What Was Theoretical Biology?
A Topic-Modelling Analysis of a Multilingual Corpus of Monographs and Journals, 1901-1971

Alexander Böhm1, Stefan Reiners-Selbach2, Jan Baedke1, Alejandro Fábregas-Tejeda1, Vera Straetmanns1, Daniel J. Nicholson3

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 discussion 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.

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 use mathematical formulas increase over time (and in which thematic contexts)?

II Corpus

Fig. 2: Sentiment works in early theoretical biology

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

Fig. 4: Current workflow of analysis.

Ill 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 (Angeli 2020) for a contrasting, embedding-based Topic Modelling approach. These results were then used as clustering for the (h) document embedding (UMAP, McMahan 2018), which we visualized in an (i) interactive scatter plot using bokeh (Bokhe Development Team 2018).

IV Results

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 linguistic presence of the vitalism-mechanism debate, are consistent with what historians of biology have recently contended (see Nicholson and Gwenn 2015; Baedke 2019; Peterson and Hall 2020).

Scan here to explore the evolution of Theoretical Biology!

References

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.