Computational approaches and the Humanities: what might await us?

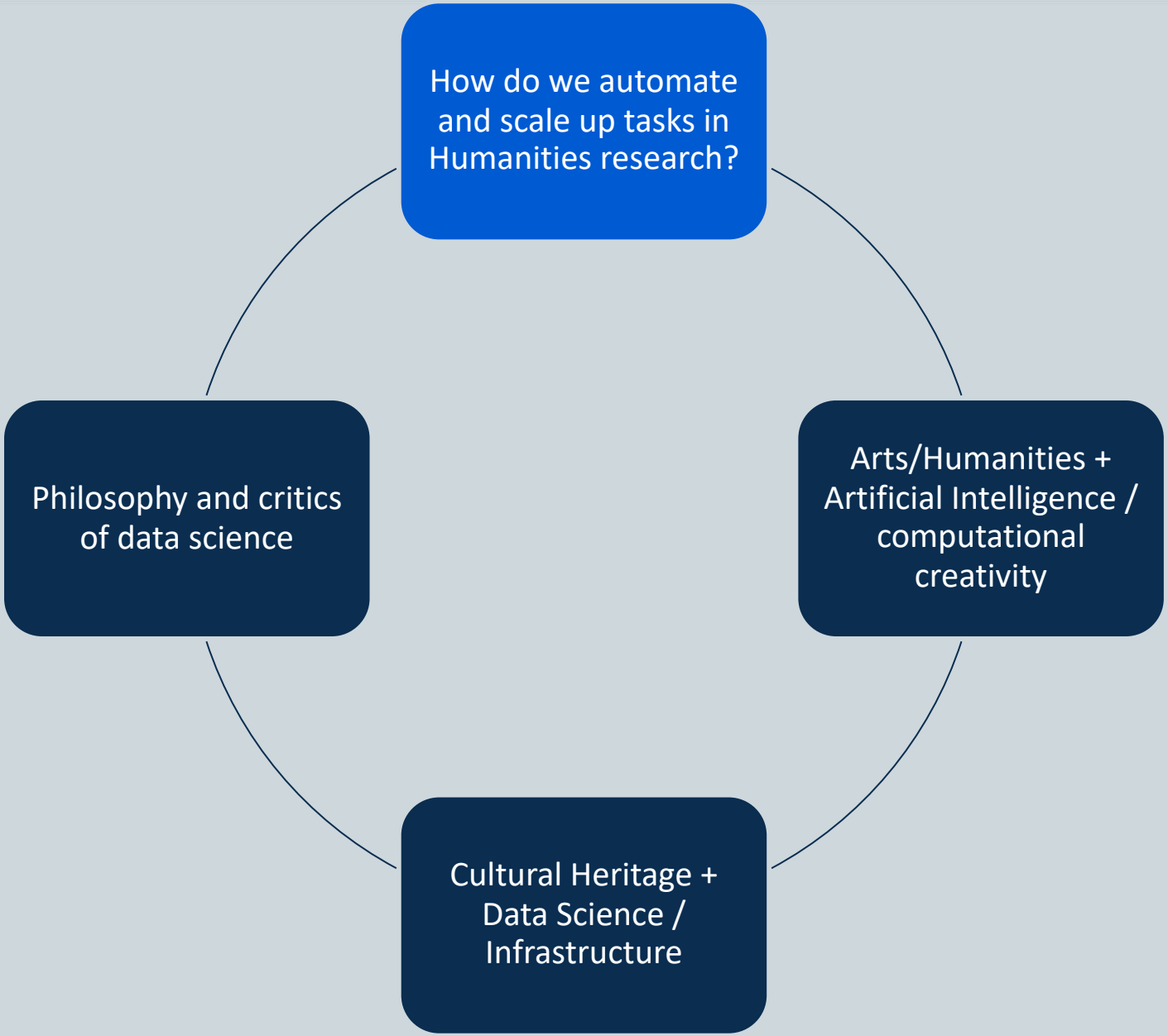
Barbara McGillivray
1 September 2022

Computational Thinking in the Humanities

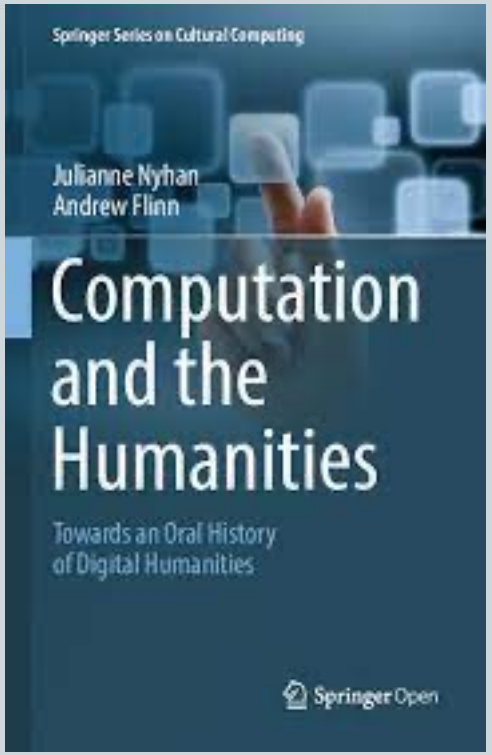
Humanistic scholarship has produced and collected data for centuries

What is different now?

Computational humanities/humanities & data science



An increasing interest



LIVING WITH MACHINES

The Alan Turing Institute

HUMANITIES & DATA SCIENCE

Workshop on Computational Methods in the Humanities 2022 (COMHUM 2022)

9 June - 10 June

« Workshop on Computational Methods in the Humanities 2018 (COMHUM 2018)

COMHUM 2022

Special track: character network construction and analysis

DHd Frankfurt/Main

COMPUTATIONAL HUMANITIES

UNIVERSITY OF LEIPZIG

The Turing white paper

McGillivray, Barbara et al. (2020). The challenges and prospects of the intersection of humanities and data science: A White Paper from The Alan Turing Institute.

Figshare. dx.doi.org/10.6084/m9.figshare.12732164

Quantitative thinking and doing

The example-based approach

Corium 'skin' is currently attested as a thematic neuter at all stages of the languages, but in Plautus' plays (e. g. *Poen.* 139: ~ 197 BC) and in Varro's *Menippeae* (*Men.* 135: 80–60 BC) there also occurs the masculine gender

Rovai (2012)

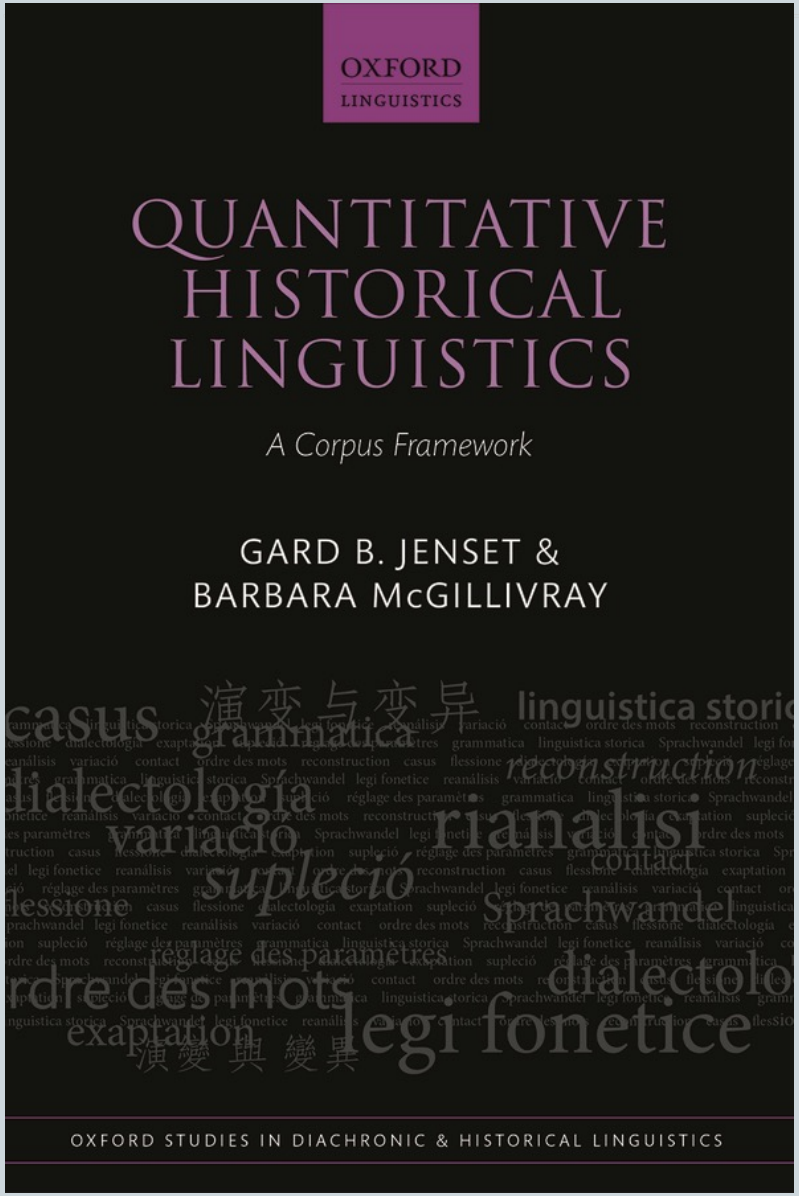
- How were the examples selected?
- Reproducibility issues

Dealing with multivariate phenomena

“especially in Sophocles and Euripides one can find relatively more subject-oriented resultatives than in the historians”. (Bentein 2012:187–188)

“the active transitive perfect (with an anterior meaning) is indeed rather uncommon in fifth century writers” (Bentein 2012:189)

Quantitative historical linguistics



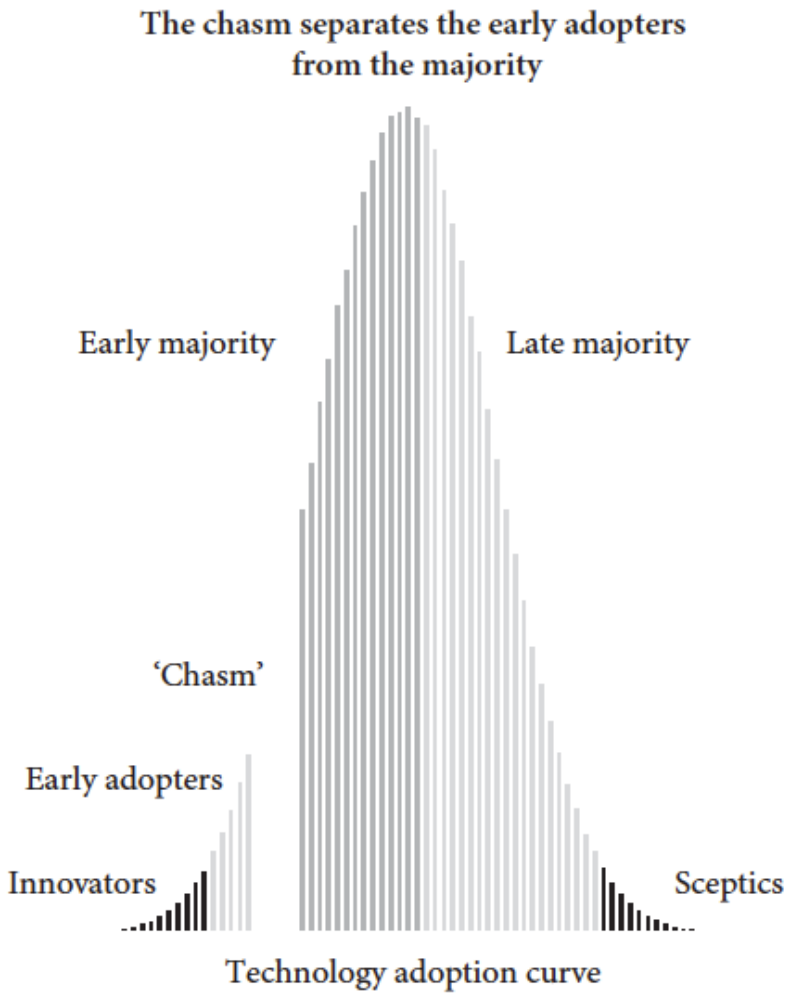


FIGURE 1.1 Technology adoption life cycle modelled as a normal distribution, based on Moore (1991, 13).

Quantitative historical linguistics

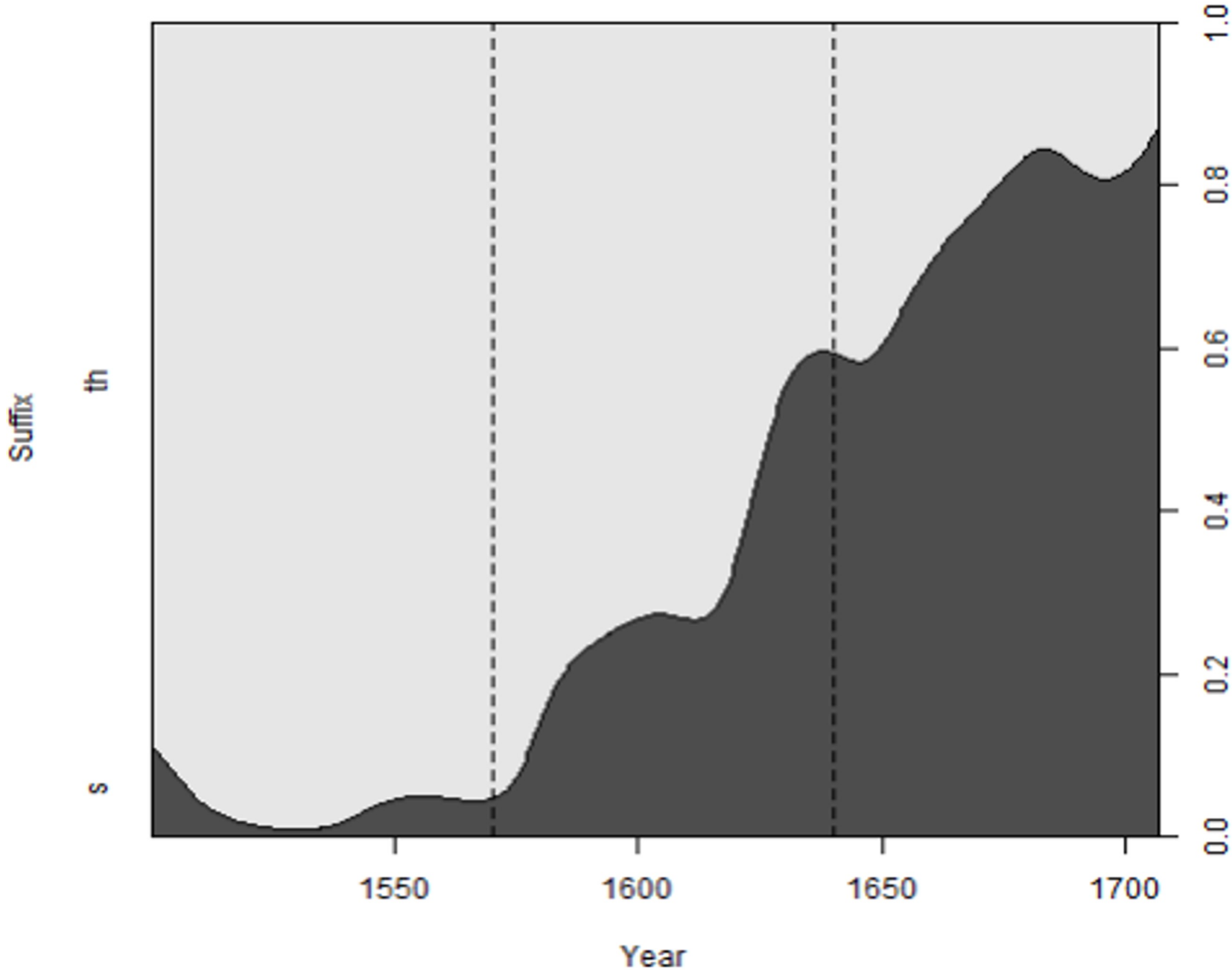
TABLE 1.3 95% confidence intervals for the percentage of quantitative papers in *Language* 2012 and the historical sample. Note that the confidence intervals do not overlap

	Proportion of quantitative papers	95% confidence interval
Language	80%	[60%, 100%]
Historical sample	40%	[28%, 52%]

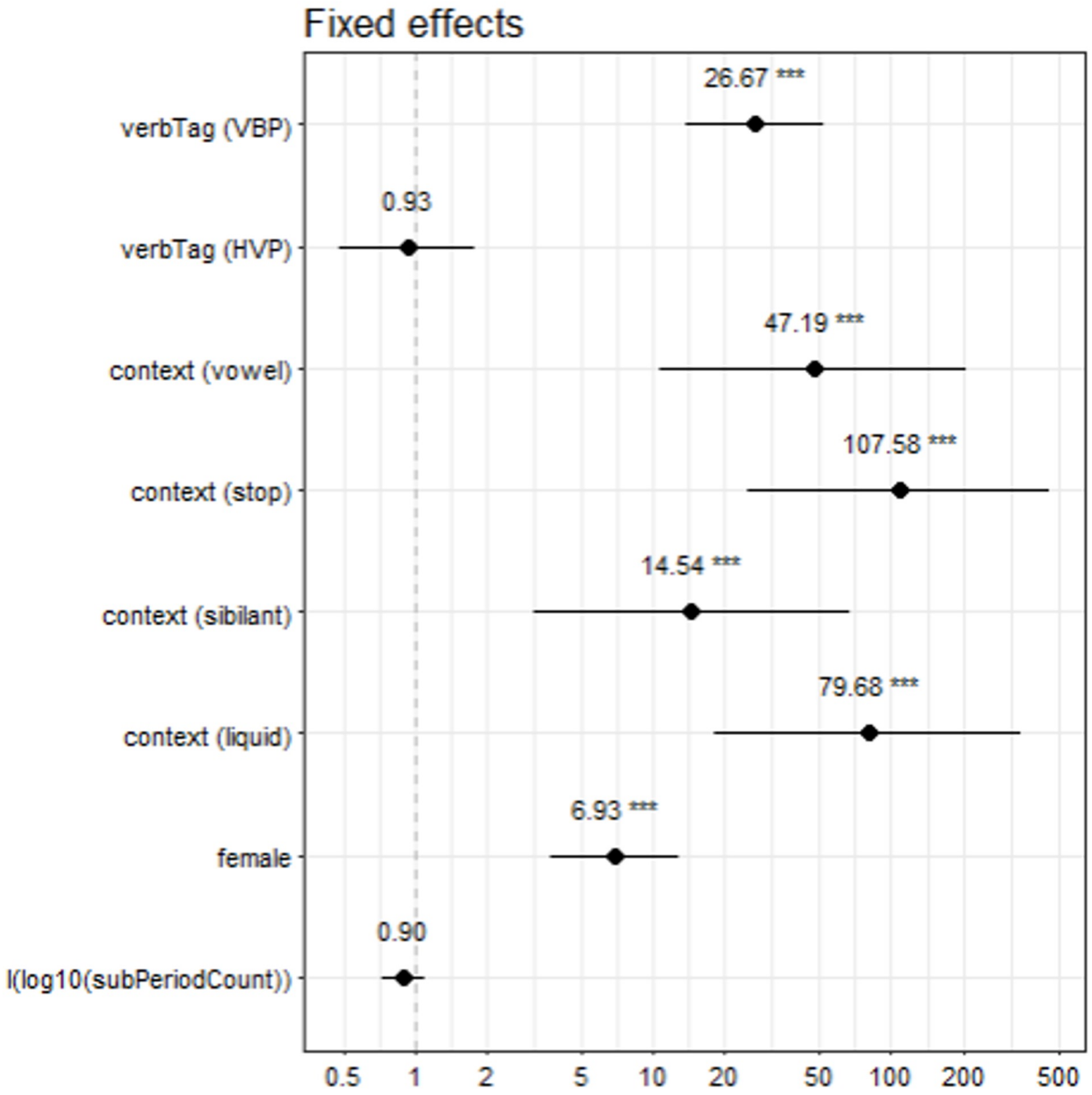
Jenset & McGillivray (2017: 29 ff.)

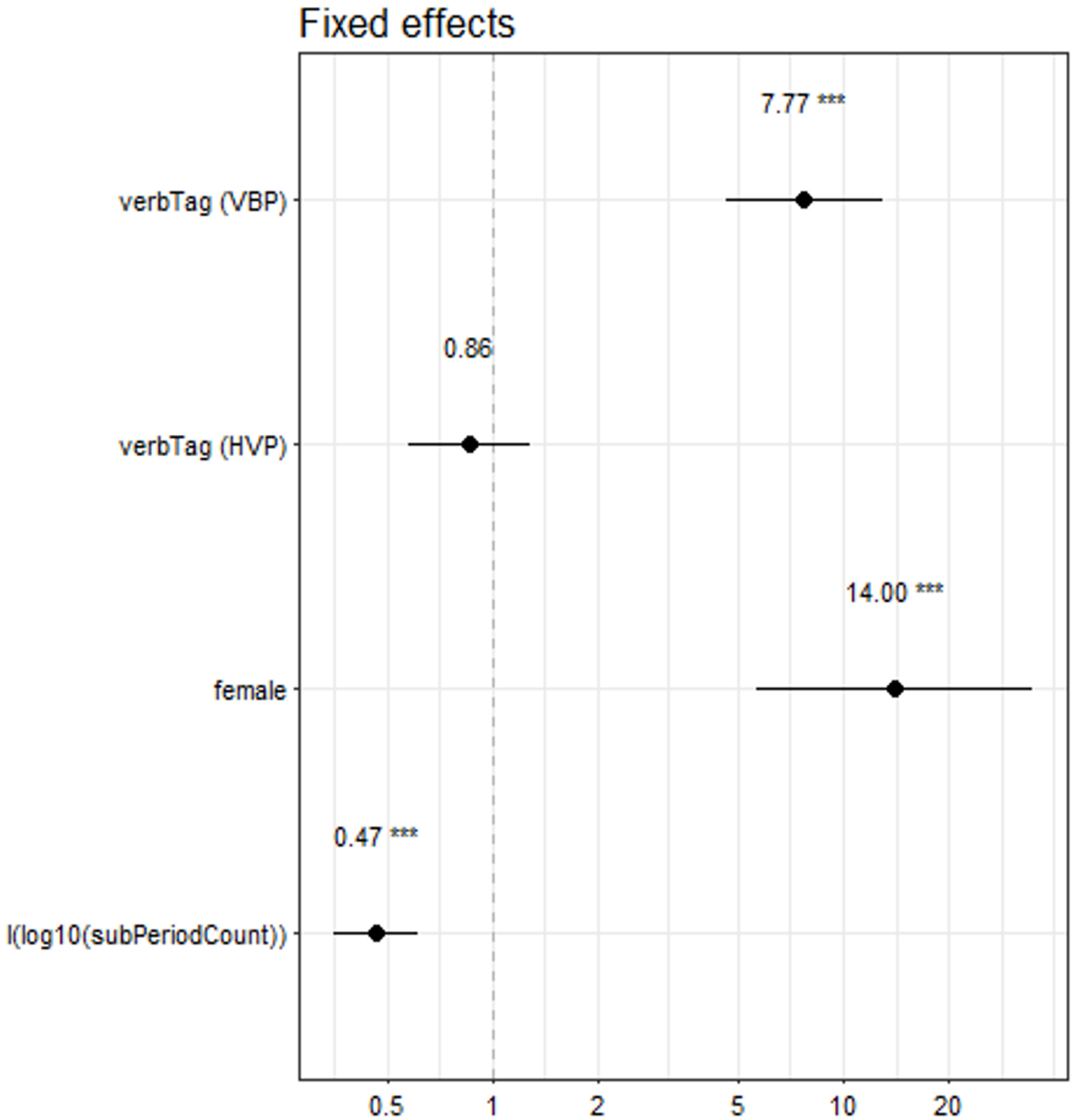
*Ford: **Has** Page any brains? **hath** he any eyes? **hath** he any thinking? Sure, they sleep; he **hath** no use of them.*

Shakespeare, *The Merry Wives of Windsor (3.2.13*



Jenset & McGillivray (2017). *Quantitative Historical Linguistics. A corpus framework* Oxford: Oxford University Press





Jenset & McGillivray (2017). *Quantitative Historical Linguistics. A corpus framework* Oxford: Oxford University Press

Research workflow

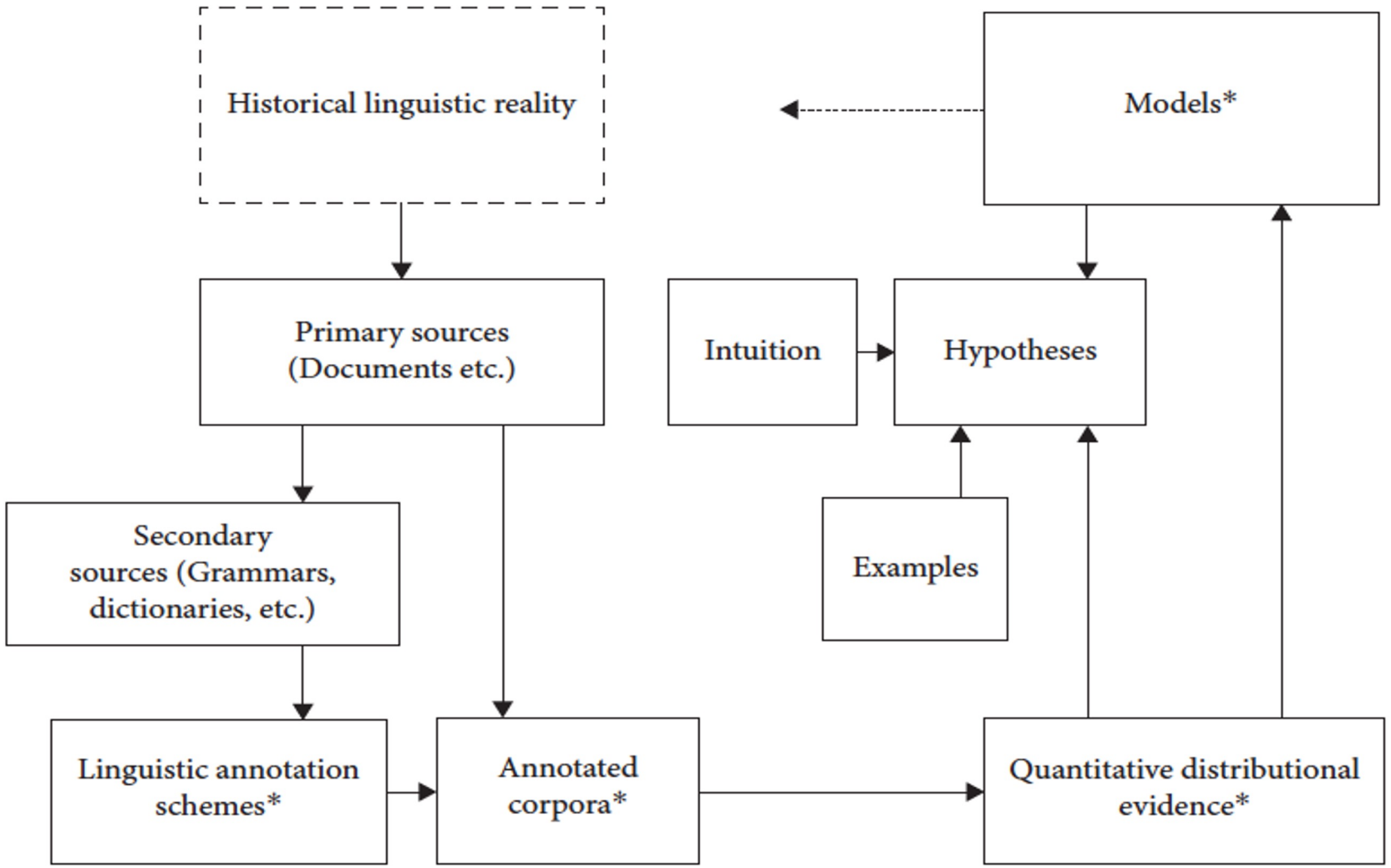


FIGURE 2.1 Main elements of our framework for quantitative historical linguistics. Boxes are entities, arrows are actions or processes; asterisks mark terms for which we use our definitions (see section 2.1.3). The dashed line from models to the (lost) historical linguistic reality implies an approximation.

Jenset & McGillivray (2017: 45)

Quantitative research and historical disciplines

- The use of quantitative methods in historical disciplines is becoming increasingly more viable
- Scholarly communities react differently to quantitative approaches
- Historical disciplines:
  - need to work with closed corpora which can only be expanded working on past records
  - focus on phenomena that change over time
  - frequent need to combine quantitative and qualitative methods
- We propose a general methodological reflection that can help in the process of conducting research in historical disciplines, by taking full advantage of quantification
- Relationship between evidence, modelling, and research practice

McGillivray, B., Wilson, J., & Blanke, T. (2019). Towards a quantitative research framework for historical disciplines. In M. Piotrowski (Ed.), *COMHUM 2018 Workshop on Computational Methods in the Humanities 2018: Proceedings of the Workshop on Computational Methods in the Humanities 2018 Lausanne, Switzerland, June 4–5, 2018* (Vol. 2314, pp. 53-58). (CEUR Workshop Proceedings). http://ceur-ws.org/Vol-2314/paper5.pdf

Modifying the framework

1. *Primary sources*: a single one, publicly registered contracts of apprenticeship.
2. *Annotated evidence*: filter formulary and construct a database.
3. *Quantitative ~~distributional~~ evidence*: numerical or categorical data.
4. *Hypotheses*: coming from examples, comparisons and previous literature.
5. *Models*: ordinary least square regression. Many alternatives, e.g. Bayesian modelling (but how to chose priors?).
6. *Relation model - historical ~~linguistic~~ reality*: evidence in favour of the double-track.

What can we say of our claim in the end?

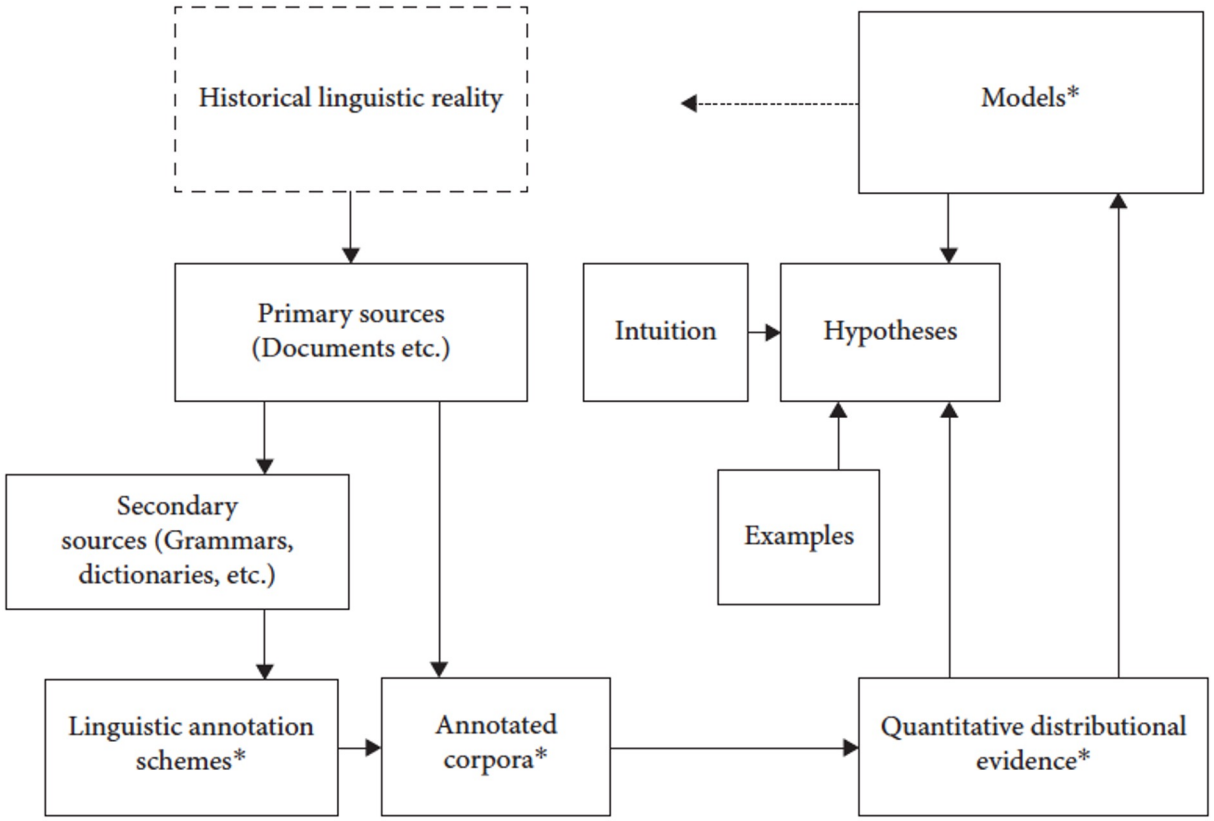


FIGURE 2.1 Main elements of our framework for quantitative historical linguistics. Boxes are entities, arrows are actions or processes; asterisks mark terms for which we use our definitions (see section 2.1.3). The dashed line from models to the (lost) historical linguistic reality implies an approximation.

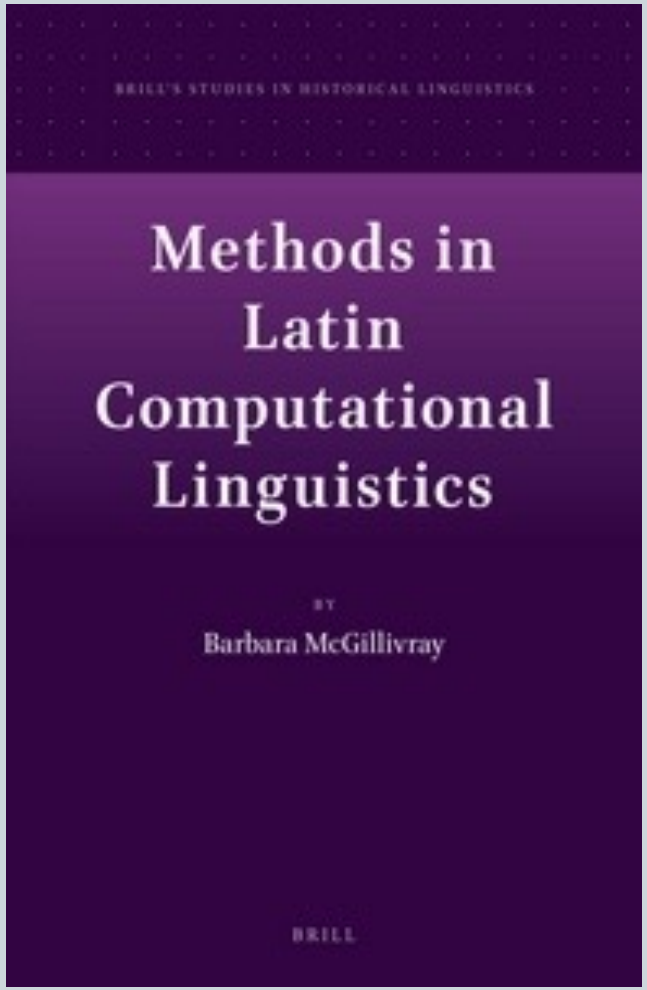
Attempts at some answers

- Modelling and hypothesis testing:
  - Generalization vs comparison
  - Representative datasets in historical linguistics
  - Hypothesis testing vs. Quantitative analysis restricts hypothesis set in history
- Evidence
  - Broader scope in history: categorical, ordinal, and numerical evidence

1. The scope of primary evidence is broader in history than linguistics
2. The scope for a purely quantitative approach is more limited in history than linguistics

Algorithmic thinking and doing

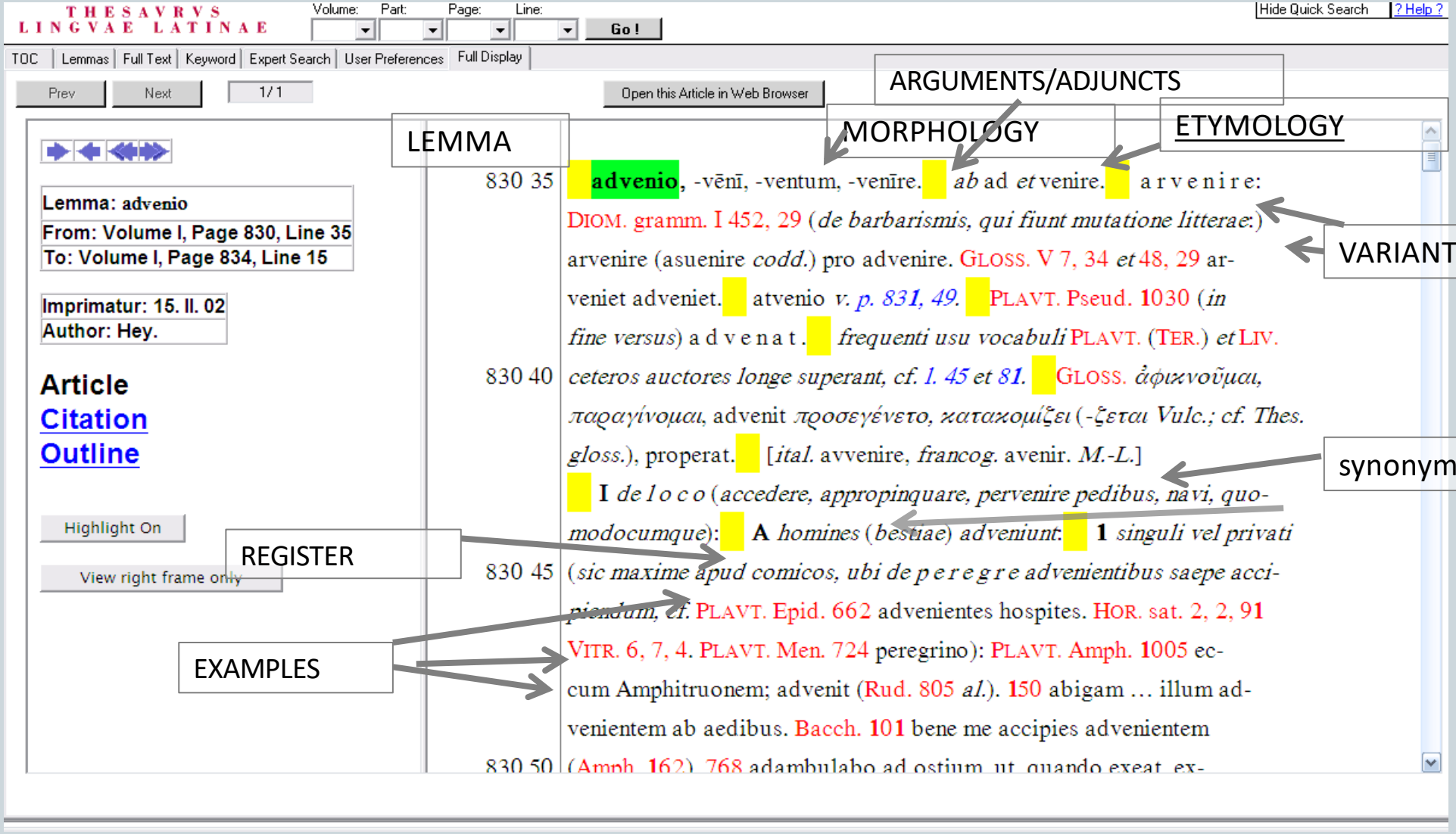
Latin Computational Linguistics



"The availability of huge amounts of data has already changed our access to information and our view on it [...], and quantitative methods are essential to handle these data; moreover, collaboration permeates so many aspects of our work, and digital means facilitate it even further."

"Historical data are no exception, and I believe that research in Latin linguistics can only progress through such an approach. Far from taking over each other's roles, historical and Computational Linguistics can enter into a truly productive symbiosis."

An example from lexicography



Thesaurus Linguae Latinae

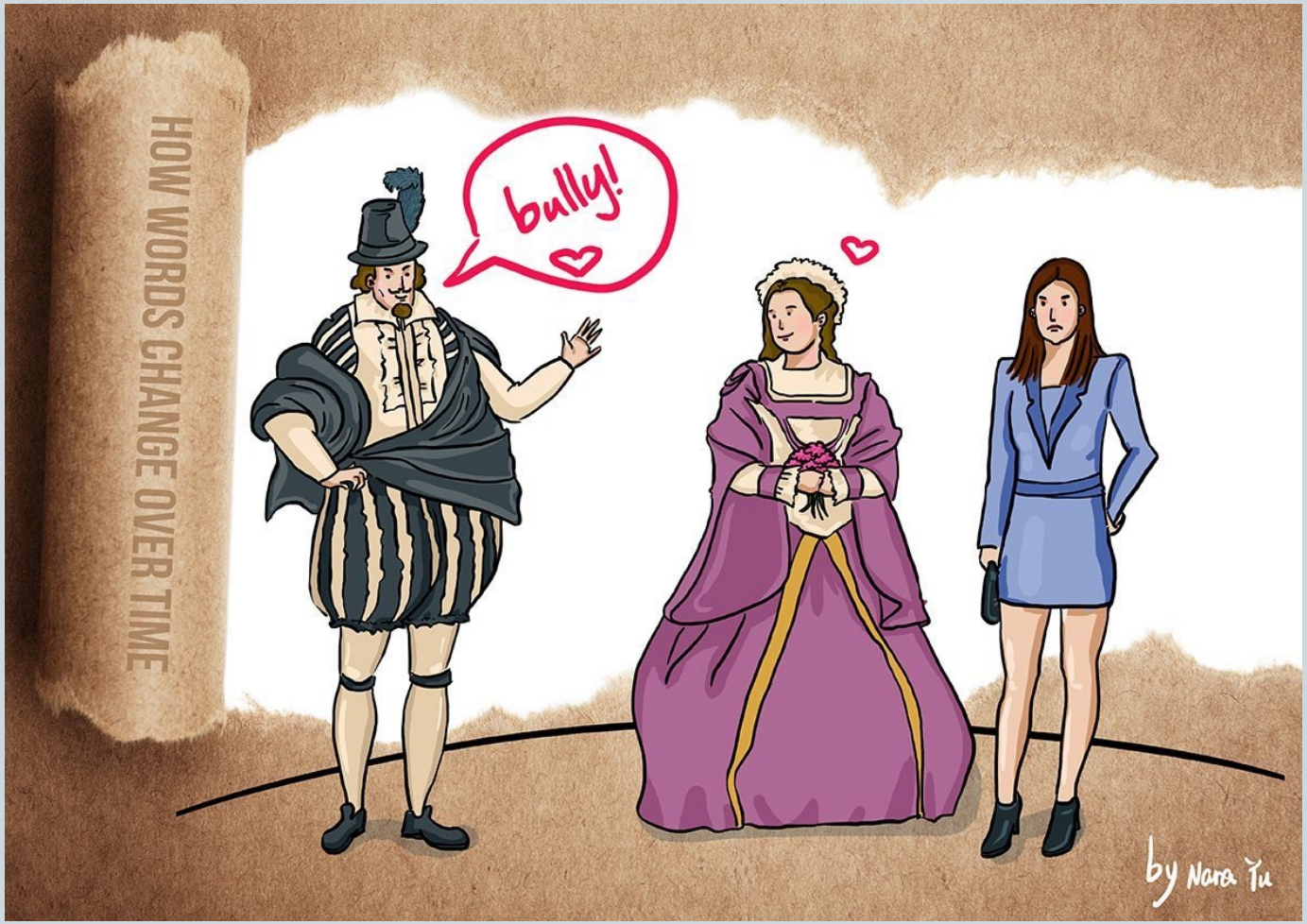
A corpus-driven lexicon

author	title	subdoc	verb	frame_fillers
Caesar	Commentarii de Bello Gallico	Book=2:chapter=2	moveo	active_Obj[accusative]{castrum}
Jerome	Vulgata	book=Apocalypse:chapter=2	moveo	active_(de)Obj[ablative]{locus},Obj[accusative]{candelabrum}
Jerome	Vulgata	book=Apocalypse:chapter=6	moveo	passive_Obj[ablative]{ventus}
Jerome	Vulgata	book=Apocalypse:chapter=6	moveo	passive_(de)Obj[ablative]{locus},Sb[nominative]{insula},Sb[nominative]{mons}
Ovid	Metamorphoses	Book=1:card=163	moveo	active_Obj[accusative]{sidus},Obj[accusative]{mare},Obj[accusative]{terra}
Ovid	Metamorphoses	Book=1:card=348	moveo	passive_Obj[ablative]{augurium},Sb[nominative]{Titania}
Ovid	Metamorphoses	Book=1:card=746	moveo	passive_Obj[ablative]{ira},Obj[ablative]{prex}
Petronius	Satyricon	text=sat:section=30	moveo	active_Obj[accusative]{gressus2}
Petronius	Satyricon	text=sat:section=30	moveo	active_Obj[accusative]{ego},Sb[nominative]{jactura}
Petronius	Satyricon	text=sat:section=64	moveo	passive_Obj[ablative]{jactura}
Propertius	Elegies	book=1:poem=2	moveo	active_Obj[accusative]{sinus}

McGillivray (2014)

Bringing together quantitative and algorithmic thinking and doing

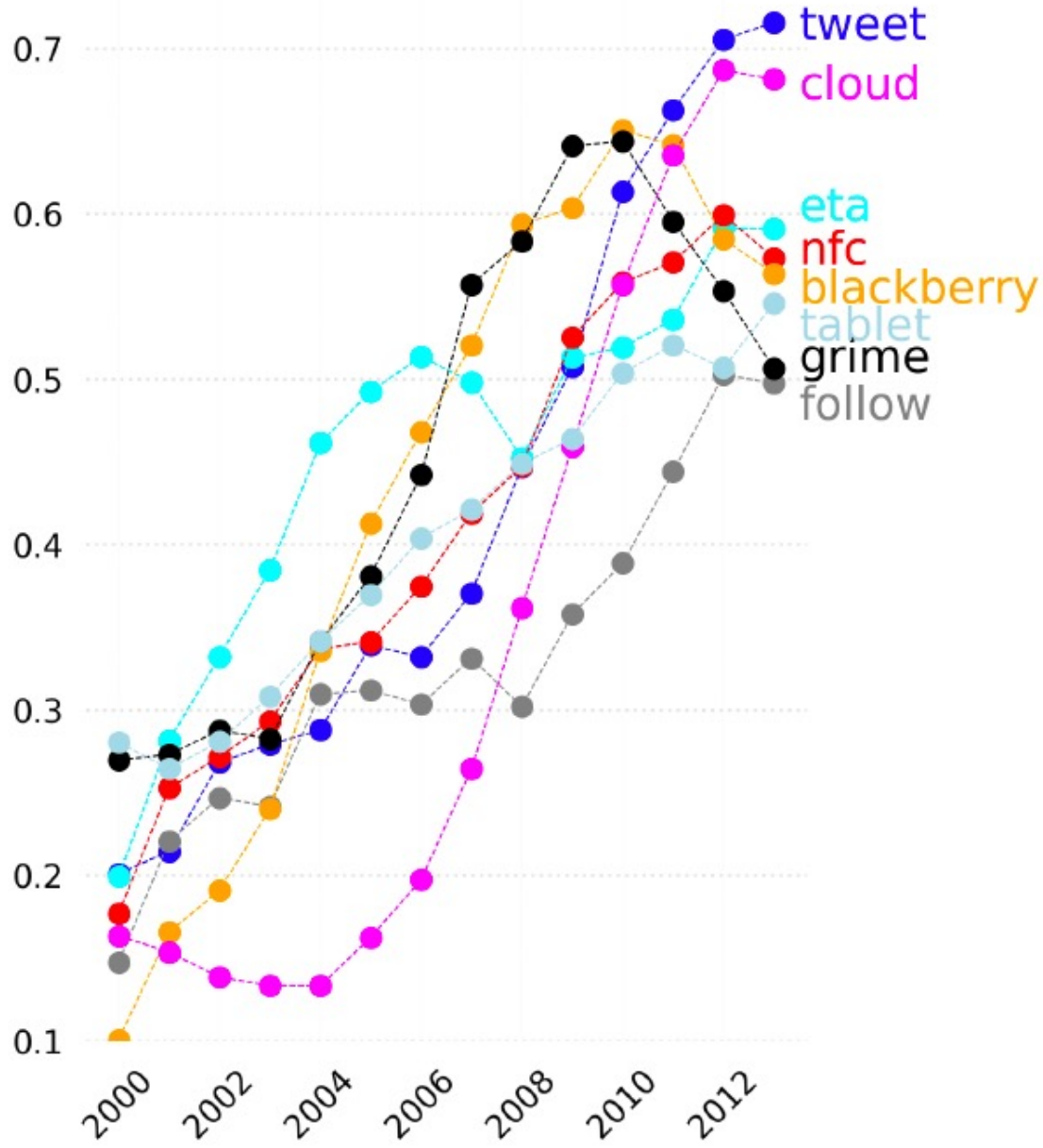
Lexical semantic change



https://blog.csoftintl.com/hq-magazine/

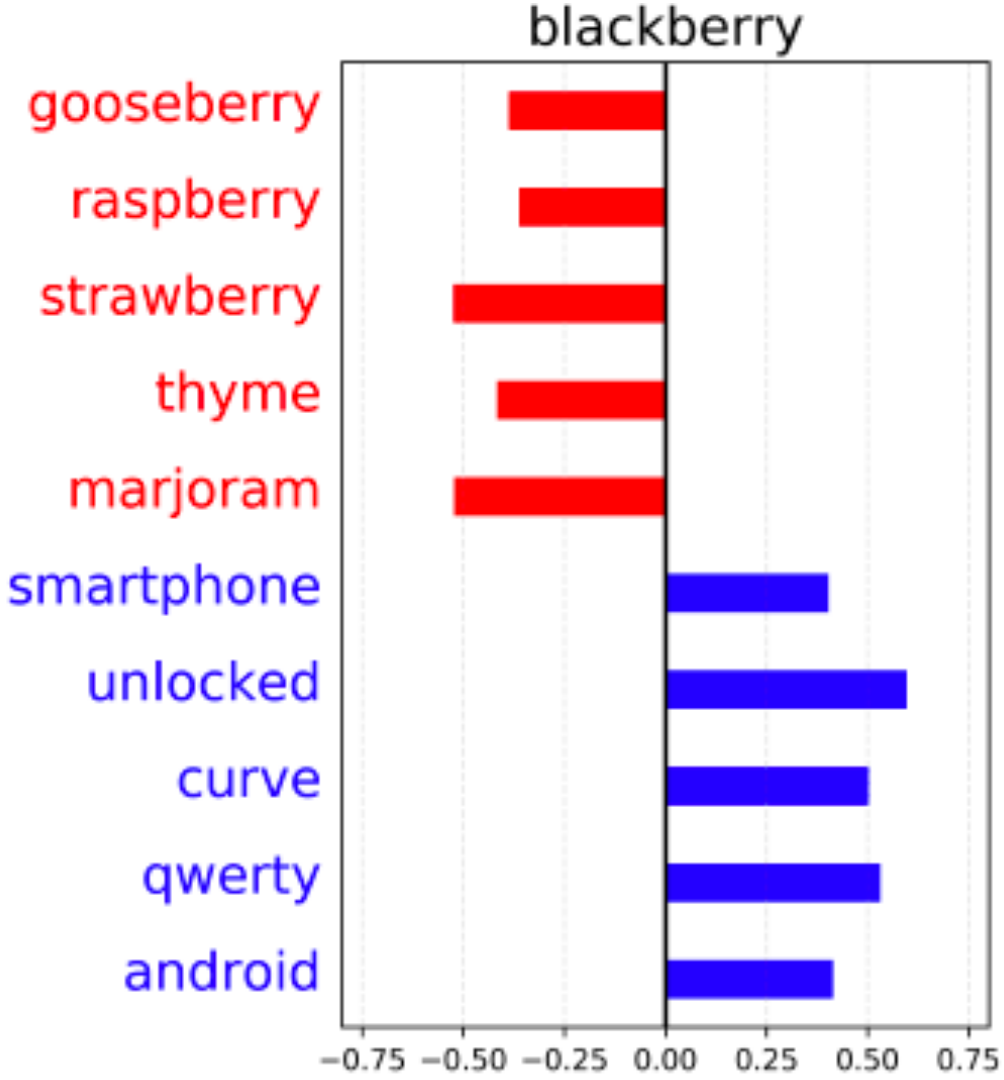
Peter Koch. 2016. Meaning change and semantic shifts. In Päivi Juvonen and Maria KoptjevskajaTamm, editors, The Lexical Typology of Semantic Shifts, pages 21–66. De Gruyter Mouton, Berlin/Boston.

UK Web archive 1996-2013

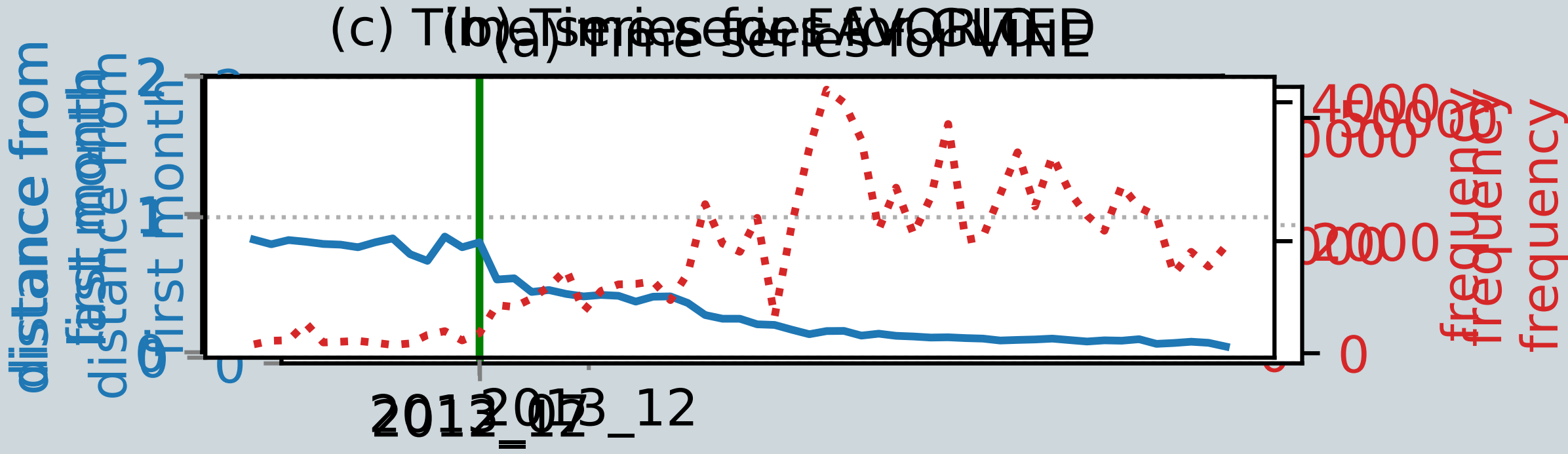


Tsakalidis, A., Bazzi, M., Cucuringu, M., Basile, P. and McGillivray, B. (2019). Mining the UK Web Archive for Semantic Change Detection. Proceedings of *RANLP 2019*

My blackberry is frozen!

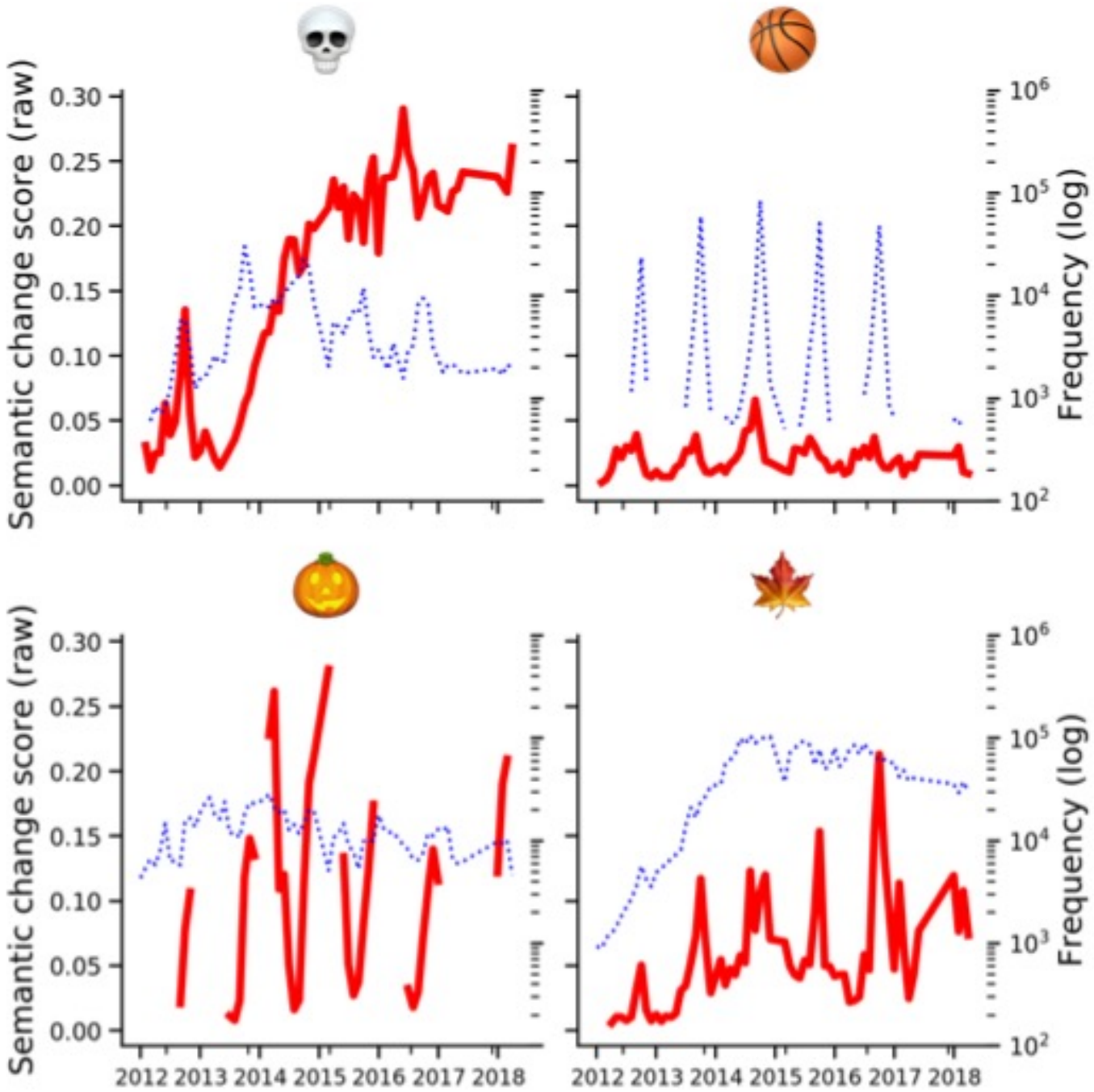


Some findings



Shoemark, P., Ferdousi Liza, F., Nguyen, D., Hale, S. and McGillivray, B. (2019).
Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings.
In *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Hong Kong, China.*

Emoji



Robertson, A., Ferdousi Liza, F, Nguyen, D., McGillivray, B. and Hale, S. (2021). Semantic Journeys: Quantifying change in emoji meaning from 2012-2018. *Emoji 2021 Workshop*

Open thinking and doing

The Open Definition

Availability and access

The Open Definition sets out principles that define "openness" in relation to **data and content**.

It makes **precise** the meaning of "open" in the terms **"open data"** and **"open content"** and thereby ensures **quality** and encourages **compatibility** between different pools of open material.

It can be summed up in the statement that:

Re-use and redistribution

> "Open means **anyone** can **freely access, use, modify, and share** for **any purpose** (subject, at most, to requirements that preserve provenance and openness)."

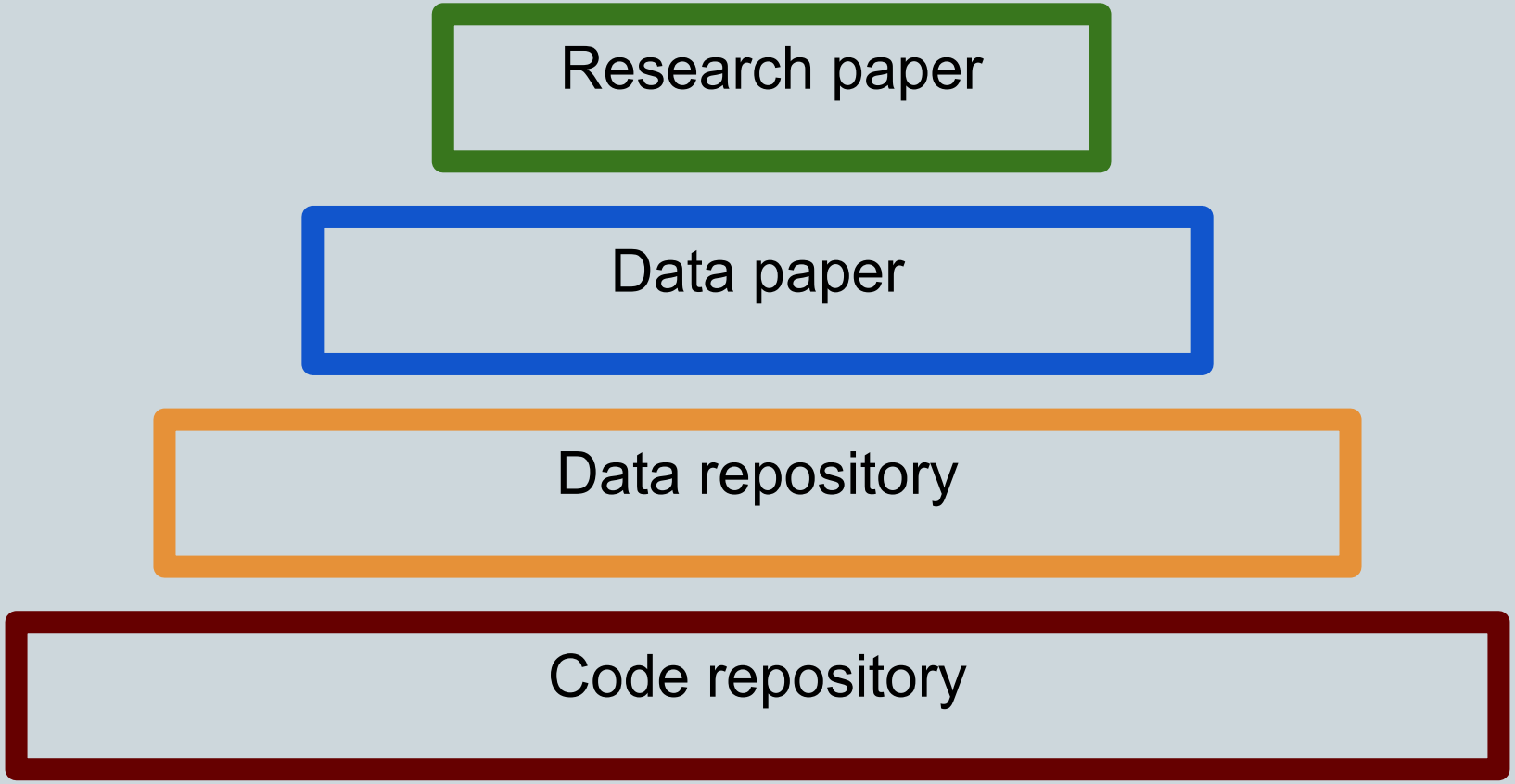
Put most succinctly:

Universal participation

> "Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**"

https://opendefinition.org/

The open access pyramid



Goal: research is transparent, reproducible and impactful

Credits: Paola Marongiu
Marton Ribary, *Surrey Open Research and Transparency Showcase 2021*

The *Journal of Open Humanities Data*

- Launched in 2015
- It is part of the '*metajournals*' family in Ubiquity Press. They publish papers about:
  - research data
  - software
  - hardware

Credits: Paola Marongiu

Open Humanities Data

"Open science commonly refers to efforts to make the output of publicly funded research more widely accessible in digital format to the scientific community, the business sector, or society more generally."

OECD (2015), "Making Open Science a Reality", *OECD Science, Technology and Industry Policy Papers*, No. 25, OECD Publishing, Paris, https://doi.org/10.1787/5jrs2f963zs1-en.

Open to society

The future

Homo in machina

“The first wave of digital humanities work was quantitative, mobilizing the search and retrieval powers of the database, automating corpus linguistics, stacking hypercards into critical arrays. The second wave is qualitative, interpretive, experiential, emotive, generative in character."

Digital Humanities manifesto 2.0, published in 2011 in Humanities Blast, blog maintained by Todd Presner (UCLA, Chair Digital Humanities Program)

Future directions

Greening research

Even more infrastructure

Work with GLAM

Textual and visual

New careers

Continuous development

zenodo

Search

Upload

Communities

April 29, 2022

Other | Open Access

A Researcher Guide to Writing a Climate Justice Oriented Data Management Plan

DHCC Information, Measurement and Practice Action Group

Editor(s)

Baker, James; Ohge, Christopher; Otty, Lisa; Walton, Jo Lindsay

Project member(s)

Alexander, Anne; Ames, Sarah; Baker, James; Cummings, James; Ho, Racelar; Isaksen, Leif; McGillivray, Barbara; Vignoles, Anna; Walton, Jo Lindsay; Winters, Jane

This ***Researcher Guide to Writing a Climate Justice Oriented Data Management Plan*** aims to enable researchers to be bold in interpreting the data management guidance.

The guide was written in Winter/Spring 2022 by the Information, Measurement and Practice Action Group of the Digital Humanities Climate Coalition.

If you have suggestions for changes or improvements, please let us know.

Preview

Future directions

Greening research

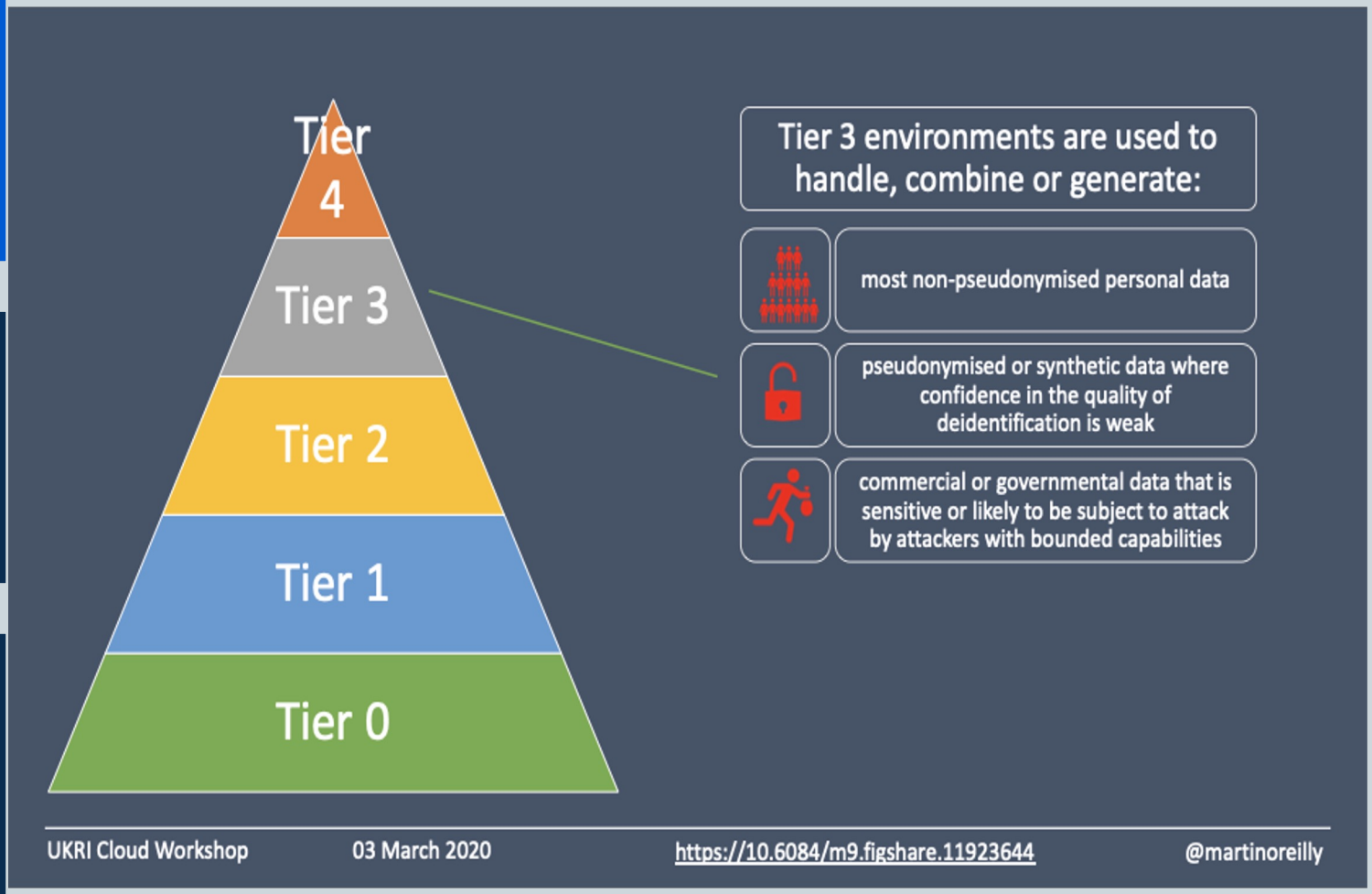
Even more infrastructure

Work with GLAM

Textual and visual

New careers

Continuous development



Future directions

Greening research

Even more infrastructure

Work with GLAM

Textual and visual

New careers

Continuous development

The BL team provided invaluable expertise on the characteristics of the collection

Often interesting projects start from the availability of a specific collection

In our case, the UK web archive dataset had been available for some time, and was looking for some data science researchers to get their hands on it

The Turing team offered the computational skills to wrangle the collection

Regular updates kept us on track

Partial overlap between our backgrounds helped our communication

Future directions

Greening research

Even more infrastructure

Work with GLAM

Textual and visual

New careers

Continuous development

2022.aclweb.org/papers

visual 2/22

Contextual Representation Learning beyond Masked Language Modeling	Zhiyi Fu, Wangchunshu Zhou,
ConTinTin: Continual Learning from Task Instructions	Wenpeng Yin, Jia Li, Caiming Xiong
Continual Few-shot Relation Learning via Embedding Space Regularization and Data Augmentation	Chengwei Qin, Shafiq Joty
Continual Pre-training of Language Models for Math Problem Understanding with Syntax-Aware Memory Network	Zheng Gong, Kun Zhou, Xin Zhao, Jing Sha, Shijin Wang, Ji-Rong Wen
Continual Prompt Tuning for Dialog State Tracking	Qi Zhu, Bing Li, Fei Mi, Xiaoyan Zhu, Minlie Huang
Continual Sequence Generation with Adaptive Compositional Modules	Yanzhe Zhang, Xuezhi Wang, Diyi Yang
Contrastive Visual Semantic Pretraining Magnifies the Semantics of Natural Language Representations	Robert Wolfe, Aylin Caliskan
Controllable Dictionary Example Generation: Generating Example Sentences for Specific Targeted Audiences	Xingwei He, Siu Ming Yiu
CQG: A Simple and Effective Controlled Generation Framework for Multi-hop Question Generation	Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, Xuanjing Huang
Cree Corpus: A Collection of nêhiyawêwin Resources	Daniela Teodorescu, Josie Matalski, Delaney Alexa Lothian, Denilson Barbosa, Carrie Demmans Epp
Cross-Lingual Ability of Multilingual Masked Language Models: A Study of Language Structure	Yuan Chai, Yaobo Liang, Nan Duan
Cross-Lingual Contrastive Learning for Fine-Grained Entity Typing for Low-Resource Languages	Xu Han, Yuqi Luo, Weize Chen, Zhiyuan Liu, Maosong Sun, Zhou Botong, Hao Fei, Suncong Zheng

Future directions

- Greening research
- Even more infrastructure
- Work with GLAM
- Textual and visual
- New careers
- Continuous development

Future directions

Greening research

Eve
infra

Work with GLAM

Text
v

New careers

Continuous development

BRIDGES

Home | About Us | Programme | Sponsorship | Venue | Registration

Friday | September 16th	
9:00 – 10:00 Madrid time zone	Keynote Speaker: Oliver Dangles
10:00 – 11:00 Madrid time zone	Room A: Symposium – Power asymmetries in academic-scientific environments: critical insights into the neoliberal production of knowledge Room B: Online poster session
11:00 – 11:30 MADRID TIME ZONE	COFFEE BREAK
11:30 – 13:30 Madrid time zone	Room A: Symposium – Power asymmetries in academic-scientific environments: critical insights into the neoliberal production of knowledge Room B: Symposium – Towards inclusive pedagogical foundations of information communication technology curriculum in digital humanities

Thank you!

Barbara McGillivray

Barbara.mcgillivray@kcl.ac.uk