International Perspectives on Corpus Technology for Language Learning

An evaluation of computational, corpus-based approaches to language learning

Martin Schweinberger
(slides @ www.martinschweinberger.de

Some background

- Lecturer in Applied Linguistics at the *University of Queensland* (UQ)
- Part-time Associate Professor II at the *Arctic University of Norway* in Tromsø (UiT).
- Background in Philosophy, English Linguistics, and Psychology with a focus on computational processing and statistical modelling of language data.
- Research focuses on corpus-based analyses of language variation and change, language acquisition, and reproducibility in the language sciences.
- Steering committee member of ATAP (*Australian Text Analytics Platform*) and board member of *The International Computer Archive of Modern and Medieval English* (ICAME)
- Co-director to the *Language Technology and Data Analysis Laboratory* (LADAL) at UQ
- Principal data science advisor to the *AcqVA Aurora Lab* at UiT

Martin Schweinberger

LADAL

AcqVA Aurora Lab

Motivation for this talk

To assess to what extent groups profit from computation
we have to be aware that different groups have distinct needs

- Students
  - Learning to communicate in a foreign language
- Teachers
  - Assisting students in learning a foreign language
  - Positive learning experience

- Researchers
  - Assisting teachers in assisting students in learning a foreign language
  - Produce reliable and reproducible knowledge that can be applied to language teaching classrooms

Opportunities, drawbacks, and issues of using tools | web apps and programming in (computer assisted) language learning and teaching

Outline

- When/where/for whom are tools useful?
  - Advantages and drawbacks of tools
- When/where/for whom is programming useful?
  - Advantages and drawbacks of programming
  - The computational revolution, reproducibility, and computer-assisted language teaching and learning

- LADAL (ATAP): upskilling platform for coding in HASS
- Analysing learner language using R
  (LADAL tutorial and interactive Jupyter notebook)

When/where/for whom are tools useful?

Advantages and drawbacks of tools

When/where/for whom are tools useful?

- Tools are at the core of DDL and CALL

  Data driven learning (DDL) defined as

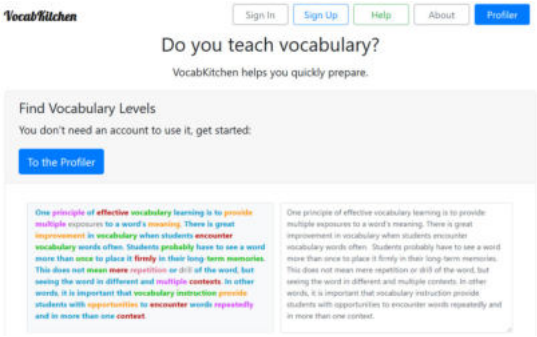
  “the use of tools and techniques of corpus linguistics for second language learning or use" (Boulton & Cobb 2017: 5)

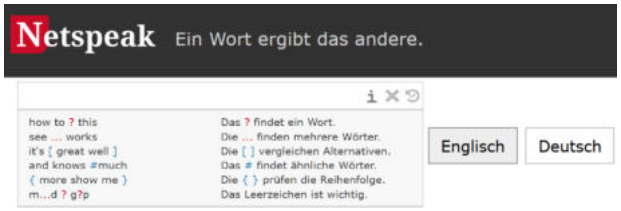
SKELL

Sketch Engine for language learning

SKETCH ENGINE

- Tools: desktop software applications and web-interfaces





When/where/for whom are tools useful?

#LancsBox: Lancaster University corpus toolbox

- Tools have many advantages
  - Relatively easy to use (a lot of bang for few bucks) and versatile – tools for many different tasks
  - Allow analysis of actual, naturally occurring language use
  - Provide examples of actual, naturally occurring language use
  - Suggestions of how to use language in a natural manner
  - Assist in creating materials, syllabus design, and speeding-up, standardizing assessment (McEnery & Xiao 2011)
  - Teaches metalinguistic skills (O'Sullivan 2007: 277) and the use of technology
- Advantages particularly for language learners and teachers
- Situation for researchers is more complex (need to know many tools, research needs to be reproducible, analyses require flexibility, etc.)

SKETCH ENGINE

SKELL

Sketch Engine for language learning

When/where/for whom are tools/apps useful?

- Disadvantages of tools
  - Many different tools and each tool requires training (with new ones coming in and others going defunct)
  - Students are trained to be users (not developers)
  - Tools can be black boxes – it's not always clear what goes on under the hood
  - Tools are versatile – but limited (there are no tools for everything)
  - Tools make replication harder (even if the tools are well documented, free, and open source)

When/where/for whom is programming useful?

Advantages and drawbacks of programming

When/where/for whom is programming useful?

Advantages of programming (see Anthony 2020)

- Allows analyses not possible with existing tools (flexibility)
- Control: putting you "in the driving seat" (Gries 2009: 12)
- Employability (learning to code)
- Reproducibility of research (Schweinberger 2021)
- Automation (scripts allow to speed up repetitive tasks)

Programming vs Coding

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

When/where/for whom is programming useful?

Disadvantages of programming

- Too complex (requires training)
- Not user-friendly (not ready to go)
- No immediate profit for language learners | teachers

- Lack of training infrastructure
- Time limits: little time to acquire additional skills
- Interests and methodology (small data sets, fine-grained qualitative analyses, manual processing)

When/where/for whom is programming useful?

The computational revolution, reproducibility, and computer-assisted language teaching and learning

The Computational Revolution

Computation is becoming increasingly important

- Computational revolution has changed all domains of life
- Language data and computational approaches to processing, transforming, analysing, and visualizing text data are and continue to be prevalent in the economy and in the humanities
- Despite this, HASS has been reluctant to integrate computational skills | programming

Machine translation

Text-2-Speech | Speech-2-Text

Voice recognition

Search engines

Content detection | summarization

Named Entity Recognition

Chat bots | question answering

Spelling correction

Replication Crisis

Controversial ongoing methodological crisis that originated in medicine (Joannidis 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)

Problem
Following comments argued that the results of this study did not support the low-replicability claim and suggested that the replicability in psychology is actually comparatively high (Gilbert et al. 2016) while again others problematized the follow up studies which showed high replicability (Anderson 2016)

Replication Crisis

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

FiveThirtyEight

Politics Sports **Science & Health** Economics Culture

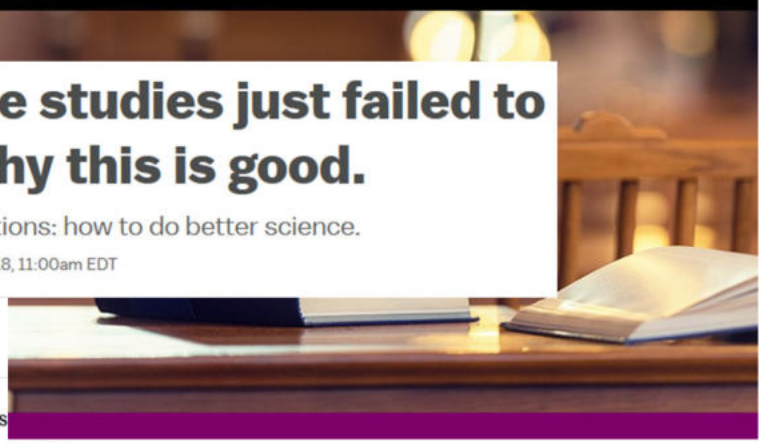
Our 2019 March Madness

DEC. 6, 2018, AT 11:19 AM

Psychology's Replication Crisis Has Made The Field Better

By Christie Aschwanden

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

(Dis)trust in Science

Can we cure the scourge of n

By Gleb Tsipursky on July 5

UMANS

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

IKE MCRAE 27 AUG 2018

PHYS.ORG Nanotechnology Physics Earth Astronomy & Space Technology

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

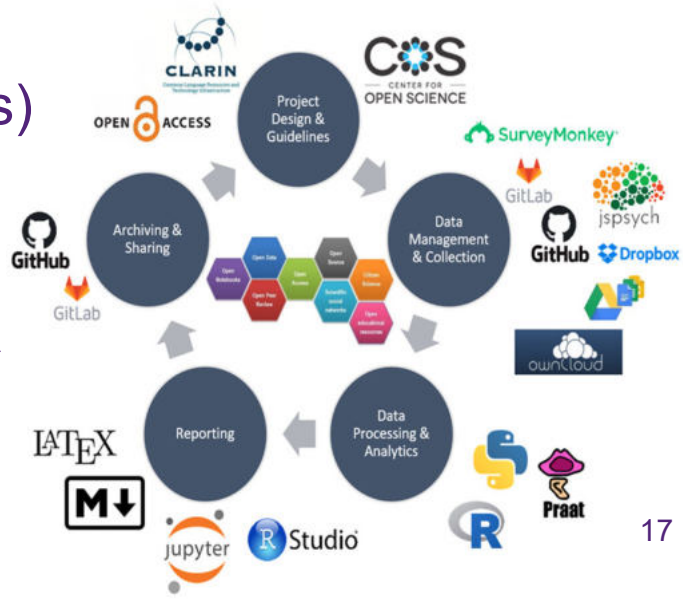
Researcher discusses the the science replication crisis

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

Replication Crisis

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)



Computational revolution, replication crisis, and language learning

In light of the replication crisis and the increasing importance of computation…

- Should CALL and teaching take the chance and integrate computational skills into CALL?
- Problems
  - Lack of expertise: few curricula in HASS teach computational skills (few people who can teach computational CALL)
  - Lack of motivation: tools are very good and flexible – complexity is a limiting factor
  - **Lack of resources**: few | no infrastructures, materials, exercises, syllabi, etc.

Language Technology and Data Analysis Laboratory **(LADAL)**

and the ***Australian Text Analytics Platform*** **(ATAP):**

upskilling platform for coding in HASS

Language Technology and Data Analysis Laboratory (LADAL)

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

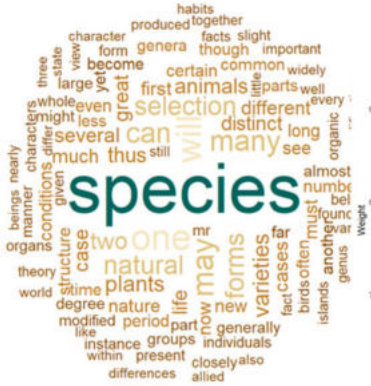
Part of ATAP – both represent upskilling infrastructures that aim at enabling **development** of skills in

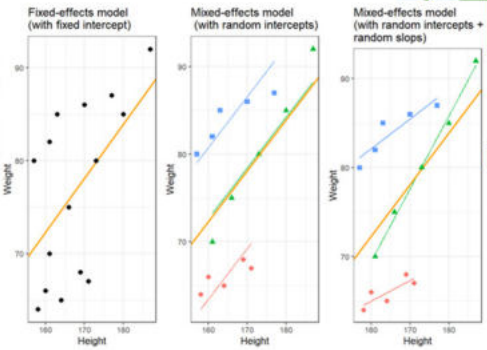
- **Text Analytics**
- Digital tools and **data management**
- **Computational methods** and (basic) **programming** skills
- **Data** extraction / transformation / processing
- Data **visualization** (including geospatial mapping and interactive web apps)
- **NLP** applications (text analytics) and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

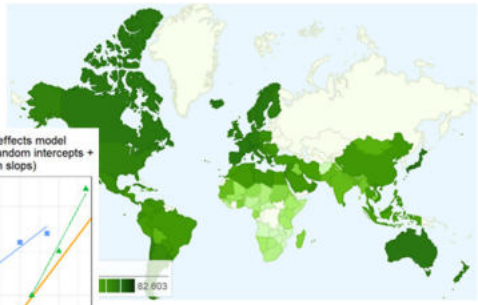
Language Technology and Data Analysis Laboratory (LADAL)

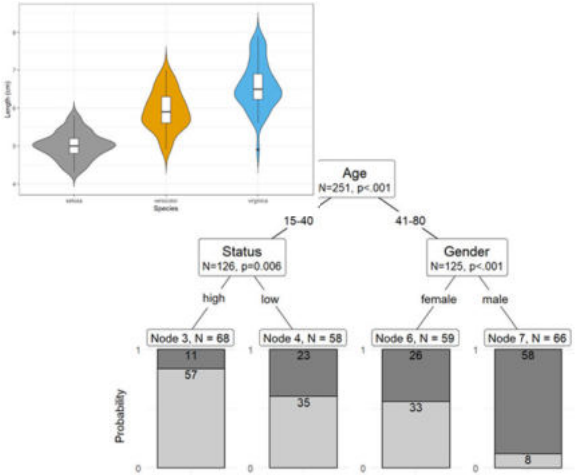
Services

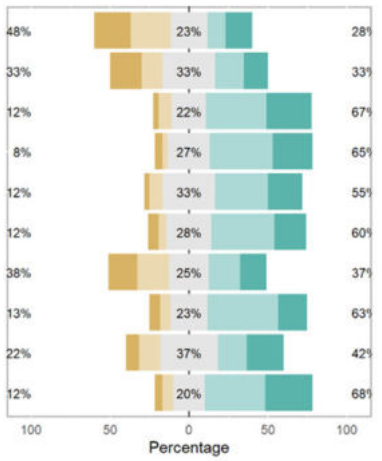
- Self-guided study **materials (**online **tutorials)** on topics relating to data extraction, processing, management, and visualization, as well as statistical analyses (learning to “code”)
- Face-to-face **consultations** and practical hands-on **workshops**
- **(Interactive notebooks)**

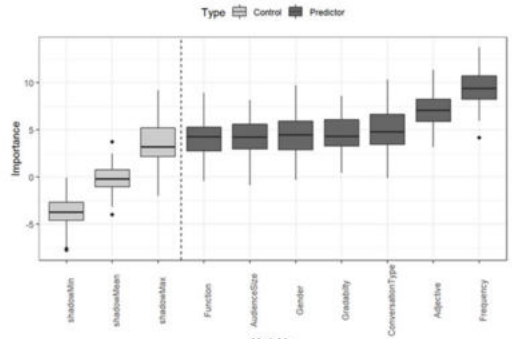


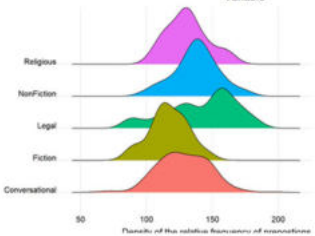


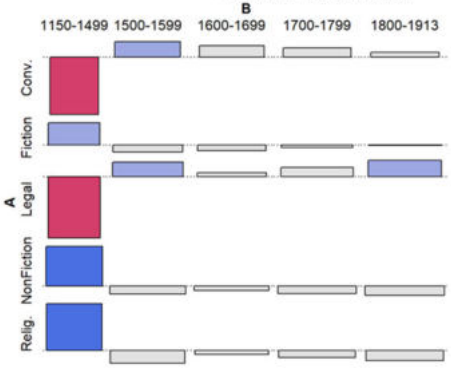


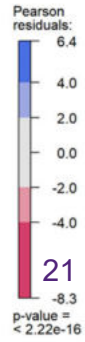










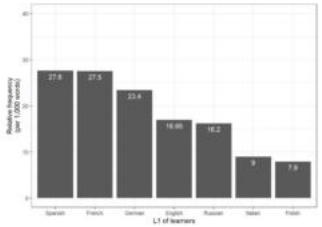


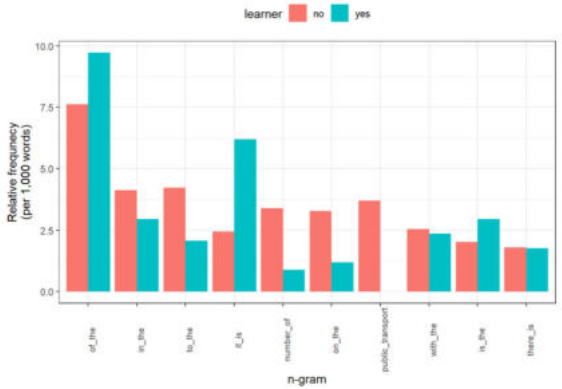
Language Technology and Data Analysis Laboratory (LADAL)

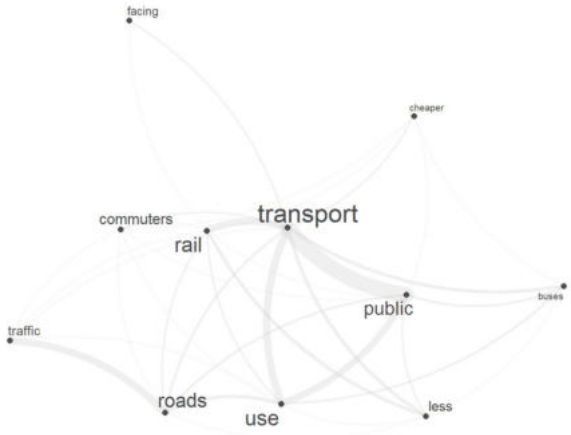
Why is LADAL relevant for computer-assisted language learning and teaching?

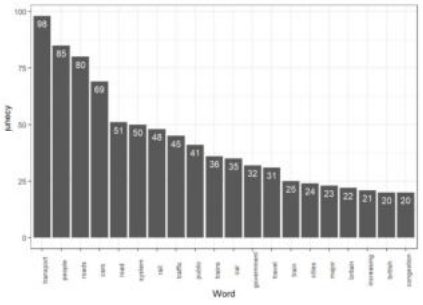
- **LADAL tutorial on *Analysing learner language with R***

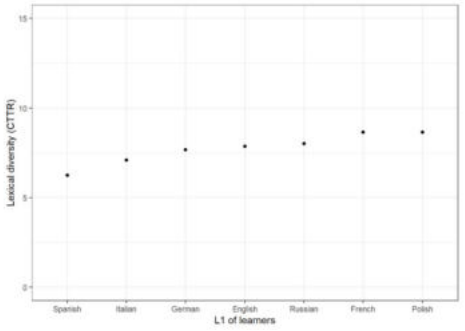
  `https://slcladal.github.io/llr.html`

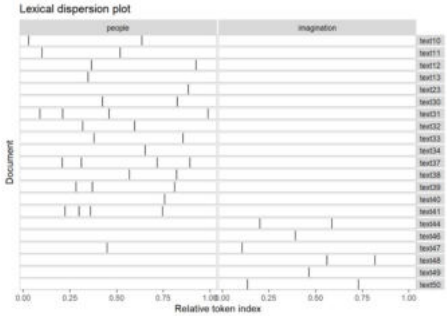


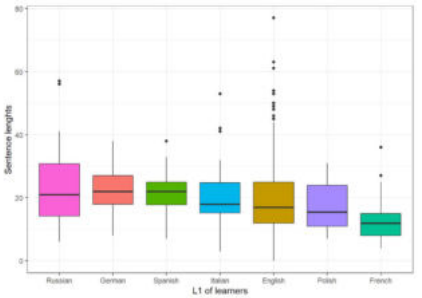












Analysing learner language using R

(LADAL tutorial and interactive Jupyter notebook)

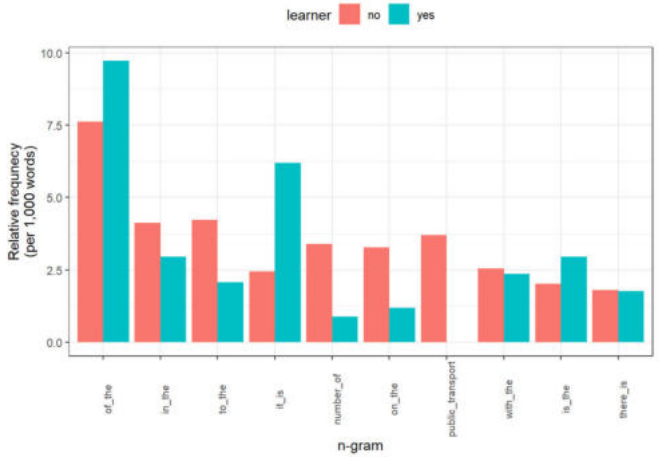
LADAL resources

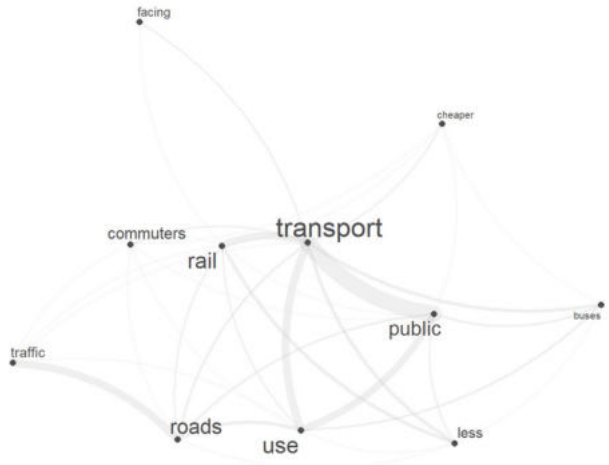
- **LADAL tutorial on *Analysing learner language with R***

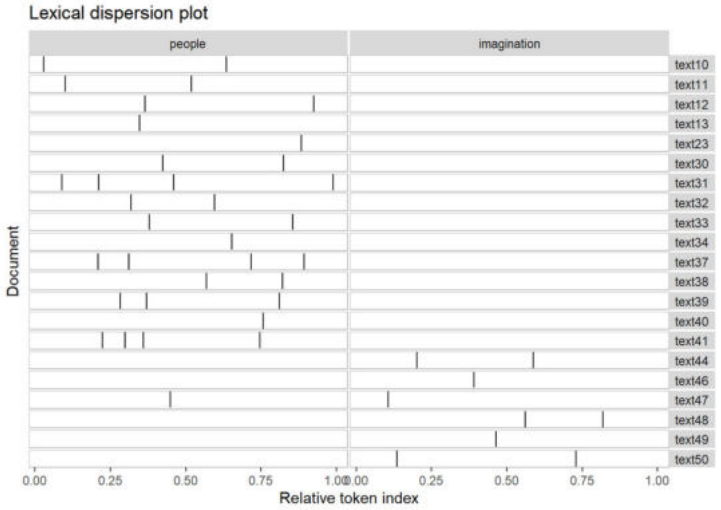
  https://slcladal.github.io/llr.html

- **LADAL interactive Jupyter notebook for *Analysing learner language with R***

  https://colab.research.google.com/drive/1aT7jbw1Gdt74irSjCqP7OLml5Ff_OzZP?usp=sharing







Discussion and Outlook

Discussion and Outlook

An evaluation of computational, corpus-based approaches to language learning

Main theses

1. Tool use is very recommended for students/teachers/researchers in computer assisted language learning and teaching as they are very easy to use, flexible, teach metalinguistic skills, and allow learners to discover actual naturally occurring language use.
2. Programming is more flexible and more difficult but it allows learners to acquire marketable skills alongside a language; for researchers they increase flexibility and reproducibility
3. Language learning and teaching as a potential vehicle for computational skills (from users to developers)?

   Depends on the context, the age of learners and teachers, the outlook of the program, and availability of training infrastructures.
4. Problem: lack of infrastructures (upskilling platforms/materials/courses) and courses that teach computational skills targeted at language teachers (LADAL/ATAP as infrastructure options)

Thank you very much for listening

and thanks to Peter and Fran

for organizing this webinar series

m.schweinberger@uq.edu.au

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

LADAL

NORGES ARKTISKE UNIVERSITET • UIT •

Logosymbol.
AcqVA Aurora Lab

References

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.

Bouton, A. & Cobb, T. (2017). Corpus use in language learning: A meta analysis. *Language Learning* 67(2): 348-393.

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.

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

O'Sullivan, I. (2007). Enhancing a process-oriented approach to literacy and language learning: the role of corpus consultation literacy. *ReCALL* 19(3): 269-286.

Schweinberger, M. (2020a). How Learner Corpus Research can inform language learning and teaching. An analysis of adjective amplification among L1 and L2 English speakers. *Australian Review of Applied Linguistics* 42(2): 196-218

Schweinberger, M. (2020b). A corpus-based analysis of differences in the use of very for adjective amplification among native speakers and learners of English. *International Journal of Learner Corpus Research* 6(2): 163-192

Schweinberger, M. (2021). Corpus Linguistics, Language Data Science, and Computational Linguistics – building bridges or splitting apart?. Plenary at ICAME 42 (Dortmund, August 21, 2021) (https://youtu.be/ZxEex7wuHYU)

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/.

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.

International Perspectives on Corpus Technology for Language Learning

An evaluation of computational,

corpus-based approaches to language learning

Martin Schweinberger

The University Of Queensland AUSTRALIA
CREATE CHANGE
LADAL

Norges arktiske universitet • UIT •
Logosymbol.
AcqVA Aurora Lab