

Sciveytor: From Corpora to Philosophy of Science

TU Eindhoven, 2020.10.6

Charles H. Pence

@pencechp • @pencelab

Outline

1. Tool development and future work
2. Current corpus
3. From corpora to philosophy of science in practice

The take-home: Empirical, corpus-based methods raise a number of meta-philosophical questions prior to our production of philosophical results.

Thanks to...



Oliver Lean



Luca Rivelli

Tool Development

Why a New Tool?

Journal articles are weird.

- They're small.
- They carry unusual metadata.
- People are used to searching/indexing them in particular ways.
- They're hard to get and come in an infinite variety of formats.

Tricks

- Have to analyze using only basic, plain text. (No guarantee we can always get anything better.)
- Ways to search/filter by journals, years (categories coming soon)
- Have to be robust to low-quality data (often entirely dependent on publishers for data quality; too many documents for manual cleaning)
- Also coming soon: citation analysis support

Current Infrastructure



The evoText Project

<https://www.evotext.org/>

[Home](#) / [Search the Database](#)

We are getting a few job failures on our new infrastructure that we're in the process of diagnosing. If you've tried jobs that have failed, please do retry them as this gives us valuable diagnostic information!

Corporate DNA

Ken Baskin, Dr Kurt Richardson | *Emergence: Complexity and Organization*, Vol. 2, No. 1 (volume-2-issue-1),

[More](#)

Editor's Note (2.1)

Michael Lissack, Dr Kurt Richardson, Dr Michael Lissack | *Emergence: Complexity and Organization*, Vol. 2, No. 1 (volume-2-issue-1),

[More](#)

Emergence

Jeffrey Goldstein, Dr Kurt Richardson | *Emergence: Complexity and Organization*, Vol. 2, No. 1 (volume-2-issue-1),

[More](#)

What can we learn from a theory of complexity?

Paul Cilliers | *Emergence: Complexity and Organization*, Vol. 2, No. 1 (volume-2-issue-1),

[More](#)

Editorial (16.4)

Peter Allen | *Emergence: Complexity and Organization*, Vol. 16, No. 4 (volume-16-issue-4),

[More](#)

Why Businesses Fail

Tom Rand, Dr Kurt Richardson | *Emergence: Complexity and Organization*, Vol. 1, No. 4 (vol1iss4),

[More](#)

Twenty-first-century Management and the Complexity Paradigm

[More](#)[Sort](#)[Save](#)

1625790 articles in database

Filter search

Authors

Jr.	16140
III	951
Bentley Glass	811
Rudolf Schmid	778
C. P. Swanson	428

Journal

Nature	363791
Proceedings of the National Academy of Sciences of the United States of America	123973
Current Science	56187
The Quarterly Review of Biology	32837
Plant	31623

Coming soon: Sciveyor!

New web interface, cleaner corpus data, new features, more speed!



Home



Search



Browse



Log In

Search...

Save as dataset Sort

Search Results

Whitehorn, James et al. 2012. "Prophylactic Platelets in Dengue: Survey Responses Highlight Lack of an Evidence Base." *PLoS Neglected Tropical Diseases* 6(6):e1716. [More...](#)

Kvitko, Brian H. et al. 2012. "Burkholderia pseudomallei Known Siderophores and Hemin Uptake Are Dispensable for Lethal Murine Melioidosis." *PLoS Neglected Tropical Diseases* 6(6):e1715. [More...](#)

Gower, Emily W. et al. 2012. "Definitions and Standardization of a New Grading Scheme for Eyelid Contour Abnormalities after Trichiasis Surgery." *PLoS Neglected Tropical Diseases* 6(6):e1713. [More...](#)

Gopal, Hemavathi et al. 2012. "Oligonucleotide Based Magnetic Bead Capture of Onchocerca volvulus DNA for PCR Pool Screening of Vector Black Flies." *PLoS Neglected Tropical Diseases* 6(6):e1712. [More...](#)

Düppre, Nádia C. et al. 2012. "Impact of PGL-1 Seropositivity on the Protective Effect of BCG Vaccination among Leprosy Contacts: A Cohort Study." *PLoS Neglected Tropical Diseases* 6(6):e1711. [More...](#)

de Souza Leoratti, Fabiana Maria et al. 2012. "Neutrophil Paralysis in Plasmodium vivax Malaria." *PLoS Neglected Tropical Diseases* 6(6):e1710. [More...](#)

Savage, Amy F. et al. 2012. "Transcript Expression Analysis of Putative Trypanosoma brucei GPI-Anchored Surface Proteins during Development in the Tsetse and Mammalian Hosts." *PLoS Neglected Tropical Diseases* 6(6):e1708. [More...](#)

Carrasco, Hernán J. et al. 2012. "Geographical Distribution of Trypanosoma cruzi Genotypes in Venezuela." *PLoS Neglected Tropical Diseases* 6(6):e1707. [More...](#)

Escadafal, Camille et al. 2012. "First International External Quality Assessment of Molecular Detection of Crimean-Congo Hemorrhagic Fever Virus." *PLoS Neglected Tropical Diseases* 6(6):e1706. [More...](#)

Zihlmann, Karina Franco et al. 2012. "Living Invisible: HTLV-1-infected Persons and the Lack of Care in Public Health." *PLoS Neglected Tropical Diseases* 6(6):e1705. [More...](#)

Katará, Gajendra Kumar et al. 2012. "Evidence for Involvement of Th17 Type Responses in Post Kala Azar Dermal Leishmaniasis (PKDL)." *PLoS Neglected Tropical Diseases* 6(6):e1703. [More...](#)

Filter Results

Author

Peter J. Hotez **45**
 Jörg Utzinger **27**
 Marleen Boelaert **25**
 Donald P. McManus **18**
 Philippe Büscher **15**
[Show more...](#)

Journal

PLoS Neglected Tropical Diseases **1550**
 Actually a Novel **3**
[Show more...](#)

Publication Date

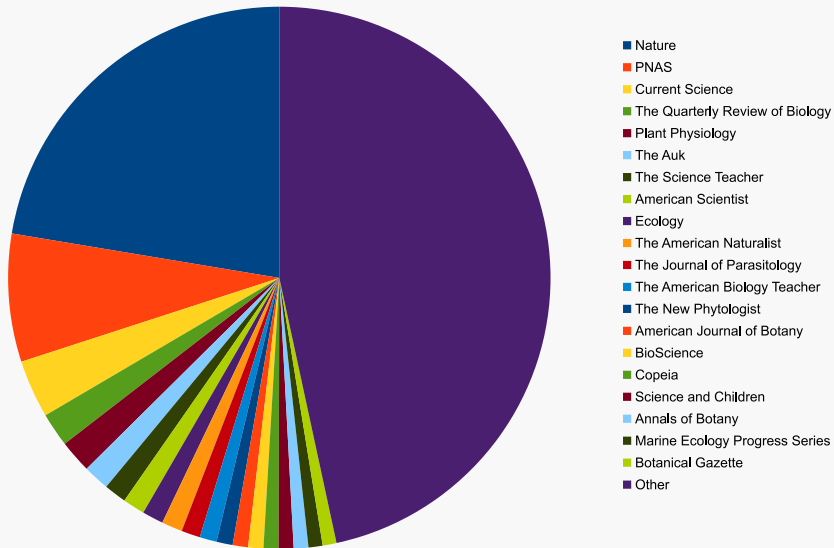
2011 **443**
 2010 **350**
 2012 **262**
 2009 **224**
 2008 **179**
[Show more...](#)

Current Corpus

The Corpus

1,625,790 documents
334 journals

Journals



Analyses

Tools available:

- General word frequency analysis (1-gram or N-gram, with tf-idf scores, configurable block creation behavior, etc.)
- Collocation analysis (like cooccurrence but for direct bigrams)
- Some simple graphing-by-date
- Export to various reference manager citation formats

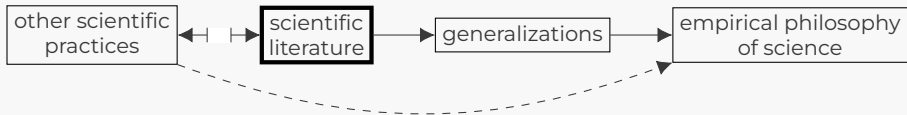
From Corpora to Philosophy of Science

Literature and Practice

A common criticism: recent philosophy of science has **done well** to move beyond the literature, which gives us (at best) a **slanted and partial** picture of scientific practice

Literature and Practice

How do we justify the empirical analysis of the scientific literature from a “post-practice-turn” perspective?



Literature and Practice

- *the logical view* – articles contain agreed-upon forms of justification for claims (e.g., Reichenbach)
- *the sociological view* – articles form part of power struggles for converts (e.g., Latour)
- *the fraud view* – articles are a reconstructed, false view of scientific argumentation (e.g., Medawar)
- *the narrative view* – articles construct a narrative into which scientific practice should be placed (e.g., Rouse)

Corpora and Generalizations

We need to:

1. Validate methods of analysis, ensuring that they're giving us empirically adequate pictures of the corpus at issue – in other words, engage with *digital humanities*
2. Explain the move from *syntactic* claims (like term frequencies or cooccurrences) to *semantic* claims (about the meanings of texts) – in other words, engage with *corpus linguistics*

Meta-Questions

The goal: We need to build coherent sets of answers to these questions which, taken together, can license our use of corpora to understand scientific practice.

Meta-Questions

As an example: combine a *social view* of the role of the literature with a *socio-linguistic* view of meaning construction.

Scientific articles do allow us access to the meanings of terms, but these are socially negotiated entities, always in flux, subject to redefinition. We need *diachronic* methods for reconstructing those meanings over the course of debates, and we will get historically- and community-relative, local knowledge.

Questions?

charles@charlespence.net

<https://pencelab.be>

@pencechp · @pencelab



The Pence Lab



Wallonie - Bruxelles
International.be

fnrs

LA LIBERTÉ DE CHERCHER