ICAME42 | tu
crossing boundaries through corpora

Corpus Linguistics, Language Data Science, and Computational Linguistics – building bridges or splitting apart?

Martin Schweinberger
www.martinschweinberger.de
m.schweinberger@uq.edu.au

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

LADAL

NORGES ARKTISKE UNIVERSITET • UIT •
Logosymbol.

AcqVA Aurora Lab

Background and Motivation

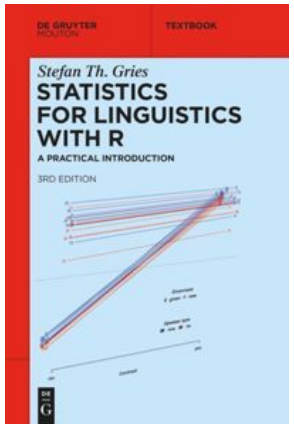
Experiences from consultation (statistics, designs | tools, data management)

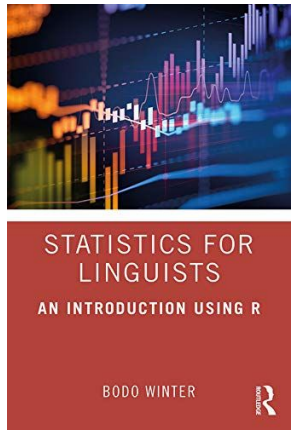
- Everybody's computer is a mess!
  - No | little training (in data management)
- Issues people struggle with
  - Identifying what someone needs is the first step
  - Workflows are often unnecessarily convoluted
  - Untidy data (Organizing data | workflows)
  - Statistics!
  - Automating labor | time intensive tasks

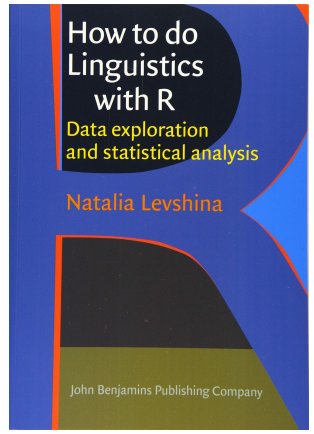
Background and Motivation

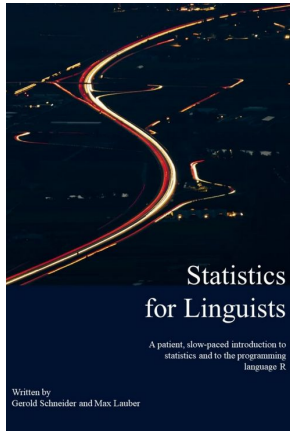
Experiences from consultation (statistics, designs | tools, data management)

- Statistics: myriad of resources
  - Books, Workshops, Bootcamps,
    Online resources (YouTube | StackOverflow |
    Quick-R, etc.)
- What about data management |
  programming | reproducibility?
  - Resources???









Blind Spot: Data Management | Reproducibility?!?

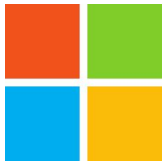
Outline

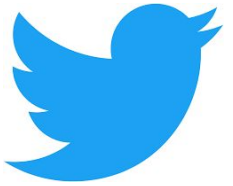
- Computational Revolution
  - Potential of Computing (in CL)
  - Drawback of Programming (in CL)
- Replication Crisis | Issue | Problem
- Training infrastructures & Case study (re: computing)
  - LADAL (Language Technology and Data Analysis Laboratory, UQ)
  - TCC (Text Crunching Center, UZH)
  - Case study: COVID-19 in the Australian Twittersphere
- Outlook (wild speculation!)
  - Split between data/methods focused CL and more discourse oriented CL?

The Computational Revolution

Computation is becoming increasingly important

- Computational revolution has changed all domains of life





- Computational approaches to processing, transforming, analyzing, and visualizing text | language data are becoming increasingly prevalent in the economy and in the humanities

Search engines

Machine translation

- Despite the quantitative turn and the computational revolution, HASS has been reluctant to integrate computational skills | programming

Text-2-Speech | Speech-2-Text

Voice recognition

Named Entity Recognition

Chat bots | question answering

Spelling correction

Content detection | summarization

History of (Programming in) CL

Corpus Linguistics represents a (early) result of the computational revolution

- Came into being when machine-readable texts became available for analysis
- Allowing to empirically test models/theories based on natural language
- Different phases (Anthony 2020)
  1. 1960s: Programming to extract concordances from texts
  2. 1980s: Ready-made user friendly tools become available (less need for programming)
  3. 2000s: Web-based corpora with in-build corpus linguistic tools
  4. 2010s: Quantitative turn (revival of computation via statistics & R...well, somewhat)

Potential of Computing

Potential of computational methods for CL (as a field)

- **New avenues for research**
  - **Innovative methods** (statistical methods, e.g., MuPDARF)
  - **Big data | multimodal data** (e.g., Trove, https://trove.nla.gov.au/)
  - **Collaboration** (interactive | multiauthor, e.g., GitHub, GoogleDocs)
  - **Reproducibility | Transparency** (e.g., GitHub, OSF)
- **Potential drawbacks**
  - Move towards application | engineering | technology
  - Move away from humanities
  - Shift in focus from language (what) to how (method)

TROVE

Billions of pieces of information: digital copies of newspapers, Government Gazettes, maps, magazines and newsletters, books, pictures, photographs, archived websites, music, interviews, letters, diaries and personal archives.

OSF

Open Science Foundation
Free, open source web environment enabling scientists to collaborate, document, archive, share, and register research projects, materials, and data.

Potential of Computing

Potential of computational methods for CL (researchers)

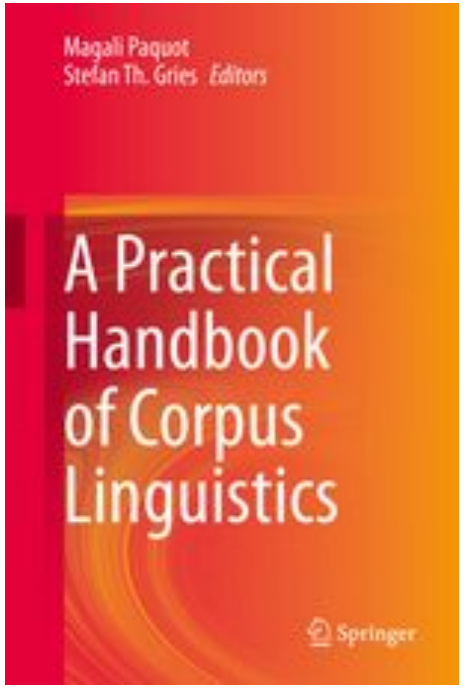
- **Versatility** of what one can do (driver's seat, Gries 2009)
- **Applicability** of skills to other domains

  (across disciplines |employability in the private sector)
- **Reproducibility** (ability to make research practices more transparent | efficient)

Quantitative turn ≠ Computational turn

- Corpus Linguistics is inherently | increasing based on frequency and probabilistic information (both in terms of theory and methodology)
- Quantitative methods are pervasive : programming is not (to the same extent)

Drawbacks for Programming



Why is programming not an integral part of Corpus Linguistics?
(e.g. Anthony 2020)

- User friendly and (web-)integrated tools limit the need to acquire programming skills
- Time limits: little time to acquire additional skills
- Teams (outsourcing)
- Interests and methodology (small data sets, fine-grained qualitative analyses, manual processing)

Computation vs Programming

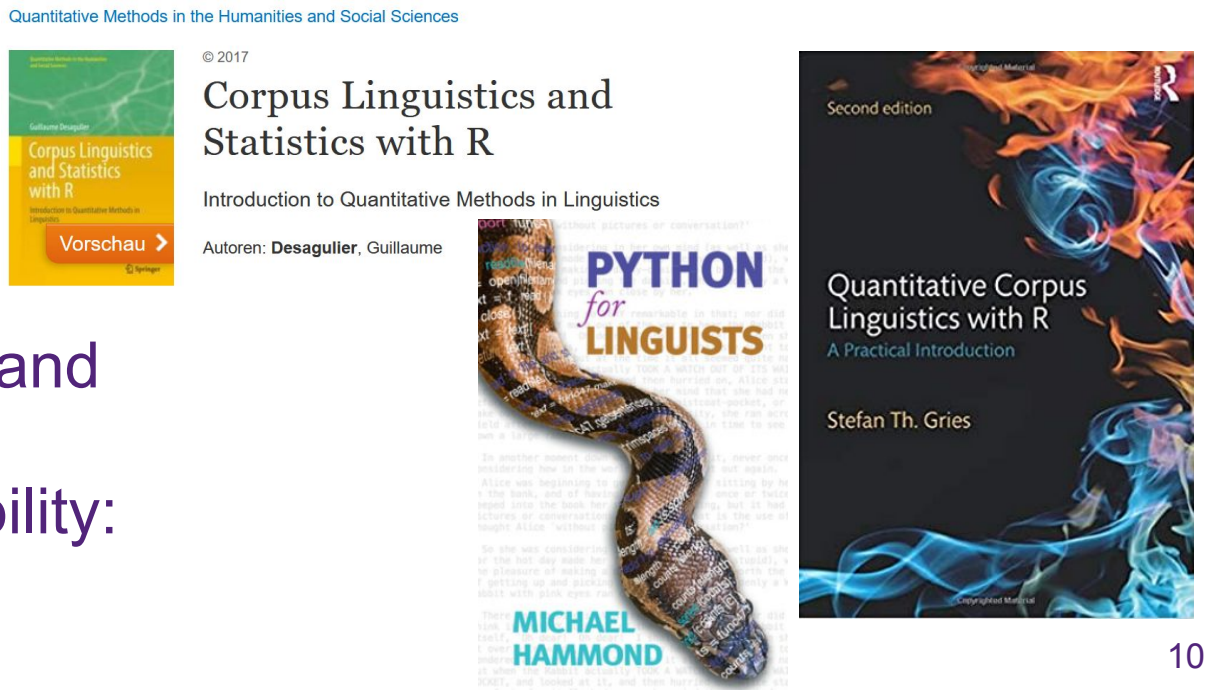
Computation, as used here, refers to the use of computers going beyond user interfaces (point & click, drag & drop tools) which includes programming as well as the integration of environments, practices, or platforms common in workflows in Computer and Data Science.

Programming in CL

So, there are very good reasons for why Corpus Linguists have not fully endorsed programming (or computational skills more generally - at least not beyond digital tools).

But what other motivations could there be that have so far received rather limited consideration?

- **Lack of training** | upskilling resources (compared to statistics)
- **Reproducibility issues** (programming allows research to keep and exact track of any analysis's steps and enables swift uncomplicated reproducibility: reproduction at the press of a button)



Replication Crisis | Issue | Problem

Replication Crisis | Issue | Problem

Controversial ongoing methodological crisis that originated in medicine (Ioannidis 2005) and swiftly expanded to STEM, the social sciences, and psychology when **replications of seminal experiments failed** - calling into question the reliability of widely accepted published research

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown.
(Open Science Collaboration 2015)


nature

Explore content ⌄ Journal information ⌄ Publish with us ⌄

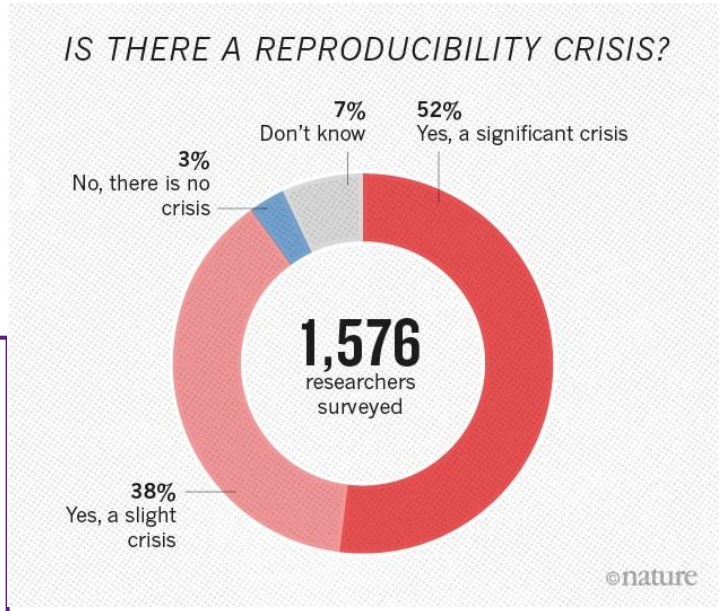
nature › news feature › article

Published: 25 May 2016

1,500 scientists lift the lid on reproducibility

Monya Baker

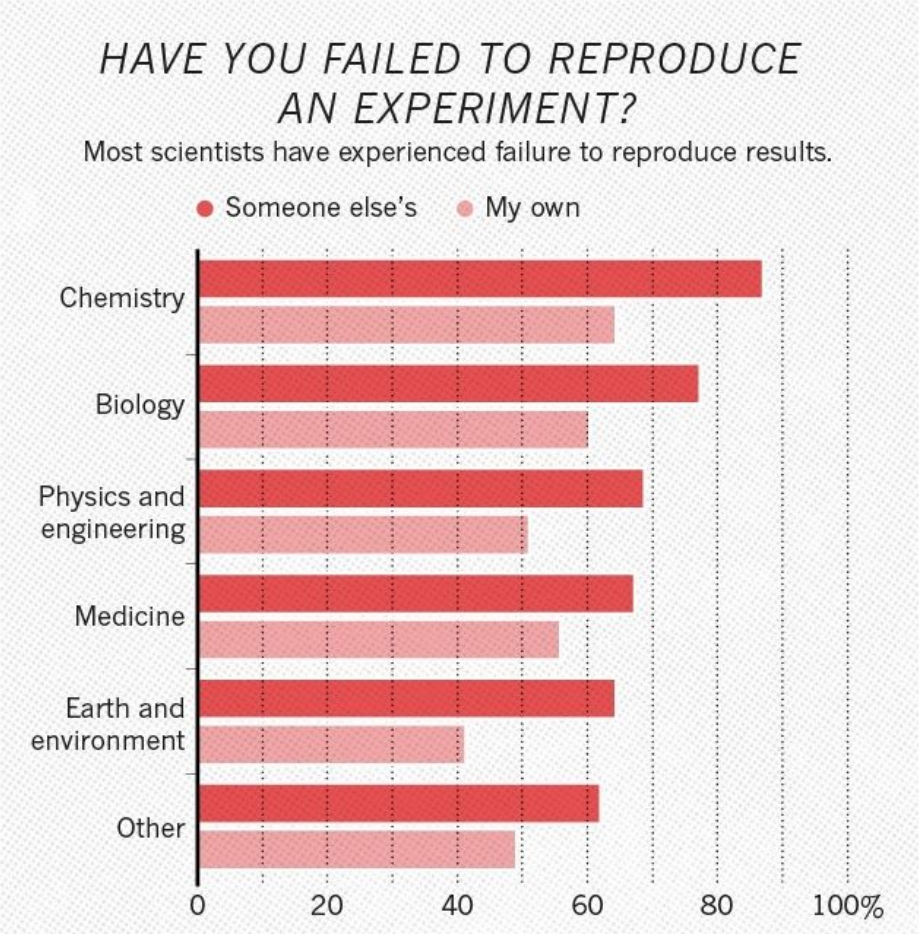
Nature **533**, 452–454 (2016) | Cite this article



"More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments."
(Baker 2016: 452)

Replication Crisis | Issue | Problem

from Baker (2016: 452)



from Baker (2016: 452)

nature

Explore content | Journal information | Publish with us

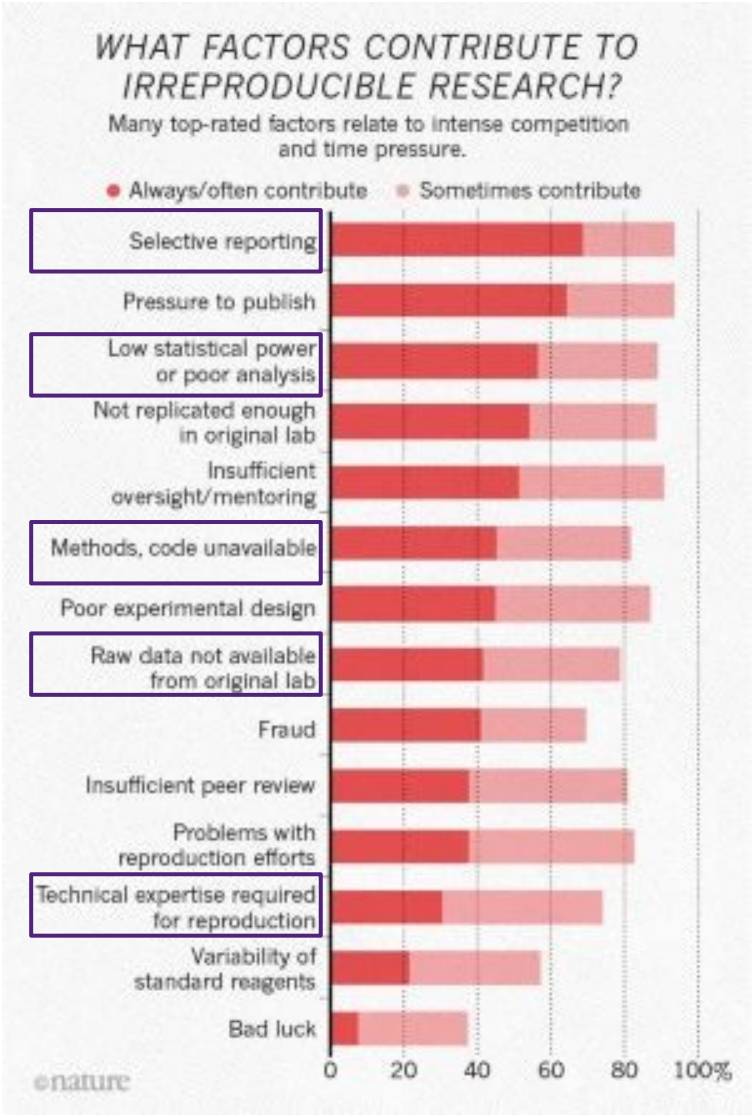
nature > news feature > article

Published: 25 May 2016

1,500 scientists lift the lid on reproducibility

Monya Baker

Nature **533**, 452–454 (2016) | Cite this article



Replication Crisis | Issue | Problem

Reproducibility

- To reproduce a study means doing the things to the same data to get the exact same results.

Replication

- Replicating a study means doing the same | (similar) things to similar data

Robustness

- Robustness refers to the results being consistent | stable across replications

True vs formal reproducibility

True reproducibility means that reproducibility is practically possible and supported while formal reproducibility means that reproduction is possible in principle but hindered by real-world restrictions (data only accessible in a specific lab | study based on blackbox tools or is accompanied by spreadsheets not code)

NOBA

Browse Content / The Replication Crisis in Psychology

The Replication Crisis in Psychology

By Edward Diener and Ro
University of Utah, Univer

American Psychological Association

MEMBERS TOPICS PUBLICATIONS & DATABASES PSYCHOLOGY HELP CENTER NEWS & EVENTS

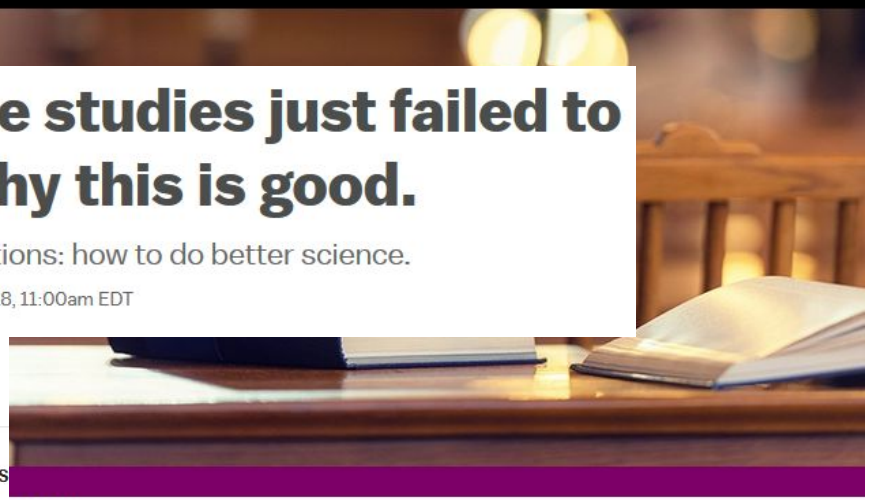
Home // Monitor on Psychology // 2015 // 10 // A reproducibility crisis?

A reproducibility crisis?

The headlines were hard to miss: Psychology, they proclaimed, is in crisis.

October 2015, Vol 46, No. 9
Print version: page 39

sciencealert



More social science studies just failed to replicate. Here's why this is good.

What scientists learn from failed replications: how to do better science.

By Brian Resnick | @B_resnick | brian@vox.com | Aug 27, 2018, 11:00am EDT

SCIENTIFIC AMERICAN.

SUSTAINABILITY EDUCATION VIDEO PODCAS

Observations

JMANS

cience's 'Replication Crisis' Has Reached ven The Most Respectable Journals, Report hows

IKE MCRAE 27 AUG 2018

FiveThirtyEight

Politics Sports **Science & Health** Economics Culture

Our 2019 March Madness

DEC. 6, 2018, AT 11:19 AM

Psychology's Replication Crisis Has Made The Field Better

By Christie Aschwanden

(Dis)trust in Science

Can we cure the scourge of n

By Gleb Tsipursky on July 5

PHYS.ORG Nanotechnology Physics Earth Astronomy & Space Technology

Home » Other Sciences » Social Sciences » November 21, 2018

Researcher discusses the the science replication crisis

November 21, 2018 by Emily Velasco, California Institute of Technology

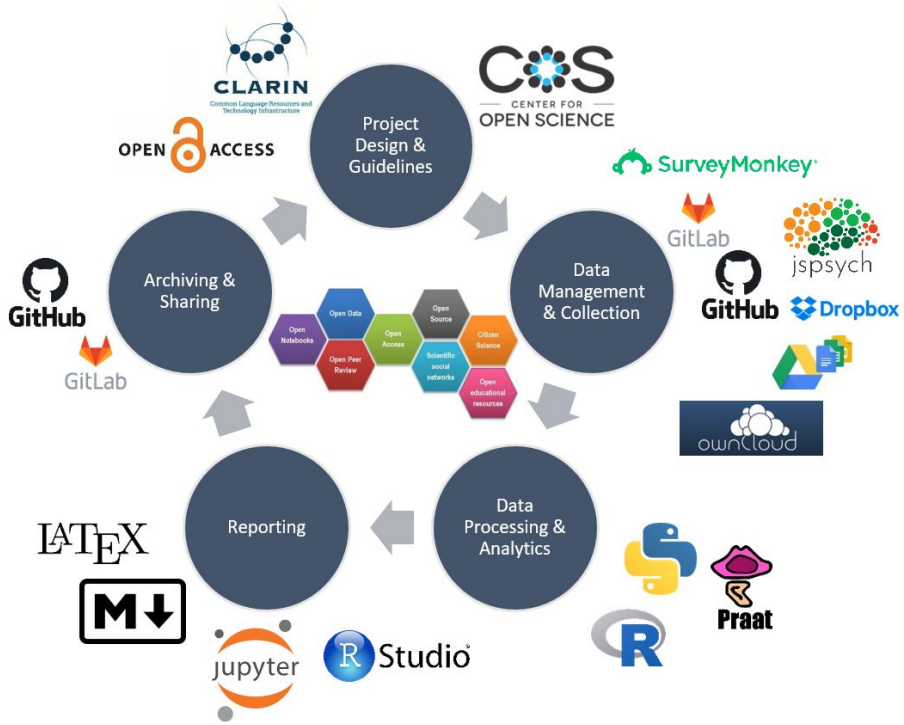
Replication Crisis | Issue | Problem

Results and Effects

- Public loss of trust in science
- Substantive efforts to improve transparency and reproducibility (in STEM and “hard” social sciences)
- Examples: increased efforts to support replication, pre-registration, and establishing a culture of sharing & infrastructures for sharing (OSF, GitHub, RNotebooks)

How to improve reproducibility | replicability | robustness

- **Data management**
  Consistency, recoverability, availability: file naming conventions, folder templates, team | lab policies, 3-2-1 rule (copies of data), bus factor (documentation)
- **FAIR data**
  Sharing data (OSF, GitHub, etc.)
- **Transparency**
  Recording workflows and version control (RStudio, **RNotebooks**, Jupyter, Markdown) commented scripts rather than (blackboxy) tools, version control (Git)
- **Practice shift**
  Pre-registration, submitting notebooks | code & data alongside papers



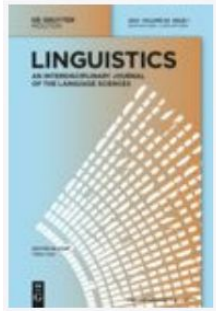
Don't see reproducibility as a burden but as a way to help and support others and yourself

Reproducibility is a lifestyle

Repercussions of the Replication Crisis in CL

Problem has been identified in (Corpus Linguistics) (recently)

- Workshops
  - ISLE 5 (London, 17-20 July, 2018, L. Sönning & V. Werner “The ‘quantitative crisis’, cumulative science, and English linguistics”
  - ISLE 6 (Joensuu, 2–5 June, 2021): M. Schweinberger & J. Flanagan “Replication and Reproducibility in English Corpus Linguistics”
  - ICAME 42 (Dortmund, 18-21 Aug. 2021): M. Schweinberger, G. Schneider & J. Flanagan “Exploring Powerful Tools to Ensure Robust and Reproducible Results in Corpus Linguistics”
- Journal publications: upcoming issue of *Linguistics*
- Efforts to improve reproducibility have started in linguistics (data citation, data sharing; not analyses)



Observation, experimentation, and replication in linguistics uistics

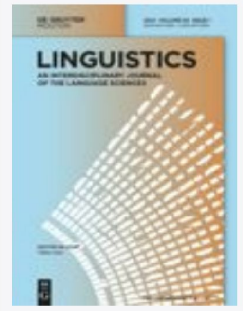
Jack Grieve 2021

Open Access Veröffentlicht von De Gruyter Mouton 6. Dezember 2017

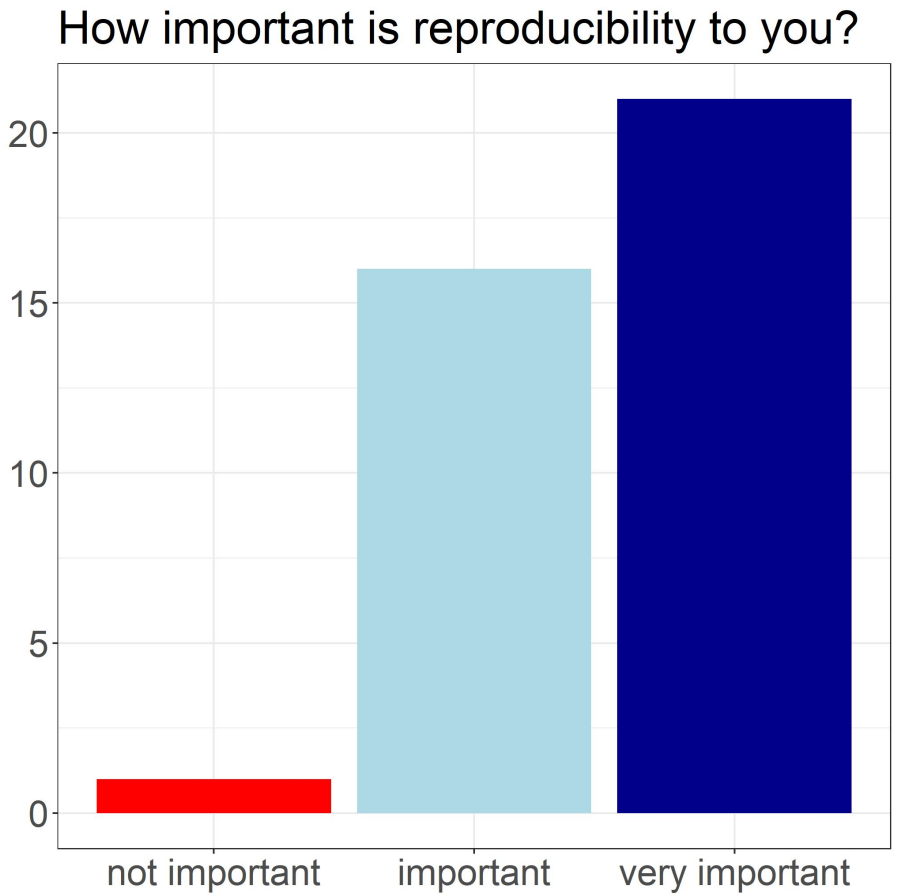
Reproducible research in linguistics: A position statement on data citation and attribution in our field

Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice und Anthony C. Woodbury

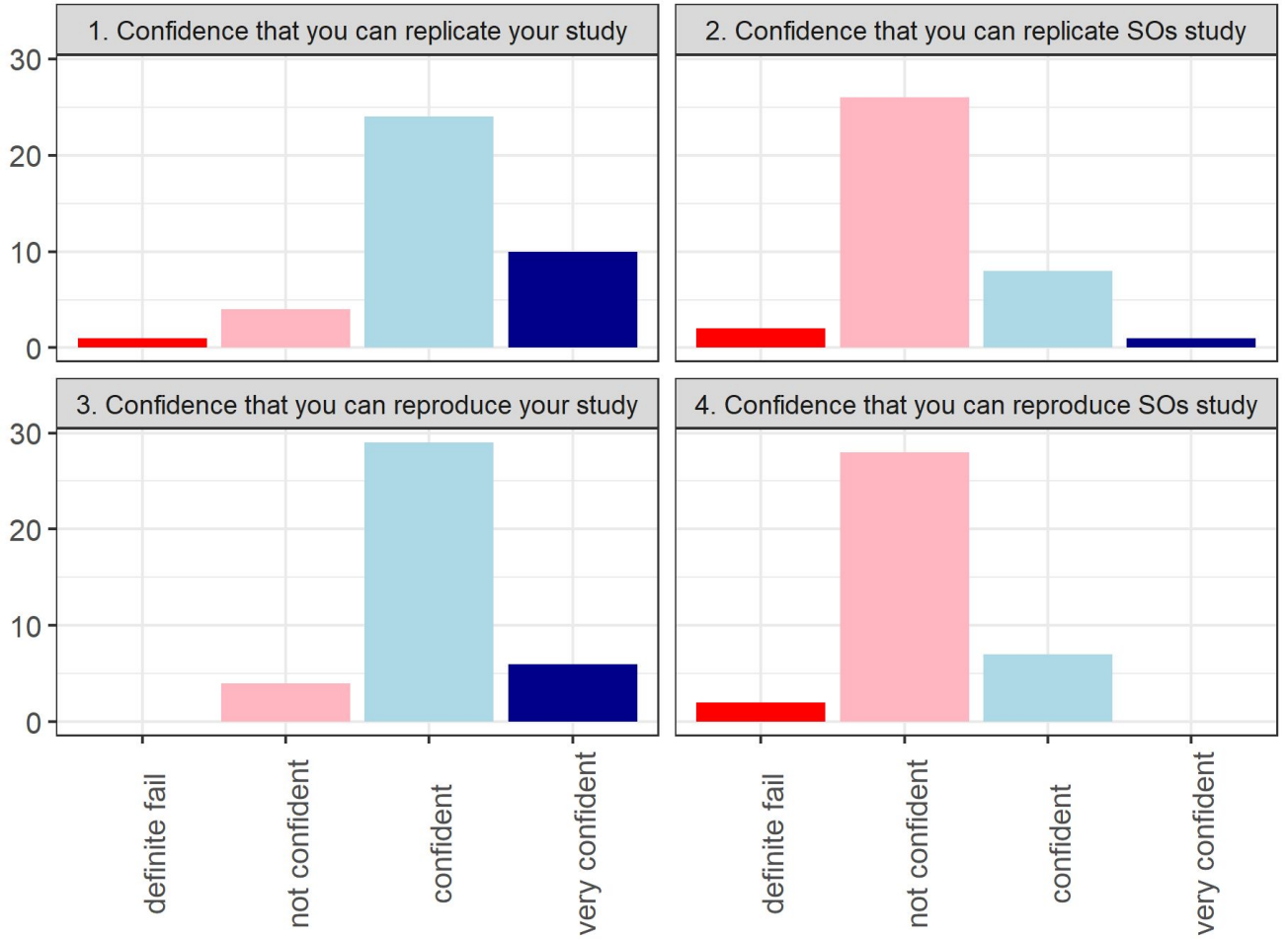
Aus der Zeitschrift Linguistics



Replication Survey



Broad support and acknowledgement that reproducibility is important



We trust ourselves but not other (others **don't TRUST** us)

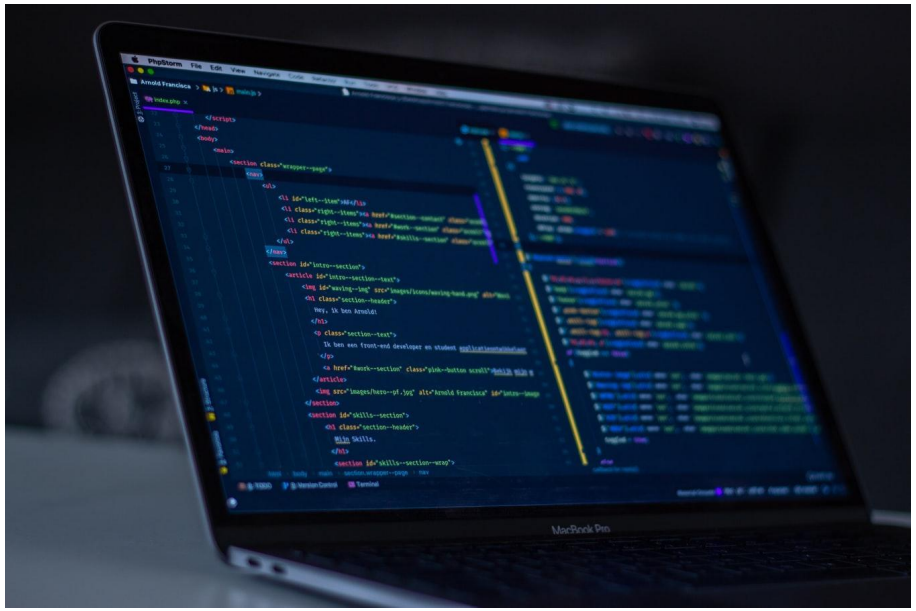
Problem | Issue

There are limited broad-range resources for HASS researchers that show how to apply computational methods to textual data and humanities topics in a transparent manner.

Resources & Infrastructure

Challenges for infrastructures

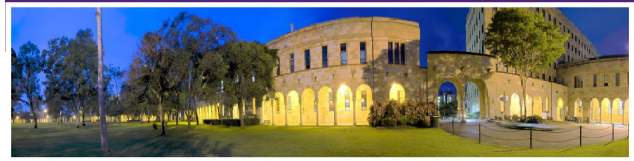
- Audiences with very different **levels of expertise**
- **Audiences** with vastly different interests, expectations, and needs
- Training is required at **different levels of specificity** (general introductions vs highly specific methods)
- Resources must meet methodological and **disciplinary variety**
- Establishing infrastructures requires **resources**
- Resources have to be user friendly | easy to use, and **intuitive**



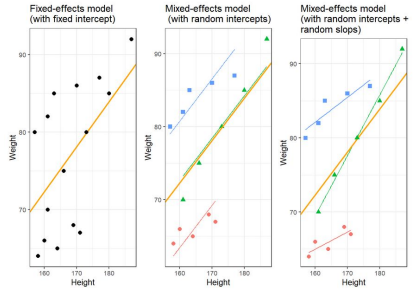
Language Technology and Data Analysis Laboratory (LADAL)

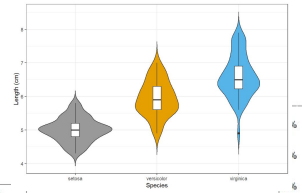
eResearch support infrastructure for computational HASS in the UQ School of Languages and Cultures

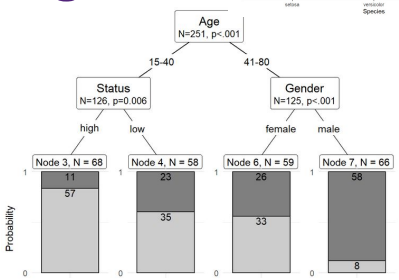
LADAL Language Technology and Data Analysis Laboratory

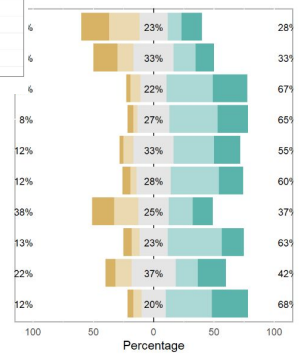


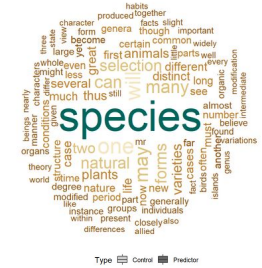
Michael Haugh (co-director of LADAL)

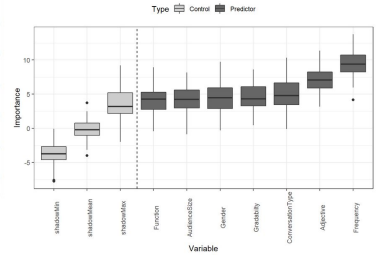




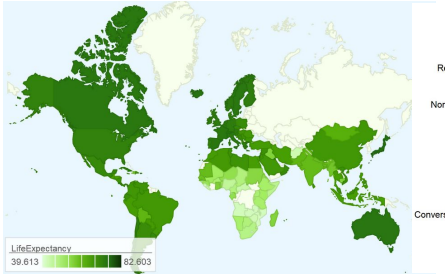


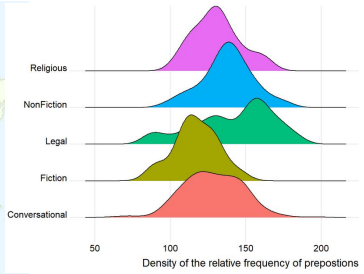


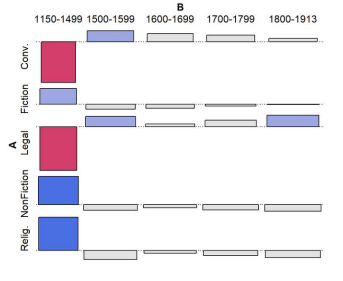




- Digital tools and **data management**
- **Computational methods** and (basic) **programming** skills
- **Data** extraction / transformation / processing
- Data **visualization** (including geospatial mapping and interactive web apps)
- **NLP** applications (text analytics) and various statistical procedures (including classification and machine learning)







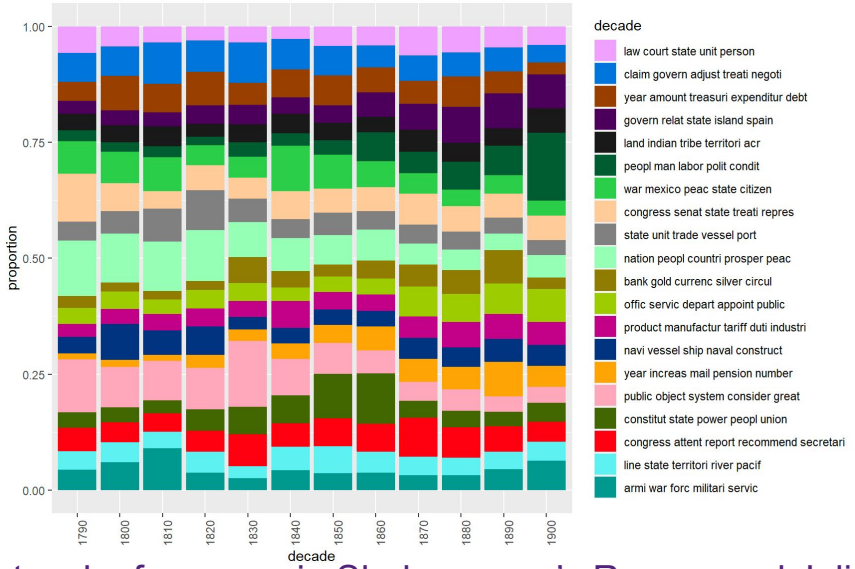
LADAL: https://slcladal.github.io/index.html

Language Technology and Data Analysis Laboratory (LADAL)

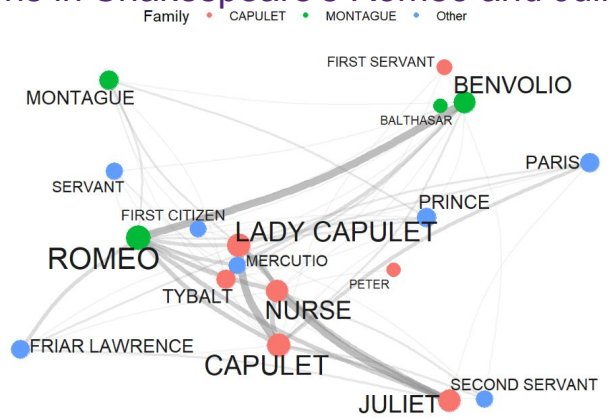
What we hope to achieve

- **Improve transparency and quality** by showcasing how to produce reproducible code)
- Enable researchers to pursue **new pathways** by using innovative methods and new types of data
- Improve **data management,** assist in making workflows **tidier**, more **transparent** and **more efficient**.
- Provide an **infrastructure for acquiring computational skills** (relevant for academia | employability for graduates)
- Showcase how **CL methods more attractive to related disciplines**

Distribution of topics in US State of the Union Addresses over time



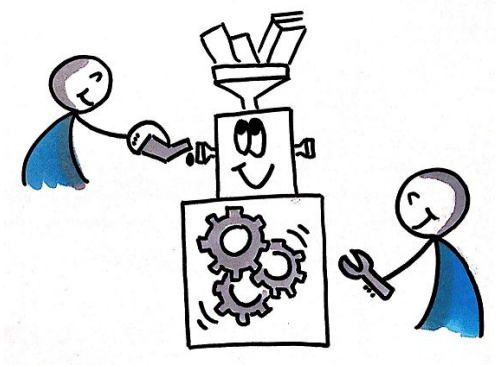
Network of persons in Shakespeare's Romeo and Juliet



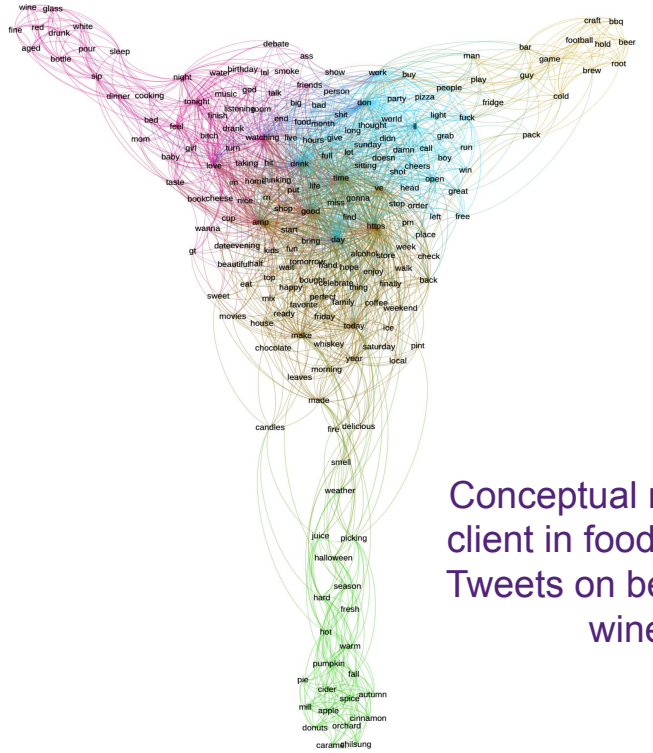
Text Crunching Center (TCC)

Service-oriented Computational Linguistics and Digital Humanities platform hosted at the University of Zurich that offers consulting, coaching, and support

- Efficient **information extraction and analysis** of large text collections (**big data**)
- **Enrichment of texts** with named entities, sentiment analysis, topic modeling, and classification, including multilingual and historical texts
- **Advice** on tools, software, and best practices
- **Help** with project applications and common projects
- (Transparent and reproducible) **Ready-made solutions**



Gerold Schneider



Conceptual map for a client in food industry: Tweets on beer, cider, wine

Case Study:

COVID19 in the Australian Twittersphere

(Schweinberger, Haugh & Hames 2021)

Case study: COVID-19 discourse in the Australian Twittersphere

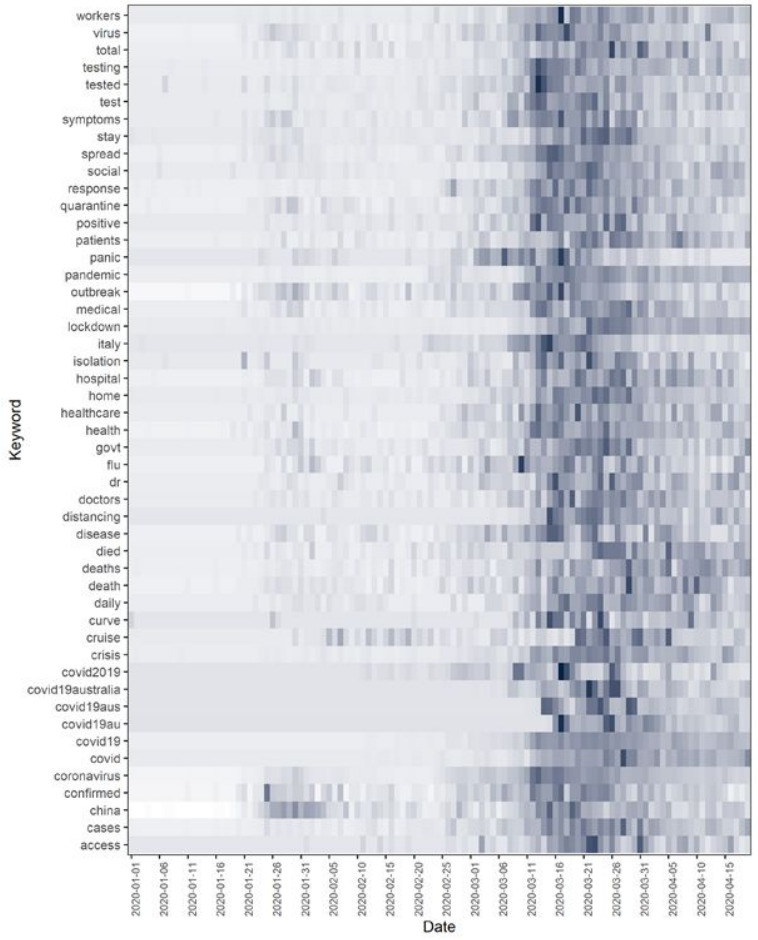
(Schweinberger, Haugh & Hames 2021)

Aim

- Showing how Corpus Linguistics can enhance (purely data-driven) text mining (done by non-linguists)
- Understanding the development and emotional shifts in the societal discourse around COVID in Australia

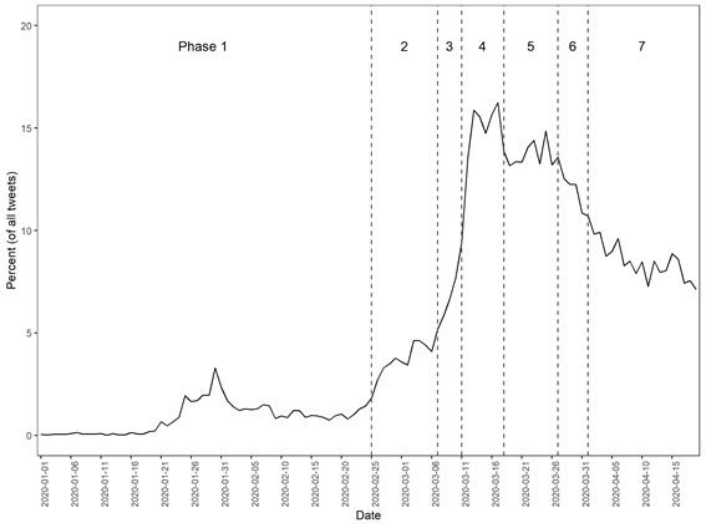
Problem

- Existing studies
  - Linguistically informed studies used qualitative approaches on rather small data sets
  - big data analytics were employed by non-linguists (discourse treated as one big undifferentiated lump)



Case study: COVID-19 discourse in the Australian Twittersphere

(Schweinberger, Haugh & Hames 2021)



Focus

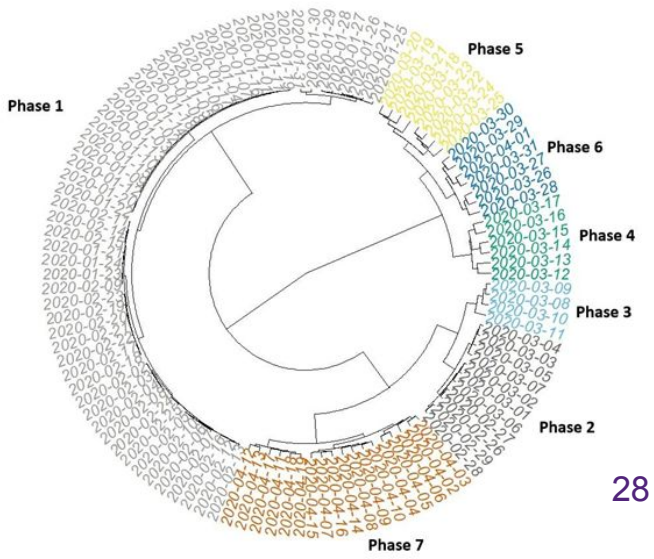
- How did it evolve and develop? (different phases in the discourse around COVID19)
- What sub-discourses form the COVID19 discourse
- What was the public's emotional response within sub-discourses?

Data

- 1 percent sample of all Australian tweets from Jan 1 to April 20

Advantages

- Combining sophisticated computational methods (e.g., PAM clustering, LDA) with a linguistically informed understanding of discourse and traditional CL methods (e.g., CCLA)

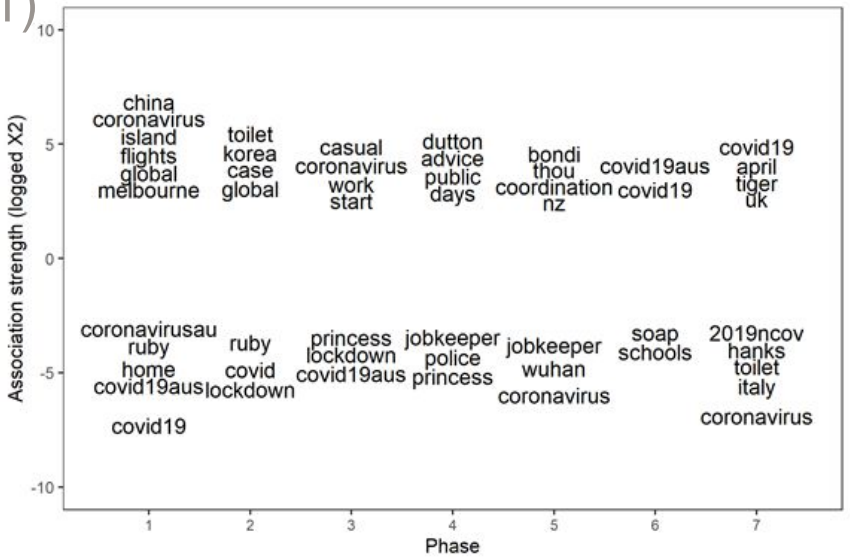


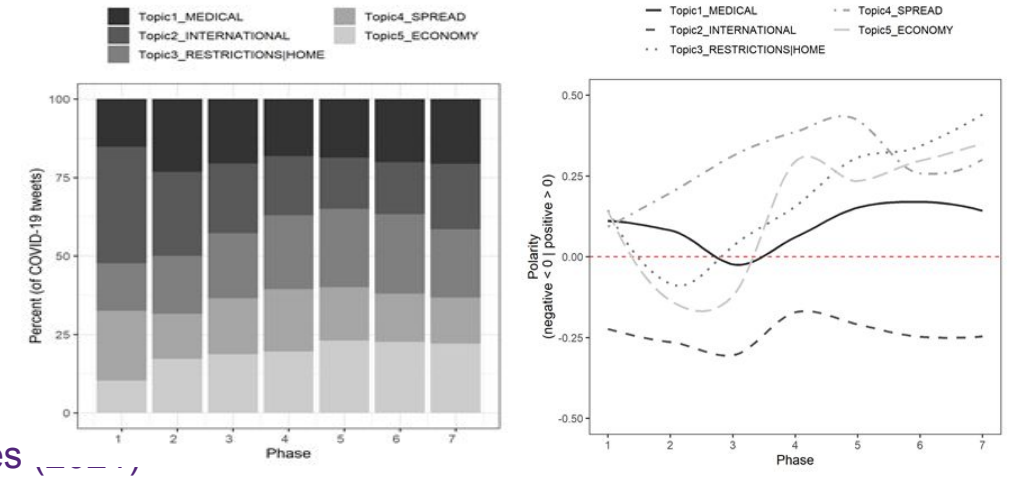
Case study: COVID-19 discourse in the Australian Twittersphere

(Schweinberger, Haugh & Hames 2021)

Outcome

- The discourse evolved in seven phases and consist of five main sub-discourses (Medical, International, Restrictions/Home, Spread, Economy)
- The discourse shifted from a focus on the outside to the inside
- Very negative when COVID19 first came to Australia but soon recovered and was increasingly positive a the pandemic spread across Australia (with the exception of the spread subdiscourse itself)
- **Important: case study for what CL can add to NLP**





Schweinberger, Haugh and Hames (

Summary

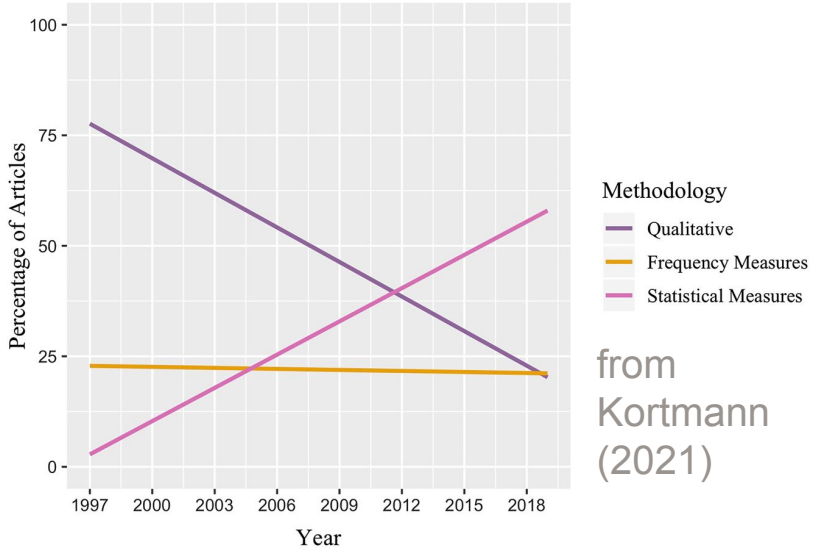
Key points

- Corpus Linguists have good reasons **not** to become Language Data Scientists | Computational Linguists (and they should not be!)
  - Our skills can support and enhance other disciplines (COVID19 case study)
- CL could profit from integrating aspects of Language Data Science | Computational Linguistics
  - Versatility
  - Applicability
  - Replicability
- Necessity for infrastructure and resources to be able to tap into this potential

Discussion and Outlook

The Quantitative Turn | quo vadis, Corpus Linguistics?

- Quantitative Turn in Linguistics: dramatic increase in the use of statistical methods (Janda 2017, Kortmann 2021) (very recommendable reflective discussion of the quantitative turn in Kortmann 2018: overall rather positive if methods are handled with care)
- Should Corpus Linguistics re-form itself: back to more linguistic description (shift in focus away from methods)? (Larsson, Egbert & Biber forthc.)
- Split akin to psychology and psychoanalysis?...



from Kortmann (2021)

Quantification is not an end in itself, but generating reliable knowledge is (replicability and transparency)

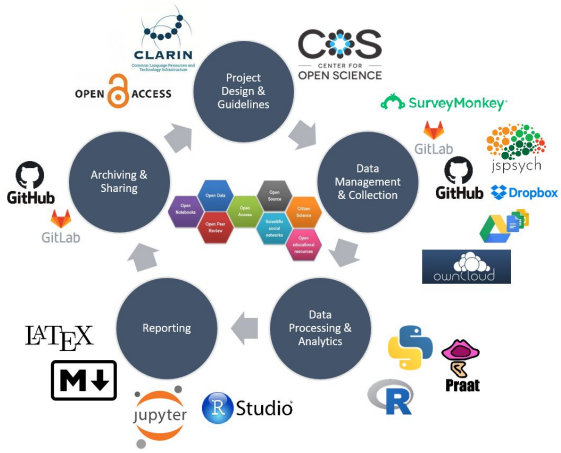
Discussion and Outlook

Corpus Linguists are and have been aware of replicability | reproducibility (issues)

- **Arppe et al. (2010)**: “Ideally, research in cognitive linguistics should be based on authentic language use, its results should be **replicable**, and its claims falsifiable. ”
- **Kortmann (2018)**: “do everything that is necessary (!) for achieving a maximum of methodological **transparency**, rigour, statistical significance, **robustness, reproducibility**, falsifiability and, ultimately, explanatory power and mileage for linguistic theory-building”
- Workshops (ISLE5, ISLE6, ICAME42)
- *Linguistics* (upcoming issue)

Communal discussion on Reproducibility: integration of tools and methods that make research more transparent and reproducible | replicable

Adopt resources and establish an infrastructure like the infrastructure for quantitative methods (books, workshops, etc.)



Discussion and Outlook

Corpus Linguists should seek collaboration and build interdisciplinary networks

- We have advantages and are more advanced than other field that contribute to Digital Humanities (we can help prevent that the wheel is re-invented over and over again)
- We can profit from adopting wheels from other disciplines (Reproducibility | Culture of Sharing)

We, as **Corpus Linguists, can contribute** by providing a more **fine-grained understanding of discourse** and we can **profit from adopting computational methods** and data management practices

ICAME42 | tu
crossing boundaries through corpora

Thank you very much

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

LADAL

NORGES ARKTISKE UNIVERSITET
UIT
Logosymbol.

AcqVA Aurora Lab

References

Anderson, C. J. et. al. (2016). Response to Comment on “Estimating the reproducibility of psychological science”. *Science* 351(6277): 1037.

Anthony, L. (2020). Programming for Corpus Linguistics. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 181-207. Berlin & New York: Springer.

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M. & Zeschel, A. 2010. Cognitive Corpus Linguistics: five points of debate on current theory and methodology. *Corpora* 5(1): 1-27.

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533: 452-454.

Desagulier, G. (2017). *Corpus linguistics and statistics with R.* Springer International Publishing.

Gilbert, D. T., King, G., Pettigrew, S., Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science".

Science 351(6277): 1037.

Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass* 3: 1–17.

Gries, S. T. (2016). *Quantitative corpus linguistics with R: A practical introduction*. Routledge.

References

Gries, S. T. (2021). *Statistics for Linguistics with R*. Berlin: De Gruyter Mouton.

Grieve, J. (2021). Observation, experimentation, and replication in linguistics. *Linguistics.* https://doi-org.ezproxy.library.uq.edu.au/10.1515/ling-2021-0094

Hammond, M. (2020). *Python for Linguists*. Cambridge: Cambridge University Press.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine* 2(8): e124.

Janda, L A. 2017. The quantitative turn. In B. Dancygier (ed.), *The Cambridge handbook of cognitive linguistics*, 498–514. Cambridge: Cambridge University Press

Kortmann, B. (2018). Reflecting on the quantitative turn in linguistics. Lunch Lecture 2017/18: Quantitative vs. qualitative approaches across sciences (11 January 2018). Accessed Aug. 13, 2021, url: https://www.frias.uni-freiburg.de/downloads/veranstaltungen/ppp-kortmann)

Kortmann, B. (2021). Reflecting on the quantitative turn in linguistics. *Linguistics* https://doi-org.ezproxy.library.uq.edu.au/10.1515/ling-2019-0046

Larsson, T., Egbert, J. & Biber, D. (forthc.). On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora* 17(1).

References

Levshina, N. (2015). *How to do linguistics with R. Data Exploration and Statistical Analysis*. Amsterdam & Philadelphia: John Benjamins.

Lundquist, H. (2010). *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

McEnery, T. & A. Hardie. (2001). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Open Science Collaboration,. (2015). Estimating the reproducibility of psychological science. *Science* 349(6251): aac4716.

Schweinberger, M., M. Haugh & S. Hames. (2021). COVID19 in the Australian Twittersphere. *Big Data & Society* 8(1): 1-17.

References

LADAL: The Language Technology and Data Analysis Laboratory, LADAL. School of Languages and Cultures, The University of Queensland, Access Aug. 13, 2021, url: https://slcladal.github.io/index.html.

AcqVQ Aurora Lab. UiT Aurora Center for Language Acquisition, Variation, and Attrition, The Arctic University of Norway, Tromsø, Access Aug. 13 2021, url: https://site.uit.no/acqvalab/.

PoLaR: Psycholinguistics of Language Representation (PoLaR) lab. UiT Aurora Center for Language Acquisition, Variation, and Attrition, The Arctic University of Norway, Tromsø, Access Aug. 13 2021, url: https://site.uit.no/polar/.

TCC: Text Crunching Center, TCC. Department of Computational Linguistics, University of Zurich, Access Aug. 13 2021, url: https://www.cl.uzh.ch/en/TCC.html.

ICAME42 | tu
crossing boundaries through corpora

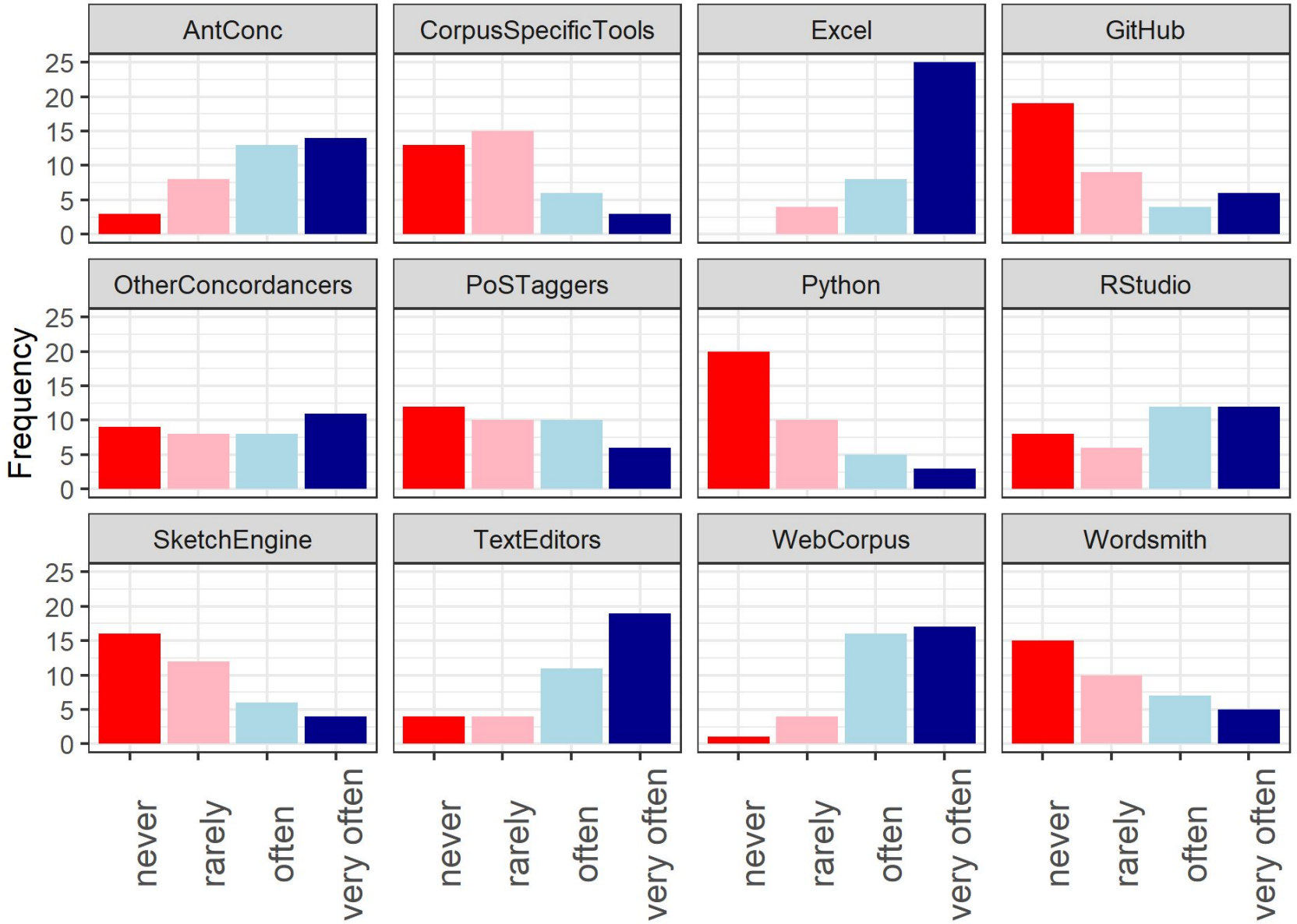
Thank you very much

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

LADAL

NORGES ARKTISKE UNIVERSITET
UIT
Logosymbol.

AcqVA Aurora Lab


AntConc
CorpusSpecificTools
Excel
GitHub
OtherConcordancers
PoSTaggers
Python
RStudio
SketchEngine
TextEditors
WebCorpus
Wordsmith
Frequency
0
5
10
15
20
25
never
rarely
often
very often

Programming in CL

Corpus Linguistics

- "[...] **study of language** based on examples of 'real life' language use." (McEnery & Wilson 2001: 1)
- An efficient way to study language use (Lundquist 2001: 1)

Computational Linguistics

- "[...] scientific study of language from a computational perspective. [...] Work in computational linguistics is in some cases motivated from a scientific perspective [...] and in other cases the motivation **may be more purely technological in that one wants to provide a working component of a speech or natural language system**." (Association for Computational Linguistics)

(Language) Data Science

- Interdisciplinary field that **uses scientific methods, processes, algorithms and systems to extract knowledge** and insights from structured and unstructured data, and **apply knowledge and actionable insights from data** across a broad range of application domains. (Wikipedia, Entry *Data Science*)