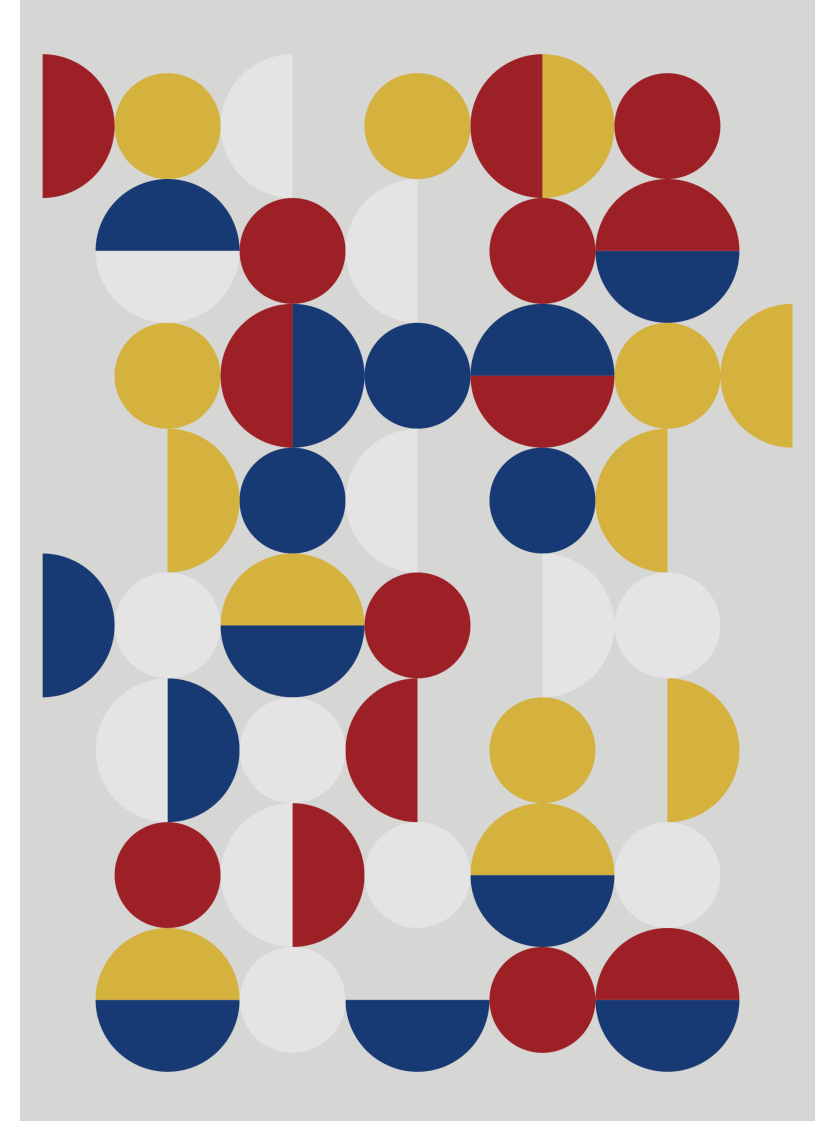


Why we need Open Science and Open Education to bridge the corpus research–practice gap

Elen Le Foll

Osnabrück University



The corpus research–practice gap



"Mind the gap" by [raghavvidya](#) (CC BY-NC-ND 2.0).

The corpus research–practice gap



"Mind the gap" by [raghavvidya](#) (CC BY-NC-ND 2.0).



Beyond the data

Analyse 1000100101000100110001101010100011
010011et010100110001110011010100101011
1 Traitement010100011000101011001101
01001Informatique010100101100100
de0101la0100011101010001



There's a lot of research out there! 489 papers in 31 years

- Can't know it all – beware “no study has ever...”
- Top journals don't give the whole picture (JCR100: 117 studies = 24%)
- Often repetitive, reinventing the wheel (but not replication)
- Reporting practices poor, inconsistent (duration, proficiency)
- Quantitative studies insufficiently robust (power, multifactorial analyses)
- Qualitative studies overly subjective (examples? q'aires ⇒ tracking?)
- Lack of mixed methods, replication studies
- Constrained by logistics
 - Longer (delayed), more ecological, outside class, after course end? (uptake)
 - Other languages, other populations, larger samples
- Researcher as teacher = niche, minority practice
- Collaboration? (teachers, courses, levels, languages, institutions, countries)
- More than just 'Does it work?', 'Do they like it?' ⇒ variation
- Comparative: EG1 vs EG2 (corpus, tools, items, profiles, training, integration – from consultation to practice and use?...)

Factors discouraging teachers' use of corpora in classroom teaching



Predominant corpus applications in higher education settings, e.g., EAP or ESP (e.g., Charles, 2014; Chen et al., 2018; Thurston & Candlin, 1998; Lee & Swales, 2006), by corpus researchers or teachers with strong research interests



Lack of suitable corpora and creative corpus-based activities for young learners (Meunier, 2019); exceptions (Kim, 2019; Crosthwaite & Stell, 2019)



Various difficulties: technical issues, suitable corpora, lack of time and confidence in operating complicated corpora (Leńko-Szymańska, 2017; Naismith, 2017; Poole, 2020; Zareva, 2017) and creation of corpus-based teaching materials



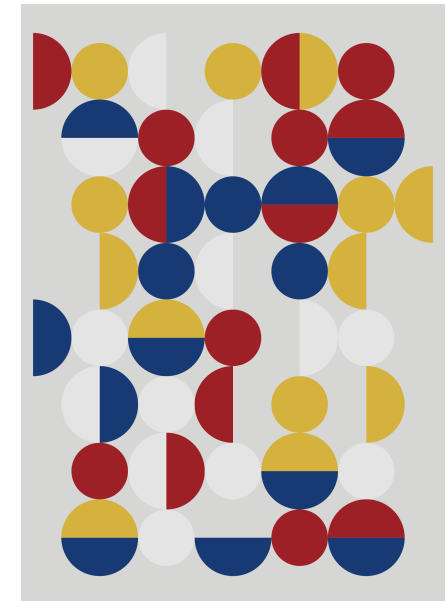
Corpus-based teacher training is largely absent from pre- and in-service teachers' education programmes or professional development (Boulton, 2017; Breyer, 2009; Callies, 2019; Chambers, 2019; Leńko-Szymańska, 2017)

Challenges of DDL

- Learners might find it technically challenging
- Time-consuming
- Learners might feel overwhelmed by the amount of data
- Better suited for learners with proficiency from an intermediate level (BUT lower level learners benefit as well) (Boulton, 2009)
- Scaffolding exercises, guidance from the teacher
- Gradually introducing DDL and direct corpus consultation
- technology in language learning – computer-anxiety (Ortega, 1997; Sullivan & Pratt, 1996)
- [...] computer anxiety is a concept-specific anxiety because it is a feeling that is associated with a person's interaction with computers (Kira & Saade, 2006, p. 32).

OPEN SCIENCE

Why we need **Open Science** and
Open Education to bridge
the corpus research–practice gap



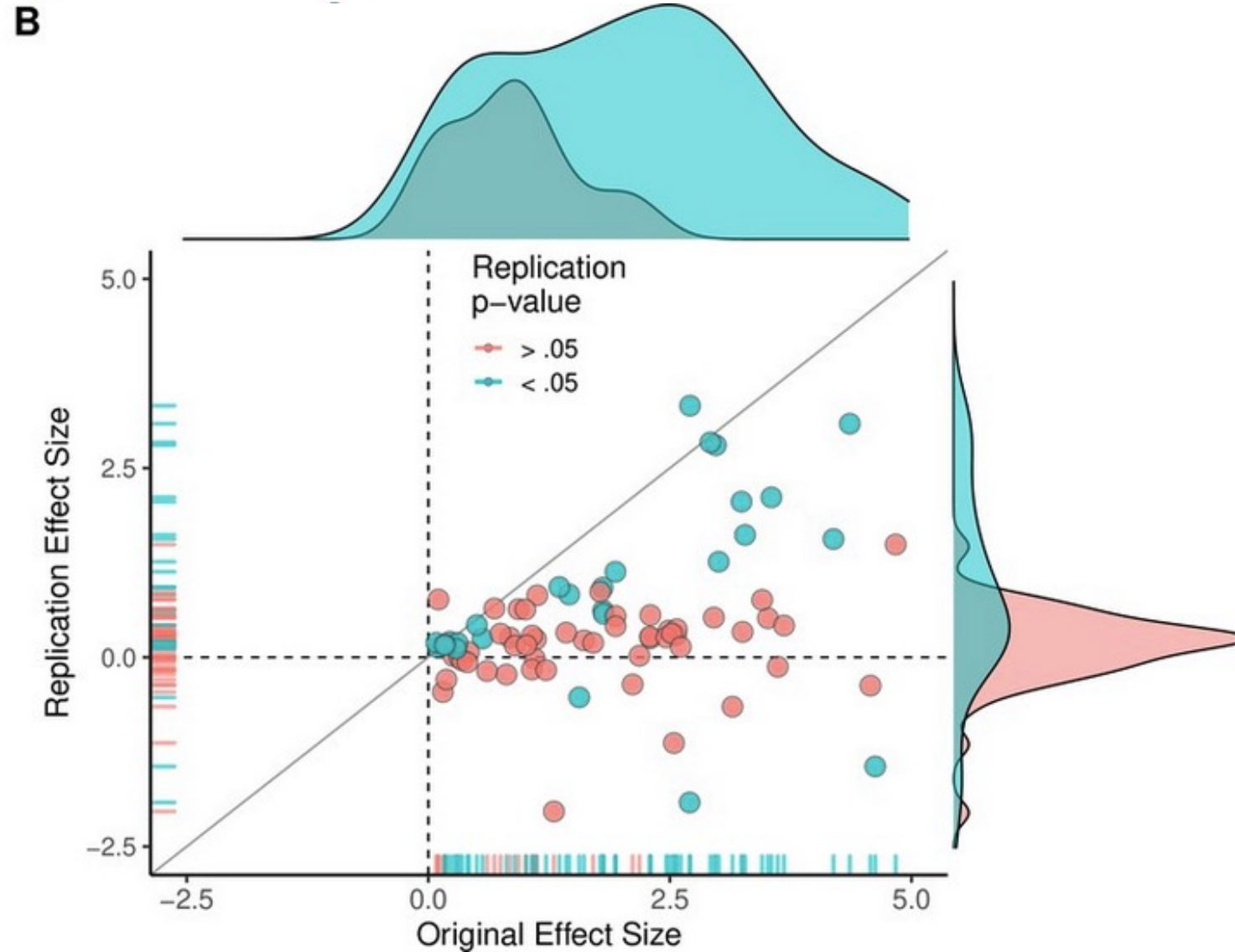
Open Science

- Replication crisis (cf. 'reproducibility crisis' and 'replicability crisis')
- Started in social psychology in the 2010s
- Prominent case: Diederik Stapel

Replication crisis

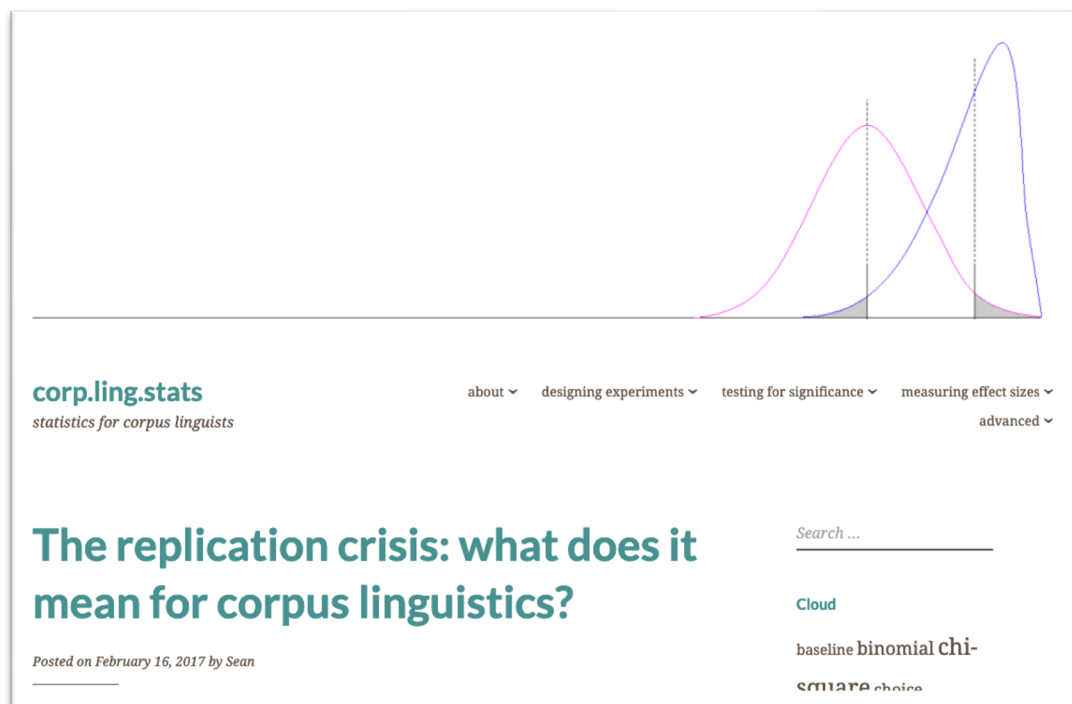
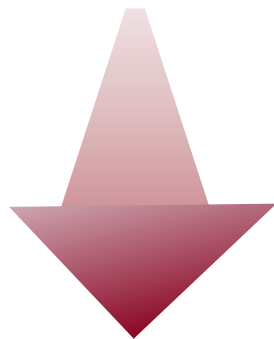
- Cancer biology research
- Intended to replicate 193 experiments, only 50 could be completed → challenges in transparency, sharing, implementation, etc.
- 92% of the replication effect sizes were smaller than original.
- Replication effect sizes were, on average, 85% smaller than the original effects.

Errington et al. (2021a) und (2021b)



Replication/reproducibility crisis

- Fraud
- Honest human errors
- “Degrees of freedom”



Wallis, Sean (2017): corp.ling.stats

Has ‘the replication crisis’ reached corpus linguistics?

👤 Tove Larsson 🕒 November 30, 2021 📁 Uncategorized

🔖 Corpus linguistics, methods, replication crisis, replication dilemma, replication studies

Discussions of ‘the replication crisis’ (by some dubbed ‘the replication dilemma’) have been ongoing in other fields such as medicine and psychology for many years (see, e.g., the [Quantitude podcast S2E11](#) for a good introduction to the topic). In short, the issue is that replication studies have revealed that the results of many (often highly influential) studies aren’t possible to reproduce. This is of course highly problematic in that failure to reproduce findings will “[undermine the credibility of theories building on them and potentially of substantial parts of scientific knowledge](#)”.

Larson, Tove (2021): Linguistics with a corpus

Best Practices in Corpus Linguistics

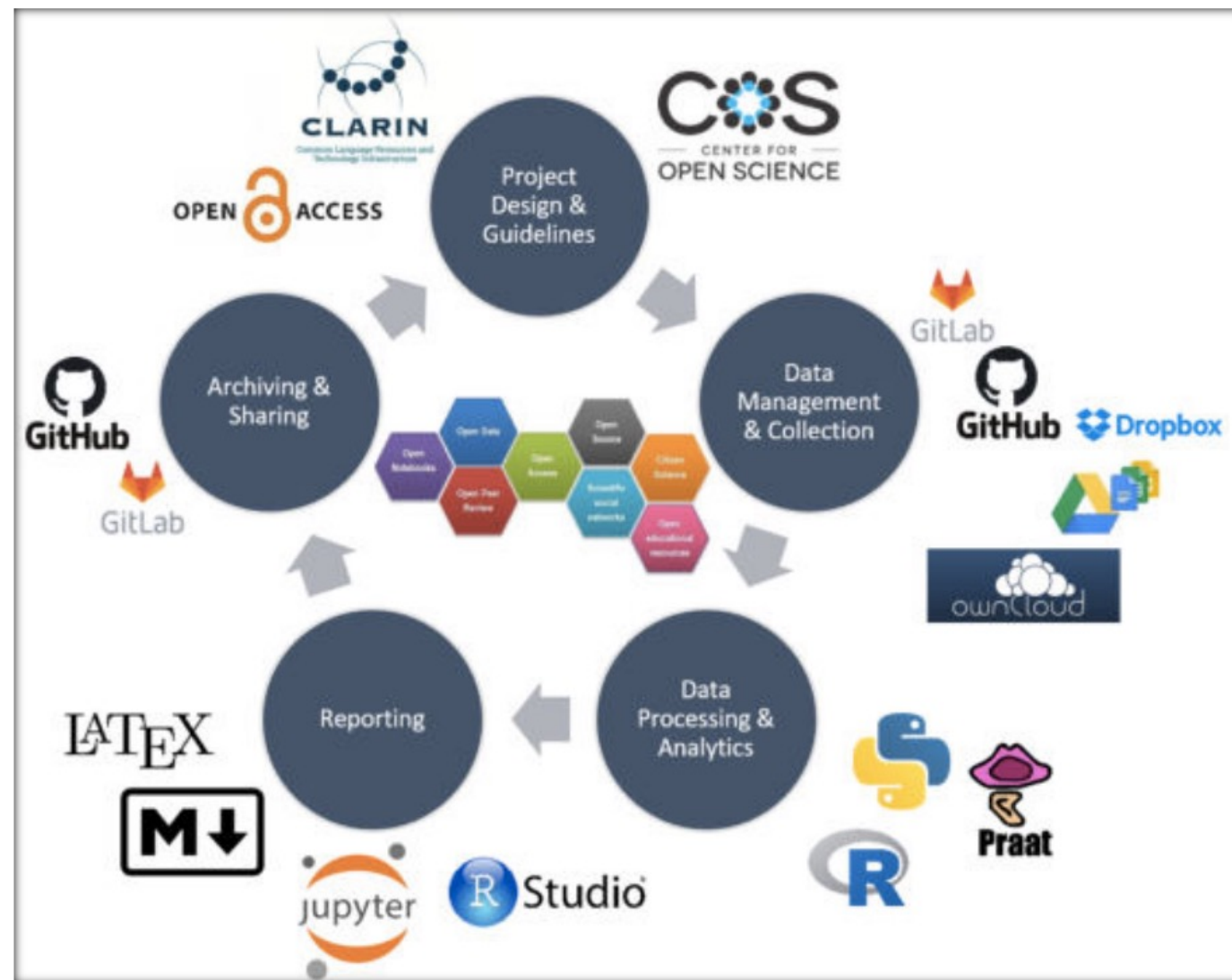
What lessons should we take from the Replication Crisis and how can we guarantee high quality in our research?

Dr. Martin Schweinberger (m.schweinberger@uq.edu.au)
available under CC license

Schweinberger, Martin (2020): ICAME 41

Open Science

- ✓ Pre-registration
- ✓ Better research designs
- ✓ Better reporting
- ✓ Sharing materials, data and code
- ✓ Encouraging replication studies



Schweinberger, Martin (2020): ICAME 41

Replication in (applied) (corpus) linguistics

<https://doi.org/10.1017/S0261444821000367>

Call for P

Language Teaching (2021), 1–9
doi:10.1017/S0261444821000367

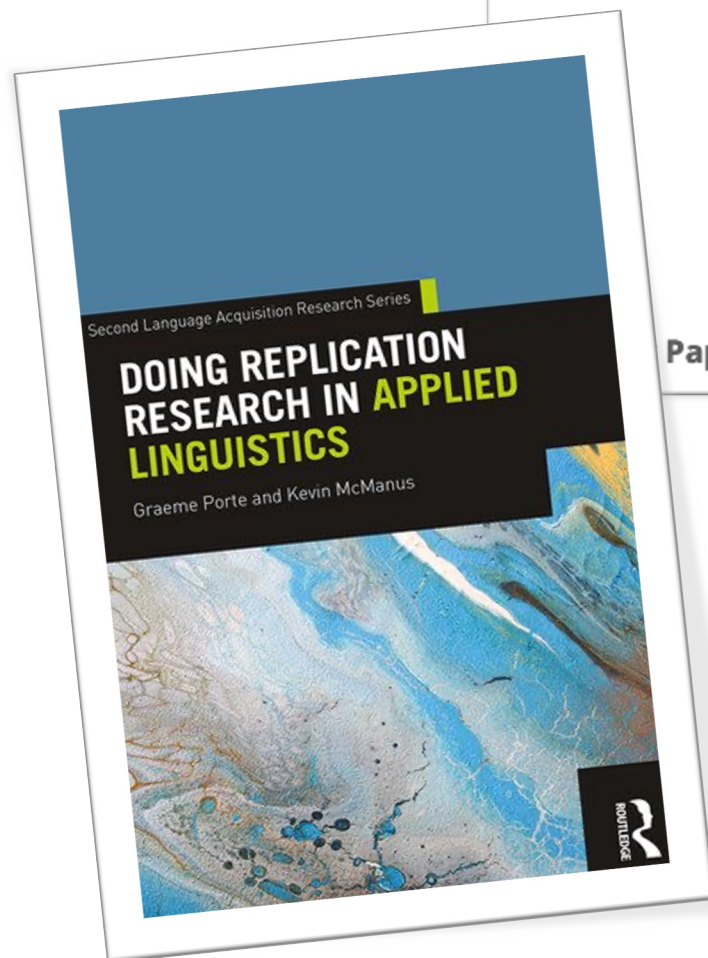
CAMBRIDGE
UNIVERSITY PRESS

REPLICATION RESEARCH

Replicating corpus-based research in English for academic purposes: Proposed replication of Cortes (2013) and Biber and Gray (2010)

Taha Omidian^{1*}, Oliver James Ballance² and Anna Siyanova-Chanturia^{1,3}

Pape



HOME ABOUT LADAL ▾ EVENTS ▾ R BASICS ▾ DATA SCIENCE BASICS ▾ TUTORIALS ▾ FOCUS STUDIES ▾ RESOURCES ▾

Introduction

1 Data Management

2 Optimizing workflows in RStudio

Citation & Session Info

ISLE 6 workshop: Replication and Reproducibility in English Corpus Linguistics

Martin Schweinberger (UQ, UiT)

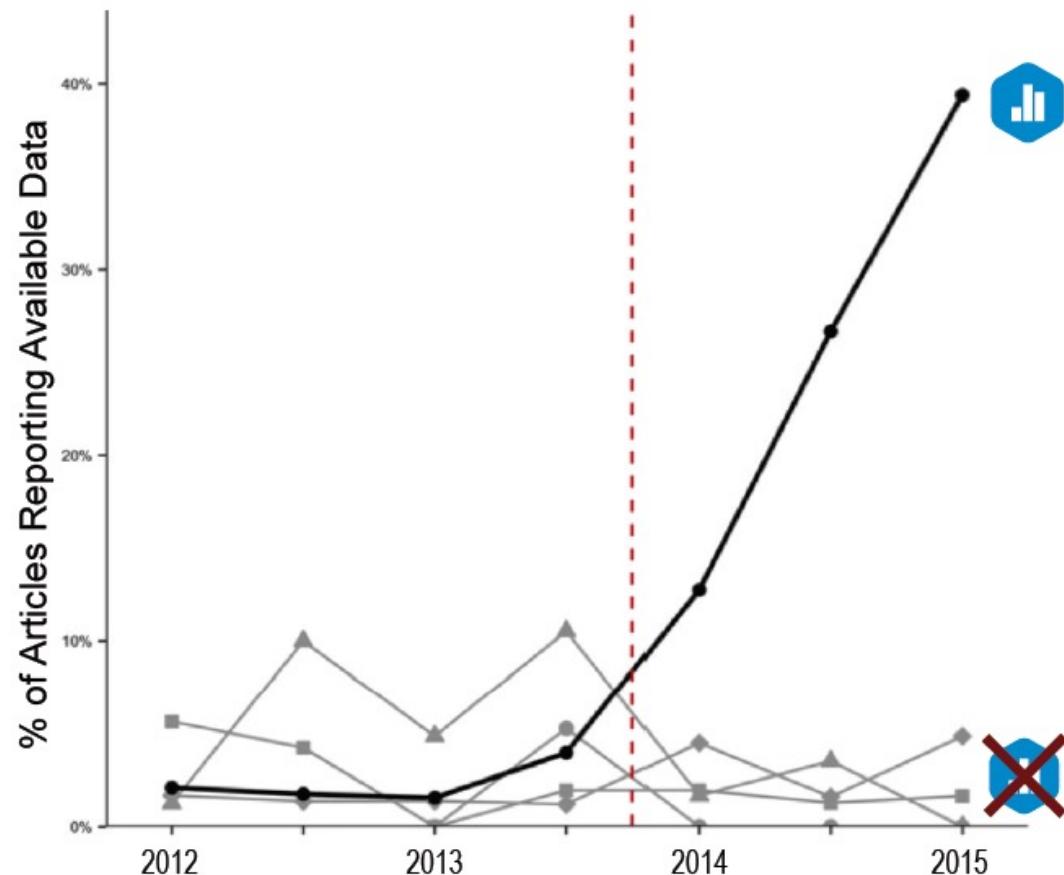
2021-11-04



https://slcladal.github.io/isle6_reprows.html

Badges

- Open Data Badge
 - Data
 - Metadata
 - Reproducible code
- Open Materials Badge
 - Materials needed for replication, e.g., questionnaires, teaching materials for pedagogical interventions, coding schemes, etc.
 - IRIS Digital Repository of Data Collection Materials (<http://www.iris-database.org>)



<https://www.cos.io/initiatives/badges>

Badges



- In the April 2019 issue of *Psychological Science* all 14 research articles received the Open Data badge.
- Crüwell et al. (2022) set out to reproduce the results of the 14 articles:
 - All 14 articles provided *some* data.
 - Only 6/14 provided analysis code or scripts.
 - Just 1/14 was exactly reproducible.
 - 3/14 essentially reproducible with minor deviations.

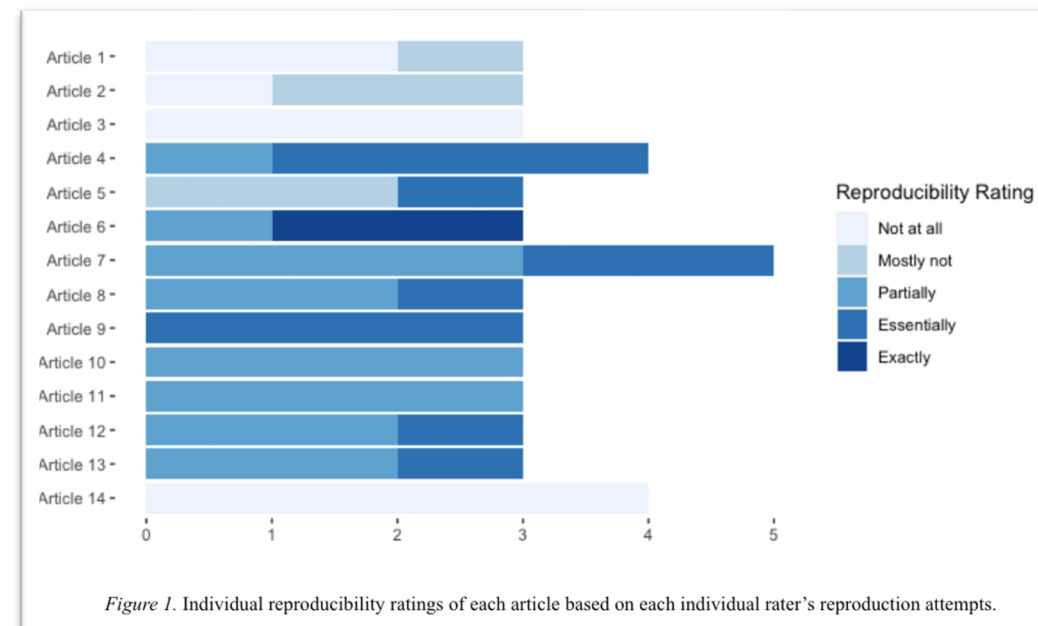
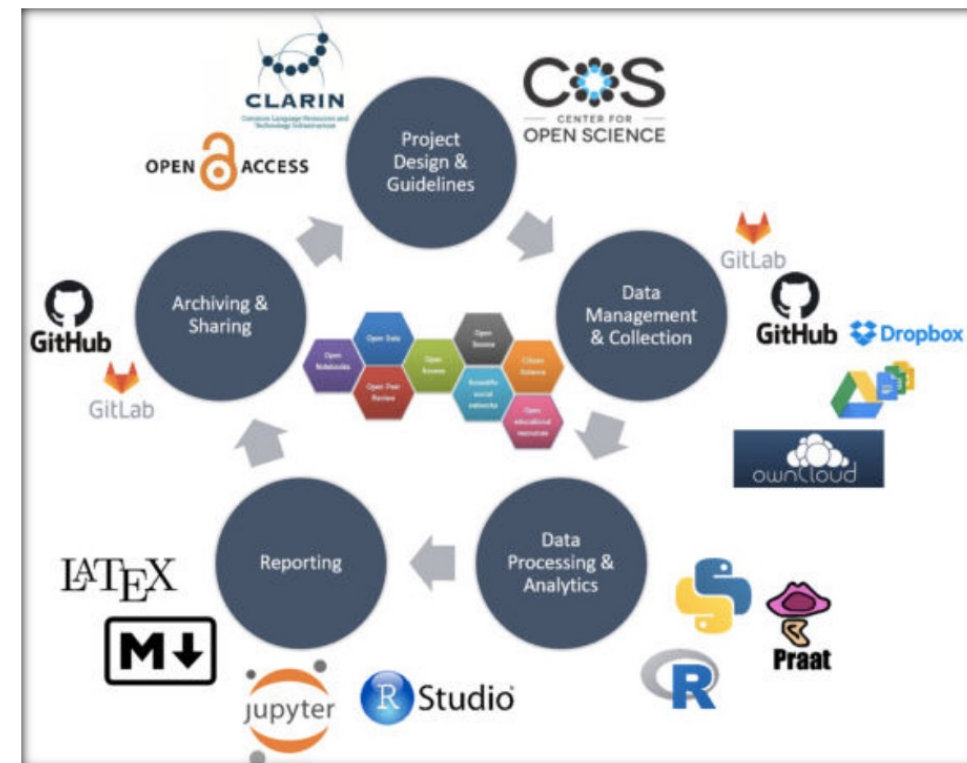


Figure 1. Individual reproducibility ratings of each article based on each individual rater's reproduction attempts.

Example 1: Textbook English (PhD thesis)

- Open Science statement
 - Open data (but not corpora for copyright reasons)
 - Published code (GPL-3.0 License)
 - Use of open source tools
- Advantages
 - Transparency
 - Reproducibility
 - Re-usability of data and code for myself (!) and others
 - Potential for more collaborations
 - Data and code can be cited.
- Risks
 - Transparency
 - Vulnerability
 - Scooping?



Schweinberger, Martin (2020): ICAME 41



- Textbook English
 - Chapter 1: Introduction
 - Chapter 2: Literature review
 - Appendix 2.1
 - Chapter 3: Research aims and data
 - Appendix 3.1
 - Appendix 3.2
 - Appendix 3.3
 - Appendix 3.4
 - Chapter 4: Exploring the progressive in Textbook English
 - Appendix 4.1
 - Appendix 4.2
 - Appendix 4.3
 - Appendix 4.4

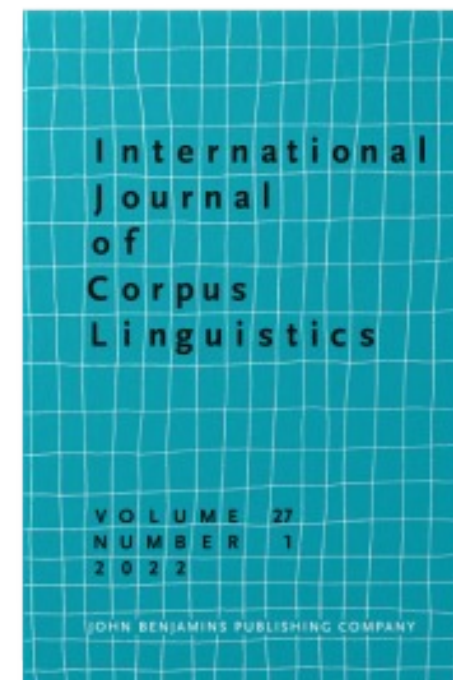
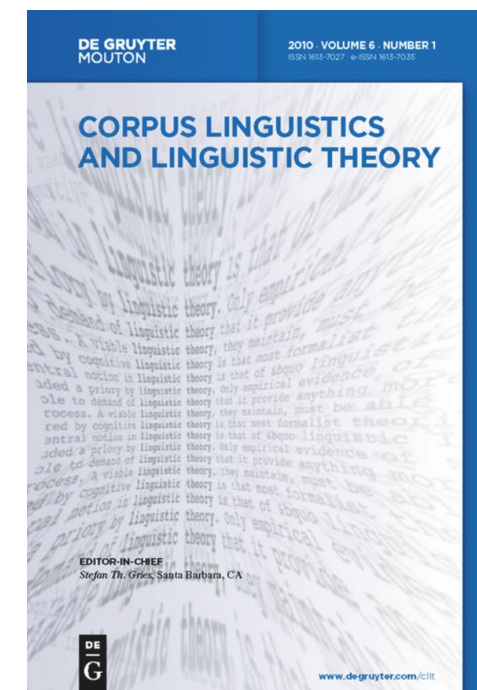
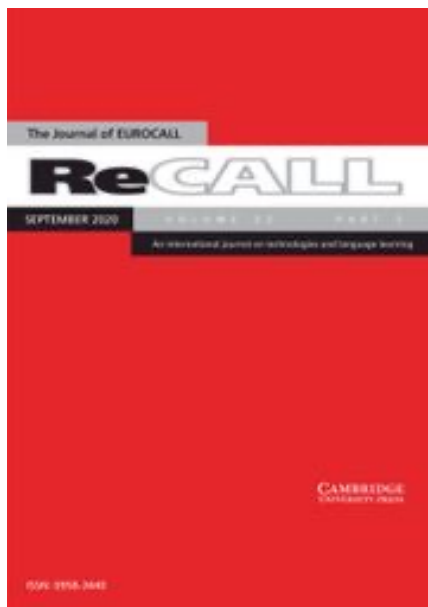
(Corpus) research accessibility

- Physical access
 - Location of research report/tools/materials
 - Financial cost
 - Format, technologies required for access
 - Physical ability to access research report/materials
- Intellectual access
 - Factors that can affect intellectual access include: information-seeking behaviours, language/dialect, literacy, education, technological literacy, cognitive ability, vocabulary, subjective views, etc.
- Social access
 - Social norms
 - (Subjective) worldviews

Burnett, Gary, Paul T. Jaeger & Kim M. Thompson (2008).

(In)accessibility of corpus research

- Hundreds of DDL publications
- Articles behind paywalls
- APCs: between 1000 EUR and 3255 USD + tax



Accessibility of corpus research

RiCL Research in Corpus Linguistics

About ▾ FirstView articles Current Archives Submissions Announcements

JOURNAL OF CORPORA
AND DISCOURSE STUDIES



Corpus linguistics

A guide to the methodology

Anatol Stefanowitsch

Textbooks in Language Sciences 7



See also: Kowalczyk, Olivia S., Alexandra Lautarescu, Elisabet Blok, Lorenza Dall'Aglio & Samuel J. Westwood. 2022. What senior academics can do to support reproducible and open research: a short, three-step guide. *BMC Research Notes* 15(1). 116. <https://doi.org/10.1186/s13104-022-05999-0>.

Open Journals (L2 Teaching Learning)

Journal of Digital Humanities <http://journalofdigitalhumanities.org/>

Journal of Interactive Technology and Pedagogy <https://jitp.commons.gc.cuny.edu/>

Language Learning & Technology <https://www.lltjournal.org/>

L2 Journal https://escholarship.org/uc/uccllt_l2/

Open Accessible Summaries In Language Studies (OASIS) <https://oasis-database.org>

Open Journal Systems <https://pkp.sfu.ca/ojs/>

Open Linguistics <https://www.degruyter.com/view/journals/opli/opli-overview.xml>

Open Journal of Modern Linguistics <https://www.scirp.org/journal/ojml/>

Reading in a Foreign Language <http://nflrc.hawaii.edu/rfl/about.html>

Second Language Research and Practice <http://www.slrpjournal.org/>

Source: Blyth, Carl S. & Joshua J. Thoms (eds.). 2021. *Open Education and Second Language Learning and Teaching: The Rise of a New Knowledge Ecology*. Multilingual Matters. <https://doi.org/10.21832/9781800411005>.

Paths of Open Access

- Preprint
 - Author submitted manuscript
- Post-print
 - Author's accepted manuscript (AAM)
 - Version of Record (VoR)



Source: open-access.network (2021), [Paths of Open Access](#) (CC BY 4.0 International)

Sherpa Romeo

About

Search

TJ List

Statistics

Help

Support Us

Contact

Admin

Welcome to Sherpa Romeo

Sherpa Romeo is an online resource that aggregates and analyses publisher open access policies from around the world and provides summaries of publisher copyright and open access archiving policies on a journal-by-journal basis.

Enter a journal title or issn, or a publisher name below:

Journal Title or ISSN

Search

Publisher Name

Search

Browse by Country

Browse by Publisher

SERVICES

Open access services from Jisc

Services to support open

SERVICE

Sherpa Services

Helping authors and

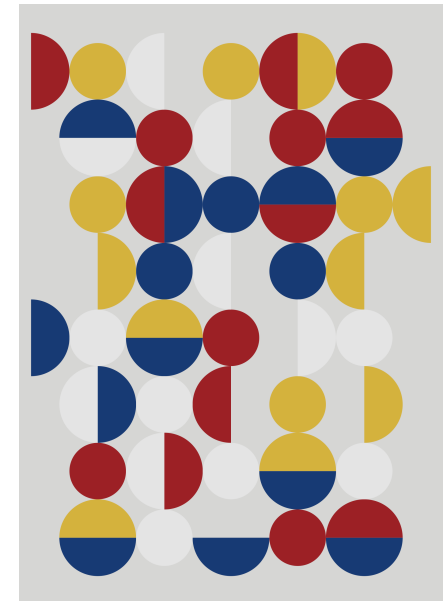
GUIDE

Managing open access costs

A guide from Jisc

OPEN EDUCATION

Why we need Open Science and
Open Education to bridge
the corpus research–practice gap



Intellectual accessibility

- Target audience?
- Format?
- Language?

How to draw an Owl.

"A fun and creative guide for beginners"

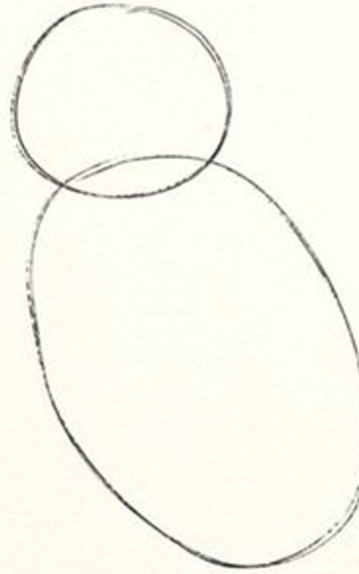
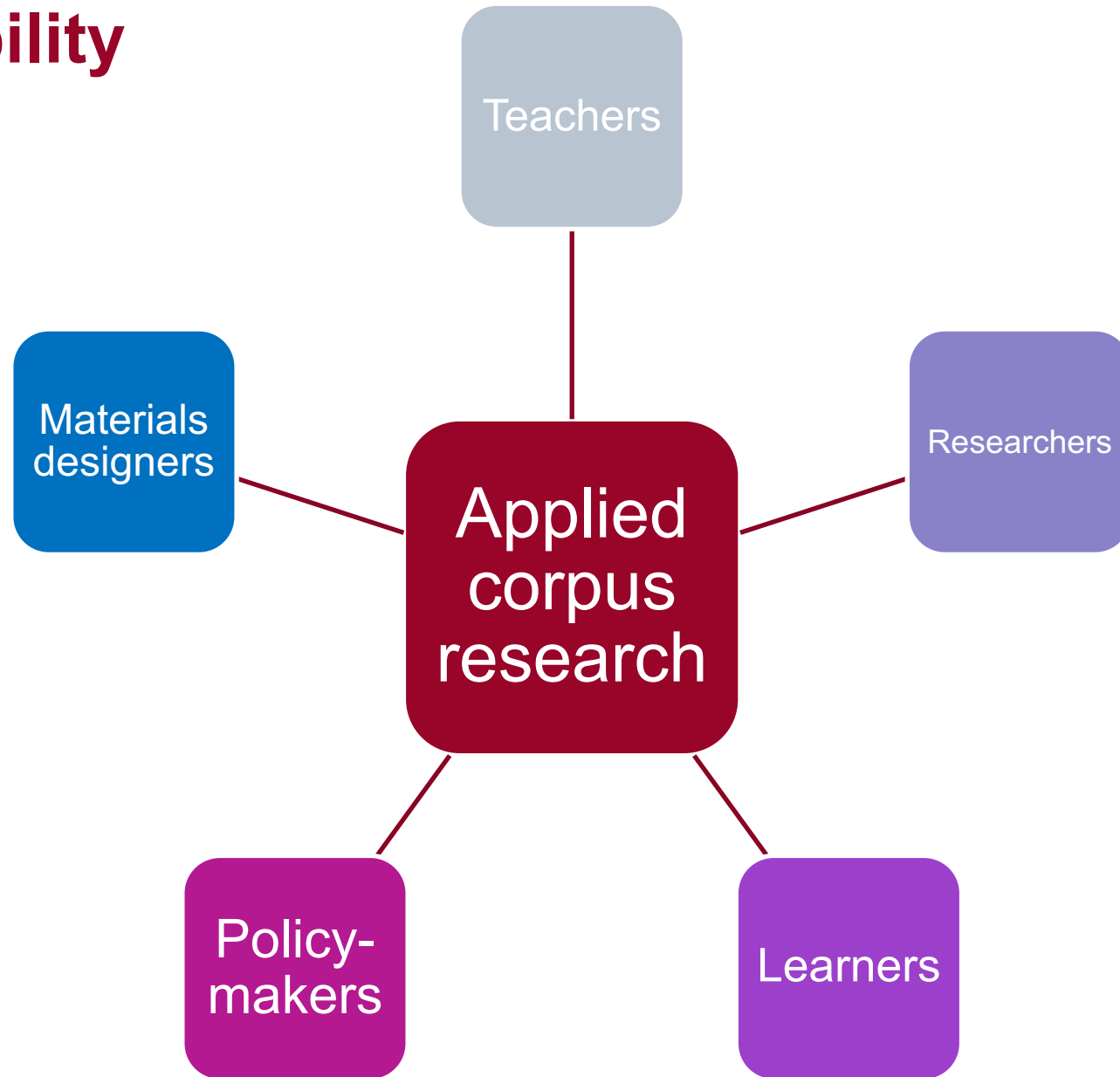


Fig 1. Draw two circles



Fig 2. Draw the rest of the damn Owl

Social accessibility



Example 2: YouTube

https://www.youtube.com/results?search_query=corpus+linguistics

Google

corpus linguistics

Corpus Linguistics: The Basics
57K views • 6 years ago
Phoneme
This is a short introduction to the idea of corpus linguistics, which should help
Subtitles
4:59

#1 Introduction to Corpus Linguistics - What is Corpus Linguistics (For Absolute Beginners)
14K views • 1 year ago
Yassine Iabdounane
Hello there! This video is made specifically for those who don't know much about corpus linguistics.
Subtitles
18:06

Corpus Linguistics for Beginners
Yassine Iabdounane
#1 Introduction to Corpus Linguistics - What is Corpus Linguistics? (For Absolute Beginners) • 18:06
#2 Introduction to Corpus Linguistics - Types of Corpora • 21:00
VIEW FULL PLAYLIST

MOOC - Corpus linguistics: method, analysis, interpretation
34K views • 8 years ago
Lancaster University

YouTube DE

corpus linguistics education

Corpus Linguistics
By Yassine Iabdounane

#1 Introduction to Corpus Linguistics - What is Corpus Linguistics? (For Absolute Beginners)
14,674 views • 9 Apr 2020
539 DISLIKE SHARE SAVE ...

Yassine Iabdounane
1.02K subscribers
SUBSCRIBE

<https://www.youtube.com/watch?v=ePD46YjYRzQ>

Example 2: YouTube

https://www.youtube.com/results?search_query=how+to+use+a+corpus

Google

how to use a corpus

SIGN IN

FILTERS

English-Corpora.org

Quick intro to using corpora
→ english-corpora.org

Basic corpus queries: First steps on english-corpora.org
9.4K views • 1 year ago
Elen Le Foll

This video was first made as a revision video for the course "Designing and Evaluating Materials for Language Teaching" ...

0:00 Welcome to this introduction to using corpora for first time users we'll be looking at the web interface English - corpora org and i...

Using a Corpus
2.4K views • 4 years ago
WUWritingCenter

Learn what a corpus is and how you can use one to help you with your English word choice in academic writing.

Subtitles

Using Corpora in the Language Classroom | The New School
64K views • 10 years ago
The New School


Randi Reppen is professor of Applied Linguistics at Northern Arizona University where she teaches in the MATESOL and Ph.D. in ...

THE NEW SCHOOL

1:17:32

Example 2: YouTube

https://www.youtube.com/playlist?list=PLAq6uhS_0brxTW99jeZjxdsIQ5BRy1Eb5



English-Corpora.org

users related resources my account upgrade

pes, looking at variation, corpus-based resources.

so download the corpora for use on your own computer.

Download	# words	Dialect	Time period	Genre
	14 billion	6 countries	2017	Web
	10.0 billion+	20 countries	2010-yesterday	Web
	279 million+	20 countries	Jan 2020-yesterday	Web
	1.9 billion	20 countries	2012-13	Web
	1.9 billion	(Various)	2014	
	1.0 billion	American	1990-2019	
	400 million	American	1810-20	
	325 million	6 countries		

Quick intro to using corpora

→ english-corpora.org

@ElenLeFoll

Elen Le Foll

Basic corpus queries: First steps on english-corpora.org

9,129 views • 18 May 2020

LIKE DISLIKE SHARE SAVE ...

Elen Le Foll

SUBSCRIBE

- 1 Quick intro to using corpora → english-corpora.org 14:43
- 2 Building your own specialised corpus → english-corpora.org 9:47
- 3 0:37
- 4 Selecting and presenting words and phrases in context → english-corpora.org 11:25
- 5 Quick intro to using corpora → Search Engine 12:39
- 6 Building your own corpus → Search Engine 8:18
- 7 Using Corpus Query Language (CQL) for the first time → Search Engine 11:27

Open Education

- Attempts to create opportunities for learners to:
 - **Access** education, open educational resources, open textbooks, and open scholarship
 - **Collaborate** with others, across the boundaries of institutions, institutional systems, and geographic locations
 - **Create** and **co-create** knowledge openly
 - **Integrate** formal and informal learning practices, networks, and identities.

https://www.wikiwand.com/en/Open_education

Accessibility of corpora and corpus tools

<https://languages-cultures.uq.edu.au/event/session/7420>

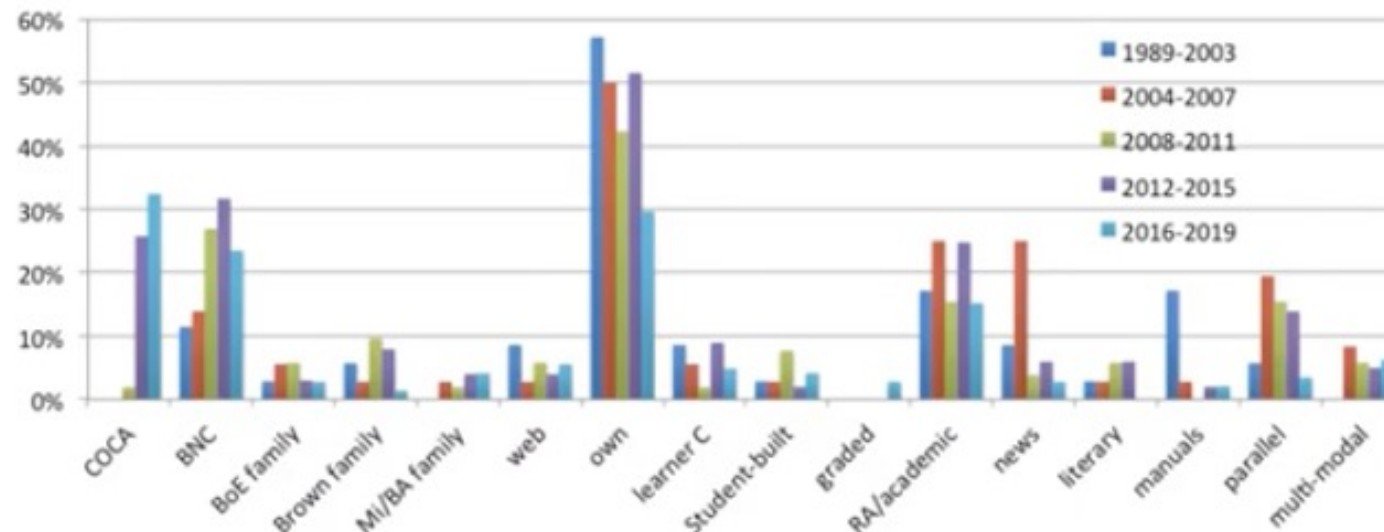


Named in 67% of hands-on studies

- 27% use more than one (1989-2003 = 8%; 2016-19 = 28%)
- 24% BNC, 20% COCA
- 42% included a local corpus (decreasing from 2004-07)
- 19% specialised corpora (e.g. academic); JCR100 = 24% vs 17%

Parallel 9%? Learner corpora 6%? Multimodal 6%? Self-compiled 4%?

Graded corpora 4%? Manuals 3%? SkELL 1%?



- Cost?
- Registration needed?
- Ease of use?
- Minimum hardware requirements?
- Minimum technical literacy needed?

Open Corpus Linguistics Education

- Corpus Linguistics MOOC (Leads: Tony McEnery & Vaclav Brezina)
- Corpora in language teaching (Moodle course) (Lead: Agnieszka Leńko-Szymańska)
- Corpus-Aided Platform for Language Teachers (Lead: Ma Qing Angel)
- Corpus for Schools (Lead: Dana Gablasova)
- Integrating Corpora (Lead: Nina Vyatkina)



Open Education Resources (OERs)



Open Educational Resources

- Teaching, learning, and research materials that are either (a) in the public domain or (b) licensed in a manner that provides everyone with free and perpetual permission to engage in the 5R activities.
 - **Retain** – make, own, and control a copy of the resource
 - **Reuse** – use your original, revised, or remixed copy of the resource publicly
 - **Revise** – edit, adapt, and modify your copy of the resource
 - **Remix** – combine your original or revised copy of the resource with other existing material to create something new
 - **Redistribute** – share copies of your original, revised, or remixed copy of the resource with others

<https://creativecommons.org/about/program-areas/education-oer/>

Corpora as Open Educational Resources (OERs)?

- Open digital DDL materials integrated with open corpora

Vyatkina, Nina. 2020. Corpora as Open Educational Resources for Language Teaching. *Foreign Language Annals* 53(2). 359–370. <https://doi.org/10.1111/flan.12464>.

Open L2 Corpora

Chinese corpus <http://corpus.leeds.ac.uk/query-zh.html>

English-Corpora <https://www.english-corpora.org/>

Lextutor (English) <https://lex tutor.ca/>

Multilingual Corpus of Second Language Speech (MuSSeL) https://l2trec.utah.edu/multi-Lingual_Speech_Corpus.php

NINJAL (Japanese corpora) <https://www.ninjal.ac.jp/english/database/type/corpora/>

Southeast Asian Languages Library (Sealang) <http://sealang.net/library/>

Talk Bank (English) <https://talkbank.org/>

Source: Blyth, Carl S. & Joshua J. Thoms (eds.). 2021. *Open Education and Second Language Learning and Teaching: The Rise of a New Knowledge Ecology*. Multilingual Matters. <https://doi.org/10.21832/9781800411005>.

Example 3: The co-creation of an OER with student teachers

Creating Corpus-Informed Materials for the English as a Foreign Language Classroom

A step-by-step guide for (trainee) teachers using
online resources

Elen Le Foll

 **Creative Commons Attribution NonCommercial**

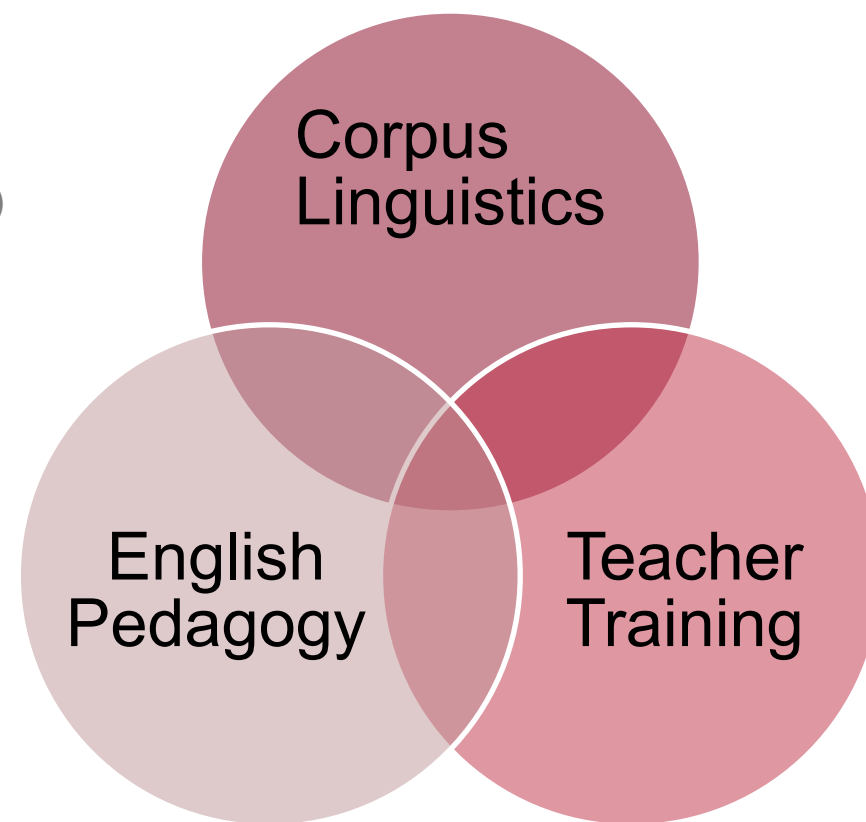
[READ BOOK](#)



<https://elenlefol.pressbooks.com/>

A project-based seminar

- For M.Ed. students training to become English teachers
- Focus on language teachers' needs (cf. Römer 2010)
- Inspired by previous similar endeavours (cf. Breyer 2009; Hüttner, Smit & Mehlmauer-Larcher 2009; Leńko-Szymańska 2014)

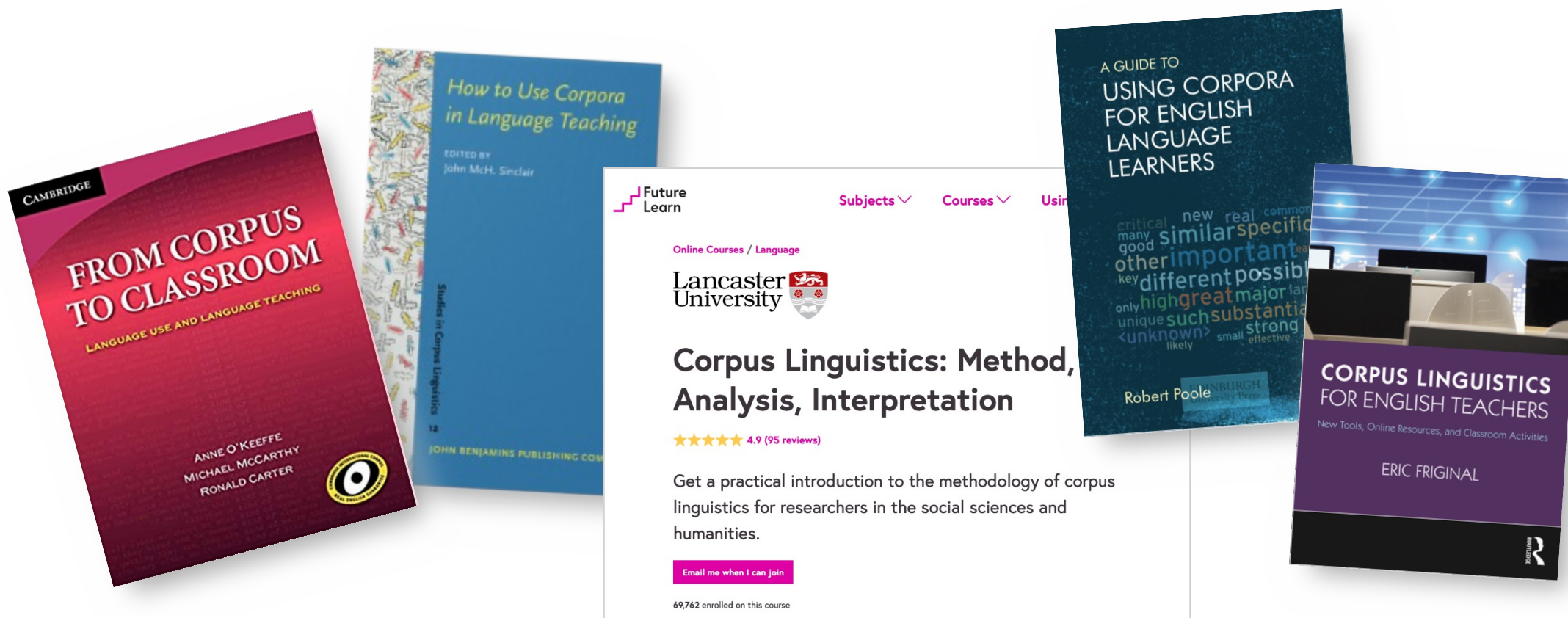


Learning objectives of the seminar

- Understand the value of using corpora in language education.
- Be able to do so autonomously using a range of tools and methods.
- Design corpus-informed ELT materials.
- Integrate corpus work in the curriculum and lesson plans.
- Explain to other teachers how to use corpora.
- Practise peer review and learn from critical feedback.

Collective project seminar aim

- Publish “*A Practical Guide to Using Corpora for English as a Foreign Language Teachers*” for teachers from around the world to use and draw inspiration from.



Process

Corpus Linguistics

- Theoretical background
- Introduction to corpora, tools and methods



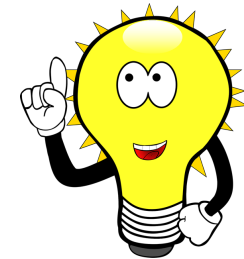
English Language Teaching

- A problem-solving approach
- Materials design
- Task design



Project idea pitches

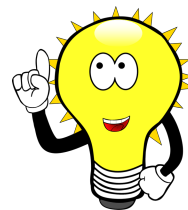
- Elevator-pitch presentations
- Discussion



Process

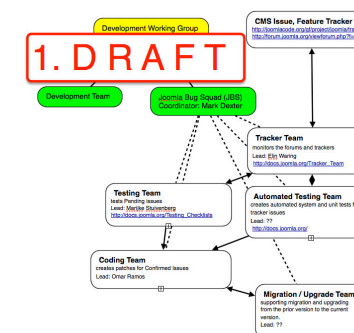
Idea development

- A problem-solving approach
- Corpus linguistics and pedagogical research



Chapter draft

- Dealing with practical issues
- Designing lesson plan and tasks



Final chapter version

- Incorporating peer feedback
- Self-reflection
- Publication



Creating Corpus-Informed Materials for the English as a Foreign Language Classroom

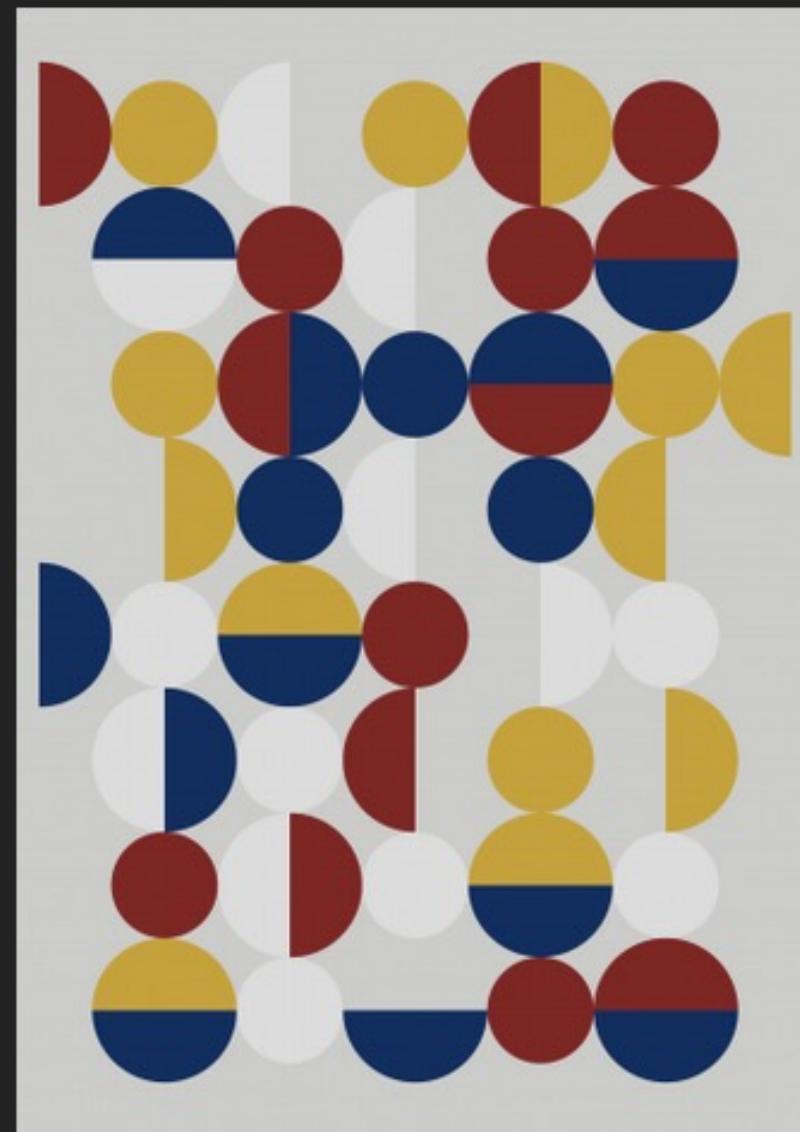
A step-by-step guide for (trainee) teachers using online resources

Elen Le Foll



Creative Commons Attribution NonCommercial

READ BOOK



<https://elenlefol.pressbooks.com/>

Enabling the 5Rs

- Available as an web-book: <https://elenlefol.pressbooks.com> (recommended version for reading)
- But also on public repositories (<https://zenodo.org/record/4992504> and <https://osnadocs.ub.uni-osnabrueck.de/handle/urn:nbn:de:gbv:700-202108205284>) as:
 - Editable XML format
 - Editable ODT format
 - HTML format (for offline use)
 - PDF versions (with and without links)

Conclusions & Outlook from OER project

Lessons learnt

- Less is more.
- Fewer tools
- Web-based only
- Short videos on basic corpus functions
- Focus on materials design
- Provide more (but not too many!) examples.
- Pedagogical knowledge often lacking.
- Being creative is very tricky for some.

What's next?

- Use of OER by lecturers and professors from universities across the world
- Option to expand, re-work, translate parts of the OER
- Delivering in-service teacher training workshops
- Cooperation with in-service teachers to try out the lesson plans and tasks.

Summary I: Why we need (more) Open Science and Open Education to bridge the corpus research-practice gap

- ✓ Collaboration
 - ✓ Strengthen methods
 - ✓ Increase sample sizes
 - ✓ Stop re-inventing the wheel
 - ✓ Involve practitioners from the word go
- ✓ Accessible research materials, code and data
 - ✓ Encourages replication, including by students, teacher-researchers and other practitioners
- ✓ More robust results
 - ✓ More likely to be generalisable
- ✓ Accessible research results
 - ✓ Increases potential audience
 - ✓ Increases potential impact

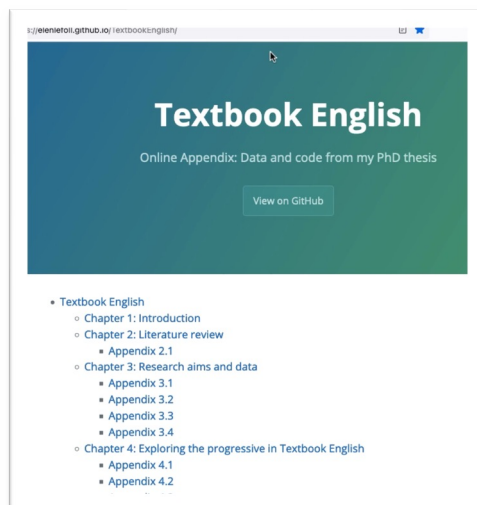
Summary II

■ Open Science

- Example 1: Online Appendix to PhD thesis

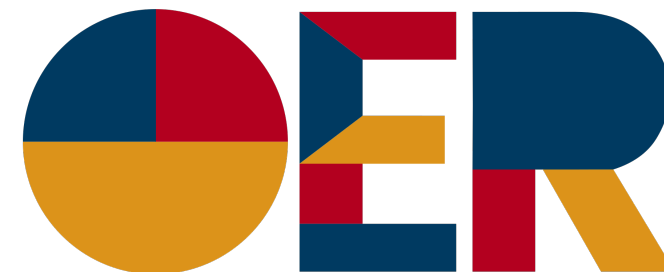
■ Open Education

- Example 2: YouTube channels with short tutorial videos on how to use corpora
- Example 3: OER on creating corpus-informed teaching materials



DISCUSSION

Do we need Open Science and Open Education to bridge the corpus research–practice gap? If so, how?



Open Educational Resources

Questions? Comments? Suggestions?

- **Elen Le Foll**
- Osnabrück University
- elefol@uos.de

