# What was Theoretical Biology?

## A Topic-Modelling Analysis of a Multilingual Corpus of Monographs and Journals, 1914-1945

**Böhm, Alexander**
alexander.boehm@rub.de
Department of Philosophy I, Ruhr University Bochum

**Reiners-Selbach, Stefan**
stefan.reiners-selbach@hhu.de
Faculty of Arts and Humanities, Heinrich-Heine-University Düsseldorf

**Baedke, Jan**
jan.baedke@rub.de
Department of Philosophy I, Ruhr University Bochum

**Fábregas Tejeda, Alejandro**
Alejandro.FabregasTejeda@ruhr-uni-bochum.de
Department of Philosophy I, Ruhr University Bochum

**Nicholson, Daniel J.**
dnicho@gmu.edu
Department of Philosophy, George Mason University

Over the course of the twentieth century, theoretical biology changed beyond all recognition. Although the field today is synonymous with mathematical biology, when it first emerged it had a drastically different agenda: to critically analyze the conceptual foundations of biology in order to resolve long-standing theoretical disputes, abstract from the 'burden of details,' and bring about the epistemic unification of biological science. The field began acquiring its now familiar mathematical character in the 1940s, as formal models became increasingly applied in different areas of biology, such as growth, ecology, genetics, and evolution. Regrettably, the early 'philosophical' period of theoretical biology has been almost completely forgotten and its existence is seldom acknowledged—let alone carefully examined (but see Nicholson & Gawne 2015, Baedke 2019). Much of this early discourse took place in a handful of book series, monographs, and journals, the majority of which were published in German (at least initially). Hence, it is perhaps not surprising that Anglophone scholars remain almost completely ignorant of this large, and surprisingly rich, body of literature.

Our aim is to analyze this forgotten corpus and rescue it from the dustbin of history. Our guiding question is: What did theoretical biology look like in the early 20th century? More specifically, we ask: (i) What were the central debates and topics? (ii) Who were the central authors and how international was the scientific community at the beginning? (iii) Can distinct language-(of-origin-)specific camps be identified in terms of the kinds of topics they addressed? (iv) What, where, and when did transitions occur in networks of authors and topics? (v) When, how, and why did the discipline develop its emphasis on formal modeling? At this early exploratory stage of the project, we operationalize these central questions mainly as a topic-modelling problem: (1) Which central topics can be identified and how does their 'share' of the documents develop? Which topic clusters can be identified? (2-3) Are certain topics dominated by particular authors, languages (of origin), and nationalities? (4) Can certain 'turning points' be identified? Additionally: (5) How steadily does the proportion of publications that use mathematical formulas increase over time? Is it gradual or rather discontinuous?

After (a) preparing and selecting documents for the corpus on a historical basis (encompassing monographs, book series and journal articles)—digitizing, and OCR-ing with tesseract where necessary—we (b) machine translate the non-German texts into German using the Google Translate API. As de Vries, Schoonvelde, and Schumacher (2018) argue for topic-modelling in general, and Malaterre (2021) for the special use-case of history of science, modern machine translations deliver useful results that are reliable for multilingual topic-modelling. Additionally, we plan to assess our translation accuracy with Malaterre's proposed "Semantic Topology Preservation Test" (2021). Then, we (c) preprocess the corpus: Following a general cleaning of common OCR-errors and stop words, we reduce the corpus to lemmatized adjectives and nouns via spaCy's POS tagging and lemmatization algorithms. We assume that the conceptual topics we aim to explore are mostly expressed in nouns and adjectives (see Jockers 2013, Malaterre et al. 2020). The preprocessed documents are then (d) analyzed with LDA topic-modelling, using gensim's MALLET-wrapper and (e) analyzed with top2vec, to cluster the documents thematically – enabling a different granularity and perspective, since top2vec does not treat the documents as bags-of-words and tends to generate few more general topics (see Angelov 2020). Finally, (f) we calculate document embeddings using UMAP and (g) visualize the embedding as an interactive scatter plot (with the option of time-period slices) with Bokeh, since the heterogeneity of our corpus does not allow for a simple linear visualization. We enrich the scatter plot with metadata for a mouse-over pop-up window, generated from the most important topics for each document, and color the documents by their top2vec cluster, complementing the visual clustering and topological distribution the document embedding shows. Thus, we create an interactive tool for exploration, hoping to motivate future research.

Moreover, we plan to utilize tesseract's equ language data to detect mathematical equations in documents. We take the use of mathematical formulas as a signal of affiliation with the mathematical side of the discourse on theoretical biology. This way, each document is assigned a gradual mathematization score. To model the mathematization of theoretical biology, we then analyze the mathematization scores per year and the scores' correlations with topics. The score can in turn be used for visual classification in the visualization by choosing different symbols for documents in the scatter plot based on their score.

## Bibliography

**Alt, W.; Deutsch, A.; Kamphuis, A.; Lenz, J. and Pfistner, B.** (1996). "Zur Entwicklung der Theoretischen Biologie: Aspekte der Modellbildung und Mathematisierung", in: *Jahrbuch für Geschichte und Theorie der Biologie* 3, pp. 7-59.

**Angelov, D.** (2020). *Top2Vec: Distributed Representations of Topics*, in: arXiv:2008.09470v1. https://arxiv.org/abs/2008.09470v1

**Baedke, J.** (2019). "O Organism, Where Art Thou? Old and New Challenges for Organism-Centered Biology", in: *J Hist Biol* 52, pp. 293–324. https://doi.org/10.1007/s10739-018-9549-4

**Blei, D. M.; Ng, A. Y.; Jordan, M. I.** (2003) "Latent Dirichlet allocation", in: *J Mach Learn Res* 3 (March), pp. 993–1022.

**Bokeh Development Team** (2018). *Bokeh: Python library for interactive visualization*. URL: http://www.bokeh.pydata.org

**De Vries, E.; Schoonvelde, M. and Schumacher, G.** (2018). "No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications", in: *Political Analysis*, 26 (4), pp. 417 – 430. https://doi.org/10.1017/pan.2018.26

**Honnibal, M.; Montani, I.; Van Landeghem, S. and Boyd, A.** (2020). *spaCy 3.1: Industrial-strength Natural Language Processing in Python*. https://spacy.io/

**Jockers, M.** (2013). *"Secret" recipe for topic modeling themes*. https://www.matthewjockers.net/2013/04/12/secret-recipe-for-topic-modeling-themes/

**Laubichler, M.** (2001). "Mit oder ohne Darwin? Die Bedeutung der darwinschen Selektionstheorie in der Konzeption der Theoretischen Biologie in Deutschland von 1900 bis zum Zweiten Weltkrieg", in: Hoßfeld U, Brömer R (eds): *Darwinismus und/als Ideologie. Verhandlungen zur Geschichte und Theorie der Biologie*, Band 6. VWB, Berlin, pp. 229–262.

**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. and Pulizzotto, D. (2019**). "What is this thing called philosophy of science? A computational topic-modeling perspective 1934–2015", in: *HOPOS*, 9 (2), pp. 215–249. https://doi.org/10.1086/704372.

**Malaterre, C.; Lareau, F.; Pulizzotto, D. and 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. and Healy, J.** (2018). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction", in: *ArXiv e-prints* 1802.03426. https://arxiv.org/abs/1802.03426v3

**Nicholson, D.J. and Gawne, R.** (2015). "Neither logical empiricism nor vitalism, but organicism: what the philosophy of biology was", in: *HPLS* 37, pp. 345–381. https://doi.org/10.1007/s40656-015-0085-7

**Noichl, M.** (2019). "Modeling the structure of recent philosophy. Synthese." https://doi.org/10.1007/s11229-019-02390-8

**Rehurek, R. and Sojka, P.** (2010). [genism]. "Software framework for topic modelling with large corpora", in: *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, pp. 45-50. https://radimrehurek.com/gensim/

**Smith, R.** (2019). *tesseract 4.1.1*. https://tesseract-ocr.github.io/