SO... COMPUTERS HAVE MASTERED PLAYING CHESS AND DRIVING CARS ACROSS THE DESERT, BUT CAN'T HOLD FIVE MINUTES OF NORMAL CONVERSATION?
PRETTY MUCH.

Source: xkcd

CS 671 NATURAL LANGUAGE PROCESSING

J. Laskar

amitabha mukerjee
iit kanpur

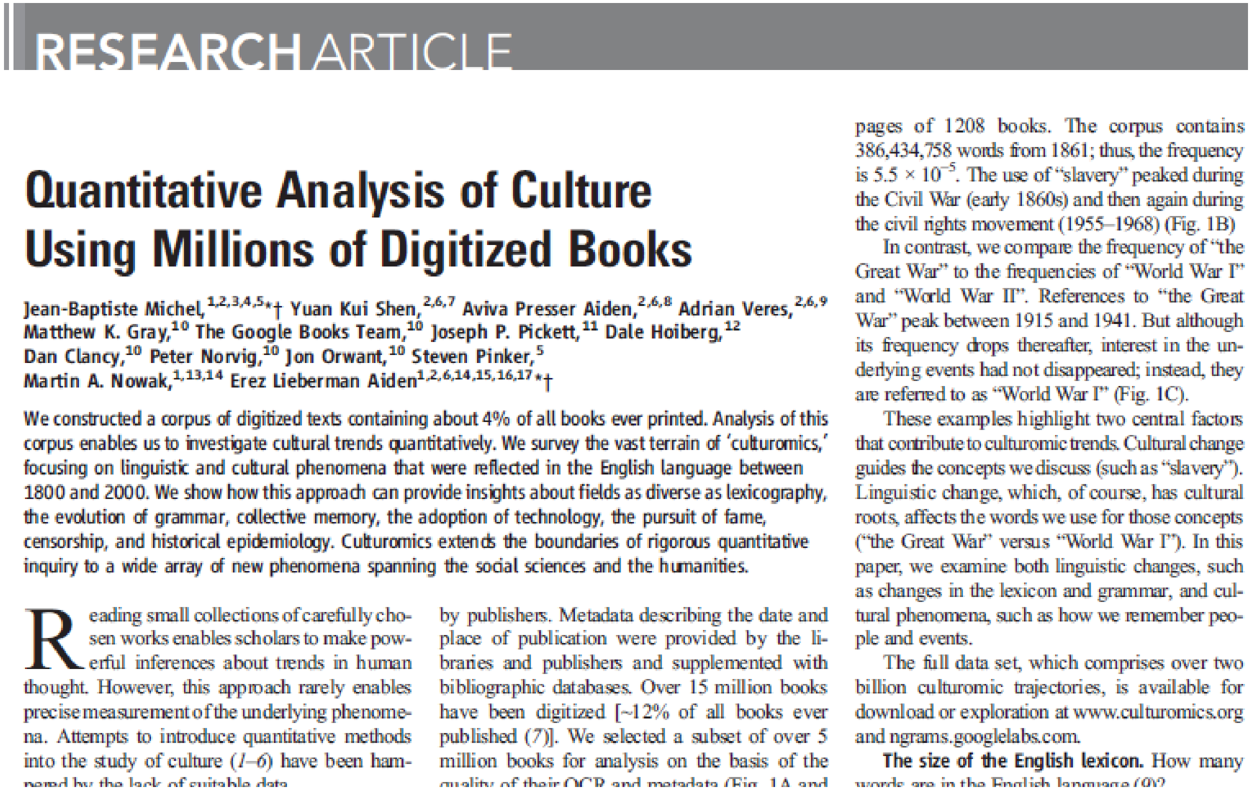
Learning Objectives

- NLP applications are expanding
    - Unstructure data >> Structured
  - Levels of Computational Models
    - Sound units (Phonemes / **Syllables**)
    - Words (Lexical Units)
    - Syntax (Morphology / Grammar)
    - Meaning (Semantics)

- Machine Learning

Culturomics

- □ Michel, Shen, Aiden,... Norvig, etal [Google/ Harvard]
- □ Science, 2011
- □ Quantitative analysis of culture using millions of digitized books



RESEARCH ARTICLE

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5]*† Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[5] Martin A. Nowak,[1,13,14] Erez Lieberman Aiden[1,2,6,14,15,16,17]*†

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

Reading small collections of carefully chosen works enables scholars to make powerful inferences about trends in human thought. However, this approach rarely enables precise measurement of the underlying phenomena. Attempts to introduce quantitative methods into the study of culture (*1–6*) have been hampered by the lack of suitable data by publishers. Metadata describing the date and place of publication were provided by the libraries and publishers and supplemented with bibliographic databases. Over 15 million books have been digitized [~12% of all books ever published (*7*)]. We selected a subset of over 5 million books for analysis on the basis of the quality of their OCR and metadata (Fig. 1A and

pages of 1208 books. The corpus contains 386,434,758 words from 1861; thus, the frequency is 5.5×10^{-5}. The use of "slavery" peaked during the Civil War (early 1860s) and then again during the civil rights movement (1955–1968) (Fig. 1B)

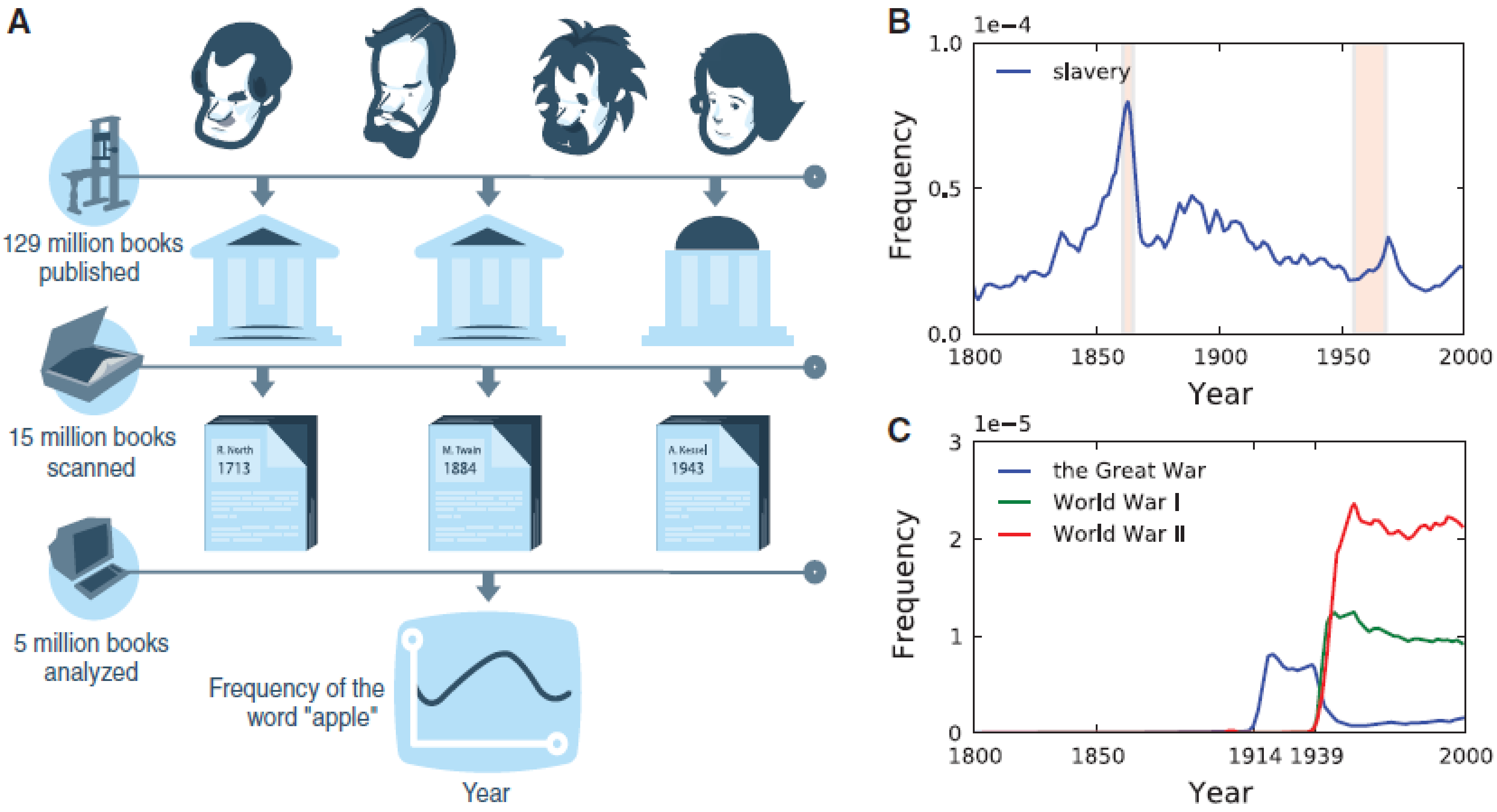
In contrast, we compare the frequency of "the Great War" to the frequencies of "World War I" and "World War II". References to "the Great War" peak between 1915 and 1941. But although its frequency drops thereafter, interest in the underlying events had not disappeared; instead, they are referred to as "World War I" (Fig. 1C).

These examples highlight two central factors that contribute to culturomic trends. Cultural change guides the concepts we discuss (such as "slavery"). Linguistic change, which, of course, has cultural roots, affects the words we use for those concepts ("the Great War" versus "World War I"). In this paper, we examine both linguistic changes, such as changes in the lexicon and grammar, and cultural phenomena, such as how we remember people and events.

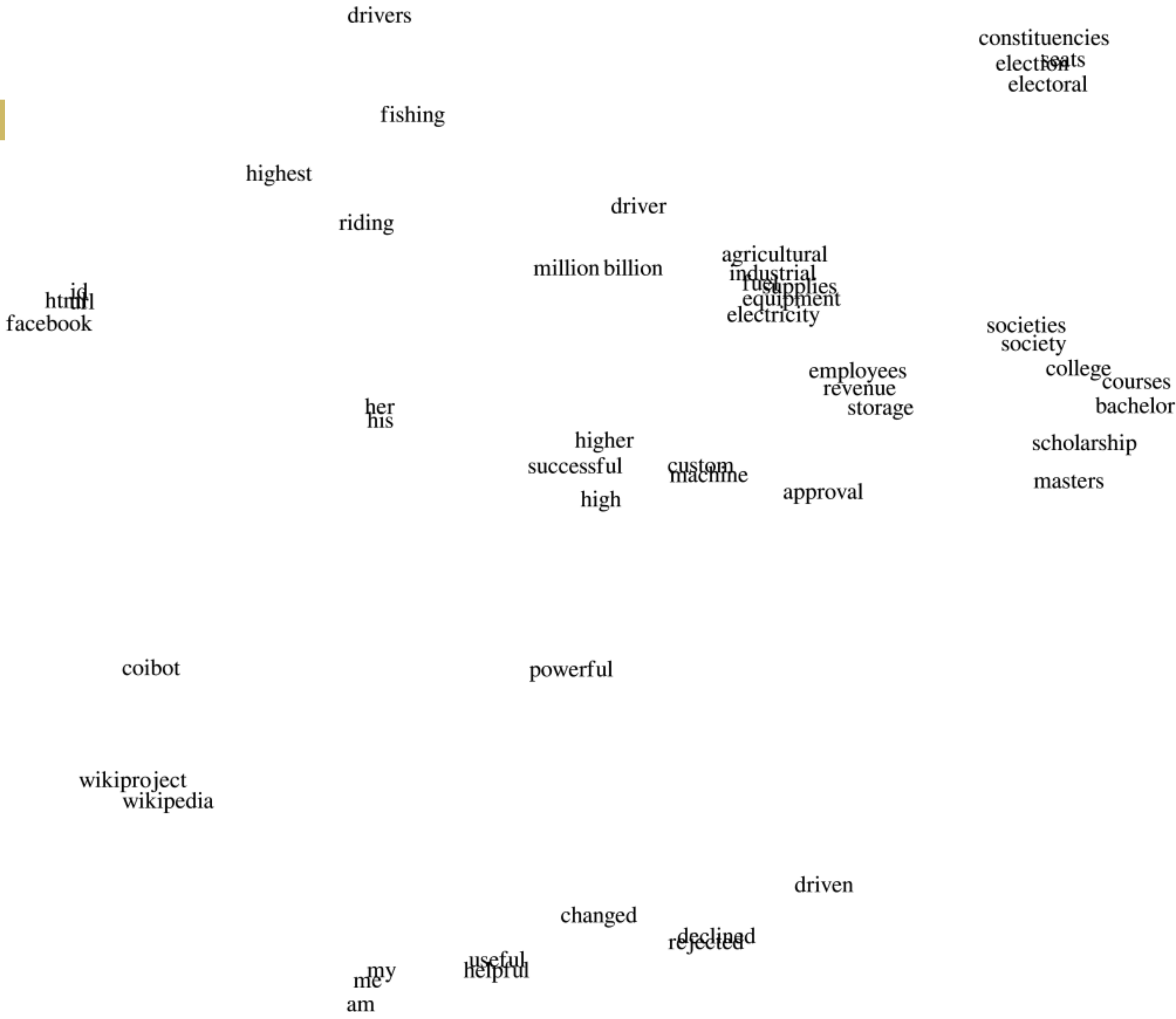
The full data set, which comprises over two billion culturomic trajectories, is available for download or exploration at www.culturomics.org and ngrams.googlelabs.com.

The size of the English lexicon. How many words are in the English language (*9*)?

Culturomics

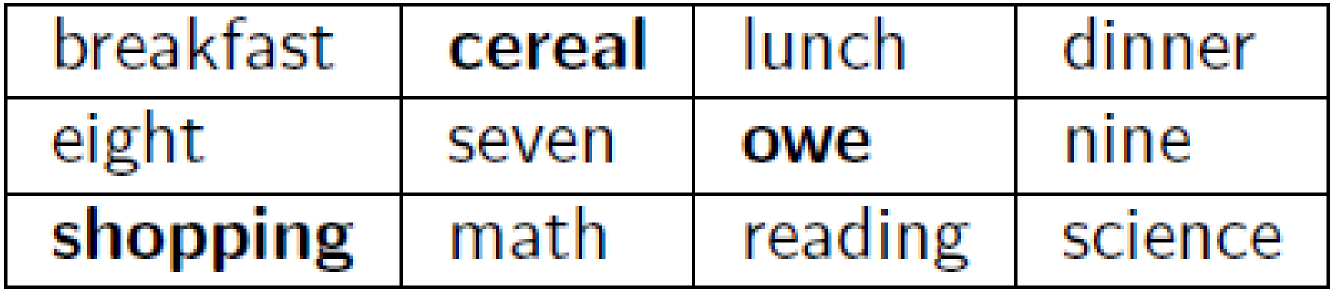


Word Vectors

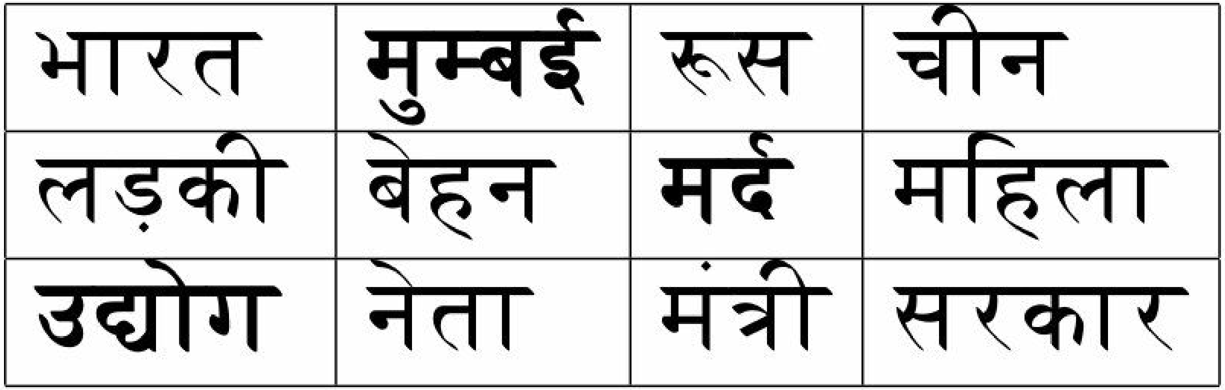


Pranjal singh 2015

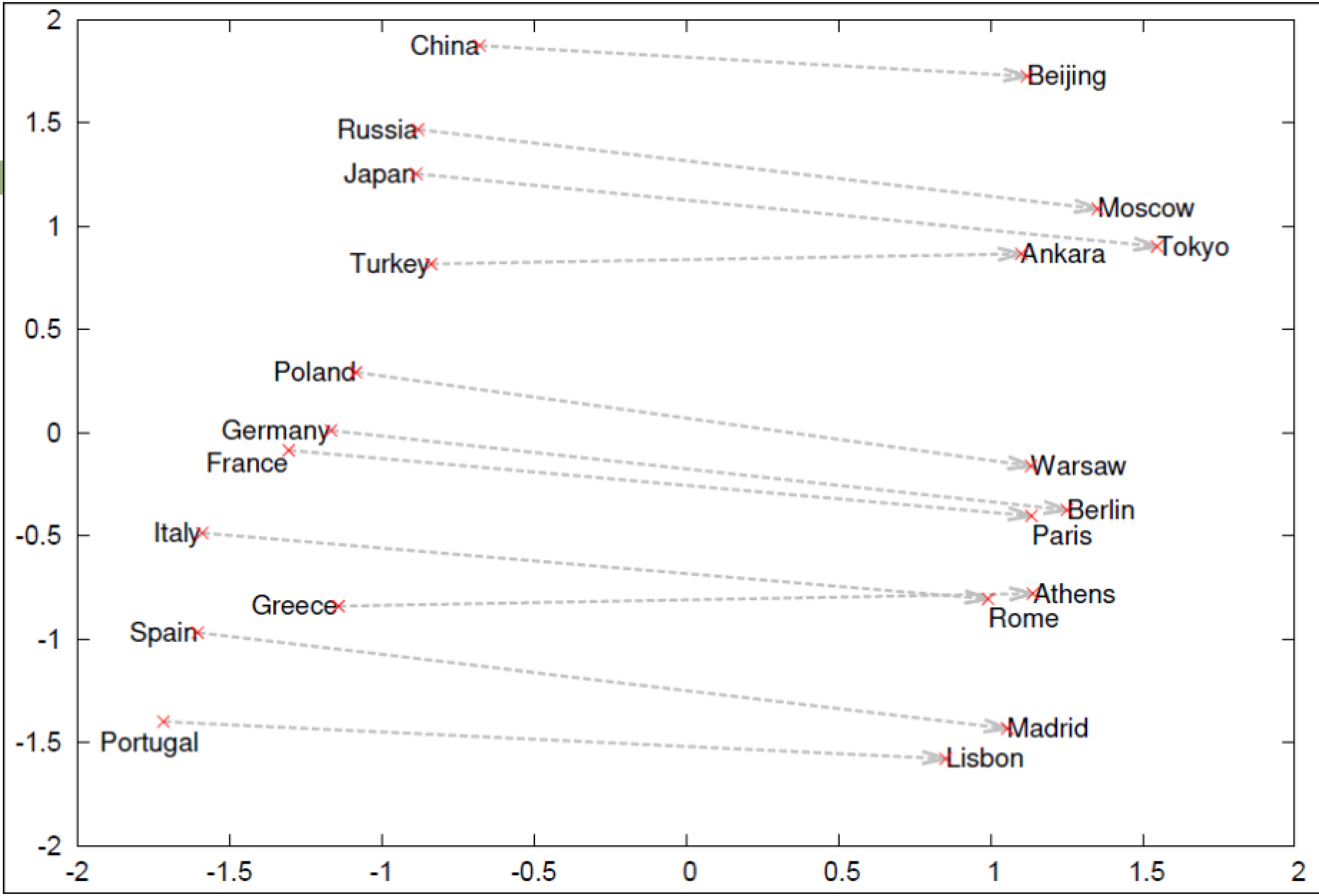
Odd One Out



breakfast	**cereal**	lunch	dinner
eight	seven	**owe**	nine
shopping	math	reading	science

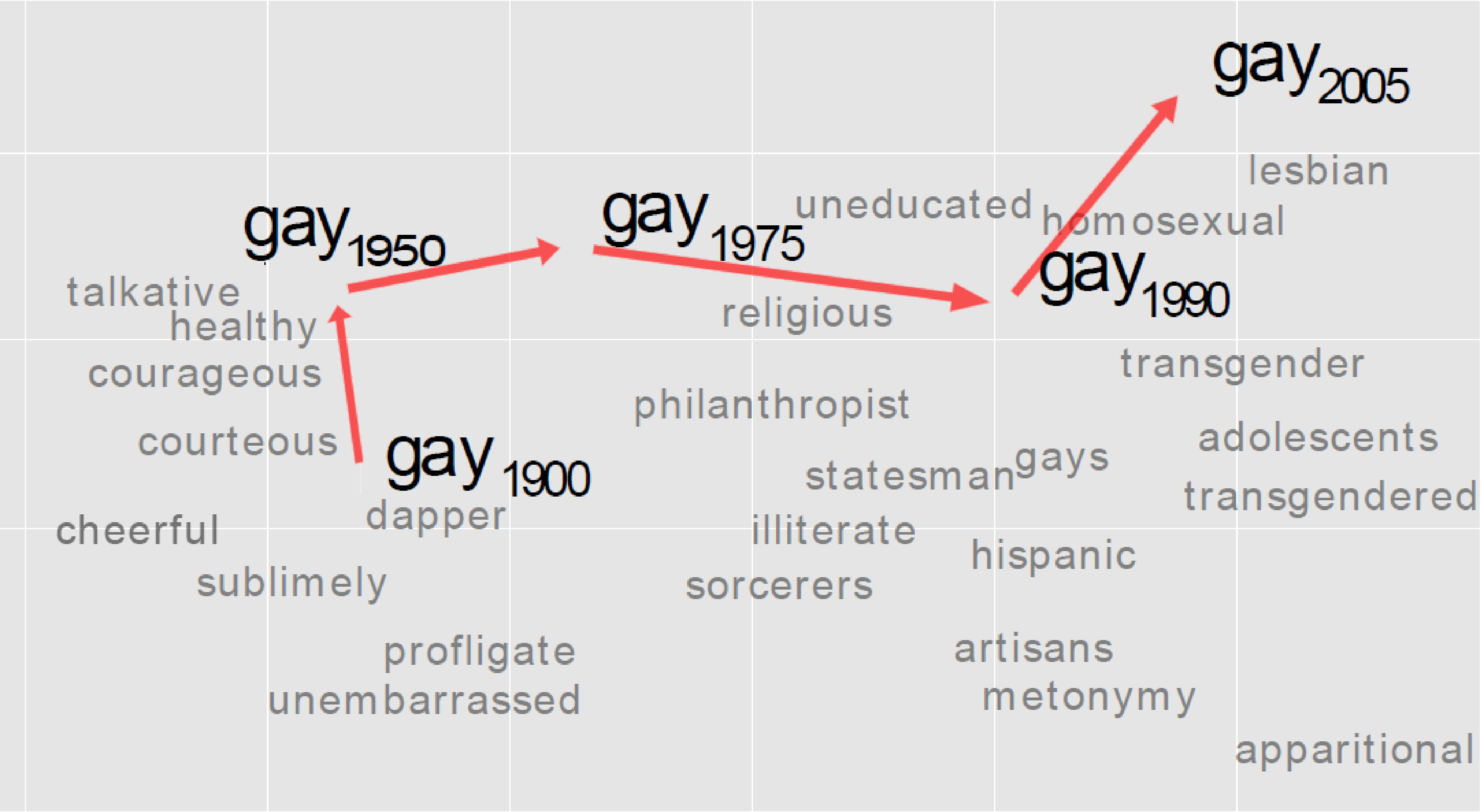


भारत	**मुम्बई**	रूस	चीन
लड़की	बेहन	**मर्द**	महिला
उद्योग	नेता	मंत्री	सरकार



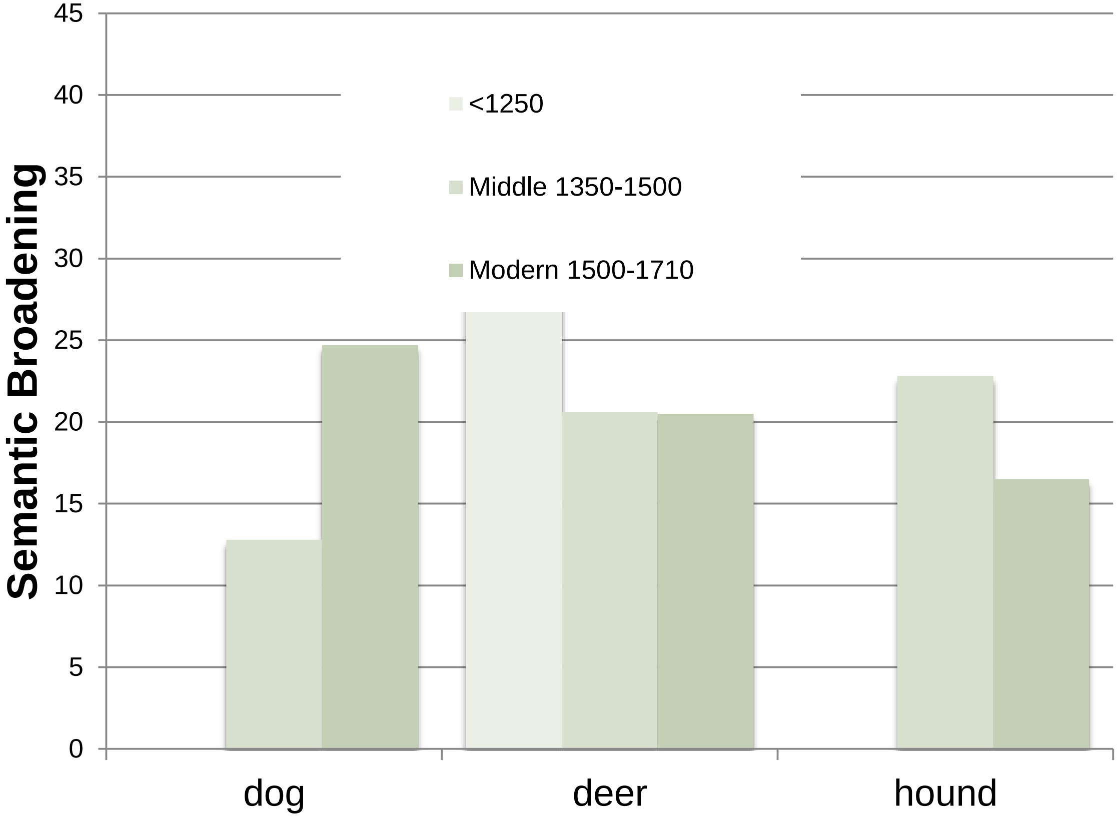
Shashwat Chandra 15

Culturomics



kulkarni-alRfou-perozzi-15_statistically-significant-linguistic-change

Historical lexicography



Sagi, Kaufmann Clark 2013

STRUCTURES IN LANGUAGE

amitabha mukerjee
iit kanpur

The magic of language

The magic of language

- You can't hold two watermelons in one hand

  - Iranian proverb

The magic of language

- Language is about conveying meaning
- Language is one-dimensional – Meaning is multi-dimensional

- Challenges
  Sounds along one-dimension express multi-dimensional aspects of reality
  - Same sounds map to different meanings [**Polysemy**]
  - Same meanings map to different sounds [**Synonymy**]

Myths about language

- **grammar** is about whether language is correct or incorrect

  It's me.

  Ganesh is at home?

  There are many small-small holes in this dress.

Myths about language

- **grammar** is about whether language is correct or incorrect

  It's me (accusative) → *"It's I"*

  Ganesh is at home? → *Is Ganesh at home?*

  There are many small-small holes in this dress.

- But how do we decide what is right?
- In Linguistics, grammar is determined based on language use.
  - descriptive, not prescriptive

Myths about language

- **grammar** is about the correct and incorrectness of language.

  Ganesh is at home? → *Is Ganesh at home?*

  It's me (accusative) → *"It's I"*

  There are many small-small holes in this dress.

- words are separated by spaces.
- how many sounds are there in English? 26

Myths about language

- **grammar** is about the correct and incorrectness of language.

  Ganesh is at home? → *Is Ganesh at home?*

  It's me (accusative) → *"It's I"*

  There are many small-small holes in this dress.

- words are separated by spaces.
- alphabets are the sounds of language

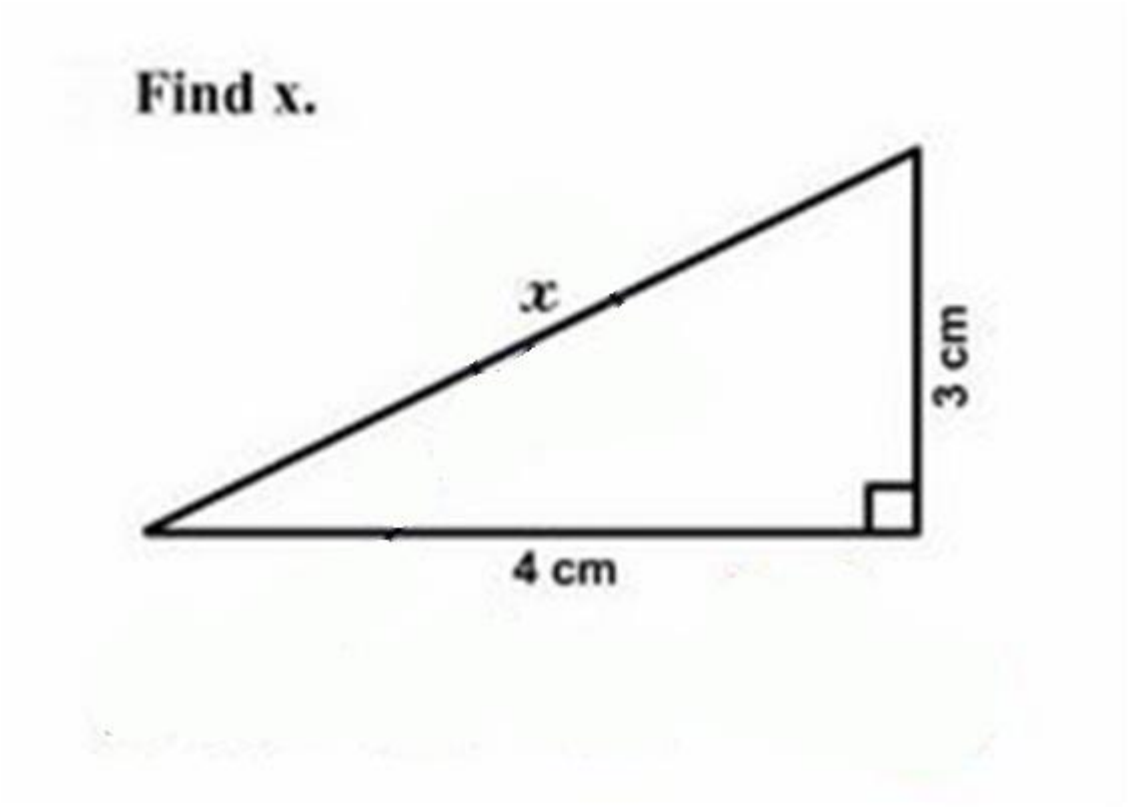
Levels of Grammar

- **Morphology** : how words are formed from smaller bits

  (*unopened* = *un* + *open* + *ed*)
- **Syntax**: how words are combined into sentences

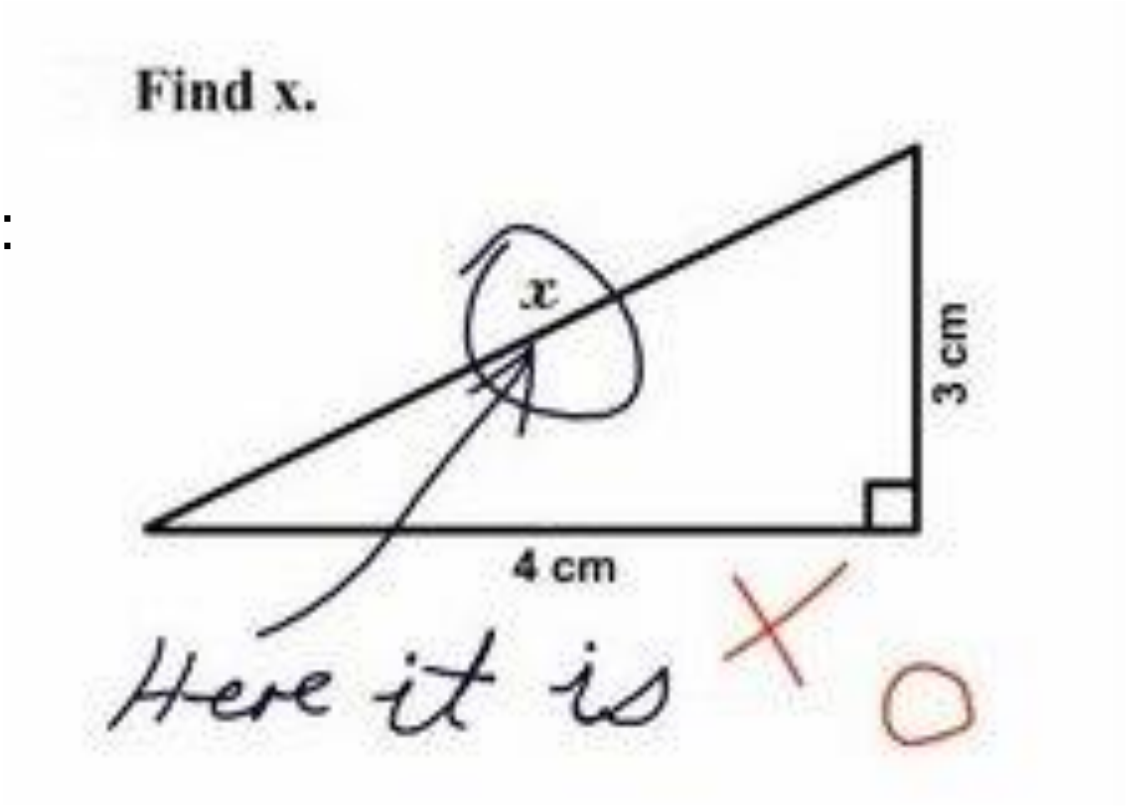
- Other levels of analysis:
  - **Phonology** : what sounds change the meaning
  - **Lexicon** : the inventory of *arbitrary* (?) words
  - **Semantics** : what language means directly
  - **Pragmatics** : what one infers from an utterance

Pragmatics: Meaning in Context



Pragmatics: Direct vs Indirect meaning

Traditional thinking:
Semantics
Direct meaning
Pragmatics
Indirect meaning



Pragmatics: Meaning in Context

Traditional levels of analysis:

- **Semantics**: composition from lexical meaning of words – "find" = detect, locate. [*direct meaning*]
- **Pragmatics**: social / contextual meaning ; [indirect meaning]

Psycholinguists:

Retrieval of pragmatic meaning is often faster

Syllables

धर्मक्षेत्रे कुरुक्षेत्रे समवेता युयुत्सवः ।

dharmakṣetre kurukṣetre samavetā yuyutsavaḥ

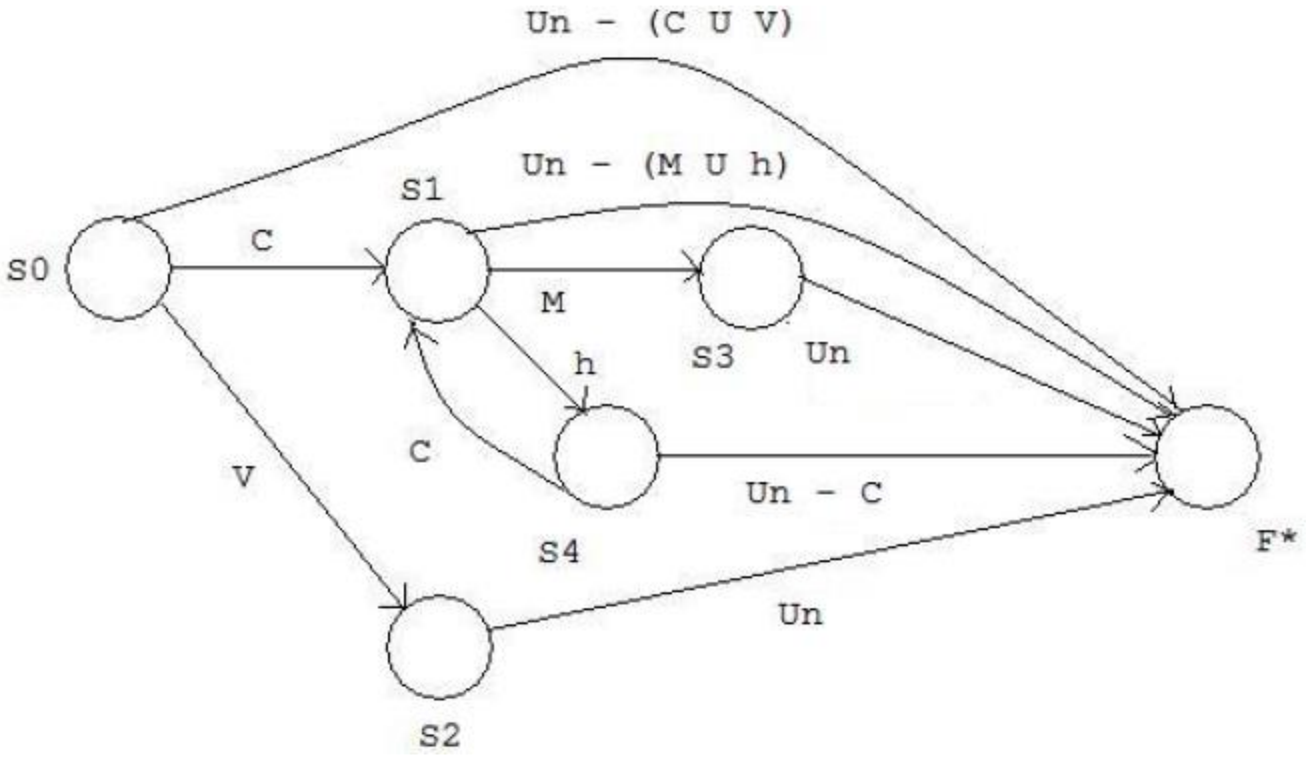
dhar + ma + kṣet + re … yu + yut + sa + vaḥ

ध र्म क्षे त्रे यु यु त्स वः ।

dha + rma + kṣe + tre … yu + yut + sa + vaḥ

Syllables

- U_n : all Unicode characters
- C : consonants
- V : vowels
- M : mAtrAs
- h : halant.



F* : failing state (sequence except last char = syllable; start next syllable with last character seen).

Syllabic writing (Katakana)

ka カ	ki キ	ku ク	ke ケ	ko コ
ma マ	mi ミ	mu ム	me メ	mo モ
tsu ツ				
ha は	hi ひ	fu ふ	he へ	ho ほ

CS 671 NLP
PHONOLOGY TO MORPHOLOGY

Language Structure: Levels

boys like girls

Language Structure: Levels

- **Phonology**
- **Lexicon**
- **Syntax [Morphology]**
- **Discourse**

- **Semantics / Compositionality**
- **Pragmatics / Discourse**

Language Structure: Levels

- **Phonology :** sounds of speech
  phoneme /b/ /oy/ /z/
- **Lexicon :** set of meaning-bearing units, **lexemes**
- **Syntax :** composing lexemes **composition**
  - **Word =** base + affixes / suffixes
  - **Phrase**: [[boys] [[like] girls]]
- **Discourse :** Boy likes girl. They meet.

NLP: Goals

Language → NLP → Decision

(NL Understanding)

NLP (MT)

Language 1 → Machine → Language 2

Translation

Situation → NLP → Language

(NL Generation)

Language Maps: Levels

- **Semantics** direct meaning
- **Pragmatics** social / implied meaning

NLP: Levels

NLP : deals with text. For languages with space-separation, deal with "orthographic words"

- **Morphology:** structures smaller than words
- **Syntax :** structures larger than words

- **Phonology:** impacts how text is written

Phonology

- Wide diversity in pronunciation and in hearing, yet we comprehend each other

- Phonetics: All possible human speech sounds
  phone
- Phonology: organization and structure of sounds of a language
  - ***Phoneme*** Minimal pair: *zip* | *sip*
    → */z/ and /s/ are different phonemes in English*

Speech sounds (phonemes)

- Which sounds change a meaning?

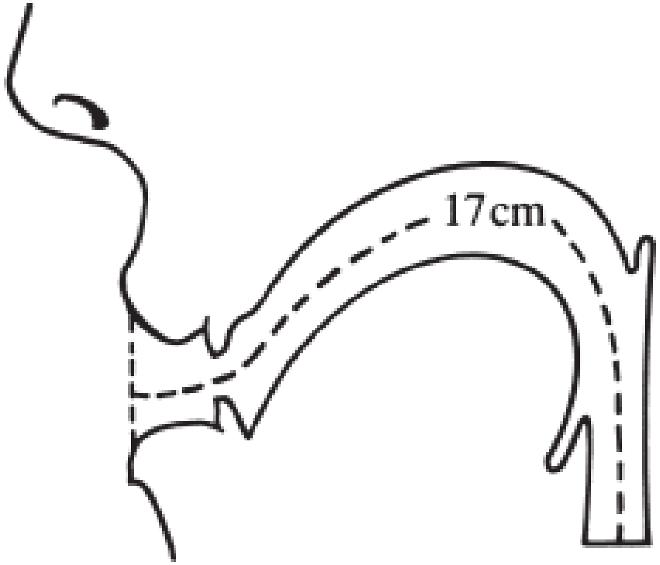
  pin, tin, kin, fin, thin, sin, shin

  dim, din, ding, did, dig, dish

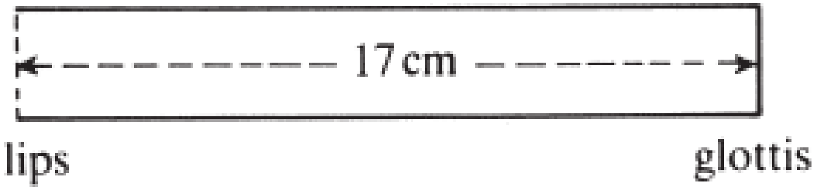
  pin, pen, pan, pun, pain, pine, pawn

- Phonemes at middle of syllable: **vowel**
  start or end: **consonant**

Vocal organs



tube model of vocal tract (for most neutral vowel)

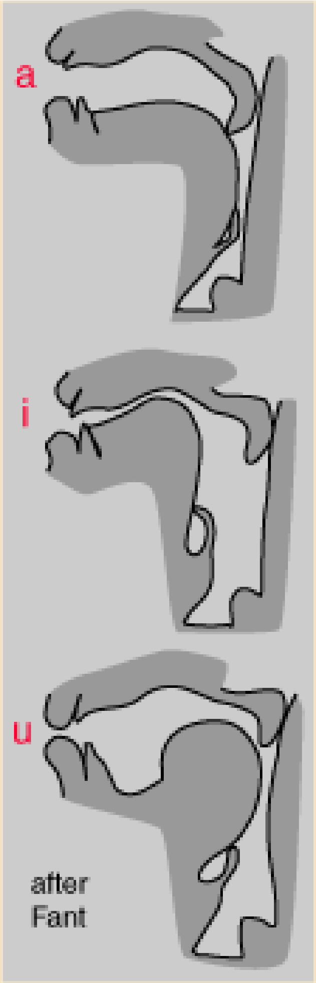


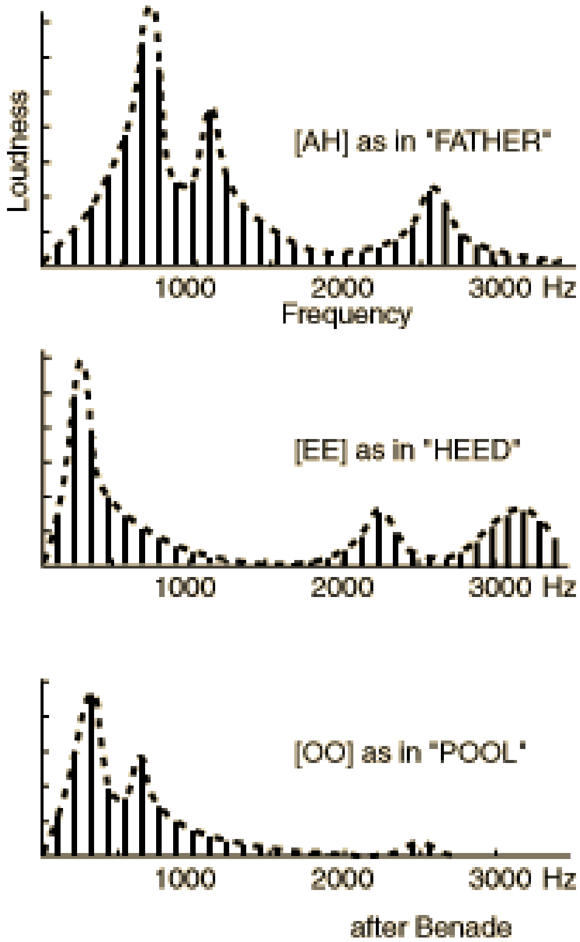
[malmkjaer 02]

Vowels : Formants

formant frequencies:
peaks in the
harmonic
spectrum of
vowel sounds

first three:
F1, F2, F3





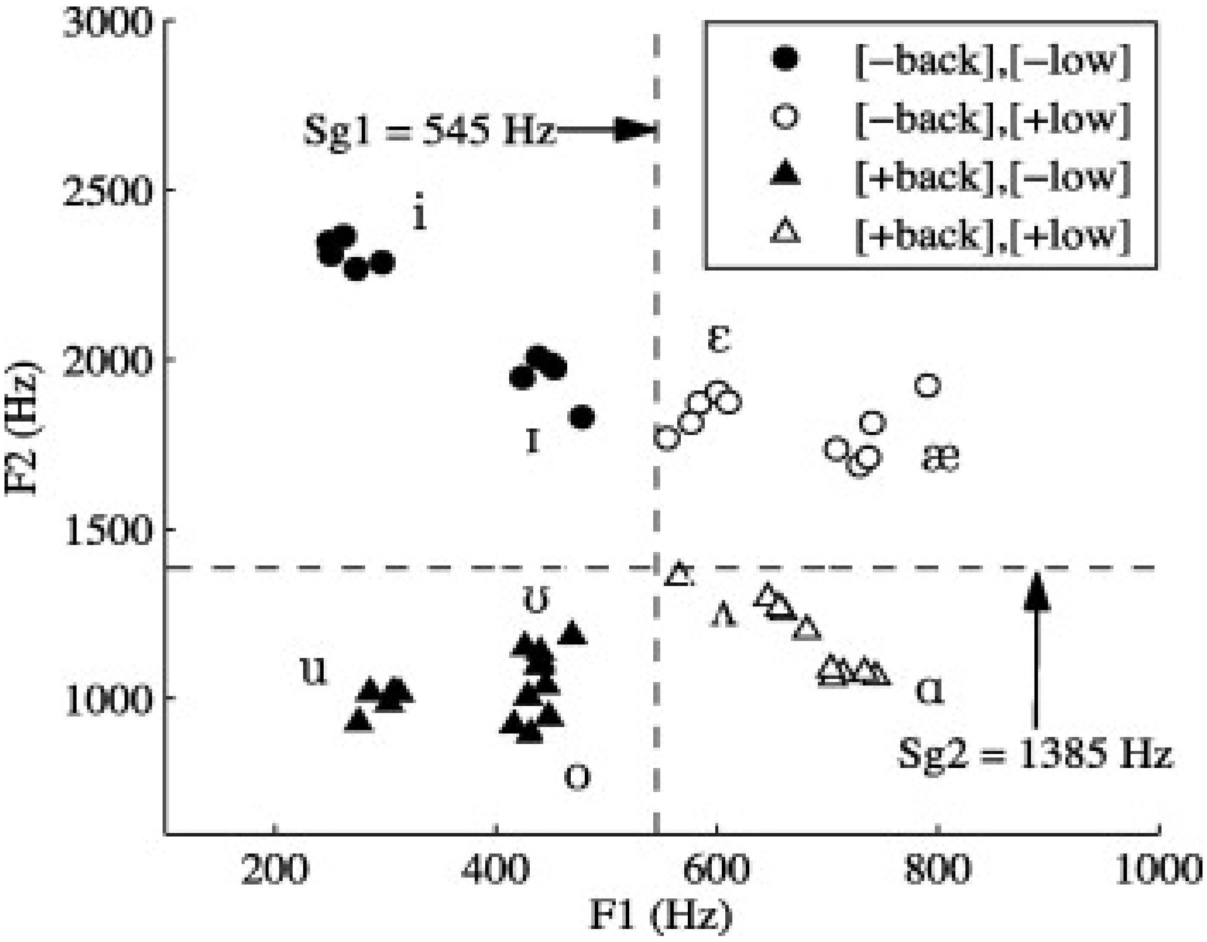
http://hyperphysics.phy-astr.gsu.edu/hbase/music/vowel.html

Vowels : Formants

Vowel space (F1,F2)

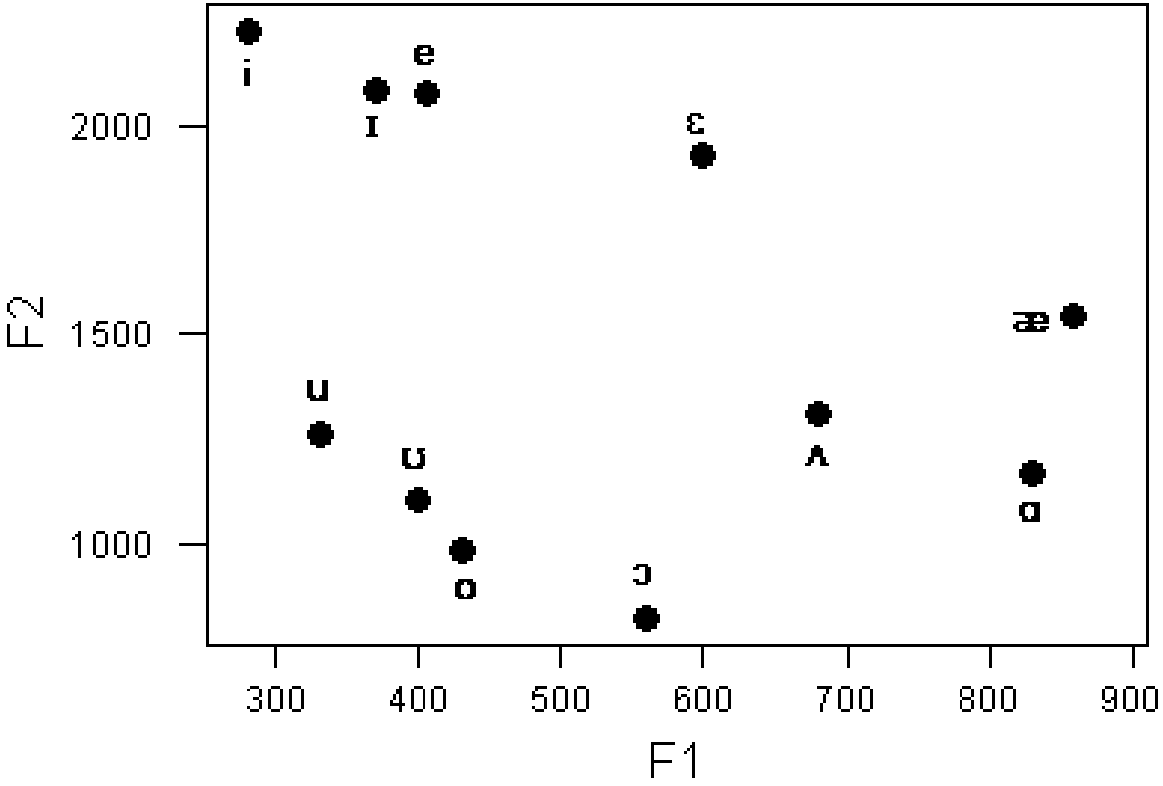
+low: F1 > 545hz;
+back : F2 > 1385hz

for a particular American English speaker (male):



[arsikere etal 11]

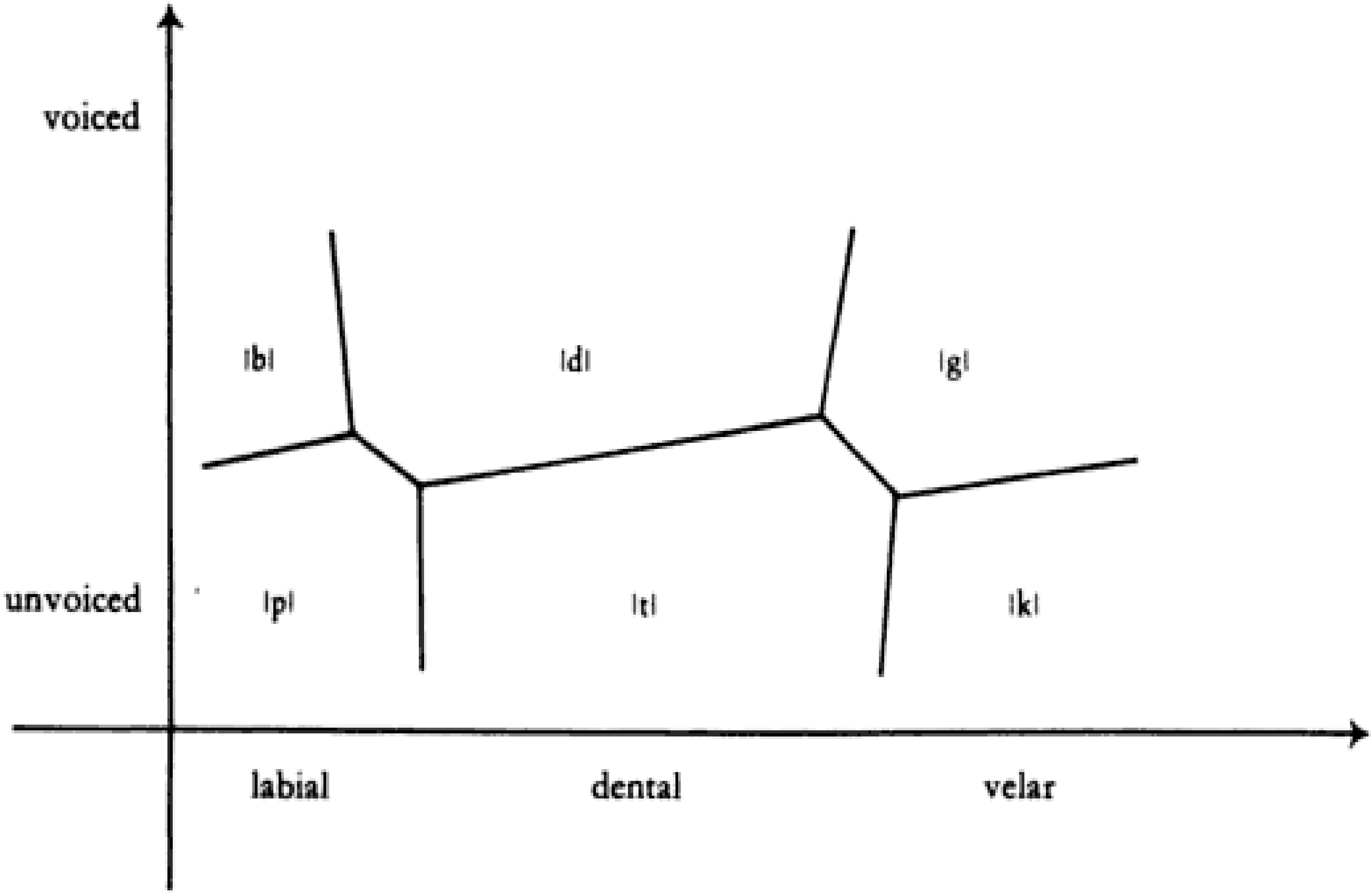
Vowels : Formants



Canadian English

http://hyperphysics.phy-astr.gsu.edu/hbase/music/vowel.html

Partitioning the speech sound space



[petitot 1989], [gardenfors 00]

Writing : Consonants

stop consonants

voiceless		voiced		nasal	
inaspirate	aspirated	in-	aspirated		
क	ख	ग	घ	ङ	[velar]
च	छ	ज	झ	ञ	[palatal]
ट	ठ	ड	ढ	ण	[retroflex]
त	थ	द	ध	न	[dental]
प	फ	ब	भ	म	[labial]

Consonants

stop consonants

voiceless		voiced		nasal	
inaspirate	aspirated	in-	aspirated		
k	kh	g	gh	N	[velar]
c	chh	j[dz]	jh[dzh]	n~	[palatal]
T	Th	D	Dh	N	[retroflex]
t	th	d	dh	n	[dental]
p	ph	b	bh	m	[labial] (bilabial)

Phonetic Notation

boys like girls

/bɔjz/ /lajk/ /gərlz/

Grammar of Phonology

“cats” → “cat” + /s/

“boys” → “boy” + /z/

Language Structures 2
Morphosyntax

Language Structure: Levels

- Phonology
- Lexicon
- Syntax [+Morphology]
- Discourse

- Prosody

- Orthography / Graphology

- Semantics / Compositionality
- Pragmatics / Discourse

Lexicon vs Grammar

- Grammar: how larger structures are assembled from smaller ones

- Smallest meaning-bearing structures = unit

- **morpheme :** less likely to appear independently

  -er , -s, -ly, -able

- **lexeme**

  cat, boy, smart, undergraduate student, cook, cooker

Grammar : Morphosyntax

“boys” → “boy” + s

[boys [like girls]]

Lexicon vs Grammar

- lexicon = mental inventory of units
  = set of all lexemes

- Is “cats” a lexeme?

cook → **cooks** : grammatical (rule-driven, inflection)
→ **cooker** : cook + er (not fully a rule; derivation)

Older thinking : lexicon is separate from grammar
at present : lexicon - grammar is a continuum

Syntax (morphosyntax)

- Regularity in how larger structures are assembled from units or smaller structures

- **morphology**

  cook-er / read-er / *-ercook
- **phrase syntax**

  smart woman / *woman smart
- **sentence syntax**

  boys like girls / girls like boys / *like boys girls