UNIVERSITÀ DEGLI STUDI FIRENZE
DISEI
DIPARTIMENTO DI SCIENZE PER L'ECONOMIA E L'IMPRESA

Working Papers - Economics

Linguistic and Methodological Divergences between Journals: an Interdisciplinary Analysis with Computational Linguistics and Topic Modeling

Filippo Pietrini

Working Paper N. 07/2025

DISEI, Università degli Studi di Firenze
Via delle Pandette 9, 50127 Firenze (Italia) www.disei.unifi.it

Linguistic and Methodological Divergences between Journals: an Interdisciplinary Analysis with Computational Linguistics and Topic Modeling

Filippo Pietrini

Dipartimento di Scienze per l'economia e per l'impresa,
University of Florence, Italy
Email: filippo.pietrini@unifi.it

April 8, 2025

Abstract

This empirical study employs computational linguistics, Latent Dirichlet Allocation (LDA) and Principal component analysis (PCA) to examine the linguistic features and the main topics of three economics journals with different aims, scopes and readership, compared to a mathematics and a sociology journal. The goal is to discern linguistic, methodological and topics divergences with a focus on how these differences reflect the variety of theoretical approaches in economic research. The findings suggest significant discrepancies along these three dimensions, underscoring the potential of combining different textual analysis tools for meta-research purposes.

Keywords— *Computational Linguistics, Topic Modeling, Journal Comparison, Empirical Analysis, Economic methodology* ***JEL Classification Codes***: Z13, B2, B5

Acknowledgments

The author acknowledges the support of colleagues and the guidance of reviewers during the development of this research. Special thanks to the research team at the University of Florence for their valuable input.

Declaration of Conflicting Interest

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical Approval and Informed Consent

No ethical approval or informed consent was required for this research as it involved analysis of publicly available data.

Data Availability Statement

The data supporting the findings of this study are available upon request from the corresponding author.

1 Introduction

Textual analysis has long been a part of economic research, with seminal works by scholars like Coase (1960), who explored the legal handling of externalities, and Friedman et al. (1963), with his identification of policy shifts using historical documents. Traditionally, such analysis involved meticulous human reading, a process insurmountable to the extensive datasets now accessible. Databases containing individual documents, such as scientific journal articles, can easily reach tens of millions, prompting a growing interest in algorithmic text analysis. It is likely that this trend will endure as additional textual data surfaces. Given the novelty of text algorithms in economics, there remains considerable uncertainty about the insights we can draw from their applications. The field lacks a unified methodological approach or even a common terminology to guide modeling decisions.

This challenge is compounded by the rapid advancements in natural language processing (NLP) technologies. Since Gentzkow et al. (2019) comprehensive review of text-as-data methods in economics, the advent of deep neural network models has revolutionized NLP by enabling the detection of subtle patterns and the extraction of semantic meaning from text (Ash and Hansen (2023)). In this paper, I use two different techniques to analyse a corpus of text: a Bayesian topic model (LDA) and deterministic scoring functions from computational linguistics. Object of this empirical analysis are scientific articles of 5 different Journals, three of economics, one of sociology and one of theoretical mathematics. The three economics journals exhibit distinct ideological tendencies and are considered leaders in their respective domains. AER, a top journal, epitomizes mainstream pluralism. JEBO stands out among top journals for economists specializing in complex systems, heavily publishing article based on computational methods, experiments and behavioral economics. Conversely, CJE is a leading journal among heterodox economists, accepting articles that also employ historical, historiographical, or hermeneutic methods. The other two journals, JAM for the hard sciences and AJS for the social sciences, serve to quantitatively assess where economics stands if these fields are used as proxies. Moreover, the inter-topic distance map visualization enables a comparison of the distances separating the three streams of economic literature represented by these journals from those separating economics as a whole from both the hard and social sciences. This exercise is, unfortunately, limited by computational power to only 328 articles over a two-year period for each journal. However, we believe that selecting two years and excluding COVID-19 related content to avoid thematic uniformity makes this dataset representative of contemporary literature trends.

The aims of the article are multifaceted: 1) to verify that the language used in the three economics journals indeed aligns with the literature streams they represent, as well as with their own "about the journal" statements. 2) To discuss the points of convergence between the results obtained with the two methods and their limitations. 3) To position economics as a science, as

represented by the selected three journals. These journals certainly do not cover all streams of economics literature, but they help map the conceptual landscape of the discipline relative to fields like mathematics and sociology. The latter two fields, which could be seen as opposites, are thus suitable for creating a kind of conceptual map of science (LDAvis visualization). They are also disciplines that we would expect to find reflected in varying degrees within economics.
The results reveal a nuanced linguistic landscape where AER's (American Economic Review) use of abstractive markers aligns it closely with JAMS's one (Journal of the American Mathematical Society), followed by JEBO (Journal of Economic Behavior and Organization) and CJE (Cambridge Journal of Economics); while AJS (American Journal of Sociology) exhibits the least use of such markers. Conversely, socio-scientific markers are most prevalent in AJS, followed by CJE, JEBO, and AER, with JAMS using them the least. Additionally, all journals show a pronounced tendency towards empirical approaches, with AER leading, followed by JEBO, AJS, CJE and JAMS. The LDA model further reveals that the topics attributable to CJE and AJS overlap, as do those typical of AER and JEBO, with JAMS standing apart.

The next section is a brief review of the literature on the exploration of language in economics with both topic modelling and computational linguistics tools. Section 3 presents the data and the method, section 4 includes both the results and their discussion. Section 5 contains some concluding reflections on the contributions these kind of textual analysis tools may give to the study of HET and expands on the results.

2 Literature review

The examination of language in economic publications has a rich history, with a growing interest in the use of computational techniques to analyze and understand the evolution of economic discourse. Contributions to this field include Goldschmidt and Szmrecsanyi (2007), who conducted a corpus-linguistic analysis to explore the rhetorical strategies employed in economics journals. They found that economists tend to favor abstract and formal language, a style that distinguishes economics from other social sciences and reflects its methodological rigor.

Backhouse et al. (1997) posed the question of what economists actually do, using quantitative data to analyze the professional practices and methodological orientations of the discipline. His paper aims at two goals: first, to review existing quantitative studies on economics and the economics profession; and second, to argue that a more rigorous quantitative approach in the field will only be achievable when historians of economic thought make better use of available databases and modern computing technology. The authors support this claim by referencing their own quantitative research on British and American economics and economists, illustrating how these resources can

enhance the depth and accuracy of such analyses.

Textual analysis has been applied in economic history: Wehrheim (2019) highlighted the raising impact of quantitative tools on the study of economic history. He applies topic modeling to 2,675 articles published in the Journal of Economic History (JEH) between 1941 and 2016 in order to highlight the impact of the cliometric revolution and he argues that topic models can enhance the methodological tools available to economic historians: "For economic historians, the three main strengths of topic models are efficiency, objectivity and quantification" (Wehrheim (2019)). Cei et al. (2022) extended the application of topic modeling to agricultural economics, analyzing the development of the field through the lens of top journals in the category. Their study demonstrated how computational methods can reveal the evolving priorities and methodological shifts within a specific subfield of economics. Similarly, Suominen and Toivanen (2016) compared unsupervised learning techniques, like topic modeling, with human-assigned subject classifications, illustrating the strengths and limitations of each approach in mapping the structure of scientific knowledge.

Cherrier (2017) contributed significantly to the understanding of classification systems in economics through her analysis of the JEL codes. Her work highlights how this nomenclatures influence the way economics is practiced and understood, with implications for how heterodox approaches are integrated into or excluded from the mainstream discourse. These insights are crucial for understanding the dynamics of orthodoxy and heterodoxy in the economics discourse.

The methodological aspects of topic modeling itself have been critically assessed by scholars such as Blei et al. (2003), who introduced LDA as a powerful tool for uncovering latent structures in large text corpora. Their foundational work in the Journal of Machine Learning Research has paved the way for its widespread application in economics and other disciplines. Building on this, Sievert and Shirley (2014) developed LDAvis, a method for visualizing and interpreting the results of topic models, which has become an essential tool for researchers using topic modeling in their analyses and it is used in the present paper.

In the context of economics, the rhetorical and methodological shifts from neoclassical dominance to a more pluralistic mainstream have been explored by Davis (2006) (as well as Ambrosino et al. (2024), already mentioned in the introduction), who argued that the growing acceptance of heterodox approaches within mainstream economics reflects a broader trend towards methodological pluralism. This trend is further supported by the work of D'Orlando (2013), who examined the role of electronic resources in promoting heterodox economics, suggesting that the digital age has opened new avenues for the dissemination and acceptance of alternative economic theories.

Ambrosino et al. (2018) explore LDA to analyze the thematic structure in economics articles from the JSTOR database. By applying topic modeling, the authors construct a historical map of the discipline, revealing trends and shifts

in economic thought. The study addresses concerns about the fragmentation of mainstream economics, noting that while the field remains unified through formalization and mathematical approaches, it has become more diverse and specialized over time. The paper focuses on the potential of LDA for studying the evolving nature of economics (so, differently from the present paper it takes a diachronic perspective), especially in a time of increasing pluralism and specialization.

Edwards et al. (2018) give further encouragement to studies such as the present one. They point out that quantitative approaches are still uncommon among historians and methodologists of economics, unlike in other fields such as science studies. Edwards' article discusses whether the historiography of economics is undergoing a "quantitative turn" and if such a shift is beneficial. The authors observe a recent rise in quantitative studies in the field and argue that this trend is promising, since these methods can complement traditional analyses, offering valuable insights for both historians and methodologists of economics.

The paper by Coats (2014) provides an in-depth examination of the historiography of economics, highlighting how economic ideas are deeply influenced by the historical contexts in which they arise. This approach underscores the necessity of contextual understanding when studying the evolution of economic thought. Similarly, the work byBogenhold (2020) extends this perspective, arguing that neglecting the history of economic thought leads to an incomplete understanding of contemporary economics. He emphasizes that acknowledging historical economic ideas as essential to modern economic inquiry is crucial for comprehensive academic analysis. Furthermore,Dupont (2017) explores the continuity of economic ideas from the past to the present, suggesting that the historical context not only shapes economic theories but also informs current economic policies and debates. Düppe and Weintraub (2018) provide a contemporary analysis of how historiographical approaches in economics have evolved, showing the various strategies employed in modern economic history writing. This comprehensive understanding of the historiography of economics is essential for appreciating the development and persistence of economic ideas through time.

Card and DellaVigna (2013) empirically analyze the characteristics of top economics journals, identifying key factors that influence publication success. Their findings offer insights into the editorial practices and citation dynamics that shape the landscape of economic research, further contributing to our understanding of what constitutes orthodoxy in economic publishing. In sum, the literature reviewed provides a comprehensive foundation for the present study, highlighting the intricate relationship between language, methodology, and theory in economics. By analyzing the linguistic features and topics of economic journals, this paper seeks to contribute to the ongoing discourse on the nature of economic knowledge and its representation in academic writing.

3 Data and Method

The methodology section employs computational linguistics tools and topic modeling via Latent Dirichlet Allocation (LDA) to analyze textual data from five distinct academic journals. This study adjusts model parameters to optimize the balance between topic coherence and differentiation, ensuring robust topic extraction. Specific attention is given to selecting model hyperparameters that accommodate the diversity of the corpus while maintaining interpretative clarity. I selected all 358 articles published in the five journals between September 2018 and January 2020, for a total of 8,229,313 tokens. Although the journals vary in the number of issues published each year and in their policies on the length of scientific articles, the documents in the corpus exhibit similar dimensions among them, as reported in 1.

Table 1: *Table of Tokens, Words, and symbols in each journal*

Name	**Tokens**	**Words**	**r.f. (r.d). of symbols**
AER	1,497,466	1,049,179	24,976.86(71.52%)
AJS	1,675,239	1,173,734	7,558.32(21.64%)
CJE	1,453,344	1,018,266	12,690.04(36.34%)
JAM	1,839,933	1,289,124	86,593.91(247.96%)
JEBO	1,763,331	1,235,454	33,773.58(96.71%)

The third column presents the relative (per million of tokens) frequency of mathematical symbols, computed through the part-of-speech tagging provided by Sketch Engine (available at the following link), in specific journals. The corresponding relative densities are in brackets. The relative frequency represents the number of occurrences of mathematical symbols per million tokens within a journal; the relative density indicates whether a specific kind of token, in this case mathematical symbols, are more prevalent in a document than in the rest of the corpus (values below 100% mean lower frequency and values above 100% mean higher frequency of math symbols in that document than in the whole corpus). This information is an initial insight into the level of formalism in the language.

The analysis conducted is synchronic and concentrates on contemporary language usage by economists in three journals that, based on initial hypotheses, exhibit considerable heterogeneity in either form, content, or both. The study excludes references to COVID-19 to mitigate bias arising from the incorporation of medical health terminology and a predictable convergence in the subjects addressed.

The CJE corpus spans from Volume 42, Issue 5, September 2018, to Volume 44, Issue 1, January 2020; AER from January 2019 (Vol. 109, No. 1) to January 2020 (Vol. 110, No. 1); JEBO from Volume 157, Pages 1-792 (January 2019) to Volume 169, Pages 1-412 (January 2020); AJS from Volume 124, Number 4, January 2019, to Volume 125, Number 4, January 2020;

and JAMS from 2019-32-01 to 2020-33-02. Special issues are absent from the dataset. Limitations in computational capacity have constrained the size of the corpora, and since not all journals publish the same number of issues, I have attempted to balance the sizes of the corpora by selecting slightly different time periods so that each journal would have a similar number of tokens. Nevertheless, having included at least six issues for each journal renders the sample representative of the published content. Additionally, covering a relatively short period of time also allows for the disregarding of any long-term shifts in a journal's focus and orientation.

It is noteworthy that the methodologies employed by heterodox economists diverge significantly from those utilized by orthodox ones, encompassing a broad array of computational methods, complex systems, agent-based models, and different production, growth, and consumption models. This distinction in the economists' toolbox will not be the focus of an in-depth analysis in this study. The first category of tools employed to analyze journals comes from the field of computational linguistics. These tools are mainly scoring functions, they allow the extraction of relative frequencies from corpora and the development of scoring such as the Keyness and the typicality (or LogDice) scores for exploring the use in context of a specific lemma. Goldschmidt and Szmrecsanyi (2007), Say (2018) and Hsu et al. (2021) used similar tools and their corpus is smaller than mine: 5,597,535 tokens. These functions enable an understanding of both the lexical characteristics of a given corpus and the semantics of specific words selected by the researcher. For a comprehensive examination of these tools, please refer to the following document: PDF.

The second tool employed is topic modeling, specifically the Latent Dirichlet Allocation model. This particular generative statistical model, known as LDA, is a scalable tool widely used in machine learning and statistics. It is recognized as a dimensionality reduction method and serves as a fully probabilistic version of latent semantic analysis. When applied to an entire corpus (represented as a list of all words appearing in each document along with their frequencies), LDA identifies probabilistic patterns or recurring themes in the text by analyzing the co-occurrence of words. It operates on the 'bag-of-words' assumption (see Blei et al. (2003)), meaning that texts are treated as collections of words without regard to grammar or word order: only the frequency of occurrence matters. LDA focuses solely on the frequency and co-occurrence of individual words, based on the premise that words related to similar topics tend to appear in similar contexts. It clusters these words into different probability distributions across a fixed vocabulary. The resulting 'topics' are essentially constellations, or groups of words, that represent underlying themes within the dataset's documents. These topics are considered latent structures, inferred from the data rather than existing prior to analysis. So in principle they are not the intentions of the researcher, but only the 'frame' of her discourse. The primary goal of LDA is to detect these themes by reverse-engineering the original intentions of the document authors, who aimed to discuss one or more specific topics Mohr and Bogdanov (2013).

LDA assumes that all documents in a given corpus share a common set of topics, focusing on words with the highest estimated frequency. However, each document displays these topics in varying proportions based on the words it contains. Notably, LDA simultaneously generates topics and associates them with the documents. This model seems particularly suited, in the case of this study, to complement the deterministic analysis based on scoring functions typical of computational linguistics. In principle LDA should bring out that framework which is not directly an expression of the author's will, but rather is the framework in which the author inserts his speech, a sort of semantic structure of the speech inferred through the words.

To enhance the accuracy and efficiency of the topic modeling process, several optimizations are implemented. Stop-word removal excludes frequent yet non-informative words (the default set of NLTK stopwords). The number of topics is manually set in one run and automatically in the other (19). In the latter case it is determined dynamically using coherence scores, ensuring an optimal balance between topic granularity and interpretability. The coherence score (Newman et al. (2010)) is adopted as a robustness check since it enhances model reliability and reducEs the risk of overfitting or underfitting the data. These automations collectively lead to a more robust and interpretable topic modeling process.

4 Results

4.1 Computational linguistics

The first result comes from a measure of lexical richness or variety: the types tokens ratio (TTR). It is the ratio between the unique word forms (including non-words) of a text (each word form is counted once) and the total number of tokens, so it is a measure of the lexical variety or richness. The closer it is to 1 the higher is the lexical richness of the *corpus*. Since I am dealing with a lot different ways of writing the same word, a more reliable measure of lexical richness is the one reported in the second line of tab 1. Where the numerator instead of the number of unique word forms (including non-words as $1, \theta, x, \sum$ and so on) is the number of unique lemmas (excluding non-words).

Table 2: *Standardized Type token ratio of each journal*

	AER	**CJE**	**JEBO**	**JAMS**	**AJS**
words' TTR	4.46%	4.49%	4.29%	2.83%	5.01%
lemmas' TTR	2.82%	3.05%	2.5%	2.15%	3.02%

AJS is the richest using the first version of the TTR while CJE is the first using the second one. JAMS being a mathematics journal, it is only natural

that it is last in this respect. We will analyse this further by categorising certain words as markers of specific attitudes.

The keywords give important information on the main contents of the various journals. For each keyword, both the keyness score and, in brackets, the relative frequency per million tokens are given in Table 3 and 4.

Table 3: *Keywords and Key-scores (relative frequencies in parenthesis) for AER and CJE Journals*

AER		**CJE**	
Token	**Key-score (r.f.)**	**Token**	**Key-score (r.f.)**
Equilibrium/a	312.23 (1151.28)	Keynesian	156.2 (263.53)
Counterfactual	146.9 (190.32)	Minsky	127.9 (154.13)
Regressions	107.7 (158.94)	Financialization	124.4 (140.37)
Shocks	106.4 (530.23)	Heterodox	110.4 (138.99)
Elasticity	105.3 (329.22)	Supermultiplier	108.3 (107.34)
Lemma	95.1 (143.58)	Leontief	79.9 (81.19)
Heterogeneity	87.4 (229.05)	Sraffa	79.9 (80.50)
Exogenous	72.5 (132.89)	Neoclassical	69.2 (136.93)
Stochastic	69.3 (213.69)	Veblen	67.5 (77.75)
Endogenous	62.6 (157.60)	Hayek	56 (94.27)
Unobserved	59.2 (82.14)	Friedman	55.4 (235.32)
Long-run	56.5 (97.50)	Capitalists	53.8 (187.84)
Covariate	51.7 (76.80)	Accumulation	52.5 (402.52)
Robustness	50.9 (127.55)	Wage	50.8 (1002.52)
Individual-level	48.5 (56.09)	Marxian	43.3 (53.67)
Marginal	44.7 (327.22)	Firms	42.6 (1311.46)

Table 4: *Keywords and Key-scores (relative frequencies in parenthesis) for JEBO, JAMS, and AJS Journals*

JEBO		**JAMS**		**AJS**	
Token	**Key-score (r.f.)**	**Token**	**Key-score (r.f.)**	**Token**	**Key-score (r.f.)**
Equilibrium	170.04 (789.40)	Theorem	366.9 (1970.72)	Desegregation	103.4 (167.74)
Aversion	109.6 (310.78)	Isomorphism	352.4 (460.34)	Homophily	84.2 (87.15)
Regression	69.2 (555.77)	Symplectic	250.3 (289.68)	Covariate	76.2 (115.80)
Lemma	676.2 (109.45)	Functor	234.3 (305.99)	Segregation	76.2 (389.80)
Spillover	59.2 (96.41)	Corollary	205.3 (408.71)	Incarceration	76.61 (308.02)
Alesina	49.8 (49.91)	Automorphism	198 (243.49)	Stratification	71.9 (154.01)
Coefficient	39.6 (335.73)	Lagrangian	181.6 (251.10)	Cohesion	64.5 (237.58)
Heterogeneity	39.2 (104.92)	Equivariant	162 (167.40)	Coethnic	63.5 (62.68)
Marginal	30.7 (224.57)	Holomorphic	147.5 (162.51)	Mobilization	62.7 (255.49)
Overconfidence	29.1 (34.03)	Homomorphism	146.2 (175.01)	Conditionality	62.2 (79.39)
Underconfident	48.4 (47.64)	Bimodule	139.5 (139.68)	Transnational	55.3 (244.74)
Subsample	28.5 (32.89)	Invariant	138.6 (339.69)	Remobilization	54 (54.32)
Optimal	27.6 (400.38)	Supercuspidal	129.2 (128.27)	Attainment	51.5 (231.01)
Empirical	27.5 (263.71)	Finitely	127 (153.81)	Enclave	51.4 (164.75)
Experiment	27.2 (775.24)	Maximal	125.5 (343.49)	Repression	51 (275.18)
Experimental	24.9 (631.76)	Submanifold	115 (119.57)	Socioeconomic	47.9 (201.76)

The JAMS Keywords are, predictably, highly technical and almost incomprehensible to a non-specialist reader. Those of AJS reveal classic sociological themes such as *stratification*, but also problems of socio-spatial segregation. The keywords of the economics journals, on the other hand, immediately reveal an affinity between AER and JEBO in the use of archetypically economic words (*equilibrium*), abstractive markers (*llemma*) and econometric terms (*regression*); *heterogenity* is also shared by the two journals and denotes that the expedient of the representative agent has been overcome, systematically or otherwise, by mainstream pluralism (Ambrosino et al. (2018)). While the CJE has no keywords in common with the other two and from this initial analysis seems to favour heterodox topics in content. Multi-keywords give perhaps more interesting insights than single ones.

Table 5: *MultiKeywords and Key-scores (relative frequencies in parenthesis) for AER and CJE Journals*

AER		**CJE**	
Token	**Key-score (r.f.)**	**Token**	**Key-score (r.f.)**
Fixed effect	311.4 (357.27)	Wage share	216.1 (219.49)
Standard error	167.8 (262.57)	Standzard error	161 (173.39)
Labor supply	129.6 (108.50)	Income distribution	147.7 (207.80)
Production function	102.1 (117.53)	Capital accumulation	143.5 (174.08)
Treatment effect	92.8 (122.87)	Capacity utilization	135.1 (156.19)
Dependent variable	80.2 (122.87)	Profit share	111.5 (125.92)
Marginal cost	54.1 (75.46)	Capital stock	111.1 (147.25)

Table 6: *MultiKeywords and Key-scores (relative frequencies in parenthesis) for JEBO and AJS Journals*

JEBO		**AJS**	
Token	**Key-score (r.f.)**	**Token**	**Key-score (r.f.)**
Dependent variable	116.6 (179.21)	Labor market	92.2 (294.88)
Price path	110.4 (110.59)	Collective action	76.3 (143.86)
Social forces	64.6 (66.26)	Peer depression	75 (74.02)
Reservation wage	106.7 (107.18)	Collective action	74.7 (143.86)
Rent extraction	98.68 (98.7)	Class origin	74.1 (74.62)
Low type	95.3 (95.27)	Skin tone	66.2 (162.36)
Ethical bank	75.2 (74.86)	Union membership	56.9 (76.41)

The multi-keywords of AER are divided between econometric and archetypically economic, those of CJE confirm the distinctly heterodox content (multi-

keywords from surplus approach), with the exception of *standard error*. JEBO also has only one entry related to econometrics, while the others are all strictly economic. *Low type* refers to a type of agents. The multi-keywords of AJS are of strictly sociological nature.

In order to better explore what kind of language distinguishes the journals from each other, I classify a number of topic markers into categories, and then descriptively analyse the relative frequencies to understand more about the vocabulary used. In tab 7 I report the results in terms of both relative frequencies (RF) and relative density (RD). It measures how typical of a given document is a token compared to how typical it is of the entire corpus. As above, a value above 100% means the token is more frequent in this text type than in the whole corpus; it is typical or specific of this text type. While below 100% vice versa. Abstract markers are indicators of a deductive logical language while socio-scientific markers of a historical method that takes cultural contextuality into account. These first two categories help to better understand the form through which the journal expresses itself. The archetypical economic markers and heterodox markers, on the other hand, are proxies for the heterodoxy or orthodoxy of the content of the journals (although *computation** could denote heterodoxy in methods, we conferred at the beginning of the article that by heterodoxy in methods here we mean the way in which one means what is scientific, and not an alternative choice of logical-formal tools). Empirical markers are a residual category that gathers many words common to all journals (except JAMS) and are a proxy for their use of econometrics.

The results show that AER and JAMS make extensive use of formalism, while JEBO CJE and AJS (in descending order) have a similar number of abstractive markers. As far as socio-scientific markers are concerned, the ranking is reversed: $AJS < CJE < JEBO < AER < JAMS$. This symmetry between abstractive and socio-scientific markers confirms the goodness of the choice of cathegories. What is surprising in this data is the huge gap between AER and CJE, as if AER does not deal with social science, whereas CJE makes extensive use of *institution*, soci** and *theor**. The extensive reference to institutionalist economics in comparison to AER and JEBO is already evident from this simple marker (a confirmation is that the economist's name *Coase* only appears in CJE with a relative frequency of 4.65 per million of tokens (Williamson with 37.78). I could have explored the economic orientations of the various journals in more detail. It quickly emerges, for example, from the relative frequency of *heuristic** that JEBO is quite behavioural and, from the keywords, that it has a focus on dynamic analysis. CJE is rather institutionalist, and AER has something of the behavioural, but less than JEBO. All this, however, is not in the interest of the article; this is information that can already be found in the aims and scopes of the journals.

Empirical markers are common to all journals except JAMS. AER makes the most use of them, followed by JEBO, followed by AJS (it is known that sociology is now strongly quantitative: Diaz-Bone and Didier (2016)), CJE and JAMS. For typically economic markers we find AER first thanks to an

Table 7: *Relative frequency (RF) and relative density (RD) of category markers across different journals*

Category Markers by words, RF is the relative frequency and RD is the relative density

Category markers	AER		CJE		JEBO		JAMS		AJS	
Words	RF	RD	RF	RD	RF	RD	RF	RD	RF	RD
assum*	982.99	145.31%	540.13	79.84%	650.47	96.16%	913.62	135.05%	287.72	42.53%
proof	235.73	52.57%	24.08	5.37%	147.45	32.88%	1,645.17	366.90%	8.95	2.00%
theorem	183.64	37.52%	21.33	4.36%	54.44	11.12%	1,970.72	402.62%	-	-
lemma	148.25	30.95%	1.38	0.29%	109.45	22.85%	1,915.83	399.95%	-	-
equation*	502.18	160.37%	593.11	189.40%	96.98	30.97%	371.21	118.54%	65.07	20.78%
model*	2,519.59	164.81%	1,550.22	101.40%	1,682.61	110.06%	179.35	11.73%	1,944.80	127.21%
Sum of abstractive markers	**4572.38**		**2730.25**		**2741.40**		**6995.90**		**2306.54**	
theor*	283.81	48.24%	1,277.05	217.04%	410.59	69.78%	355.45	60.41%	706.17	120.02%
histor*	162.94	42.60%	617.20	161.34%	195.65	51.15%	8.70	2.27%	982.55	256.85%
cultur*	46.75	24.93%	197.48	105.32%	76.56	40.83%	-	-	627.37	334.60%
soci*	454.77	24.41%	2,319.48	124.48%	1,092.82	58.65%	84.79	4.55%	5,491.16	294.69%
institution*	99.50	26.30%	844.95	223.36%	273.35	72.26%	0.54	0.14%	747.95	197.72%
justice	4.01	6.67%	19.95	33.24%	28.92	48.18%	-	-	243.55	405.71%
context*	170.29	72.80%	270.41	115.60%	153.69	65.70%	52.72	22.54%	542.61	231.96%
Sum of socio-scientific markers	**1172.78**		**5525.52**		**2231.58**		**502.2**		**9341.36**	
data	1,744.28	168.93%	937.84	90.83%	981.66	95.07%	122.83	11.90%	1,531.12	148.29%
effect*	2,474.85	148.58%	1,680.95	100.92%	2,503.78	150.32%	63.59	3.82%	1,806.31	108.45%
evidence	709.20	159.50%	397.02	89.29%	710.02	159.69%	1.09	0.24%	457.25	102.84%
empirical	400.01	135.13%	452.75	152.95%	340.27	114.95%	-	-	345.62	116.76%
result	1,718.24	134.29%	997.70	77.98%	1,904.92	148.89%	684.81	53.52%	1,126.41	88.04%
observ*	1,101.19	165.85%	404.58	60.93%	1,060.49	159.72%	229.36	34.54%	558.13	84.06%
method*	304.51	120.02%	406.65	160.27%	161.63	63.70%	97.83	38.56%	343.83	135.51%
sample	831.40	159.30%	341.97	65.52%	729.30	139.74%	19.02	3.64%	735.42	140.91%
experiment*	453.43	89.70%	78.44	15.52%	1,701.33	336.56%	5.98	1.18%	212.51	42.04%
significant*	458.11	95.18%	531.88	110.50%	922.69	191.70%	10.87	2.26%	510.37	106.03%
statistic*	435.40	125.90%	326.83	94.50%	531.38	153.65%	7.07	2.04%	459.04	132.73%
Sum of empirical markers	**11630.22**		**5256.61**		**10657.87**		**1234.47**		**8117.25**	
equilibri*	1,151.95	260.93%	354.36	80.27%	784.31	177.66%	3.26	0.74%	2.39	0.54%
rational*	341.91	172.30%	185.78	93.62%	257.47	129.75%	125.00	62.99%	99.69	50.24%
optimal*	649.10	280.55%	45.41	19.63%	437.24	188.98%	44.57	19.26%	7.76	3.35%
market	1,646.78	180.12%	1,551.59	169.70%	826.84	90.44%	-	-	802.87	87.81%
price*	1,271.48	192.10%	1,003.20	151.56%	1,126.28	170.16%	1.63	0.25%	57.31	8.66%
agent*	926.23	235.76%	143.12	36.43%	805.86	205.12%	-	-	129.53	32.97%
maximiz*	171.62	215.30%	24.77	31.07%	182.61	229.08%	4.89	6.14%	19.10	23.96%
utility	509.53	269.82%	48.85	25.87%	381.66	202.11%	0.54	0.29%	27.46	14.54%
allocat*	247.75	205.73%	92.89	77.14%	210.40	174.71%	-	-	68.05	56.51%
efficien*	251.09	176.46%	108.71	76.40%	303.97	213.62%	9.78	6.88%	49.55	34.82%
Sum of arch. economic markers	**7360.33**		**3560.67**		**4959.64**		**189.67**		**1444.08**	
heterogen*	332.56	240.91%	94.95	68.79%	202.46	146.66%	-	-	85.36	61.84%
complex*	80.80	34.80%	236.01	101.63%	60.11	25.89%	593.50	255.58%	148.64	64.01%
computation*	62.77	157.97%	23.39	58.87%	15.31	38.53%	75.00	188.75%	20.30	51.08%
evolution*	80.80	83.33%	283.48	292.34%	79.96	82.46%	28.81	29.71%	42.38	43.71%
Kaleck*	1.34	3.90%	192.66	562.22%	-	-	-	-	-	-
supermultiplier	-	-	108.71	566.23%	-	-	-	-	-	-
Sraff*	-	-	108.03	566.23%	-	-	-	-	-	-
financialisation	7.16	4.47%	898.62	561.08%	-	-	-	-	-	-
crisis	309.19	177.81%	469.95	270.26%	55.58	31.96%	-	-	111.63	64.19%
instab*	14.69	30.92%	211.24	444.59%	12.48	26.26%	2.72	5.72%	20.89	43.97%
distributi*	846.10	160.21%	765.13	144.88%	617.01	116.83%	145.66	27.58%	364.72	69.06%
Marx*	-	-	432.79	523.77%	5.10	6.18%	-	-	-	-
pluralis*	0.67	9.64%	15.14	218.55%	2.84	40.94%	-	-	17.31	249.93%
Keynes*	28.72	14.61%	1,031.41	524.91%	35.73	18.18%	-	-	7.16	3.65%
green	67.45	123.89%	66.05	121.34%	56.14	103.13%	32.07	58.90%	55.51	101.97%
environment*	192.99	115.51%	198.85	119.01%	232.51	139.16%	3.26	1.95%	227.43	136.12%
feminis*	0.67	19.63%	8.84	262.89%	-	-	-	-	8.36	245.62%
Sum of heterodox markers	**2017.91**		**4383.40**		**1089.90**		**881.30**		**1379.94**	

abundant use of *equilibrium*, optimal*, agent* and *utility*, while *price* and *market* are also shared by other journals. With respect to these markers, the descending ranking is $AER < JEBO < CJE < AJS < JAMS$. The heterodox markers, on the other hand, suggest that the CJE is the distinctly more heterodox journal. The decreasing ranking is $CJE < AER < AJS < JEBO < JAMS$. AER and JEBO have a considerable number of occurrences among the heterodox markers due to 3 words: *heterogen**, which confirms the existence of mainstream pluralism and is one of the so-called 'belt' hypotheses (protective belt according to Lakatos (2014)), i.e. criticisable, attackable and substitutable, which helps the mainstream to see/report itself as open and continuously progressing. The 'protective belt' of a theory consists of auxiliary hypotheses that can be modified or adjusted in response to new evidence or criticisms (such as perfect rationality/information, the representative agent, perfect competition and other similar abstractions), thus protecting the 'hard core' from direct falsification. An example of direct falsification of the hard core of mainstream economic theory was the Cambridge contrversy won by the English Cambridge, which destroys the notion of capital as it is used today; this result has simply disappeared from the debate. The same fate befell Debreu's critique (Debreu (1974), Sonnenschein (1972), Mantel (1974)) of his own theory. The Sonnenschein-Mantel-Debreu theorem is dangerous to the mainstream according to Hahn (1975) and still valid today according to Rizvi (2006).

The other two words that increase the heterodox markers for AER and JEBO are *environment**, which is now a common theme, and *distirub**, which denotes interest in distirbutive issues, though not necessarily on the macroeconomic level. Finally, it is interesting to point out that the CJE and AJS are the only ones who also deal with feminist issues by talking about them directly; whereas financialisation is a topic only touched upon by the CJE. The results in Table 7 are plotted in Figure 1.

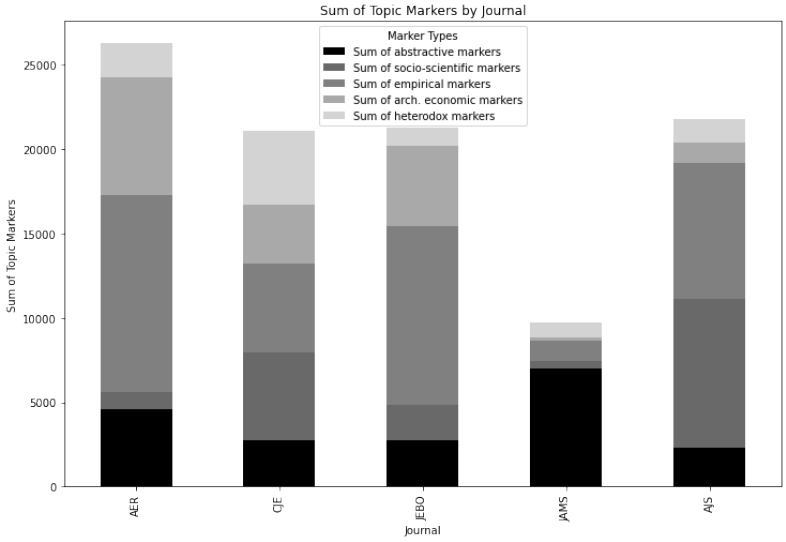

Figure 1: *The histogram summarizes 7: the relative frequencies of topic markers*

A useful tool for the purpose of this research is word collocation analysis. The scoring function used in this regard is the LogDiceScore (link). It measures how typical of a specific document (or corpus, or any unit of analysis) a collocation is, i.e. a node-collocate pair, where the node is the word under consideration and the collocates are all the words next to it. Below, I applied this scoring function to typically economic words, to compute the different uses made of them by the analyzed journals. The collocational graph should be red as follows: the higher is the distance from the center the lower is the typicality score (LogDice, which in this case represents the strength of a given link between two words in a given journal); the circle size indicates the absolute frequency of the collocation; each shade is a different grammatical relation between the node and the collocate; finally the slice's size approximates the number of collocates in that specific grammatical relationship with the node.

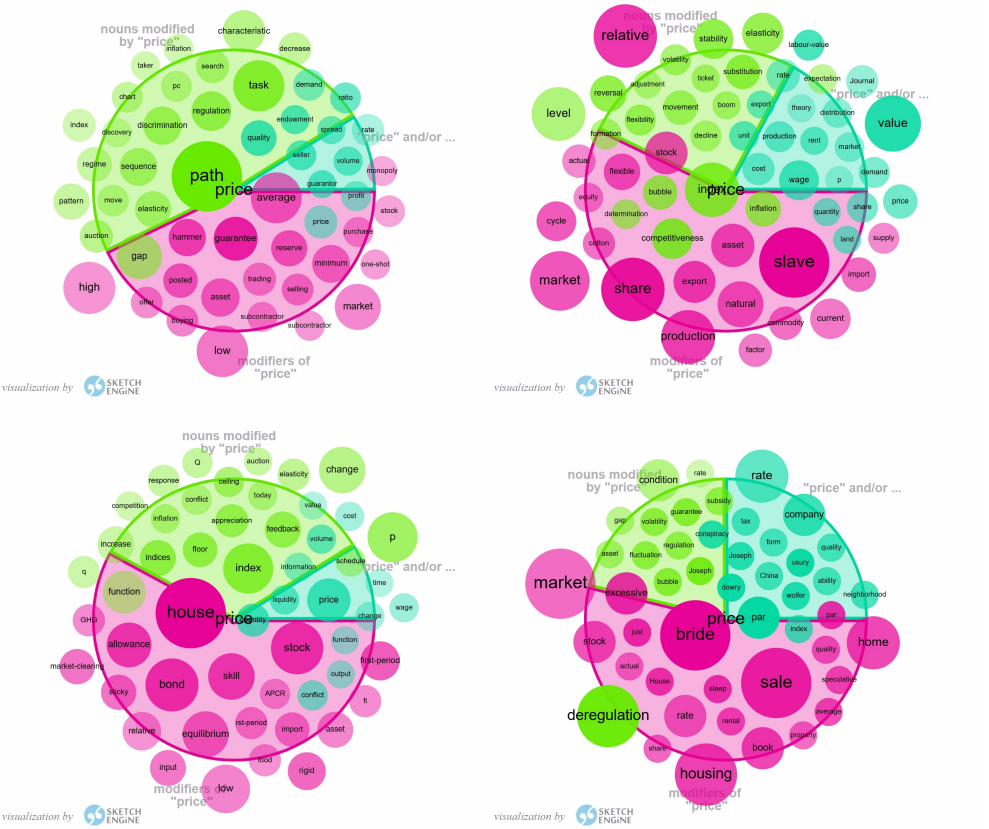

Figure 2: *The collocations of the word* price *(in clockwise sense) in JEBO, CJE, AJS and AER*

JEBO confirms the focus on dynamic aspects with *path* as a stronger collocation than *price*. AJS contains a bit of a mix of the collocations of the other journals. It is noticeable how the CJE links to *price* a series of names referring to crisis, instability and cycles (*fluctuation, volatility, bubble* and *speculative*). AER has no particularly surprising collocations, *sticky* falls into that mainstream pluralism already mentioned.

Figure 3: *The collocations of* agent *(clockwise) in AJS, AER, JEBO and CJE*

The collocations of *agent* are quite predictable and in line with what has been found so far. AJS seems to have a wider semantic range of the word than the other 3 journals, all of which use it with the same meaning.

Figure 4: *The collocations of* policy *(clockwise) in JEBO, CJE, AER and AJS*

The collocations of *policy* show the focus on the environment of JEBO and AJS. AER maintains an archetypically economic setting with *optimal* among the collocations. JEBO also has *optimal* among the collocations and also has *redistributive*. CJE and AJS are the only ones that mention *neoliberal policy*, showing an aptitude for theoretical discussion that was already evident from the keywords and the use of socio-scientific markers.

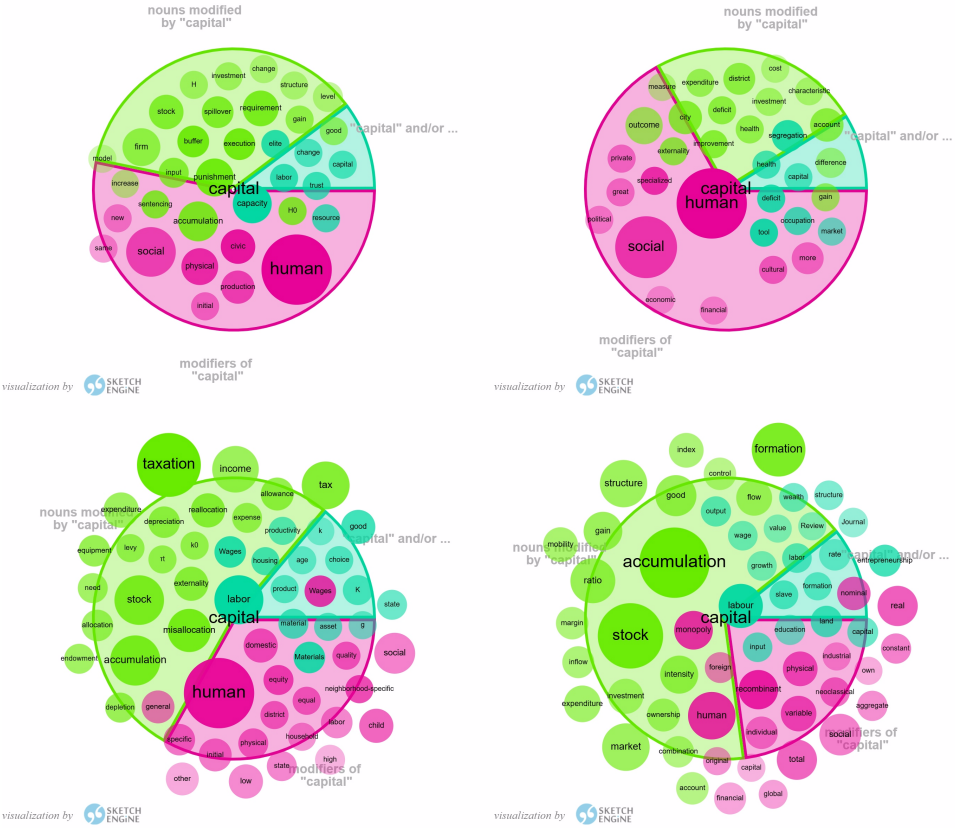

Figure 5: *The collocations of* capital *(clockwise) in JEBO, AJS, CJE and AER*

All journals make use of the collocation *human capital* although CJE less than the others.

Figure 6: *The collocations of* labor *(clockwise) in JEBO, AJS, AER and* labour *in CJE*

Similarly, the collocations of the word *labour* highlight how CJE also deals with labour in terms of conflict or in Marxist terms, and how AJS instead deals with organised struggles, trade unions, etc. AER and JEBO, on the other hand, speak of labour in a less political way and using terms that suggest a more technical approach.

The reduction of language to numbers is a simplification that makes one lose not only interpretation but also information with respect to standard text's exegesis. Relative frequency in itself does not tell us how a word is semantically used. For some, however, we can do something: for example, *rational** can be either positive, negatively connoted (*rationalism, rationalist*) or neutral (*rationalisation, rationalised*). Let us then see the relative frequencies of word forms in the various journals for *rational**. Obviously, on many words (such as *model*, *institutional*, *neoclassical* and also on the word form *rational*) there remains doubt as to the meaning they are associated to.

Let us begin by seeing which journals such a wild card is typical of, according to the measure reported in the column 'relative densitiy' in tab:5, column 3.

Table 8: *Relative frequency of the wild card* rational* *in each journal*

Journal	**Relative frequency**	**Relative density**
AER	341.91	172.30%
JEBO	257.47	129.75%
CJE	185.78	93.62%
JAMS	125.00	62.99%
AJS	99.69	50.24%

*rational** is typical in AER and JEBO, while it is not in other journals, but how is it used? The word forms can help us understand its semantic usage.

Table 9: *Relative Frequencies of the Word forms of* rational*

Word	**JEBO**	**AER**	**CJE**
rational	150.28	91.49	83.94
rationality	59.55	37.40	24.08
rationally	8.51	4.67	6.19
Rational	7.94	40.07	11.70
Rationality	7.37	10.02	1.38
rationalize	6.81	11.35	-
rationalized	5.67	4.67	2.06
rationale	4.54	11.35	29.59
rationalizing	1.70	0.67	0.69
rationalization	1.70	2.00	0.69
rational-play	0.57	-	-
rational-expectations	0.57	20.03	1.38
Rationally	0.57	2.00	-
rationales	0.57	1.34	5.50
rational-Expectations	0.57	1.34	-
rational-agent	0.57	-	-
rationalizable	-	88.15	-
Rationalizability	-	4.01	-
Rationalizable	-	3.34	-
rationalizes	-	3.34	-
rationalizability	-	1.34	-
Rationale	-	0.67	-
rationality-type	-	0.67	-
rationalitytype	-	0.67	-
Rationalize	-	0.67	-
Rational-Expectation	-	0.67	-
rationalisation	-	-	3.44
rationalizations	-	-	2.75
rationalised	-	-	2.06
rationalism	-	-	1.38
rational-choice	-	-	0.69
Rationalism	-	-	0.69
rationalisations	-	-	0.69
Rationales	-	-	0.69
Rationalization	-	-	0.69

*Rational** is used in a neutral and negative sense only by CJE, while AER and JEBO use it in a positive sense. only *rationalization* could also have a negative sense, but the concordance reveals that it is used positively in both

AER and JEBO.

4.2 Cosine similarity (TF-IDF and Word2Vec)

Another useful measure for comparing the content of texts is cosine similarity. Cosine similarity is a metric used to measure the similarity between two non-zero vectors that, in the context of text analysis, often represent the vector forms of two documents or sentences. The key idea is to assess the similarity between these vectors based on the angle between them, rather than their size. This is particularly useful in text analysis because it focuses on the direction of the vector, capturing the essence of the content rather than the absolute frequencies of lemmas. To calculate cosine similarity, each text is transformed into a vector in a multidimensional space, where each dimension corresponds to a unique word in the combined vocabulary of the two texts. The components of these vectors are the relative frequencies of the words (TF-IDF or term frequency). The cosine similarity is then calculated as the cosine of the angle between these two vectors, using the formula: $\text{cosine similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$, where the numerator is the dot product of the vectors and the denominator is the product of their magnitudes (euclidean norms). The resulting value ranges from -1 to 1, where 1 indicates identical vectors (i.e., the texts are very similar), 0 indicates orthogonality (no similarity), and -1 indicates completely opposite vectors (dissimilarity).

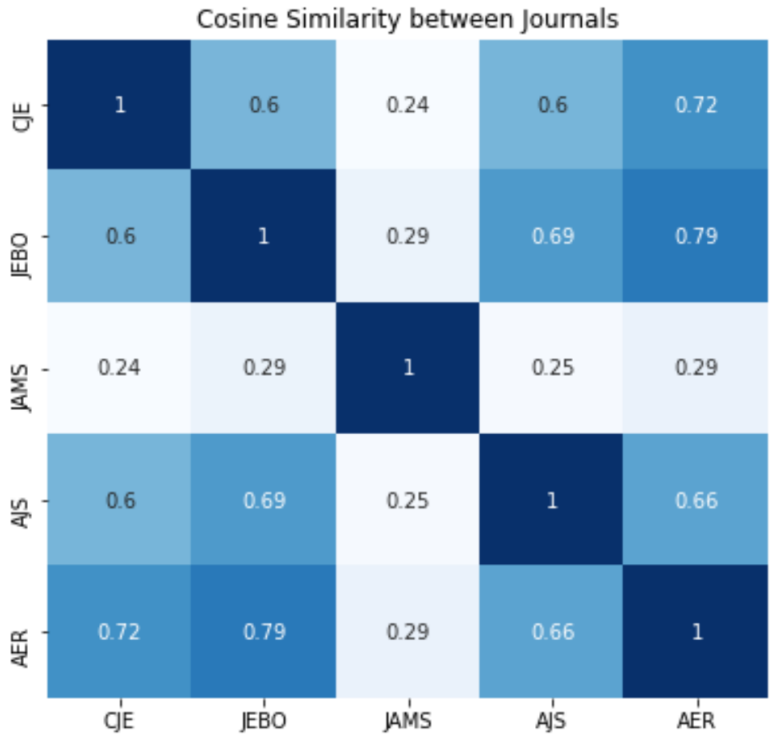

Figure 7: *The grid of pairwise TF-IDF Cosine Similarity between the journals*

The figure above shows the pairwise cosine similarity between the various journals. Surprisingly, CJE is more similar to AER than to the others and is equidistant from AJS and JEBO. The latter is very close to AER in word choices; it is also the most similar to AJS. JAMS is equidistant and quite different from all of them. As they are all scientific journals they are likely to share some default words.

A complementary measure of cosine similarity is the one based on Word2Vec word embeddings represented in 8. Unlike TF-IDF, which captures only word frequencies, Word2Vec encodes the meaning and semantic context in which words are used. While TF-IDF represents documents through a weighted frequency matrix, Word2Vec utilizes pre-trained word vectors (in this case Google news300) to map each term into a continuous high-dimensional space, allowing it to capture latent relationships between words. As a result, documents that share similar concepts but use different vocabulary can still be considered semantically close. This distinction is particularly evident in the similarity heatmaps: TF-IDF primarily reflects lexical overlap, emphasizing journals with a similar distribution of word frequencies, whereas Word2Vec captures deeper conceptual connections. For instance, JAMS appears distinct in both approaches, but for different reasons, its vocabulary significantly diverges in TF-IDF, while its thematic structure is unique in Word2Vec. Similarly, CJE maintains a moderate distance from mainstream economics and

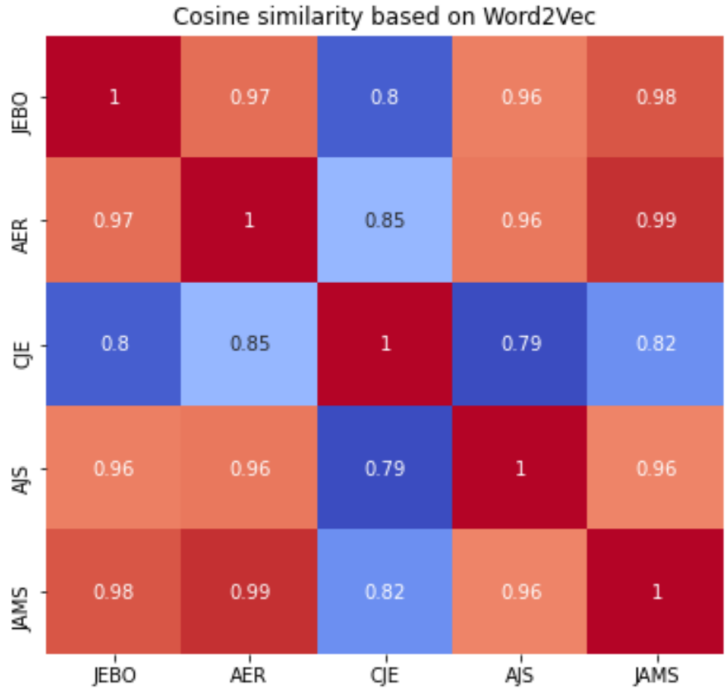

Figure 8: *The grid of pairwise Word2Vec Cosine Similarity between the journals*

sociology journals in both models; however, in TF-IDF, its similarity scores indicate a closer lexical alignment, while in Word2Vec, the greater semantic distance suggests a more nuanced conceptual deviation. These findings highlight the complementary nature of the two methods: while Word2Vec is more effective for capturing thematic and semantic relationships, TF-IDF provides a precise measure of textual similarity based on explicit word occurrences. A hybrid approach that integrates both methods could offer a more comprehensive understanding of disciplinary boundaries and textual coherence across academic literature.

Both in the following topic modelling and in 7 and 8, besides the stopwords, some words attributable to the 'noise' of the pdf downloaded from the journals' sites ('journal','downloaded','american','review','università','licensed','degli','studi', and others) have been removed.

4.3 LDA model

The results of the LDA model are presented below, starting with the result with 3 topics ($k = 3$) which is a way of forcing the model to 'match' the journals in order to see which ones are similar. This is followed by results with 5 and with 35 topics. Unless otherwise specified, all runs are done with α and β (also called η) optimised to the available data. (e.g. $\alpha =' auto'$ and

$\beta =' auto')$.

There is no common procedure to derive the number of topics (see Rhody (2012)), or a test supporting a precise choice of the parameters, especially when topic modeling is employed to explore the content of a dataset, and not for prediction (Mimno and Blei (2011)). The coherence score used in the robustness section is an example of unsupervised, of automated choice of the number of topics.

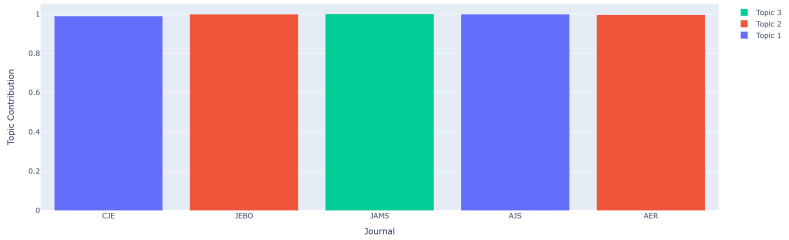

Figure 9: *Topics distribution in the 5 journals with 3 topics*

The topic distribution across journals with $k = 3$ (number of topics) reveals that CJE and AJS deal with the same topic, as do JEBO and AER, while JAMS deals with a topic of its own. The figure below also provides a measure of the distance between topics via a multidimensional scaling (MDS) of the LDA output. Although similar to PCA (Principal Component Analysis, used for example by Tusset (2021)), MDS differs in that it is based on a dissimilarity matrix between pairs of points rather than a variance/covariance matrix. Furthermore, MDS does not require the original distances to be linear or Euclidean, allowing more complex relationships between data to be modelled.

The visualisation type in 9 called LDAvis was developed by Sievert and Shirley (2014). Thanks to λ it is possible to choose whether to display words with higher frequencies, but common to many topics, or words exclusive to a certain topic.

λ is a measure of 'relevance' or 'typicality' of a term to a topic and rests on the possibility of linearly combining the probability ϕ of a term w to topic k and its exclusivity or 'lift', defined as the ratio of a term w's probability p within the topic to its marginal probability across the corpus. Relevance depends on the value to be attributed to a parameter λ, ranging from 0 to 1, that determines the relative weight assigned to the log of the two components, the probability in the corpus and lift. The relevance index r of term w for topic k depends on λ and takes the following form:

$$r(w, k \mid \lambda) = \lambda \log(\varphi_{wk}) + (1 - \lambda) \log \left(\frac{\varphi_{wk}}{p_w} \right) \tag{1}$$

λ is the relative weight assigned to the *Log* of the two components: the probability of the term given the topic and the probability of the term in the whole corpus.

Sievert and Shirley (2014) suggest setting $\lambda = 0.6$. If it is close to 0, the figure shows words that are potentially rare in the whole corpus, but exclusive (or very typical) of that specific topic. Whereas with values close to 1, it highlights words with an higher frequency of occurrence in the whole corpus, but which could not be exclusive of a given topic.

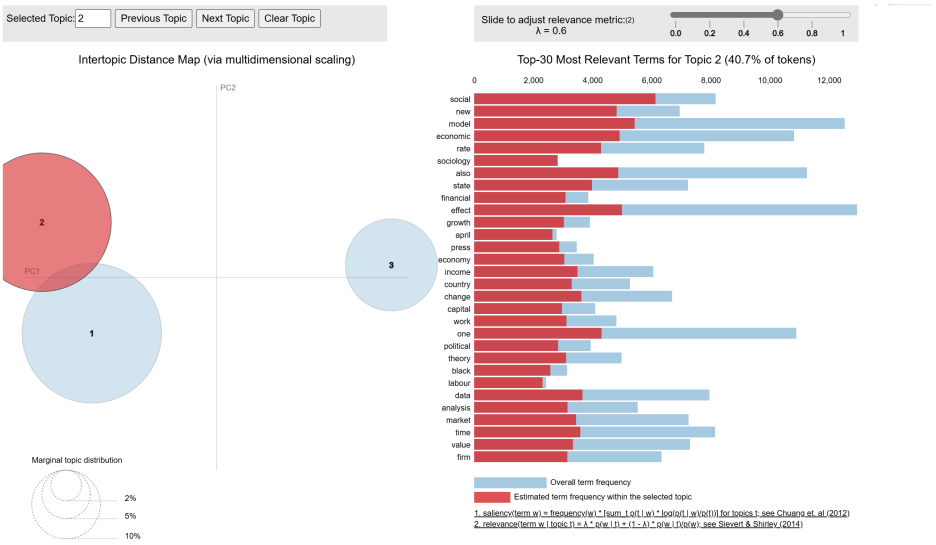

Figure 10: *Map of 3 topics. LDAvis visualization*

Each circle on the map represents a single topic. The distance between the circles reflects how far apart the topics are semantically. The closer the circles are, the more related the topics are (i.e. they share similar words). Thus topics that are far apart in the two-dimensional plane are thematically distinct. The size of the circles (or bubbles) represents the marginal distribution of the topic, i.e. how dominant each topic is in the corpus. Larger bubbles indicate more representative topics in terms of frequency. The axes of the two-dimensional plane (PC1 and PC2) represent the two principal components (Principal Component 1 and 2). They are reduced and synthetic dimensions of the topic space, where each component is a linear combination of the original variables (words in the topics). PC1 and PC2 represent the synthetic dimensions that best distinguish between topics in terms of words used in the texts.

Topic 2 can be attributed for its content to JEBO and AER, while topic 1 to CJE and AJS. These two topics overlap, meaning that the social sciences journals deals with overlapping topics. The fact that topics overlap does not contradict the topic distribution in the 8, where the words characterising a topic are chosen with a low λ and thus have a high probability of being associated with a specific topic.

With $\lambda = 0.6$ the first 20 words of each topic are:

- Topic 1: effect, model, result, economic, behavior, table, cost, price, treatment, subject, percent, optimal, p, equilibrium, one, agent, level, b, two, also, c, estimate, data, panel, game, information, time, market, variable, econ. Can be interpreted as "economic and behavioural models: cost analysis, pricing and market dynamics"

- Topic 2: social, new, model, economic, rate, sociology, also, state, financial, effect, growth, april, press, economy, income, country, change, capital, work, one, political, theory, black, labour, data, analysis, market, time, value, firm. Can be broadly labelled as "socio-economic models and market dynamics: analysis of political, financial and labour factors".

- Topic 3: x, f, n, g, k, let, v, p, r, u, w, see, h, b, theorem, lemma, c, z, may, j, proof, restriction, apply, map, set, q, e, di, group, ξ. Which form the "geometry and algebraic topology topic".

It is noteworthy that topic 2 is closer to topic 3 (attributable to JAMS) than topic 1. Topic 1 could be broadly defined as heterodox economics and sociology, in particular growth theory, financialisation, Keynesian theory and racial segregation (an issue that seems dear to AJS). Topic 2 is an archetypical economics topic. Topic 3 instead can be labelled as geometry and algebraic topology.

Only words from topic 1 are displayed. The red part of the frequency bar indicates the estimated frequency of the word within the reference topic, while the blue part indicates the frequency in the rest of the corpus. It can considered a measure of the typicality of the word with respect to a topic.

Note that the bubble numbers in the maps (9,11,13) do not correspond to the histogram's order in the topic distribution (8,10,12). LDAvis visualisation does not group or assign topics to documents but simply divides the entire corpus into topics, it is up to the researcher to interpret what a given set of words is about. In this case it appears that many words are unambiguously attributable to specific journals. But the content of the histograms and that one of the map are profoundly different.

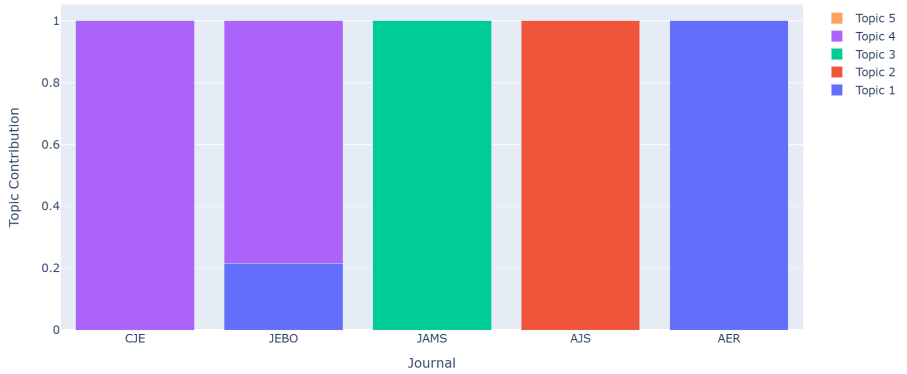

Figure 11: *Topics distribution in the 5 journals with 5 topics*

10 shows that each journal deals with a different topic.

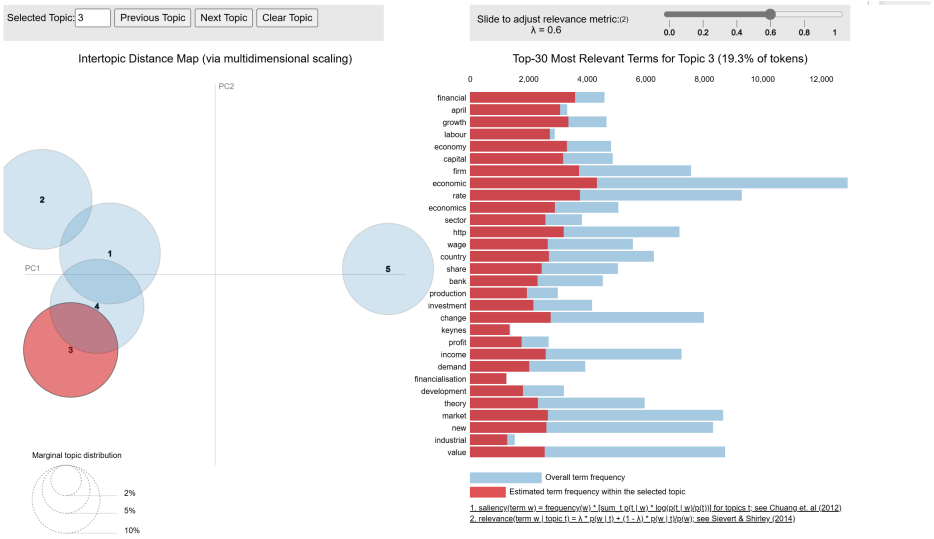

Figure 12: *Map of 5 topics' LDA output*

The topic distribution reveals that JEBO shares topics with both CJE and AER, suggesting it occupies a middle ground between the two. This may reflect JEBO's interdisciplinary scope or methodological affinities with both journals. The other journals are each strongly associated with a single, distinct topic, highlighting clear thematic separation across the rest of the corpus. Here the list of the first 20 words for each topic with $\lambda = 0.6$. The words of each topic are reported below:

- Topic 1: [behavior, econ, treatment, effect, c, subject, organization,

et, result, task, game, experiment, table, p, economic, b, θ, l, level, e, variable, participant, model, one, individual, h, payoff, yes, σ, offer]

- Topic 2: [social, sociology, black, white, state, new, group, network, racial, model, press, neighborhood, school, data, work, tie, police, effect, research, study, also, men, york, practice, political, job, one, woman, analysis, prison]

- Topic 3: [financial, April, growth, labour, economy, capital, firm, economic, rate, economics, sector, http, wage, country, share, bank, production, investment, change, Keynes, profit, income, demand, financialisation, development, theory, market, new, industrial, value]

- Topic 4: [percent, firm, model, panel, effect, estimate, shock, market, economic, tax, data, figure, equilibrium, cost, result, online, first, price, child, market, table, column, agent, information, rate, fixed, time, belief, year, show]

- Topic 5: [x, f, g, k, n, let, v, theorem, lemma, r, w, u, h, proof, z, map, p, restriction, see, apply, ip, edt, apr, isomorphism, copyright, prepared, b, redistribution, j, c].

Results below are with $k = 35$ (35 topics).

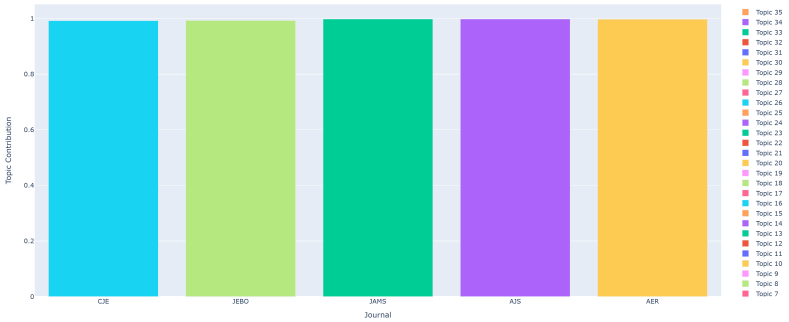

Figure 13: *35 topics*

With 35 topics each journal deals with little more than one topic, which suggests that the content covered in each journal is quite singular.

I also tried it with 20 topics and the result was substantially the same as in the case with 35. Since there are not 3 different disciplines in the corpus, it seemed appropriate to raise the number of topics with respect to studies that only concern economics such as that of Ambrosino et al. (2018).

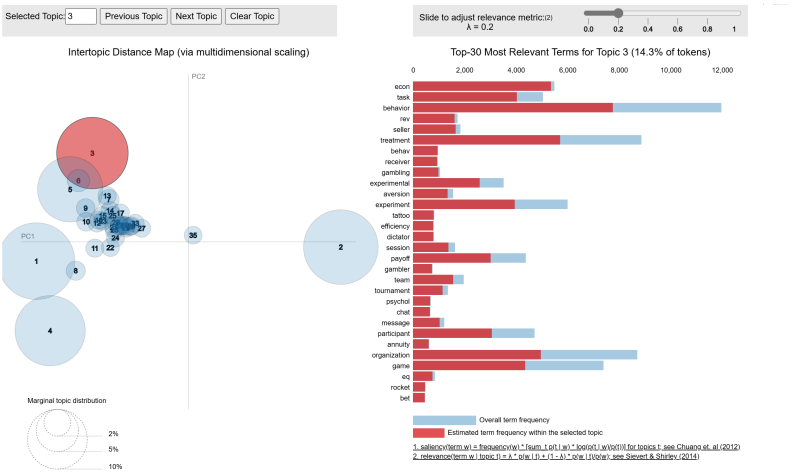

Figure 14: *35 topics inter distance map*

Even with 35 topics there are clearly topics that are more important than others in the whole corpus and seems to correspond (both from the fact that they are 5 and from their content) to the corpus' documents (the journals). The narrower circles in the picture are words that lie halfway between the social science topics and the mathematical one. Topic 35 is exactly halfway through and consists of this list of words ($\lambda = 0.2$ here because with 35 topics I preferred to visualize words very typical of each topic): β, fubini, zx, yampolsky, technicality, dujardin, epimorphism, hee, encyclopaedia, secant, cancelled, honda, ambrosio, imaginary, teristic, reformulation,spreading, lurie, nt, kobayashi, lc, finer, cattani, satellite, knot, interpolating, endow. It can be labelled as advanced mathematics and algebraic geometry. "lc" stands for "log-canonicity".

Again, the map seems polarised by the distinction between hard sciences and social ones. The complete list of the words of the dominant topics in the corpus (the largest bubbles) is given below, one can easily infer which journals have most influenced each topic in the LDAvis output.

The topic highlighted in red is the one attributable to JEBO and the words that make it up ($\lambda = 0.2$) are shown in figure 13.

4.4 PCA analysis

Since LDA shows relationships between topics themselves, and not between documents, in order not to rely only on personal interpretation of topics, a principal component analysis was also run.

Principal Component Analysis (PCA) is a dimensionality reduction technique that identifies the directions in the data with the highest variance. The first principal component (PCA1) represents the direction along which the variance is maximized, capturing the most significant differences among observations. The second principal component (PCA2) is orthogonal (or in-

dependent) to PCA1 and accounts for the next largest portion of variance. Orthogonality ensures that each principal component captures unique, non-overlapping information. These components are derived as linear combinations of the original features, in this case, the topic distributions from the Latent Dirichlet Allocation (LDA) model. In other words the two PCAs are the projection of the data along the two eigenvectors associated with the highest eigenvalues of the data varaince/covariance matrix. The original data matrix is $W \in \mathbb{R}^{25\times 5}$. Whose entries are the probability distributions of topics per document. The variance covariance matrix is $\Sigma \in \mathbb{R}^{25\times 25}$. Each element (σ_{ij}) of the covariance matrix represents the covariance between topic (i) and topic (j), computed across the 5 documents. Each principal component in a PCA is a linear combination of the original variables (in this case, topics), where the weights of the combination are given by the coefficients of the corresponding eigenvector. Formally, if the dataset consists of n variables $X_1, X_2, \ldots, X_n$, then the first and second principal components (PCA1 and PCA2) can be written as:

$$\text{PCA1} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n \tag{2}$$

$$\text{PCA2} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2n}X_n \tag{3}$$

where the vectors

$$\mathbf{a}_1 = [a_{11}, a_{12}, \ldots, a_{1n}], \quad \mathbf{a}_2 = [a_{21}, a_{22}, \ldots, a_{2n}]$$

are the eigenvectors associated with the first and second largest eigenvalues of the covariance matrix, respectively. As the eigenvectors define the directions of maximum variance in the data the corresponding eigenvalues indicate how much variance is explained by each component.

In the PCA space, journals that are positioned farther apart differ more in their topic composition, while those closer together share a more similar thematic structure. The interpretation of PCA1 and PCA2 does not have a fixed meaning but is instead data-driven, reflecting the axes that best separate the topic distributions across journals. The amount of variance explained by PCA1 and PCA2 is quantified (PCA1: 60.42%, PCA2: 39.58%) to assess how well the two-dimensional representation retains the original information. In this case the total variance is explained by the two components. The number of topics is the optimal one, according to the coherence score discussed in the next subsection.

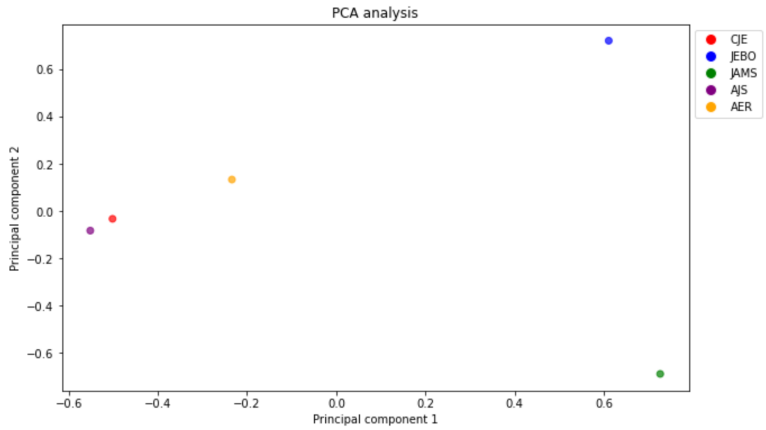

Figure 15: *PCA analysis based on topic distribution (number of topics set via the coherence score) in the 5 journals*

From the PCA plot we observe that CJE, AJS and just a bit further AER, are positioned relatively close to each other, indicating a degree of thematic proximity in their topic distributions. This suggests that, despite being from different disciplines, these journals may share conceptual frameworks or methodological approaches that lead to some semantic similarity. On the other hand, JAMS is the most distant from the others, positioned at the lower right corner. This separation suggests that its thematic structure significantly diverges from the other journals, likely due to its focus on mathematical and statistical methods rather than social science discourse. An interesting observation is JEBO, which appears slightly detached from the main cluster and positioned towards the upper right. This may indicate that its research themes, while still related to economics and social sciences, have some distinct characteristics that differentiate it from the rest. In order to better understand which ones, I refer to the results in subsection 4.1 from computational linguistics methods.

Given the limited number of documents (five, one for each journal) in the dataset, the Principal Component Analysis (PCA) was conducted setting the number of topics to five. This choice aligns with the dataset size and provides a more reliable and interpretable outcome compared to larger topic numbers, which could lead to overfitting or spurious results. The corresponding PCA results are illustrated in 16.

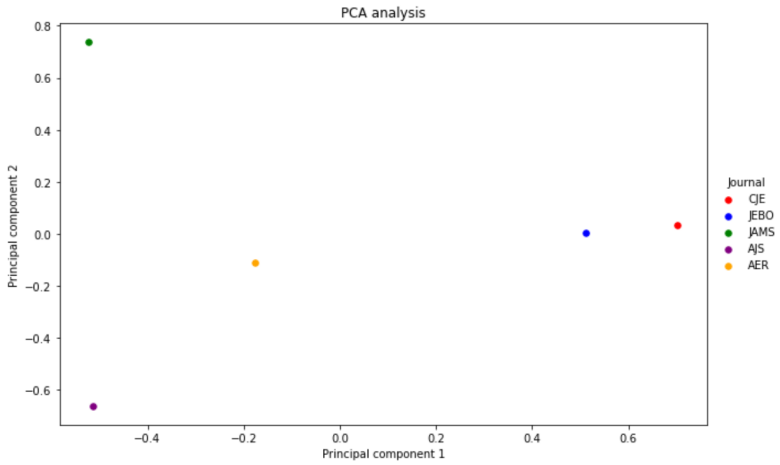

Figure 16: *PCA analysis based on topic distribution (number of topics manually set and equal to 5) in the 5 journals*

It is noteworthy that the results from the LDA Davis map shown in 12 are consistent with the PCA analysis conducted with five topics. Specifically, JAMS and AJS represent two opposite poles, while the economics journals are situated intermediately, consistent with their theoretical and methodological orientations.

These findings demonstrate how PCA can effectively complement topic modeling techniques by providing a global view of document relationships based on their topic distributions. It is important to note that the axes in the LDA visual map (often labeled as PC1 and PC2, as in 14, 12 and 10) are not true principal components in the statistical sense, as in Principal Component Analysis. Instead, they are synthetic dimensions produced through a technique called Multidimensional Scaling (MDS), which aims to preserve the pairwise distances between topics in a two-dimensional space. These distances are typically computed using a dissimilarity measure such as Jensen-Shannon divergence applied to the topic-word distributions. The resulting 2D coordinates are chosen so that topics that are more similar (i.e., that share more words or have overlapping distributions) appear closer together on the map, while more distinct topics are farther apart. Thus, the purpose of these axes is purely visual: to approximate the high-dimensional relationships between topics in a human-interpretable format. They do not represent directions of maximum variance, nor do they involve eigenvectors or eigenvalues of a variance/covariance matrix, as is the case in pure PCA analysis.

4.5 Robustness

In order to test the robustness of the model, LDA is run also with the hyperparameters (α and β) of the Dirichlet distribution set to high (14) and low (13) values, inducing a less sparse and more sparse topics distribution respectively.

In the first case, with high α, there is a higher probability of having more uniform topics, and with high β more words are plausible to be associated to each topic, resulting in less sparse topics. Setting both hyperparameters to high values increases the likelihood of observing more topics per document; conversely, the opposite is true when the hyperparameters are low.Mimno et al. (2009) argue that tuning these hyperparameters is crucial for the model's robustness, which is one of the reasons why I included a data-driven approach. However, the impact of setting extreme hyperparameter values is explored . As the number of topics increases, the model becomes more sensitive to hyperparameter adjustments. Therefore, I fixed the number of topics at 35 in both 13 and 14.

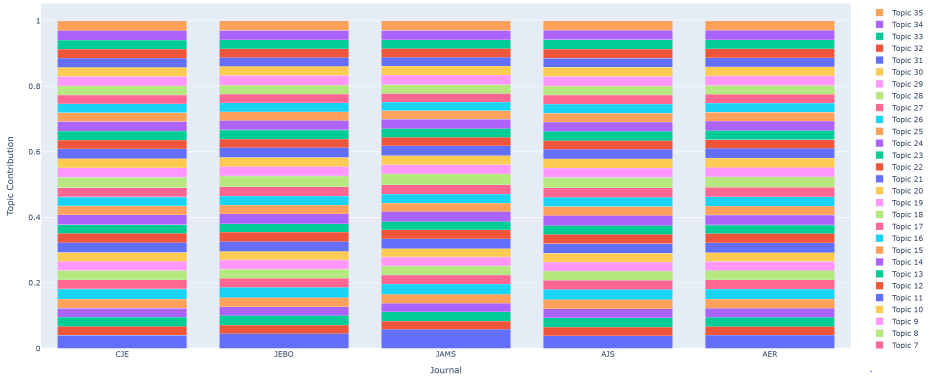

Figure 17: *35 topics with* $\alpha = 1, \beta = 1$

You can see that the topics are almost equally distributed, except for JAMS which has a higher percentage of the topic in blue, whose words are mainly mathematical symbols.

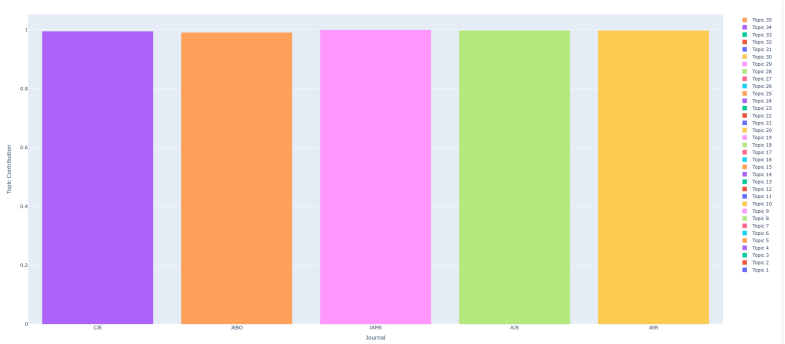

Figure 18: *35 topics with* $\alpha = 0.000001, \beta = 0.000001$

With a low value of the hyperparameters the output is equal to the unsupervised version of figure 12.

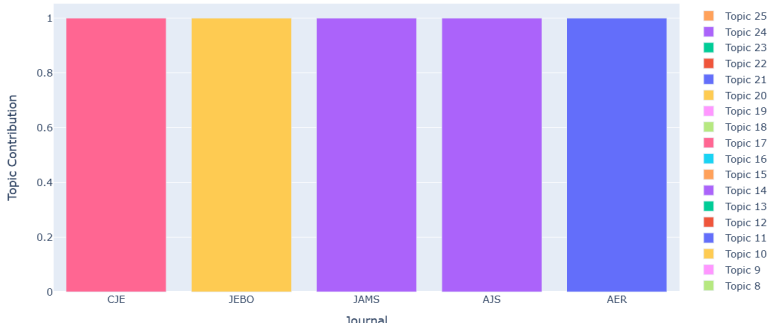

Figure 19: *Topic distribution with the coherence score method to find the optimal number of topics (25)*

Using the coherence score to automatically find the optimal number of topics does not change the results, as seen in 19. Notice that the two purples are different among them, the code randomly assigns colors to the 25 topics, and similar colors do not indicate closer proximity of topics.

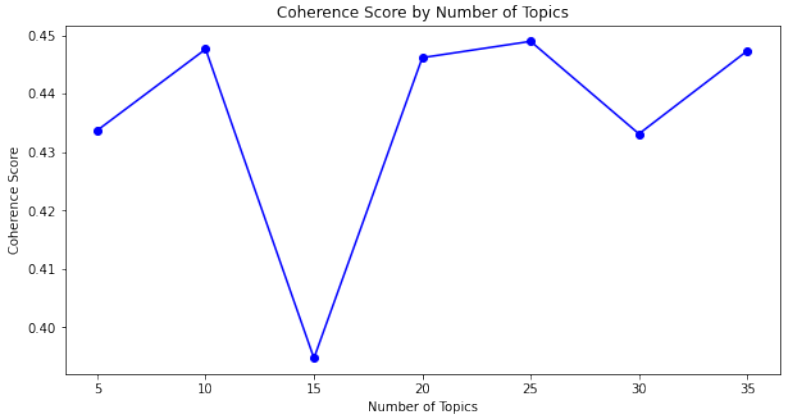

Figure 20: *Coherence score for each number of topics*

20 represents the relationship between the Choerence score and the number of topics.

5 Conclusions

Mixed methods look promising even if the results are not particularly surprising. They show that JAMS deals with a topic of its own, while the other four journals deal broadly with the same topics. This result simply reflects the division between hard sciences and social sciences. JEBO is the closest to JAMS, followed by AER, CJE and AJS in that order, if the topic interpretation is correct. Among the economics journals, the CJE is the most sociological, as one might expect. Indeed, heterodox economics maintained

closer contact with general sociology than mainstream pluralism, which, if it has approached sociology, has done so in more its quantifiable aspects, such as network analysis (the work of Mark Granovetter is a cornerstone) or spatial models alla Schelling.

The analysis revealed clear distinctions in the usage of language across the economics journals (AER, JEBO, CJE) compared to those in other disciplines (JAMS in mathematics and AJS in sociology). Notably, AER and JEBO shared similarities in the use of formal economic terminologies and models, reflecting their orientation towards mainstream economic theories and quantitative methodologies. This is in contrast to CJE, which predominantly utilized heterodox economic terms and concepts, aligning with its editorial mission to foster diverse economic discussions.

The distinctions are particularly salient when considering the application of LDA, which highlighted not only the thematic focus of each journal but also the underlying methodological preferences. For instance, the prevalent use of empirical and econometric terms in AER and JEBO suggests a strong emphasis on data-driven research, which is less pronounced in CJE where historical and socio-economic contexts are more prominently featured.

These overlaps not only highlight the potential for methodological borrowing across disciplines but also suggest that economic research can benefit from incorporating broader social and mathematical perspectives to address complex economic phenomena more comprehensively. The use of computational methods, such as LDA and computational linguistics, has demonstrated substantial utility in uncovering latent semantic structures within large text corpora. This approach is particularly beneficial in settings where traditional qualitative methods may fall short due to the sheer volume and complexity of the data. The ability of LDA to distill complex datasets into comprehensible thematic structures offers a powerful tool for researchers to systematically explore and compare vast amounts of academic literature.

Moreover, the findings emphasize the importance of methodological rigor and transparency in computational research. The variations in topic coherence and relevance across different model settings underscore the need for careful calibration and validation of computational models to ensure their reliability and applicability to economic research.

While this study provides valuable insights into the linguistic and methodological landscapes of economic journals, it is not without limitations. The scope of journals and articles analyzed, while comprehensive, is not exhaustive. Future research could expand the corpus to include a broader array of journals and interdisciplinary fields to examine more diverse linguistic and ideological trends.

Additionally, the dynamic nature of academic disciplines suggests that linguistic and methodological trends may evolve over time. Longitudinal studies could provide deeper insights into how economic discourse has changed in response to shifting academic, social, and economic contexts.

A critical aspect of this study is the examination of how results from two

distinct methodologies, computational linguistics and Latent Dirichlet Allocation (LDA), intersect and reinforce each other. The convergence of these methods is evident in the way they both highlight the distinct thematic and ideological focuses of the journals analyzed. For example, both methodologies underscored the predominance of formal economic modeling in AER and JEBO, as well as the emphasis on heterodox and sociologically-oriented themes in CJE. This consistency bolsters the validity of the findings, indicating that despite their different operational mechanisms, both approaches are capable of capturing the fundamental linguistic patterns that define these journals.

Furthermore, the intersection of these methodologies provides deeper insights into the structure and coherence of the discourse within each journal. While computational linguistics offers a granular view of the frequency and distribution of specific linguistic markers, LDA presents a broader thematic landscape that these markers collectively contribute to. This dual perspective is particularly valuable in identifying the underlying themes that may not be immediately apparent through a single methodological lens. For instance, the prevalence of socio-economic markers identified through computational linguistics in CJE aligns with the broader themes of economic heterodoxy and social context revealed by LDA, confirming the journal's commitment to a diverse and interdisciplinary approach.

The points of contact between the results of the two methodologies not only validate the individual findings but also enhance our understanding of how different linguistic elements come together to shape scholarly discourse. This methodological synergy offers a robust framework for future research, suggesting that a combined approach can provide a comprehensive and nuanced analysis of complex textual data in academic research.

Latent Dirichlet Allocation (LDA) leverages the automated, unsupervised nature of topic detection, allowing it to uncover latent or hidden structures within economic works that might otherwise remain invisible. The purpose of topic modeling is not to classify journals per sè, as each journals in principle could contain a mixture of topics, nor to strictly group them into clusters or fields, but rather to analyse the frame of the discourse, the way the writer implicitly or unconsciously or strategically maybe, characterize the discourse.

In fact, the basic assumption of this approach is that the choice of words is in itself (independently of the logic of the sentence and of the discourse) a way of giving a shape to the discourse, a latent structure in fact. In this way, LDA facilitates the comparison of 'economics as discourse' (Amariglio (1990), McCloskey (1983), Samuels (1992)) with other sciences, introducing a more radical concept of 'conversation' than what is traditionally captured by bibliometric methods. For instance, an exercise in which topic modelling could be useful is linking specific historical periods to the spread of specific theories, hypotheses or frames of discourse.

The assumption that differences in the semantic content of topics reflect the ideological setting is strong and requires further research beyond economics, into other social sciences, hard sciences and humanities. At least,

topic modeling, as an unsupervised technique, is useful to assist in searching, browsing, and summarizing large archives, and it can challenge human-assigned metadata or subject classifications like JEL codes or other institutional classification. Studies suggest that automated classification methods are better at identifying new areas of knowledge (see for instance Suominen and Toivanen (2016)).

Perhaps most importantly, there is evidence that shifts in the language of economic documents, particularly in semantic terms, mirror changes in approaches and attitudes. Studies of this nature, if supported by rigorous research in the history of economics, can significantly impact our understanding of how economic knowledge disseminate and evolves (Ambrosino et al. (2018)). Topic modeling, in particular, can aid scholars in applying quantitative historical semantics to economics (Klaes (2017)). The convergence between computational linguistics and LDA results is evident as both methodologies consistently highlight similar thematic orientations across the journals. For instance, both approaches reveal a stronger presence of formal economic modeling terms in AER and JEBO, and a greater emphasis on socio-economic contexts in CJE. This methodological harmony enhances the credibility of our findings, suggesting that integrating these tools can offer a comprehensive view of disciplinary languages and methodologies.

Moreover the integrated use of topic modelling, computational linguistics techniques and traditional text exegesis, can strengthen the scientificity of an interpretation in the field of HET; and can help trace theoretical trajectories of both economics as a whole and of its sub fields. This is closely related to the need for economists to collaborate with economic methodologists and philosophers of science to carry out historical/scientometric analyses and to question the scientifcity of the methods adopted in social sciences.

The insights gained from this study advocate for the adoption of mixed computational methods across other disciplines to uncover hidden thematic structures and methodological preferences.

References

Amariglio, J. (1990). Economics as a postmodern discourse. In *Economics as discourse: An analysis of the language of economists*, pages 15–64. Springer.

Ambrosino, A., Cedrini, M., and Davis, J. B. (2024). Today's economics: one, no one and one hundred thousand. *The European Journal of the History of Economic Thought*, 31(1):59–76.

Ambrosino, A., Cedrini, M. A., Davis, J. B., Fiori, S., Guerzoni, M., and Novarese, M. (2018). Topic modeling, long-term trends and the transition from classical to modern economics. *The European Journal of the History of Economic Thought*, 25(3):409–434.

Ash, E. and Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, 15(1):659–688.

Backhouse, R., Middleton, R., and Tribe, K. (1997). 'economics is what economists do', but what do the numbers tell us? *but what do the numbers tell us*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bogenhold, D. (2020). Economic thought and history: The neglected legacy of the german historical school. *Journal of Economic Issues*, 54(1):1–18.

Card, D. and DellaVigna, S. (2013). Nine facts about top journals in economics. *Journal of Economic Literature*, 51(1):144–161.

Cei, L., Defrancesco, E., and Stefani, G. (2022). Evolution of agricultural economics: Insights from topic modeling. *European Review of Agricultural Economics*, 49(4):749–777.

Cherrier, B. (2017). Classifying economics: A history of the jel codes. *Journal of Economic Literature*, 55(2):545–546.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, 3:1–44.

Coats, A. B. (2014). Historiography of economics: Methodology, history and policy. *Journal of the History of Economic Thought*, 36(3):247–259.

Davis, J. B. (2006). The turn in economics: neoclassical dominance to mainstream pluralism? *Journal of institutional economics*, 2(1):1–20.

Debreu, G. (1974). Excess demand functions. *Journal of mathematical economics*, 1(1):15–21.

Diaz-Bone, R. and Didier, E. (2016). Introduction: The sociology of quantification-perspectives on an emerging field in the social sciences. *Historical Social Research/Historische Sozialforschung*, pages 7–26.

D'Orlando, F. (2013). The role of electronic resources in promoting heterodox economics. *Journal of Economic Methodology*, 20(4):403–421.

Dupont, B. (2017). Continuity in economic ideas: the history of economic thought as a political argument. *European Journal of the History of Economic Thought*, 24(4):651–672.

Düppe, T. and Weintraub, E. R. (2018). *The evolution of historiographical approaches in economics*. Duke University Press.

Edwards, J., Giraud, Y., and Schinckus, C. (2018). A quantitative turn in the historiography of economics?

Friedman, M., Schwartz, A. J., et al. (1963). *Money and business cycles*. Bobbs-Merrill Company, College Division.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Goldschmidt, R. and Szmrecsanyi, B. (2007). Exploring rhetorical strategies in economics: A corpus-linguistic analysis. In *Corpus Linguistics 2007 Conference*.

Hahn, F. H. (1975). Revival of political economy: The wrong issues and the wrong argument. *Economic Record*, 51(135).

Hsu, C., Yu, T., and Chen, S.-H. (2021). Narrative economics using textual analysis of newspaper data: new insights into the us silver purchase act and chinese price level in 1928–1936. *Journal of Computational Social Science*, 4(2):761–785.

Klaes, M. (2017). Quantitative approaches to historical semantics in economics. In *22nd annual conference of the European Society for the History of Economic Thought (ESHET), University of Antwerp*, pages 18–20.

Lakatos, I. (2014). Falsification and the methodology of scientific research programmes. In *Philosophy, Science, and History*, pages 89–94. Routledge.

Mantel, R. R. (1974). On the characterization of aggregate excess demand. *Journal of economic theory*, 7(3):348–353.

McCloskey, D. N. (1983). The rhetoric of economics. *Journal of economic literature*, 21(2):481–517.

Mimno, D. and Blei, D. (2011). Bayesian checking for topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 227–237.

Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 880–889.

Mohr, J. W. and Bogdanov, P. (2013). Introduction—topic models: What they are and why they matter.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

Rhody, L. M. (2012). Topic modeling and figurative language.

Rizvi, S. (2006). The sonnenschein-mantel-debreu results after thirty years. *History of Political Economy*, 38.

Samuels, W. J. (1992). *Essays on the Methodology and Discourse of Economics*. Springer.

Say, W. D. C. L. (2018). Has homo economicus evolved into homo sapiens from 1992 to 2014. *Big Data in Computational Social Science and Humanities*, page 117.

Sievert, C. and Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70.

Sonnenschein, H. (1972). Market excess demand functions. *Econometrica: Journal of the Econometric Society*, pages 549–563.

Suominen, A. and Toivanen, H. (2016). Map or maze? unsupervised and supervised learning for analysing contemporary science. *Journal of Informetrics*, 10(3):633–648.

Tusset, G. (2021). Plotting the words of econophysics. *Entropy*, 23(8):944.

Wehrheim, L. (2019). Economic history goes digital: Topic modeling the journal of economic history. *Cliometrica*, 13:83–125.

A Appendix

Journal statements are reported below. They assess that CJE has an heterodox approach, JEBO focuses on computational method and complex systems and AER is of general interest.

Cambrdige journal of Economics: "The Cambridge Journal of Economics, founded in the traditions of Marx, Keynes, Kalecki, Joan Robinson and Kaldor, welcomes contributions from heterodox economics as well as other social science disciplines. Within this orientation the journal provides a focus for theoretical, applied, interdisciplinary, history of thought and methodological work, with strong emphasis on realistic analysis, the development of critical perspectives, the provision and use of empirical evidence, and the construction of policy. The Editors welcome submissions in this spirit on economic and social issues including, but not only, unemployment, inflation, the organisation of production, the distribution of the social product, class conflict, economic underdevelopment, globalisation and international economic integration, changing forms and boundaries of markets and planning, and uneven development and instability in the world economy." (Link reported here)

Journal of Economic Behavior and Organization: "The Journal of Economic Behavior and Organization is devoted to theoretical and empirical research concerning economic decision, organization and behavior and to economic change in all its aspects. Its specific purposes are to foster an improved understanding of how human cognitive, computational and informational characteristics influence the working of economic organizations and market economies and how an economy's structural features lead to various types of micro and macro behavior, to changing patterns of development and to institutional evolution. Research with these purposes that explore the interrelations of economics with other disciplines such as biology, psychology, law, anthropology, sociology, finance, marketing, political science, and mathematics is particularly welcome. The journal is eclectic as to research method; systematic observation and careful description, simulation modeling and mathematical analysis are all within its purview. Empirical work, including controlled laboratory experimentation that probes close to the core of the issues in theoretical dispute is encouraged." (Link reported here).

American Economic Review: "The AER is a general-interest economics journal. Established in 1911, the AER is among the nation's oldest and most respected scholarly journals in economics. The journal publishes 12 issues per year containing articles on a broad range of topics." (Link reported here).