# Richer Corpus Annotation

Digital Humanties for Computational Linguists

Nerbonne, Tübingen

WS 2023-2024

Week 2-3

# Old desideratum in DH

- Annotations revealing who is speaking, who is being addressed, when, where, …
  - Needed to support literary analysis, e.g., Hamlet’s speech vs. Ophelia’s, Polonius’s, …
  - Standard Generalized Markup Language (SGML)
  - Text Encoding Initiative (TEI)
- SGML was technically complex (difficult to parse), TEI based on XML, simplified SGML
- Parlamint based on TEI

# TEI looks like HTML

```
=== Hallo Welt! ===
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Hallo Welt!</title>
            </titleStmt>
            <publicationStmt>
                <p>Demo für Wikipedia</p>
            </publicationStmt>
            <sourceDesc>
                <p>Originales Werk, keine Vorlage</p>
            </sourceDesc>
        </fileDesc>
    </teiHeader>
    <text>
        <body>
            <p>Hallo Welt!</p>
        </body>
    </text>
</TEI>
```

# TEI is flexible

- DIGITAL HUMANISTS annotate for research purposes
- Used a lot for literary digitization projects
  - tei-c.org/activities/projects/
    - Deutsches Textarchiv, www.deutschestextarchiv.de/
    - Dig. Bibl. Nederlandse Letteren, dbnl.org
    - Eng. Poetry DB, Virginia
    - Nat’l Corpus, Poland, Slov. lit., Old French, …
    - Specialized corpora, e.g., Emily Dickinson
  - Still difficult anticipating research needs
- Unexpectedly useful elsewhere

# Richer than n-gram viewer

- Google n-gram view does make use of metadata, especially date of publication
- But TEI goes further

# Parliamentary proceedings

- Often digitally archived (somehow!) by requirement of law
- ParlaMint (CLARIN project) gathered corpora of different EU countries
  - … and converted all to common format
  - Based on TEI!
    - Erjavec, T., et al. (2022) "The ParlaMint corpora of parliamentary proceedings." *Language Resources and Evaluation.* 1-34.
- Demonstrates added value of rich annotation

# Parlamint

- Common Language Resources and Technology Infrastructure (CLARIN) sponsored, 7/20-5/21
  - www.clarin.eu/content/parlamint
  - Modest funding, but 17 parliaments, 16 lg., $5\times10^8$ wd.!
  - Goal – support observation & analysis
    - Special focus on COVID-19
- Two CONCORDANCERS (NoSketch/KonText)
  - … showing a word in different contexts
    - … Social Mobility and Child Poverty Commission identified the 30 …
    - … near enough on fuel poverty , and I want to
- Data & programs on GitHub

# Construction

- Differed in diff. countries
  - Scraping websites, retrieving from existing corpora, using an API (Croatia), downloading from servers, …
  - Converting to TEI XML, or through XLST, or via conversion scripts
  - Correcting (a bit)
  - Preprocessing via NLP, incl morphosyntactic annotation and NER
- Important goal: support of comparisons among parliaments, e.g., during epidemics
- Future: linking to resources such as Wikipedia

# (Preliminary) studies

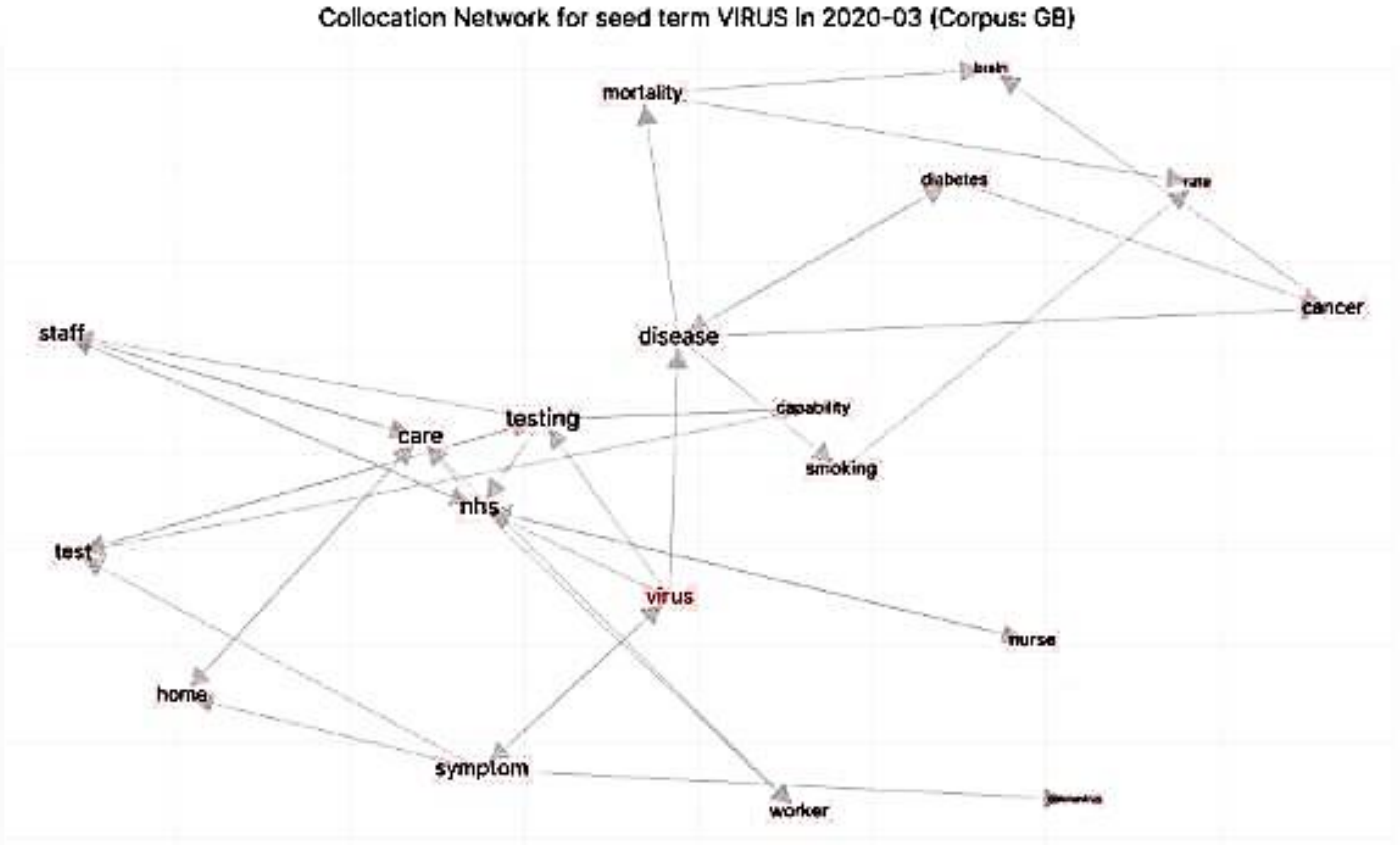
- Marta Kołczyńska “Parliamentary debates in COVID”
  - Report from CLARIN DH Hackathon
  - https://dhhackathon.wordpress.com/2021/05/28/parliamentary-debates-in-the-covid-times/

- Taking slides from Kołczyńska’s presentation

# Collocation network

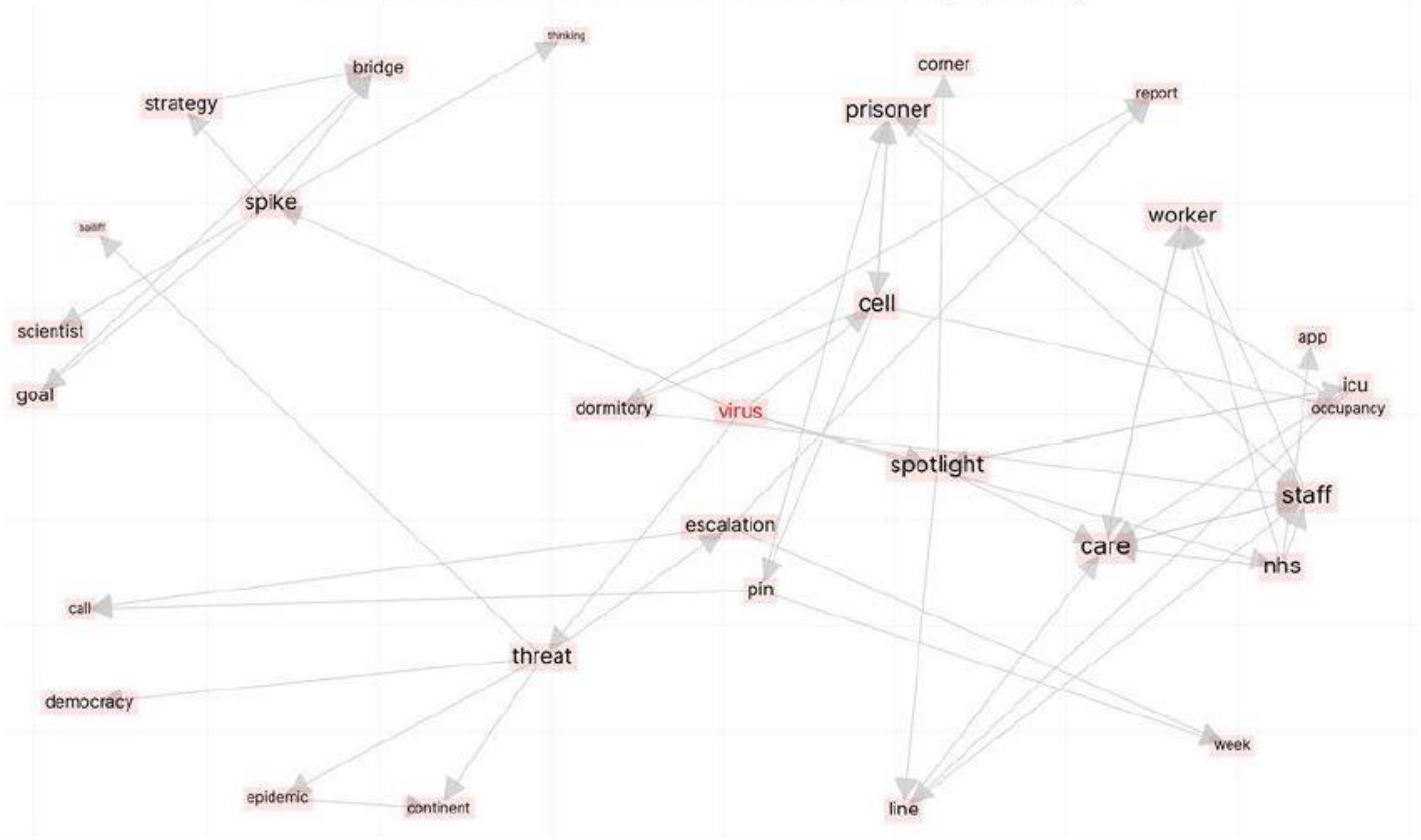
- Collocations: word pairs that tend to co-occur
  - Within a given range (#words)
  - Different in frequency
    - *love* w. *in* (*love* very likely to be preceded by *in*)
    - *affair* w. *love (affair* likely to be preceded by *love)*
  - Different in direction
    - *in* w. *love* (*in* is less likely to be near *love*)
    - *love* w. *affair (love* less likely to be near *affair)*

- Tool in #LancsBox



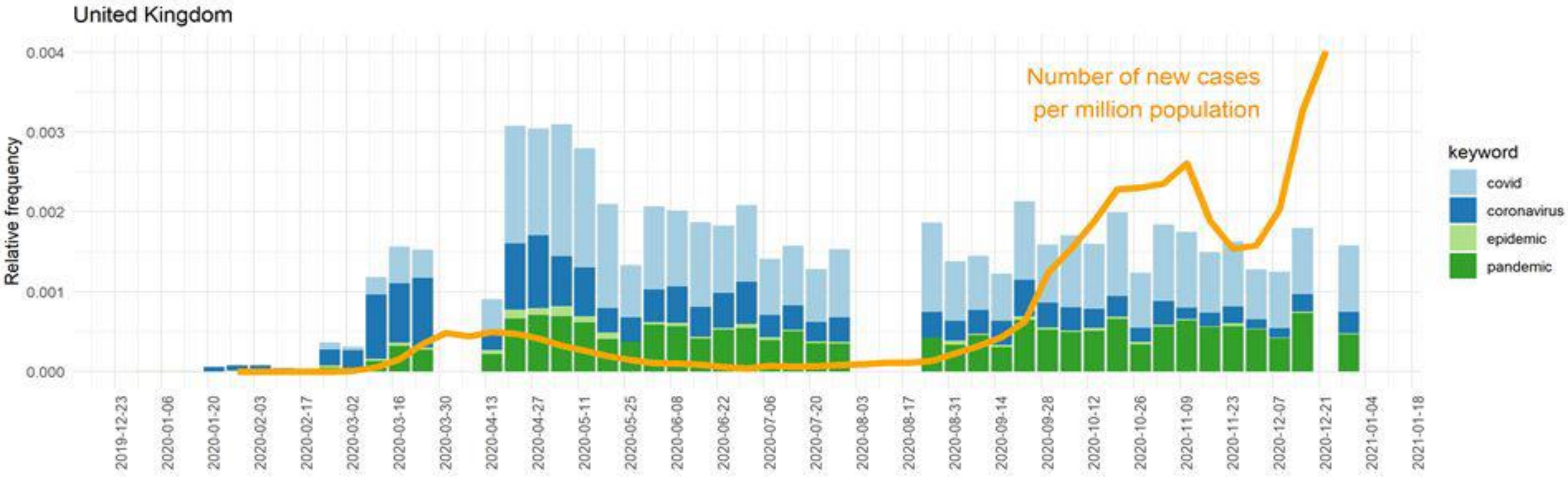
Collocation Network for seed term VIRUS in 2020-03 (Corpus: GB)

Collocation Network for seed term VIRUS in 2020-04 (Corpus: GB)



# <u>How much</u> was <u>what</u> discussed?

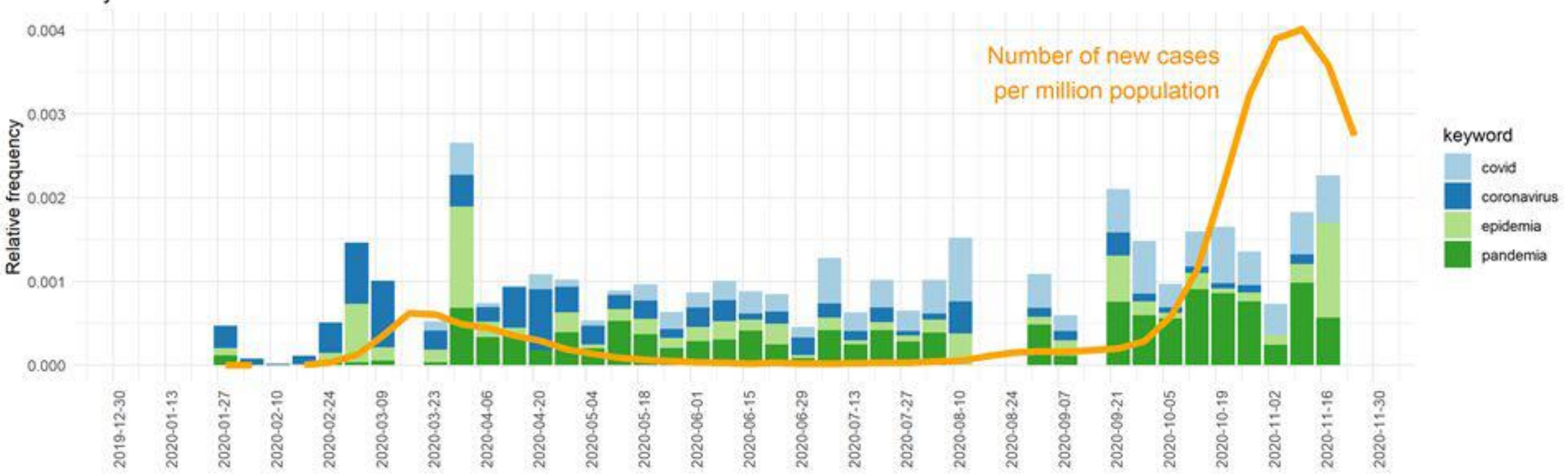
- › Simple word frequency (as in Culturomics)
- › Overlaying epidemic incidence



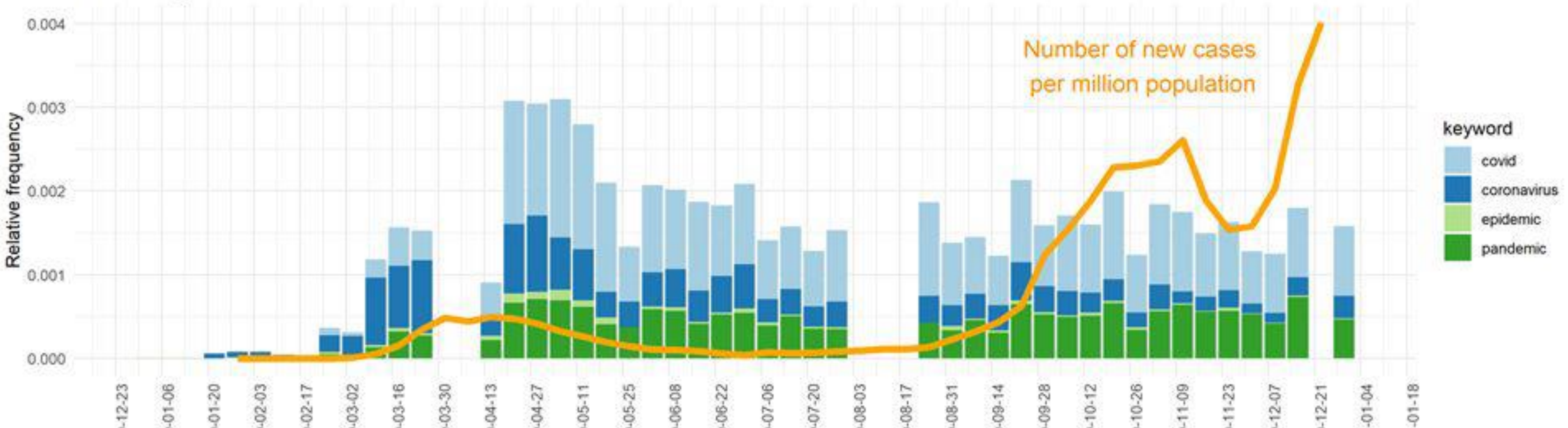
# Comparisons possible

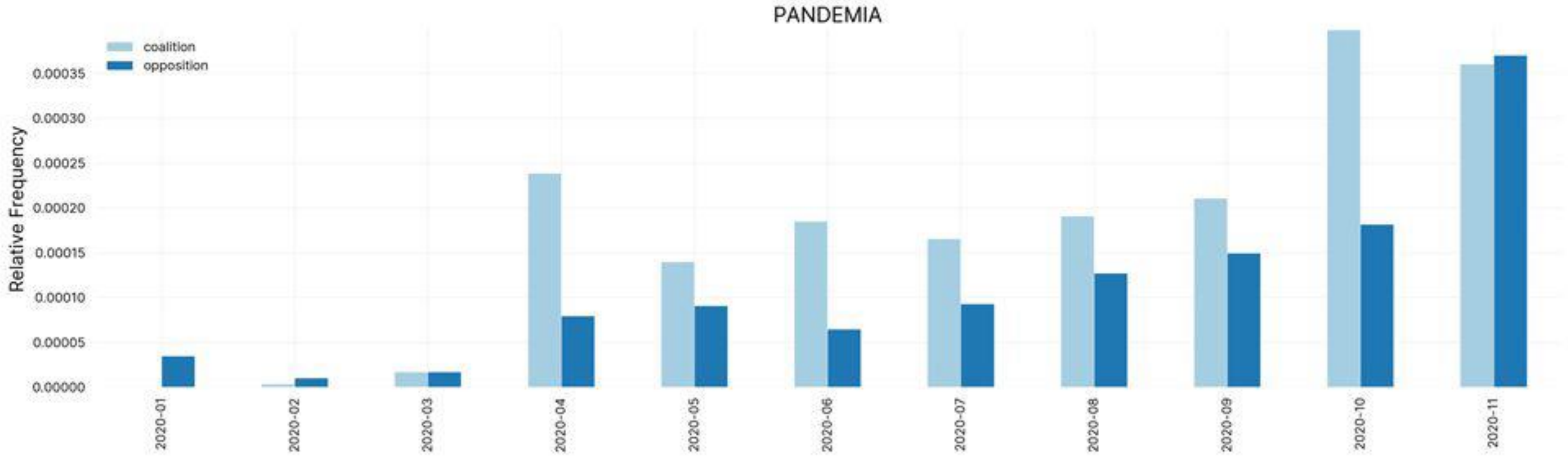
- Between countries (Italy vs. UK)
- Between governing parties and opposition (Italy)
- Between men and women in parliament (from Miguel Pieters’ Master thesis, Data Science, Amsterdam)

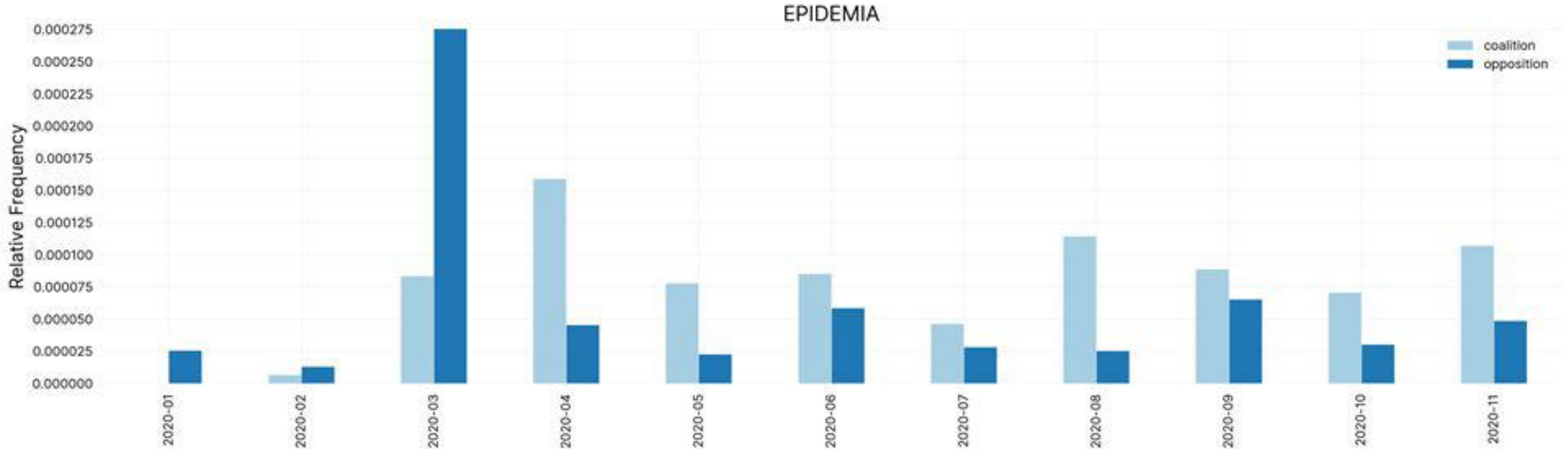
Italy



United Kingdom

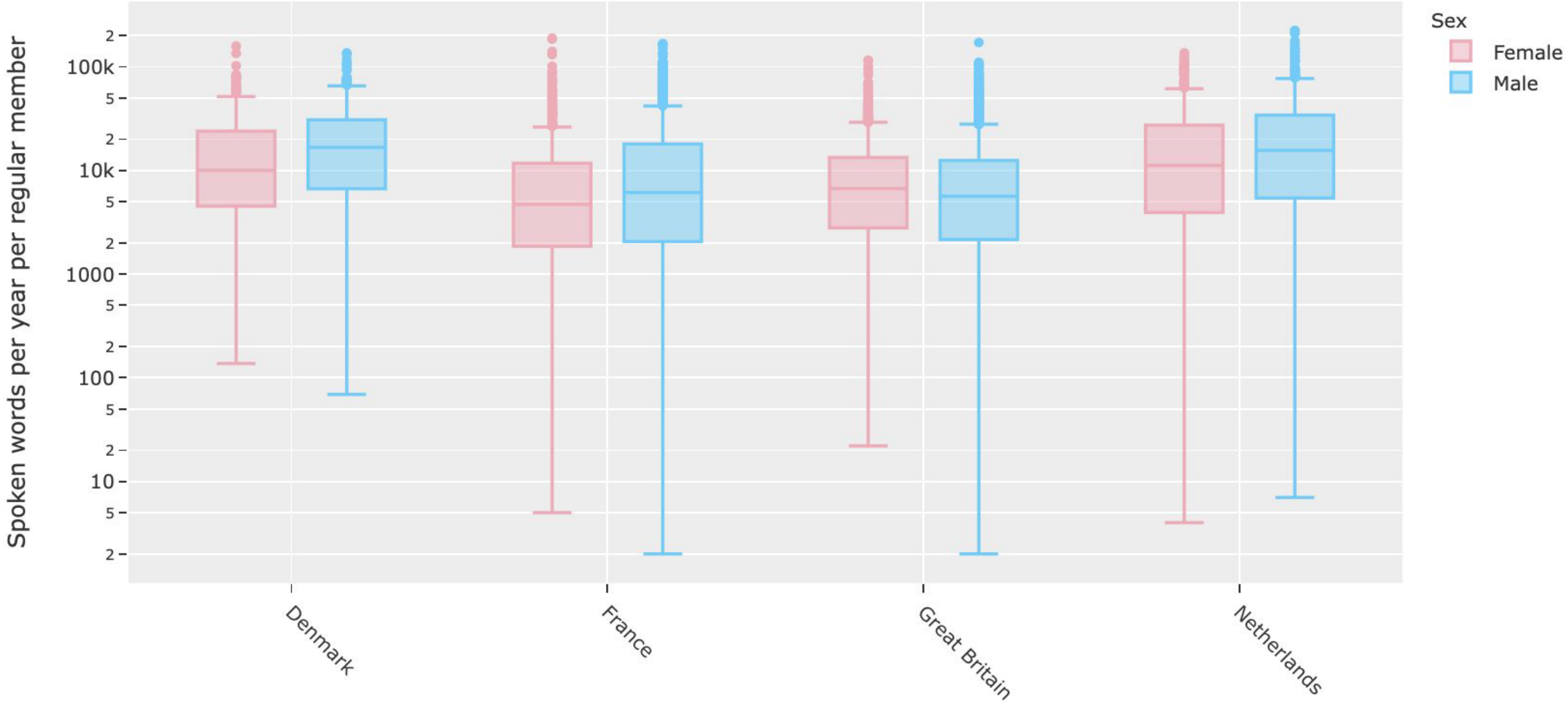



PANDEMIA
coalition
opposition
Relative Frequency
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11


EPIDEMIA
coalition
opposition
Relative Frequency
0.000275
0.000250
0.000225
0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11

# Men vs. women

Words/yr/gender, regular members (not ministers)



# Tutorial on using Parlamint

- Darja Fišer & Kristina Pahor de Maiti
  - https://sidih.github.io/voices/index.html
  - Using Slovenian debates as an example
  - People interested in this sort of work or these sorts of questions, should follow the tutorial.

- Lots of queries involve creating SUBCORPORA, for more subtle questions, this will be needed

# Richer Corpus Annotation

- Includes information on speakers (or writers), place/time of speech, structure of document, …
  - Enables richer automatic analysis
- ParlaMint is a corpus of 17 EU Parliament Proceedings, annotated for speakers’ status (MP vs. minister), party, role as part of governing party or opposition, and gender
  - In a TEI derivative, a rich annotation system