

International Perspectives on Corpus Technology for Language Learning

### An evaluation of computational,

## corpus-based approaches to language learning

### Martin Schweinberger (slides @ www.martinschweinberger.de



CREATE CHANGE









## 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



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







THE UNIVERSITY





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

Tools: desktop software applications and web-interfaces



7

SKETCH

-	1.1.1.1		-		
Do you tea	ch vo	ocabula	ary?		
VocabKitchen he	lps you o	quickly prep	are.		
Find Vocabulary Levels					
You don't need an account to use it, get started:					
To the Profiler					
To the Profiler					
To the Profiler One principle of effective variability learning is to provide	0	ne principle of et	fective vocabular	y lewning is to provide	
To the Profile One principle of effective socilatory fearing is to provide studying reporters to a word's managing. There is great increasing the socialized science science.	0	ne principle of et utiple exposure	Tective vocabular to a word's mea	y learning is to provide ring. There is great underst encoder	
To the Profiler One principle of effective secalablery learning is to provide methyle operanes to a word's among There is great suppressent in wordballery when students accessite wordballery ward chars. Students publicly learn to see a so	ord vo	ne principle of et uitiple exposurer sprovensent in vo scabulary words	fective vocabular to a word's mea cabulary when st often. Students p	y learning is to provide ning. There is great adents encounter robably have to see a w	vord
To the Profiler One principle of effective weakladary learning in to provide multiple operators to a send smaller. There is great improvement is weakladary there is defined as executive methodology used on the Statistic privately from the set that share more partners when the set the set to set to the share more partners when the set to set to set to the share more partners when the set to set to set to the share more partners when the set to set to set to the share more partners when the set to set the set to set to the share more partners when the set to set the set to set to the share more partners when the set to set to set to set to the share more partners when the set to set to set to set to the share more partners when the set to set to set to set to the set to set to the set to set to the set to set to the set to set to set to set to set to set to set to the set to set to set to set to set to set to set to the set to set to set to set to set to set to set to the set to set to the set to set to set to set to set to set to set to set to	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	we principle of et uttiple exposure scaladary words scaladary words is dear order to	fective vocabular to a word's rea cabulary when st often. Students p place it filmly in presentation	y learning is to provide ring. There is great unbably have to see a w faint long-term merced with the second to	vord les.
To the Profiler One principle of effective excluding learning to to provide employee exclusion of the exclus		ne principle of et altiple exposure sprovement in v calobacy words care than since the state does not mea wing the word in	fective vocabular to a word's rear cabulary when st often "Sudents p place it firmly in mere repetition different and mu	y learning is to provide ning. There is great saferts encounter couldy have to see a w their long-term memor or shift of the word to see a logic contexts. In-safer	voi d leg. e
To the Profiler One principle of effective excelutory leaving to to provide multiple company to a work another, there is good multiple company to a work another and another to excelutory means that work of the family in their to be the means that are not to place. If family in their topic of the off the order one good and the means mere specified or the off of the order to easing the work in different and multiple contexts. In other earise, in its impact that workshold primetation and other off of the order off of the order off off off order off of easing the work in different and multiple contexts. In other earise, its impact that workshold primetation predict	ord of	me principle of et utiple exposure proversent in vo scabolary words doe than since to this does not mea sing the word in ords, it is imports	fective vocabular to a word's mai cabulary when st others. Students p place it firmly in a more expetition in more expetition officient and mi mit that socabula	) learning is to provide: ning. There is grout saferts encounter table) have to see a w their long-term memori or will of the word, but lique cardinate. In-affer by instruction provide	void las, t

[	Netspeak	Ein Wort ergibt das andere.		
ĺ		i×9		
	how to ? this see works it's [ great well ]	Das ? findet ein Wort. Die finden mehrere Wörter. Die [ ] vergleichen Alternativen.	Englisch	Deutsch
	and knows #much { more show me } md ? g?p	Das # findet ähnliche Wörter. Die ( ) prüfen die Reihenfolge. Das Leerzeichen ist wichtig.		

English-Corpora.org 🚷					
hove con				ccount upgrade	
and the second se				20100	
The most widely used online corpora: guided lour, ow	rview, search by	pes, variation, v	rtual corpora (c	uick overview) and	
The links below are for the online interface. But you ca	in also 🕐 down	foad the corpor	a for use on you	r own computer.	
Corpus (orither access)	Download	# words	Dialect	Time period	Genre(s)
Web: The Intelligent Web-based Corpus	0	14 billion	6 countries	3917	web
News on the Web (NOW)	0	13.7 billion+	20 countries	2010-yesterday	Web: New
Global Web-Based English (GloWbE)	0	1.9 billion	20 countries	2012-13	Web (Incl big
Wikipedia Corpus	0	1.9 billion	(Various)	2014	Witipedia
Coronavirus Corpus	0	1.23 billion+	20 countries	Jan 2020-yesterday	Web: New
Corpus of Contemporary American English (COCA)	0	1.0 billion	American	1990-2019	Balanced
Corpus of Historical American English (COHA)	0	475 million	American	1820-2019	Balanced
The TV Corpus	0	325 million	6 countries	1950-2018	TV shows
The Movie Corpus	0	200 million	6 countries	1930-2018	Movies
	-	And an Art of	A second state	and and	

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

- 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.)



Netspeak	Ein Wort ergibt das andere.		
	i×9		
how to 7 this see works it's [ great well ]	Das 7 findet ein Wort. Die finden mehrere Wörter. Die [] vergleichen Alternativen.	Englisch	Deutsch
( more show me ) md ? g?p	Des # findet ähnliche Worter. Die ( ) präfen die Reflenfolge. Des Leerzeichen ist wichtig.		



THE UNIVERSITY



**#LancsBox: Lancaster University corpus toolbox** 



## 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 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



YAHOO!

Machine translation

Text-2-Speech | Speech-2-Text

nte

Voice recognition

Named Entity Recognition

Spelling correction 14

 Despite this, HASS has been reluctant to integrate computational skills | programming

Content detection | summarization

Chat bots | question answering

Search engines



## **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)

15



## **Replication Crisis**

#### 💧 NOBA

Browse Content / The Replication Crisis in Psychology

#### SCIENCEalert

## The Replication Crisis in Psychology

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

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

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



By Christie Aschwanden



17

## **Replication Crisis**

PEW

TREND ARTICLE

bruary 9 2021 B

**Trend Magazine** 

**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)



RStudio

jupyter

IAT<sub>E</sub>X

M↓



## 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)







THE UNIVERSITY

text11 text12 text13 text23 text30 text30 text31 text33 text34 text34

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

• LADAL tutorial on Analysing learner language with R





## 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

facing

https://colab.research.google.com/drive/1aT7jbw1Gdt74irSjCgP70Lm15F











24



### **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



CREATE CHANGE









## 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: <a href="https://site.uit.no/acqvalab/">https://site.uit.no/acqvalab/</a>.

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



## International Perspectives on Corpus Technology for Language Learning

## An evaluation of computational,

## corpus-based approaches to language learning

## Martin Schweinberger



CREATE CHANGE



