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

J. Laskar

CS 671
NATURAL LANGUAGE PROCESSING

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

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel, 1,2,3,4,5,6† Yuan Kui Shen, 3,4,7 Aviva Presser Aiden, 2,3,4,6 Adrian Veres, 3,4,6 Matthew K. Gray,* The Google Books Team, 10 Joseph P. Pickett, 11 Dale Holberg, 12 Dan Clancy, 10 Peter Norvig, 10 Jon Orwant, 10 Steven Pinker, 5 Martin A. Nowak, 3,11,14 Erez Lieberman Aiden 1,2,4,15,16,17,18†

We constructed a corpus of digitized texts containing about 1% 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 librarians 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 \(3.5 \times 10^{-5}\). 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 facts 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 2 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?
Culturomics

Diagram A: 129 million books published, 15 million books scanned, 5 million books analyzed. Frequency of the word "apple" over time.

Diagram B: Frequency of the word "slavery" over time, with peaks around 1850 and 1950.

Diagram C: Frequency of "the Great War," "World War I," and "World War II" over time, with peaks around 1914-1918 and 1939-1945.
### Odd One Out

<table>
<thead>
<tr>
<th>breakfast</th>
<th>cereal</th>
<th>lunch</th>
<th>dinner</th>
</tr>
</thead>
<tbody>
<tr>
<td>eight</td>
<td>seven</td>
<td>owe</td>
<td>nine</td>
</tr>
<tr>
<td><strong>shopping</strong></td>
<td>math</td>
<td>reading</td>
<td>science</td>
</tr>
</tbody>
</table>

### Hindi Words

<table>
<thead>
<tr>
<th>भारत</th>
<th>मुम्बई</th>
<th>रूस</th>
<th>चीन</th>
</tr>
</thead>
<tbody>
<tr>
<td>लड़की</td>
<td>बहन</td>
<td>मर्द</td>
<td>महिला</td>
</tr>
<tr>
<td>उद्योग</td>
<td>नेता</td>
<td>मंत्री</td>
<td>सरकार</td>
</tr>
</tbody>
</table>
Culturomics

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

Semantic Broadening

- Dog: <1250, Middle 1350-1500, Modern 1500-1710
- Deer: <1250, Middle 1350-1500, Modern 1500-1710
- Hound: <1250, Middle 1350-1500, Modern 1500-1710

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
  
  \( \text{unopened} = \text{un} + \text{open} + \text{ed} \)

- **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
Find x.
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
धर्मक्षेत्रे कुरुक्षेत्रे समवेता युयुत्सवः ।
dharmakṣetre kurukṣetre samavetā yuyutsavaḥ

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

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
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 $\rightarrow$ NLP $\rightarrow$ Decision
(NL Understanding)

Language 1 $\rightarrow$ Machine $\rightarrow$ Language 2
Translation

Situation $\rightarrow$ NLP $\rightarrow$ 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)
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: \( F1 > 545\text{hz} \);
+back: \( F2 > 1385\text{hz} \)

for a particular American English speaker (male):

arsikere etal 11
Vowels : Formants

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

Canadian English
Partitioning the speech sound space

[petitot 1989], [gardenfors 00]
stop consonants

<table>
<thead>
<tr>
<th>voiceless</th>
<th>voiced</th>
<th>nasal</th>
</tr>
</thead>
<tbody>
<tr>
<td>inaspirate</td>
<td>aspirated</td>
<td>in- aspirated</td>
</tr>
<tr>
<td>क ख ग घ ङ</td>
<td>च छ ज झ ञ</td>
<td>[velop]</td>
</tr>
<tr>
<td>च छ ज झ ञ</td>
<td>ट ठ ड ढ ण</td>
<td>[palatal]</td>
</tr>
<tr>
<td>ट ठ ड ढ ण</td>
<td>त थ द ध न</td>
<td>[retroflex]</td>
</tr>
<tr>
<td>प फ ब भ म</td>
<td>[labial]</td>
<td></td>
</tr>
</tbody>
</table>
# Consonants

<table>
<thead>
<tr>
<th>stop consonants</th>
<th>voiceless</th>
<th>voiced</th>
<th>nasal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>inaspirate</td>
<td>aspirated</td>
<td></td>
</tr>
<tr>
<td>k</td>
<td></td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>kh</td>
<td></td>
<td></td>
<td>[velar]</td>
</tr>
<tr>
<td>g</td>
<td></td>
<td>gh</td>
<td>N</td>
</tr>
<tr>
<td>jh[dzh]n~</td>
<td></td>
<td></td>
<td>[palatal]</td>
</tr>
<tr>
<td>j[dz]</td>
<td></td>
<td></td>
<td>[retroflex]</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td>chh</td>
<td>N</td>
</tr>
<tr>
<td>chh</td>
<td></td>
<td></td>
<td>[labial] (bilabial)</td>
</tr>
<tr>
<td>T</td>
<td></td>
<td>Th</td>
<td>N</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>Dh</td>
<td>N</td>
</tr>
<tr>
<td>t</td>
<td></td>
<td>th</td>
<td>n</td>
</tr>
<tr>
<td>d</td>
<td></td>
<td>dh</td>
<td>n</td>
</tr>
<tr>
<td>p</td>
<td></td>
<td>ph</td>
<td>m</td>
</tr>
<tr>
<td>b</td>
<td></td>
<td>bh</td>
<td>m</td>
</tr>
</tbody>
</table>
Phonetic Notation

boys like girls

/bojz/ /lajk/ /gərlz/
"cats" → "cat" + /s/

"boys" → "boy" + /z/
Language Structures 2
Morphosyntax
Language Structure: Levels

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

- Semantics / Compositionality
- Pragmatics / Discourse

- Prosody
- Orthography / Graphology
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