DFG Deutsche Forschungsgemeinschaft

ROTO The Return of the Organism in the Biosciences

What Was Theoretical Biology?

A Topic-Modelling Analysis of a Multilingual Corpus of Monographs and Journals, 1901-1971

Alexander Böhm[1], Stefan Reiners-Selbach[2], Jan Baedke[1], Alejandro Fábregas-Tejeda[1], Vera Straetmanns[1], Daniel J. Nicholson[3]

Abstract: The early 'philosophical' period of theoretical biology, before the field became synonymous with mathematical biology, has been almost totally forgotten—let alone carefully examined. Much of this discourse took place in a handful of book series, monographs, and journals, the majority of which were initially published in German. Our aim is to rescue this multilingual corpus from the dustbin of history. Our guiding question is: **What did theoretical biology look like in the early 20th century?** We utilize LDA topic modelling (after machine translating where necessary), top2vec, and document embeddings to create an interactive tool for the exploration of this corpus, which allows us to analyze the thematic development of theoretical biology during the 20th century, paying particular attention to the field's declining interest in philosophical disputes and its increasing emphasis on formal modelling.

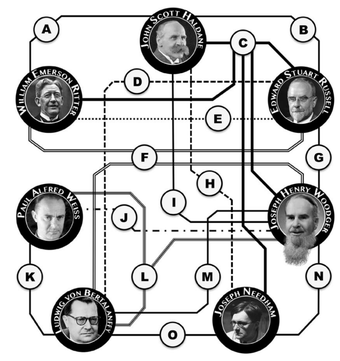


Fig. 1: Some central figures in early 20th century theoretical biology
(for details, see Nicholson & Gawne 2015)

I Aims

At this early exploratory stage of the project, operationalization via **topic modelling**:

(1) Central debates and topics: Which key topics can be identified and how does their 'share' (i.e. probability distribution) in the documents develop over time? Which topic clusters can be identified?

(2) Central authors and structure of scientific community: Are certain topics dominated by particular authors, languages (of origin), and nationalities? What, where, and when did transitions occur in the networks of authors and topics?

(3) Development of formal modelling: How steadily does the proportion of publications that uses mathematical formulas increase over time (and in which thematic contexts)?

II Corpus

Fig. 2: Seminal works in early theoretical biology

Reinke 1901 | Schaxel 1919 | Mayer 1934 | *Acta Biotheoretica*

Fig. 3: Constitution of the corpus (in number of words, by publication series, per decade)

III Methods

After scanning and OCR (Smith 2019) of the selected sources, we **(a)** machine translated part of the corpus using deepL (Kutylowski 2017) to achieve linguistic homogeneity (German). We ran a **(b)** layout analysis using layoutparser (Shen 2021) to calculate the ratio of text areas to mathematical-formula areas, which we used as a mathematization score. We pre-processed and cleaned our text data using re and spaCy (Honnibal 2020). Then we used MALLET (McCallum 2002) for LDA-Topic Modelling, which we visualized diachronically. Additionally, we used top2vec (Angelov 2020) for a contrasting, embedding-based Topic Modelling approach. These results were then used as clustering for the **(h)** document embedding (UMAP, McInnes 2018), which we visualized in an **(k)** interactive scatter plot using bokeh (Bokeh Development Team 2018).

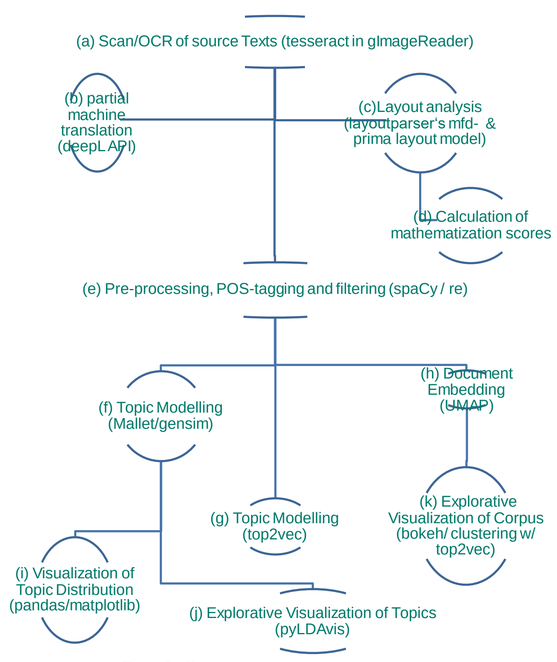


Fig. 4: Current workflow of analysis.

IV Results

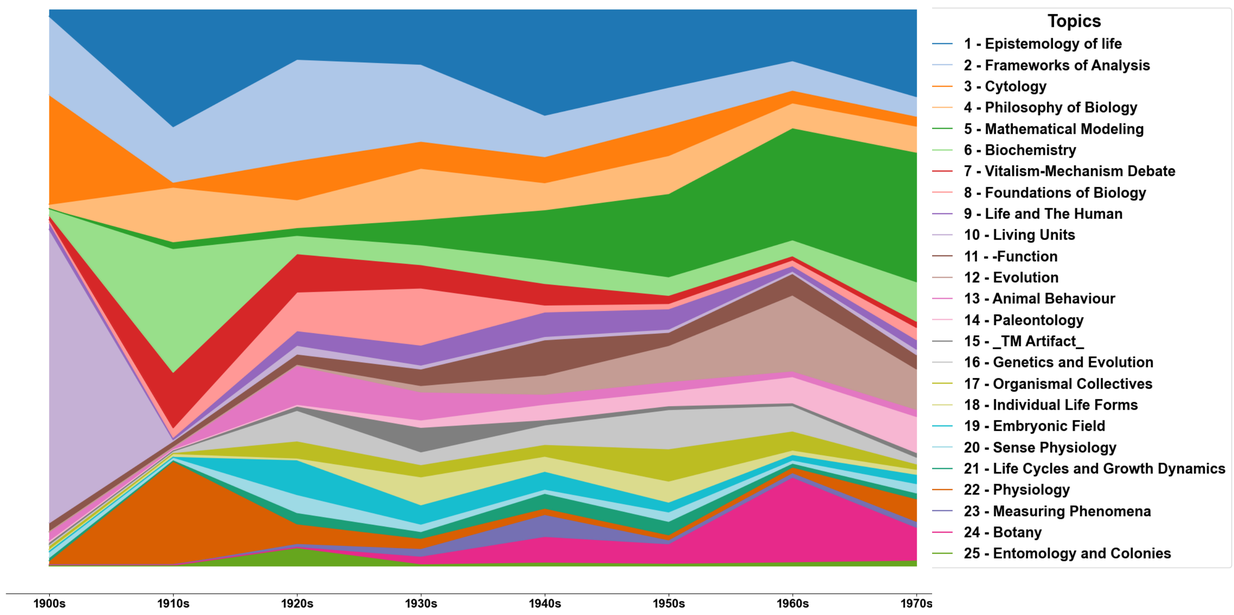


Fig. 5: Development of Topic Model proportion per decade

Results (continued)

- **Topic modelling: Central debates in theoretical biology (Fig. 5)**
  - (i) Our preliminary survey highlights the importance of the **'organism' concept** in early (German-speaking and international) twentieth-century theoretical biology.
  - (ii) In contrast to the field as it exists today, which largely consists of mathematical modelling, theoretical biology in its early decades focused in substantial part on **philosophical themes and controversies** (e.g., conceptual and theoretical foundations for biology and the vitalism-mechanism dispute; topics 1, 2, 4, 7, 8).
  - (iii) The most important biological debates **interlinked evolutionary with developmental (and temporal) issues** (see clusters 3, 12, 21), something that was typical of German-speaking biology at the time.
  - (iv) **Attempts to mathematize and formalize biology** (topic 5) were already present in the earliest work in theoretical biology, but these were largely decoupled from other central philosophical concerns (e.g., about the concept of the organism).
- **Historical development of theoretical biology (Fig. 5):**
  - (i) ***Decreasing interest:*** While bio-philosophical topics (4) remain a part of the field until 1970, there is a marked decline in interest from the 1940s onwards in the conceptual foundations of biology and the vitalism-mechanism dispute (topics 2, 7, 8)
  - (ii) ***Increasing interest:*** **Mathematical modelling and formal approaches** (topic 5) increase from the 1940s onwards (see the 'mathematization index' below); other increasing trends concerned **evolution** (12) from the 1940s onwards, likely a reaction to the rise of population genetics and the forging of the Modern Synthesis (the absence of evolution in earlier decades is surprising nevertheless).
  - (iii) ***Constant interest:*** There is a steady interest in many topics of lesser prominence, such as **'Genetics'** (16), even though gene-related discussions increased over time as a result of the rise of molecular biology.
- **Modelling degrees of mathematization (Fig. 6)**

To address the progressive formalization of theoretical biology over time, we have introduced a 'mathematization score' that will allow us, in future works, to assess when exactly and in the context of which topics and debates did mathematical models and tools came to be widely adopted. Our initial results, however, proved to be inconclusive. The pre-trained models for LayoutParser produced mixed results (see, for instance, topics 2 and 5).

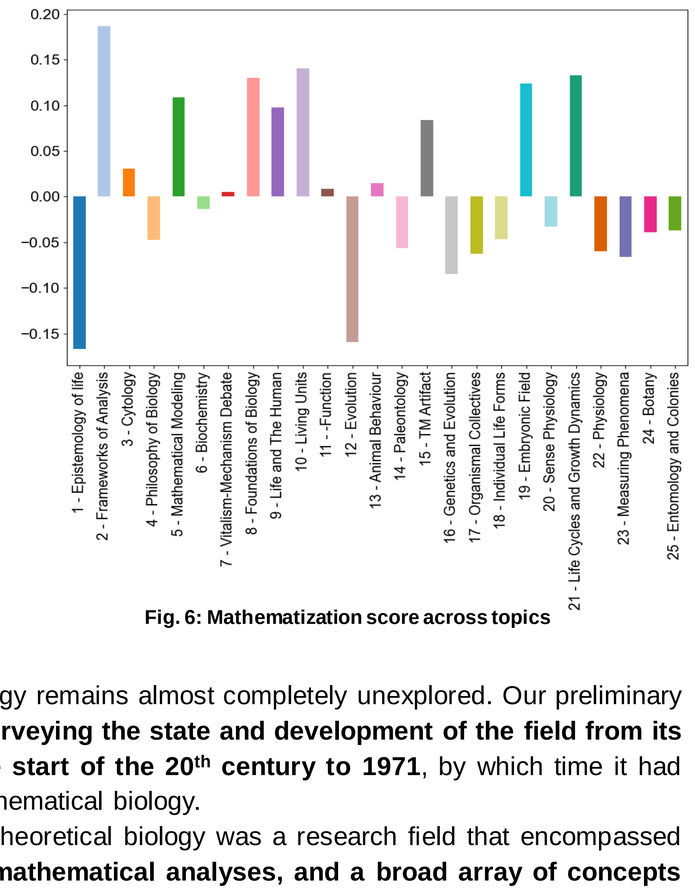


Fig. 6: Mathematization score across topics

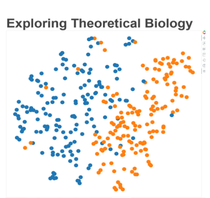
V Conclusions

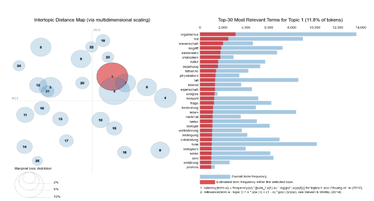
I. The history of theoretical biology remains almost completely unexplored. Our preliminary analysis begins the work of **surveying the state and development of the field from its inception in Germany at the start of the 20th century to 1971**, by which time it had become synonymous with mathematical biology.

II. For most of the 20th century, theoretical biology was a research field that encompassed **philosophical discussions, mathematical analyses, and a broad array of concepts and topics** from different biological disciplines which were explored by an international community of scholars.

III. Some of our results, such as the **prominence of the organism concept** in early twentieth-century biology and the **lingering presence of the vitalism-mechanism debate**, are consistent with what historians of biology have recently contended (see Nicholson and Gawne 2015; Baedke 2019; Peterson and Hall 2020).

Scan here to explore the evolution of Theoretical Biology!





Affiliations

1 Department of Philosophy I, Ruhr University Bochum;
2. Faculty of Arts and Humanities, Heinrich Heine University Düsseldorf;
3. Department of Philosophy, George Mason University

RUHR UNIVERSITÄT BOCHUM RUB

hhu Heinrich Heine Universität Düsseldorf

GEORGE MASON UNIVERSITY

Emmy Noether-Programm Deutsche Forschungsgemeinschaft DFG

References

Angelov, D. (2020): Top2Vec: Distributed Representations of Topics, arXiv:2008.09470v1. https://arxiv.org/abs/2008.09470v1
Baedke, J. (2019). O organism, w here art thou? Old and new challenges for organism-centered biology. *J Hist Biol*, 52(2), pp. 293–324.
Blei, D. M, Ng, A. Y., Jordan, M. I. (2003): Latent Dirichlet allocation. *J Mach Learn Res*, 3 (March), pp. 993–1022.
Bokeh Development Team (2018). Bokeh: Python library for interactive visualization URL http://w ww.bokeh.pydata.org
De Vries, E., Schoonvelde, M., Schumacher, G. (2018): No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis*, 26 (4), pp. 417 – 430.
Honnibal, Matthew ; Montani, Ines; Van Landeghem, Sofie; Boyd, Adriane (2020): spaCy 3.1: Industrial-strength Natural Language Processing in Python. https://spacy.io/
Kutylow ksi, J. *et. al.* (2017): DeepL. https://w ww.deepl.com/
Malaterre, C. (2021): Topic-modeling of multilingual non-parallel corpora: Applying machine-translation to a philosophy of science corpus. Talk at the DS² 2021 online Conference, March 16, 2021. https://youtu.be/FTzmpNYZs3E
Malaterre, C., Chartier, J.F., Pulizzotto, D. (2019). What is this thing called philosophy of science? A computational topic-modeling perspective 1934–2015. HOPOS, 9(2), pp. 215–249. https://doi.org/10.1086/704372.
Malaterre, C., Lareau, F., Pulizzotto, D., St-Onge, J. (2020): Eight journals over eight decades: a computational topic-modeling approach to contemporary philosophy of science. *Synthese*. https://doi.org/10.1007/s11229-020-02915-6
McCallum, A. K. (2002): "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu.
McInnes, L., Healy, J. (2018): UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426. https://arxiv.org/abs/1802.03426v3
Nicholson, D.J., Gaw ne, R. (2015) Neither logical empiricism nor vitalism, but organicism: w hat the philosophy of biology w as. *Hist Philos Life Sci*, 37(4), pp. 345–381.
Noichl, M. (2019). Modeling the structure of recent philosophy. *Synthese*. https://doi.org/10.1007/s11229-019-02390-8
Peterson, E.L., Hall, C. (2020). "What is Dead May Not Die": Locating Marginalized Concepts Among Ordinary Biologists. *J Hist Biol*, https://doi.org/10.1007/s10739-020-09618-1
Rehurek, R., Sojka, P. (2010): [genism]. Softw are framew ork for topic modelling w ith large corpora. In: Proceedings of the LREC 2010 w orkshop on new challenges for NLP framew orks, pp. 45-50. https://radimrehurek.com/gensim/
Shen, Z. *et. al.* (2021): LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis, arXiv:2103.15348. https://arxiv.org/abs/2103.15348
Smith, R. (2019): tesseract 4.1.1. https://tesseract-ocr.github.io/