MAGKS

Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 07-2016

Jochen Lüdering and Peter Winker

Forward or Backward Looking? The Economic Discourse and the Observed Reality

Coordination: Bernd Hayo • Philipps-University Marburg
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

Forward or Backward Looking? The Economic Discourse and the Observed Reality*

Jochen Lüdering[†] Peter Winker[‡]

January 2016

Is academic research anticipating economic shake-ups or merely reflecting the past? Exploiting the corpus of articles published in the Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik) for the years 1949 to 2010, this pilot study proposes a quantitative framework for addressing these questions. The framework comprises two steps. First, methods from computational linguistics are used to identify relevant topics and their relative importance over time. In particular, Latent Dirichlet Analysis is applied to the corpus after some preparatory work. Second, for some of the topics which are closely related to specific economic indicators, the developments of topic weights and indicator values are confronted in dynamic regression and VAR models. The results indicate that for some topics of interest, the discourse in the journal leads developments in the real economy, while for other topics it is the other way round.

*The authors want to thank Andreas Schiermeier, Peter Reifschneider, and Christoph Funk for assistance with obtaining and preprocessing the data. Moreover, we are thankful for the funding provided by the publisher Lucius & Lucius.

[†]Center for International Developmental and Environmental Research (ZEU), Justus-Liebig-Universität Gießen, e-Mail: jochen.luedering@zeu.uni-giessen.de Phone: (+49) 641 99 12718

[‡]Department of Statistics and Econometrics and Center for International Developmental and Environmental Research (ZEU), Justus-Liebig-Universität Gießen, e-Mail: peter.winker@wirtschaft.uni-giessen.de

1. Introduction

What drives the selection of topics in economic research? Given freedom of research in the public higher education system, the research agenda is not a result of political decision making. Thus, it might be driven by personal interest, perspectives of gains in reputation, traditions handed-down from the doctoral supervisor, networks, job perspectives, tasks in economic policy advice etc. While all listed arguments might be relevant for almost all fields of science, empirical social sciences such as economics might be subject to a further driver – reality. Following the financial crisis, we have seen a regained interest in financial market stability and credit rationing (e.g., Turner et al. 2010), in the sequel, public debt came back on the agenda (e.g., Burret, Feld, and Köhler 2013). Given the current influx of refugees in Germany, it does not appear too far-fetched to predict that the economic analysis of causes and effects of migration will see a renaissance in the near future.

The link between reality and economic research might be triggered by different mechanisms including some of the aforementioned ones. We do not strive to identify these drivers and their weight, but address a much more modest intellectual goal, namely the identification of the evolution of research interests over time and its interaction with developments in the real economy. Possibly, economists have rational expectations about future economic developments and, consequently, focus their research on topics which are to become relevant. Alternatively, they just observe the economic situation and try to explain it ex post. Besides identifying relevant research topics, our aim is to find out which direction of the links between economic science and economic reality is prevailing.

Given that the evolution and deployment of research fields takes time, such an analysis requires a sufficiently long observation period. For the quantitative research approach taken in this contribution, it implies that a long sample of data is required. Long time series on key economic indicators become increasingly available (see e.g., Rahlf 2016), but have to be treated carefully given a substantial number of structural breaks over the last 150 years.

The task of finding quantitative information about what topics economists focused over time or, at least, what their expectations might have been about key indicators, is even more challenging. Besides some business cycles indicators including qualitative information about business expectations at the individual level, such times series are not available neither at individual nor at aggregate level.

To close this gap we employ a quantitative analysis on the discourse in scientific journals in economics. We start with a pilot study on the discourse in the Journal of Economics and Statistics for the period between 1949 and 2010. Thereby, we assume that a foreseeable development, in particular one which is considered as a problem, results in an increase of scientific publications with a focus on the particular problem prior to the actual development. This holds true if economists became aware of the problem early enough. If this is not the case, publications in economic journals will only be published after the problem has occurred. Obviously, the latter also applies for shocks which are hardly predictable.

Our empirical approach comprises two stages: First, we have to identify topics discussed in the Journal of Economics and Statistics and their relative importance over time. Second, we have to establish a link between the importance attached to certain topics and the actual development of the economic reality which they might reflect. While the second step makes use of standard approaches from econometric time series analysis, the first step relies on tools from computational linguistics, which more recently, also made their way into economic analysis. However, establishing a link between topic weights and real data appears to be a novel contribution. To the best of our knowledge, only the recent contributions by Hansen, McMahon, and Prat (2014) and Larsen and Thorsrud (2015) follow a similar approach. Hansen, McMahon, and Prat analyze the impact of increased transparency on the functioning of central banks and ultimately on monetary policy using the minutes and transcripts of the Federal Open Market Committee. Larsen and Thorsrud examine the impact of "news" on the business cycle, based on a Norwegian business newspaper followed over a period of 9000 days. In contrast to our analysis, both articles consider a rather short time span and, in the case of Hansen Hansen, McMahon, and Prat (2014) are based on a narrowly defined text corpus.

Topic models are a mean to classify the content in a large text corpus. The algorithms endogenously identify so called *topics*, which are not necessarily topics in the semantic sense, but rather clusters of words which often appear jointly in a text. Our approach to topic modeling closely follows Griffiths and Steyvers (2004), who also work on a corpus of scientific literature. In their contribution, the authors introduce Gibbs sampling as an algorithm to Latent Dirichlet Analysis (LDA). They classify articles in the *Proceedings of the National Academy of Science of the United States* (PNAS) using LDA. Similarly, Hall, Jurafsky, and Manning (2008) analyze the history of ideas in the field of computational linguistics. In their paper, they provide a convincing visualization of the rise of probabilistic topic models in computational linguistics. Grün and Hornik (2011) conduct an analysis for the Journal of Statistical Software and provide the R Package *topicmodels*[1] along with programming examples. Our own implementation builds partly on their code.

In the second step of the analysis, we use the probabilities assigned to each volume of the Journal for specific topics resulting from the LDA as input for dynamic regression models. In a univariate model for explaining an economic time series related to a topic, both leads and lags of this input variable are used as potentially explanatory variables. This allows us to assess the dynamic dependencies between the relevance of a topic in papers of the Journal with the development of real economic data related to the topic. As a robustness check, we also specify and estimate bivariate VAR-models for the two variables of interest and conduct Granger causality tests on these models.

The remainder of the paper is organized as follows. First, Section 2 introduces the text corpus obtained from the Journal of Economics and Statistics and the economic time series used for the further analysis. The following Section 3 provides a short explanation of topic modeling. The application of this method and the results obtained for the Journal of

[1] See `https://cran.r-project.org/web/packages/topicmodels/index.html`

Economics and Statistics are subject of Section 4. In Section 5, the dynamic interrelation between the importance of topics and the respective economic time series is analyzed. Section 6 provides concluding remarks and an outlook for further research.

2. Text Corpus and Economic Data

2.1. Text Corpus

For the purpose of our pilot study, we concentrate on a single economic journal with a close link to economic science in Germany. This permits to assume that scientists publishing in this journal might rather focus more on Germany than on other economies, which simplifies the second step of the analysis. Furthermore, in order to identify developments over sensible time spans, a journal existing already for a long time period was required.

For these reasons, we selected the Jahrbücher für Nationalökonomie and Statistik (also Journal of Economics and Statistics). The journal has been appearing regularly since 1863, with a few exceptions (e.g. during the second World War). There have been 235 volumes to date (2015), with currently one volume being published per year. Due to the effort required for the preparation of text data used in the quantitative analysis, the analysis is restricted to the period from 1949 to 2010. It will be left for future research to include more volumes.

We had access to the scanned images of all volumes except the most recent ones via `digizeitschriften.de`. The meta data regarding the volumes considered for the present analysis are provided in Table 3 in Appendix B. For obvious reasons, the index volumes are excluded from the analysis. No volumes of the journal appeared in 1957 and 1974. Each of the two years was succeeded by a year with two volumes (1958 and 1975). For the periods 1967 to 1970 and 1971 to 1973, there were volumes covering two years. To disentangle the volumes covering multiple years, we made use of the dates on the covers of the single issues to assign the included articles to a calender year. As a consequence, volumes 181 to 183, as well as 186, 187, 191 and 192 were allotted on two years. It turned out that all issues of volume 190 appeared in 1976 and volume 188 covered a period of three years. Allocating the individual issues to calender years solved the case of missing data for 1974. Unfortunately, this did not provide a solution for the case of missing data in 1957. The observations for this year were later imputed by calculating the mean of the value for the preceding and succeeding year.

The source format differed among the volumes. From the year 2000 onward, we had access to digital publications. Older volumes were obtained only as scanned PDF files from `digizeitschriften.de`. We used *Abby Finereader 12 Corporate* to perform Optical Character Recognition (OCR) and turn the documents into text files as the quality of the already existing text files was not sufficient for the purpose of our analysis. The OCR Software retained the formatting of the headlines which we used to break the journal up into single articles which are the units of our analysis. Finally, we used manual labor to clean the texts, e.g. by removing tables, footnotes and equations.

A few further preparatory steps were necessary to come up with the final text corpus for the topic modeling algorithm. For the implementation of these steps, we closely followed Grün and Hornik (2011) by employing the text mining infrastructure supplied in the *R* package *tm*[2]. In particular, the following steps were performed:

- The German language features a large number of grammatical forms of words. Considering every single case would greatly inflate the vocabulary (the set of words which forms the basis for the application of LDA). By stemming of words the different grammatical forms of the same word are reduced to an identical stem. *"The stem is the part of the word that is common to all inflected variants"*, as wikipedia[3] puts it. We apply the stemming algorithms *SnowballC* (setting "german") to produce final word stems. The main feature is the removal of suffixes of words. Consequently,

[2]See `https://cran.r-project.org/web/packages/tm/index.html` for details about the package.
[3]`https://en.wikipedia.org/wiki/Word_stem`, retrieved December 8th, 2015.

kaufen,*kaufe*,*käufer* are all reduced to the stem *kauf*. The algorithm removes the umlaut *ä* and replaces it by the vowel *a* as umlauts are often used in forming the plural form (Example: Ball $\rightarrow$ Bälle). Unfortunately, these transformations come with a certain loss of information, which to some degree is intended (e.g. removal of plural forms) but also has unintended consequences (*fordern* (request) and *fördern* (support) become indistinguishable).

- All superfluous blanks, newline and tabulator codes, numbers and punctuation marks were removed.
- We removed all German stopwords, i.e. the most common words that would otherwise dominate most of the topics without being linked to specific content. We use the list of stopwords shown in Appendix A as supplied by the R package *tm*.
- Finally, we only considered terms which, after stemming, consisted of five to twenty characters to further reduce the size of the vocabulary. This measure also helped to remove foreign language stopwords, in particular English words (e.g. he, she, it, you) and long compounded words (e.g. *Nasenspitzenwurzelentzündung*[4]), which could potentially bias the result.

For all remaining terms, the relative importance of the term for a specific article i is calculated. This measure is named *term frequency inverse document frequency (tf-idf)* and is defined for term j as:

$$\text{tf-idf}_{j,i} = \frac{\text{frequency of term } j \text{ in the single article } (\text{tf}_i)}{\text{frequency of term } j \text{ in the whole corpus (df)}}. \qquad (1)$$

For the analysis we select all terms which are prominent in individual articles, rather than exhibiting a high background noise (high *df* value). Hence, for every term j the mean of its $\textit{tf-idf}_{j,i}$ values across all articles is calculated. Following Grün and Hornik (2011) we use the median (in our case 0.004) over all $\textit{tf-idf}_{j,i}$ values ($\forall i, j$) as a cut-off and only include

[4]Example for the "intractable problems of compound words" in German from the original description of the Snowball algorithm `http://snowball.tartarus.org/texts/germanic.html`.

terms with mean values across documents larger than this median.

As final output of these preparatory steps we obtain a document term matrix, providing the number of occurrences f_{ij} of each selected term j (column) in each article i (row). For a total of $D = 2\,675$ articles, the number of terms (word stems) considered for the further analysis is $|V| = 22\,171$. The matrix $\mathbf{F} = (f_{ij})$ serves as input for the topic modeling algorithm.

2.2. Economic Data

As will be described in more detail in Section 4, topic modeling results in a substantial number of topics. While many of them can be given an intuitive interpretation as they refer to specific economic theories or institutions, only a small number corresponds closely to some key economic indicators which are also available as long time series, and hence qualify for the second step of our quantitative approach. For this second step, we select data for the five economic issues inflation, trade (net-exports), public debt, unemployment rate and interest rates.

Inflation

The longest time series we could obtain out of these five domains is the German inflation rate. The German statistical office compiled a very long time series of the consumer price index (CPI), which, due to limited data quality for the early years including the run-up to the hyperinflation after 1918, is only available on request.[5] The price index data covers the period from 1881 to 2009. Due to the period of hyperinflation, there is no price data for 1922 and 1923. In 1948, there are separate values for the first and second half of the year, following the introduction of the D-Mark in June 1948.

[5]Statistisches Bundesamt (2013). Preise – Verbraucherpreisindex Lange Reihe von 1881.

From the price index the year by year inflation rate is calculated as

$$\text{inflation}_t = \frac{CPI_t}{CPI_{t-1}} \cdot 100 - 100\,. \tag{2}$$

Note that we make use of a traditional growth rate given that the approximation by log-differences might deviate substantially for the periods of high inflation or hyperinflation.

Figure 1 provides a plot of the inflation rate. Obviously, the plot is dominated by the period of German hyperinflation during the 1920s, which dwarfs all other spikes in inflation rates. However, this does by no means imply that inflation was not an issue at any other period in time. A second period of monetary instability followed after the second world war. Afterwards, the high inflation period in the 1970s stands out. In the early years, there have been frequent periods of deflation, which were most pronounced in the interwar period. Since the 1940s deflation has become very rare.

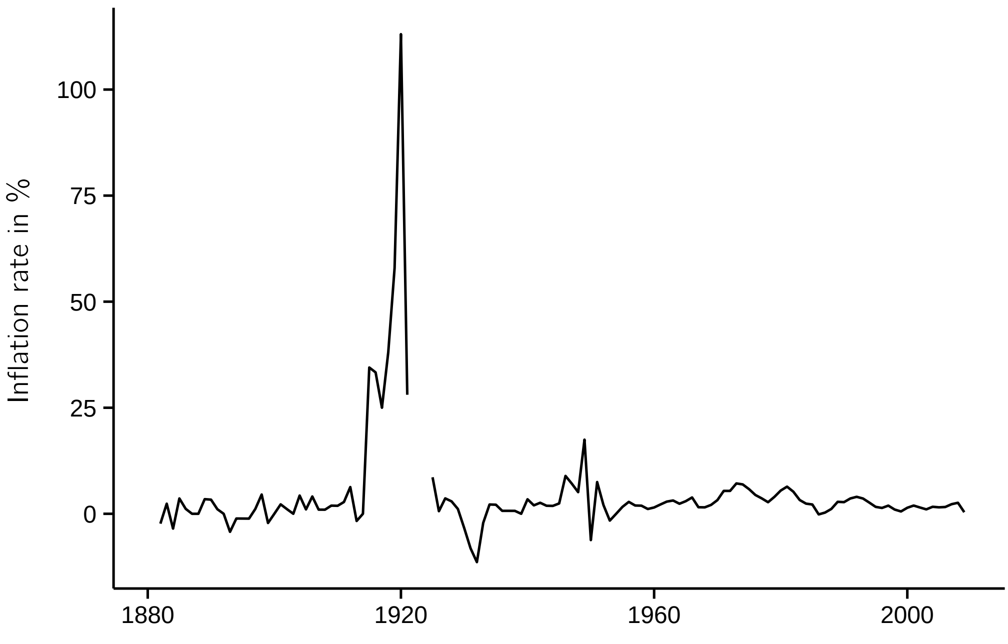


Figure 1: The German inflation rate 1881 – 2009

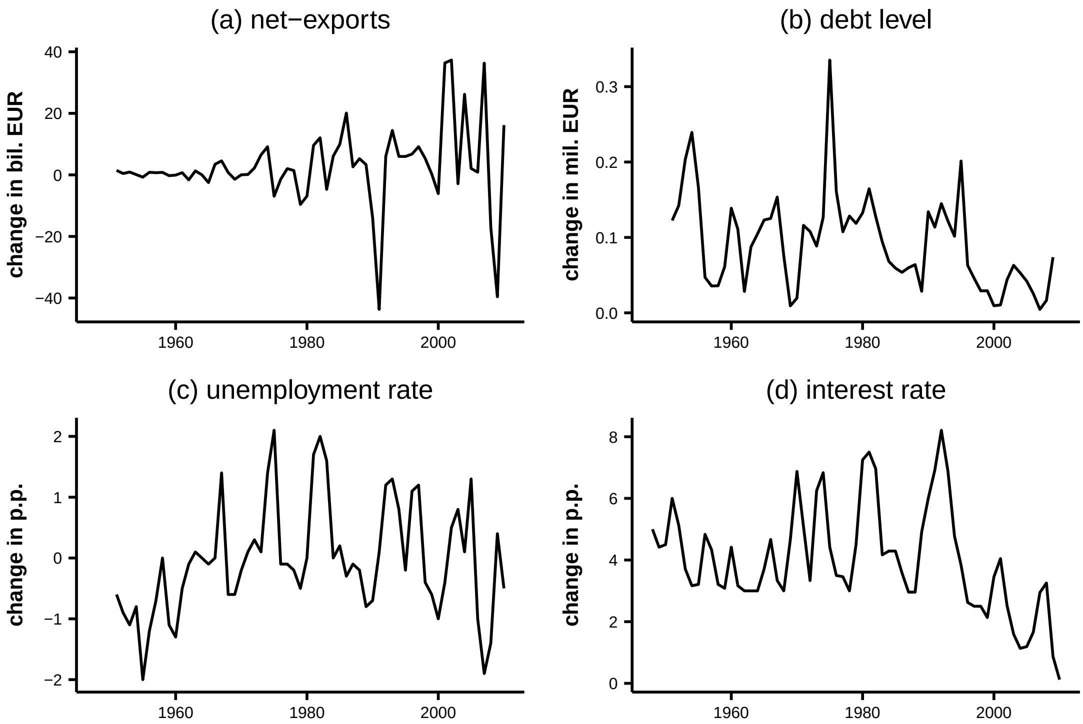


Figure 2: Real economic indicators

Trade

Trade is operationalized as the German net exports, for which data are available from the German statistical office at a yearly frequency since 1950.[6] We rescaled the original data to billion Euro in order to operate with a similar scale as for the other economic indicators. As the data series is non-stationary, the econometric analysis will be based on the differentiated series. Along with debt, the unemployment rate and the interest rates, a time series plot of net-exports is shown in Figure 2.

[6] *Außenhandel: Zusammenfassende Übersichten für den Außenhandel (Endgültige Ergebnisse)* Fachserie 7 Reihe 1, Issue 2013 from December 2014.

Debt

Data for the German public debt is available since 1950.[7] To mitigate structural breaks in the time series due to changes in the data collection methodology and the presence of non-stationarity, the nominal changes in total debt are used in this analysis.

Unemployment Rate

The German unemployment rates (in percentage terms) were retrieved from the GENESIS Database of the German Statistical Office. From 1950 to 1990, the data coverage is limited to West Germany. From 1991 onward, data for the whole of Germany are used. Since the series is non-stationary, the first differences are used for the econometric analysis.

Interest Rates

The time series of interest rates is available from *Deutsche Bundesbank*.[8] From July 1st, 1948 until 1998, the "Diskontzinssatz", which widely served as a base for financial contracts, was used. It is available at monthly frequency. With the transfer of the authority of the monetary policy to the ECB, this rate was replaced by the so called *Basiszins*. It is still available at a monthly frequency, which is only adjusted every six months. Therefore, we use the yearly average of the interest rate for the econometric analysis.

3. Topic Modeling the Economic Discourse

This section will introduce the basic idea of topic modeling in its application to the economic discourse. The method exhibits two major aspects. First, it analyzes text without imposing a priori keywords or categories. Instead, clusters of terms appearing together frequently

[7] Statistisches Bundesamt Fachserie 14 Reihe 5: Finanzen und Steuern - Schulden des Öffentlichen Gesamthaushalts Table 1.1.1.

[8] `http://www.bundesbank.de/Navigation/DE/Statistiken/Zeitreihen_Datenbanken/Makrooekonomische_Zeitreihen/makrooekonomische_zeitreihen_node.html` retrieved 01.11.2015.

(topics) emerge endogenously. Second, the method has a sound statistical background, allowing the application of standard estimation and inference procedures. We will briefly sketch the historical background of the method, the theoretical model, estimation procedures and the evaluation of modeling outcomes in this section.

3.1. Historical Development of Topic Models

Topic models originate from the field of information retrieval, where these methods were developed in order to render text electronically searchable. Early applications relied on the tf-idf (term frequency - inverse document frequency) classification, which is still used as a preprocessing step in modern approaches. The classification is based on a simple counting procedure to represent the importance of a term in a document (the term frequency) in relation to the importance of the term in the entire corpus. This method represents a text corpus comprising an arbitrary number of documents as a matrix (term-by-document matrix), with values for any given term in a vocabulary. A standard reference for these early methods is Salton and McGill (1986).

Deerwester et al. (1990) developed *Latent Semantic Analysis (LSA)*[9] as a more sophisticated method to overcome shortcomings of the tf-idf classification. *Probabilistic Latent Semantic Analysis (pLSA)* introduced a sound statistical foundation to topic modeling (Hofmann 1999). Being based on the likelihood principle, it also defines a generative model of the data (on term level). Building up on pLSA, Blei, Ng, and Jordan (2003) extended the method to *Latent Dirichlet Allocation*, by adding a probabilistic generative model at the level of the documents. In spite of several extensions made in recent years, e.g. the *correlated topic model (CTM)* (Blei and Lafferty 2007) allowing for correlation of topics across documents and the recent introduction of *TopicMapping* by Lancichinetti et al. (2015), which adds the idea that documents can be described as networks of terms, LDA remains the state of the art in topic modeling. The original estimation method by Blei,

[9] In the context information retrieval, the method is often called Latent Semantic Indexing (LSI).

Ng, and Jordan (*Variational Expectation Maximization*), was rather slow, which proved to be problematic in some areas (e.g. commercial applications in information technology). As a consequence, Bayesian methods as suggested by Griffiths and Steyvers (2004) became widely used.

3.2. The Theoretical Model of LDA

At the core of LDA lies an abstract theoretical model which describes how documents are created. Thereby, it is assumed that a document is "a mixed bag of words", which implies that the order of words within a document is ignored and just the frequencies of words are considered. In practice, the documents are obtained from machine readable sources (either short as a tweet or long like a journal article) and are usually altered by a preparatory data cleaning step.[10] Each document is assumed to be made up from several topics which determine the probability of each term from the vocabulary to be included in the document.

To be more precise about this data generating mechanism, we introduce some notation. First, we have a *vocabulary* V comprising all terms considered in the analysis. The size of this vocabulary is denoted by $|V|$. Each document is given by a vector of terms $\boldsymbol{w} = (w_1, \ldots, w_N)$, where N denotes the length of the document. We do without a document specific index, as the following discussion will focus on a single document. Only in the final step, the results for single documents will be aggregated for the corpus comprising all documents.

It is assumed that each term in the text has its origin in some topic. Thereby, topic k, $k = 1, \ldots, K$ is represented by a vector of probabilities $\boldsymbol{\beta_k} = (\beta_{k,1}, \ldots, \beta_{k,|V|})$ assigned to each term in the corpus, i.e. it is characterized by some terms being more frequent than others. For example, if the corpus includes the terms "inflation", "debt" and "growth", a topic with probabilities $(0.9, 0.05, 0.05)$ might be associated with inflation, while a topic with probabilities $(0.05, 0.6, 0.345)$ might be focused on the nexus between public debt and

[10] See Subsection 2.1 for a description of this preprocessing step for the current application.

growth. The $K \times |V|$ matrix resulting from stacking all $\boldsymbol{\beta_k}$ vectors is denoted as $\boldsymbol{\beta}$. The vector $\boldsymbol{z} = (z_1, \ldots, z_N)$ denotes the vector of topics giving rise to the terms in $\boldsymbol{w}$. All z_n, $n = 1, \ldots, N$ come from the set of all topics of size K, which is the same for all documents in the corpus. Typically, the z_n are not all different, but rather concentrate on a few topics for a specific document. Furthermore, also the assignment of a term to a topic is not unique as the same term might belong to several topics.

Now, the generative process of a document in the LDA model can be described as follows (see also Grün and Hornik (2011)): First, a categorial probability distribution $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ is randomly chosen which describes the relevance of topics within the document.[11] In particular, for all $k = 1, \ldots, K$, $\theta_k \in [0, 1]$ and $\sum_{k=1}^{K} \theta_k = 1$. Next, for each term w_n, $n = 1, \ldots, N$ in the document, a single topic z_n is randomly selected according to the probability distribution $\boldsymbol{\theta}$. Then, according to the probability distribution on the terms of z_n, i.e. $\boldsymbol{\beta}_{z_n}$, the term w_n is drawn.

In conclusion, given all topics and the probabilities of these topics for a document, the random process of generating the document can be described. However, the aim of the analysis is rather the reverse: Only the documents forming the corpus and, consequently, the vocabulary are available, while we are interested in identifying topics and, for each document, the relevance (probability) of each topic. How this can be achieved by adding some assumptions on the generative process is described in the following subsection.

3.3. Estimation of LDA models

The theoretical model presented in the previous subsection requires a substantial number of parameters, namely $K \cdot |V|$ probabilities $\beta_{k,i}$ and the number of documents times K probabilities θ_k. Given these parameters, the probability of the observed documents can be calculated. However, as in a standard maximum likelihood setting, we are interested

[11] In the literature on LDA modeling, this distribution is often labeled as a multinomial distribution which is adequate assuming just the outcome of one draw.

in “reversing” the argument and obtaining estimates of the parameters given the observed documents, i.e., we are searching those parameter settings making it most likely to observe our documents and, consequently, determine topics and their relevance for individual documents endogenously. Given the number of parameters and the functional interdependencies between the $\beta_{k,i}$ and the θ_k, it turns out that a straightforward maximum likelihood approach is not feasible without imposing additional constraints and, possibly, using alternative optimization/estimation procedures (Griffiths and Steyvers 2004, p. 5229).

A first simplifying assumption consists in considering the categorial distribution $\boldsymbol{\theta}$ as a random draw from a uniform Dirichlet distribution with scaling parameter α, i.e., $\boldsymbol{\theta} \sim \mathrm{Dir}(\alpha)$ (Blei, Ng, and Jordan 2003). Then, for given parameters α and $\boldsymbol{\beta}$, the probability of observing a specific $\boldsymbol{\theta}$, a set of N terms $\boldsymbol{w}$ and corresponding topics $\boldsymbol{z}$ is given by (Blei, Ng, and Jordan 2003, p. 996):

$$p(\boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{z}) = p(\boldsymbol{\theta}|\alpha) \prod_{n=1}^{N} p(w_n|z_n, \boldsymbol{\beta}) p(z_n|\boldsymbol{\theta})\,. \tag{3}$$

Integrating over the random vector $\boldsymbol{\theta}$ and summing over the components of $\boldsymbol{z}$ results in the marginal distribution for a single document. Finally, by calculating the product of the marginal properties of all documents of the corpus, the probability of the corpus is obtained. Despite the simplification by considering $\boldsymbol{\theta}$ as a random draw from $\mathrm{Dir}(\alpha)$, maximum likelihood estimation still does not appear to be feasible Griffiths and Steyvers (2004, p 5229).

To overcome this problem, a variety of (approximate) estimation procedures have been suggested. The original procedure is a variant of the *expectation maximization* (EM) algorithm, the so called *variational expectation maximization* (VEM), which is the method suggested by Blei, Ng, and Jordan (2003) when introducing LDA. Shortly afterwards, Griffiths and Steyvers (2004) proposed the use of Gibbs sampling, which – according to the authors – exhibits faster convergence. This method has become widely applied and is used for our empirical application. For a comparison of the two methods see Welling, Teh, and Kappen (2008).

The method requires using a further assumption regarding the data generating process. It is assumed that the term distribution is a random draw from a Dirichlet distribution with parameter δ. Using this assumption, the probability for a single document according to Equation (3) could be obtained by integrating out β and θ, which might be done separately as β appears only in the first and θ in the second term. As is shown by Griffiths and Steyvers (2004, p 5229), the resulting expressions still do not allow for a direct calculation. Therefore, they propose to apply a Markov Chain Monte Carlo approach. They provide details on how the probability for each topic is updated based on the distribution for all other topics. In a way, the Markov chain is constructed to converge to the target distribution by repeated sampling from the target distribution. After the Markov chain has converged, the predictive term (β) and topic (θ) distributions can be obtained.

3.4. Model Validation

There are two approaches to compare topic models with respect to the choice of the number of topics. One approach focuses on fitting the model on a subset (e.g. 90%) of data and evaluating the fit for the remaining data. However, Chang et al. (2009) argue that comparing this approach does not result in models which are appealing to human judgment.

A more convenient method is introduced by Griffiths and Steyvers (2004, p. 5231). Ideally one would compare the different models based on the likelihood as a function of the number of topics K, which involves summing over all possible assignments of words to topics. Griffiths and Steyvers (2004, p. 5231) circumvent the resulting computational issue by approximating the likelihood by the harmonic mean of a set of values which are calculated from samples provided by the Gibbs sampling algorithm. For reasons of completeness it shall be mentioned that there is also some criticism regarding this method of approximation.

See Buntine (2009), as well as Wallach et al. (2009) for an overview on the critique and alternative approximation methods.

While yielding good results in our analysis, selecting a model by maximization of the likelihood occasionally leads to an unreasonable large number of topics, which become hard to interpret. Consequently, authors sometimes deviate from the estimated number of topics according to the likelihood approach to allow for a more straightforward interpretation of the topics. Nonetheless, in our case estimating the number of topics according to this method results in topics allowing for a meaningful interpretation and stable results.

4. Taking LDA to the Data

Due to the size of the data set we estimated the topics using Gibbs sampling. Repeating the estimation for a number of topics (K) between 2 and 1000, we found K=165 to be the optimal choice based on the harmonic mean method. Finding an optimal number of topics through maximization is subject to some difficulties as the function is not smooth. It shall be noted that the original application by Griffiths and Steyvers (2004) was aimed at providing a rough estimate of the magnitude of the value for K.

Apart from K, some parameters had to be chosen a priori. We stick to $\alpha = 1/K$ and $\delta = 0.1$, chosen according to the literature (Griffiths and Steyvers 2004). Afterwards the Markov chain is run for 2000 iterations, which we, following Grün and Hornik (2011, p 10), assume sufficient for it to converge. From the resulting 165 topics we select five topics that, to our understanding, are the ones most closely related to the economic indicators introduced in Section 2.2. These topics are presented in Figure 3, where the font size of the terms indicate their relative importance within a topic. The discussion of trade (topic 1) is the only topic that is primarily discussed using the English language, which might be explained to some degree by the international interest in trade itself. The German language equivalent of the topic can be found in Appendix D. The topic concerned with debt (topic 22) appears to be centered on loans given to companies and individuals. The terms describing sovereign debt (e.g. *Staatsschulden*) are part of the topic but do not show up with the same frequency. The stem *arbeitslos* is the single most significant term in topic 56 (unemployment), all other somewhat significant words are compound words closely related to employment. Topic 144 is concerned with the discussion of inflation and the inflation rate, also the term *Phillipskurve* shows up prominently describing a theoretical framework in which inflation is often discussed. The fifth and last topic considered is based around the term *zinssatz* [en: interest rate]. The discussion also encompasses finance (*geldmarkt* [en: money market]) and macroeconomics (Preisniveau [en: price level], Liquiditätsfalle [en: liquidity trap]).

Even though Figure 3 shows only a small subset of the 165 topics, each of the topics can be attributed to a particular idea or debate in economics. Appendix D shows additional topics which appear to be closely related to the ones used here. Any further attempt to derive the stories behind these topics in greater detail should also involve qualitative analysis, i.e. a careful reading of those documents exhibiting high probabilities for the topic of interest.

5. The Relationship Between Discourse and Economic Data

5.1. Univariate Dynamic Model

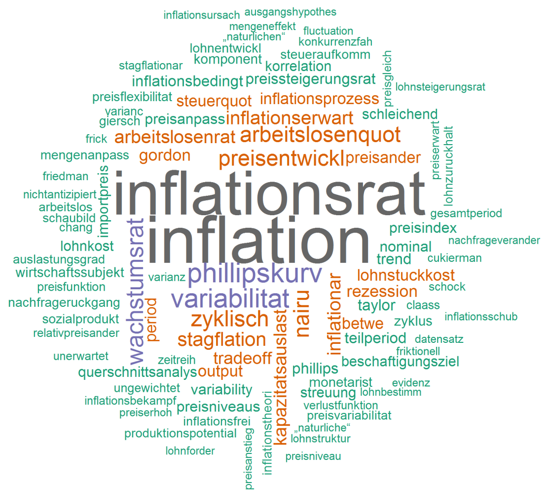
The relationship between a real economic indicator and the logarithm of the sum over probabilities for the corresponding topic of all documents in a given year is estimated by linear regression models. The results are shown in Table 1. The augmented distributed lag model includes both leads and lags of the topic indicator as explanatory variables for the current value of the economic indicator. Statistically significant parameters for lagged values indicate that the scientific discussion on the topic precedes changes in the economic variable, while statistically significant leads point at a scientific discussion following the

(a) Topic 1 (Trade)

(b) Topic 22 (Debt)

(c) Topic 56 (Unemployment)

(d) Topic 144 (Inflation)



(e) Topic 161 (Interest Rate)

Figure 3: Identified key topics

developments of economic indicators. The model selection procedure considers all models with lagged values up to three years and leading values for up to three years. From all possible 1024 subsets of these potential explanatory variables, the selected model is the one minimizing the Akaike Information Criteria (AIC).

Table 1 summarizes the estimation results. Each column provides the estimated coefficients for a particular economic variable exhibited in the row labeled "dependent variable", while the number of the corresponding topic is shown in the next row. The row labeled topic without a number corresponds to the parameter for the instantaneous effect. For the following rows, positive numbers in parentheses indicate leads, while negative numbers indicate lags. We also allow for lagged values of the endogenous variables and a deterministic linear trend.

For all models, we find a link between the importance of the topic in the scientific discussion and the observed economic indicator with the adjusted R^2 ranging from 0.23 to 0.61. However, it turns out that there are differences across models with respect to the role of lagged, leading and contemporaneous effects. The use of logarithms for the topic probabilities allows us to interpret the coefficients[12] β as semi-elasticities: A 1% percent increase in the discussion on the topic is associated in a change by $0.01 \times \beta$ units of the indicator.

One needs to be careful when interpreting the time lag as indicated by the regression analysis. While the analysis treats the time of publication and the time of writing as identical, in reality there is, in most cases, a notable publication lag between those two dates.

Without taking into account a potential publication lag, the results of the dynamic regression models in Table 1 allow for the following conclusions. More discussion of the topic related to inflation in the past has no significant effect, and contemporaneous discussion appears even negatively related to the actual inflation rate. In contrast, the link between

[12]Note: We use β by convention, it is not to be confused with the parameter of the same name from the LDA model.

Table 1: Regression results

Number	(1)	(2)	(3)	(4)	(5)
Dependent Var.	Inflation	d(NetExp)	d(debt)	d(unemp)	d(interest)
Topic	Topic 144	Topic 1	Topic 22	Topic 56	Topic 161
Constant	3.4544***	−86.9928**	0.2761***	0.8935	−0.1028
	(1.2546)	(40.197)	(0.0478)	(0.7827)	(0.8609)
Topic	−0.3077*	−5.7066**		0.2593**	
	(0.1778)	(2.4449)		(0.1218)	
Topic (1)		−5.6708**	0.0102		0.2587*
		(2.1887)	(0.0073)		(0.1455)
Topic (2)				0.3052**	0.1977
				(0.1172)	(0.1452)
Topic (3)	0.5810***	−0.7422			
	(0.1807)	(2.1619)			
Topic (-1)				−0.3956**	−0.2402
				(0.1247)	(0.1598)
Topic (-2)		5.2182**	0.0208***		
		(2.5121)	(0.0077)		
Topic (-3)	0.1385				−0.2309
	(0.1675)				(0.1514)
Endogenous (-1)	0.5653***		0.4146***	0.7148***	0.1874
	(0.0971)		(0.1195)	(0.1221)	(0.1164)
Endogenous (-2)			0.1402	−0.3032**	−0.4741***
			(0.1200)	(0.1297)	(0.1250)
lin. Trend		0.5482***			
		(0.2024)			
adj R^2	0.606	0.2275	0.3010	0.4505	0.2635
F	22.174	4.299	6.9225	10.0214	4.3407
P(F)	0.0	0.002	0.0001	0.0000	0.0013
N	56	57	56	56	57

Note: standard errors in parenthesis
*, ** and *** indicate significance at 10%, 5% and 1% levels respectively
All topic probabilities in logarithms

an increase in inflation and future discussion of the topic is statistically significant, but the absolute size of the effect is moderate: An increase in current inflation by 0.05810 percentage points would correspond to an increase of the topic weight by 10% three years later. For net-exports, a statistically significant positive effect of past discussion of the topic on current values is found, while a change in net exports today rather seems to reduce future discussion of the topic as indicated by the statistically significant negative coefficient for Topic (1). For debt, only a statistically significant positive effect of past discussion of the topic (2 years ago) with its actual change is found, while for unemployment again links in both directions are found. While past discussion appears to be negatively correlated with the unemployment rate, the nexus becomes positive for the discussion in the future. There is only a weak link between an increase in the interest rate and an increase of the relevance of the corresponding topic in the future.

5.2. VAR-Model

As a robustness check for our empirical results, we also estimate VAR-models, which allow conducting tests for Granger causality (Lütkepohl 2007, pp 102f). A major advantage of this model class is that both variables under consideration are treated as endogenous, while all explanatory variables are lagged values of these endogenous variables. On the downside, the model does not allow for an explicit modeling of contemporaneous dependencies, which only show up through a correlation of error terms.

For the VAR model, we do not consider holes in the lag structure (Winker 2000) to check whether the subset selection procedure applied for the dynamic model in Subsection 5.1 has a qualitative impact on the findings. The lag length for the VAR model is selected based on the AIC with maximum lag length of six years. The lag lengths used for the VAR models for the different variables are reported in the last row of Table 2.

The VAR model and the corresponding Granger causality tests are illustrated using the first pair of variables from our application, i.e. inflation rate (infl_t and the weight of topic 144 (t144_t) over time. The current values of both, the inflation rate and the topic weight are modeled as depending on their own past values and the past values of the other variable. Given that the optimum lag length according to AIC for this pair is three years, the VAR model is given by:

$$\text{infl}_t = \alpha_{1,1}\text{t144}_{t-1} + \ldots + \alpha_{1,3}\text{t144}_{t-3} + \alpha_{1,4}\text{infl}_{t-1} + \ldots + \alpha_{1,6}\text{infl}_{t-3} + \varepsilon_{1,t} \quad (4)$$

$$\text{t144}_t = \alpha_{2,1}\text{t144}_{t-1} + \ldots + \alpha_{2,3}\text{t144}_{t-3} + \alpha_{2,4}\text{infl}_{t-1} + \ldots + \alpha_{2,6}\text{infl}_{t-3} + \varepsilon_{2,t} \quad (5)$$

Given that the explanatory variables are the same for both equations, the system of equations can be estimated by conducting a simple OLS regression for each equation separately (Lütkepohl 2007, p 72). The estimated parameters jointly reflect the intertemporal dependencies between the two variables. Therefore, considering individual parameters and testing their statistical significance is not very informative. Instead, the idea of Granger causality is to check whether the lagged values of the other variable have jointly a statistical significant influence beyond the dynamics already reflected by the lagged values of the endogenous variable itself. For example, in equation (4), testing the null hypothesis that $\alpha_{1,1}$, $\alpha_{1,2}$ and $\alpha_{1,3}$ are all zero corresponds to the statement that the past development of the topic weights has no additional explanatory power for the current inflation rate going beyond the information already contained in the past development of inflation rate itself (i.e., in the parameters $\alpha_{1,4}$, $\alpha_{1,5}$ and $\alpha_{1,6}$). This null hypothesis is labeled as "topic 144 is not Granger causal for inflation". It is tested by means of a Wald test. The test statistics asymptotically follows a χ^2-distribution with the number of degrees of freedom corresponding to the lag length of the selected model. Table 2 provides the test statistics and the marginal p-values for all variable pairs and both directions of potential Granger causality.

Overall, the results of Granger causality testing are consistent with those found for the single equation models with the exception of net exports. Let us consider the models one by one.

Table 2: Test for Granger Causality

	(1)	(2)	(3)	(4)	(5)
Econ. Var.	inflation	d(netexp)	d(debt)	d(unemp)	d(interest)
Topic	144	1	22	56	161
			topic → reality		
χ^2	3.0901	2.4045	11.416	6.2377	0.3179
p-value	0.3779	0.3005	0.0096	0.0442	0.8530
			topic ← reality		
χ^2	9.4342	1.4324	0.882	2.5116	4.9948
p-value	0.0240	0.4885	0.8297	0.2848	0.0822
Lag length (max. 6)	3	2	3	2	2

For the link between the inflation rate and topic 144, we find a lag length of three which is identical to the one in the single equation setting. We cannot reject the null hypothesis that scientific discussion (i.e. topic 144) does not Granger cause the inflation rate, while the reverse hypothesis has to be rejected at the 5% level. This implies that scientific discussion is affected by developments in the inflation rate, but not vice versa.

The model for the change of net exports and topic 1 differs from the single equation model by a lag length of two instead of three. Furthermore, for the single equation model, a deterministic time trend and the current value of the topic weight were found to be influential. This contemporaneous effect is reflected in a high correlation of error terms between the two equations of the VAR model, but does not affect the Granger causality test which might explain the missing evidence for a link in both directions. The VAR model for the change of debt includes three lags, while the single equation model exhibited only a maximum of two leads and lags. Nevertheless, again a significant impact of past discussion in economic science of topic 22 on current changes in debt level in the sense of Granger causality is found, while the actual development of debt does not exhibit Granger causality on the discourse. For unemployment, both models suggest a lag length of two. While the single equation model suggests dependencies in both directions, the null hypothesis of no Granger causality could be rejected at the 5% level only for the influence of past discussion regarding the topic on current changes of unemployment. Finally, for the interest rate, the VAR model comprises only two lags, while the maximum lag and lead length found in the single equation model was three. Nevertheless, the qualitative findings are again similar. While the past discourse on the interest topic is not Granger causal for the current changes in interest rates, the null hypothesis of no Granger causality running from interest rate changes to the extent of scientific discussion about interest rates is rejected at the 10% level though not at the 5% level.

6. Conclusion

It was demonstrated, how Latent Dirichlet Analysis of scientific publications in economics and actual economic developments can be put in perspective. To this end, first the corpus of articles published in the Journal of Economics and Statistics between 1949 and 2010 was analyzed by means of the Latent Dirichlet Analysis resulting in an endogenously created list of relevant topics. Most of the topics found make sense from a semantic point of view and a substantial part of them can be given an immediate interpretation related to economic theory, economic institutions or developments of economic variables.

The second step of the analysis concentrated on those topics which are found to be closely linked to economic indicators available for a long enough time period to allow for a dynamic econometric modeling at annual frequency. This econometric analysis was conducted both with single equation models allowing for lags and leads of the topic weights to have an impact on the current value of the economic variable of interest as well as with VAR models. For all five variables under consideration (inflation rate, net-exports, debt, unemployment rate and interest rate), a relevant – and mostly also statistically significant – link between scientific discussion in the journal and real developments could be found. However, the direction of the influence along the time dimension is not uniform across the models. While a lead of economic discussion with respect to the realization of variables is found for debt and unemployment, the temporal dependency is the other way round for inflation and interest rates, while the dependency appears to be most pronounced contemporaneously for trade.

While the proposed two-stage quantitative approach appears promising as an additional method for analyzing the development of economic thought over time, it will have to be extended in future work in various directions. First, the constraint on a single journal caused by available resources for digitalization, text recognition and data preparation has to be overcome by extending the analysis to other journals being published over a long period. Second, although the application of a specific implementation of LDA using Gibbs sampling for the estimation works well for the present corpus, it has been reported in the literature that the robustness of these methods is limited. Therefore, further research should be devoted on improving the modeling and estimation procedure. Third, our econometric analysis at the second step of the analysis does not take into account the fact that the topic weights are generated data, which might have an impact on the inference in the second step. It does not appear obvious to us, however, how the uncertainty from the first step might be modeled statistically without using a bootstrap approach which, however, does not appear to be feasible with available computational resources given the high computational complexity of the first step. Finally, and most importantly, the purely quantitative approach used here does not represent a substitute to classical hermeneutic analysis, it rather provides a complementary method to detect relevant fields of research (topics) and how they developed over time putting them in perspective to real economic developments.

References

Blei, David M. and John D. Lafferty (2007). "A correlated topic model of Science". In: *Annals of Applied Statistics* 1.1, pp. 17–35.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning Research* 3, pp. 993–1022.

Buntine, Wray (2009). "Estimating Likelihoods for Topic Models". In: *Advances in Machine Learning*. Vol. 5828. Lecture Notes in Computer Science, pp. 51–64.

Burret, Heiko T., Lars P. Feld, and Ekkehard A. Köhler (2013). "Sustainability of Public Debt in Germany – Historical Considerations and Time Series Evidence". In: *Journal of Economics and Statistics (Jahrbücher fuer Nationalökonomie und Statistik)* 233.3, pp. 291–335.

Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei (2009). "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Neural Information Processing Systems*.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman (1990). "Indexing by latent semantic analysis". In: *Journal of the American Society for Information Science* 41.6, pp. 391–407.

Griffiths, Thomas L. and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5228–5235.

Grün, Bettina and Kurt Hornik (2011). "topicmodels: An R Package for Fitting Topic Models". In: *Journal of Statistical Software* 40.13, pp. 1–30.

Hall, David, Daniel Jurafsky, and Christopher D. Manning (2008). "Studying the History of Ideas Using Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08, pp. 363–371.

Hansen, Stephen, Michael McMahon, and Andrea Prat (2014). *Transparency and Deliberation within the FOMC: A Computational Linguistics Approach*. CEP Discussion Papers dp1276. Centre for Economic Performance.

Hofmann, Thomas (1999). "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 50–57.

Lancichinetti, Andrea, M. Irmak Sirer, Jane X. Wang, Daniel Acuna, Konrad Körding, and Luís A. N. Amaral (2015). "High-Reproducibility and High-Accuracy Method for Automated Topic Classification". In: *Physical Review X* 5.011007.

Larsen, Vegard H. and Leif Anders Thorsrud (2015). *The Value of News*. Working Papers 0034. Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

Lütkepohl, Helmut (2007). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

Rahlf, Thomas (2016). "The German Time Series Dataset 1834–2012". In: *Journal of Economics and Statistics (Jahrbücher fuer Nationalökonomie und Statistik)* 236.1. *forthcoming*.

Salton, Gerard and Michael J. McGill (1986). *Introduction to Modern Information Retrieval*. New York, USA: McGraw-Hill.

Turner, Adair et al. (2010). *The Future of Finance: The LSE Report*. London School of Economics & Political Science. URL: `https://harr123et.wordpress.com/`.

Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov, and David Mimno (2009). "Evaluation Methods for Topic Models". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09, pp. 1105–1112.

Welling, Max, Yee W. Teh, and Bert Kappen (2008). "Hybrid variational/MCMC inference in Bayesian networks". In: *Proceedings on the 24th Conference on Uncertainty in Artificial Intelligence*.

Winker, Peter (2000). “Optimized Multivariate Lag Structure Selection”. In: *Computational Economics* 16.1-2, pp. 87–103.

Appendices

A. German stopwords

The following *stopwords* are removed from the vocabulary. The list is supplied by the r package *tm*.

aber alle allem allen aller alles als also am an ander andere anderem anderen anderer anderes anderm andern anderr anders auch auf aus bei bin bis bist da damit dann der den des dem die das daß derselbe derselben denselben desselben demselben dieselbe dieselben dasselbe dazu dein deine deinem deinen deiner deines denn derer dessen dich dir du dies diese diesem diesen dieser dieses doch dort durch ein eine einem einen einer eines einig einige einigem einigen einiger einiges einmal er ihn ihm es etwas euer eure eurem euren eurer eures für gegen gewesen hab habe haben hat hatte hatten hier hin hinter ich mich mir ihr ihre ihrem ihren ihrer ihres euch im in indem ins ist jede jedem jeden jeder jedes jene jenem jenen jener jenes jetzt kann kein keine keinem keinen keiner keines können könnte machen man manche manchem manchen mancher manches mein meine meinem meinen meiner meines mit muss musste nach nicht nichts noch nun nur ob oder ohne sehr sein seine seinem seinen seiner seines selbst sich sie ihnen sind so solche solchem solchen solcher solches soll sollte sondern sonst über um und uns unse unsem unsen unser unses unter viel vom von vor während war waren warst was weg weil weiter welche welchem welchen welcher welches wenn werde werden wie wieder will wir wird wirst wo wollen wollte würde würden zu zum zur zwar zwischen

B. Tables

Table 3: List of volumes

Vol	Year	Vol	Year	Vol	Year	Vol	Year	Vol	Year	Vol	Year	Vol	Year
1	1863	38	1882	74	1900	111	1918	147	1938	184	1970	221	2001
2	1864	39	1882	75	1900	112	1919	148	1938	185	1971	222	2002
3	1864	40	1883	76	1901	113	1919	149	1939	186	71/72	223	2003
4	1865	41	1883	77	1901	114	1920	150	1939	187	72/73	224	2004
5	1865	42	1884	78	1902	115	1920	151	1940	188	1975	225	2005
6	1866	43	1884	79	1902	116	1921	152	1940	189	1975	226	2006
7	1866	44	1885	80	1903	117	1921	153	1941	190	~~75/~~76	227	2007
8	1867	45	1885	81	1903	118	1922	154	1941	191	76/77	228	2008
9	1867	46	1886	82	1904	119	1922	155	1942	192	77/78	229	2009
10	1868	47	1886	83	1904	120	1923	156	1942	193	1978	230	2010
11	1868	48	1887	84	1905	121	1923	157	1943	194	1979		
12	1869	49	1887	85	1905	122	1924	158	1943	195	1980		
13	1869	50	1888	86	1906	123	1925	159	1944	196	1981		
14	1870	51	1888	87	1906	124	1926	160	1944	197	1982		
15	1870	*r*	*1888*	88	1907	125	1926	161	1949	198	1983		
16	1871	52	1889	89	1907	126	1927	162	1950	199	1984		
17	1871	53	1889	90	1908	127	1927	163	1951	200	1985		
18	1872	54	1890	91	1908	128	1928	164	1952	201	1986		
19	1872	55	1890	92	1909	129	1928	165	1953	*202^r*	*1986*		
20	1873	56	1891	93	1909	130	1929	166	1954	203	1987		
21	1873	57	1891	94	1910	131	1929	167	1955	204	1988		
22	1874	58	1892	95	1910	132	1930	168	1956	205	1988		
23	1874	59	1892	96	1911	133	1930	169	1958	206	1989		
24	1875	60	1893	97	1911	134	1931	170	1958	207	1990		
25	1875	61	1893	98	1912	135	1931	171	1959	208	1991		
26	1876	62	1894	99	1912	*r*	*1931*	172	1960	209	1992		
27	1876	63	1894	100	1913	136	1932	173	1961	210	1992		
28	1877	64	1895	101	1913	137	1932	174	1962	211	1993		
29	1877	65	1895	102	1914	138	1933	175	1963	212	1993		
30	1878	66	1896	103	1914	139	1933	176	1964	213	1994		
31	1878	67	1896	104	1915	140	1934	177	1965	214	1995		
32	1879	68	1897	105	1915	141	1935	178	1965	215	1996		
33	1879	69	1897	106	1916	142	1935	179	1966	216	1997		
34	1879	70	1898	107	1916	143	1936	180	1967	217	1998		
35	1880	71	1898	108	1917	144	1936	181	67/68	218	1999		
36	1881	72	1899	109	1917	145	1937	182	68/69	219	1999		
37	1881	73	1899	110	1918	146	1937	183	69/70	220	2000		

Note: The volumes marked *r* are index volumes, only 202 carries a volume number.

Notes on the list of volumes

181 Issue 4 was the first to appear in 1968 (March)

182 Issue 4–5 was the first to appear in 1969 (March)

183 Issue 5 was the first to appear in 1970 (February)

184 completely appeared in 1970

185 completely appeared in 1971

186 Issue 3 was the first to appear in 1972 (February)

187 Issue 2 was the first to appear in 1973 (January)

188 Issue 1 Appeared in 1973 (December), Issue 2–5 appeared in 1974 (January to November), Issue 6 appeared in 1975 (February)

190 All issues appeared in 1976 (contrary to the available meta data)

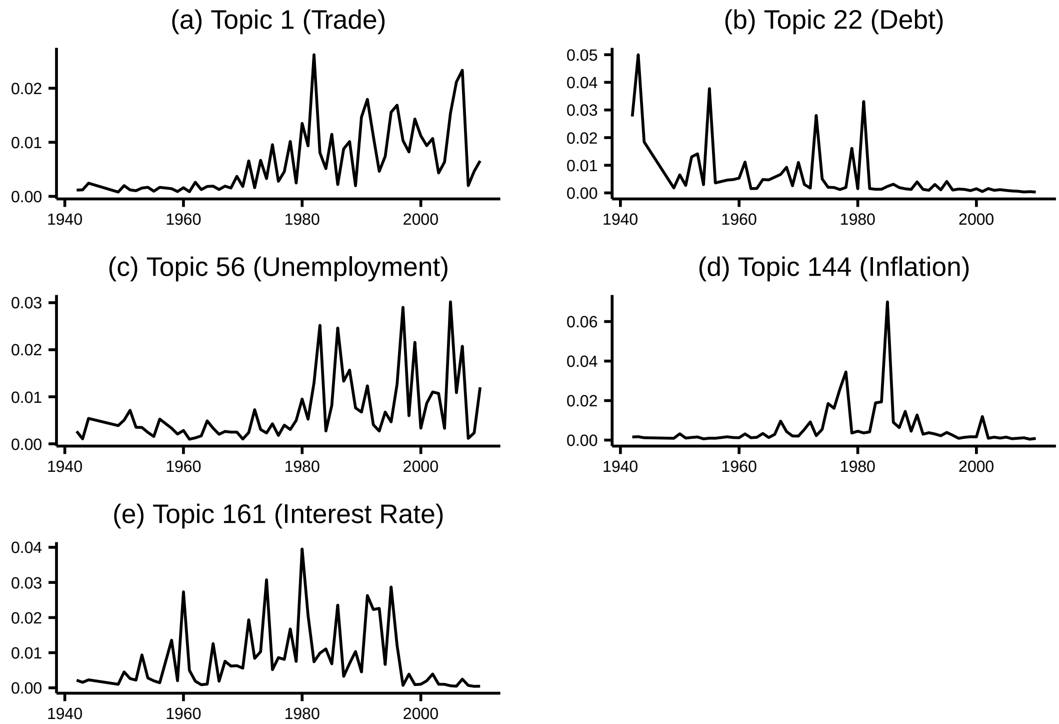
191 Issue 4 was the first to appear in 1977 (February)

192 Issue 5 was the first to appear in 1978

C. Topic Probabilities

The Figure 4 shows the development of probabilities for the key topics between 1948 and 2010.

Figure 4: Topic probabilities



D. Further Topics

The following pages show additional topics identified by the LDA algorithm. In addition to the key topics used in the analysis, there are further topics in the field of inflation (Figure 5), trade (Figure 6), debt (Figure 7) unemployment (Figure 8) and interest rates (Figure 9). This list of of fields is far from being exhaustive. There are a variety of other topics discussed in the journal (see examples in Figure 10), which are not easily operationalized as the discussion of capitalism and Marxism (Topic 100) or may not very interesting from an economic point of view (e.g. “terms describing a table” in Topic 165).

While Topic 144, which we used in the analysis, is narrowly focused on inflation and the inflation rate, there are further topics related to inflation (Figure 5), Topic 119 is concerned with *geldpoliti* [en: monetary policy], as well as money supply and expansionary policy. Topic 134 is concerned with shocks, with inflation being a prominent term. Topic 142 is the English language equivalent to Topic 119 (monetary policy). Figure 6 shows further topics associated with international trade. The German equivalent (topic 36) to the topic we selected (Topic 1) is centered around “ausland” and “inland” [en: foreign and domestic] and not as narrow as the english original. Topic 44 is loosely concerned with trade, with terms “handelspoliti” [en: trade policy] and “aussenhandelstheori” [en: theory of international trade] popping into the eye. Price differentiation [ger: preisdifferenzier], product [ger: erzeugnis] as well as terms relating to foreign and domestic are at the center of topic 86. Figure 7 and Figure 8 show additional topics related to debt and unemployment respectively. Apart from topic 191, which is concerned with interest rates in the narrow sense and consequently used in our analysis, only Topic 120 (Figure 9) appears to be somewhat related but talks more about central banking.

In the regression analysis it would be possible to combine two or more topics, which makes the analysis broader. Prior research has shown that this does not improve our results. It can be assumed that narrow topics are best at reflecting narrow economic ideas.

(a) Topic 119

(b) Topic 134

(c) Topic 142

Figure 5: Estimated topics related to inflation

(a) Topic 36

(b) Topic 44

(c) Topic 86

Figure 6: Estimated topics related to trade

(a) Topic 48

(b) Topic 112

(c) Topic 156

Figure 7: Estimated topics related to debt

(a) Topic 76

(b) Topic 104

(c) Topic 105

(d) Topic 124

Figure 8: Estimated topics related to unemployment

(a) Topic 120

Figure 9: Addtional estimated topic related to interest rates

(a) Topic 100

(b) Topic 165

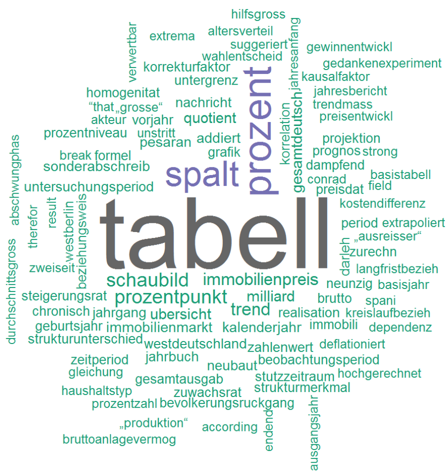


Figure 10: Example for “unrelated topics” estimated by the algorithm