The ABC of computational Text Analysis

02: Text as Data

Alex Flückiger
05 March 2020

Outline

- recap + reading
- methodological ground 😬
- first computational investigation

Recap last Lecture

computer as ...

- ... an intelligent device
- ... a tool for analysis

datafication

- abundance of data
- exploit new form of data

Reading

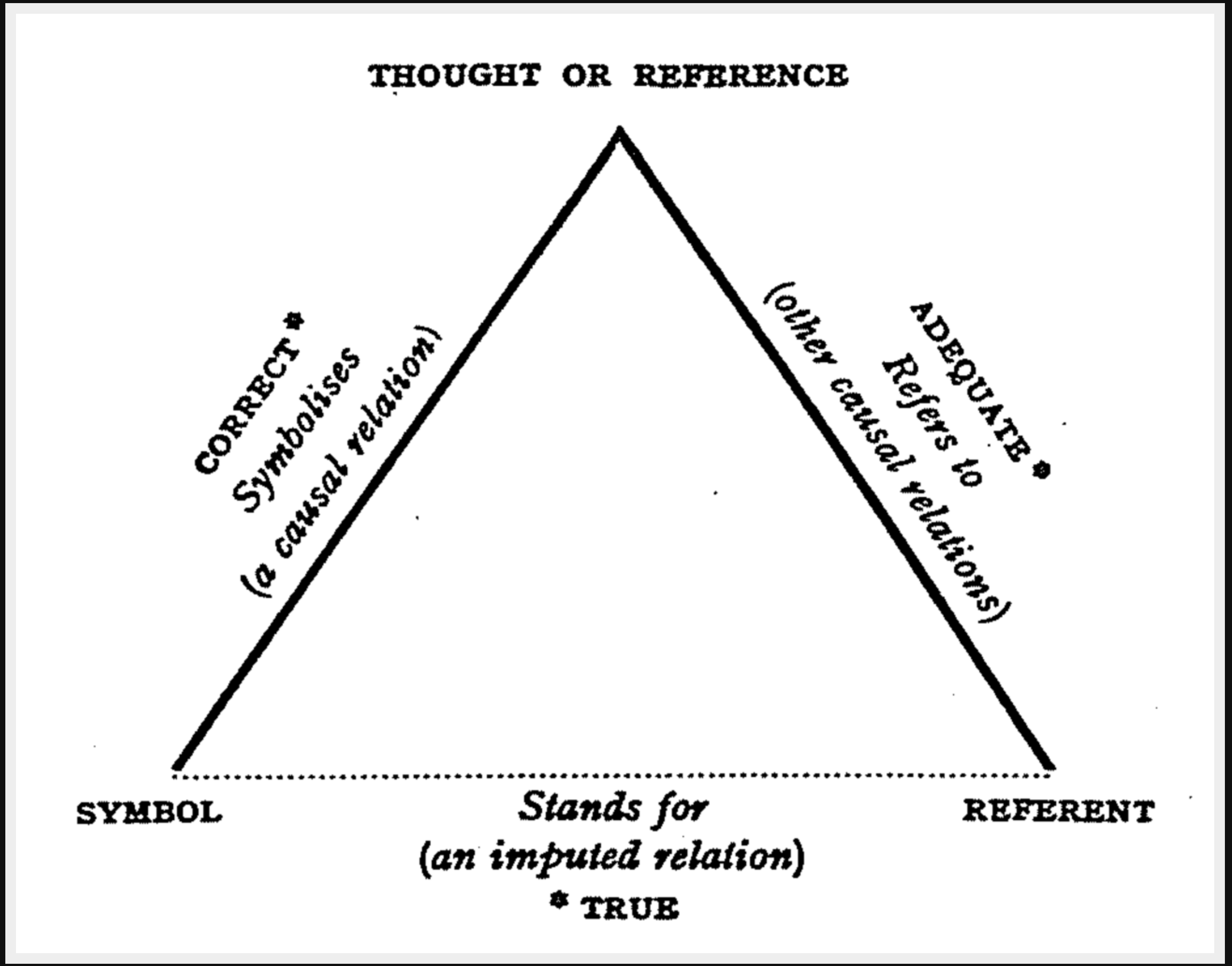
Computational Social Science (Lazer et al. 2009)

- data-driven
- network analysis + text analysis
- historical perspective vs. real-time dynamics
- limited access to data

Semiotic Triangle

Loose coupling between

- World
- Cognition
- Language



Semiotic Triangle (Ogden and Richards 1923)

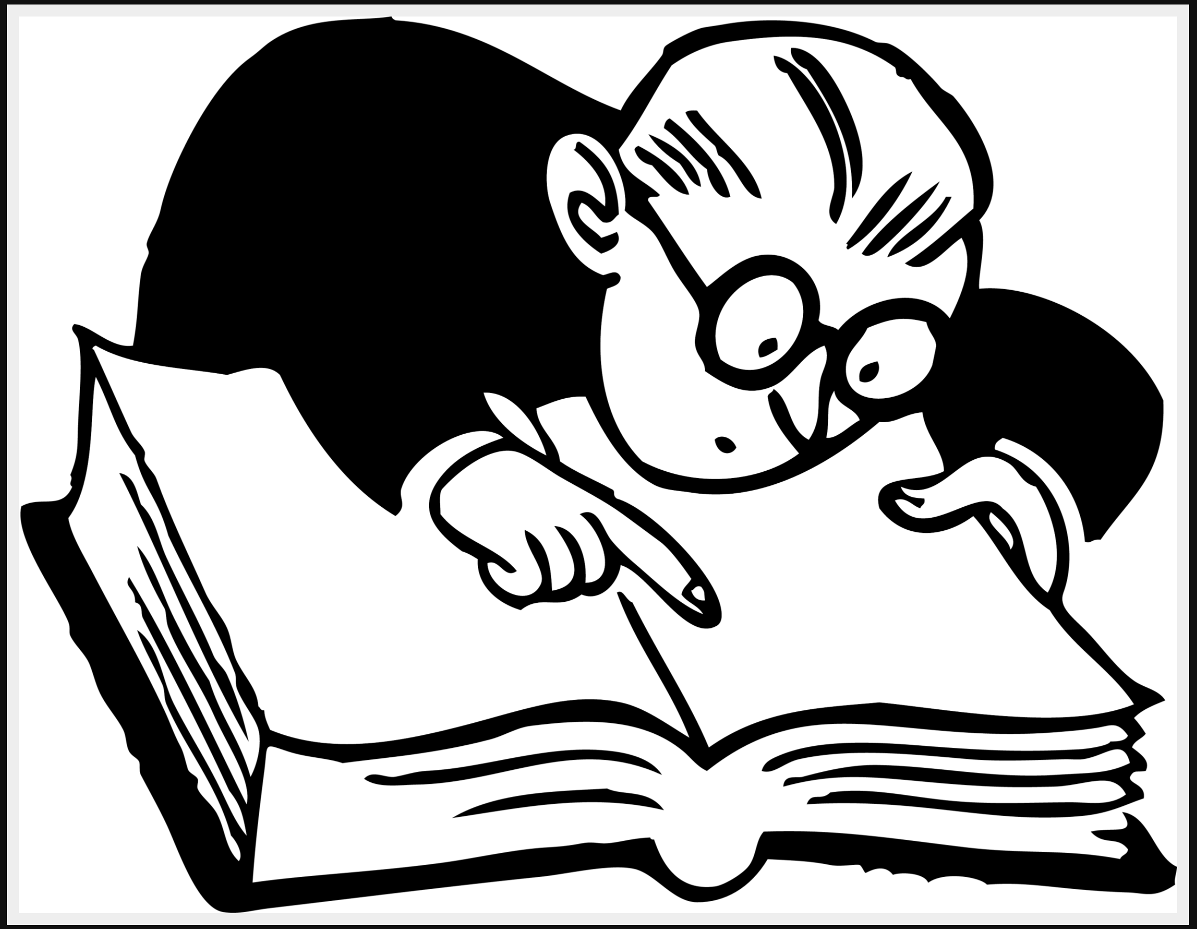
> «Language shapes the way we think,
> and determines what we can think about.»

— Benjamin Lee Whorf

Working with Texts

A micro and macro perspective I

individual cases vs. collective trends



close reading (source)

distant reading (source)

A micro and macro perspective II

non-scalability vs. abstraction

tons of text (source)

meaning of numbers (source)

From micro to macro…
…and back again

Two Research Paradigms

data exploration vs. hypothesis testing

- add nuance
- develop new narratives
- verify hypothesis

Remember 👍

Quantification and qualitative analysis go well together.

Numbers do not talk, never.

Text as Data

- synonymy
- ambiguities
- compositonality of meaning
- agnostic, discrete symbols
- unstructured, messy data

(see also Grimmer and Stewart 2013)

Data Formats

In-class task: File types

- What file formats do you know?
- Open files of different types in a text editor.

  Which look nice?

File formats

- machine-readability

  raw: txt, csv, tsv
  formatted: docx, pdf, html, xml

- open vs. proprietary
- digital sustainability

Let’s dive into it!

Ngram Viewer

Google Ngram Viewer

- historical perspective with n-grams
- 5.2 million books
- rise and fall of cultural ideas and phenomena

More information

In-Class Task: Environmental Discourse

questions about environmental discourse

Refine your Queries

check out case-sensitiveness, wildcards (*) and operators 🤓

Operator	Description
`+`	sums multiple expressions into one
`-`	subtracts the expression on the right from the expression on the left, giving you a way to measure one ngram relative to another.
`/`	divides the expression on the left by the expression on the right, which is useful for isolating the behavior of an ngram with respect to another.
`*`	multiplies the expression on the left by the number on the right, making it easier to compare ngrams of very different frequencies. (Be sure to enclose the entire ngram in parentheses so that * isn’t interpreted as a wildcard.)

Remember 👍

Has the language evolved over time or the social perception?

Both, most likely.

Similarly, language may vary across regions and communities.

No Culturomics but Meaning-Making

phenomena in collective memory

- semantic drifts
- lexical shifts

Read, read, read to complement stats with context!

Questions of Interpretation

Example: interpretation of decreasing frequency

- loosing interest
- becoming an established fact
- new reference

  The Great War → *World War I*

A word of caution

The unknowns of Google Ngram Viewer

- indexed books

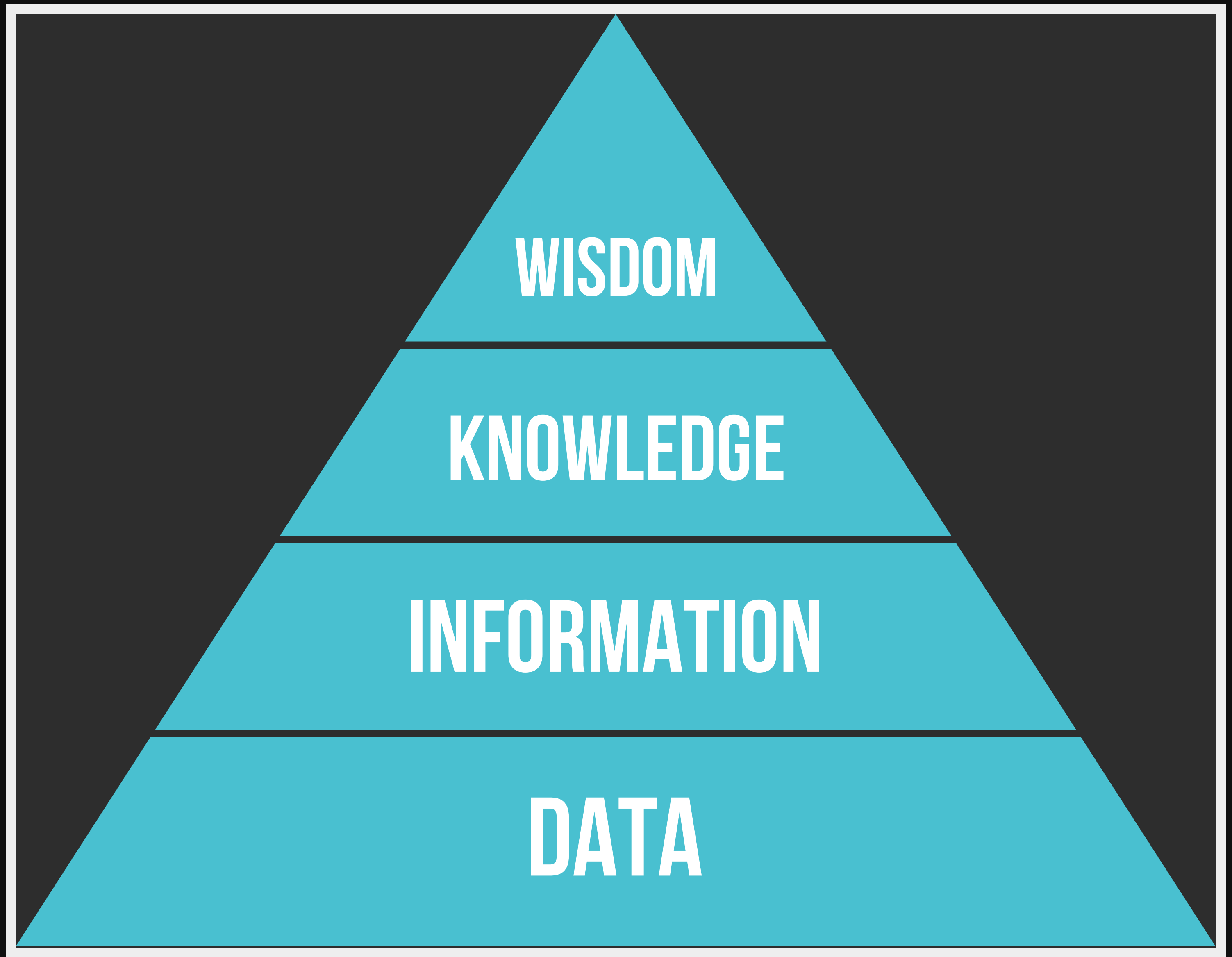
  genre, authors, quantity

- artifacts of digitalization

use alternative: bookworm HathiTrust

Interacting, not mapping

It is a lense, not a map.



DIKW pyramid (Wikipedia)

References

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. "Computational Social Science." *Science* 323 (5915): 721–23. https://doi.org/10.1126/science.1167742.

Ogden, Charles Kay, and Ivor Armstrong Richards. 1923. *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism. Supplementary Essays by B. Malinowski and F.G. Crookshank*. New York: Harcourt. http://books.google.com?id=i3MIAQAAIAAJ.