ANALYSING OPEN-ENDED SURVEY RESPONSES IN FINNISH

Adeline Clarke and Maria Valaste

BNU workshop on Survey Statistics
26-30 August 2024, Poznań, Poland

CONTENTS

- Introduction
  - Background and motivation
- Finnsurveytext package
  - Demos

Presentation materials (including demo code):

https://github.com/DARIAH-FI-Survey-Concept-Network/Workshop-on-Survey-Statistics-2024_finnsurveytext

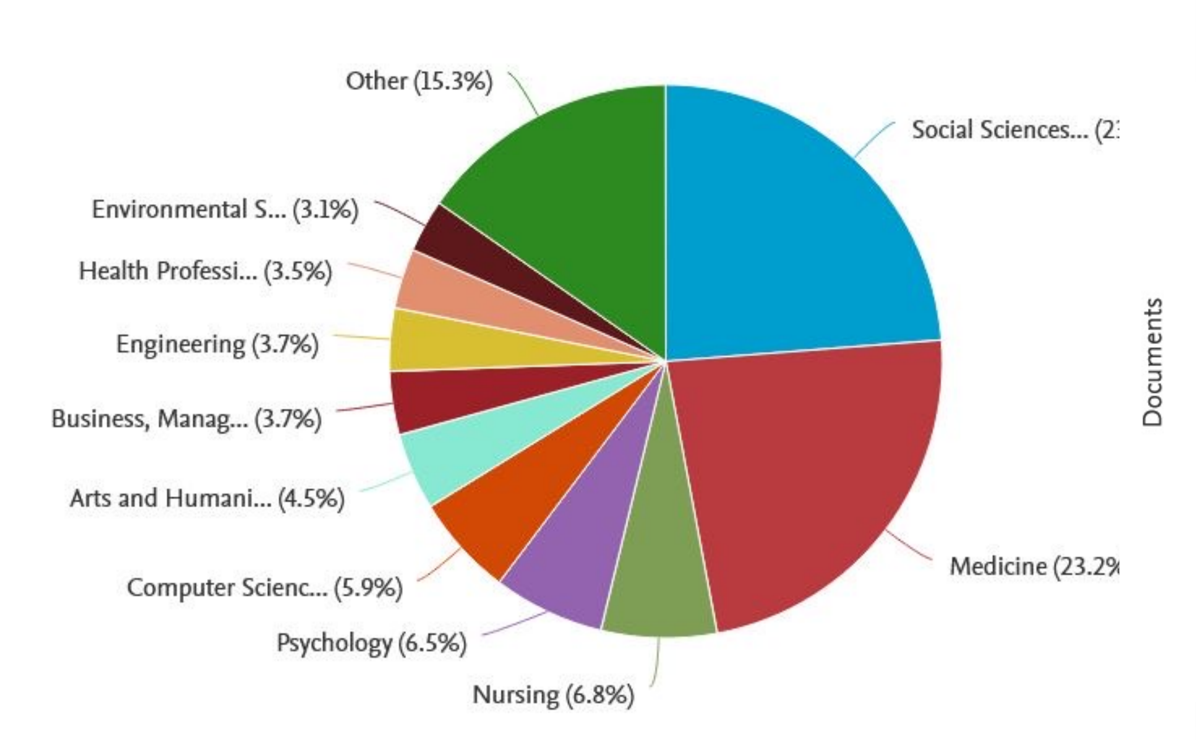
MOTIVATION

- Open-ended questions are an **important** but **challenging** way to obtain informative data in surveys.
    - Open-ended question data usually requires **extra time investment** (Fielding et al., 2013), but open-ended questions are particularly useful if researchers do not want to constrain respondents' answers to **pre-specified selections**. Open-ended questions allow respondents to provide diverse answers based on their experience, and some answers are probably never thought of by researchers. (He & Schonlau, 2021.)
    - *Hypothesis: Sometimes these divergent experiences may bring to view completely new, emerging societal phenomena.*
- There's limited support for conducting qualitative analysis on **Finnish** open-ended survey responses, so open responses tend not to be utilized properly
- Our aim is to build tools for text data that work with Finnish language with sufficient ease and to support explorative analysis of open responses
    - Integrating tools with R workflows
    - R package to visualize, describe and analyze

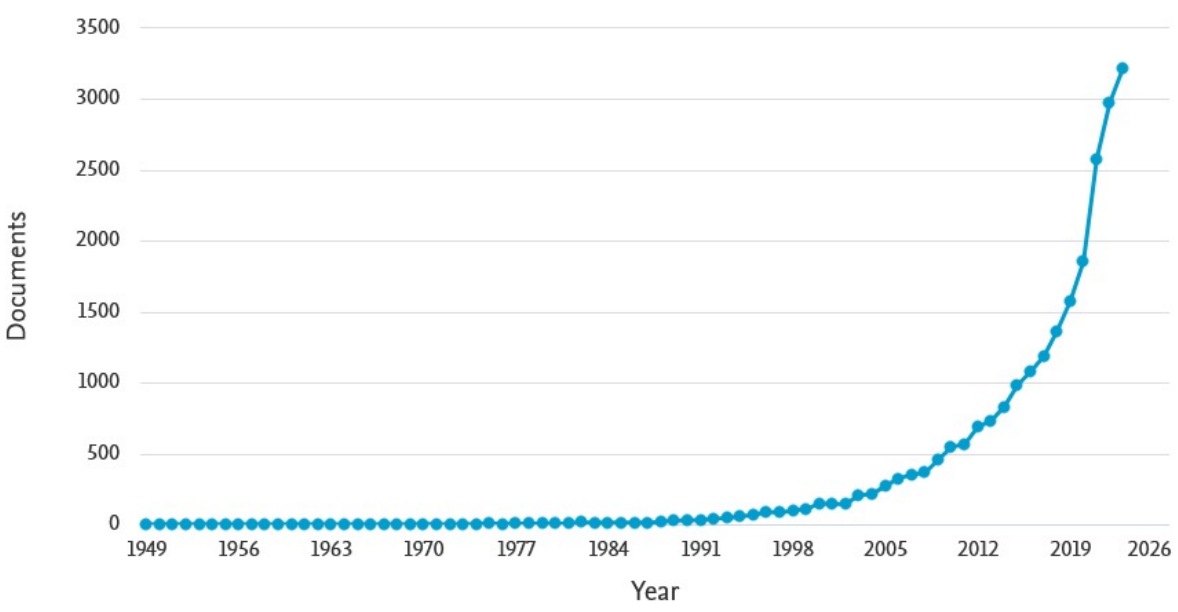
BIBLIOMETRIC ANALYSIS

Scopus: all fields includes open+ended and question and survey

Documents by subject area



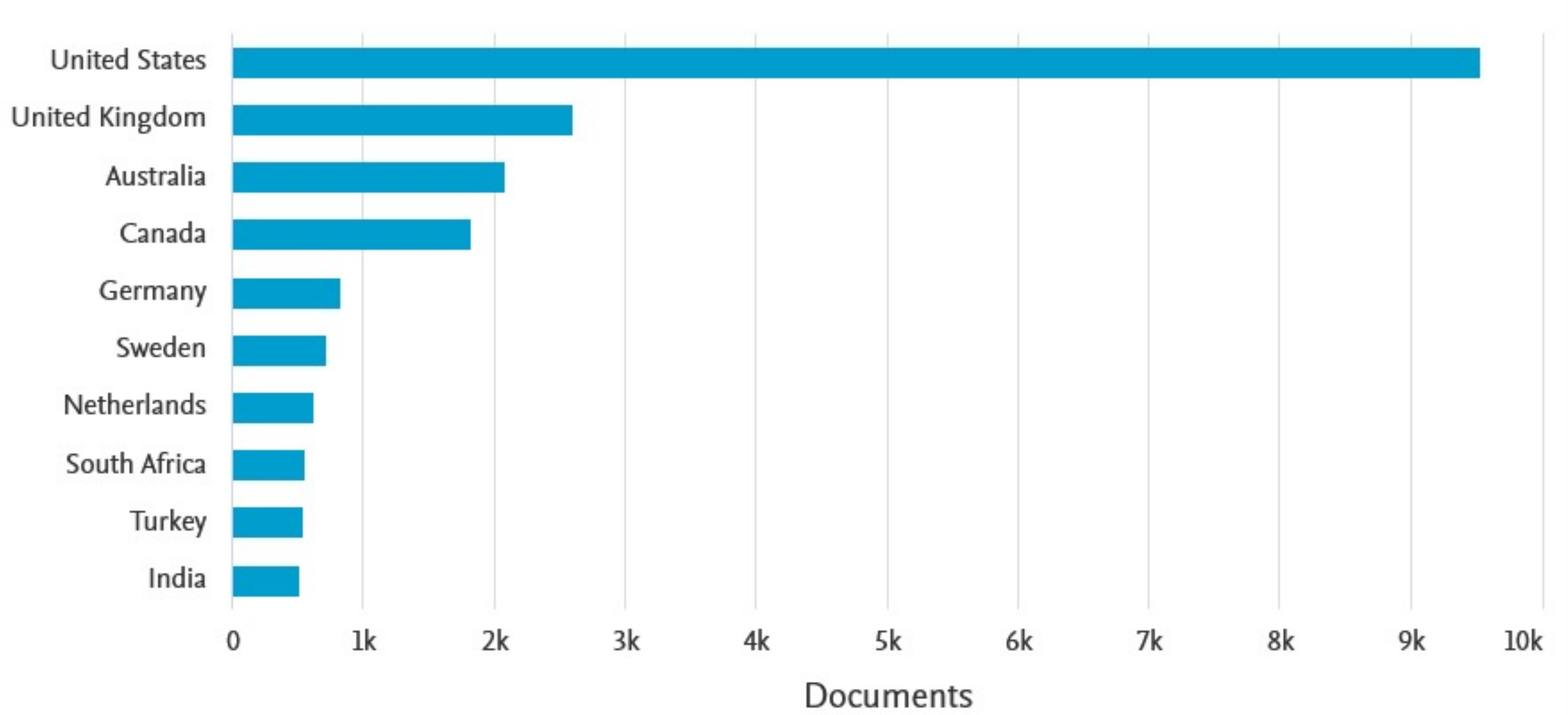
Documents by year, All fields



BIBLIOMETRIC ANALYSIS

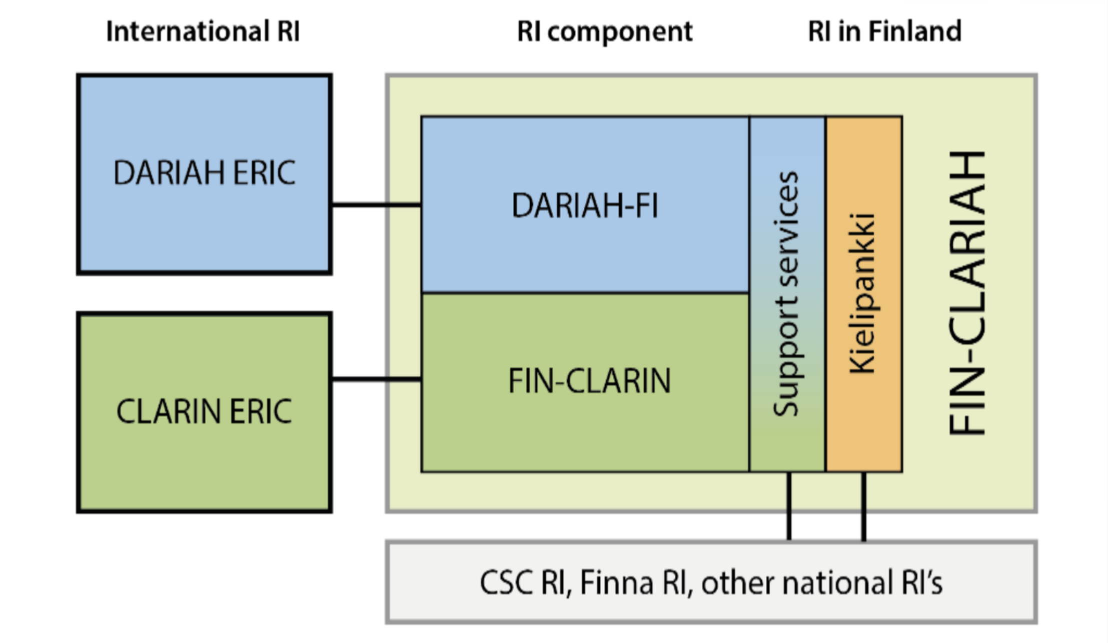
Scopus: all fields includes open+ended and question and survey

By country



BACKGROUND

- FIN-CLARIAH is the premier Finnish digital research infrastructure (RI) for Social Sciences and Humanities (SSH) comprising two components, FIN-CLARIN and DARIAH-FI
- The project involves all Finnish universities with research in SSH
- Project aims is to ensure that a digital transformation happens in an orderly fashion without duplication of efforts or reinventing the wheel
- Funding periods:
  I. 01.01.2022 – 31.12.2023
  II. 01.01.2024 – 31.12.2025



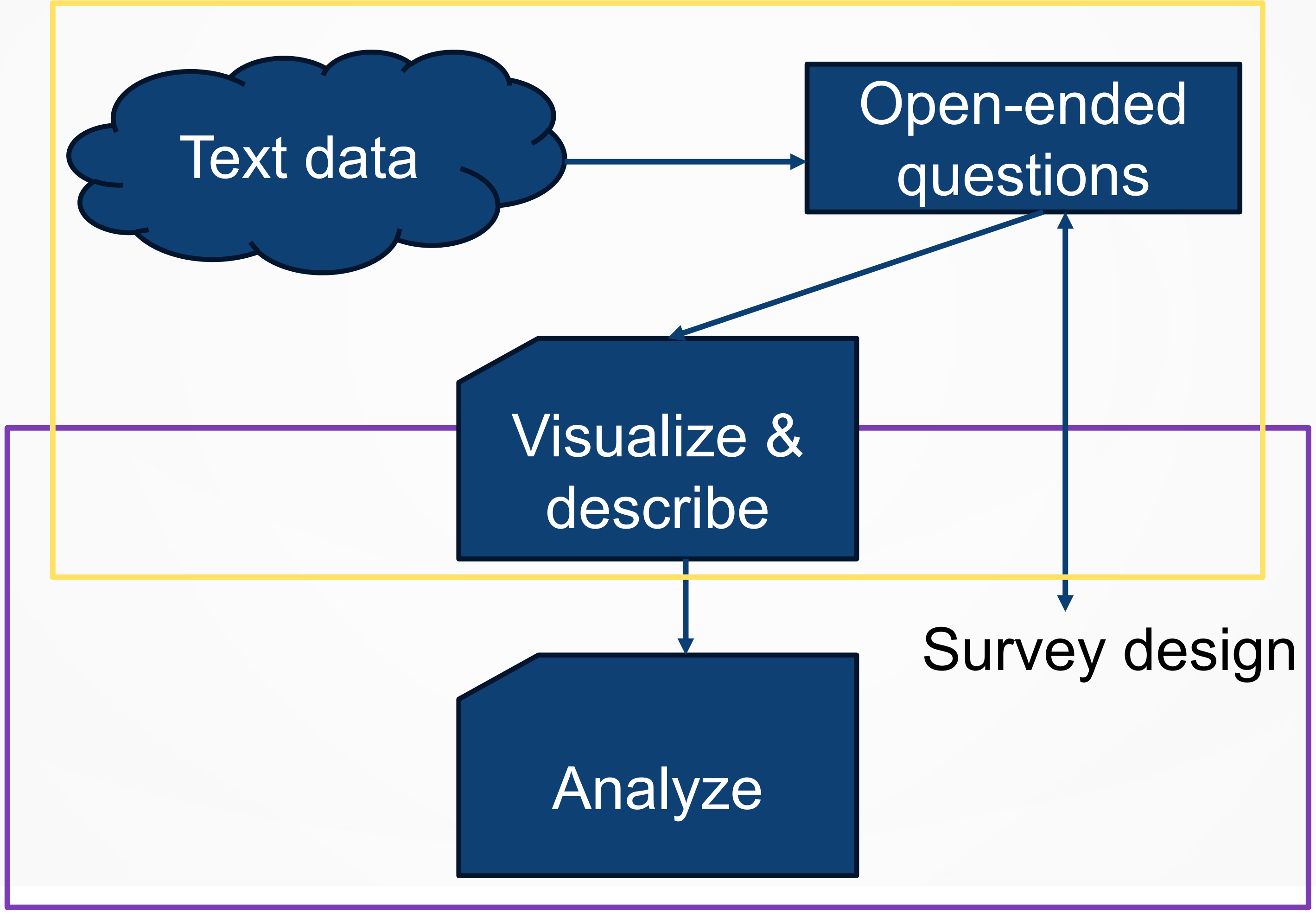
FUNDING PERIODS: 01.01.2022 - 31.12.2023 & **01.01.2024 – 31.12.2025**

01/2022 –
12/2023

finnsurveytext

FST

01/2024 –
12/2025



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

FINNSURVEYTEXT PACKAGE

- The finnsurveytext package can be found on the CRAN here: CRAN: Package finnsurveytext
- Package website: https://dariah-fi-survey-concept-network.github.io/finnsurveytext/
  - The website contains a number of tutorials covering the package including one about using languages other than Finnish with the package: Extra-AnalysingOtherLanguages
- To learn more about *TextRank* – the unsupervised algorithm used to within our Concept Network to rank keywords in responses – you may want to look at paper TextRank: Bringing Order into Text (Mihalcea & Tarau, EMNLP 2004)
- The released version of finnsurveytext can be installed from the CRAN: install.packages("finnsurveytext")

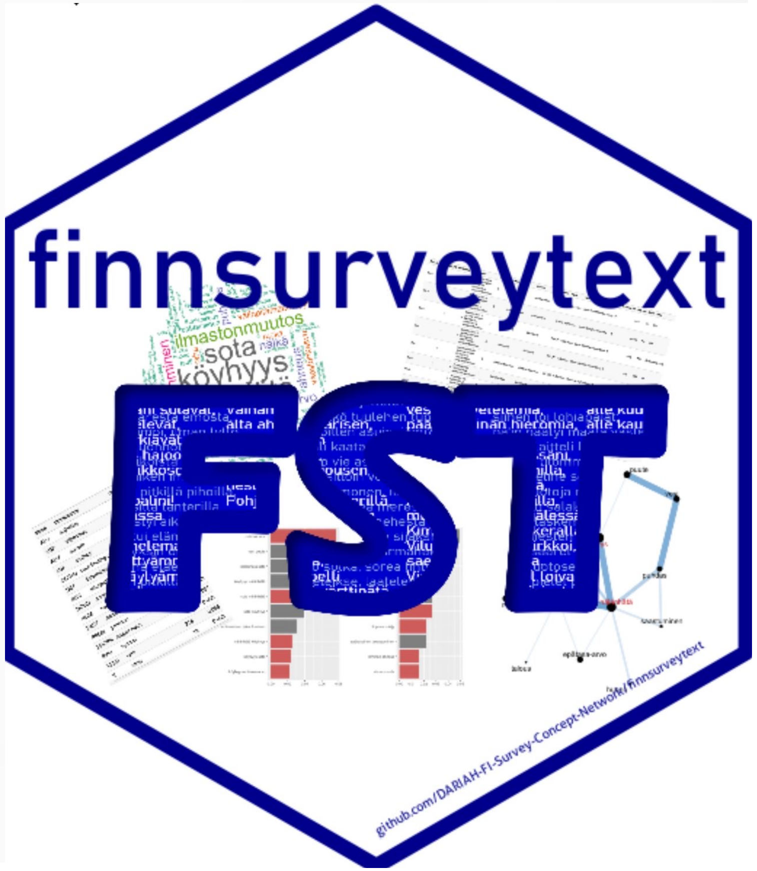
HOW TO USE THE PACKAGE

Install and load package from CRAN:

```
> install.packages("finnsurveytext",
                    type = "source")
> library(finnsurveytext)
```

Run (BETA) RShiny app:

```
> runDemo()
```



FINNSURVEYTEXT DEMOS

- Demo part 1: Introduction, Data Preparation and Summary Tables
- Demo part 2: Wordclouds and N-Grams
- Demo part 3: Concept Network
- Demo part 4: Comparison functions
- Demo part 5: R Shiny App

LANGUAGE-SPECIFIC PARAMETERS FOR FST_PREPARE()

Language	language	model	stopword_list
Estonian	'et'	'estonian-edt' **OR** 'estonian-ewt'	'stopwords-iso'
Finnish	'fi'	'finnish-ftb' **OR** 'finnish-tdt'	'stopwords-iso' **OR** 'snowball' **OR** 'nltk'
Latvian	'lv'	'latvian-lvtb'	'stopwords-iso'
Lithuanian	'lt'	'lithuanian-alksnis' **OR** 'lithuanian-hse'	'stopwords-iso'
Polish	'pl'	'polish-lfg' **OR** 'polish-pdb' **OR** 'polish-sz'	'stopwords-iso'
Ukrainian	'uk'	'ukrainian-iu'	'stopwords-iso'

EXAMPLE FST_PREPARE()

```
df <- fst_prepare(data = survey_data,                 # Reqd
                  question = 'open-ended qn',         # Reqd
                  id = 'ID’,                          # Reqd
                  model = 'polish-lfg',               # Reqd
                  stopword_list = 'stopwords-iso',    # Reqd
                  language = 'pl',                    # Reqd
                  weights ='weight',                  # Optional
                  add_cols = 'col1, col2',            # Optional
                  manual = FALSE,                     # DEFAULT
                  manual_list = ''                    # DEFAULT
)
```

EXAMPLE FST_PREPARE_SVYDESIGN()

```
df <- fst_prepare(svydesign = survey,                  # Reqd
                  question = 'open-ended qn’,          # Reqd
                  id = 'ID',                           # Reqd
                  model = 'latvian-lvtb',              # Reqd
                  stopword_list = 'stopwords-iso',     # Reqd
                  language = ‘lv',                     # Reqd
                  use_weights = TRUE,                  # Optional
                  add_cols = 'col1, col2',             # Optional
                  manual = FALSE,                      # DEFAULT
                  manual_list = ''                     # DEFAULT
                  )
```

If you try the package, we would welcome your feedback.

REFERENCES

Fielding, J., Fielding, N., & Hughes, G. (2013). Opening up open-ended survey data using qualitative software. Quality & Quantity, 47(6), 3261–3276. https://doi.org/10.1007/s11135-012-9716-1.

He, Z., & Schonlau, M. (2021). Coding Text Answers to Open-ended Questions: Human Coders and Statistical Learning Algorithms Make Similar Mistakes. Methods, Data, Analyses, 15(1), Article 1. https://doi.org/10.12758/mda.2020.10.

Rada Mihalcea & Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.