Synthese (2022) 200:242
https://doi.org/10.1007/s11229-022-03722-x

ORIGINAL RESEARCH

The early days of contemporary philosophy of science: novel insights from machine translation and topic-modeling of non-parallel multilingual corpora

Christophe Malaterre[1] · **Francis Lareau**[2]

Received: 1 July 2021 / Accepted: 27 April 2022 / Published online: 31 May 2022

Abstract

Topic model is a well proven tool to investigate the semantic content of textual corpora. Yet corpora sometimes include texts in several languages, making it impossible to apply language-specific computational approaches over their entire content. This is the problem we encountered when setting to analyze a philosophy of science corpus spanning over eight decades and including original articles in Dutch, German and French, on top of a large majority of articles in English. To circumvent this multilingual problem, we use machine-translation tools to bulk translate non-English documents into English. Though largely imperfect, especially syntactically, these translations nevertheless provide correctly translated terms and preserve the semantic proximity of documents with respect to one another. To assess the quality of this translation step, we develop a "semantic topology preservation test" that relies on estimating the extent to which document-to-document distances have been preserved during translation. We then conduct an LDA topic-model analysis over the entire corpus of translated and English original texts, and compare it to a topic-model done over the English original texts only. We thereby identify the specific contribution of the translated texts. These studies reveal a more complete picture of main topics that can found in the philosophy of science literature, especially during the early days of the discipline when numerous articles were published in languages other than English.

Keywords Machine translation · Topic model · LDA · Philosophy of science · History of philosophy of science · Multilingual corpus

✉ Christophe Malaterre
malaterre.christophe@uqam.ca

[1] Département de philosophie and Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal (UQAM), 455 Boulevard René-Lévesque Est, Montreal, QC H3C 3P8, Canada

[2] Département d'informatique, Université du Québec à Montréal (UQAM), 201 av. Président-Kennedy, Montreal, QC H2X 3Y7, Canada

1 Introduction

Computational textual analyses such as topic modeling are well proven methodologies to investigate the semantic content of large corpora.[1] To be effective, most of these methodologies require corpora to be homogeneous in terms of languages. In practice however, many corpora include texts that are written in several languages, making it impossible to apply these methodologies across such multilingual corpora. This is precisely the problem we encountered when setting to analyze a philosophy of science corpus of 16,917 articles: starting in the 1930s and spanning over eight decades, this corpus was largely in English but also included some 1016 articles that were written in Dutch, German or French. Being relatively marginal, these non-English texts could be set aside and computational tools deployed on the (much larger) English-language portion of the corpus, as done in (Malaterre et al., 2020). Yet, the relatively high proportion of non-English-language texts in the earlier decades of the discipline of the philosophy of science raised questions about the reliability of an English-only approach. We thereby investigated different means to gain novel insights on pre-WWII philosophy of science by applying a diverse set of computational tools not just to the English portion of the corpus but also to the non-English portion which, at the time, accounted for over half of the publications in specialized journals.

The multilingual characteristic of textual data is of course a well-known issue for computational textual analyses, notably topic-modeling. Solutions to this issue can be sorted into two broad families of approaches. The first one is to develop topic-modeling algorithms that directly work on the original multilingual corpora while finding ways to bridge languages.[2] In cases where corpora include parallel or comparable texts (e.g. texts that are simultaneously available in different languages as high quality translations of one another such as proceedings of the European parliament, or texts in different languages that are supposed to roughly exhibit the same distributions of themes or subjects though not being exact translations of one another, as in Wikipedia), language bridging can be done by aligning topics at a sentence or document level (De Smet & Moens, 2009; Mimno et al., 2009; Zhao & Xing, 2007). In cases where corpora are not aligned (e.g. corpora that include articles from journals accepting publications in multiple languages), other resources can be used to bridge languages. Some have proposed to use multilingual dictionaries or other lexical resources (Boyd-Graber & Blei, 2009; Jagarlamudi & Daumé, 2010; Zhang et al., 2010). Others have proposed to use concept trees, for instance inferred from WordNet or based on user input (Boyd-Graber & Blei, 2009; Hu, Boyd-Graber, et al., 2014; Hu, Zhai, et al., 2014). Still others combine data from text alignment and lexical resources (Hu et al., Hu, Zhai, et al., 2014). While many of these approaches build on probabilistic latent topic models such as probabilistic Latent Semantic Analysis (pLSA) (Landauer et al., 1998) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003), other recent developments rely on other

[1] For instance, see (Pence & Ramsey, 2018) for an overview in the context of the philosophy of science.

[2] Such multilingual topic-modeling approaches have been used, among others, to infer semantic similarities between terms of different languages, notably with a view to improving machine translation but also to classify documents in different languages, to investigate sentiment analyses or to retrieve information within documents written in different languages.

approaches to topic-modeling, for instance anchor-based topic models (Yuan et al., 2018).

The second broad family of approaches to the challenges raised by multilingual corpora consists in harnessing the power of machine translation to convert the multilingual textual data into monolingual data (typically English) and use existing monolingual computational tools to analyse the textual content. Machine translation can be done for complete texts or only for specific terms such as those that figure in a term-document matrix (TDM) (Lucas et al., 2015). Some have proposed to combine machine translation with multilingual topic modeling algorithms, for instance by translating just a subset of the corpus, thereby constituting a parallelized sub-corpus that makes it possible in turn to learn term pairings that can then be projected over the rest of the corpus (Pruss et al., 2019). However, the use of machine translation raises questions of its own: Are machine-translated texts good enough for (some types of) computational textual analyses despite not being as polished as expert-based translations? Do they lead to similar results as expert-based translations? Does the choice of machine-translation service, for instance Google Translate over DeepL or Microsoft Translator, affect results? Comparisons of TDMs and of topic-models obtained from machine-translated texts and from expert-based translations have shown high similarities (Vries et al., 2018). Studies based on LIWC (Linguistic Inquiry Word Count) indices, which count words in psychologically meaningful categories, have also shown a high similarity of results between machine-translated texts and expert-based translations (Windsor et al., 2019). Some have however noted a slight degradation of results when machine translation is conducted not over the full-texts but only over the terms that of the TDMs, while differences in terms of machine learning tool remained minimal (Ruder et al., 2019).

Weighting the pros and cons of these two broad families of approaches, we opted for full-text machine-translation and the "vanilla" version of LDA topic-modeling. The size of the non-English portion of our corpus being quite limited, full-text machine-translation costs were reasonable. Full-text translations in English also appeared as a better alternative to TDM-only translation, not only to secure more reliable results as shown by (Ruder et al., 2019) but also to provide a means to go deeper into the textual content: in particular, full-texts access is much needed to interpret topics in an accurate and meaningful way; full-texts translations also offer the potential for more in-depth ad hoc analyses if needed. Furthermore, a preliminary analysis of our original corpus revealed numerous OCR and encoding issues among texts of the first half of the nineteenth century. Interestingly, translation tests revealed that machine translation is able to fix most of these issues, thereby proving a translated corpus of a much better quality than the original one. Also, since Google Translate had already been assessed in the context of LDA topic-modeling (Vries et al., 2018), we also opted for Google Translate. Yet, since the corpus was non-parallel, we lacked means for checking the quality of the translation and its effect on TDMs and topic-models. This is the reason why we devised a specific test of "semantic topology preservation": our objective was to somehow get an indication of the extent to which the machine-translation step preserved some key properties of the semantic space that mattered for bag of words computational approaches such as topic-modeling.

As a result, our contribution here is twofold. First, on a methodological level, we assess the reliability of implementing machine translation for the purpose of topic-modeling. Second, on a content level, we examine the history of the philosophy of science from the 1930s till the late 2010s. In particular, we compare topic models with and without the translated texts. The paper is organized as follows. We first describe the data. We then present the research design we followed along its two major methodological stages: machine translation and topic modeling. Third, we describe the results that were obtained. These results concern in particular machine-translation quality assessments, synchronic and diachronic topic-modeling and comparisons with English-only topic-modeling.[3] Finally, we discuss both the methodology and the results that were obtained.

2 Data

The corpus consists of the full-text content of eight major philosophy of science journals from 1931 (the first issue of the earliest published journal) up until 2017: the *British Journal for the Philosophy of Science (BJPS)*, the *European Journal for Philosophy of Science (EJPS)*, *Erkenntnis, International Studies in the Philosophy of Science (ISPS)*, the *Journal for General Philosophy of Science (JGPS,* founded in the 1970s as *Zeitschrift für allgemeine Wissenschaftstheorie)*, *Philosophy of Science (PS)*, *Studies in History and Philosophy of Science Part A (SHPSA)* and *Synthese*.[4] Overall, the corpus comprises 16,917 articles, including 1016 in languages other than English (i.e. about 6%) (see Table 1).[5] These articles amount to over 65 million word occurrences (for an average word count of about 3900 words per article). Out of the 1016 non-English articles, 72% are in German, 17% in Dutch and 11% in French. Some 37% of these articles were published before WWII. In these early days of the philosophy of science, non-English articles represented about 54% of publications in the selected journals. The content of non-English articles is therefore of a high interest when it comes to understanding the philosophy of science during its nascent phase, in particular through the lenses of the first two publication venues in the field, *Erkenntnis* (founded in Germany) and *Synthese* (in The Netherlands). In the second half of the twentieth century, the share of non-English articles substantially decreased over time,

[3] The topic model can be explored on https://philscitopics.uqam.ca/.

[4] The articles were downloaded from JSTOR and the publishers Internet platforms (Elsevier, Oxford University Press, Springer, Taylor and Francis and University of Chicago Press) between May and June 2018. Philosophy of science is of course published in many other venues, the entirety of which we cannot hope to cover. These include other journals, be they more general philosophy journals (e.g. *Mind*), more specialized philosophy of science journals (e.g. *Studia Logica, Hyle, Biology and Philosophy*), or even science journals (e.g. *Bioscience*). Philosophy of science is also published in many non-English languages (e.g. *Principia, Epistemologia, Philosophia Scientiae, Theoria*) and in numerous books and edited volumes. By selecting 8 of the major general philosophy of science journals in English language, including some of the earliest ones, our objective is to provide a representative perspective on the thematic content of philosophy of science and its evolution over the past 8 decades.

[5] To the extent feasible, we only retained articles with research content. We thereby excluded book-reviews, editorials, errata, and very short texts such as discussion notes (less than 4 000 characters). Some very minor differences in article numbers compared to (Malaterre et al., 2020) are due to language detection methods (15,901 English articles vs 15,899 previously).

Table 1 Corpus articles, by journal and by language, with publication periods

Journals	Publication	Articles per language (from start-date until 2017)				Total
(Alphabetic order)	Periods	English	German	Dutch	French	
British Journal for the Philosophy of Science (BJPS)	1950–present	1862	–	–	–	1862
Erkenntnis	1931–1940; 1975–present	1867	251 (§)	–	9 (§)	2127
European Journal of Philosophy of Science (EJPS)	2011–present	156	–	–	–	156
International Studies in the Philosophy of Science (ISPS)	1986–present	560	–	–	–	560
Journal for General Philosophy of Science (JGPS)	1970–present	515	410	–	4	929
Philosophy of Science (PS)	1934–present	4604(*)	1	–	–	4605
Studies in History and Philosophy of Science – Part A (SHPSA)	1970–present	1420	1	–	–	1421
Synthese	1936–1939, 1946–1949, 1955–present	4917	74 (‡)	171 (§)	95 (‡)	5257
Total		15,901	737	171	108	16,917

(*) including proceedings of the Philosophy of Science Association meetings; (§) mostly before 1940; (‡) mostly before 1960.

down to 2% from the 1990s onward, resulting in an average of about 6% over the 8 decades. However, one journal (*Journal for General Philosophy of Science* founded in Germany in 1970) still published about 44% of non-English articles at that time. As a result, the inclusion of such articles in the corpus may also affect the topical profile of that particular journal (even if only marginally affecting the overall topic distribution for this period) (Fig. 1).

3 Methods

The research design we followed is articulated around two main stages (see Fig. 2). The first one concerns the translation of the non-English articles into English and the assessment of the translation quality, especially for the purposes of bag-of-word textual analyses. The second stage includes the topic-modeling of the entire corpus,

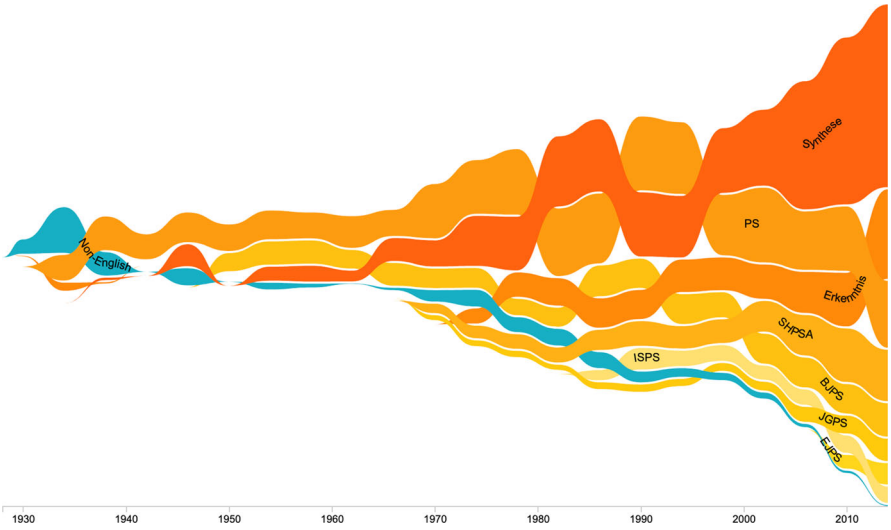


Fig. 1 Corpus articles over time, per journal for articles in English (orange streams), with a specific stream for non-English articles independently of publication journals (blue stream) (width of each stream proportional to the number of articles; streams sorted by decreasing size at each 4-year time-period; width of *SHPSA* in the 2010s corresponds to roughly 80 articles per year; visualization: RAWGraphs (Mauri et al., 2017))

its analysis from both synchronic and diachronic perspectives, and its comparison with a previous topic-modeling done on the English-only portion of the corpus as in (Malaterre et al., 2020). Altogether, these two main stages comprise 7 distinct methodological steps.

3.1 Machine translation and translation quality assessment (Stage A)

We organized all journal articles and their metadata into a Python Pandas dataframe. As a first step (Step 1), to identify the non-English articles, we used three automatic language-detection methods (language metadata when available in the downloaded documents, langid and langdetect packages) and extracted all documents for which non-English was detected at least once.[6] The corpus was split into four sub-corpora (English, German, Dutch, French). Each non-English sub-corpus was sent to Google Translate by chunks of ~ 25,000 characters (closest dot after 25,000 characters).[7] Translation results were then reassembled into articles.

[6] https://pypi.org/project/langid/ (Lui & Baldwin, 2012); https://pypi.org/project/langdetect/ (Shuyo, 2010). We found out that differences in language prediction between methods could be triggered by articles that were written with significant sections in different languages (for instance, an article in German with large portions of cited text in French and in English). This was notably the case for 5 articles written in several languages and that were removed for steps 2 and 3 of the methodology (but kept for all of stage B): 2 articles that were initially classified as English (from *BJPS* and *JGPS*), 1 as French (*Erkenntnis*), 1 as Dutch (*Synthese*) and 1 as German (*Erkenntnis*).

[7] Google Translate requirements; Google API (translate_v2 from google.cloud) accessed on 30 march 2020 at https://cloud.google.com

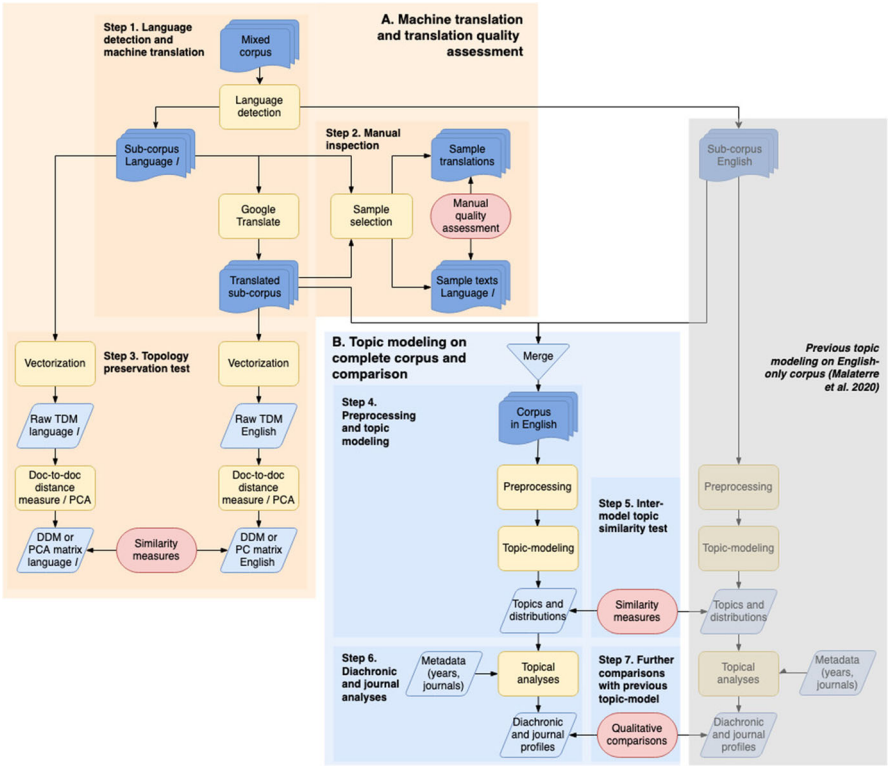


Fig. 2 Research design. Two major stages divided into seven methodological steps (Textual corpora in dark blue, data in light blue, operations in light orange, data comparisons in red, previous study in shaded grey)

To assess the quality of the machine translation step, we started with a manual inspection (Step 2). We randomly selected 10 texts in each non-English language and for each we inspected their first 500 words. In particular, we tracked three types of problems in the texts with possible impact on computational textual analyses: (i) spelling issues in the original text that induced issues in the translation (e.g. problems resulting from OCR and encoding); (ii) inaccurate terms that were introduced by the translation, and (iii) spelling issues that were present in the original text and that have been corrected through machine-translation. We measured these three types of issues to add a quantitative perspective on our manual translation quality check.

In addition, to provide a systematic translation assessment over the whole non-English language, we chose to compare the relative distances between documents before and after translation (Step 3), the rationale being that documents that were close to one another before translation should also remain so after (in other words, the structure or topology of the document-term spaces should be preserved). Hence the following "semantic topology preservation test". We first constructed the document-term matrices of the three sub-corpora in their original languages (Dutch, German, French): $\mathbf{W_l} = [w_{ij}]_{M(l)\times N(l)}$, where $l \in \{\text{Dutch, German, French}\}$, $M(l)$ is the size

Table 2 Matrix dimensions throughout the "semantic topology preservation test"

Language l	Dutch	German	French
Number of articles $N(l)$	170	736	107
Number of terms in the original lexicon $M(l)$	48,270	212,132	27,384
Number of terms in the translated lexicon $M^*(l)$	31,005	90,866	20,981
Reduced dimension from SVD	100	100	100

5 Articles written with multiple languages (for instance in German with large quotes from original French and English texts) were removed for the semantic topology preservation test

of the lexicon of the sub-corpus of language l, $N(l)$ is the number of articles of the sub-corpus of language l, and w_{ij} is the frequency of word i in article j belonging to the sub-corpus of language l (see Table 2). In parallel, we built the three document-term matrices of the translations: $\mathbf{W}^*_{\boldsymbol{l}} = [w^*_{ij}]_{M^*(l)\times N(l)}$ (where l is the original language, i.e., Dutch, German or French, $M^*(l)$ is the size of the lexicon of the sub-corpus consisting of the English translations of articles initially in language l, $N(l)$ is the number of articles in language l, and w^*_{ij} is the frequency of word i in article j belonging to that same sub-corpus. We then took two parallel approaches. (1) For assessing matrix similarity with the Mantel and RV approaches, we calculated the document-document distance matrices $\mathbf{D}_{\boldsymbol{l}}$ and $\mathbf{D}^*_{\boldsymbol{l}}$ of dimensionality $N(l)$ x $N(l)$ by measuring the documents pairwise Euclidian or cosine distances; we then investigated the similarity of the distance matrices $\mathbf{D}_{\boldsymbol{l}}$ and $\mathbf{D}^*_{\boldsymbol{l}}$ for each language l by calculating their Mantel and RV coefficients. (2) For assessing matrix similarity with the Procrustes approach, we followed (Peres-Neto & Jackson, 2001) and conducted a principal component analysis (PCA) before calculating the Procrustes coefficient of the resulting matrices.[8]

[8] The Mantel, RV, and Procrustes approaches aim at assessing the similarity between two matrices. The Mantel approach, originally introduced in epidemiology by Mantel (1967) and now widely used in ecology (Legendre & Legendre, 2012), consists in calculating the correlation between two distance matrices. Given $\mathbf{D}$ and $\mathbf{B}$ two N x N distance matrices, given $\overline{d}$ and $\overline{b}$ the means of their respective off-diagonal elements, their Mantel coefficient is: $r_M = \left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(d_{i,j}-\overline{d})(b_{i,j}-\overline{b})\right] / \sqrt{\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(d_{i,j}-\overline{d})^2\right]\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(b_{i,j}-\overline{b})^2\right]}$. The RV coefficient arises as a generalization of Pearson's correlation (Escoufier, 1973). It is computed as the ratio of the covariance to the square-rooted product of the variances: $RV = \text{trace}\left\{\mathbf{S}^{\mathrm{T}}\mathbf{T}\right\} / \sqrt{\text{trace}\left\{\mathbf{S}^{\mathrm{T}}\mathbf{S}\right\} \times \text{trace}\left\{\mathbf{T}^{\mathrm{T}}\mathbf{T}\right\}}$ where $\mathbf{S}$ is a positive semi-definite matrix (i.e., $\mathbf{S}$ is such that there exists a matrix $\mathbf{X}$ such that $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$). Finally, the Procrustes approach consists in applying a Procrustean superimposition that scales and rotates matrices so as to maximize their fit (Mardia et al., 1979, pp. 416–419). The sum of the squared residuals between configurations in their optimal superimposition can then be used as a metric of similarity (Peres-Neto & Jackson, 2001). For each coefficient, statistical tests and comparisons to the null hypothesis were performed by means of random permutations on rows and columns. Packages: https://www.rdocumentation.org/packages/vegan/versions/2.4-2/topics/mantel, https://www.rdocumentation.org/packages/FactoMineR/versions/2.4/topics/coeffRV, https://www.rdocumentation.org/packages/vegan/versions/2.4-2/topics/procrustes

3.2 Topic-modeling on complete corpus and comparison (Stage B)

For the topic-modeling task (Step 4), we followed the same methodological steps as in (Malaterre et al., 2020) (since one of our objectives was to compare results). As a preliminary step, we assembled the sub-corpus of English articles together with the three translated sub-corpora (in Python). We then preprocessed the data: we removed terms such as determinants, prepositions or pronouns with part-of-speech (POS) tagging and lemmatized the textual data in order to reduce the number of word variants[TreeTagger package (Schmid, 1994) with Penn TreeBank POS tagging (Marcus et al., 1993)]. In addition, we filtered out words that occurred in fewer than 50 sentences in the corpus. These operations resulted in a lexicon of 23,762 distinct terms distributed among over the 16,917 articles. For the topic-modeling itself, we used the Latent Dirichlet Allocation (LDA) algorithm, following (Blei et al., 2003) and (Griffiths & Steyvers, 2004),[9] with a number of topics K set to 25 as in the previous topic-modeling of (Malaterre et al., 2020) (hyperparameters similarly set to the same values: alpha = 0.4 and eta = 0.01).[10] Implementing the LDA algorithm on that basis thereby resulted in 25 topics (i.e., 25 sets of terms with their respective probabilities) and the probability distributions of these topics in each one of the 16,917 articles of the corpus. Topic interpretation was done by examining the most probable words in each topic and by retrieving the articles in which a given topic was in turn the most probable. This resulted in attributing meaningful labels to all 25 topics. In addition, topics were grouped into clusters on the basis of both expert knowledge of the field and their correlation within the corpus documents.

To assess the extent to which the addition of the non-English sub-corpora affected the overall topic-modeling compared to the English-only model of (Malaterre et al., 2020), we measured how similar topics from the new topic-model were to topics from the previous one (Step 5). More specifically, given the topic-word probability distribution matrix of the new topic-model $\mathbf{\Phi} = [\Pr(w|z)]_{MxK}$ (where $\Pr(w|z)$ is the probability of finding word w in topic z, M the size of the lexicon and $K = 25$) and of the previous one $\mathbf{\Phi}^{\dagger} = [\Pr(w^{\dagger}|z^{\dagger})]_{M^{\dagger}xK}$, we measured the Hellinger pair-wise distances between topics z and $z^{\dagger}$ as the distance between their respective probability distribution vectors $\Pr(w|z)$ and $\Pr(w^{\dagger}|z^{\dagger})$ for all words $w = w^{\dagger}$.[11] The result of this "inter-model topic similarity test" is a distance matrix that makes it possible to assess

[9] https://pypi.org/project/lda/

[10] LDA is a generative statistical model that finds out optimal probability distributions of words in topics and of topics in documents provided a number K of topic is chosen beforehand. The choice of $K = 25$ in the previous topic modeling was done by generating a set of models with different values of K and comparing them based on expert judgment (see (Malaterre et al., 2020, Sect. 2)). LDA-generated topics are sets of words with their probabilities. Once sorted by decreasing order of probability, these terms are usually informative of the semantic content of each topic. Retrieving texts in which topics are the most probable helps in formulating a meaningful interpretation of each topic, in particular when its most likely terms may appear ambiguous. Sometimes, depending on the objectives of the study, it can be helpful to increase the number of topics so as to capture contextual variations of terms of interest (for instance, an earlier topic-model of the single journal *Philosophy of Science* resulted in identifying 5 topics on explanation out of 126 meaningful topics; see (Malaterre et al., 2019, p. 225)).

[11] Hellinger distance from Gensim package (https://pypi.org/project/gensim/). We thank an anonymous reviewer for this suggestion.

the extent to which the topics of the new model (on the complete corpus) are similar or not to topics of the previous model (on the English-only texts).

Adding publication years and journals to the model made it possible to compute topic probability distributions for given time-periods or per journal or both (Step 6). Given document-topic probability distribution matrix $\mathbf{\Theta} = [\Pr(z|d)]_{K\text{x}N}$ (where $\Pr(z|d)$ is the probability of finding (words that express) topic z in document d, and $N = 16{,}917$), one can average the $\Pr(z|d)$ for all documents d that belong to a given time period or to a given journal or both. This step therefore led to a diachronic topic model for the entire corpus as well as (diachronic) journal profiles. Similarly to what had been done before, the length of time-periods was set to 4 years, which makes it possible to average out accidental yearly fluctuations while still providing a fine-enough temporal granularity.

Finally, we qualitatively compared these results with those of the previous topic-model (Step 7). This was done both for the overall diachronic topic models and per journal. This led us to examine in more details the pre-WWII period for which the proportion of non-English was significant.

4 Results

In what follows, we present in turn results obtained at each one of the seven methodological steps.

4.1 Machine translations

The output of the machine translation step consists of the English translation of the three sub-corpora (originally in German, Dutch and French). As mentioned earlier, most of these texts concern the earlier period of the philosophy of science (pre-WWII) and specifically two journals: *Erkenntnis* (in German) and *Synthese* (in Dutch). The bulk of texts in French are from *Synthese* in the 1950s–1960s, while a significant share of *JGPS* articles from the 1970s up until 2000s were in German. One peculiarity of the corpus as we collected it was a relatively high number of OCR or encoding issues. As a result, numerous question marks and misspellings plagued the original texts, especially before the 1960s. In such cases, machine translation can be of significant help. For instance, text segments such as "VAssociation fran?aise des Sciences politiques" or "l'histoi re" (which should actually be "l'Association française des Sciences politiques" and "l'histoire") were correctly translated into "the French Association of Political Sciences" and "history". While we could not find any simple way of systematically identifying OCR or encoding issues, we measured the impact of machine translation on the presence of question marks. In particular, we measured the number of question marks that were present at the beginning or inside words[12] (as in "?ventuellement" or in "fran?aise") and at the end or outside words[13] (as in "entende? ment" or " ?"), both in the original texts and in their English translations, and also

[12] Regex = "\\? + \\p{L}".

[13] Regex = "\\? + [^\\p{L}\\?]".

Table 3 Effect of machine translation on question marks

Sub-corpora	Question marks	In original texts	Average per text	In translations	Average per text	Reduction
Non-English sub-corpus (1016 texts)	Inside words	316,049	310.5	77	0.1	99.9%
	Outside words	97,145	95.4	83,013	81.5	14.6%
	Total	413,194	405.9	83,090	81.6	79.9%
English sub-corpus (15,901 texts)	Inside words	52,976	3.3	–	–	–
	Outside words	253,798	16.0	–	–	–
	Total	306,774	19.3	–	–	–

in English texts for reference (see Table 3).[14] The data show a very high number of question marks in original non-English texts (indeed much higher than in original English texts) that was drastically reduced through machine translation by a factor of about 80% (question marks inside words were practically all eliminated). This is a significant benefit of machine translation.

4.2 Translation assessment: manual quality inspection

Manual assessment of the machine translations aimed at specifically identifying issues with potential impact on bag-of-word analyses. Therefore, we did not focus on style or syntactic issues in the translated texts but sought issues at word level. In particular, we identified in sample texts: (i) anomalies present both in the original text and in the translation (likely due to OCR and encoding issues, such as those identified in the previous section); (ii) anomalies introduced by the translation while not present in the original text, (iii) anomalies present in the original text but corrected through machine-translation (see Table 4 for examples). Estimates of the frequency of these anomalies indicate that there were, on average, 12.2 anomalies per 100 words in the original texts (Table 5, type (i) and type (iii) anomalies). Machine translation corrected about 77% of these word anomalies, while introducing some 11% new anomalies, hence a net improvement of 66%. Machine translation therefore proves to be very faithful at a word level (as already known (e.g., Vries et al., 2018)). Furthermore, in cases where texts are plagued by OCR and encoding issues, machine translation proves as a powerful ally to improve terminological content. Therefore, from a pure bag-of-word perspective, machine translation is likely to improve the accuracy of analyses.

[14] The number of question marks inside words in English texts (3.3) appears to be due to the presence of words in foreign languages (e.g., citations in German or Greek), algebraic expressions, OCR errors and encoding errors, and predominantly in *Synthese*, *Erkenntnis* and the *GJPS*.

Table 4 Extracts of samples of texts in original languages with corresponding machine translations

Article	Original text	Machine translation
Bourgin, Georges (1956) Neutralité. *Synthese* 10: 265–269	[…] Le mot et le concept de neutralit? doit aux circonstances actu elles de la politique internationale qu'on s'arr?te sur sa ***significa tion***. Et d'abord, sans ?laborer une ?tude vraiment s?mantique du terme, il convient d'indiquer tout de suite qu'il se rattache au pronom ind?fini latin ?neuter", ?ni l'un ni l'autre". C'est ? la m?me origine qu'est ***d?*** le mot neutralisme, dont nous examinerons rapidement aussi le sens. […]	[…] The word and the concept of neutrality owes to the current circumstances of international politics that we stop on its ***significance***. And first, without developing a really semantic study of the term, should immediately indicate that it is related to the indefinite Latin pronoun neuter ", neither neither". It is the same origin that is ***he*** word neutralism, whose meaning we will also quickly examine. […]
Piaget, Jean (1956) Méthode axiomatique et méthode opérationnelle. *Synthese* 10: 23–43	Nous examinerons les relations entre les m?thodes axiomatique et op?rationnelle selon trois cat?gories de probl?mes: un probl?me psy chologique, d'abord, car ces deux sortes de m?thodes, tout en ?tant utilis?**es** dans un but logique, c'est-?-dire normatif, constituent par ailleurs des d?marches de la pens?e, donc des faits psychologiques; un probl?me de relation entre la psychologie et la logique, ensuite, car les op?rations constituent le point de contact entre ces deux disciplines; un probl?me ***?pist?mologique***, enfin, car ces m?mes questions sont significatives du point de vue de la connaissance en g?n?ral	We will examine the relations between the axiomatic and operational methods according to three categories of problems: a psychological problem, first, because these two kinds of methods, while being used **es** with a logical aim, that is to say a normative one, constitute in addition steps of the thought, therefore psychological facts; a problem of relation between psychology and logic, then, because the operations constitute the point of contact between these two disciplines; a ***pistemological*** problem, finally, because these same questions are significant from the point of view of knowledge in general
Rougier, Louis (1939) La relativité de la logique. *Erkenntnis* 8: 193–217	La logique est d?finie comme l'art de bien conduire sa pens?e, Fart de raisonner avec justesse. Raisonner c'est montrer que **cer?** taines propositions sont n?cessairement vraies ? supposer que d'autres ***propositions***, appel?es **pr?misses**, soient tenues pour telles. C'est dans les sciences du raisonnement, c'est-?-dire en math?ma thiques, que la logique, pour la premi?re fois, est entr?e en action	Logic is defined as the art of properly conducting one's thinking, the art of reasoning with accuracy. To reason is to show that **cer?** some propositions are necessarily true? assume that other ***proposals***, called **presumptions**, are held to be such. It is in the sciences of reasoning, that is to say in mathematics, that logic, for the first time, came into action

Three types of anomalies have been identified, with possible impact on computational textual analyses: (i) anomalies that are present both in the original text and in the translation (in bold); (ii) anomalies that are introduced by the translation and that were not present in the original text (in bold italics), (iii) anomalies that were present in the original text and that have been corrected through machine-translation (underlined)

Table 5 Number of word anomalies per hundred words (phw) in the machine translation excerpts when compared to original texts

Document original language	Type (i) anomalies (present both in original text and translation) (phw)	Type (ii) anomalies (introduced by the translation) (phw)	Type (iii) anomalies (corrected with machine-translation) (phw)	Number of articles in sub-corpora
German	3.0	1.4	9.6	736
Dutch	2.2	1.5	6.1	170
French	2.2	0.9	13.1	107
Estimated average*	2.8	1.4	9.4	1013

Based on manual analysis of 10 sample texts in each language
*Prorated to number of articles in each sub-corpus
As in Table 2, 5 articles written with multiple languages (for instance in German with large quotes from original French and English texts) were removed from the corpus for these analyses

4.3 Translation assessment: semantic topology preservation test

The results of the similarity tests that were carried on between the DDMs of the original texts and of their translations are summarized in Table 6. Though slightly varying depending on language and on similarity measures (Mantel, Procrustes or RV), the results show a very high degree of similarity between DDMs before and after machine translation. This very high degree of similarity indicates that machine translation preserves very well the structure of semantic spaces: documents that are close to one another in terms of their word frequencies in their original language also

Table 6 Results of similarity tests between DDMs of the original texts and of their machine translations

Similarity measures	Dutch	German	French
Mantel *r* (a)	0.86*	0.87*	0.90*
Mantel *r* (b)	0.97*	0.95*	0.98*
RV coefficient (a)	0.78**	0.80**	0.85**
RV coefficient (b)	0.97**	0.94**	0.97**
Procrustes correlation	0.97*	0.92*	0.96*

Results with cosine distance tend to be inferior to results with Euclidian distance. This could be explained by remaining noise patterns in documents and the fact that cosine distance, contrary to Euclidian distance, tends to conflate documents with similar terms but different lengths (Aggarwal, Hinneburg, and Keim 2001; Francois, Wertz, and Verleysen 2005). In any case, note that our final goal here was to use distances not for measuring the similarity of documents within a corpus, but for building distance matrices of different set of documents (these matrices being then subjected to similarity analyses)
(a) cosine distance; (b): Euclidian distance; *: $p = 0.001$ (999 permutations); **: $p < 0.001$

tend to be close to one another in their machine translation versions. In so far as bag-of-words textual analyses rely on term frequencies, they should therefore give similar results when conducted on either the original texts or their machine translations. For instance, if two documents were to express similar topic distributions in their original language (and therefore be close to another in their word distribution patterns), then their machine translations should also express similar topic distributions. The semantic topology preservation test therefore gives an a posteriori estimate of the reliability of machine translation for bag-of-words approaches. It is especially useful for non-parallel corpora for which there is no gold-standard translation to which machine translations might be compared (remember that (Vries et al., 2018) checked machine translations against gold-standard human-made translations). Comforted by random manual inspections (as seen above), the semantic topology preservation test expands our level of confidence in the usability of the complete translated sub-corpora for our topic-modeling purposes.

4.4 Topic model: synchronic perspective

The 25 topics provide a high-level perspective on the main research themes that characterize the discipline of the philosophy of science. These topics can be examined through the lens of their top-words (as in Table 7) but also through their correlations with one another in documents (as depicted in Fig. 3). A first set of five topics that concern philosophy of language and logic (Cluster A). On the logic side, topics TRUTH, FORMAL and MATHEMATICAL appear the most correlated, with such top-terms as "logic", "truth", "mathematical", "set" or "function". The philosophy of language is denoted by topics SENTENCE and LANGUAGE, with such top-terms as "language", "sentence", "meaning" as well as "context" or "reference". Nearby, cluster B includes three correlated topics that are indicative of epistemology and theory of knowledge: EPISTEMOLOGY is characterized by such keywords such as "belief" or "knowledge", while ARGUMENT concerns argumentation and argument making and SCIENTIFIC-THEORY dwells on the notion of scientific theory, but also the question of realism. Several topics about probability and confirmation are found in cluster C. Top-terms of PROBABILITY include "probability", "measure" or "chance" while CONFIRMATION is characterized by "law", "hypothesis", "evidence", "inductive" or "confirmation". A third topic is that of EXPERIMENT, especially correlated with PROBABILITY, and revealed by top-terms such as "datum", "experiment" or "test". Cluster D is constituted by the sole topic AGENT- DECISION, whose signature are terms such as "agent", "action", "decision", "game" or "choice" (hence quite naturally correlated with PROBABILITY).

A larger cluster is cluster E that includes topics about philosophy of mind and philosophy of biology: MIND, NEUROSCIENCES and PERCEPTION feature keywords such "behavior", "state", "mental", "perception" but also "process", "cognitive" or "information" and "representation". On the other hand, EVOLUTION revolves around terms such as "selection", "organism" or "species" which are typical concepts targeted in the philosophy of biology. Cluster F includes topics PROPERTY, CAUSATION and EXPLANATION. While PROPERTY may refer to ontological research themes (through keywords such as "property", "world" or "object"), CAUSATION is clearly about causation, as

Table 7 List of all 25 topics with top-10 keywords (topic labels include cluster letter, topic name and topic ID)

Topic	Top-10 words
A- Formal (4)	set; function; relation; define; definition; structure; order; model; theory; class
A- Language (17)	language; sentence; term; meaning; concept; use; statement; logical; mean; word
A- Mathematical (15)	mathematical; mathematics; number; proof; axiom; geometry; theory; object; point; line
A- Sentence (7)	sentence; context; use; say; reference; content; name; true; semantic; speaker
A- Truth (23)	logic; truth; true; proposition; sentence; logical; formula; follow; rule; world
B- Arguments (22)	argument; claim; say; question; make; view; reason; fact; case; point
B- Knowledge (21)	belief; knowledge; epistemic; believe; know; case; evidence; reason; justification; true
B- Scientific- theory (1)	theory; scientific; theoretical; empirical; realism; realist; truth; science; true; claim
C- Confirmation (20)	law; hypothesis; statement; evidence; theory; condition; inductive; problem; confirmation; fact
C- Experiment (12)	datum; experiment; value; use; test; result; experimental; model; hypothesis; method
C- Probability (9)	probability; measure; value; give; chance; case; function; distribution; degree; frequency
D- Agent- decision (8)	agent; action; decision; game; choice; act; utility; strategy; moral; preference
E- Evolution (5)	selection; population; organism; evolutionary; gene; biological; individual; group; evolution; specie
E- Mind (11)	behavior; state; mental; action; psychological; human; function; psychology; person; child
E- Neurosciences (13)	system; information; process; cognitive; level; mechanism; state; representation; structure; function
E- Perception (10)	object; experience; perception; see; color; perceptual; visual; content; red; image
F- Causation (19)	causal; cause; event; effect; causation; condition; case; variable; time; occur
F- Explanation (16)	model; explanation; explain; account; explanatory; phenomenon; use; case; system; provide
F- Property (2)	property; world; object; physical; relation; kind; entity; part; identity; exist
G- Particles (3)	theory; energy; law; particle; electron; atom; physical; physic; chemical; system
G- Quantum (14)	time; state; space; quantum; system; theory; particle; physical; field; point

Table 7 (continued)

Topic	Top-10 words
H- CLASSICS (24)	motion; body; force; newton; law; galileo; earth; move; light; time
H- HISTORY (0)	work; time; man; history; new; year; make; life; century; write
H- PHILOSOPHY (6)	world; nature; knowledge; concept; experience; kant; sense; thing; idea; reality
H- SOCIAL (18)	science; scientific; social; research; scientist; philosophy; knowledge; problem; history; practice

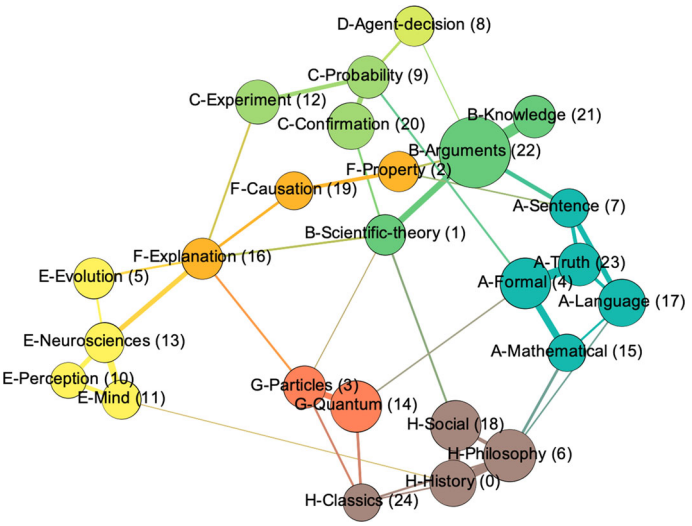


Fig. 3 Correlation graph of the 25 topics grouped into 8 clusters (nodes represent topics; color based on thematic cluster; size proportional to topic probability in the complete corpus; thickness of edges indicate topic correlation within documents; node labels include cluster letter, topic name and topic ID); visualization: Gephi (Bastian et al., 2009), with Multigravity ForceAtlas for layout rendering)

indicated by such keywords as "cause", "event" and "effect", and EXPLANATION concerns research themes about scientific explanation and models ("model"; "explanation", "account"). Note how EXPLANATION is correlated with NEUROSCIENCES, possibly though the notion of mechanistic explanation (with "mechanism" in NEUROSCIENCES and "explanation" in EXPLANATION). Cluster G is more about philosophy of physics, with the topic PARTICLES ("energy", "electron", "chemical") more centred on philosophical questions about thermodynamics, particle physics or chemistry, and the topic QUANTUM ("time", "state", "space", "quantum") about modern physics, including relativity and quantum mechanics. Finally, cluster H appears as a broader cluster of topics that are more of a historical and social nature. CLASSICS includes terms that denote both classical physics and the history of science, with key-terms

such as "motion", "force", "newton", "galileo" or "earth" (Note that this topic is also correlated with philosophy of physics topics: PARTICLES and QUANTUM); PHILOSOPHY, with top-terms such as "world", "nature", "concept", "experience" or "kant" could denote mainstream philosophy themes (some of which might also be construed as history of philosophy). The two other topics are HISTORY and SOCIAL. While HISTORY is somehow delicate to interpret due to the presence of ubiquitous key-terms such as "work", "time", "man" or "history", SOCIAL appears to capture themes that relate to the social dimension of science (with top-terms such as "science", "social", "research" or "practice").

4.5 Topic similarity with previous modeling

To assess the extent to which the addition of the translated sub-corpora affected the English-only topic model, we measured the pair-wise distance between topics (as probability distribution vectors over words) of the complete model with those of the English-only model (denoted by a †). As can be seen on Fig. 4, the results of this "inter-model topic similarity test" show that the topics of both models are very much aligned: there is clear matching between 23 of the 25 topics of the complete model with topics of the English-only model. The two problematic cases concern topic GAME-THEORY† (which appears closer to EXPERIMENT than any other topic of the new model, while EXPERIMENT better maps to CONFIRMATION† compared to any other topic of the previous model) and TIME† (which maps slightly better to CAUSATION compared to QUANTUM, though it had previously shown to be more correlated to physics topics). Yet there are differences in the topic models, and the matching is not bijective. In particular, this shows in slight differences in topic distribution into thematic clusters. For instance, the two topics CONFIRMATION† and PROBABILITY† of cluster C† appear to have been reshuffled into three topics: CONFIRMATION, PROBABILITY and EXPERIMENT. Similarly, the three topics EVOLUTION†, MIND† and NEUROSCIENCES† (cluster E†) have been somehow reorganized and replaced by four topics: EVOLUTION, MIND, NEUROSCIENCES and PERCEPTION (cluster E); in particular, MIND† appears to have been split into MIND and PERCEPTION. A comparable change concerns the historical-social topics which increased from 3 in cluster H† to 4 in H. On the other hand, some topics in the previous models seem to have somehow shrunk in the new model. This is notably the case for AGENT-DECISION† and GAME-THEORY† that could best be matched only to AGENT-DECISION (though, as noted above, the matching is not unambiguous), but also, and more significantly, for topics that concern the philosophy of physics, which decreased from 4 in the previous model to 2 in the new model: in particular, QUANTUM seems to have absorbed both QUANTUM MECHANICS† and RELATIVITY†, and somehow a significant portion of TIME†, the other going to CAUSATION as discussed above. These changes in topic models give an indication that the added translated sub-corpora are likely heavier on topics about confirmation and experiment, mind and perception, and topics of a historical-social nature than the English-only corpus. Conversely, they are less concerned with philosophy of physics topics.

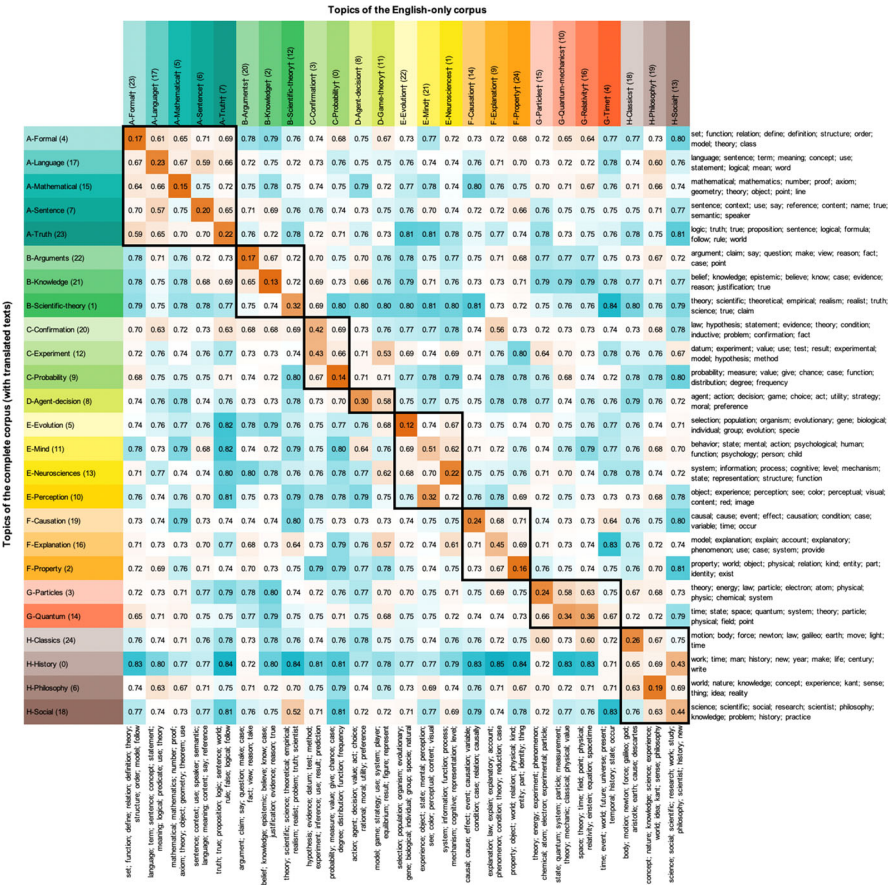


Fig. 4 Topic matching between full-corpus model and English-corpus model (Rows correspond to topics of the English-corpus model, sorted by category as in (Malaterre et al., 2020), with top-10 words; lines correspond to topics of the full-corpus model, sorted by category based on matching, with top-10 words; values are the Hellinger distances between topics considered as word-probability vectors)

4.6 Topic evolution from 1931 till 2017 and journal profiles

Considering article publication years, topic probabilities can be averaged per time-period. This simple calculation makes it possible to generate a diachronic picture of topical evolution over time in the complete corpus (Fig. 5). Perhaps one of the most striking features is the relative dominance of historical-social topics from the 1930s throughout the 1950s and then their steady decrease (topics of Cluster H, top-portion of Fig. 5). This is especially true of topics PHILOSOPHY and HISTORY which, before the 1950s, denote a special style of writing philosophy papers that tended to disappear after the 1950s (see sample papers in Fig. 6c). The professionalization of the discipline of the philosophy of science in the 1960s is likely to be an explanatory factor, but also the editorial choices made by certain journals such as *Philosophy of Science* (Howard,

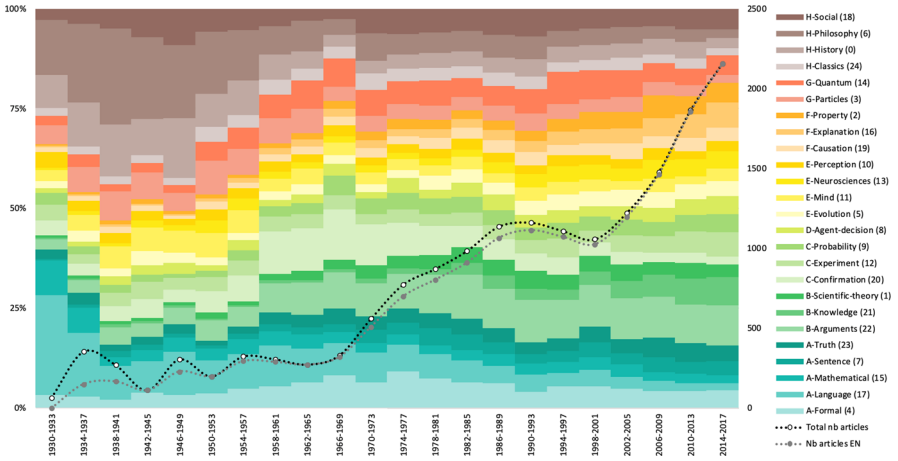


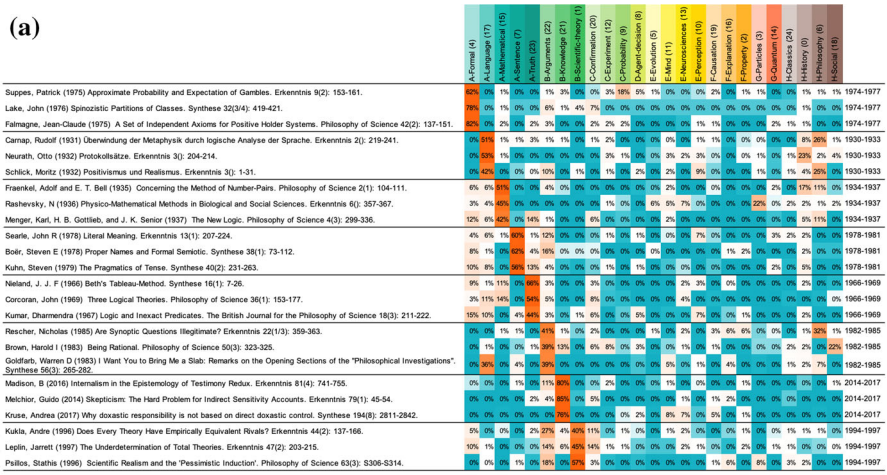
Fig. 5 Evolution of topic probability between 1930 and 2017 (probabilities represented by colored bars, left-side *y* axis; number of articles represented by dotted lines, right-side *y* axis; 4-year time-periods on the *x* axis)

2003; Reisch, 2005 Ch. 5). Another marked trend is the decrease of logic and especially philosophy of language topics throughout the 1930s, followed by an increase of these topics up until the 1970s, and then a slow but rather regular decrease up until now (topics of cluster A, bottom portion of Fig. 5). This trend likely mirrors the relative significance of logical empiricism in the early days of the philosophy of science, and the interests of neo-positivists later on (Hardcastle & Richardson, 2003).[15]

Topics that populate the middle section of the graph have generally tended to increase over time, but there are notable differences among them. For instance, topics that concern the philosophy of physics (cluster G) have remained relatively steady over time, though one notices a recent decrease since the 2000s. But note how PARTICLES was generally more significant before the 1970s, while it is the other way around for QUANTUM, showing a shift in interest away from particle physics and thermodynamics to quantum mechanics (see sample texts in Fig. 6c). Topics of cluster H have slightly increased a first time from the 1950s up until the 1970s, and again from the 1980s onward, starting with CAUSATION, and closely followed by PROPERTY (which notably depicts debates on realism) and EXPLANATION (likely due to an increased interest in the explanatory role of models—see sample texts in Fig. 6c). Among the topics of cluster E, that relate to the philosophy of biology and the philosophy of mind, there is a noticeable presence of EVOLUTION in the 1940s (especially denoting work on life and organization), then a more marked increase in the 1980s onward (this time typically with research on the theory of evolution by natural selection). The three topics of philosophy of mind exhibit slightly different diachronic patterns. While PERCEPTION appears relatively constant, MIND was relatively more present before the 1950s (covering research themes about mind in general and psychology), while NEUROSCIENCES

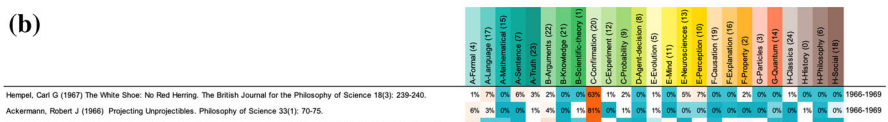
[15] On the more general role of logic in philosophy, see (Bonino et al., 2020) for a quantitative corpus-based approach. (Noichl, 2019) provides an all-encompassing view of field of philosophy, also based on quantitative approaches.

(a)



Suppes, Patrick (1975) Approximate Probability and Expectation of Gambles. Erkenntnis 9(2): 153-161.	82%	0%	1%	0%	0%	1%	3%	0%	0%	3%	16%	5%	1%	0%	0%	0%	2%	0%	1%	1%	0%	0%	1%	1%	1%	1974-1977
Lake, John (1976) Spinozistic Partitions of Classes. Synthese 32(3/4): 419-421.	78%	0%	1%	0%	0%	6%	1%	4%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	1974-1977
Falmagne, Jean-Claude (1975) A Set of Independent Axioms for Positive Holder Systems. Philosophy of Science 42(2): 137-151.	82%	0%	2%	0%	2%	3%	0%	0%	2%	2%	2%	0%	0%	0%	0%	1%	1%	0%	0%	0%	2%	1%	1%	0%	0%	1974-1977
Carnap, Rudolf (1931) Überwindung der Metaphysik durch logische Analyse der Sprache. Erkenntnis 2(): 219-241.	0%	51%	1%	1%	3%	1%	1%	0%	1%	2%	0%	0%	0%	0%	0%	1%	1%	0%	0%	0%	0%	0%	8%	26%	1%	1930-1933
Neurath, Otto (1932) Protokollsätze. Erkenntnis 3(): 204-214.	0%	53%	1%	0%	0%	0%	0%	0%	0%	3%	1%	0%	0%	3%	2%	3%	0%	0%	0%	1%	1%	0%	23%	2%	4%	1930-1933
Schlick, Moritz (1932) Positivismus und Realismus. Erkenntnis 3(): 1-31.	0%	42%	0%	0%	0%	10%	0%	1%	0%	1%	0%	2%	0%	2%	0%	9%	0%	0%	0%	1%	0%	1%	4%	26%	0%	1930-1933
Fraenkel, Adolf and E. T. Bell (1935) Concerning the Method of Number-Pairs. Philosophy of Science 2(1): 104-111.	6%	6%	81%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%	2%	0%	17%	11%	0%	1934-1937
Rashevsky, N (1936) Physico-Mathematical Methods in Biological and Social Sciences. Erkenntnis 6(): 357-367.	3%	4%	45%	0%	0%	0%	0%	0%	0%	1%	0%	0%	6%	5%	7%	0%	0%	0%	0%	0%	22%	0%	2%	2%	1%	1934-1937
Menger, Karl, H. B. Gottlieb, and J. K. Senior (1937) The New Logic. Philosophy of Science 4(3): 299-336.	12%	6%	42%	0%	14%	1%	0%	0%	6%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%	5%	11%	0%	1934-1937
Searle, John R (1978) Literal Meaning. Erkenntnis 13(1): 207-224.	4%	6%	1%	60%	1%	12%	0%	0%	0%	1%	0%	1%	0%	0%	0%	7%	0%	0%	0%	0%	3%	2%	2%	0%	0%	1978-1981
Boër, Steven E (1978) Proper Names and Formal Semiotic. Synthese 38(1): 73-112.	8%	1%	0%	62%	4%	16%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%	0%	0%	0%	2%	0%	1978-1981
Kuhn, Steven (1979) The Pragmatics of Tense. Synthese 40(2): 231-263.	10%	8%	0%	56%	13%	4%	0%	0%	0%	1%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	5%	0%	0%	1978-1981
Nieland, J. J. F (1966) Beth's Tableau-Method. Synthese 16(1): 7-26.	9%	1%	11%	0%	66%	3%	0%	0%	3%	0%	0%	0%	0%	2%	3%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	1966-1969
Corcoran, John (1969) Three Logical Theories. Philosophy of Science 36(1): 153-177.	3%	11%	14%	0%	54%	5%	0%	0%	8%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1966-1969
Kumar, Dharmendra (1967) Logic and Inexact Predicates. The British Journal for the Philosophy of Science 18(3): 211-222.	15%	10%	0%	4%	44%	3%	0%	1%	6%	0%	0%	5%	0%	0%	0%	7%	0%	0%	1%	0%	0%	0%	2%	0%	0%	1966-1969
Rescher, Nicholas (1985) Are Synoptic Questions Illegitimate? Erkenntnis 22(1/3): 359-363.	0%	0%	1%	1%	0%	41%	1%	0%	2%	0%	0%	0%	1%	0%	2%	0%	3%	6%	6%	0%	0%	0%	1%	32%	1%	1982-1985
Brown, Harold I (1983) Being Rational. Philosophy of Science 50(3): 323-325.	0%	1%	0%	0%	1%	39%	13%	0%	6%	8%	0%	3%	0%	0%	1%	0%	1%	0%	0%	0%	0%	2%	2%	0%	22%	1982-1985
Goldfarb, Warren D (1983) I Want You to Bring Me a Slab: Remarks on the Opening Sections of the "Philosophical Investigations". Synthese 56(3): 265-282.	0%	36%	0%	4%	0%	39%	0%	0%	0%	0%	0%	0%	0%	5%	0%	1%	0%	4%	0%	0%	0%	1%	2%	7%	0%	1982-1985
Madison, B (2016) Internalism in the Epistemology of Testimony Redux. Erkenntnis 81(4): 741-755.	0%	0%	0%	1%	0%	11%	80%	0%	0%	2%	0%	0%	0%	1%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	1%	2014-2017
Melchior, Guido (2014) Skepticism: The Hard Problem for Indirect Sensitivity Accounts. Erkenntnis 79(1): 45-54.	0%	0%	0%	0%	0%	2%	4%	85%	0%	5%	0%	0%	0%	0%	0%	1%	0%	0%	1%	2%	0%	0%	0%	0%	0%	2014-2017
Kruse, Andrea (2017) Why doxastic responsibility is not based on direct doxastic control. Synthese 194(8): 2811-2842.	0%	0%	0%	0%	0%	0%	76%	0%	0%	0%	0%	2%	0%	8%	7%	0%	5%	0%	1%	0%	0%	0%	0%	0%	0%	2014-2017
Kukla, Andre (1996) Does Every Theory Have Empirically Equivalent Rivals? Erkenntnis 44(2): 137-166.	5%	0%	0%	0%	2%	27%	4%	40%	11%	0%	1%	0%	0%	1%	2%	0%	0%	0%	1%	0%	0%	1%	1%	1%	0%	1994-1997
Leplin, Jarrett (1997) The Underdetermination of Total Theories. Erkenntnis 47(2): 203-215.	10%	1%	0%	0%	0%	14%	6%	45%	14%	1%	1%	0%	0%	0%	2%	1%	0%	0%	2%	0%	1%	1%	0%	0%	0%	1994-1997
Psillos, Stathis (1996) Scientific Realism and the 'Pessimistic Induction'. Philosophy of Science 63(3): S306-S314.	0%	0%	0%	1%	0%	18%	0%	57%	3%	0%	0%	0%	0%	0%	0%	0%	1%	6%	0%	8%	0%	3%	2%	0%	0%	1994-1997

(b)



Hempel, Carl G (1967) The White Shoe: No Red Herring. The British Journal for the Philosophy of Science 18(3): 239-240.	1%	7%	0%	6%	3%	2%	0%	0%	63%	1%	2%	0%	1%	0%	5%	7%	0%	0%	2%	0%	0%	1%	0%	0%	0%	1966-1969
Ackermann, Robert J (1966) Projecting Unprojectibles. Philosophy of Science 33(1): 70-75.	6%	3%	0%	0%	1%	4%	0%	1%	81%	0%	1%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	1966-1969
Baumer, William H (1968) Confirmation Still without Paradoxes. The British Journal for the Philosophy of Science 19(1): 57-63.	6%	0%	0%	0%	1%	10%	0%	0%	75%	1%	0%	0%	0%	2%	0%	1%	0%	1%	1%	0%	1%	0%	0%	0%	0%	1966-1969
Mayo, Deborah G and Aris Spanos (2006) Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. The British Journal for the Philosophy of Science 57(2): 323-357.	1%	0%	0%	0%	0%	11%	0%	0%	8%	84%	8%	0%	0%	2%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	4%	2006-2009
Sprenger, Jan (2009) Evidence and Experimental Design in Sequential Trials. Philosophy of Science 76(2): 637-649.	0%	0%	0%	0%	0%	16%	0%	0%	1%	53%	17%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	2006-2009
Ward, Andrew and Pamela Jo Johnson (2008) Addressing Confounding Errors When Using Non-Experimental, Observational Data to Make Causal Claims. Synthese 163(3): 419-432.	1%	0%	0%	0%	1%	2%	1%	0%	0%	70%	7%	0%	0%	0%	0%	0%	14%	0%	0%	0%	0%	0%	1%	0%	2%	2006-2009
Van Fraassen, Bas C (1981) A Problem for Relative Information Minimizers in Probability Kinematics. The British Journal for the Philosophy of Science 32(4): 375-379.	8%	0%	0%	0%	0%	7%	0%	0%	5%	0%	58%	1%	1%	0%	1%	8%	0%	0%	0%	0%	2%	0%	9%	0%	0%	1978-1981
Williams, P. M (1980) Bayesian Conditionalisation and the Principle of Minimum Information. The British Journal for the Philosophy of Science 31(2): 131-144.	15%	2%	1%	0%	0%	6%	0%	0%	4%	1%	58%	0%	0%	0%	0%	0%	0%	1%	0%	0%	2%	0%	0%	0%	0%	1978-1981
Armendt, Brad (1980) Is There a Dutch Book Argument for Probability Kinematics? Philosophy of Science 47(4): 583-588.	7%	0%	0%	0%	0%	2%	4%	0%	0%	0%	82%	0%	0%	0%	1%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	1978-1981
Wagner, Elliott O (2015) Conventional Semantic Meaning in Signalling Games with Conflicting Interests. The British Journal for the Philosophy of Science 66(4): 751-773.	0%	5%	0%	4%	0%	0%	0%	0%	0%	2%	6%	80%	8%	1%	6%	0%	0%	2%	0%	0%	2%	0%	0%	0%	0%	2014-2017
De Jaegher, Kris and Robert Rooij (2014) Game-Theoretic Pragmatics Under Conflicting and Common Interests. Erkenntnis 79(Supplement 4): 769-820.	0%	1%	0%	7%	2%	2%	0%	0%	0%	6%	4%	75%	1%	0%	0%	0%	1%	0%	0%	0%	0%	1%	0%	0%	0%	2014-2017
Vanderschraaf, Peter (2016) In a Weakly Dominated Strategy Is Strength: Evolution of Optimality in Stag Hunt Augmented with a Punishment Option. Philosophy of Science 83(1): 29-59.	6%	0%	0%	0%	0%	2%	0%	0%	0%	1%	0%	76%	8%	0%	0%	0%	0%	0%	0%	3%	3%	0%	0%	0%	0%	2014-2017
Dunlop, W. R (1944) Organization. Philosophy of Science 11(2): 171-177.	1%	3%	0%	1%	0%	0%	0%	0%	0%	0%	1%	1%	30%	6%	7%	0%	0%	0%	0%	14%	0%	1%	8%	18%	9%	1942-1945
Beatty, John (1984) Chance and Natural Selection. Philosophy of Science 51(2): 183-211.	0%	0%	1%	0%	0%	13%	0%	1%	0%	0%	10%	0%	60%	0%	0%	2%	2%	0%	1%	0%	2%	1%	2%	0%	0%	1982-1985
Thompson, Paul (1983) Tempo and Mode in Evolution: Punctuated Equilibria and the Modern Synthetic Theory. Philosophy of Science 50(3): 432-452.	0%	0%	0%	0%	0%	6%	0%	7%	1%	0%	0%	0%	78%	0%	0%	0%	2%	3%	0%	0%	0%	2%	0%	0%	0%	1982-1985
Leverett, Hollis M (1948) The Mind We Measure and Its Dimensions. Philosophy of Science 15(1): 39-46.	8%	5%	1%	0%	0%	1%	0%	1%	0%	19%	0%	0%	3%	30%	0%	0%	0%	0%	0%	0%	2%	0%	0%	28%	1%	1946-1949
McGill, V. J and Livingston Welch (1946) A Behaviorist Analysis of Emotions. Philosophy of Science 13(2): 100-122.	3%	0%	0%	0%	0%	2%	0%	0%	0%	1%	0%	7%	2%	56%	0%	2%	1%	0%	0%	0%	0%	0%	12%	0%	1%	1946-1949
Tatarkiewicz, Wladyslaw (1949) Psychological Hedonism. Synthese 8(8/9): 409-425.	1%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	10%	0%	30%	0%	0%	0%	0%	0%	0%	0%	0%	28%	25%	0%	1946-1949
Horgan, Terence and John Tienson (1994) A Nonclassical Framework for Cognitive Science. Synthese 101(3): 305-345.	4%	3%	3%	2%	0%	6%	3%	0%	7%	0%	0%	0%	0%	0%	40%	0%	0%	0%	3%	0%	9%	0%	0%	0%	0%	1994-1997
McLaughlin, B. P and T. A. Warfield (1994) The Allure of Connectionism Reexamined. Synthese 101(3): 365-400.	3%	1%	0%	0%	2%	15%	0%	0%	0%	8%	0%	2%	0%	1%	59%	1%	0%	3%	0%	0%	2%	0%	3%	0%	0%	1994-1997
Bub, Jeffrey (1994) Testing Models of Cognition through the Analysis of Brain-Damaged Performance. The British Journal for the Philosophy of Science 45(3): 837-855.	16%	2%	0%	0%	1%	2%	1%	0%	6%	6%	0%	0%	0%	7%	59%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1994-1997
Lowe, E. J (1986) The Topology of Visual Appearance. Erkenntnis 25(3): 271-274.	1%	1%	1%	0%	1%	14%	1%	0%	1%	1%	2%	1%	0%	0%	0%	61%	0%	0%	2%	0%	10%	1%	0%	1%	0%	1986-1989
Clark, Austen (1989) The Particulate Instantiation of Homogeneous Pink. Synthese 80(2): 277-304.	10%	0%	0%	0%	0%	5%	1%	1%	1%	2%	0%	0%	0%	0%	7%	62%	0%	0%	6%	1%	1%	1%	0%	0%	0%	1986-1989
Strawson, Galen (1989) Red and 'Red'. Synthese 78(2): 193-232.	0%	15%	0%	3%	0%	18%	0%	0%	0%	0%	0%	0%	1%	1%	0%	58%	0%	0%	2%	0%	0%	0%	0%	1%	0%	1986-1989

Fig. 6 Text samples with topic probabilities, **a** categories A and B, **b** categories C, D and E, **c** categories F, G and H

became more dominant in the 1990s (notably focusing on cognition and the neural system of the brain). As for the sole topic of cluster D, AGENT-DECISION, its relative probability in the corpus has slightly fluctuated over the years, yet shows a steady increase since the late 1990s (with a special focus on game-theory and evolutionary stable strategies).

Among the three topics of Cluster D, one of the most striking patterns is that of CONFIRMATION which peaked in the 1960s–1970s and then decreased quite drastically. Research themes typically include questions on the problem of induction and confirmation paradoxes (see sample texts in Fig. 6b). On the other hand, EXPERIMENT fluctuated all throughout the period, but increased more markedly since the 2000s (for

(c)

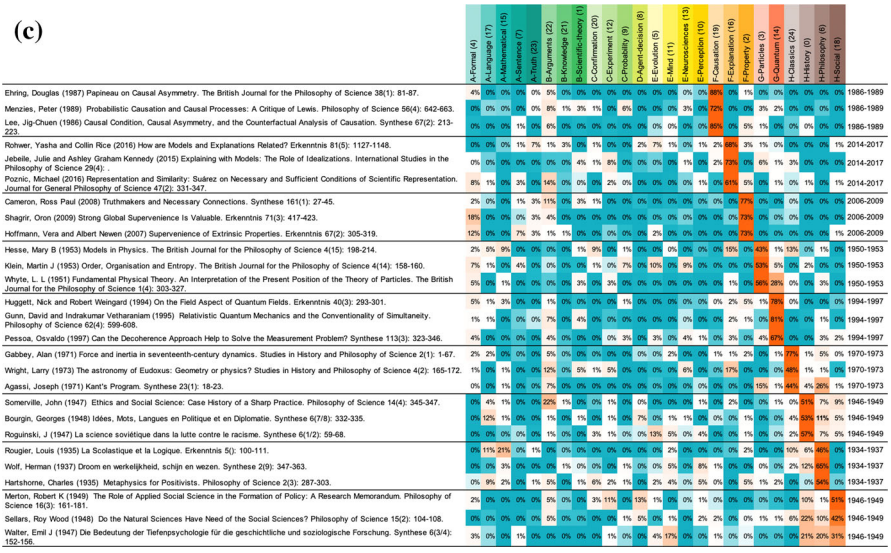


Fig. 6 continued

instance with works on experiment, testing and evidence). In the meantime, PROBABILITY remained relatively constant since the 1960s, notably with such developing themes as Bayesianism or the Dutch-book argument. Finally, the epistemology-related cluster (cluster B) appears to have steadily increased in probability all throughout the eight decades of the corpus. ARGUMENTS is the most probable topic overall, denoting research on rationality and argumentation (as can be seen from some of the sample texts of Fig. 6a), but probably also due to the high level of generality of some its terms. THEORY increased in probability especially throughout the 1980s–1990s, with work on scientific theories, underdetermination or realism for instance. Finally, KNOWLEDGE, which centrally denotes research on epistemology, emerged in the 1970s and has increased ever since.

Averaging now topic probabilities per journal—instead of per time-period—results in topical journal profiles, in other words, the probability distributions of specific topics in articles of given journals over their complete publication timespan (Fig. 7). One striking feature of these profiles is the extent to which *Erkenntnis* and *Synthese* stand out with a significant share of philosophy of language, logic and epistemology topics (clusters A and B, altogether in the vicinity of 45%) while the *JGPS* and *SHPSA* are characterized by a relatively stronger representation of historical-social topics (cluster H at around 35–40%). This is coherent with the logical empiricist anchorage of *Erkenntnis*, which was founded by Carnap and Reichenbach in the 1930s and relaunched by Hempel in the 1970s, and with the specialty in logic of Hintikka who served as editor of *Synthese* for nearly 40 years (Malaterre et al., 2020). *SHPSA* on the other hand is known for its editorial policy of publishing articles of a more historical (or even sociological) nature. In between these two extremes, the four other journals tend to have more similar profiles, though slight differences persist. *PS* for instance

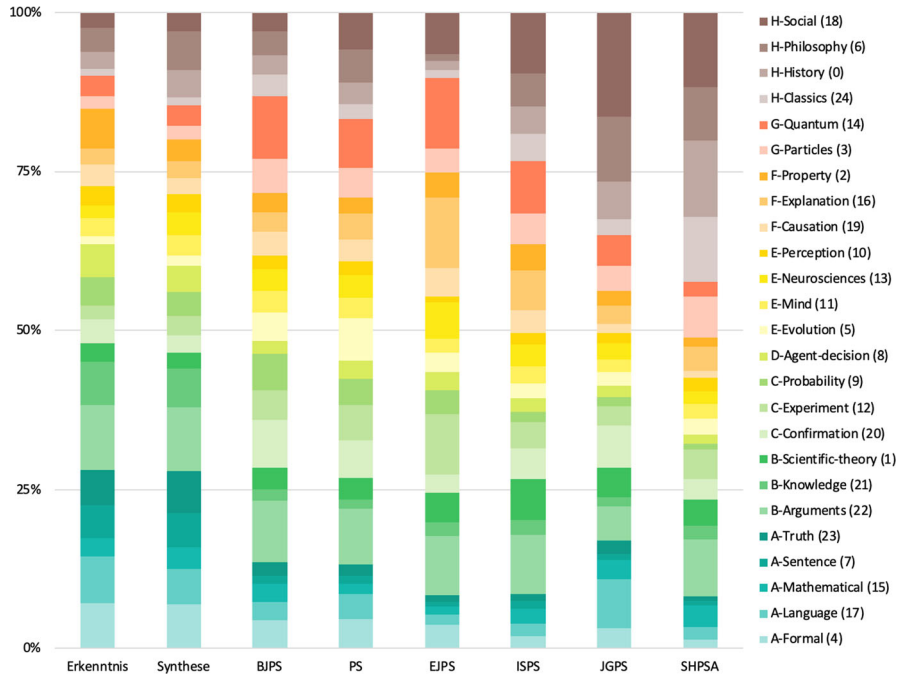


Fig. 7 Journal topical profiles (average topic probabilities per journal, sorted by cluster)

tends to have a comparatively higher share of topics that relate to the philosophy of mind and of biology. The *EJPS* appears as the journal with the smallest share of philosophy of language and logic topics (one should remember however that the journal was only founded in 2011—see Table 1 above). It also displays one of the highest shares of philosophy of physics topics (cluster G), very much like the *BJPS*. *ISPS* on the other hand appears fairly balanced, still with a comparatively high proportion of epistemology-related topics (cluster B) and historical-social topics (cluster H).

4.7 Diachronic and journal comparisons with previous modeling

As can be seen on Fig. 8, the addition of the translated sub-corpora to the English-corpus only marginally changed the diachronic picture. This could be expected given the high similarity of the topic models (see Fig. 5). Nevertheless, this makes it possible to increase the scope of the topic model, notably by adding another 4-year period and by being more exhaustive in the early days of the philosophy of science.

As mentioned in the Data section (see Table 1), most articles of the translated sub-corpora concerned *Erkenntnis* and *Synthese* in the first half of the twentieth century, and then the *JGPS* in the 1970s through the 1990s. Figure 9 focuses on the diachronic topical profiles of these three journals. For *Erkenntnis*, the addition of the translated sub-corpora mostly modified the pre-WWII period, resulting in an increased probability of philosophy of language and logic topics (cluster A), as well

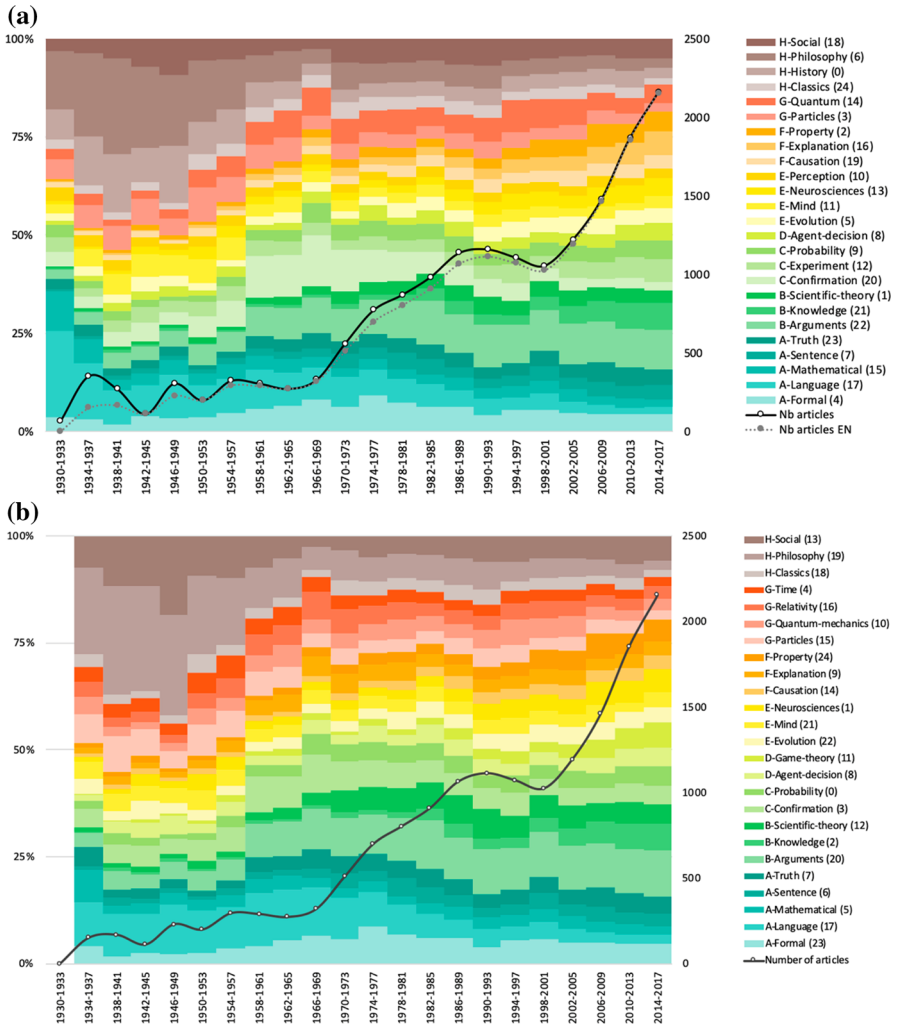


Fig. 8 Side by side comparison of the diachronic topic models obtained with the complete corpus (Panel a) and with the English-only corpus (Panel b; with permission)

as epistemology-related topics (cluster B). On the other hand, philosophy of physics topics (cluster G) slightly decreased in probability. Changes also concern the pre-WWII period for *Synthese*. The translated sub-corpora increased quite significantly the probability of historical-social topics (cluster H), while decreasing the probability of the philosophy of language and logic topics (cluster A). Contrary to *Erkenntnis*, *Synthese* was clearly not a venue for logical empiricist research work in its early days. Its profile however significantly changed in the 1950s (note that such changes cannot be attributed to Hintikka who only became editor in 1965). For the *JGPS*, the translated sub-corpora mainly concerned the 1970s up until the 1990s. Their impact can be seen

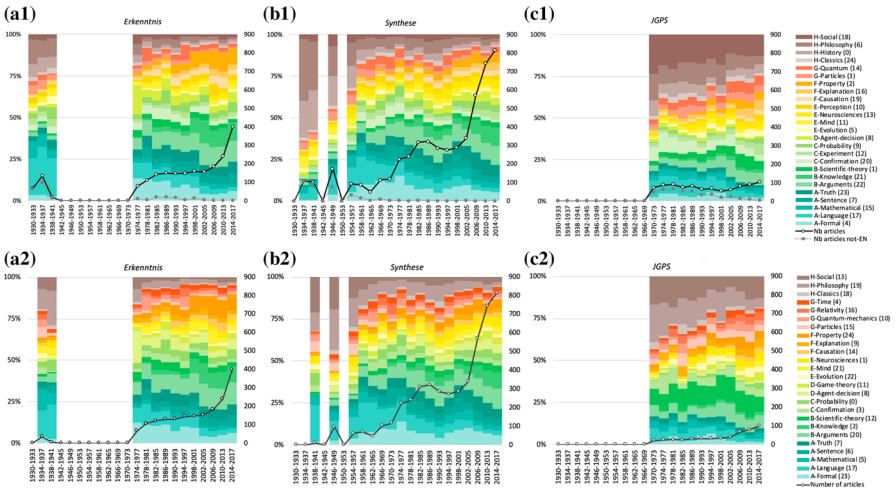


Fig. 9 Comparison of journal profiles for *Erkenntnis*, *Synthese* and the *JGPS*, resulting from the complete-corpus topic model (panels **a1**, **b1**, **c1**) and from the English-only topic model (panels **a2**, **b2**, **c2**; with permission)

especially on the historical-social topics (cluster H) whose probability increased, as well as on the philosophy of language and logic topics, which also increased slightly in probability. On the other hand, the probability of epistemology-related topics (cluster B) tended to decrease.

The addition of the non-English sub-corpora provide a more accurate picture of the research themes that pervaded the philosophy of science, especially in its early decades through the publications of *Synthese* and *Erkenntnis*. To provide more details on this period, we examined more closely the topical profiles and the authors of the three journals that existed before WWII: *Synthese*, and *Erkenntnis* in Europe and *Philosophy of Science* in the USA. More specifically, we calculated the average topic probability per author in articles published up until 1941 (Fig. 10). The results show that each journal was characterized at that time by quite unique authorship patterns. The leaning of *Erkenntnis* towards philosophy of language and logic is explained by the strong contributions of leading logical empiricists. For instance, Reichenbach, Carnap, Neurath, Schlick, Frank, Hempel—all members of the Vienna circle (Giere & Richardson, 1996; Richardson & Uebel, 2007)—are among the most significant contributors to the topic LANGUAGE, and so is Ajdukiewicz, a key figure in the Lwów–Warsaw school of logic (Woleński, 2020). Neurath's and Frank's publications also contained a significant share of the general historical and social topics HISTORY and PHILOSOPHY (unlike those of Carnap, Hempel or Ajdukiewicz). Note also how Reichenbach's articles contributed to several other topics, notably of cluster C, including CONFIRMATION, PROBABILITY and EXPERIMENT. Sample articles and their topical distributions help make sense of these findings, as shown in Fig. 11.

A radically different set of authors and their contributions explain the dominance of general historical-social topics in pre-WWII *Synthese*. Schoenmaekers and Groot for instance were prolific authors who contributed much to topics PHILOSOPHY and

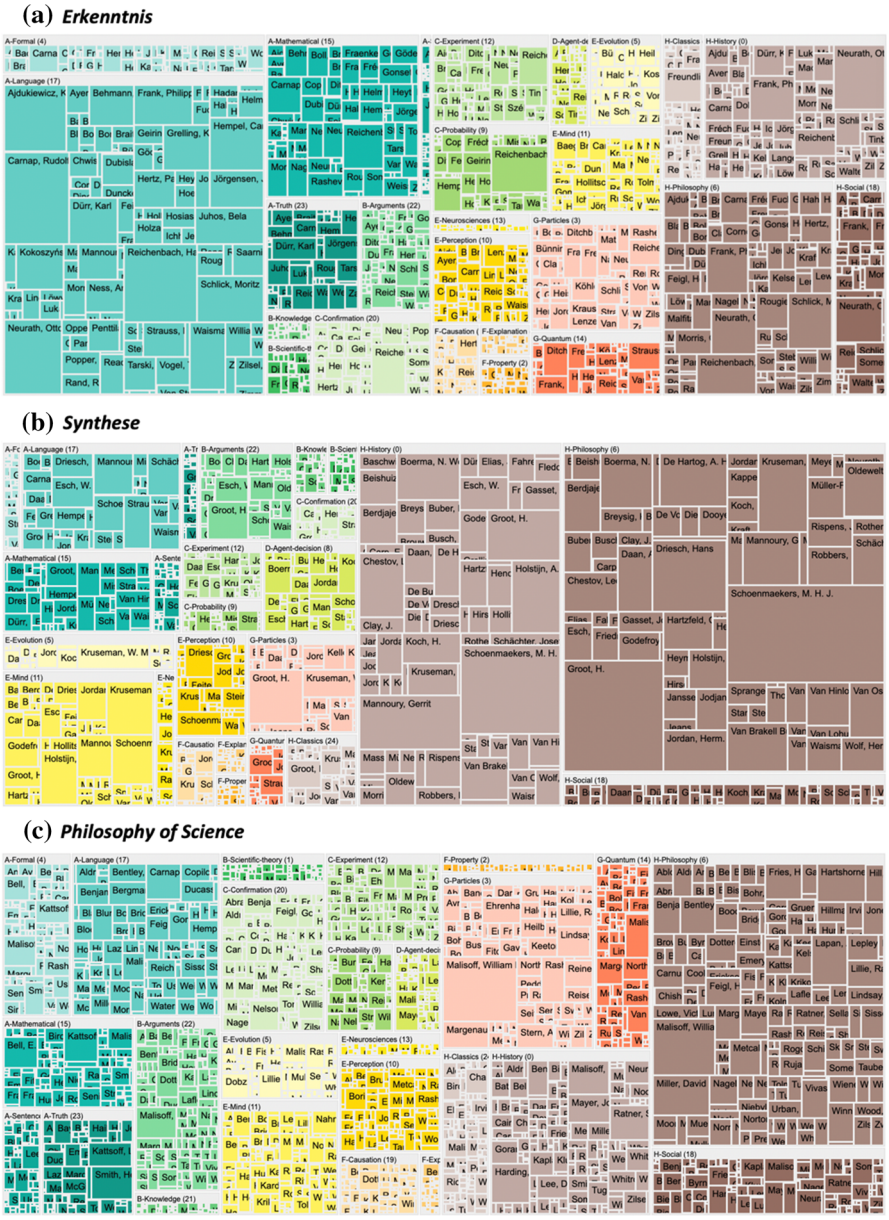


Fig. 10 Pre-WWII topic-author hierarchical graphs for **a** *Erkenntnis*, **b** *Synthese*, **c** *Philosophy of Science* (for each journal, author contributions to each topic are represented by parallelograms whose areas are proportional to the average topic probability in their pre-1941 articles; probabilities of multi-authored papers were equally divided among authors; the largest parallelograms therefore represent the strongest contributing authors to a given topic; for each topic, the sum-area of author contributions represents the overall topic probability for all pre-1941 articles in the corpus; due to graphical constraints, only authors with the highest topical contributions are readable on the graph; complete data available in the Supplementary Information files)

Fig. 11 Examples of pre-WWII articles and their topic probabilities (articles sorted by journal and by author alphabetical order; translated titles into brackets; topics sorted by category; color shading corresponding to topic probabilities in articles)

HISTORY with diverse articles on matter, thinking, cosmogony but also on beauty and sin, among others (see Fig. 11). One of the early collaborators of *Synthese*, Mathieu H. Schoenmaekers was an eccentric mathematician, member of the De Stijl (The Style) avant-garde artistic group (Johnson, 2016). He was also a member of the theosophical society (van Berkel, 2013), like the founder of the journal, Vuysje (Bloembergen, 1969). Actually, many of the early collaborators were all members of the theosophical society. This was the case of Dutch philosopher and mathematician Gerrit Mannoury whose articles also contributed much to these same topics PHILOSOPHY and HISTORY, but also to topic LANGUAGE related to philosophy of language and logic, which can be explained by Mannoury's central role in the International Society for Significs (ISS or significic circle), a Dutch counterpart of the Vienna circle (Stegeman, 1992). The German biologist and philosopher Hans Driesch also contributed to topic PHILOSOPHY, for instance with writings on the a priori, and, like Schoenmaekers, was a theosophist (Guillemain, 2013). Another biologist, Willem M. Kruseman also contributed to more biology- and mind-related topics EVOLUTION and MIND. An admirer of Driesch, he too was a member of the theosophical society and of the ISS (Bloembergen, 1969), which thereby appear to have both played a very central role in setting the pre-WWII topical content of the journal.

The third journal published at that time, *Philosophy of Science*, displays a topical profile also heavily leaning towards topics HISTORY and PHILOSOPHY, though not as

markedly as *Synthese* (see Fig. 10c). The journal also appears less dominated by a specific set of authors than both *Erkenntnis* and *Synthese*. A major figure however emerges, that of William Malisoff, founder of the journal and its first editor from 1934 till 1947 (Howard, 2003). Malisoff addressed a broad range of questions, ranging from monads to freedom, and from emergence to quantum mechanics (see Fig. 11). This explains his contributions to many different topics, notably PHILOSOPHY and HISTORY, but also PARTICLES or QUANTUM. Five other figures emerge: David Miller contributed much to PHILOSOPHY as well as Charles Hartshorne, the known philosopher of religion and metaphysician (Dombrowski, 2020); on the other hand, logicians such as Louis Katsoff or Henri Smith (Anellis, 2005) and Rudolf Carnap, who had emigrated in 1935 to the United States (Giere & Richardson, 1996) contributed to philosophy of language and logic-related topics MATHEMATICAL, TRUTH and LANGUAGE. *Philosophy of Science* contributors otherwise appear more numerous than those of *Erkenntnis* or *Synthese*, and more scattered across the full range of topics.

5 Discussion

Machine translation services offer tempting solutions for non-parallel multilingual corpora. In this paper, we investigated the possibility of assessing the quality of machine translation for the purpose of bag-of-words analyses. Both our manual examinations of the translations and the results of our semantic topology preservation test concur with (Reber, 2019; Vries et al., 2018) and provide good reasons to trust machine translation for such analyses as topic-modeling. Machine translation thereby broadens the scope of corpora that might be investigated through text-mining approaches and offers solutions to counter corpora selection biases (which have usually favored English-language corpora). This is particularly true for studies of non-parallel corpora with multiple languages.

Here we chose Google Translate services for consistency with (Vries et al., 2018), yet as (Reber, 2019) have shown, other machine translation services offer equally valid solutions, especially from the point of view of bag-of-words analyses. Following Google Translate requirements, we submitted texts by chunks of less than 30,000 characters (complete sentences). As we subsequently found out, articles sometimes used multiple languages at length (we identified 5 such texts out of the 16,917 of the corpus, in addition of numerous other articles that include short non-English citations). In such cases, the submitted chunks were also heterogeneous and the translation results left out some sentences untranslated (depending on the language detected by Google Translate). A solution could be to first run a language detection at the sentence level, then segment the texts into language-homogeneous chunks of max 30,000 characters depending on detected languages.

As we found out, machine translation is also of great help for fixing OCR-related or encoding issues which may plague older texts. In the case of our particular corpus, the translated texts appeared to have been substantially improved compared to their original versions, meaning that were textual analyses be run on the original texts, they would be far less faithful to the semantic content of these texts compared to textual analyses run on the translations. Of course, solutions already exist for automatically

correcting words in texts (e.g., Kukich, 1992; Volk, Furrer, and Sennrich 2011), yet machine translation provides this as an interesting side-benefit.

Another interesting feature of machine translation is the fact that it tends to preserve the order of words. In particular, words that are close to one another in the original text (for instance within a 5-term distance of one another, or in the same sentence) would also have their translated terms in the vicinity of one another.[16] This is also supported by the results of our semantic topology preservation test. This means that machine translated texts should not only be adequate for bag-of-word textual analyses but also for more fine-grained analyses that rely on word ordering in texts, such as collocation, co-occurrence or even sentiment analyses. At this stage however, the extent to which this can be done remains to be investigated.

Another debate is whether it could be sufficient to only machine translate the terms of the document-term matrices (DTMs) instead of the full-texts. As (Reber, 2019) showed, topic-modeling done on translated DTMs provides very good results, yet not as good as full-translations. In our case, our primary motivation was also to be able to consult the translated text in order to better interpret topics in light of their textual context. This is also a strong advantage when it comes to making sense of the diachronic evolution of topics or of their distribution according to metadata such as journal names in our case.

Non-parallel corpora raise an acute challenge when it comes to validating the outcome of machine translation, as there are no "gold-standard" reference texts (contrary to parallel corpora). To address this problem, we devised a two-step approach that relies on manual checks of randomly selected texts and deploys a "semantic topology preservation test" on the complete corpora. Our reasoning was that a proper translation should preserve the distribution of texts in their semantic vector spaces before and after translation. In other words, texts that are close to one another in their original language should also have their respective translations close to one another.[17] Our results—at least on the present corpus and in German, Dutch and French languages—indicate this to be the case. However, passing such a test is only a necessary condition for a reliable translation and is not, strictly speaking, sufficient.[18] This is where manual inspections come into play. Together with the semantic topology preservation test, they increase the level of confidence of the translations in cases where no reference translation is available. Note that conducting such a test on parallel corpora which have already been tested for joint machine translation and topic-modeling—as in (Reber, 2019; Vries et al., 2018)—could further evaluate the merits of this test.

Here, our primary motivation for including non-English texts to our philosophy of science corpus was to be able to carry out a more accurate topic-modeling, especially

[16] Of course, due to different syntactic word order rules in different languages, one should not expect the exact word order to be preserved in translations. Our point here concerns the approximate collocation of words in n-term windows as investigated for instance by co-occurrence analyses.

[17] Note that this is also the case when a term in the original language is translated by different terms in the target language depending on context. In other words, the semantic topology preservation test assesses the contextual consistency of translations.

[18] As suggested by a reviewer, we could conceive of a defective translator which would replace each properly translated term by its following term in a dictionary. This would result in an inadequate translation that would pass the semantic topology preservation test. Hence the importance of manual checks.

in the early days on the discipline (when non-English texts accounted for more than half of all publications). We limited ourselves to eight of the most central journals, but of course philosophy of science research is also published in many other journals, often more specialized ones depending on scientific disciplines or geographic regions/languages, and in numerous monographies or edited volumes. Our results should therefore be interpreted in light of this corpus-related limitation, though the representativeness of the selected journals lends confidence that the topical trends we observed indeed captured meaningful disciplinary patterns. Further research could be done by adding other philosophy of science journals to the existing corpus, notably journals that publish in non-English languages (e.g., Spanish, French, Chinese) so as to assess possible linguistic or geographic specificities in this domain.

As with many other textual analyses, topic-modeling requires parameter choices to be made, most notably the number K of topics. Here we deliberately chose $K = 25$ topics so as to facilitate comparisons with the English-only topic model of (Malaterre et al., 2020). Such a value also has the advantage of offering a fairly coarse-grained view which suits the purpose of sketching a disciplinary portrait over the course of eight decades. Obviously, finer-grained topic-models (i.e., with higher K values) would offer more details, but the choice of granularity ultimately depends on research questions (e.g., Grimmer & Stewart, 2013). One important consideration is also the interpretability of topics. Though LDA is a generative model which builds up topics from corpora, expert knowledge is needed to interpret the topics and the other results of the modeling. The use of topic modeling algorithms therefore does not free the researcher from the task of providing such an informed contextual framework.

Here we were confronted to the specific issue of comparing two topic models: before and after inclusion of the non-English texts. To this aim, we proposed to run an "inter-model topic similarity test" to assess the extent to which topics of one model matched topics of the other. This method could be used in other contexts whenever there is a need for comparing topics from different models (e.g., to assess the robustness of a topic modeling when random seeds are changed or when a fraction of documents are omitted; to investigate changes in topics when the number K of topics is changed; or to compare topics from different models fitted on specific time-slices of a given corpus or corpus subsets based on other metadata). The test could be developed further, notably by defining aggregate goodness-of-fit measures at topic, topic-cluster and model levels, possibly considering also the directionality of the fit (e.g., a single number indicating the degree to which each topic in a first model has a unique corresponding topic in a second model, and/or vice-versa).[19]

Though the addition of the translated texts did not significantly alter the synchronic topic model, it nevertheless affected the early decades of the diachronic topic-model and, somehow, the topical profiles of three journals (*Erkenntnis*, *Synthese* and the *JGPS*). To a certain extent, this could be expected: the translated texts only amounted to some 6% of the total corpus, though this share rose to 54% before WWII. Furthermore, the added texts were in journals whose English publications had already been included, thereby ensuring some content continuity of the translated texts. However, there was no warranty that the topics would actually be aligned, and this is precisely

[19] We thank an anonymous reviewer for highlighting this point.

such unalignment that the results show at the finer scale of specific temporal windows and specific journals.

While topic-modeling provides a descriptive view of the topical content of a corpus, and thereby an empirical basis for specific claims about that corpus, it cannot explain the observed facts such as diachronic shifts in topic probability. This is again an area that the researcher has to fill in with specific knowledge of the field, or that can lend itself to further investigations. In our particular case, explanations of changes in topical distribution over the years may include at least five sets of factors. Changes may be (a) researcher-driven, for instance when a researcher evaluates a research question to be of higher or lesser epistemic value, or when research agenda are chosen for other pragmatic reasons such as institutional fit or career motives; (b) journal-driven, in particular through editorial policies set by journal editorial boards (Giere, 1996; Hardcastle & Richardson, 2003), but also as a consequence of broader corporate strategies about journal portfolio management or competitive position; (c) driven by disciplinary dynamics, for instance the professionalization of the discipline (Howard, 2003), the maturing of a field, the emergence of sub-disciplines, but also sociological dynamics that result in some topics being more trendy than others at a certain point in time, like in the rest of science, or receiving or not the attention of leading scholars (Dewulf, 2020); (d) driven by extra-disciplinary dynamics in science, for instance the emergence of pressing epistemic issues raised by novel scientific developments (e.g., artificial intelligence), by new scientific disciplines (e.g., synthetic biology), or by the scientific practice at large (e.g., the replication problem in science or the development of open-source science); finally (e) driven by extra-scientific factors, including science policies, funding agencies (Vaesen & Katzav, 2019) or broader historical and sociological factors (e.g., WWII, the cold-war, brain drain etc.) (Reisch, 2005). Identifying the driving forces behind topic probability distributions goes beyond what topic-models can do. This requires further investigations and the development of theoretical frameworks specific to the questions at stake.

6 Conclusion

The results in this paper are two-fold. First, we examined the possibility of using machine translation for the purpose of bag-of-word analyses such as topic modeling. While (Vries et al., 2018) benchmarked Google Translate for topic modeling with the aid of parallel corpora, we investigated solutions to algorithmically assess the quality of machine translation in the case of non-parallel corpora. This led us to devise a "semantic topology preservation test" that assesses the extent to which machine translation preserved the relative proximity of texts in their semantic vector spaces. Together with manual examination of random texts, this procedure provides an indication of the appropriateness of the machine translated texts for bag-of-word analyses. We implemented the test on sub-corpora in three languages: German, Dutch and French. We also found out that machine translation may provide the side benefit of correcting OCR and encoding errors, thereby providing an enhanced textual content compared to the original texts. The second main facet of this paper concerns the topic modeling itself and the impact of including the translated texts. The results show that these

non-English texts did not significantly change the overall topic model though they noticeably altered the topical distribution in the pre-WWII decades, thereby giving a somehow modified view of the philosophy of science at that time, leaning much more heavily on historical and social topics than later on. The results also show changes in the topical profiles of some journals, mostly *Erkenntnis* and *Synthese* for the pre-WWII period and the *JGPS* from the 1970s till the 1990s. This opens up new questions, in particular on the specific style of pre-WWII philosophy of science, on some its main authors and on a finer-grained analysis of its research-questions. More detailed analyses of other specific topical changes could also be done, for instance by implementing more fine-grained topic models or in relationship with close-reading approaches.

Acknowledgements The authors are grateful to JSTOR, Elsevier, Oxford University Press, Springer, Taylor and Francis, and University of Chicago Press for providing access to journal articles for text-mining purposes. Special thanks are due to Martin Léonard for developing the topic-model web-browser, to Pedro Peres-Neto for providing guidance with matrix similarity measures, to Sari Lemable and Frédérick Deschênes respectively for Dutch and German translation checks, to Rens Strijbos for insights about W. M. Kruseman, and to Charles Pence and Luca Rivelli for their invitation to submit to this special issue. The authors also thank the audiences of a 2020 TEC seminar at UQAM, of the DS^2-2021 conference and of the 2021 CSHPS congress for comments on an earlier version of the manuscript. They also thank the reviewers at *Synthese* for their very valuable comments. C.M. acknowledges funding from Canada Foundation for Innovation (Grant 34555) and Canada Research Chairs (CRC-950-230795). F.L. acknowledges funding from the Fonds de recherche du Québec – Société et culture (FRQSC-276470).

Author contributions CM and FL jointly conceived the study. CM analyzed the results, wrote and revised the manuscript. FL prepared the corpus, wrote the code and revised the manuscript.

Supplementary information A technical appendix with code and data, including data for graphs is available on https://zenodo.org/record/6484582 (https://doi.org/10.5281/zenodo.6484582). The topic model can be explored on https://philscitopics.uqam.ca/

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Database Theory — ICDT 2001*, edited by Jan Van den Bussche and Victor Vianu, (Vol. 1973, pp. 420–34). Lecture Notes in Computer Science. Berlin: Springer. https://doi.org/10.1007/3-540-44503-X_27.

Anellis, I. H. (2005). Smith, Henry Bradford (1882–1938). In J. R. Shook & R. T. Hull (Eds.), *The dictionary of modern American philosophers.* Thoemmes Continuum.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *International AAAI conference on weblogs and social media*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research, 3*(March), 993–1022.

Bloembergen, S. (1969). Dr. Willem Marius Kruseman 1902–1969. *Methodology and Science: Interdisciplinary Journal for the Empirical Study of the Foundations of Science and Their Methodology* March. Retrieved from https://achterderug.nl/pageauteurs_libel.php?id=Kruseman&id2=W.M.

Bonino, G., Maffezioli, P., & Tripodi, P. (2020). Logic in analytic philosophy: A quantitative analysis. *Synthese*. https://doi.org/10.1007/s11229-020-02770-5

Boyd-Graber, J., & Blei, D. (2012). Multilingual topic models for unaligned text. https://arxiv.org/abs/1205.2657

De Smet, W., & Moens, M. F. (2009). Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the 2nd ACM workshop on Social web search and mining* (pp. 57–64).

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), 417–430. https://doi.org/10.1017/pan.2018.26

Dewulf, F.(2020). The institutional stabilization of philosophy of science and its withdrawal from social concerns after the second world war. *British Journal for the History of Philosophy* (pp. 1–19). https://doi.org/10.1080/09608788.2020.1848794.

Dombrowski, D. (2020). Charles Hartshorne. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/hartshorne/.

Escoufier, Y. (1973). Le Traitement Des Variables Vectorielles. *Biometrics, 29*(4), 751–760. https://doi.org/10.2307/2529140

François, D., Wertz, V., & Verleysen, M. (2005). Non-Euclidean metrics for similarity search in noisy datasets. In *ESANN* (Vol. 2005, pp. 339–344).

Giere, R. N. (1996). From Wissenschaftliche Philosophie to Philosophy of Science. In *Origins of Logical Empiricism*, edited by Ronald N. Giere and Alan W. Richardson, (pp. 335–354). Minnesota Studies in the Philosophy of Science, v. 16. Minneapolis: University of Minnesota Press.

Giere, R. N., & Richardson, A. W. (Eds.). (1996). *Origins of logical empiricism. Minnesota studies in the philosophy of science* (Vol. 16). University of Minnesota Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297.

Guillemain, H. (2013). *La Méthode Coué: Histoire d'une pratique de guérison au XXe siècle*. Média Diffusion.

Hardcastle, G. L., & Richardson, A. W. (Eds.). (2003). *Logical empiricism in North America. Minnesota studies in the philosophy of science* (Vol. 18). University of Minnesota Press.

Howard, D. (2003). Two left turns make a right: On the curious political career of North American philosophy of science at midcentury. In *Logical Empiricism in North America*, edited by Gary L. Hardcastle and Alan W. Richardson, (pp. 25–93). Minnesota Studies in the Philosophy of Science, v. 18. Minneapolis: University of Minnesota Press.

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning, 95*(3), 423–469. https://doi.org/10.1007/s10994-013-5413-0

Hu, Y., Zhai, K., Eidelman, V., & Boyd-Graber, J. (2014b). Polylingual tree-based topic models for translation domain adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers) (pp. 1166–1176).

Jagarlamudi, J., & Daumé, H. (2010). Extracting multilingual topics from unaligned comparable corpora. In *European Conference on Information Retrieval* (pp. 444–456). Springer.

Johnson, M. (2016). *De Stijl (1917–1932). Routledge encyclopedia of modernism* (1st ed.). London: Routledge.

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR), 24*(4), 377–439.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284.

Legendre, P., & Legendre, L. (2012). *Numerical ecology*. Elsevier.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis, 23*(2), 254–277.

Lui, M., & Baldwin, T. (2012). Langid. Py: An off-the-Shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*, (pp. 25–30).

Malaterre, C., Chartier, J.-F., & Pulizzotto, D. (2019). What is this thing called philosophy of science? A computational topic-modeling perspective, 1934–2015. *HOPOS: the Journal of the International Society for the History of Philosophy of Science, 9*(2), 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

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research, 27*(2), 209–220.

Marcus, M. P., Marcinkiewicz, M. A., & Santorin, B. (1993). Building a large annotated corpus of English: The penn treebank. *Computational Linguistics, 19*(2), 313–330. https://doi.org/10.21236/ADA273556
Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. Academic Press.
Mauri, M., Elli, T., Caviglia, G., Uboldi, G., & Azzi, M. (2017, September). RAWGraphs: a visualisation platform to create open outputs. In *Proceedings of the 12th biannual conference on Italian SIGCHI chapter* (pp. 1–5). Cagliari, Italy: ACM Press. https://doi.org/10.1145/3125571.3125585.
Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., & McCallum, A. (2009, August). Polylingual topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 880–889).*Volume 2 - EMNLP '09*, 2:880. Singapore: Association for Computational Linguistics. https://doi.org/10.3115/1699571.1699627.
Noichl, M. (2019). Modeling the structure of recent philosophy. *Synthese*. https://doi.org/10.1007/s11229-019-02390-8
Pence, C. H., & Ramsey, G. (2018). How to do digital philosophy of science. *Philosophy of Science, 85*(5), 930–941. https://doi.org/10.1086/699697
Peres-Neto, P. R., & Jackson, D. A. (2001). How well do multivariate data sets match? The advantages of a procrustean superimposition approach over the mantel test. *Oecologia, 129*(2), 169–178.
Pruss, D., Fujinuma, Y., Daughton, A. R., Paul, M. J., Arnot, B., Szafir, D. A., & Boyd-Graber, J. (2019). Zika discourse in the americas: A multilingual topic analysis of Twitter. *PLoS ONE, 14*(5), e0216922. https://doi.org/10.1371/journal.pone.0216922
Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods and Measures, 13*(2), 102–125.
Reisch, G. A. (2005). *How the cold war transformed philosophy of science: To the icy slopes of logic*. Cambridge University Press.
Richardson, A., & Uebel, T. (2007). *The Cambridge companion to logical empiricism*. Cambridge University Press.
Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research, 65*, 569–631.
Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, (pp. 44–49). Manchester.
Shuyo, N. (2010). Language Detection Library for Java. http://code.google.com/p/language-detection/.
Stegeman, J. H. (1992). *Gerrit Mannoury: A bibliography*. Tilburg University Press.
Vaesen, K., & Katzav, J. (2019). The national science foundation and philosophy of science's withdrawal from social concerns. *Studies in History and Philosophy of Science Part A, 78*(December), 73–82. https://doi.org/10.1016/j.shpsa.2019.01.001
van Berkel, K. (2001). Schoenmaekers, Mathieu Hubertus Josephus. In *Biografisch Woordenboek van Nederland V* (pp. 462-464).*Instituut voor Nederlandse Geschiedenis*. Retrieved from http://resources.huygens.knaw.nl/bwn1880-2000/lemmata/bwn5/schoenma.
Volk, M., Furrer, L., & Sennrich, R. (2011). Strategies for reducing and correcting OCR errors. In *Language Technology for Cultural Heritage*, (pp. 3–22). Springer.
Windsor, L. C., Cupit, J. G., & Windsor, A. J. (2019). Automated content analysis across six languages. *PLoS ONE, 14*(11), e0224425. https://doi.org/10.1371/journal.pone.0224425
Woleński, J. (2020). Lvov-Warsaw School. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2020. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2020/entries/lvov-warsaw/.
Yuan, M., Van Durme, B., & Ying, J. L. (2018). Multilingual anchoring: Interactive topic modeling and alignment across languages. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 11. Montréal.
Zhang, D., Mei, Q., & Zhai, C. (2010). Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (pp. 1128–1137).
Zhao, B., & Xing, E. (2007). HM-BiTAM: Bilingual topic exploration, word alignment, and translation. *Advances in Neural Information Processing Systems, 20*, 1689–1696.