

The ABC of Computational Text Analysis

#2 Text as Data

Alex Flückiger
Faculty of Humanities and Social Sciences
University of Lucerne

Outline

- questions
 - assignment, website, course
- recap last lecture
- methodical foundation 😊
- first computational text analysis ✨

Recap last lecture

computer as ...

- ... an intelligent device
- ... a tool for a **new social science**

datafication

- abundance of data
- exploit new form of data

Why analyzing texts?

Ceci n'est pas une banane



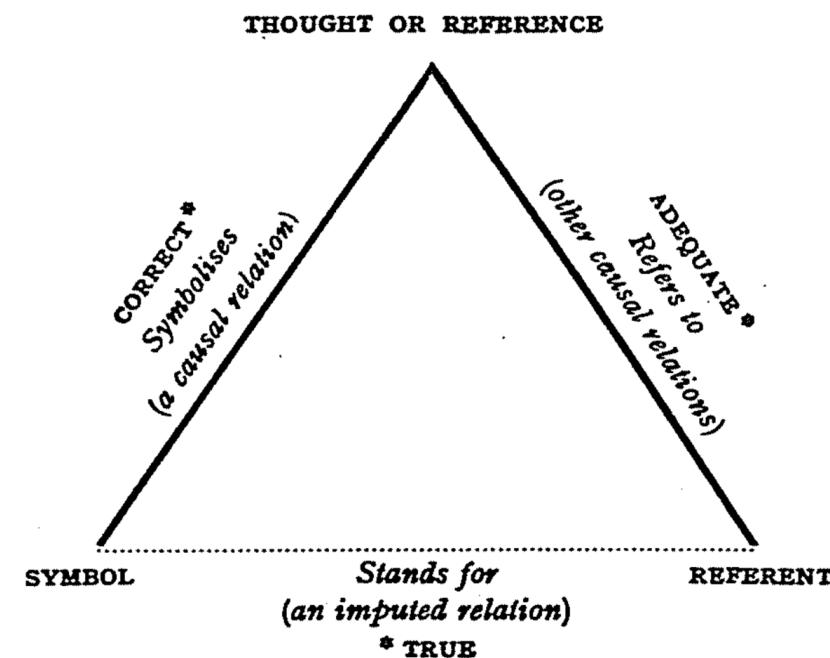
Max Gruber / [Better Images of AI](#)

Semiotic Triangle

Loose coupling between

- World
- Cognition
- Language

synonyms, ambiguity



Semiotic Triangle (Ogden and Richards 1923)

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

—**Benjamin Lee Whorf**

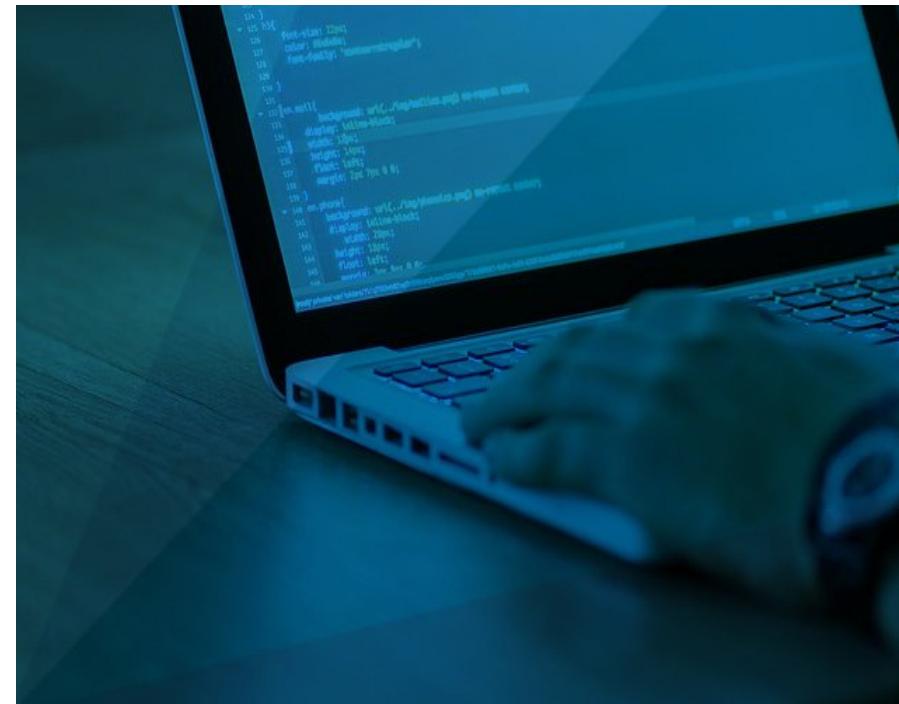
Working with Texts

A micro and macro perspective I

Identifying trends beyond individual cases



Close reading to understand a text in depth



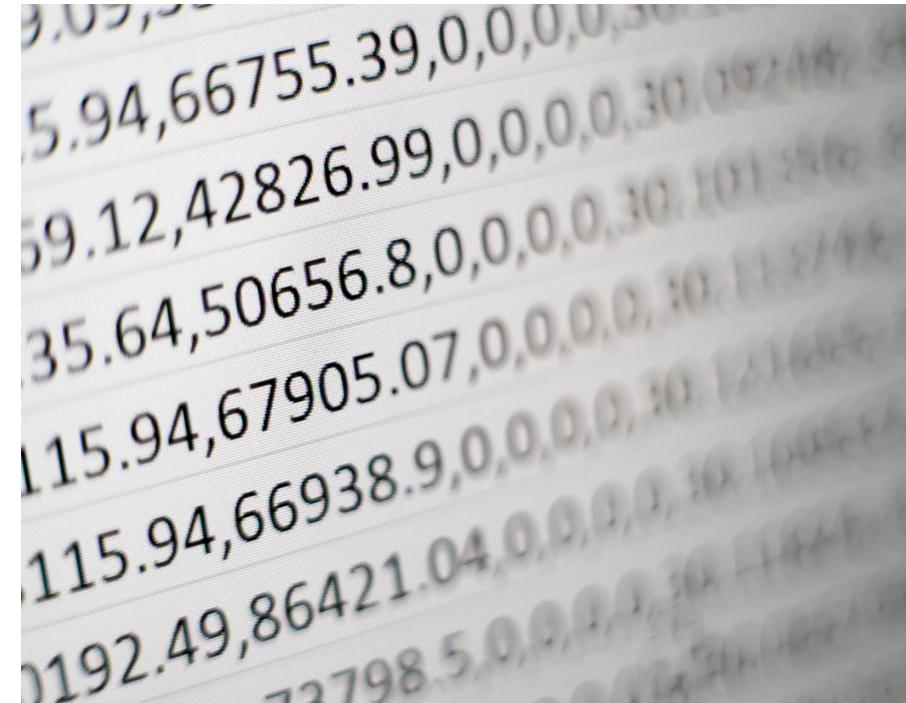
Distant reading to analyse trends across texts (Moretti 2000)

A micro and macro perspective II

Scale leads to abstraction



Too big to analyse by manual means



What does these abstract numbers represent? And what is abstracted away?

From micro to macro
...and back again



Two research paradigms

data exploration vs. hypothesis testing (Grimmer, Roberts, and Stewart 2021)

- add nuance
- develop new narratives
- verify hypothesis

Numbers do not talk



Thus, quantification and qualitative analysis go well together.

Text as Data

Text is challenging for computers due to

- synonymy
- ambiguities
- compositionality of meaning
- discrete symbols
- unstructured, messy data

(see also Grimmer and Stewart 2013)

Unstructured Text?



Collection > Documents > Paragraphs > Sentences >
Words



Challenging structure of texts does not imply no structure.

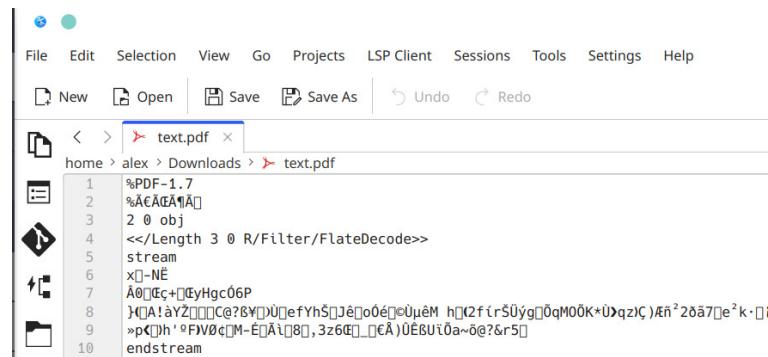
Data Formats

In-class Task: File Types

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

File Types

- any filename consists of name + suffix
 - suffix defines the file type
 - e.g. task.txt
 - machine-readability
 - raw: .txt .csv .tsv ...
 - formatted: .docx .pdf .html .xml ...
 - open vs proprietary
 - digital sustainability



File Management



Use meaningful names

- no spaces/umlauts
 - only: alphanumeric, underscore, hyphen, dot
- versioning using date
 - e.g. `task_20240229.pdf` instead of `task_new_final.pdf`

Let's dive into it! 

Counting ngrams

Google Ngram Viewer (Michel et al. 2011)

- historical perspective with ngrams
- >5.2 million books
- rise and fall of cultural ideas and phenomena

In-Class Task: Investigate the environmental discourse

- What other terms have been used to describe nature?
e.g. environment
- What environmental issues are debated the strongest? When?
e.g. nuclear power plant
- Are there any differences between languages? Between corpus versions?
i.e. similar words with non-equivalent curves over time



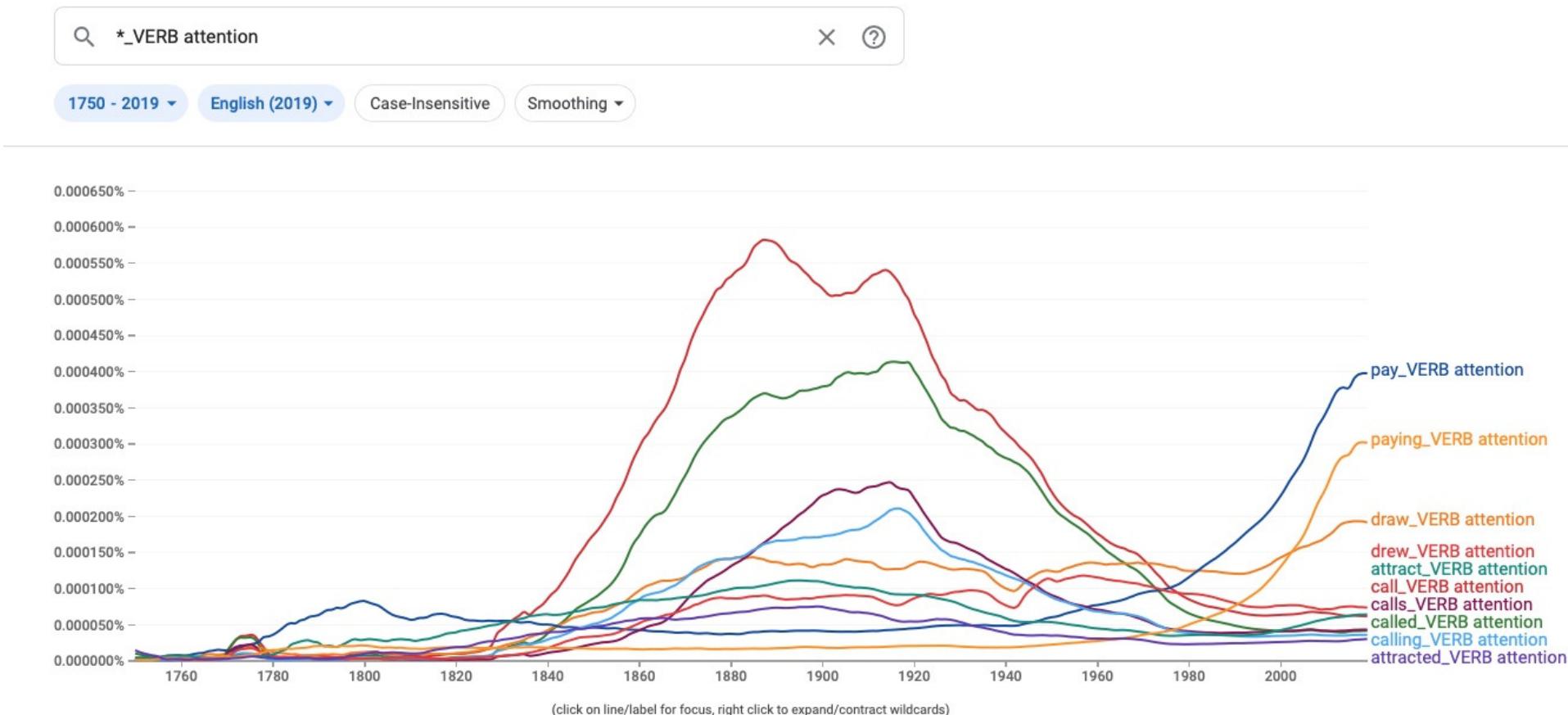
What do you conclude from your observations?

Refine your queries

Check out case-sensitiveness, wildcards (*) and operators 😎

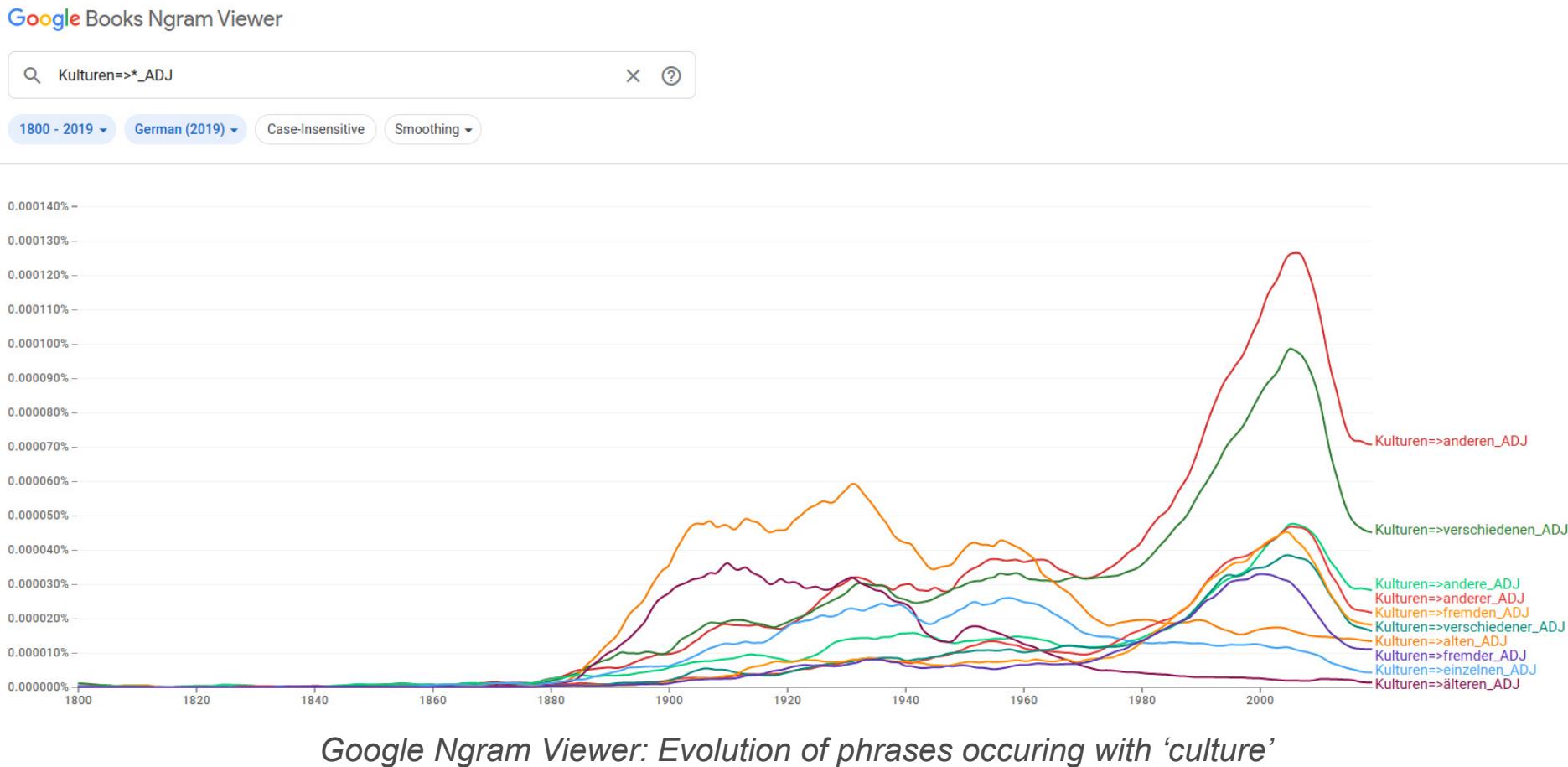
Operator	Description
+	sums multiple expressions to aggregate trends .
-	subtracts an expression from another to measure one ngram relative to another .
/	divides the expression by another one for isolating the behavior of an ngram with respect to another .
*	multiplies the expression by a number to compare ngrams of very different frequencies. (Enclose the ngram in parentheses so that * isn't interpreted as a wildcard.)

The raise of the ngram pay attention



Google Ngram Viewer: Evolution of the phrase 'attention'

The raise of the ngram different culture



Has the language evolved over time or
the social perception? 🤔

Likely both.

Similarly, language may vary across regions and communities.

No Culturomics but meaning-making

Phenomena in collective memory

- semantic drifts (meaning)
- lexical shifts (frequency)

Read, read, read to complement **stats** with context!

Interpretation

Potential reasons of decreasing frequency

- loosing interest
- becoming an established fact
- new reference

The Great War → World War I

- news values and media cycles
- selection of data sources

A word of caution

The unknowns of Google Ngram Viewer

- index of books
 - genre, authors, quantity
- artifacts of digitalization



use better alternative: [bookworm HathiTrust](#)

A large pile of jigsaw puzzle pieces in various colors and patterns, including yellow, blue, red, green, and orange. The pieces are scattered across the entire frame, creating a textured background.

Research in
practice means
organizing

The Zen of organizing

How a computational approach helps

- code as documentation allowing for criticism 
- efficient automation 
“don’t repeat yourself”
- less error-prone  
- reproducible 

Prepare your system

1. backup files + update system 
2. start installation with this [guide](#) 

Reading

Required

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. "Computational Social Science." *Science* 323(5915):721–23.

(via OLAT)

Optional

Graham, Shawn, Ian Milligan, and Scott Weingart. 2015. *Exploring Big Historical Data: The Historian's Macroscope*. Open Draft Version. Under contract with Imperial College Press.

[online](#)



Questions?

References

- Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24 (1): 395–419.
<https://doi.org/10.1146/annurev-polisci-053119-015921>.
- 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>.
- Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82. <https://doi.org/10.1126/science.1199644>.
- Moretti, Franco. 2000. "Conjectures on World Literature." *New Left Review* 1: 54–68.
<http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature>.
- 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. <https://books.google.com?id=i3MIAQAAIAAJ>.