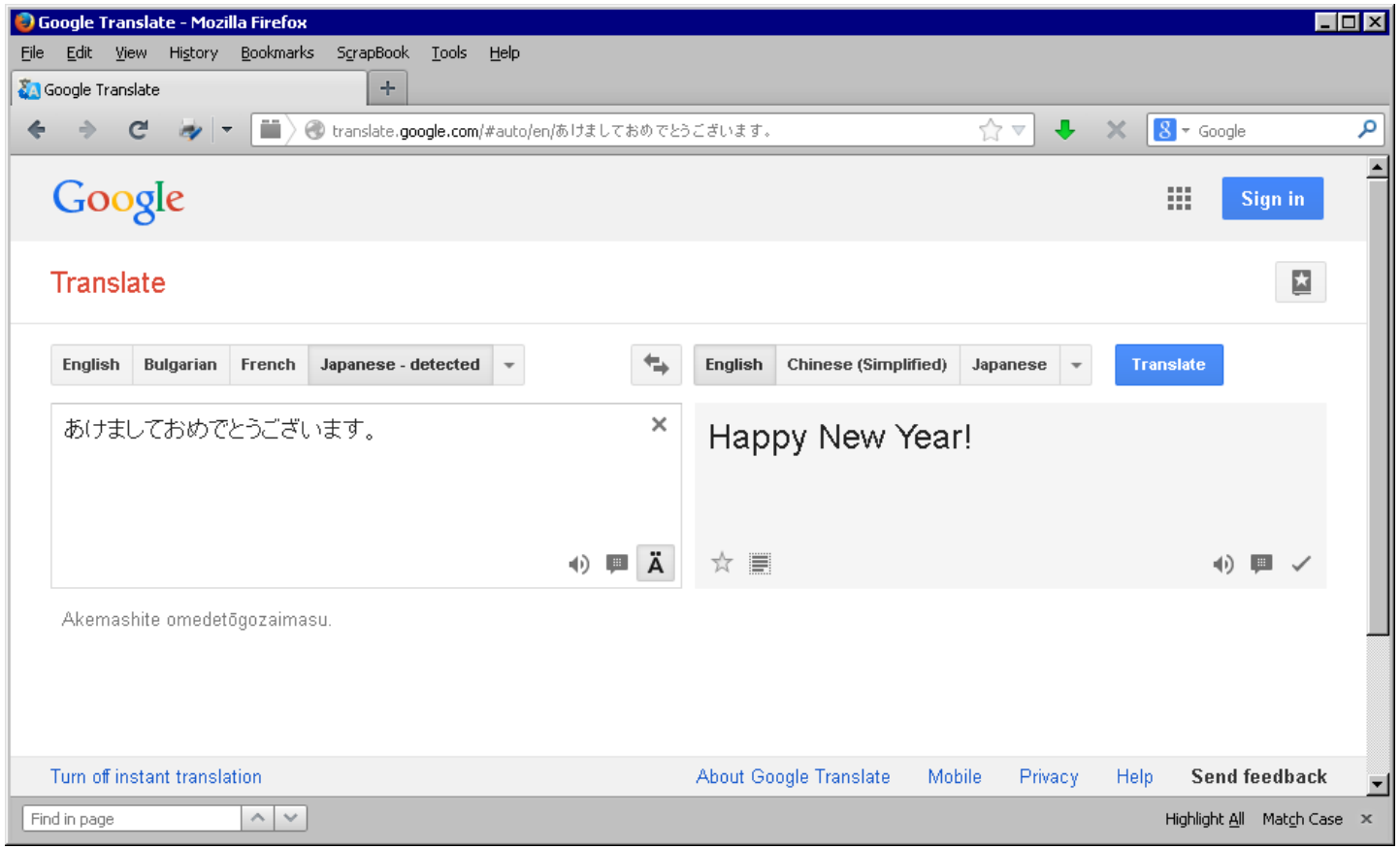
http://www.epinions.com/review/apple_iphone_5c_latest_model_16gb_graphite_unlocked_smartphone/content_640679317124

Sentiment Analysis



Machine Translation

- あけましておめでとうございます。
- Happy New Year!



Elephants are social animals. They live with their families, give hugs and call each other by using their trunks as trumpets. They also might know how to help each other.

In a recent elephant study by researchers from the United States and Thailand, pairs of giant animals learned to work together to get some ears of corn. Other animals, especially some primates, are already known to work together to complete tasks, but now elephants have joined the club. Perhaps the finding is not too surprising: Scientists suspect that elephants, with their big brains and survival savvy, may be among the smartest animals on the planet.

Joshua Plotnik, who worked on the study, told Science News that the animals didn't just learn a trick. Instead, the ways the elephants behaved show that they understand how working together brings benefits to everyone involved. Plotnik is a comparative psychologist now at the University of Cambridge in England. Psychology is the study of behaviors and mental processes, and comparative psychologists study how animals other than humans behave.

Les éléphants sont des animaux sociaux. Ils vivent avec leur famille, faire des câlins et appeler les uns les autres en utilisant leurs troncs trompettes. Ils pourraient également savoir comment aider les uns les autres.

Dans une étude récente d'éléphants par des chercheurs des États-Unis et la Thaïlande, des paires d'animaux géants ont appris à travailler ensemble pour obtenir des épis de maïs. D'autres animaux, en particulier des primates, sont déjà connus pour travailler ensemble pour accomplir des tâches, mais maintenant, les éléphants ont rejoint le club. Peut-être le résultat n'est pas trop surprenant: Les scientifiques soupçonnent que les éléphants, avec leurs gros cerveaux et de bon sens de survie, peut-être parmi les plus intelligents des animaux sur la planète.

Joshua Plotnick, qui a travaillé sur l'étude, dit Nouvelles de la Science que les animaux n'ont pas seulement appris un truc. Au lieu de cela, les moyens les éléphants se comportent montrent qu'ils comprennent comment travailler ensemble apporte des avantages à toutes les personnes impliquées. Plotnik est un psychologue comparative maintenant à l'Université de Cambridge en Angleterre. **La psychologie est l'étude des comportements et des processus mentaux**, et étude comparative des psychologues comment les animaux autres que les humains se comportent.

https://student.societyforscience.org/article/theres-no-i-elephant

Elephants are social animals. They live with their families, give hugs and call each other by using their trunks as trumpets. They also might know how to help each other.

In a recent elephant study by researchers from the United States and Thailand, pairs of giant animals learned to work together to get some ears of corn. Other animals, especially some primates, are already known to work together to complete tasks, but now elephants have joined the club. Perhaps the finding is not too surprising: Scientists suspect that elephants, with their big brains and survival savvy, may be among the smartest animals on the planet.

Joshua Plotnik, who worked on the study, told Science News that the animals didn't just learn a trick. Instead, the ways the elephants behaved show that they understand how working together brings benefits to everyone involved. Plotnik is a comparative psychologist now at the University of Cambridge in England. Psychology is the study of behaviors and mental processes, and comparative psychologists study how animals other than humans behave.

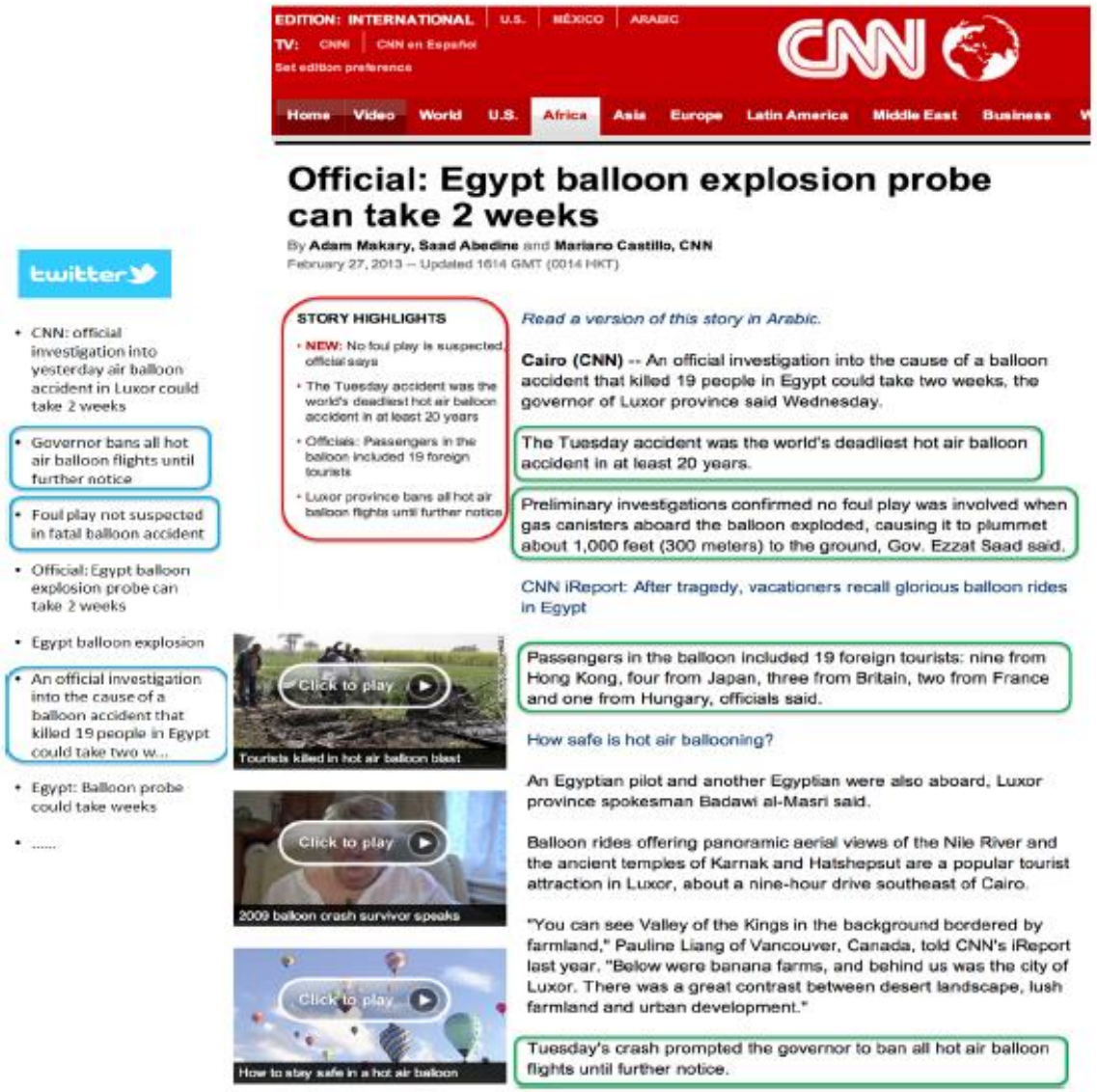
Les éléphants sont des animaux sociaux. Ils **vivent** avec leur famille, **faire** des câlins et **appeler** les uns les autres en utilisant leurs troncs trompettes. Ils pourraient également savoir comment aider les uns les autres.

Dans une étude récente d'éléphants par des chercheurs des États-Unis et la Thaïlande, des paires d'animaux géants ont appris à travailler ensemble pour obtenir des épis de maïs. D'autres animaux, en particulier des primates, sont déjà connus pour travailler ensemble pour accomplir des tâches, mais maintenant, les éléphants ont rejoint le club. Peut-être le résultat n'est pas trop surprenant: Les scientifiques soupçonnent que **les éléphants**, avec leurs gros cerveaux et de bon sens de survie, **peut-être** parmi les plus intelligents des animaux sur la planète.

Joshua Plotnick, qui a travaillé sur l'étude, dit **Nouvelles de la Science** que les animaux n'ont pas seulement appris un truc. Au lieu de cela, les moyens les éléphants se comportent montrent qu'ils comprennent comment travailler ensemble apporte des avantages à toutes **les personnes** impliquées. Plotnik est un psychologue **comparative** maintenant à l'Université de Cambridge en Angleterre. La psychologie est l'étude des comportements et des processus mentaux, et **étude comparative des psychologues** comment les animaux autres que les humains se comportent.

https://student.societyforscience.org/article/theres-no-i-elephant

Single Document Summarization



[Wei and Gao 2014]

Multi Document Summarization

Health Benefits

- Eating a diet rich in vegetables and fruits as part of an overall healthy diet may reduce risk for heart disease, including heart attack and stroke.
- Eating a diet rich in some vegetables and fruits as part of an overall healthy diet may protect against certain types of cancers.
- Diets rich in foods containing fiber, such as some vegetables and fruits, may reduce the risk of heart disease, obesity, and type 2 diabetes.
- Eating vegetables and fruits rich in potassium as part of an overall healthy diet may lower blood pressure, and may also reduce the risk of developing kidney stones and help to decrease bone loss.
- Eating foods such as vegetables that are lower in calories per cup instead of some other higher-calorie food may be useful in helping to lower calorie intake.

Nutrients

- Most vegetables are naturally low in fat and calories. None have cholesterol. (Sauces or seasonings may add fat, calories, or cholesterol.)
- Vegetables are important sources of many nutrients, including potassium, dietary fiber, folate (folic acid), vitamin A, and vitamin C.
- Diets rich in potassium may help to maintain healthy blood pressure. Vegetable sources of potassium include sweet potatoes, white potatoes, white beans, tomato products (paste, sauce, and juice), beet greens, soybeans, lima beans, spinach, lentils, and kidney beans.
- Dietary fiber from vegetables, as part of an overall healthy diet, helps reduce blood cholesterol levels and may lower risk of heart disease. Fiber is important for proper bowel function. It helps reduce constipation and diverticulosis. Fiber-containing foods such as vegetables help provide a feeling of fullness with fewer calories.
- Folate (folic acid) helps the body form red blood cells. Women of childbearing age who may become pregnant should consume adequate folate from foods, and in addition 400 mcg of synthetic folic acid from fortified foods or supplements. This reduces the risk of neural tube defects, spina bifida, and anencephaly during fetal development.
- Vitamin A keeps eyes and skin healthy and helps to protect against infections.
- Vitamin C helps heal cuts and wounds and keeps teeth and gums healthy. Vitamin C aids in iron absorption.

Summary

Eating vegetables is healthy.

Caption Generation

www.cs.toronto.edu/~nitish/nips2014demo/results/98756067.html

Results

Tags

- ballplayers
- gloved
- sweeps
- fencers
- pushups

Nearest Caption in the Training Dataset

a person wearing kendo martial arts armor stands with his hand on his practice sword in a room with other martial arts people .

Generated Captions

- there are two men who appear to be practicing martial arts .
- two men are playing a game of martial arts .
- a man standing in front of men holding a basketball game .
- a man in jeans with a sword in a basketball game .
- two men playing a game of martial arts .

Visual Question Answering

https://visualqa.org

Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (CVPR 2017)

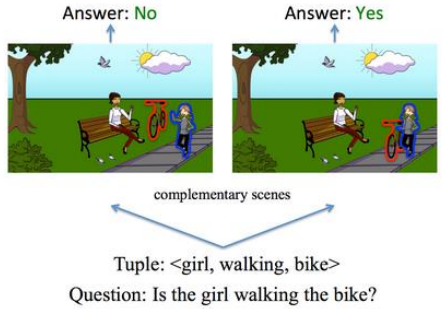
Download the paper

BibTeX

Yin and Yang: Balancing and Answering Binary Visual Questions (CVPR 2016)

Download the paper

BibTeX



VQA: Visual Question Answering (ICCV 2015)


What color are her eyes?
What is the mustache made of?


How many slices of pizza are there?
Is this a vegetarian pizza?

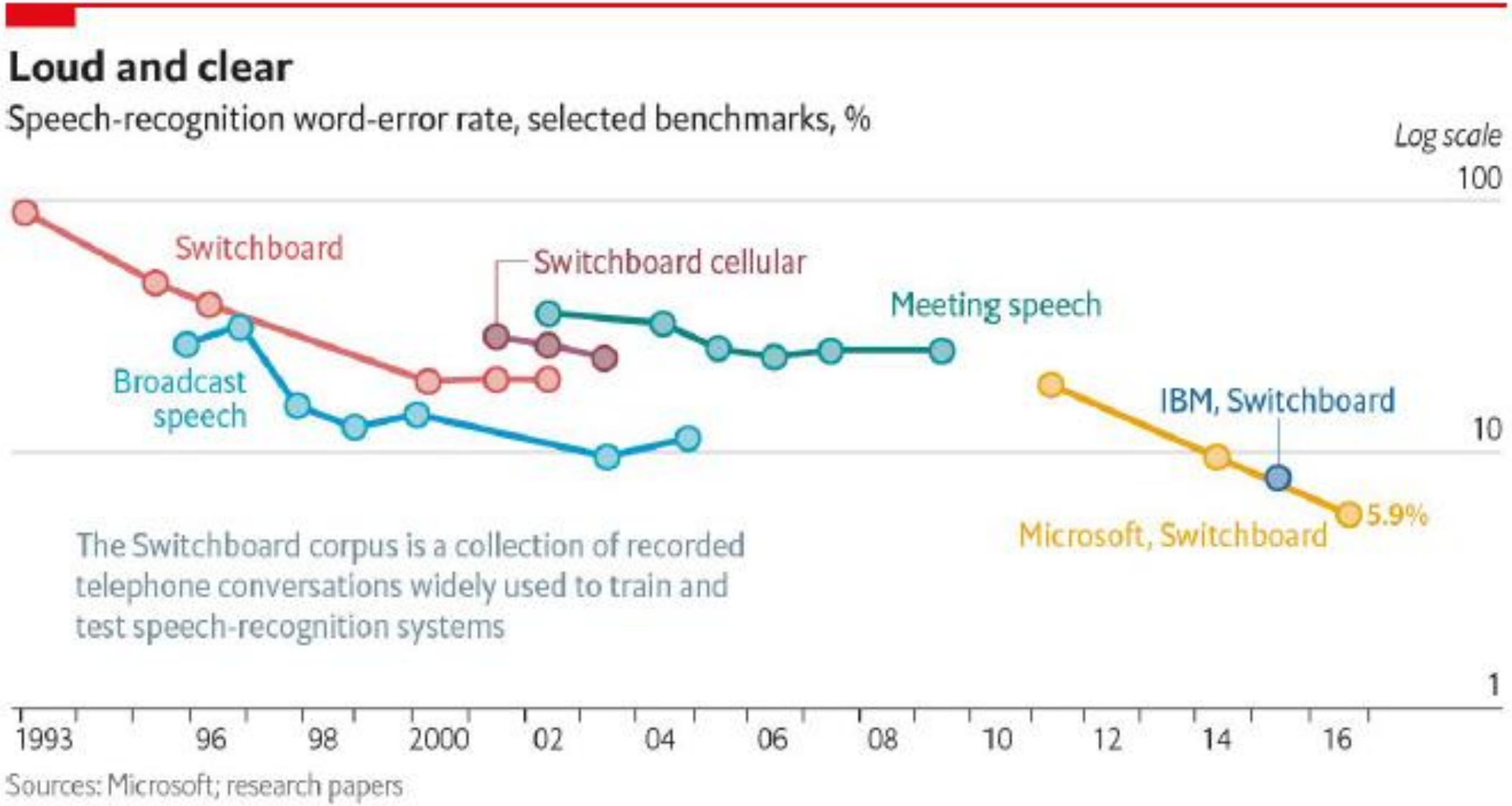

Is this person expecting company?
What is just under the tree?


Does it appear to be rainy?
Does this person have 20/20 vision?

Conversational Agents

https://www.oreilly.com/library/view/iphone-the-missing/9781449372781/ch04.html

Speech Recognition



Entailment and Paraphrasing

ID	TEXT	HYPOTHESIS	TASK	VALUE
1586	*The Republic of Yemen is an Arab, Islamic and independent sovereign state whose integrity is inviolable, and no part of which may be ceded.*	*The national language of Yemen is Arabic.*	QA	True
1076	*Most Americans are familiar with the Food Guide Pyramid– but a lot of people don't understand how to use it and the government claims that the proof is that two out of three Americans are fat.*	*Two out of three Americans are fat.*	RC	True
1667	*Regan attended a ceremony in Washington to commemorate the landings in Normandy.*	*Washington is located in Normandy.*	IE	False
2016	*Google files for its long awaited IPO.*	*Google goes public.*	IR	True
2097	*The economy created 228,000 new jobs after a disappointing 112,000 in June.*	*The economy created 228,000 jobs after dissapointing the 112,000 of June.*	MT	False
893	*The first settlements on the site of Jakarta were established at the mouth of the Ciliwung, perhaps as early as the 5th century AD.*	*The first settlements on the site of Jakarta were established as early as the 5th century AD.*	CD	True
1960	*Bush returned to the White House late Saturday while his running mate was off campaigning in the West.*	*Bush left the White House.*	PP	False
586	*The two suspects belong to the 30th Street gang, which became embroiled in one of the most notorious recent crimes in Mexico: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.*	*Cardinal Juan Jesus Posadas Ocampo died in 1993.*	QA	True

Table 1. Examples of Text-Hypothesis pairs

Ido Dagan, Oren Glickman and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge

Discourse Analysis

- Anaphoric relations:

*1. Mary helped Peter get out of the car. **He** thanked **her**.*
2. Mary helped the other passenger out of the car.
*The man had asked **her** for help because of **his** foot injury.*

Tom appeared on the sidewalk with a bucket of whitewash and a long-handled brush. He surveyed the fence, and all gladness left him and a deep melancholy settled down upon his spirit. (Tom Sawyer)

Dialogue Systems

- I would like to make a reservation at Sorrento.
- For when?
- 8 pm Friday night.
- We only have availability for 7 pm and 10 pm.
- Sorry, these times don't work for me.

NLP

Introduction to NLP

153

Preprocessing

Text Preprocessing

- Removing non-text
  - ads, javascript
- Dealing with text encoding
  - e.g., Unicode
- Sentence segmentation
- Normalization
  - labeled/labelled, extra-terrestrial/extraterrestrial, extra terrestrial
- Stemming
  - computer/computation
- Morphological similarity
  - car/cars
- Capitalization
  - Now/NOW, led/LED
- Phrase and named entity extraction, foreign words
  - USA/usa, MIT/mit

Text Preprocessing

- Types vs. Tokens
  - To be or not to be
- Tokenization:
  - ALS vs. A.L.S.
  - Paul’s, Willow Dr., Dr. Willow, New York, ad hoc, can’t
  - “The New York-Los Angeles flight” vs. “Minneapolis-St.Paul”
  - Numbers, e.g., (888) 555-1313, 1-888-555-1313
  - Dates, e.g., Jan-13-2012, 20120113, 13 January 2012, 01/13/12
  - URLs

Word segmentation into morphemes

- Arabic:

  كتاب

- Japanese:

  この本は重い。

  (kono hon ha omoi)

- German:

  Finanzdienstleistung = financial services

- Chinese:

  电视 (television)

  电 (diàn = electric) 视 (shì = to look at)

Text preprocessing

ニューヨーク (New York) は、アメリカ合衆国ニューヨーク州にある都市

- Kanji, Katakana, Hiragana, Rōmaji, (numbers)
- Nyūyōku wa, Amerikagasshūkoku nyūyōku-shū ni aru toshi

Sentence segmentation into words

- 金属製品製造の日立金属は１９日、世界最大手の鉄鋳物メーカー「ワウパカ　ファウンドリー　ホールディングス」（米国・デラウェア州）を米投資ファンドから買収し、完全子会社にすると発表した。買収額は１３億ドル（約１３３０億円）で、１０月中にも手続きを終える。

Sentence boundary recognition

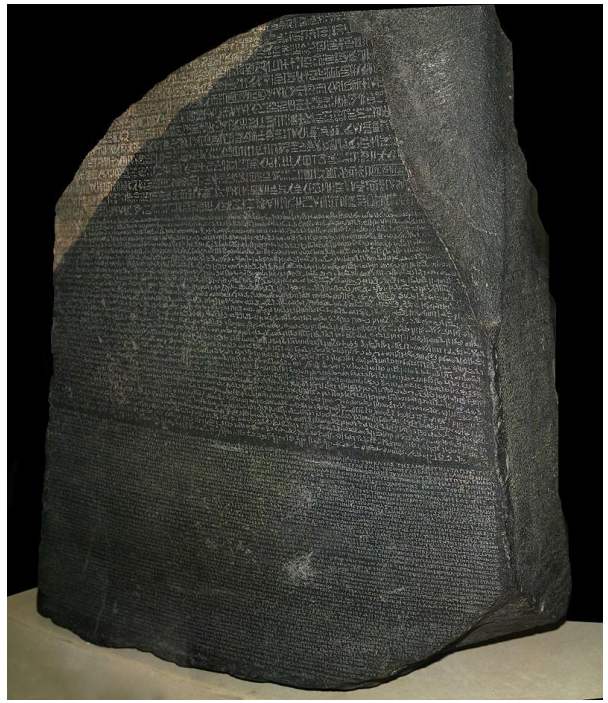
- Classification problem
- Features
  - punctuation
  - formatting
  - fonts
  - spacing
  - capitalization
  - case
  - use of abbreviations, e.g., Dr., a.m.
- Example
  - If there is no space after a period, don’t assume that there is a sentence boundary

NLP

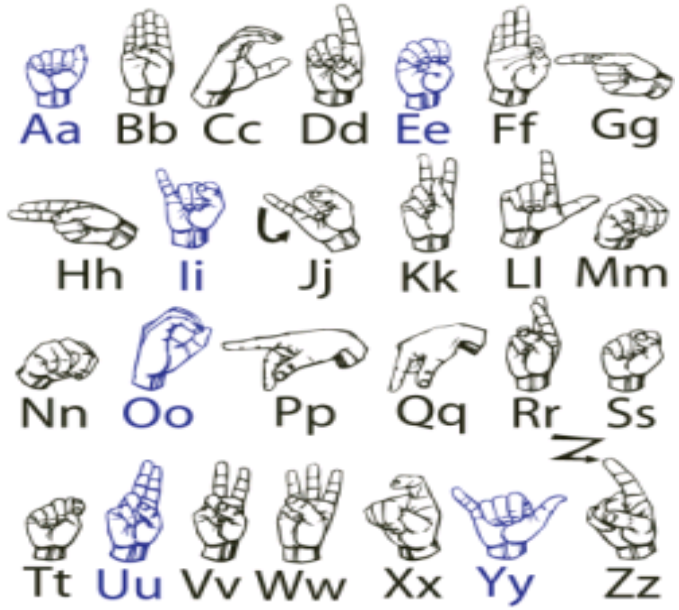
Introduction to NLP

141

A Brief Overview of Languages and Linguistics

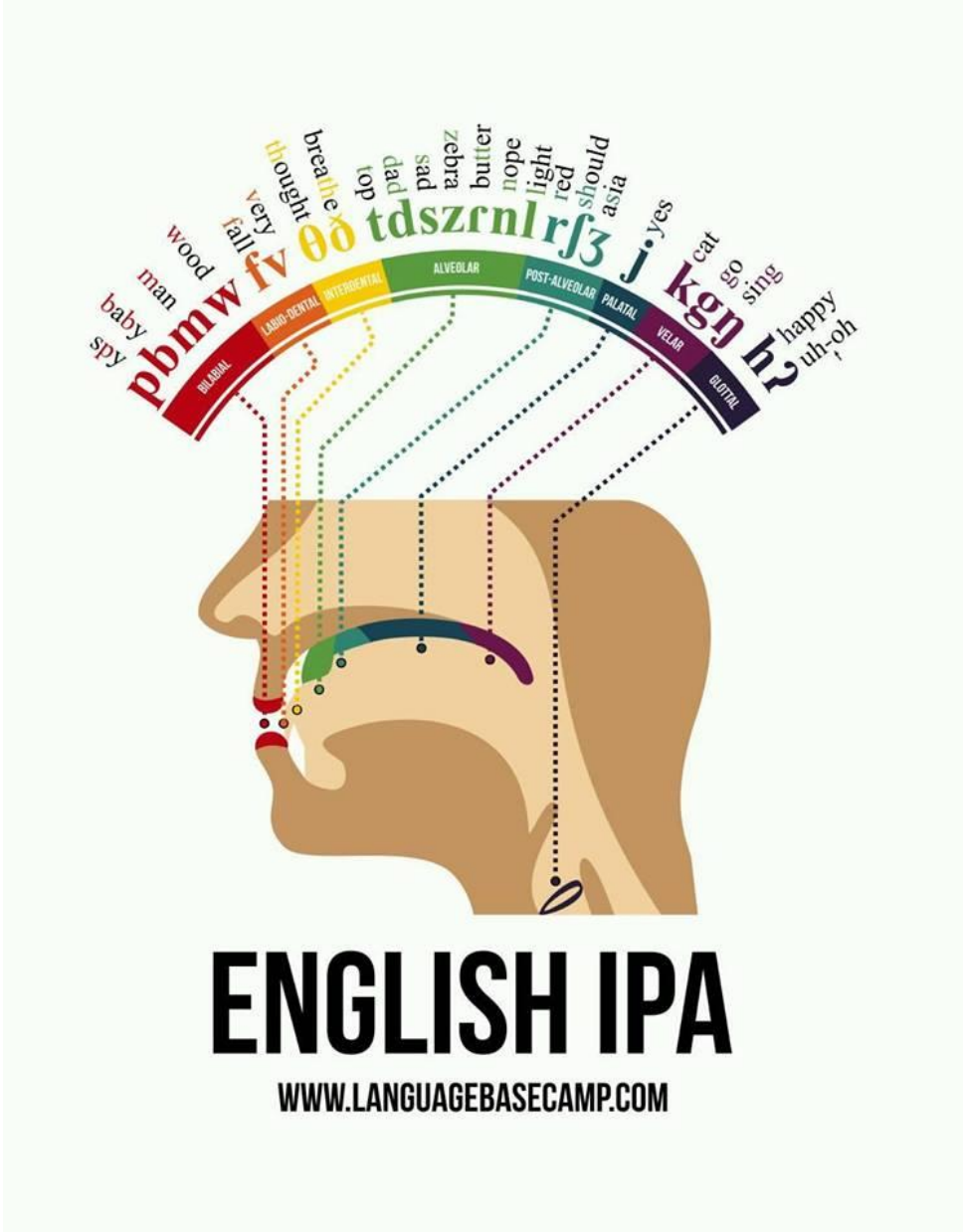


خ	ح	ج	ث	ت	ب	ا
kha	haa	jiim	thaa	taa	baa	alif
ص	ش	س	ز	ر	ذ	د
saad	shiin	siin	zaay	raa	thaal	daal
ق	ف	غ	ع	ظ	ط	ض
qaaf	faa	ghayn	ayn	thaa	taa	daad
ي	و	ه	ن	م	ل	ك
yaa	waaw	ha	nuun	miim	laam	kaaf



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



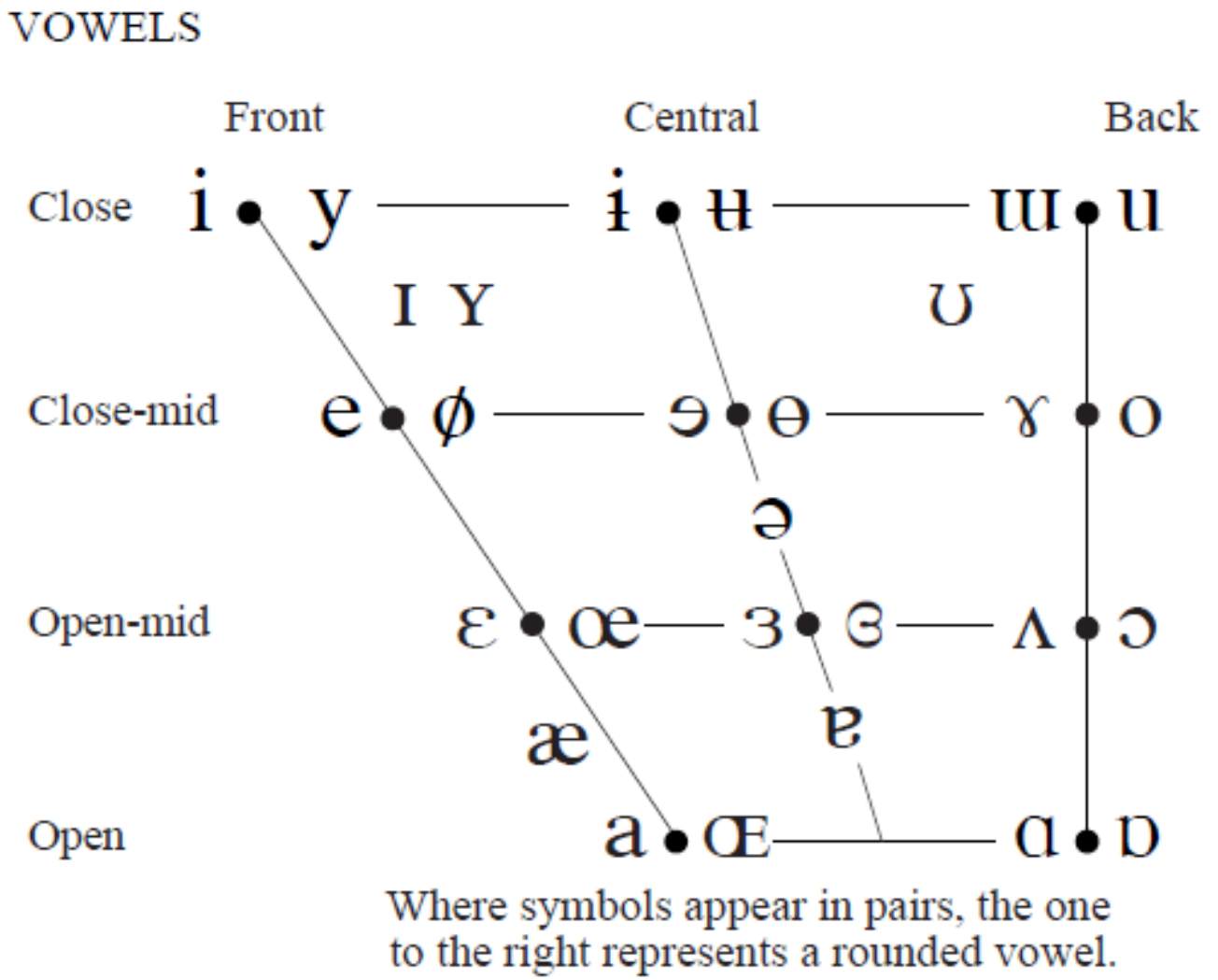
IPA Chart (consonants)

CONSONANTS (PULMONIC)

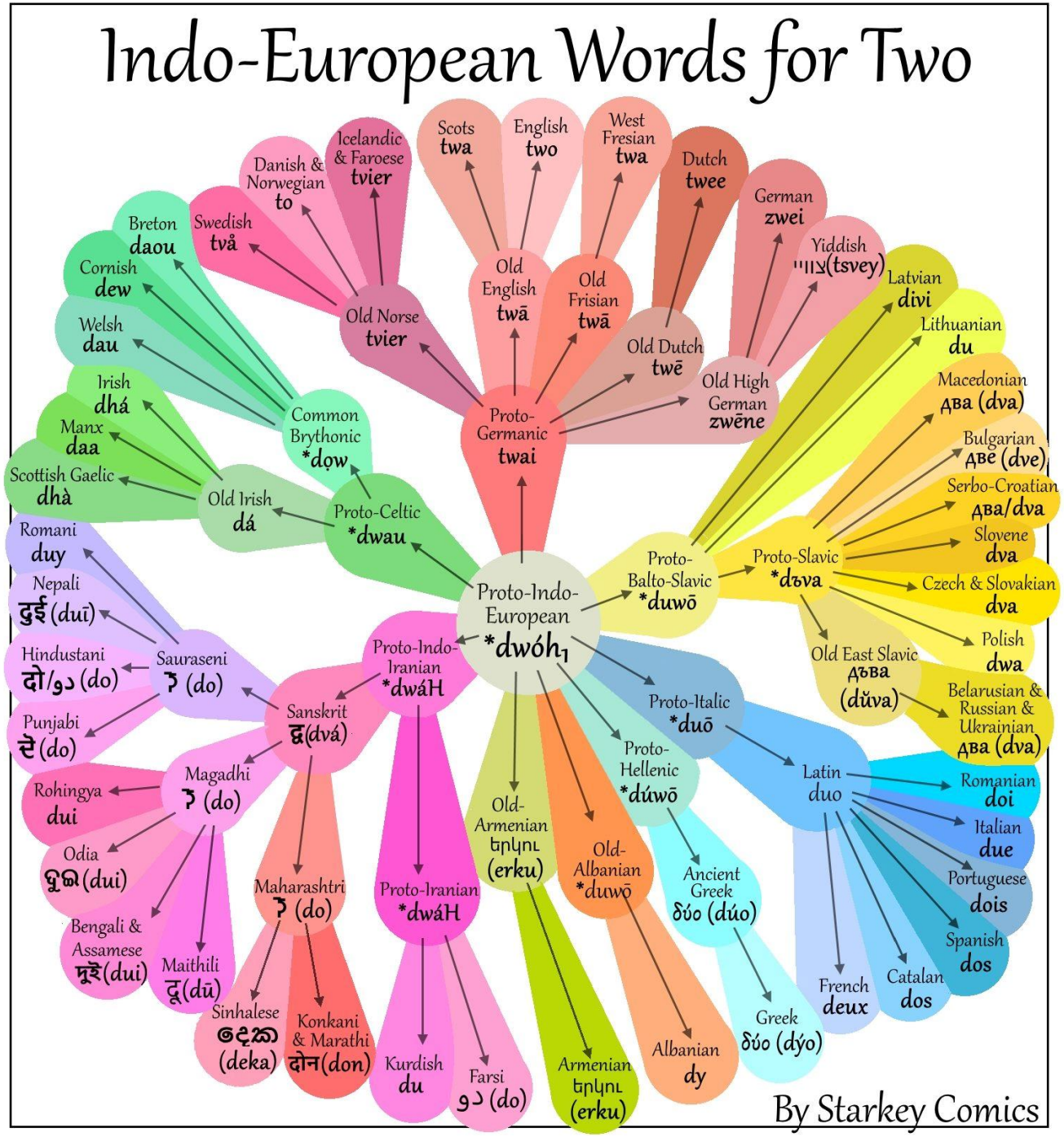
	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N		
Trill	B			r					R		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant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

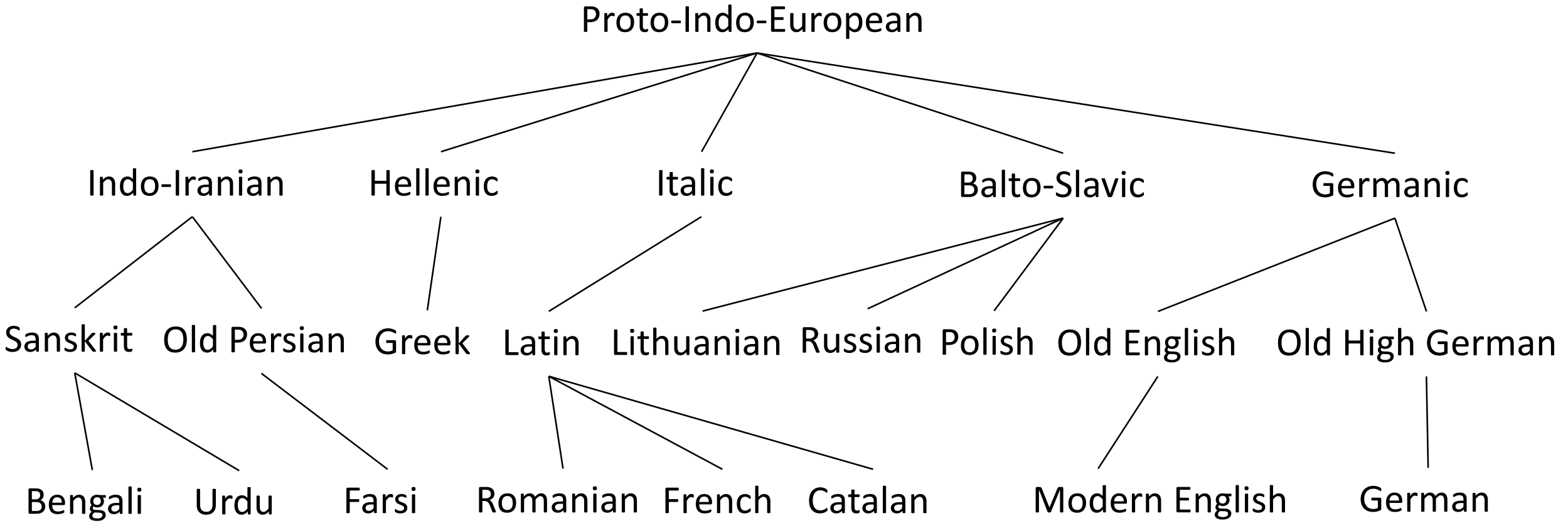
IPA Chart (vowels)



By IPA (http://www.langsci.ucl.ac.uk/ipa/ipachart.html) [CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons


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

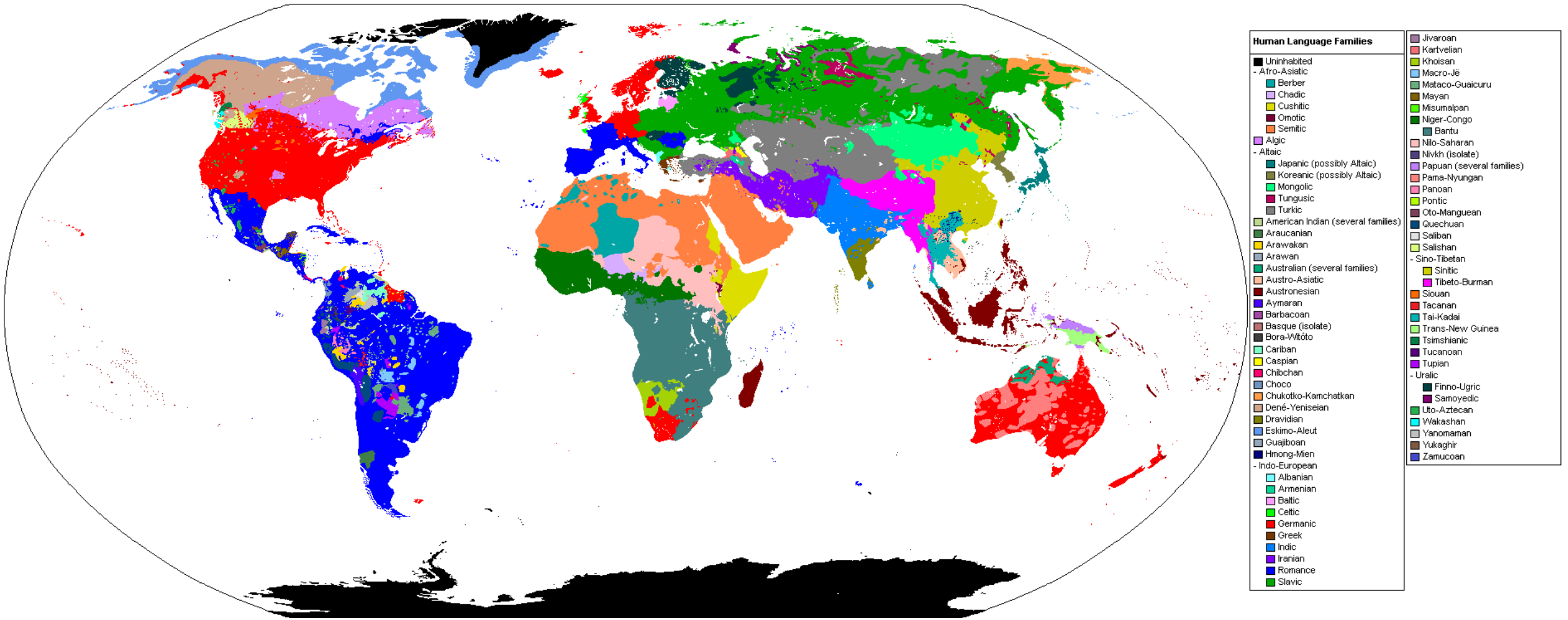
Some Indo-European languages



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

NLP

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, amimígo = 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"?

Answer

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

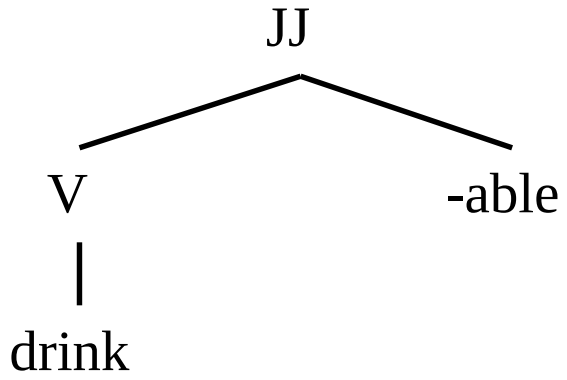
  edu-*freakin’*-cation

  * 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)

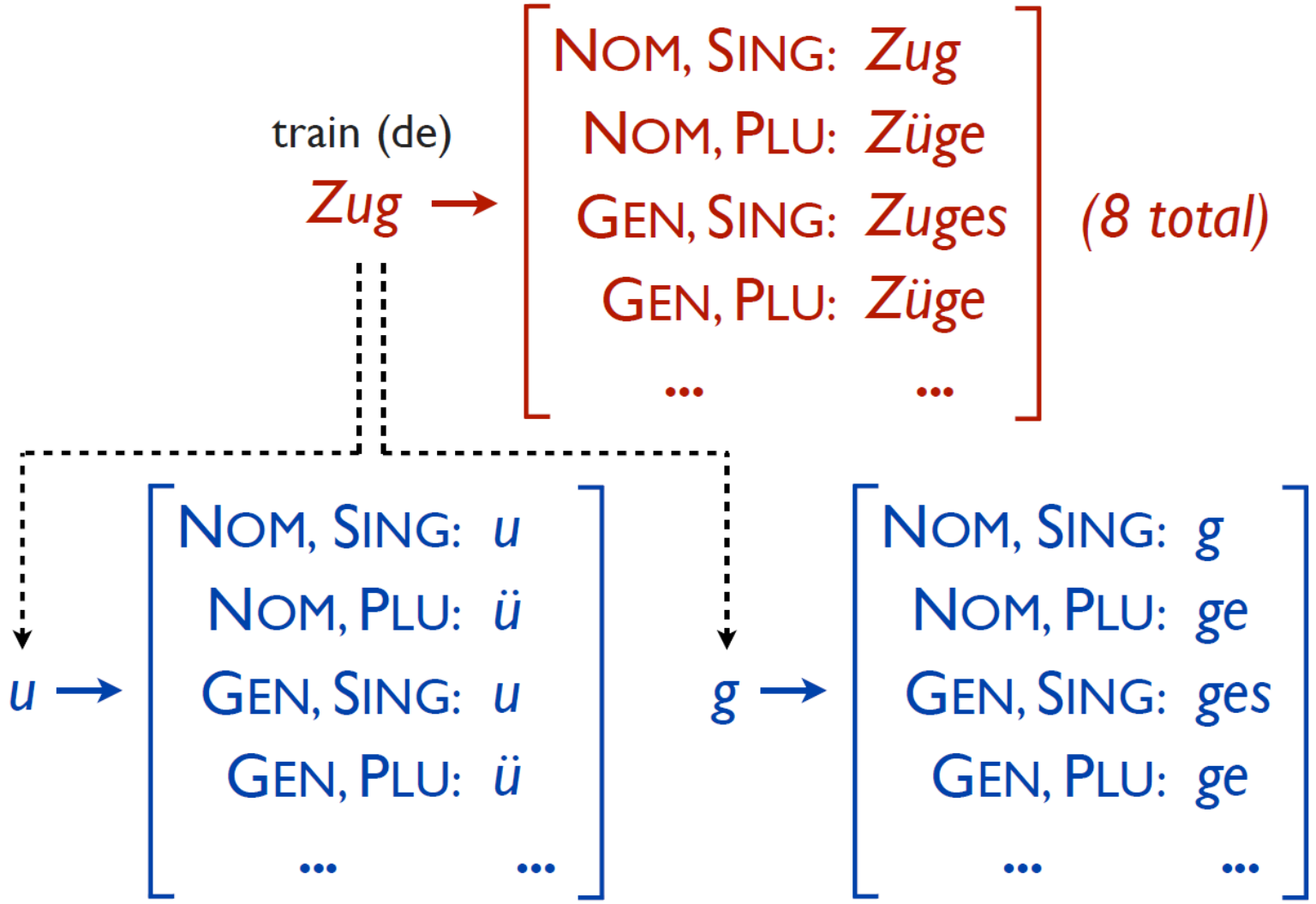
Noun and Pronoun Declension in Latin

	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				

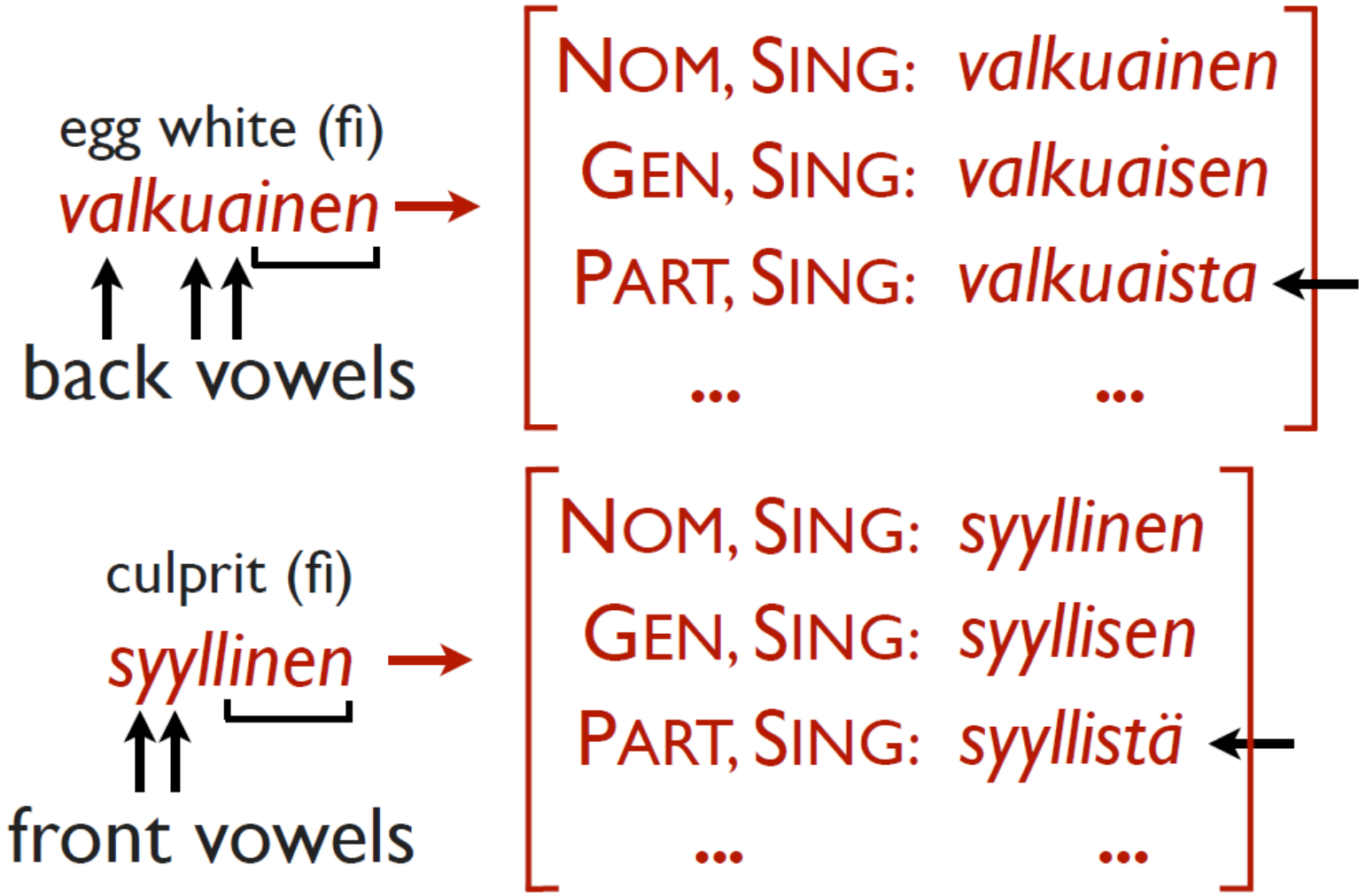
https://en.wikipedia.org/wiki/Latin_declension

Automatic Inflection



Durrett and Denero 2013]

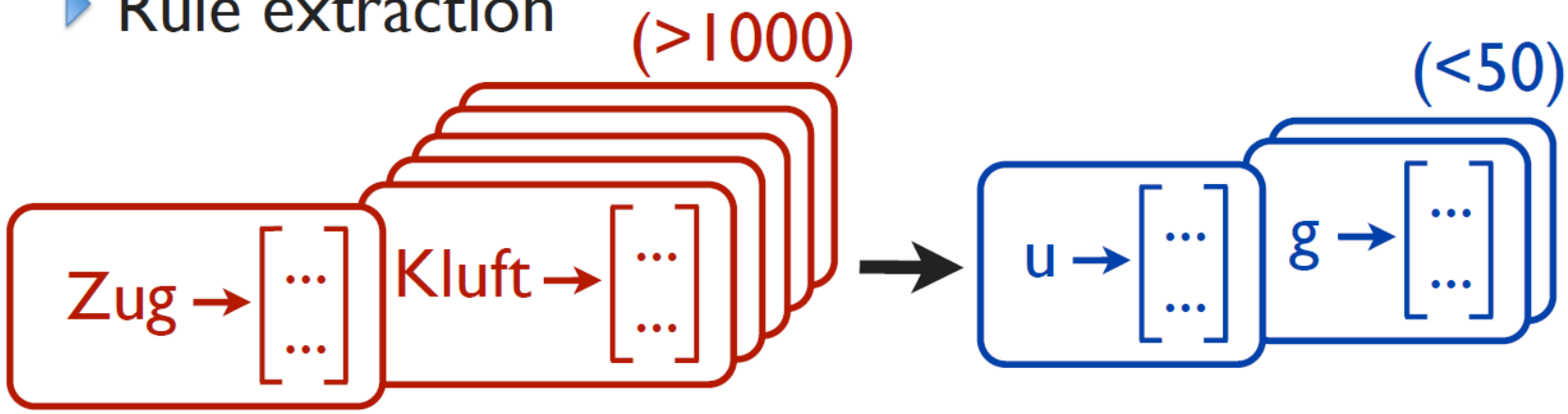
Automatic Inflection



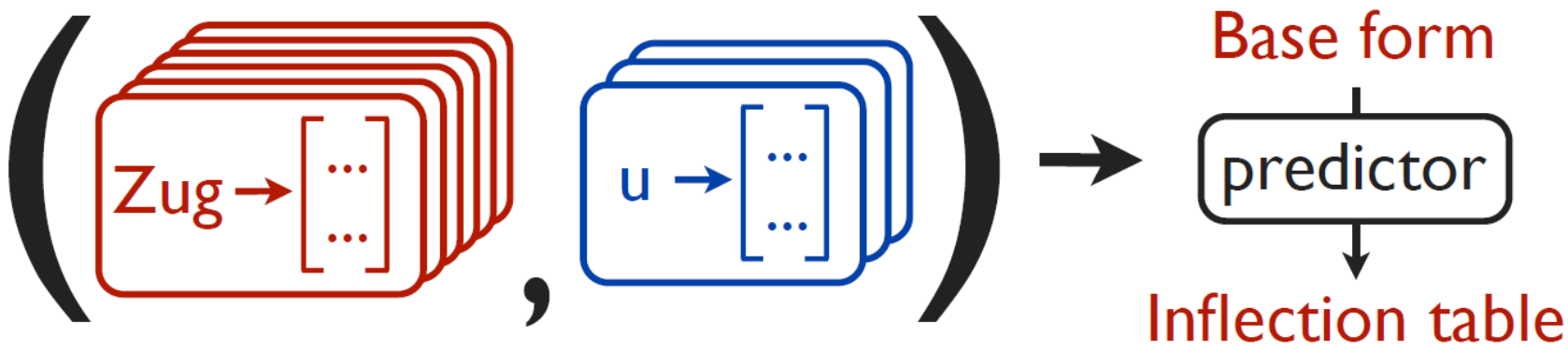
Durrett and Denero 2013]

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

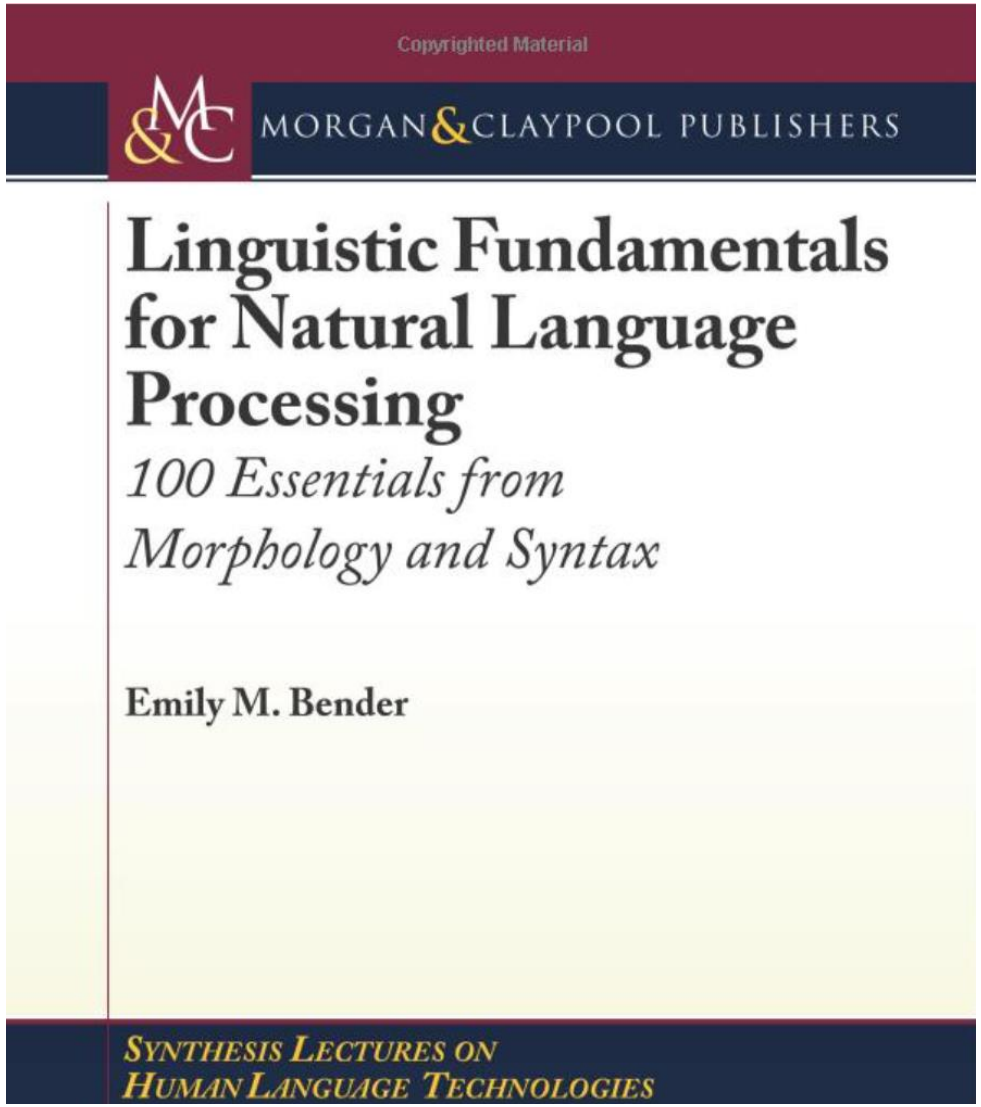
Example from Kemal Oflazer

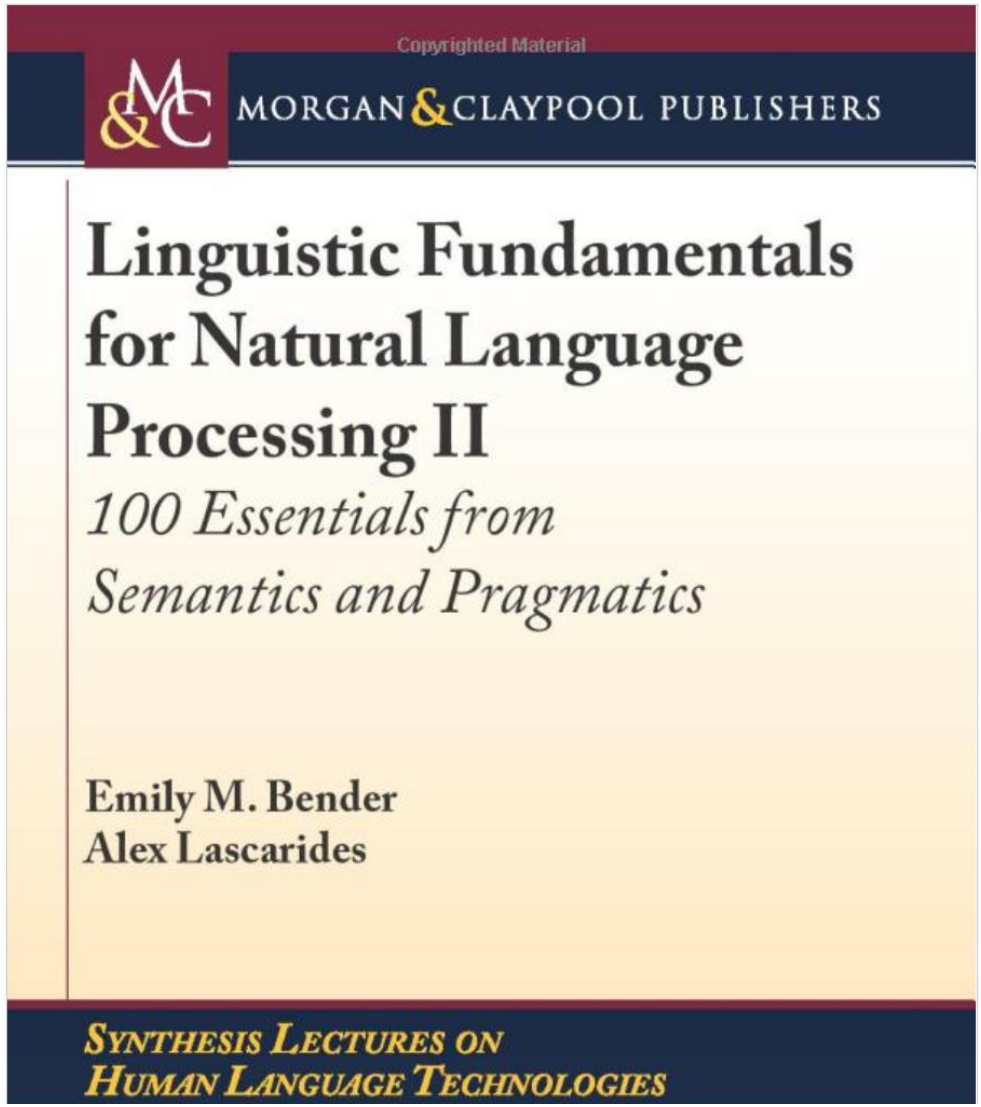
Turkish Vowel Harmony

	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





NLP

Introduction to NLP

144

Word Distributions

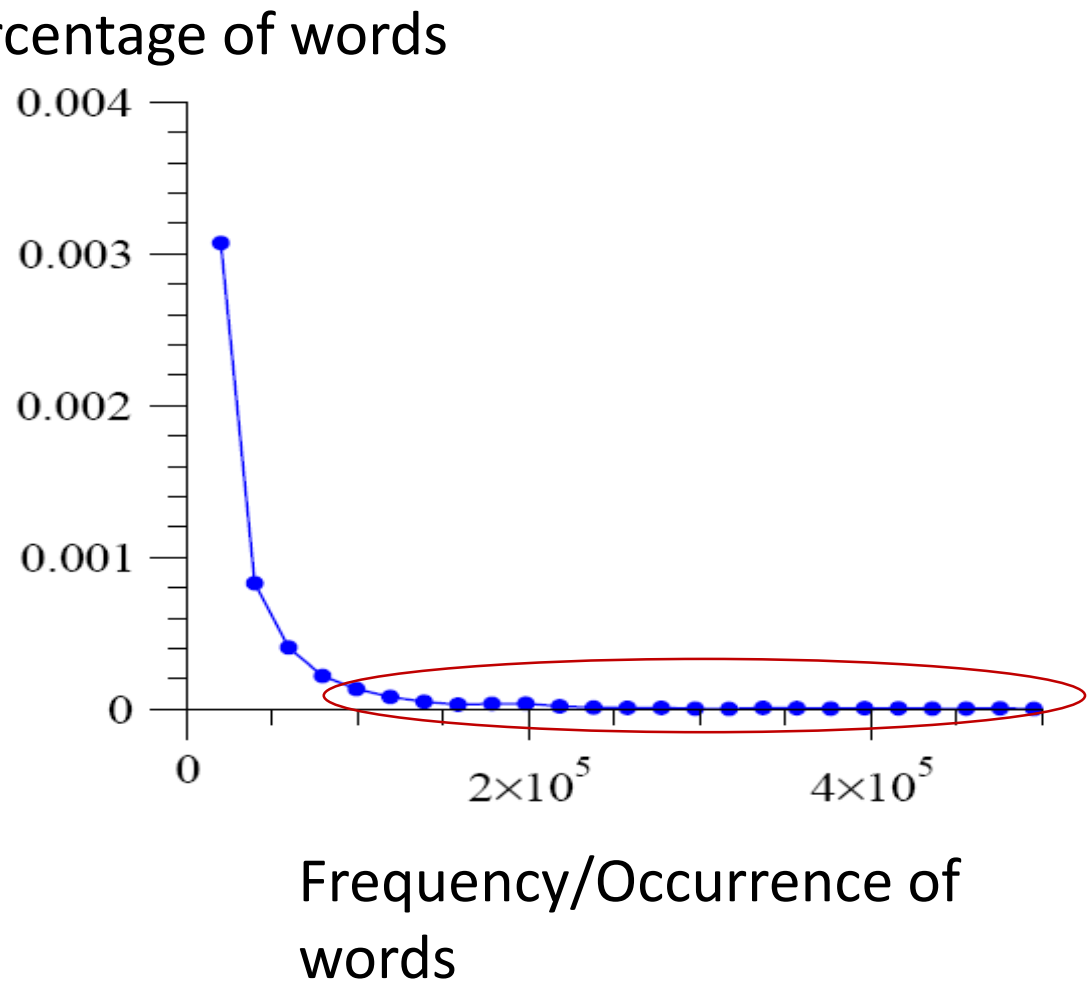
Word Distributions

- 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

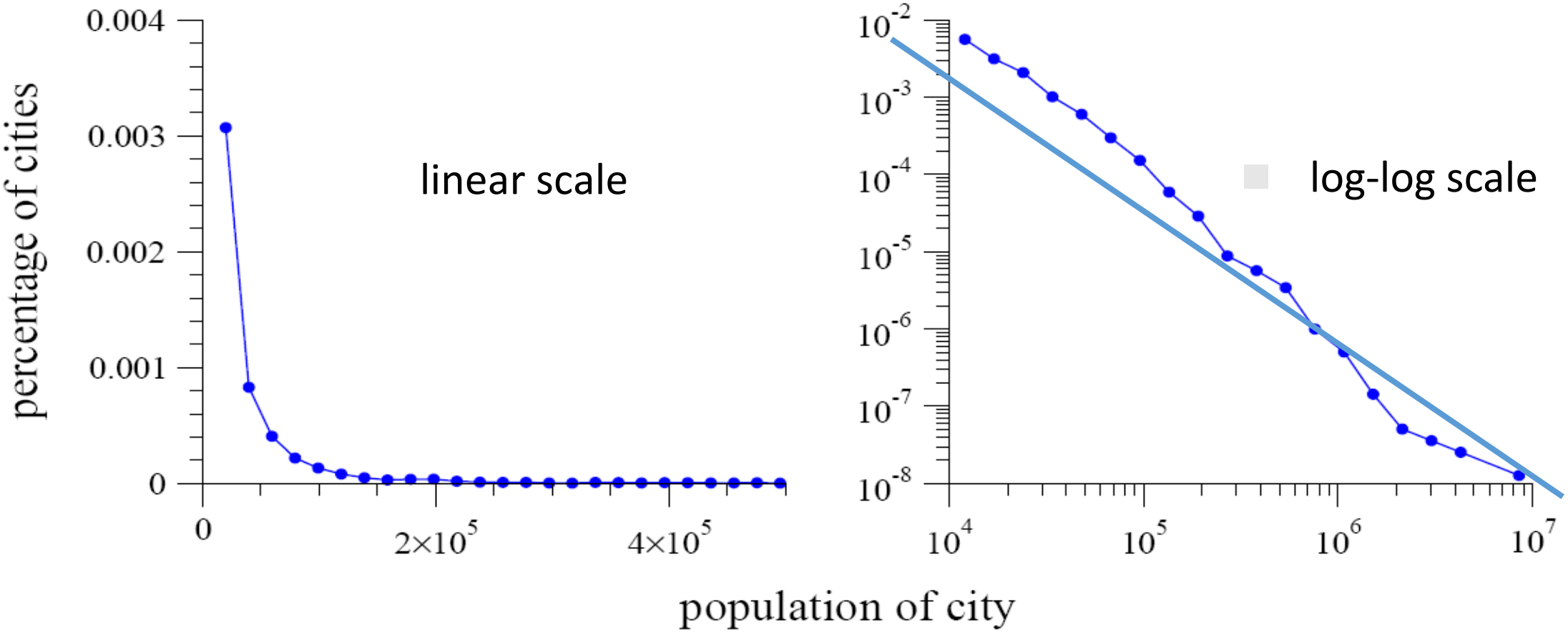
Power-law Distribution



- 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)

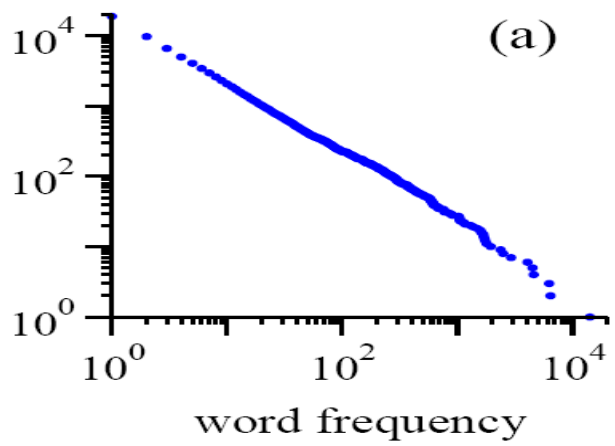
α : scaling exponent, or power law exponent

- Straight line on a log-log plot

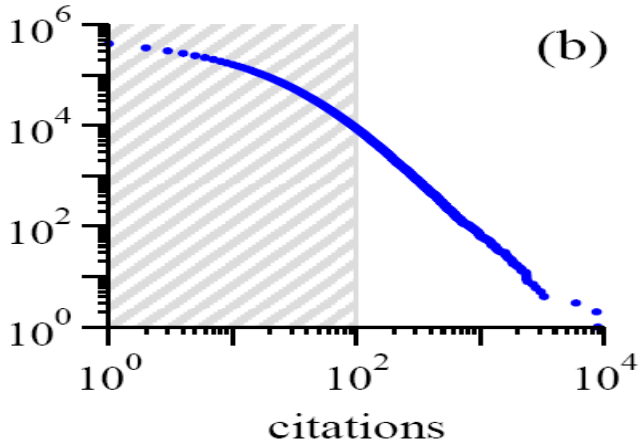
$$\ln\big(p(x)\big) = c - \alpha\, ln(x)$$

Power Laws Are Seemingly Everywhere

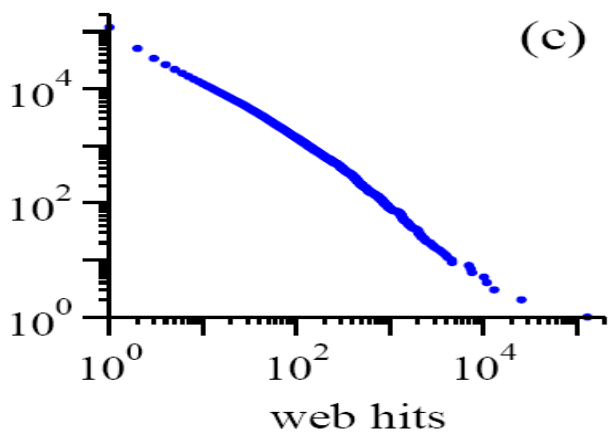
note: these are cumulative distributions



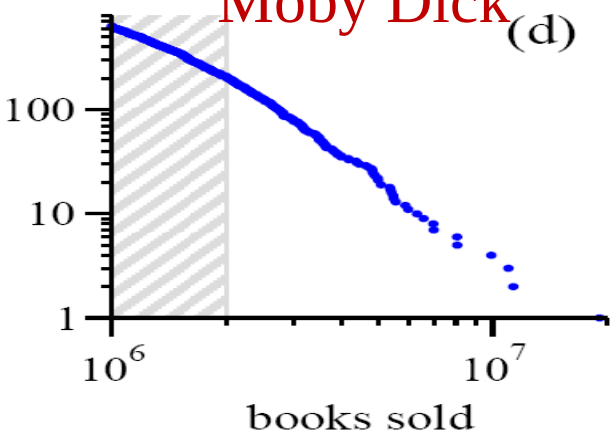
Moby Dick



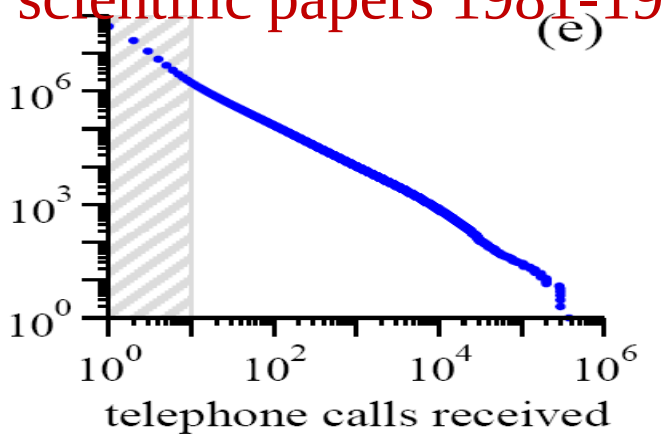
scientific papers 1981-1997



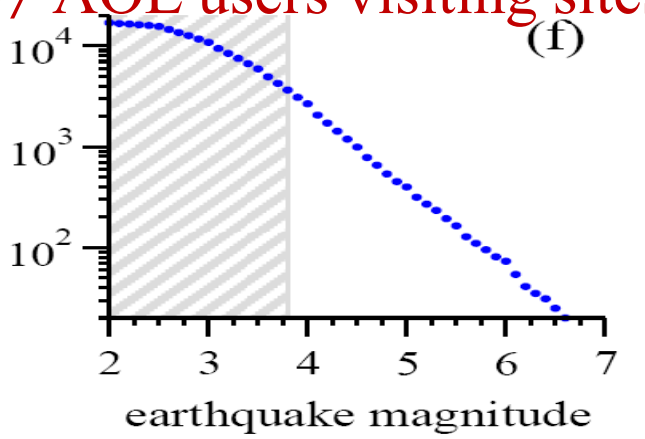
AOL users visiting sites ‘97



bestsellers 1895-1965



AT&T customers on 1 day

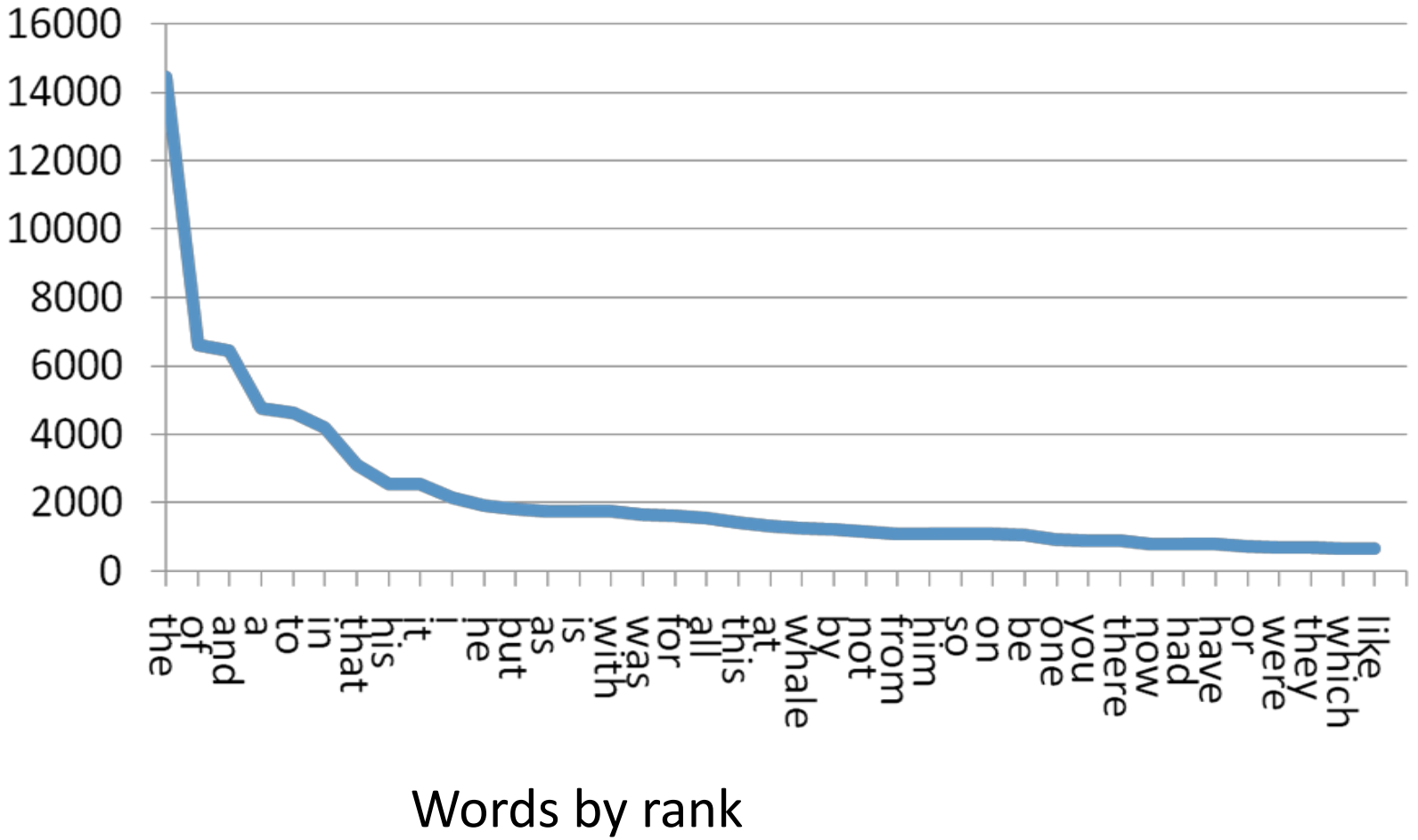


California 1910-1992

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

Rank x Frequency ≈ 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

NLP

Text Similarity

313

Spelling Similarity:

Edit Distance

Spelling Similarity

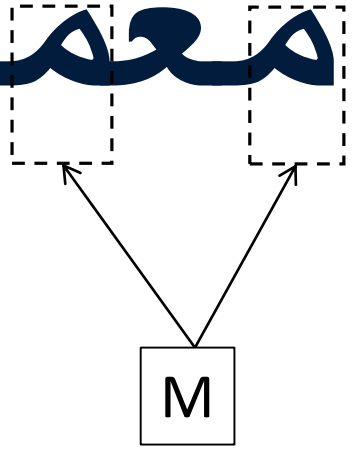
- Typos:
  - Brittany Spears -> Britney Spears
  - Catherine Hepburn -> Katharine Hepburn
  - Reciept -> receipt
- Variants in spelling:
  - Theater -> theatre

Who is this?

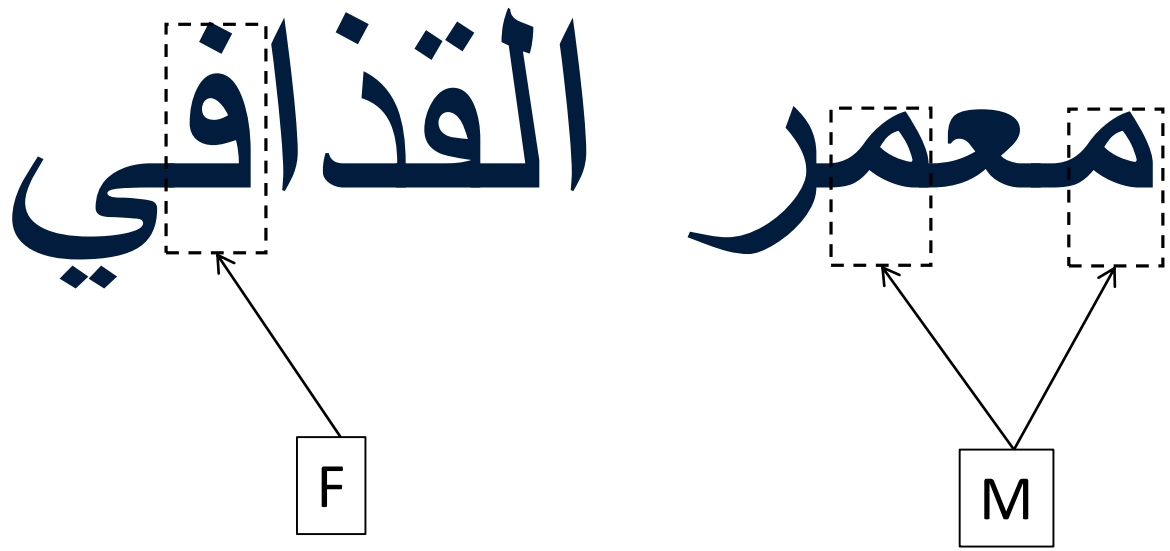
معمر القذافي

Hints

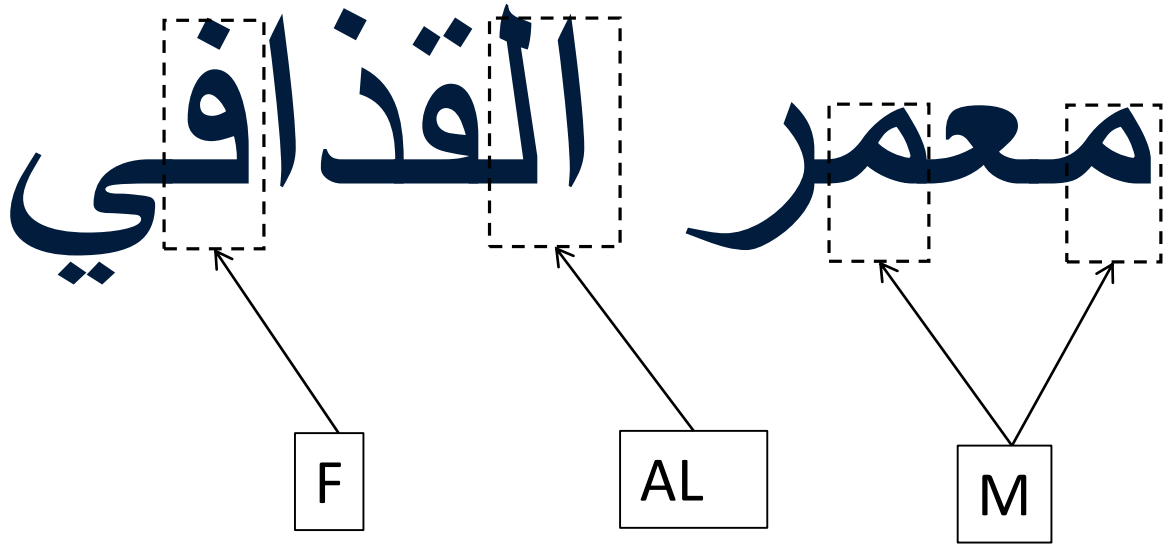
معمر القذافي



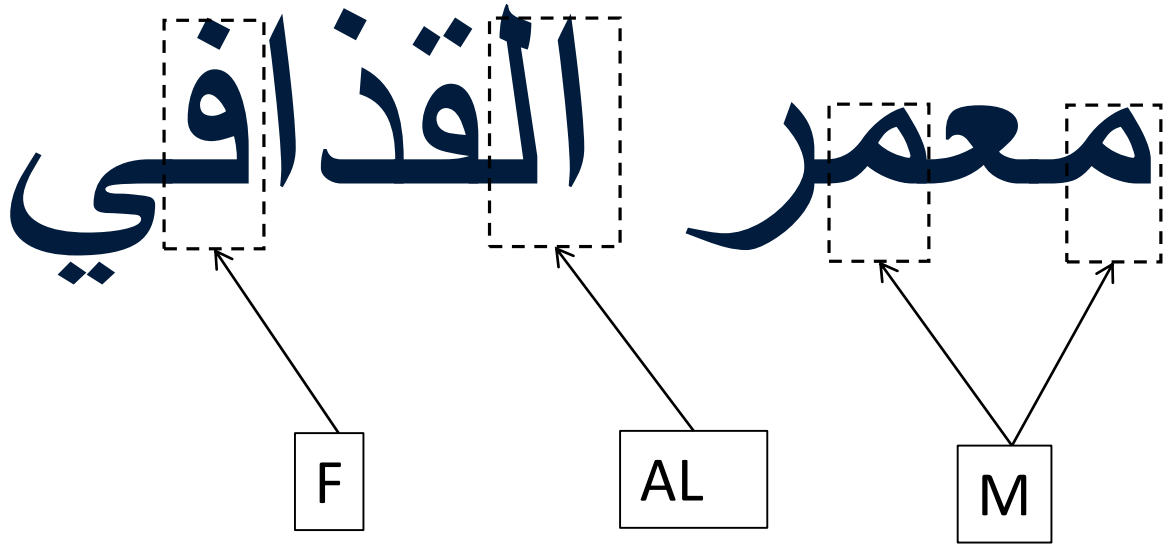
Hints



Hints



Hints



Muammar (al-)Gaddafi, or **Moamar Khadafi,** or …

Quiz

How many different transliterations can there be?

m u o a m mm a e r	el al El Al ø	Q G Gh K Kh a e u d dh ddh dhdh th zz a f ff i y

A lot!

m u o a m mm a e r		el al El Al ø		Q G Gh K Kh a e u d dh ddh dhdh th zz a f ff i y		
8	x	5	x	360	=	14,400

Edit Operations

- Insertion/deletion
  - behaviour - behavior
- Substitution
  - string - spring
- Multiple edits
  - sleep - slept

Levenshtein Method

- Based on dynamic programming
  - Insertions, deletions, and substitutions usually all have a cost of 1.

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	5								

Recurrence relation

- Definitions
  - $s_1(i)$ – i^{th} character in string s_1
  - $s_2(j)$ – j^{th} character in string s_2
  - $D(i,j)$ – edit distance between a prefix of s_1 of length i and a prefix of s_2 of length j
  - $t(i,j)$ – cost of aligning the i^{th} character in string s_1 with the j^{th} character in string s_2

- Recursive dependencies

$$D(i,0)=i$$
$$D(0,j)=j$$
$$D(i,j)=\min[$$
$$\quad D(i-1,j)+1$$
$$\quad D(i,j-1)+1$$
$$\quad D(i-1,j-1)+t(i,j)$$
$$\quad]$$

- Simple edit distance:

$$t(i,j)=0 \ \textit{iff} \ s_1(i)=s_2(j)$$
$$t(i,j)=1, \textit{otherwise}$$

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1							
r	2								
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1						
r	2								
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2						
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2						
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	4

Edit Transcript

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	4

Other Costs

- Damerau modification
  - Swaps of two adjacent characters also have a cost of 1
  - E.g., Lev(“cats”,”cast”) = 2, Dam(“cats”,”cast”) = 1

Quiz

- Some distance functions can be more specialized.
- Why do you think that the edit distances for these pairs are as follows?
  - Dist (“sit clown”,“sit down”) = 1
  - Dist (“qeather”,”weather”) = 1, *but* Dist (“leather”,”weather”) = 2

Quiz Answers

- Dist(“sit down”,”sit clown”) is lower in this example because we want to model the type of errors common with optical character recognition (OCR)
- Dist(“qeather”,”weather”) < Dist(“leather”,”weather”) because we want to model spelling errors introduced by “fat fingers” (clicking on an adjacent key on the keyboard)

Quiz: Guess the Language

AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAAACT
TTCAACAATGGATCTCTTGGTTCCGGC

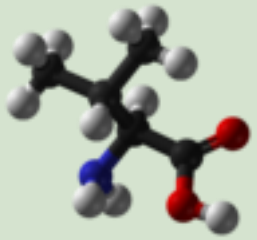
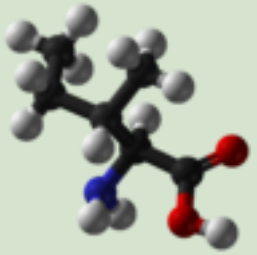
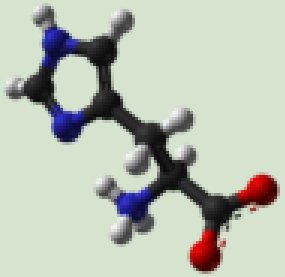
Quiz Answer

- This is a genetic sequence (nucleotides AGCT)

>U03518 Aspergillus awamori internal transcribed spacer 1 (ITS1)
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAAACT
TTCAACAATGGATCTCTTGGTTCCGGC

Other uses of Edit Distance

- In biology, similar methods are used for aligning non-textual sequences
  - Nucleotide sequences, e.g., GTTCGTGATGGAGCG, where A=adenine, C=cytosine, G=guanine, T=thymine, U=uracil, "-"=gap of any length, N=any one of ACGTU, etc.
  - Amino acid sequences, e.g., FMELSEDGIEMAGSTGVI, where A=alanine, C=cystine, D=aspartate, E=glutamate, F=phenylalanine, Q=glutamine, Z=either glutamate or glutamine, X="any", etc. The costs of alignment are determined empirically and reflect evolutionary divergence between protein sequences. For example, aligning V (valine) and I (isoleucine) is lower-cost than aligning V and H (histidine).

Valine	Isoleucine	Histidine
		

External URLs

- Levenshtein demo
  - http://www.let.rug.nl/~kleiweg/lev/
- Biological sequence alignment
  - http://www.bioinformatics.org/sms2/pairwise_align_dna.html
  - http://www.sequence-alignment.com/sequence-alignment-software.html
  - http://www.ebi.ac.uk/Tools/msa/clustalw2/
  - http://www.animalgenome.org/bioinfo/resources/manuals/seqformats

NLP