# Lexicon-based Approaches

COST Action Training School in Computational Opinion Analysis – COpA

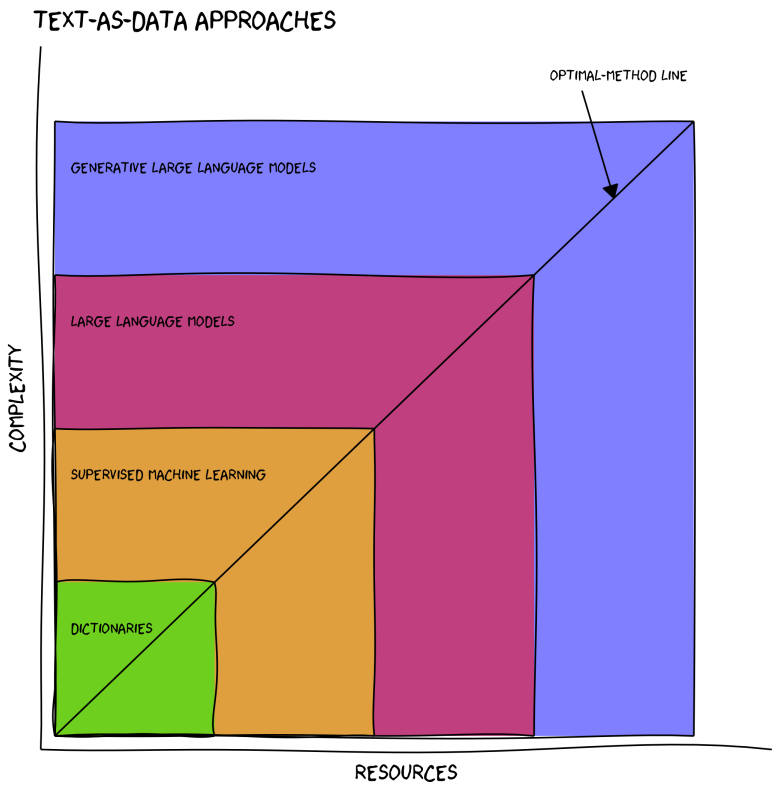
Johannes B. Gruber | VU Amsterdam

# Introduction

# What are Lexicon/Dictionary-based Approaches?

- Lexicon/Dictionary: the words in a language and their meaning
- Lexicon/Dictionary-based approaches: simply count how often pre-defined words appear to infer meaning of text
- Wordcounts are usualy used to categorise text (e.g., non-/relevant, positive/negative, a-/political)
- To infer category from count, researchers define mapping function (e.g., N positive terms > N negative terms = positive text)
- Like ‘normal’ dictionaries: several forms of the word carry same meaning, expressed through wildcards (e.g., econom*) or regular expressions (e.g., econom.+) (matches economists, economic, and so on)

# Deciding on the Right Method




Whether you've chose the optimal method is not always clear beforehand. But you should consider a few things like how complex your concept is, how many resources you have and whether you know what you want to measure already.

Also see Fig 1 from J. Grimmer and Stewart (2013) and Figure 1 in Boumans and Trilling (2015)

# Why choose Lexicon/Dictionary-based Approaches?

- Fully transparent even without technical knowledge
- Lightweight to run, even on enormous data sets
- Easy to implement it for nonconsumptive research (e.g., Google Books let's you search, but not read/consume books)
- Valid choice under 3 conditions (Atteveldt, Trilling, and Arcíla (2021)):
    1. Variable we want to code is manifest and concrete rather than latent and abstract: names of actors, specific physical objects, specific phrases, etc., rather than feelings, frames, or topics.
    2. All synonyms to be included must be known beforehand.
    3. And third, the dictionary entries must not have multiple meanings.

# Examples


OP1N1ONOP
P1N1ONOP1N1
1ONOP1N1ONO
NOP1N1ONOP1
NOP1N1ONO

# 1. Non-Consupmtive Research with Google Books

Taken from Duneier (2017): Ghetto: The Invention of a Place, the History of an Idea

**RQ**: How did the meaning of ghetto change over time? **Method**: Non-Consumptive Research with the Google Books Ngram Viewer

# Exercise 1: Create your own plot

- see notebooks

# 2. Sentiment Analysis

This part is taken from Atteveldt, Trilling, and Arcíla (2021) Chapter 11.2.

- see notebooks

# Exercise 2: Discuss the results

- Based on the example text below, which issues do you see arise from this approach to measuring sentiment?

Hm. Where do I start? I usually ignore whatever rating IMDb has when looking up a movie because I think I might like it anyway or whatever and I should at least give it a chance, but this time I wish I'd paid attention.

I know some people liked it, and I'm not trying to say that they shouldn't. It was semi-amusing at some parts. But if you're like me and you don't like watching cats prancing around in the undergrowth for 20 minutes, random fast motion cloud scenes, dogs barking in cages for another 20 minutes set to 'thrilling' music, and close-ups of faces while people are speaking, then you might want to avoid this movie. The actors were either positively wooden or way over the top, and the film quality was awful, fuzzy and grainy and bland and not in an artistic way at all. And I know that we were supposed to think that Carol was not just a crazy maniac with a gun shooting innocent people with this weird religious psychosis going on, but... well, she doesn't really convince me otherwise. In fact, I ended up really disliking her crazy character. And what was up with the souls in space? I understand this is a fantasy movie, but come on.

I will say, the angel at the end was freaking creepy. It was the creepiest thing in the whole movie, WAY more creepy than the Darth Maul lava-face demon. I give them props for that scene, it was good. But not good enough to actually see the movie. And the opening credits were great, but don't be fooled! I would've rather they used whatever money went into those credits to make the movie better.

Bah. I wish I hadn't bought this for even the $2 that I paid for it, I could have bought a candy bar instead. :/

- This COST action is about opinions. Would you say:
  a. the provided data contains opinions?
  b. the approach we looked at is suitable to measure these opinions?

# 3. Build your own dictionary

- Coming up with a comprehensive list of words that describe the cateogries of interest is hard for humans
- But easy for computers! (King, Lam, and Roberts 2017)
- You can annotate or otherwise categorise documents and see which words are unique (or much more common) in one set
- Justin Grimmer, Roberts, and Stewart (2022) call this “fictitious prediction” problem: “The goal isn’t the prediction itself, but identifying the words that are the most effective predictors.”

# Some Research Examples

# Dictionary Use for API Data Retrieval

For Langer and Gruber (2021) we retrieved data from Twitter. For that we had to build an extensive dictionary. You can find that here:

https://journals.sagepub.com/doi/suppl/10.1177/1940161220925023/suppl_file/sj-pdf-1-hij-10.1177_1940161220925023.pdf

# Actor Identification

If you concept is concrete, dictionaries can still be the best option. Like in Langer and Sagarzazu (2018), where simple dictionaries were used to identify individual politicians and the party that they belong to. The research question whether individuals were mentioned more often in news than their parties can be answered using a dictionary in a valid way and given that this works quickly and spending very few resources, I consider it the optimal method.

# Issues

- The more terms we add to our dictionary, the more false positives we will get
- Building a good dictionary is a lot of work (complexity-resource plot):
  - Negation and bag-of-word issues ("not good" will be counted as positive + modifiers such as "very good")
  - "great" should be more positive than "good"
- Negative image of dictionaries in academia
  - Many negative examples where dictionaries were applied often outside of the domain they had been developed
  - Wrong believe that popular off-the-shelf dictionaries do not need validation
  - Many papers that show that dictionaries do not perform as well as machine learning: e.g. Van Atteveldt, Van Der Velden, and Boukes (2021); González-Bailón and Paltoglou (2015); Boukes et al. (2020)

Now that you know about dictionaries, remember to apply them only under some

circumstances:

0. When no other method is available, e.g., in data retrieval or nonconsumptive research
1. Variable we want to code is manifest and concrete rather than latent and abstract: names of actors, specific physical objects, specific phrases, etc., rather than feelings, frames, or topics.
2. All synonyms to be included must be known beforehand.
3. And third, the dictionary entries must not have multiple meanings.

# References


Atteveldt, Wouter van, Damian Trilling, and Carlos Arcíla. 2021. *Computational Analysis of Communication: A Practical Introduction to the Analysis of Texts, Networks, and Images with Code Examples in Python and R*. Hoboken, NJ: John Wiley & Sons. https://cssbook.net.

Boukes, Mark, Bob Van De Velde, Theo Araujo, and Rens Vliegenthart. 2020. "What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools." *Communication Methods and Measures* 14 (2): 83–104. https://doi.org/10.1080/19312458.2019.1671966.

Boumans, Jelle W., and Damian Trilling. 2015. "Taking Stock of the Toolkit." *Digital Journalism* 4 (1): 8–23. https://doi.org/10.1080/21670811.2015.1096598.

Duneier, Mitchell. 2017. *Ghetto: The Invention of a Place, the History of an Idea*. First paperback edition. New York: Farrar, Straus; Giroux.

González-Bailón, Sandra, and Georgios Paltoglou. 2015. "Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources." {SSRN} {Scholarly} {Paper}. Rochester, NY. https://papers.ssrn.com/abstract=2558788.

Grimmer, J., and B. 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.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton Oxford: Princeton University Press.

King, Gary, Patrick Lam, and Margaret E. Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61 (4): 971–88. https://doi.org/10.1111/ajps.12291.

Langer, Ana Ines, and Johannes B. Gruber. 2021. "Political Agenda Setting in the Hybrid Media System: Why Legacy Media Still Matter a Great Deal." *The International Journal of Press/Politics* 26 (2): 313–40. https://doi.org/10.1177/1940161220925023.

Langer, Ana Ines, and Iñaki Sagarzazu. 2018. "Bring Back the Party: Personalisation, the Media and Coalition Politics." *West European Politics* 41 (2): 472–95. https://doi.org/10.1080/01402382.2017.1354528.

Van Atteveldt, Wouter, Mariken A. C. G. Van Der Velden, and Mark Boukes. 2021. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms." *Communication Methods and Measures* 15 (2): 121–40. https://doi.org/10.1080/19312458.2020.1869198.