Bergamini, Enrico; Mourlon-Druol, Emmanuel

Working Paper

Talking about Europe: Exploring 70 years of news archives

Bruegel Working Paper, No. 04/2021

Provided in Cooperation with:
Bruegel, Brussels

Suggested Citation: Bergamini, Enrico; Mourlon-Druol, Emmanuel (2021) : Talking about Europe: Exploring 70 years of news archives, Bruegel Working Paper, No. 04/2021, Bruegel, Brussels

This Version is available at:
https://hdl.handle.net/10419/237626

WORKING PAPER | ISSUE 04/2021 | 2 MARCH 2021

TALKING ABOUT EUROPE: EXPLORING 70 YEARS OF NEWS ARCHIVES

ENRICO BERGAMINI AND EMMANUEL MOURLON-DRUOL

This paper quantitatively explores news coverage on the subject of 'Europe' in three different countries and three newspapers: France (*Le Monde*), Italy (*La Stampa*) and Germany (*Der Spiegel*). We collected and organised large web-scraped datasets covering the period 1945 to 2019. After ensuring the quality of the archives, we identified articles referring to 'European' news while leaving aside national and other non-European news, based on a mix of keyword matching, large-scale natural language processing and topic identification on the full text of news articles. Once articles were classified and datasets labelled, we performed a time-series analysis, detecting salient events in European history, across France, Germany and Italy. We analysed these events in light of the evolution of European cooperation and integration since 1945. We found that the most important events in post-war European history are easily identifiable in the archives and that European issues have gathered substantially greater attention since the early 1990s.

Keywords: European public opinion, big data, machine learning, digital humanities, digital history, topic modelling, media analysis, European history, natural language processing, large text analysis, distant reading

Enrico Bergamini (enrico.bergamini@unito.it) is a PhD candidate at Collegio Carlo Alberto, Turin
Emmanuel Mourlon-Druol (emmanuel.mourlon-druol@bruegel.org) is a Non-Resident Fellow, Bruegel, and Professor of International Economic History, University of Glasgow

The authors thank Catarina Midões for her fundamental inputs and suggestions. The authors would like to thank Amandine Crespy, Francois Foret, Michael Leigh, N. Piers Ludlow, Francesco Papadia, Niclas Poitiers, Giuseppe Porcaro, Stefanie Pukallus, André Sapir and Guntram Wolff for their valuable comments and input. We thank Kurt Jansson and Ulla Siegenthaler at *Der Spiegel* for their kind and helpful collaboration. Furthermore, we thank Estelle Bunout, Marten During and Frédéric Clavert for their helpful comments, and for inviting us to present an early version of this work at the C2DH workshop, as well as the workshop's participants.

Recommended citation:
Bergamini, E. and E. Mourlon-Druol (2021) 'Talking about Europe: exploring 70 years of news archives', *Working Paper* 04/2021, Bruegel

bruegel

1 Introduction: research question and related literature

This paper aims to contribute to the understanding of Europe as reflected in European media. The question of how frequently the media have talked about Europe throughout the course of European integration has evident relevance. Media attention is important in forming public opinion, and allowing citizens to form their ideas. Especially since the 2010s, when arguably the European unification process suffered a setback compared to the acceleration of the 1990s, public discourse about Europe, in national media outlets, has become even more relevant.

But how can we measure media interest in European affairs? The digitalisation of public archives has enabled the application of statistical analysis and natural language processing. Statistical and machine-learning based techniques now make it possible to analyse large amounts of historical data. The combination of digitised databases and computational techniques, at the intersection between humanities and computer science, has helped answer questions from different fields: history, journalism, public opinion and economics (Broersma and Harbers, 2020).

Analysis of the public debate around given issues in European integration has long been of interest to scholars of European studies, in particular historians. Some have studied the press based on a specific case study, such as Mathias Häussler who analysed the debate about Europe in the British popular press through an analysis of the *Daily Express* and *Daily Mirror* of the early 1960s (Häussler, 2014). Häussler stressed the opposing narratives of the two newspapers: the *Express* opposed European integration, while the *Mirror* favoured the UK's entry to the European Economic Community.

Diez Medrano used the 'quality press' in Germany, Spain and the UK to shed light on these countries' different attitudes to European integration. The methodology is based on a close-reading of "*a sample of newspaper editorials and opinion pieces published in British, German, and Spanish quality newspapers between 1946 and 1997, the year I ended my fieldwork*" (Diez Medrano, 2003, 16). In the case of the German press, Diez Medrano analysed even years only for the *Frankfurter Allgemeine Zeitung*, and odd years only for *Die Zeit*, and concentrated on op-ed articles only (Diez Medrano, 2003, 267–69). Reading these articles allowed him to code them into several categories (attitude to European integration and to transfer of sovereignty being two of the most relevant). He used a sample of 90 articles from the *Frankfurter Rundschau* between 1950 and 1995 to verify his findings, as this journal is more overtly leftist and regionalist (*FAZ* being conservative-liberal and *Die Zeit* liberal). The analysis allowed Medrano to draw conclusions on the negative and positive comments about European integration (Diez Medrano, 2003, 106–56).

Another strand of literature focuses on press coverage of European summits. Jan-Henrik Meyer analysed the media coverage of European summits between 1969 and 1991, focusing on five important summits: The Hague (December 1969), Paris (December 1974), Brussels (December 1978), Luxembourg (December 1985) and Maastricht (December 1991) (Meyer, 2010). Meyer's analysis covered the press in three countries from shortly before the summits took place until shortly after. Two newspapers were selected for each country: *Le Monde* and *Le Figaro* for France, *The Guardian* and *The Daily Telegraph* for the UK, and *Süddeutsche Zeitung* and *Frankfurter Allgemeine Zeitung* for Germany. Meyer explicitly chose a selected number of cases *"to reduce the amount of data"* (Meyer 2010, 43). In this paper we embrace the opposite methodological choice: we analyse a vast amount of data to give a robust statistical basis to our analysis. One of Meyer's research questions is partly related to our project, since he examined whether Europe as a polity became a point of reference for the media. However, he was not looking at the media focus on European news in contrast to national news, and because of his methodological choices, he could not draw conclusions over the entire post-war period.

Still partly focusing on summits, Christiane Barth and Patrick Bijsmans analysed the public debate about European integration at the time of the discussion of the Maastricht treaty (Barth and Bijsmans, 2018). Their analysis focused on newspaper articles published during, three days before, and three days after five important summits: Dublin (June 1990), Maastricht (December 1991), Lisbon (June 1992), Copenhagen (June 1993) Corfu (June 1994); and during, before and after four important referendums: Denmark (June 1992 and May 1993), Ireland (June 1992), France (September 1992). The analysis included four print media sources: *Frankfurter Allgemeine Zeitung*, *Süddeutsche Zeitung*, *The Times* and *The Guardian*. They counted relevant articles, and identified the main framing used in the debates through a close reading of the articles. In contrast to this approach, our research allows an unprecedentedly exhaustive analysis, covering all articles published in print media. Barth and Bijsmans were not looking at the news focusing on the Maastricht debate in proportion to the totality of news in the relevant journals.

Finally, another stream of research focuses not on articles but on journalists working on Europe (Baisnée, 2003; Herzer, 2017). Herzer focused on the journalists themselves (who he calls *euro-journalists*) and their advocacy for the cause of European integration. Herzer also based his research on the newspapers, through an *"extensive reading of European integration coverage,"* from the French, Italian, German and British press between the 1950s and 1970s (Herzer, 2017, 11). Herzer's analysis was carried out *"using keyword searches in online databases or were consulted on microfilm,"* rather

than through a systematic digital search like ours, which can provide a quantitative indication of the significance of the coverage of European issues relative to coverage of national issues (Herzer, 2017, 11).

There are some, more recent examples of text-analysis exercises based on digital-native samples. The work of Küsters and Garrido (2020) is closest to our methodological approach, in that it makes use of structural topic modelling, but its focus is different. The authors focused on *Die Zeit* only, and on how the idea of a 'European South' emerged in German public discourse. Boomgaarden *et al* (2010) studied the factors that explain news coverage about the EU from 1990 to 2006. They took a two-step approach to identification of items of 'EU news'. First, the article had to have at least one reference to the EU or any of its institutions. Second, the relevant articles were weighted according to their place in the newspaper, and the number of references to the EU they include. Such a methodology would not be applicable for our corpus reaching as far back as 1945, as we explain below.

Baker *et al* (2016) created an index of Economic Policy Uncertainty (EPU), based on the uncertainty around economic policy measures in the news. This instrument, complementary to the more traditional indexes of uncertainty based on stock market movements, has proven to be a useful high-frequency indicator for uncertainty. Müller and Hornig (2020) expanded on this index, by clustering with machine learning-based techniques the different components of EPU, giving a new fundamental role to their intersection. The technique employed uncovers different and possibly latent topics in texts.

Our paper takes a similar approach, but focuses on the history of the coverage of the European Union in the news. In a similar context, Müller *et al* (2018) explored the dynamics of the *"blame game"*, looking at which countries were blamed for the financial crisis. In the aftermath of the financial crisis of 2008, different media in different countries reported different reactions around blame. Vliegenthart *et al* (2008) investigated the effects of European media presence at the aggregate level, in the context of the European unification process. They explored the role of framing news in terms of benefits or conflict, and how this framing affects public opinion and support for the European project. They found that media coverage, especially if framed in terms of conflict, has a significant effect on public opinion. Public opinion was, in this case, proxied by survey data on trust in European institutions and perceived benefits from EU membership.

The applications of Natural Language Processing (NLP) methods in political economy papers are obviously not limited to archival news sources. In the context of European interests, Greene and Cross (2017) explored the political agenda of the European Parliament, uncovering topics in textual data.

Our computational analysis makes use of longer-spanning newspaper archives than previous exercises in the literature, ranging from the end of the Second World War to the end of the 2010s. In section 2, we briefly explain how we built a large database of media articles, intended to proxy for the media narrative, and how we made it as consistent across time and as comparable as possible across countries. We selected outlets from three of the founding members of the European Union: *La Stampa* (Italy), *Der Spiegel* (Germany) and *Le Monde* (France). We explain this choice in more detail in the next section.

After discussing in the second section the value of the construction and organisation of a large and long-spanning database, in the third section, we discuss how we identified articles dealing with Europe to isolate them from our corpus. The ambition to cover a long chronological range, from 1945 until 2019, renders impossible close reading of the sources, often used to study the press. To identify European discourse through 'distant reading' as opposed to 'close reading' (Moretti, 2015), we developed an exploratory methodology based on NLP techniques, applied to our dataset of roughly 13 million full-text newspaper articles. Identifying and classifying these articles, given the margins of error of machine learning techniques, presents multiple challenges, spanning from quality and consistency of the data, to computational limits due to the size of the data. We discuss these issues in detail in section 3. Exploiting these techniques, the fourth section investigates how the frequency of reporting of European news has changed over time. Conceptually, we can define a measure of frequency of European news pieces as E/T, where E is the number of European news articles, while T is the total number of news articles. We can investigate how this variable evolved over time and across countries. By looking at peaks and troughs in news reporting, we can also identify milestones in European history.

2 Data collection and preparation

We used three criteria to choose our news sources. First, the sources should belong to the six founding members of the European Coal and Steel Community, then European Economic Community, later transformed into the European Union. Second, the sources should be of importance in the national media landscape. Third, the most limiting criterion was that sources should have a digitised archive spanning back as far as the post-war period. The countries selected are France, Germany and Italy, for which we respectively selected *Le Monde*, *Der Spiegel* and *La Stampa* as data sources. The newspapers, daily in the case of France and Italy, weekly in the case of Germany, are of recognised and established importance in the countries' media landscapes. The historical archives of these newspapers have been digitised and are publicly available online.

The digitisation of raw scanned sources is a time- and resource-intensive process. Typically, this procedure first recognises the boundaries of an article on the page, be it by recognising the spaces across columns or, in older formats, the lines between them. Next, Optical Character Recognition (OCR) is applied to the images. OCR technologies allow us to obtain digital text from images containing scanned or photographed text. The result of these two steps is a machine-readable text, on which text analysis such as ours is possible.

During the years analysed, the newspapers went through substantial transformation in terms of sections and format more generally. Furthermore, the digitisation process and software introduced a large number of errors. Collecting, cleaning and organising this material required complex and careful work, in three languages, involving also computationally intensive machine-learning applications. Annex 1 sets out in detail the procedure we followed to collect, clean and ensure the quality and consistency of the data.

3 Identifying the 'European' corpus

The size of these news archives obviously makes large-scale manual labelling of single articles too resource intensive to be possible. Thus, we relied on automated techniques to recognise the topics discussed in the articles. In this section, we describe our methodology to identify 'European' news. The methodology has two steps. Compared to studies which rely on aggregation of news, building a consistent sample was of utmost importance for us. Annex 1 details how we did it. After ensuring the archives are as time-consistent as possible, we attempted to isolate the relevant subset of news dealing with Europe. For example, the archives contain information about sports, which may explicitly refer to Europe, but are outside the scope of our analysis, which considers the 'European' topic in its commonly understood institutional, political and economic sense. Our methodology is based on a mix of keyword-matching and latent topics discovery, which allows us to identify the subset of articles pertaining to Europe, while excluding false positives.

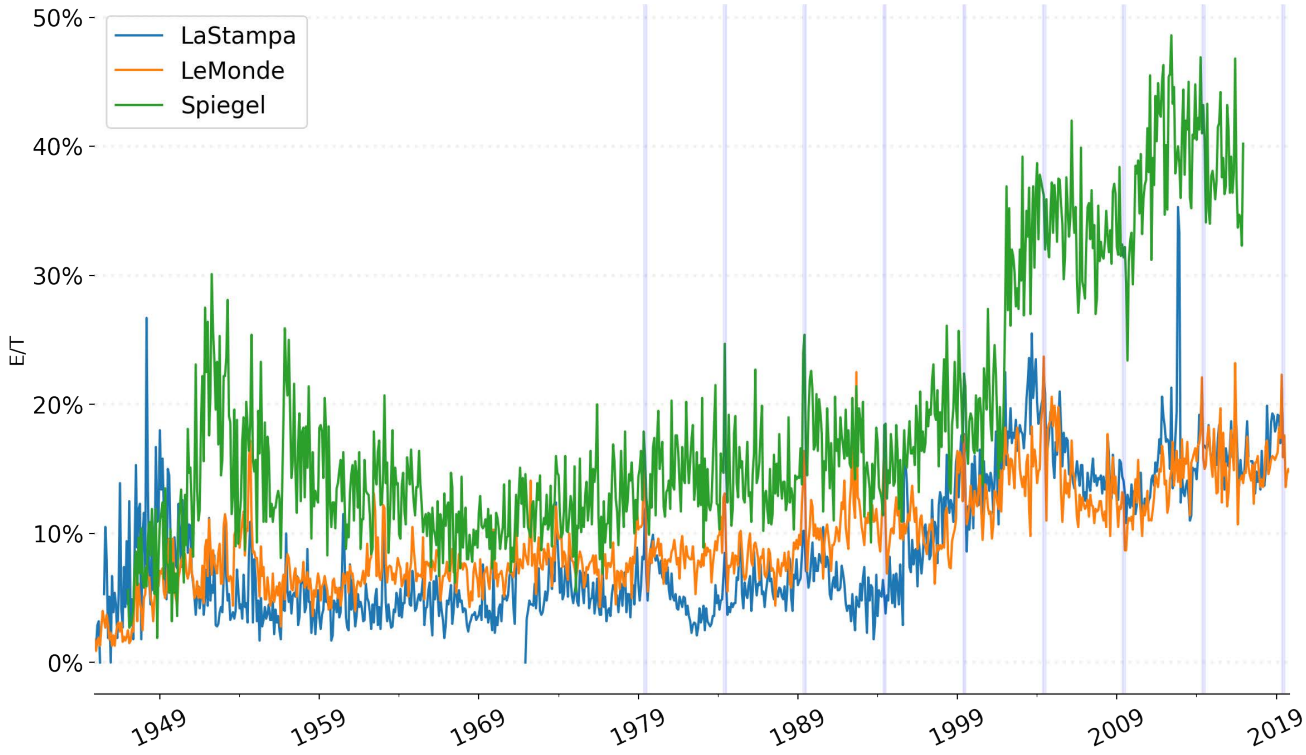
3.1 Keyword scores

We began with simple keyword matching. We put together a list of keywords for the concepts and institutions related to European integration and the concept of Europe, in French, German and Italian. For example, we looked for words including 'Europe', 'European Commission' and acronyms such as 'BCE' (working both in Italian and French for Banca Centrale Europea/Banque centrale européenne; in English ECB, European Central Bank) or 'CECA' (same, with Comunità Economica Carbone e Acciaio/Communauté européenne du charbon et de l'acier; in English ECSC, European Coal and Steel Community). The full list of keywords and acronyms is available in Annex 2. We relied on wildcards and regular expressions to account for cases (so that 'europea', and 'europeo,' 'européen' all match).

The search produced a cumulative score: in the first instance, every word match increases the score of an article by one point. We assumed that an article cannot be about 'Europe' if it does not contain a word in our list: having a word in the list is a necessary but insufficient condition for the article be about Europe. We believe this is a fair assumption, as our keyword list is purposefully broad. However, as a first defence against false negative, we assigned a penalty score to words such as 'euro' or 'European': their contribution to the score is 0.5 instead of one. In addition, to further protect against false positives, we choose a cut-off threshold, as a rule of thumb, of 3 for the keyword score. Articles below this threshold were assigned only to the 'T' (Total) count, but above the threshold they added to the count 'E' (European). We aggregated the results into monthly time series, and calculate E/T, ie

European news as a share of total news. As discussed in section 4, spikes in the series, seen in Figure 1, correspond to historically relevant events regarding Europe, for instance European elections.

Figure 1: E/T estimate based on matching keywords.



Source: Bruegel.

Vertical blue bars in Figure 1 signal European elections: we can observe how the frequencies increased in all newspapers in the months around elections. Many other events are already identifiable at this stage, as further discussed in section 4.

In terms of level, we can see a structural break starting in the mid-1990s. It is reasonable to link this to the Maastricht Treaty of 1992 and the subsequent acceleration of the process of European integration. However, we cannot rule out that the increase in level, despite the penalty score and cut-off threshold, is due to remaining false positives. The use of the adjective 'European', and the use of 'euro' as a currency indeed became ever more frequent and intense after 1992 and therefore could lead to false positives. This seems to be the most relevant problem, not accounted for by just lowering the weight of these words in the keyword count. Furthermore, another source of noise is likely to be sports. While this is also an interesting part of public opinion, and could be considered 'European' news, as mentioned above, our analysis focuses on Europe as a mostly institutional, political and economic concept.

3.2 Topic modelling and sub-topic filtering

While the false negatives, assuming that keywords are comprehensive, are not worrisome, false positives inflating the numerator 'E' are. Keyword matching allowed us to subset the full corpus to obtain news articles talking about Europe, largely defined by the keywords score. However, we wished to refine this subset further, by means of natural language processing. We also faced a definitional problem: what is the concept of 'Europe' that we are trying to define? European institutional, political or economic issues might be mentioned in the news, without necessarily being the direct subject of the news. We make use of Latent Dirichlet Allocation, a form of topic modelling (Blei *et al*, 2003; Blei and Lafferty, 2006). Topic modelling is a probabilistic technique aiming at inferring latent topics from a corpus of texts. It assumes that any collection of texts can be understood as a blend of latent topics. A 'topic' is identified by a collection of salient words, which in turn can be interpreted as a coherent topic. Our corpus, therefore, can be seen as a collection of different topics. We could thus build a document-topic matrix, which expresses the probability that each article belongs to each topic.

Our dataset is large, in terms of dictionaries (single words), number of documents, and potential number of topics. We rely on the MALLET[1] toolkit for implementation of topic models (McCallum, 2002). MALLET performs superiorly with large datasets, as well as a large numbers of topics (more than 100), compared to other widely used software such as Gensim[2]. Under the hood, the improved performance is attributed to a different underlying model of sampling: while Gensim relies on Variational Bayes Sampling, MALLET makes use of Gibbs sampling. Interested readers may explore the technical differences in available documentation. For the scope of this paper, given the size of our dataset, the most relevant difference is software performance.

LDA is an unsupervised model, that is, topics are free to form independently, and, although we suspect certain topics of being blended in the corpus, we do not nudge the models in this or in any other direction. The main limitation of the LDA model is that it does not take into consideration the time dimension in our datasets. This implies that the topics recognised are not time-varying, despite our data's strong time variation. Structural Topic Models (Roberts *et al*, 2019) and Dynamic Topic Models (Blei and Lafferty, 2006; Jähnichen *et al*, 2018) overcome this gap. However, because of computational limitations arising from the size of our datasets, we did not implement these models.

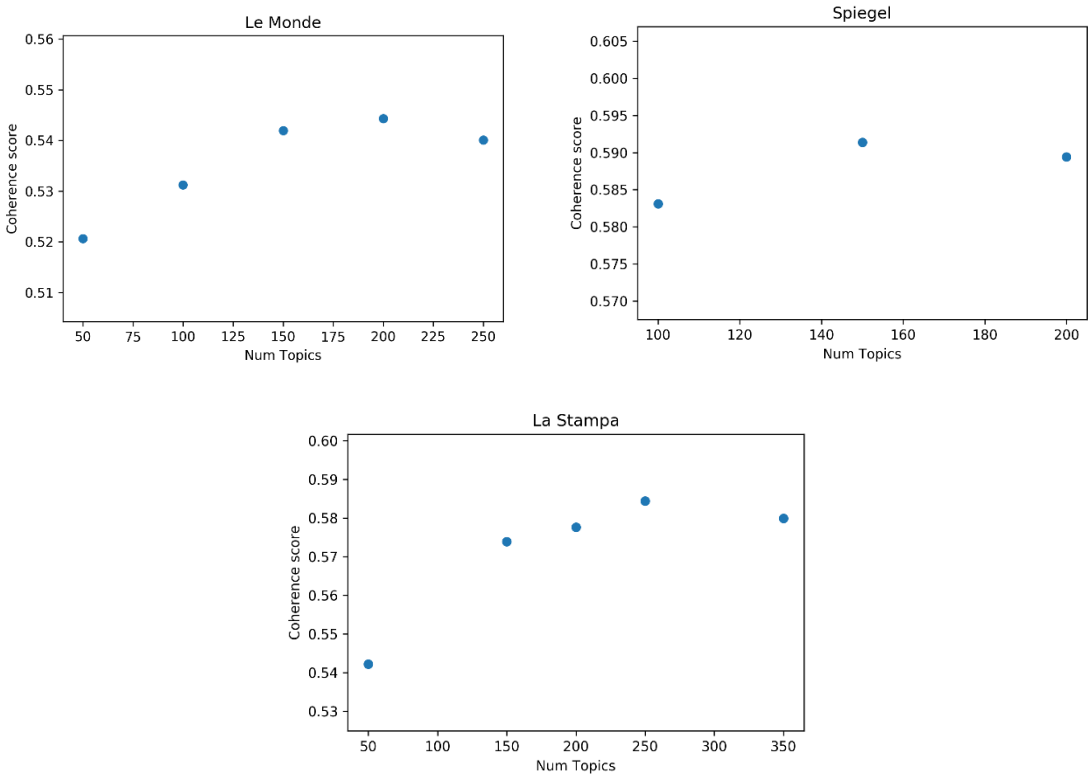
[1] Documentation available at: http://mallet.cs.umass.edu/topics.php.
[2] Documentation available at: https://radimrehurek.com/gensim/.

Future work should explore the time variation in the topics. We discuss the issue in further detail in section 5.

We estimate the topic models for each of the three archives separately: our analysis is language dependent, and requires building three different models. One of the key parameters to tune when fitting an LDA model is the number of topics potentially present in the corpus. *A priori*, this number is unknown, particularly when dealing with unstructured data. We run the same model while varying the number of topics and use a common evaluation metric, the coherence score (CV metric), to choose the best performing one.

In Figure 2, we visualise the coherence score for the three LDA models. We plot the coherence score on the vertical axis against an increasing number of topics on the horizontal axis. For each dataset, we choose the coherence-maximising number of topics[3].

Figure 2: Coherence score and number of topics for the *Le Monde*, *Der Spiegel* and *La Stampa*



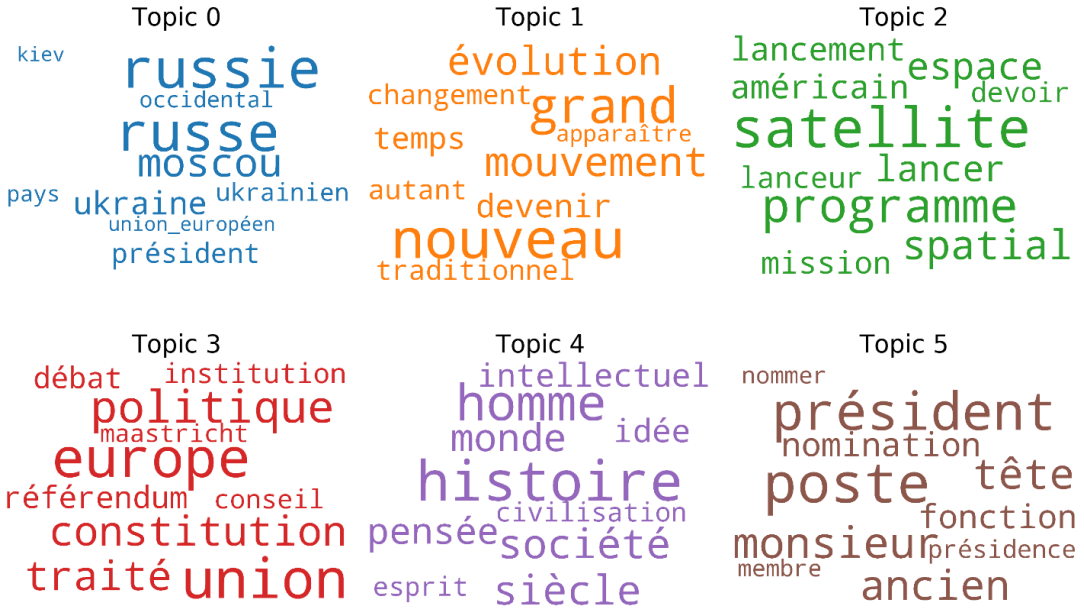
Source: Bruegel.

[3] We present here only the coherence-maximising attempts between 100 and 300 topics, although initial attempts worked with different orders of magnitude (10, 500 or 1000).

Once we have optimised the model over the number of topics, we estimate the LDA model for each of the archives. After this estimation, we can exploit the probability distributions in the document-term matrices. We can then perform an automated labelling: assigning each document to a topic. For both simplicity and computational limitations, we chose a single-topic assignment: we picked the highest probability score from the rows of the document-topic matrix. This means that one article of the archives can belong to only one of the topics: namely the one assigned the highest probability by the model. Future research should aim at exploring the nuances present in these probability distributions.

At this stage, we have, alongside the information about the cumulative keyword scores, also each article's topic, as determined by the LDA. We exploited this labelled dataset to summarise the information about the topics and combine it with the keyword scores. We created a dataset of one topic per row; for each topic, we reported the salient words describing it, as well as the mean value of the keywords scores matched from Section 3.1 of articles in this topic, to facilitate topic interpretation. In Figure 3 we plot, for example, some of the topics present in *Le Monde*'s archive.

Figure 3: Selected topics appearing in the *Le Monde* archive



Source: Bruegel.

The examples presented in this picture are only a small subset of the 200 topics that we ought to interpret from the LDA model for *Le Monde*. Given that the LDA modelling was performed on the subset of 'European keyword'-matching articles, determining which topic truly belongs to 'E' and which is not

a small exercise. We ordered the topics-summary tables by mean keyword scores. The most 'European' topics should appear at the beginning.

From the examples in Figure 3, only one topic (Topic 3) clearly refers to the institutional and political reality of the European Union. The information extrapolated from Topic 0 clearly refers to the countries on the eastern border of Europe. This topic might contain articles referring to the Cold War, as well as to the most recent tensions with Russia over Ukraine's political situation.

We performed this exercise in order to filter out from the 'E' sample news articles about sports, business or economics, which do not refer to Europe, and to deflate it from matching common words (for example the currency euro). However, in the case of more nuanced topics, the articles may refer to Europe in some cases. We can imagine, for example, an article about Europe's foreign policy. Since we associate with each article both a topic and a score based on keyword matching, we combined this information to capture these nuances. We mapped our topics into three categories: definitely belonging to 'E', might be 'E', or definitely not 'E'. The articles belonging to topics assigned to the last category were excluded from our subset. Sports-related topics fall into this category. The uncertain topics are often related to international relations (such as Topic 0 of the example), news of financial nature, or more vague topics like Topic 1, 4 and 5 of the example. From this category, we only keep in the 'E' subset those articles with a keywords score higher than the average score of the articles belonging to that topic.

In Table 1 we present the top five topics ordered by the average score of keywords within each article, for each newspaper. In Annex 3, we report the full tables of the topics isolated and labelled with each of the four categories into 'European', 'not included', 'international politics', or 'uncertain'.

Table 1: Top 5 topics in terms of the average keywords scores, for each newspaper

Label	Topic ID	Average keywords per article	Keywords
European	Der Spiegel	7.2	ewg, brüssel, europäisch, gemeinschaft, brüsseler, europa, land, prozent, frankreich, gemeinsam
European	Der Spiegel	6.8	prozent, wirtschaft, staat, land, regierung, unternehmen, deutschland, wachstum, investition, hoch
European	Der Spiegel	6.7	merkel, schäuble, kanzlerin, angela_merkel, politik, seehofer, koalition, union, deutschland, berlin
International politics	Der Spiegel	6.3	england, britisch, brite, london, groß_britannien, britische, londoner, englisch, engländer, europa

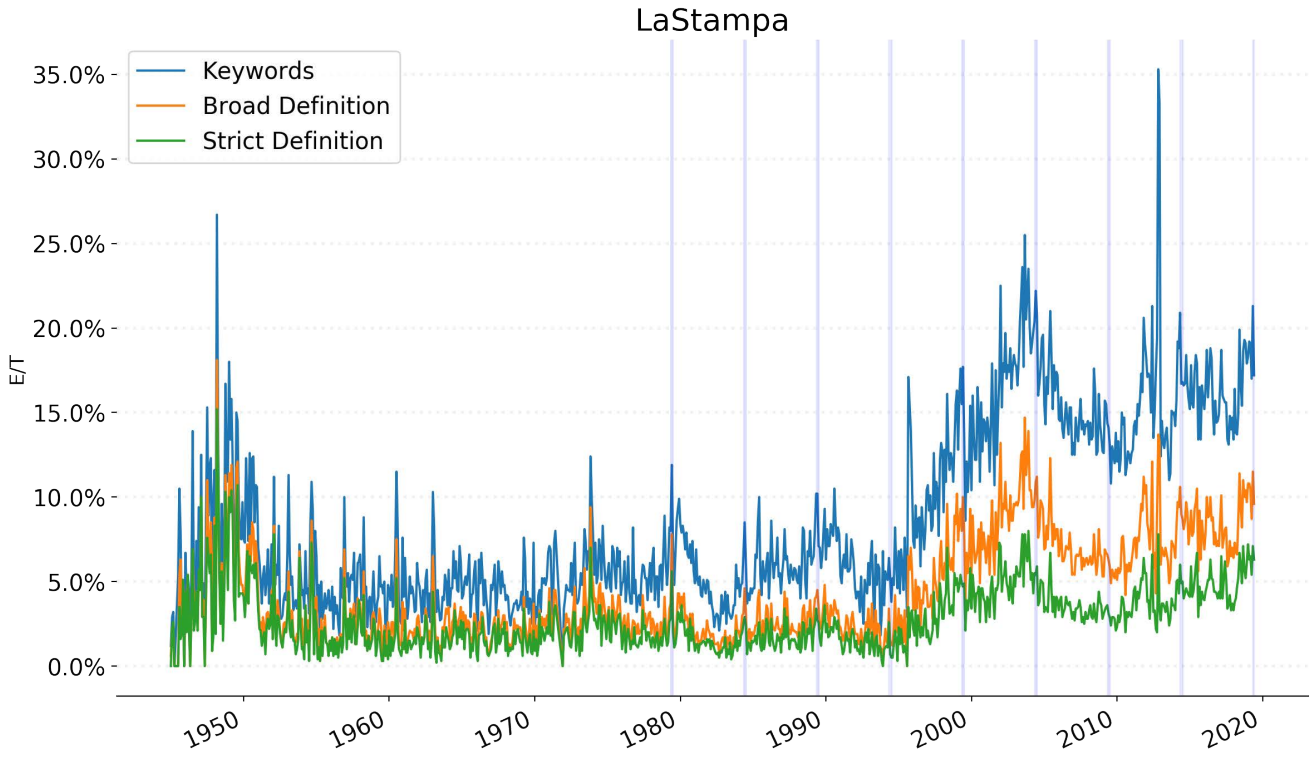
International politics	Der Spiegel	6.1	russland, usa, westen, land, europa, nato, eu, moskau, welt, obama
European	Le Monde	5.3	président, angela_merkel, nicolas_sarkozy, dirigeant, chef_etat, bruxelles, juncker, vouloir, françois_hollande, monsieur_sarkozy
European	Le Monde	5.0	commercial, accord, etat_unis, commerce, négociation, pays, omc, gatt, libre_échange, marché
European	Le Monde	4.8	système, exemple, cas, principe, forme, technique, méthode, existe, formule, pratique
European	Le Monde	4.5	ue, pays, union_européen, union, membre, etat_membre, adhésion, élargissement, bruxelles, candidat
European	Le Monde	4.1	conseil, comité, conférence, membre, réunion, travail, ministre, réunir, organisation, représentant
European	La Stampa	5.0	europeo, unione, europa, paesi, comune, parlamento, strasburgo, comunità, nazionale, integrazione
European	La Stampa	4.0	presidente, europeo, vertice, unione, ue, ciampi, europa, paesi, presidenza, prodi
European	La Stampa	3.6	italia, deficit, economico, economia, pil, ciampi, crescita, politico, tesoro, finanziario
European	La Stampa	3.5	cee, comunità, comunitario, commissione, europeo, bruxelles, paesi, italia, lira
European	La Stampa	3.4	accordare, trattativa, volere, soluzione, chiedere, compromettere, proporre, incontrare, accettare, negoziare

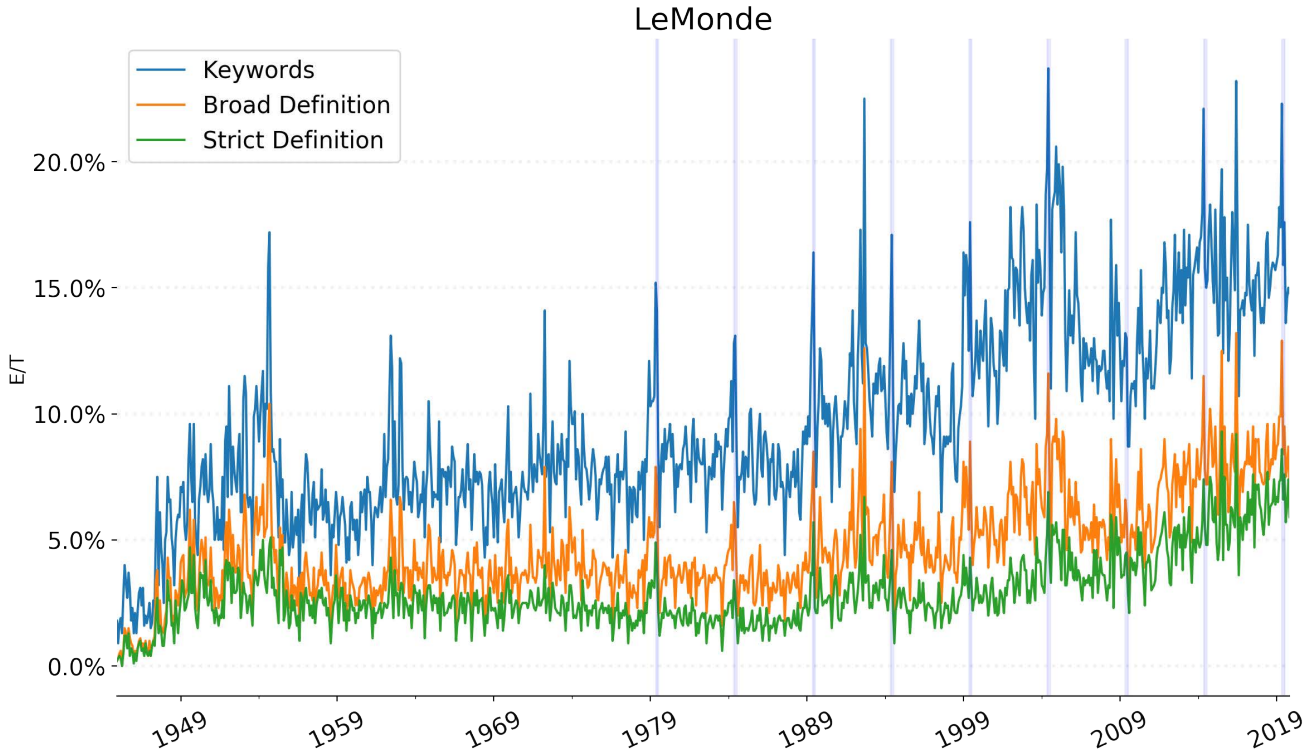
Source: Bruegel.

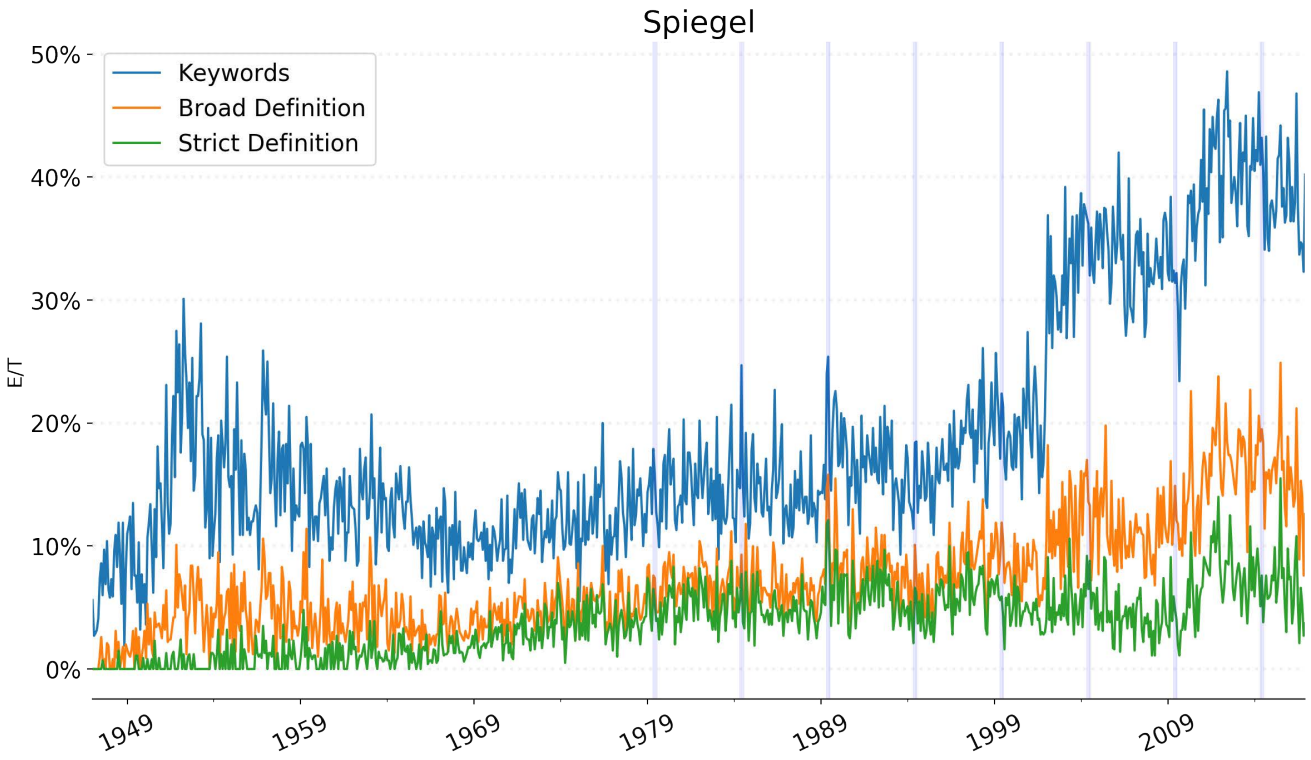
The dividing line between an article talking about Europe and one about international politics is blurred. For computational reasons, we have to attribute each article to only one topic. Hence, we try and flag uncertain topics that are more likely to belong to the denominator and find a compromise between false positive and false negatives.

We plot, in the following figure, different E/T ratios, derived from our methodology. Alongside the simple keywords matching (blue line), we build two new E/T frequencies, reflecting the filtering out of topics. The E/T under the 'strict definition' (green line) excludes altogether the articles belonging to topics classified as uncertain, while the 'broad definition' (brownish line) includes the articles from the uncertain topics only if they have a higher-than-average keyword score. This is our preferred measure of the E/T ratio.

Figure 4: Three interpretations of the E/T ratio for *La Stampa*, *Le Monde* and *Der Spiegel*







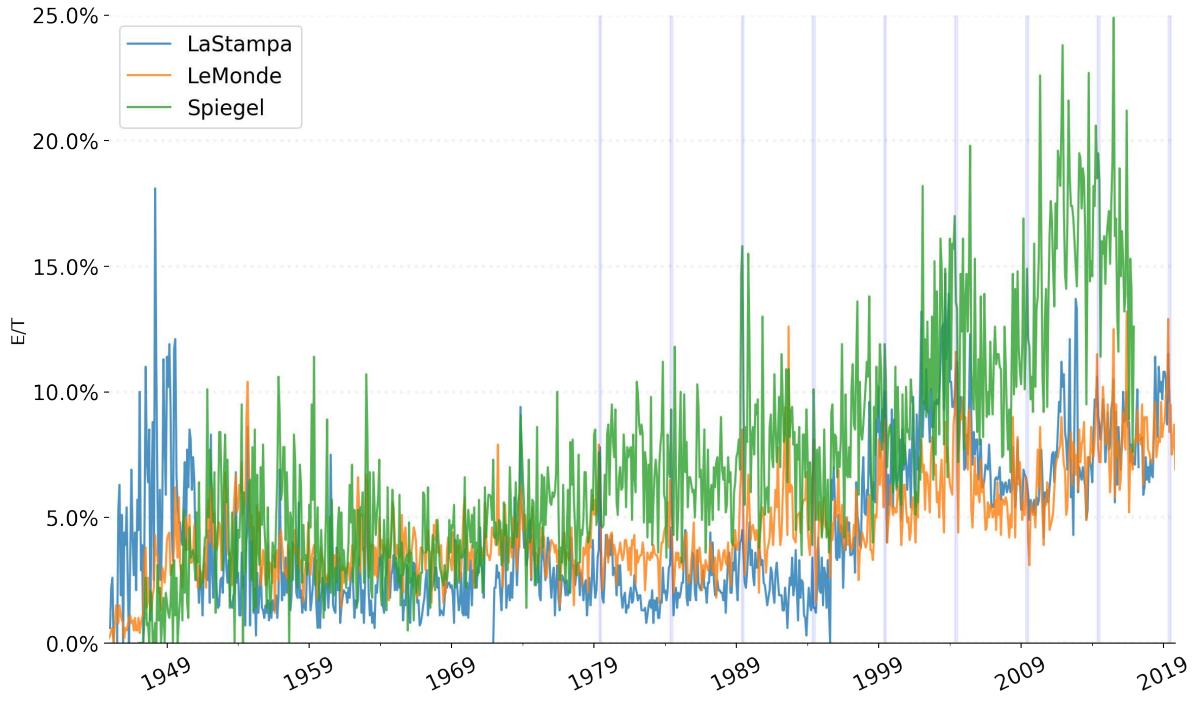
Source: Bruegel.

We can observe how the deflation of the numerator mostly happens in the post 1990s period, reflecting our suspicion that mentions of the euro as a currency were particularly frequent, substantially affecting the final results.

In conclusion, we started from a largely unstructured and unlabelled database. Our methodology combined an *a-priori* definition, purposefully broad, about one topic, to isolate a subset within the broader corpus. Mixing this definition with a latent topic analysis helped us to optimise the trade-off between false positives and false negatives. While introducing qualitative judgement and interpretation of the topics in the analysis might be a source of error, we believe that this methodology might be useful in more precisely identifying an *a-priori* known topic, within unstructured corpora, particularly when dealing with large amounts of data.

Similarly to Figure 1, we depict in Figure 5 monthly time series of the frequency of European news, again highlighting European election dates in blue.

Figure 5: E/T estimate after topic modelling



Source: Bruegel.

The results of the exercise are reassuring since most of the peaks correspond to European events, and the rather sudden jump in level of the mid-1990s is now deflated.

4 Europe in the news: trends and events detection

As mentioned, we selected the filtered version of the E/T ratio as our preferred measure ('broad' definition in the previous charts). We thus have three time series for *Le Monde*, *Der Spiegel* and *La Stampa*. We rely on the monthly aggregation in Figure 5. Comparability across sources is quite difficult as the outlets vary greatly. While the national sections of *La Stampa* and *Le Monde* seem to move together, the values are generally higher in the case of *Der Spiegel*. This can be attributed to the difference in format, addressed readership, and editorial choices. In spite of these differences, we observe that the attention paid to 'Europe' has been growing over time. Particularly, the positive trend seems to have accelerated in correspondence with most of the major developments in post-war European cooperation and integration. Since we cannot provide a close reading of the ca. 13 million articles in the dataset, we analyse the evolution of the trend and how this relates to the historical evolution of the EEC/EU.

We look at the deviations from the trend, and particularly the peaks. In order to consistently and robustly flag as 'events' the peaks which exceed a robust threshold, we relied on a simple Z-score algorithm (Brakel, 2020). For every time series, we estimated a moving average of 6 months, and a threshold of one standard deviation. All the peaks in the distribution exceeding (positively) one standard deviation above the moving average are marked as events.

We show that this 'distant reading' correctly identified European-related news over national news in two opposite but complementary ways. The first is to compare what we could expect to appear (that is, European-related events that should have 'made the news') with the E/T variable and see whether there is a match. The second is to isolate what are statistical peaks in the E/T ratio across several newspapers, and check whether they match events in the evolution of post-war European cooperation and integration.

As to the first element, Figure 6 plots two types of European-related events that we expect to have resulted in increases of E over T: the European direct elections every five years since 1979, and some of the most famous European Council meetings since 1974. Both European elections and European Council meetings are traditional focal points of European affairs, and often used as reference points for analyses in the press, as explained in the first section.

Concerning the second element, Table 2 shows the date (year and month) of the events that have been identified as common peaks across *La Stampa* and *Le Monde*, that we briefly interpret. We implement the Z-score peak detection on all the three time series and compile a list of the events contemporaneously happening in all the three newspapers[4]. For ease of presentation, we used monthly rather than weekly granularity: monthly observations already detect 30 events, while weekly observations lead to more than 200. The peak detection algorithm derives a rolling average on the time series, and flags as an event a deviation from this moving average of more than one standard deviation. While the results are promising for *Le Monde* and *La Stampa* (both daily newspapers), *Der Spiegel*'s events seem to match less than the other two. We attribute the problems to differences in editorial format, and further research should uncover these differences. At a monthly granularity, we flag the events, and identify which dates match in all three outlets.

Out of the 30 common events statistically detected, only four dates (April 1950, December 1963, January and May 1971) remain unclear: either there is no obvious reason that can be found to explain

[4] We implement this at a monthly granularity. The events presented include those common for *La Stampa* and *Le Monde* for the years 2016-2019, because in these dates *Der Spiegel* data was not consistently available.

the simultaneous peak, or there are several correlated elements, but we considered it would be an overstretch to attribute to them, alone, the peak. All other events clearly match important evolutions, or landmarks, in European integration since 1945. These events are of two types. First, the events that are in part the 'expected events' we had plotted in Figure 6. In addition to one European election (June 2004) and seven European Council meetings (June 1977, December 1985, June 1991, June 2004, March 2007, October 2014 and March 2017), we found referendums (Lithuania/Slovakia on EU accession in May 2003 and the Irish 'no' to the Treaty of Lisbon in June 2008), and Treaty-related events (signature of the Western Union Treaty in March 1948 and rejection of the European Defence Community by the French national assembly in August 1954). Second, we found some of the traditional milestones of European integration: de Gaulle's veto of the UK's entry to the EEC (January 1963), creation of the European Monetary System (December 1978), release of the Delors Report on Economic and Monetary Union (EMU, April 1989), and reform of the Stability and Growth Pact (September and November 2011).

Table 2: Relevant events in both *La Stampa* and *Le Monde*, and their interpretation

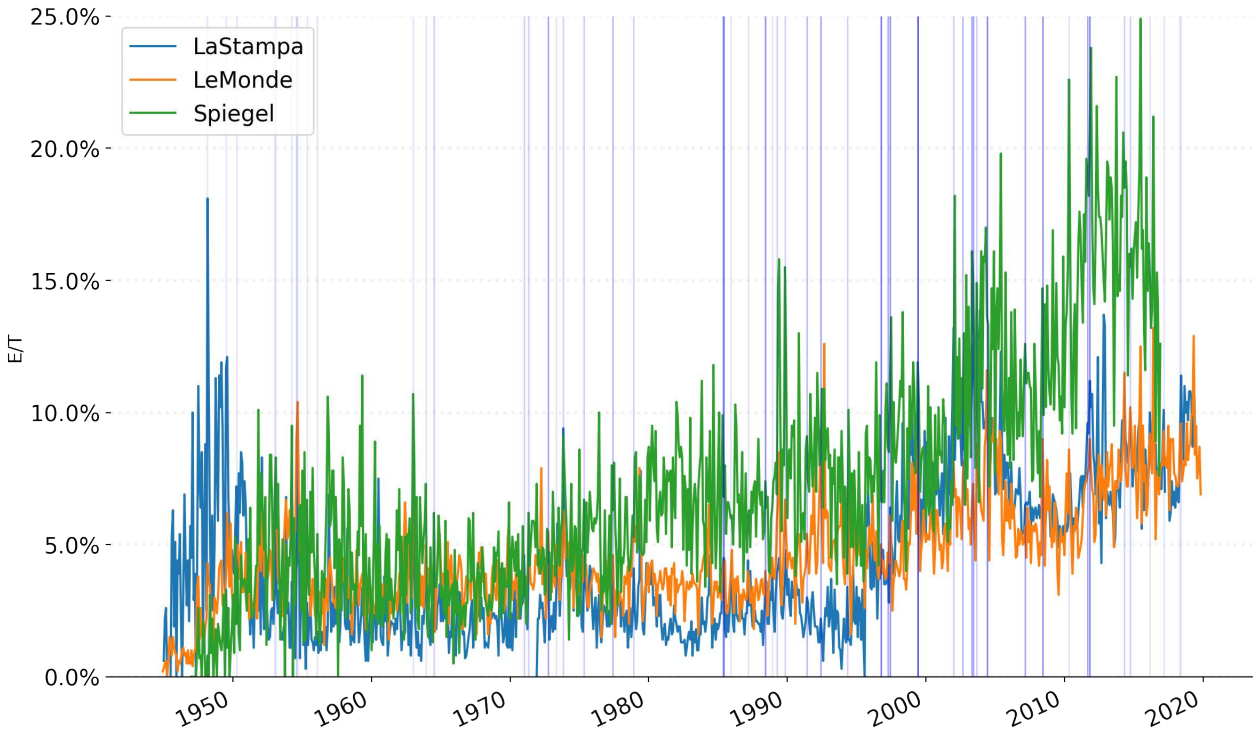
1948-03	Signature of the Western Union Treaty (Belgium, France, Luxembourg, the Netherlands and the United Kingdom)	1989-11	Fall of the Berlin Wall
1950-04	Difficult: talks at the Council of Europe	1991-06	European Council in Luxembourg
1953-02	Common market for coal and iron ore put into place	1994-05	Establishment of the European Investment Fund and inaugural conference for the Stability Pact for Central and Eastern Europe
1954-08	French national assembly rejects European Defence Community	1997-04	Unclear (Commission Green paper 'Partnership for a new organisation of work' and cooperation agreements with Cambodia and Laos)
1963-01	De Gaulle vetoes UK application to EEC	1997-06	European Council in Amsterdam
1963-12	Unclear	2003-05	Referendums in Lithuania and Slovakia on joining the EU; EU-Canada and EU-Russia summits
1971-01	Unclear (France presides EEC Council and Yaoundé Convention enters into force)	2004-06	European elections; European Council in Brussels
1971-05	Unclear (Monetary Compensatory	2007-03	European Council in Brussels

	Amounts introduced in Common Agricultural Policy)		
1973-11	Oil shock since October 1973	2008-06	Irish referendum on the Treaty of Lisbon (53.4% against); European Council meeting
1977-06	European Council in London	2011-09	Reform of the Stability and Growth Pact; State of the Union address
1978-12	Agreement on the creation of the European Monetary System	2011-11	Draghi takes over as ECB president; G20 in Cannes; 'Six pack' adopted
1985-05	G7 in Bonn and commemoration of the end of the Second World War	2014-10	European Council in Brussels; European Parliament approves new Commission
1985-12	European Council in Luxembourg	2017-03	European Council meeting
1987-03	30th anniversary of the signature of the Treaties of Rome	2018-05	GDPR enforceable
1989-04	Release of the Delors Report on EMU		

Source: Bruegel.

If we examine the weekly granularity, resulting in more than 200 co-events, then other expected milestones appear. This includes several institutional milestones, such as the signature of the Treaties of Rome, summits of heads of state and government (including not only the summit in The Hague in December 1969 that gave a new impetus to European integration, but also the summits in Paris of 1972 and Copenhagen of 1973, and several European Councils after 1974), and European elections (in particular the first direct elections in 1979, but also subsequent ones). From the 2010s, the weekly granularity also allows us to see the Five Presidents' report on EMU published in June 2015 or the agreement on the Iranian nuclear programme reached a month later, in July 2015. In short, moving from monthly granularity to weekly granularity further reinforces the result that the news articles computationally identified as European-related indeed reflect an event that is relevant in the evolution of European integration.

Figure 6: Relevant events reflected in the three newspapers



Source: Bruegel.

Finally, to investigate the correlation of our measure of media presence with public support for Europe, we performed a simple econometric test, by referring to the model employed by Vliegenthart *et al* (2008): we regressed public opinion about Europe on our E/T ratio.

For the measure of public opinion, we relied on the indicator of preference for Europe developed in Andreson (2018), despite a much more limited timespan than ours. This indicator expresses support for Europe, and shows an opposite trend compared to ours, decreasing in time, and particularly in the past 20 years. We constructed a panel dataset for the three countries, including these two variables, unemployment and GDP growth, all aggregated at the yearly level.

As expected, the simple correlation figure between the E/T ratio and this public opinion indicator is -0.55. We further explored this correlation by running a simple fixed effects model, using two of the controls employed by Vliegenthart *et al* (2008). The dependent variable was again the indicator for

support for Europe, E/T was our independent variable of interest, and we controlled for unemployment rate and GDP growth rate.

Table 3: Fixed effects estimation of media effects on preference for Europe

	Preference for Europe
GDP growth	0.011***
	(0.004)
Unemployment	-0.010**
	(0.004)
E/T	-1.688***
	(0.263)
Time fixed effects	Yes
Observations	96
R2	0.402
Adjusted R2	0.369
F Statistic	20.151*** (df = 3; 90)

Notes: *** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

Source: Bruegel.

The negative correlation stays significantly also with the controls, confirming the significance of our measure of press coverage, across the three countries. Thus, the increased frequency of European news in the three selected newspapers does not signal a more positive attitude towards Europe, but rather more attention paid to it, accompanied by a less favourable attitude.

5 Conclusions

This large-scale and unprecedented 'distant reading' analysis of the digitalised archives of three European newspapers over more than 70 years and 13 million articles has allowed us to identify the overall rising share of European news in printed media. Our study thus considerably widens and systematises the scope of previous studies on the press, thanks to the use of a novel quantitative methodology. This allows us to observe the growth of coverage of European news relative to national news, with an acceleration from the mid-1990s. Our study not only confirms that European elections and summits have been traditionally high points of media coverage since the end of the Second World War, it also identifies some key milestones in European integration, from the rejection of the European Defence Community in 1954 to the euro-crisis summits of the 2010s.

Three areas would deserve further subsequent study and improvement. The use of a novel, web-scraped and unstructured methodology presents obvious challenges that we tried to address to the best of our knowledge, but that would benefit from further exploration.

First, the frequency indicator constructed in this paper depends on the composition of the archives. We tried to ensure that the web-scraped material was as consistent as possible. However, it is likely that the digitisation process, while making available a large amount of data, also introduced errors in terms of the quality of the material. In our view, simple factors such as reliable information about the number of pages per edition, and the page to which each article belongs, could be one of the most enriching (and relatively less problematic) sources of information to improve qualitatively the archives.

Second, the difference in sources was a major challenge in this study, as shown in the time-series analysis results. In this sense, the introduction of counterfactuals and addition of different sources could be a way to improve the exercise. Notably, looking into non-European newspapers, with long-spanning and relevant archives, such as the *New York Times*, might yield interesting comparisons, and give robustness to our results, both in terms of trends and events.

Third, in terms of modelling, our application of LDA classification could be further improved. The first and most important development that future research should address is the inclusion of the time dimension into the models. Our LDA implementation assumes models to be atemporal. There exist, however, implementations of LDA which allow for topics to vary over time. Notably, Dynamic Topic models (DTM; Blei, 2006), allow words within topics to evolve over time, and observe their evolution. This could be used in conjunction with the literature in European studies (in particular history, political

science, and sociology) analysing the evolution of representations of Europe, and how the perceptions about European integration changed over time, especially across three different countries, each with its own idiosyncrasies. The implementation, given the size of our data, is quite challenging, although we believe that it might unlock great potential for understanding evolving narratives in the media around the concept of Europe. Conceptually, Europe is an evolving identity, and modelling the time dimension of its media representation could yield very interesting results. A second development would be to integrate a topic variation in the time-series perspective. Mapping the topics across languages is clearly a difficult exercise and would probably require a multilingual implementation. Subdividing E/T frequencies into single-topic frequencies could surely help to interpret phenomena underlying the movements. The coevolution of topics 'within' the denominator is as interesting as the coevolution with topics composing the denominator. Employing dynamic network analysis techniques and semantic structure analysis (Hoffman *et al*, 2018), could help address very relevant questions, such as: has the topic of Europe semantically moved closer to economics than to politics? What is the centrality of immigration within the 'European' topic? How did media framings evolve in the long term?

Finally, a discussion about the relationship between media representation and the medium itself is beyond the scope of this paper. Evidently, though, the media ecosystem has been subject to dramatic changes over the course of the past 70 years. Two comparative analyses could integrate and complement our research question. First, understanding the representation of Europe on television would be of great interest. Second, given how the introduction of digital media has deeply changed news consumption, adding to the analysis the role of social media and web-native content could be highly valuable.

References

Baisnée, O. (2003) 'La Production de l'actualité Communautaire. Eléments d'une Sociologie Du Corps de Presse Accrédité Auprès de l'Union Européenne (France/Grande-Bretagne)', Institut d'Etudes Politiques de Rennes

Baker, S.R., N. Bloom and S.J. Davis (2016) 'Measuring Economic Policy Uncertainty', *The Quarterly Journal of Economics* 131 (4): 1593–1636, https://doi.org/10.1093/qje/qjw024

Barth, C. and P. Bijsmans (2018) 'The Maastricht Treaty and Public Debates about European Integration: The Emergence of a European Public Sphere?' *Journal of Contemporary European Studies* 26 (2): 215–31, https://doi.org/10.1080/14782804.2018.1427558

Blei, D.M. and J.D. Lafferty (2006) 'Dynamic Topic Models', in *Proceedings of the 23rd International Conference on Machine Learning*, 113–120, ICML '06, Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, https://doi.org/10.1145/1143844.1143859

Blei, D.M., A.Y. Ng and M.I. Jordan (2003) 'Latent Dirichlet Allocation', *The Journal of Machine Learning Research* 3 (null): 993–1022

Boomgaarden, H.G., R. Vliegenthart, C.H. de Vreese and A. Schuck (2010) 'News on the move: exogenous events and news coverage of the European Union', *Journal of European Public Policy* 17 (4): 506-526

Brakel, J-P. (2020) 'Peak Signal Detection in Realtime Timeseries Data - Smoothed Z-Score Algorithm (Peak Detection with Robust Threshold)', https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data/2264036222640362

Diez Medrano, J. (2003) *Framing Europe: Attitudes to European Integration in Germany, Spain, and the United Kingdom*, Princeton University Press

Greene, D. and J.P. Cross (2017) 'Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach', *Political Analysis* 25 (1): 77–94, https://doi.org/10.1017/pan.2016.7

Haeussler, M. (2014) 'The Popular Press and Ideas of Europe: The Daily Mirror, the Daily Express, and Britain's First Application to Join the EEC, 1961–63', *Twentieth Century British History* 25 (1): 108–31, https://doi.org/10.1093/tcbh/hws050

Herzer, M. (2017) 'The Rise of Euro-Journalism : The Media and the European Communities, 1950s-1970s', European University Institute, http://cadmus.eui.eu/handle/1814/48767

Hoffman, M.A., J-P. Cointet, P. Brandt, N. Key and P. Bearman (2018) 'The (Protestant) Bible, the (Printed) Sermon, and the Word(s): The Semantic Structure of the Conformist and Dissenting Bible', 1660–1780', *Poetics* 68 (June): 89–103, https://doi.org/10.1016/j.poetic.2017.11.002

Küsters, A. and E. Garrido (2020) 'Mining PIGS. A structural topic model analysis of Southern Europe based on the German newspaper Die Zeit (1946-2009)', *Journal of Contemporary European Studies*, 28 (4): 477-493

McCallum, A.K. (2002) 'MALLET: A Machine Learning for Language Toolkit', available at http://mallet.cs.umass.edu

Meyer, J-H. (2010) *The European Public Sphere. Media and Transnational Communication in European Integration 1969-1991*, Stuttgart: Franz Steiner Verlag

Moretti, F. (2015) *Distant Reading*, London: Verso

Müller, H. and N. Hornig (2020) 'A New Uncertainty Perception Indicator (UPI) – Concept and First Results', *Dortmund Center for Data-Based Media Analysis (DoCMA)* no. 1: 39

Müller, H., G. Porcaro and G. von Nordheim (2018) 'Tales from a crisis: diverging narratives of the euro area', *Policy Contribution* 2018/03, Bruegel, https://www.bruegel.org/2018/02/tales-from-a-crisis-diverging-narratives-of-the-euro-area/

Roberts, M.E., B.M. Stewart and D. Tingley (2019) 'Stm: An R Package for Structural Topic Models', *Journal of Statistical Software* 91 (2), https://doi.org/10.18637/jss.v091.i02

Vliegenthart, R., A. Schuck, H.G. Boomgaarden and C.H. De Vreese (2008) 'News Coverage and Support for European Integration, 1990–2006', *International Journal of Public Opinion Research* 20 (4): 415–39, https://doi.org/10.1093/ijpor/edn044

Annex 1: Data collection and preparation

Data collection

We built *ad-hoc* web scrapers to crawl the archives and create machine-readable datasets. We relied on a variety of libraries and techniques, mostly written in the *Python* and *Bash* programming languages. The sources of collection were different. While the *Le Monde* and *Der Spiegel* archives were consistently formatted in a digitised form, the *La Stampa* archive consists of two, overlapping, web archives[5].

After assessing the feasibility, we built web crawlers that collected more than 13 million full-text articles, from the three sources. While the weekly *Der Spiegel* consisted of 310,000 articles, *Le Monde's* archive consisted of 2.8 million articles. *La Stampa's* archives, when combined, contained 9.9 million articles. The differences in volumes do not only have to do with type of source, but also with the different composition of the newspapers. Necessarily, the size of the *Der Spiegel* archive (a weekly newspaper) was much smaller than the two daily outlets. *La Stampa's* archive in particular, as we will detail, outweighs the other two in size, including different archives (local and national news).

Web scraping allows mining of lists of webpages that compose internet archives, in order to construct structured, machine-readable datasets. We collected all the available information about each single article. Our databases include, among other information, the titles, dates, and full text of each article.

After the data collection procedure, and especially considering the volume of the web archives of interest, we performed quality tests on the resulting datasets, to make sure that the web scrapers did not fail in collection and did not generate duplicate entries.

As we detail in the next section, the quality of novel and web-scraped data is often worrisome, and ruling out scraping errors is of great importance. In fact, unlike other similar studies, our analysis focuses on one newspaper per country, rather than on an aggregation of different archives. Aggregating several archives might help to ensure the consistency of the samples, by reducing the biases from errors (such as missing data, digitisation errors etc.) with respect to a single archive. In our case, historical digitised archives, despite being a unique source, are harder to handle. Their composition depends on the quality of the digitised material, on the composition of the various

[5] The first available at archivio.lastampa.it, while the second available at archiviolastampa.it

sections of the newspaper, on the editions, and several other factors that vary greatly in time, and which are often not obvious, even after the data collection.

The next section precisely illustrates the structure of the archives, and outlines the cleaning methods, as well as the pre-processing steps performed on the text of the articles. Section 3 presented the methodology for identifying the European topics within the archives and the subsequent analysis.

Cleaning unstructured, big, text-data

After collecting the data, we obtained large datasets of 2.8 million articles for *Le Monde*, 310,000 articles for *Der Spiegel*, and 9.9 million articles from the *La Stampa* archives. Given the complexity of our sources, and the uncertainty about their structure, we explored the composition of these archives, and tried to understand the online sample to the best of our knowledge before performing any analysis. We summarise the steps taken to clean the samples, described in this section, in the following table:

1) Text pre-processing:
a. Removal of punctuation, numbers and stop-words
b. Bigram and trigram models
c. Lemmatisation
2) Duplicates removal
3) Consistency of sections (over time)

Text pre-processing

Text obtained via OCR suffers from errors, especially when the raw source of the data is old and complex. We believe we largely accounted for these errors when we applied the machine-learning methodologies (next section). However, before applying any model to our large text sources, we applied the text pre-processing steps, which are common practice in natural language processing.

We implemented three language-tailored versions of pre-processing steps, all relying on the models provided by the Python library *Spacy*[6]. Our implementation removes punctuation, numbers, and common words (stop words). Subsequently, we transformed every word in lower case. Finally, we applied another common step of text pre-processing: lemmatisation. This process reduces each word to a *lemma*, the base form of a word, and is implemented for the three newspapers respectively, in French, German and Italian.

[6] Documentation available at: https://spacy.io/.

The intuition is that reducing the declinations, conjugations and variations of words to their roots, will reduce the dictionary size and the complexity of the database. The loss of semantic meaning is not problematic for the topic models that we apply, as they rely on the bag-of-words assumptions, as we explain in section 3.

Subsequently, we also added bigrams and trigrams. Bigrams and trigrams join together words which often occur at the same time: the words 'European' and 'union' will not only be considered singularly in our dictionary, but they will also be understood as a single entity 'European Union'.

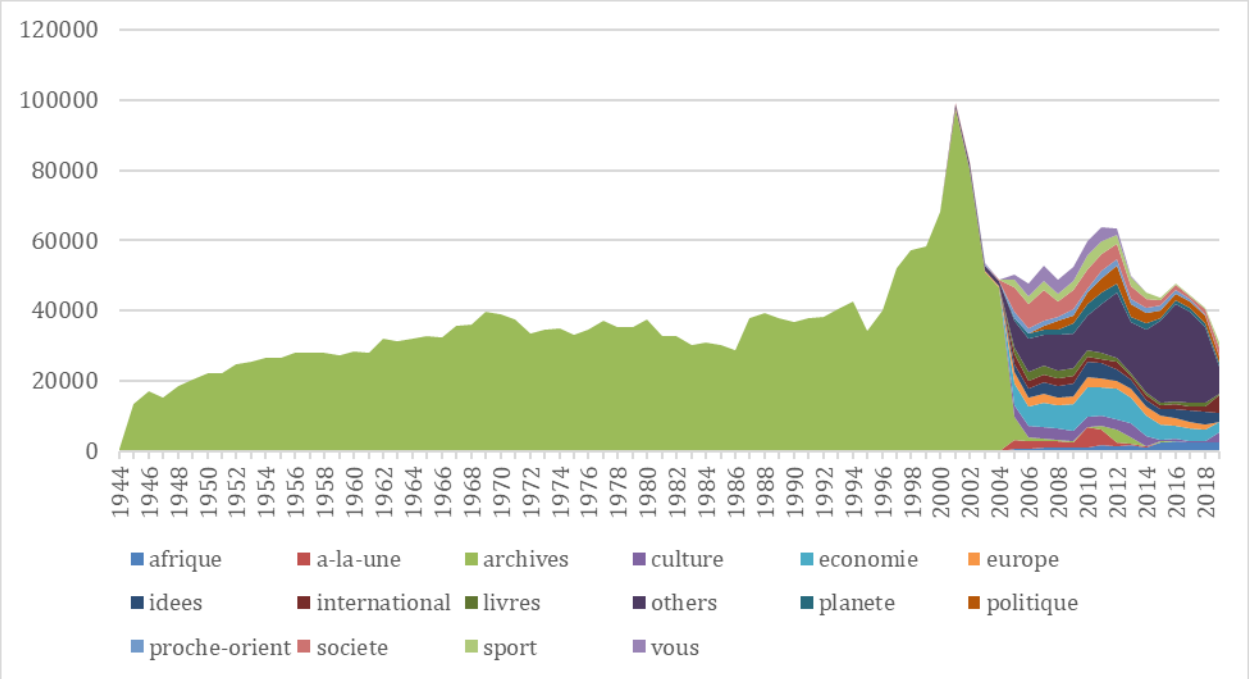
Ensuring the consistency of the samples

The estimation of frequency measures depends as much on the numerator as on the denominator. Other analyses based on common sources of big news data (GDAL or Lexis Nexis) often rely on aggregations of articles from a larger number of media sources. In this sense, this makes the composition of the news samples less worrisome. In our case, the exploration of news archives' composition is essential.

Obviously, newspapers have changed as a product, especially when thinking about time and format. In our case, the delicate cleaning of the sources is crucial to ensure that the sample of articles remains as consistent as possible across the 70 years timespan. In fact, wanting to build a frequency measure of 'European' news over the total (E/T), it is of utmost importance to understand the structure of the databases. Sudden changes in the E/T ratio could be driven by a change in the archival composition of the articles, rather than an increase in the number of news articles related to the topic.

Figure A.1 illustrates the volume of *Le Monde's* archive. In this case, the archives can be subdivided into sections only after 2004, while in the previous period we only have a single 'historical archive' category.

Figure A.1: *Le Monde*'s archive 1945-2019

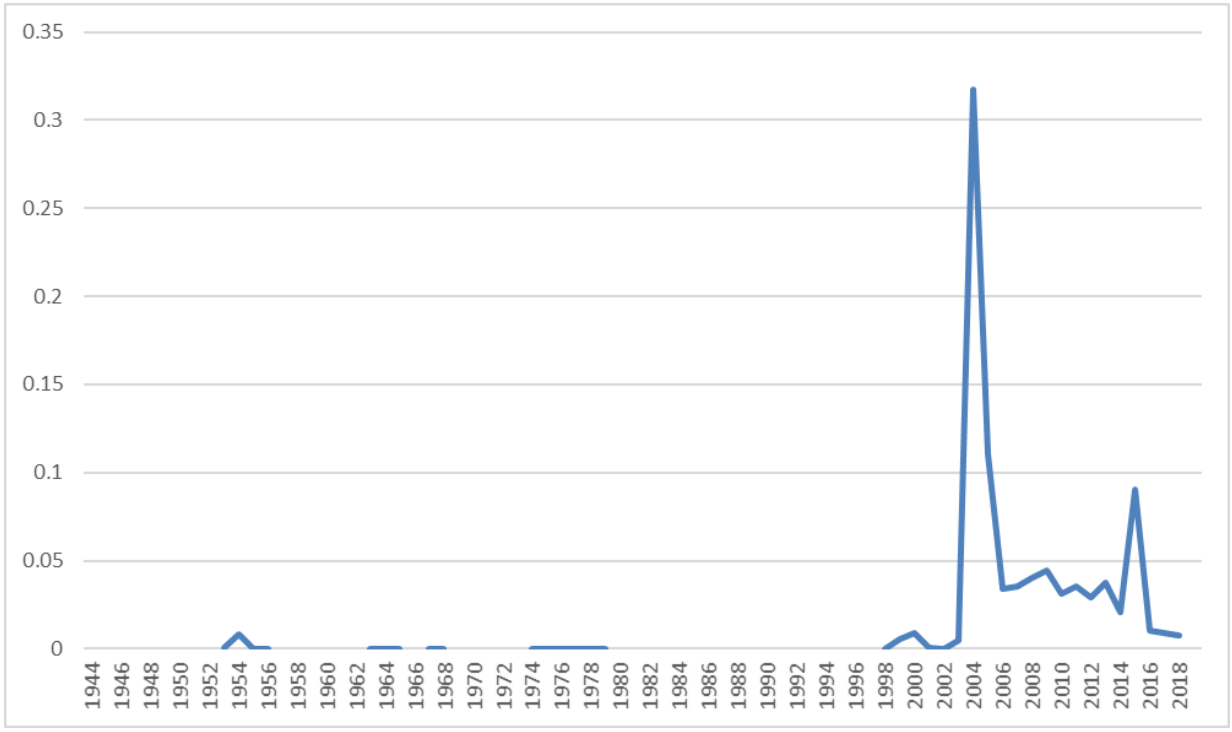


Source: Bruegel.

The structural change in most archives happens around the 1990s, or at the beginning of the 2000s, around the same time that newsrooms developed digitally-native websites, or digitised and merged previous content on the web. Therefore, it is logical to think that the digitisation process can create duplicate entries particularly around those times.

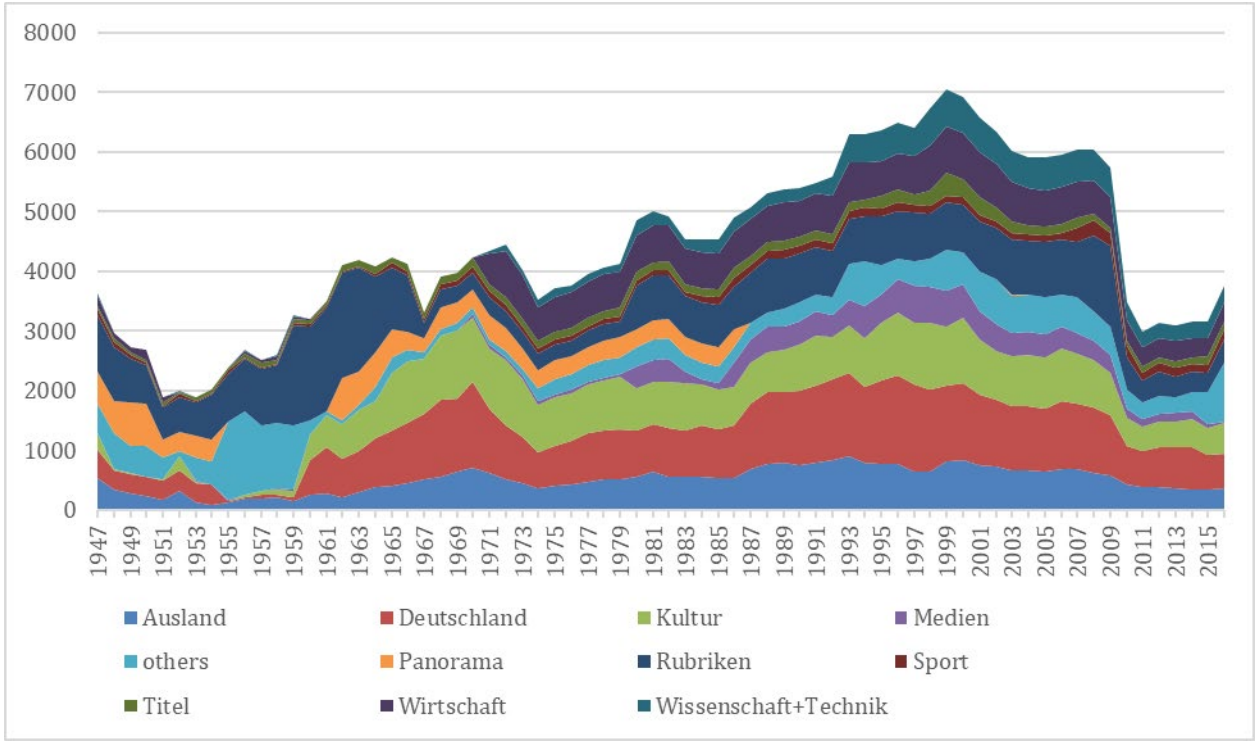
Here, the 2001's spike in volume is likely driven by duplicated content. While a single URL under the archive's internet domain is what we consider a 'unique' article during the scraping phase (hence dropping duplicate links), it is possible that the digitisation procedure produced pairs of almost identical digital articles for the same 'physical' article.

Figure A.2: Percentage of exact duplicates in *Le Monde*'s archive



Source: Bruegel.

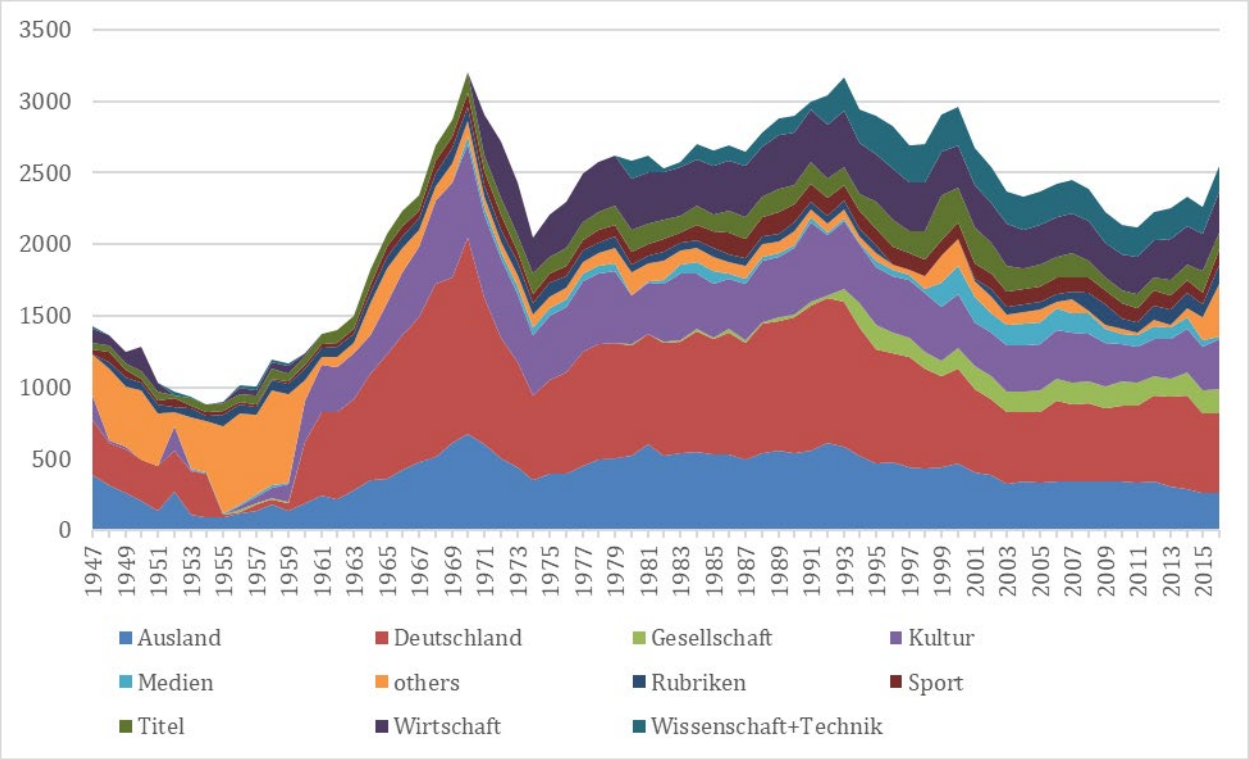
Figure A.3: Raw archive in *Der Spiegel*, by Section



Source: Bruegel.

Conversations with *Der Spiegel* archivists revealed how the drop could be due to a merger of smaller articles into single articles, during the digitisation phase. This represents a clear problem for the consistency of our sample. As the sudden drop in 'T' could lead to an artificial, disproportionate increase in the frequency of European news relative to the total, given that the smaller articles are less likely to talk about Europe, and that the drop is not homogenous across sections. In order to avoid this drop and error, we dropped the very short articles, below 350 words, creating a subset archive of only longer articles. As shown in Figure A.4, this cleaning resulted in a much more consistent archive, in which the 2009 drop disappears. As we present in the next section, compared to the first attempt at E/T[7], this archive yielded a much more consistent frequency ratio.

Figure A.4: Cleaned *Der Spiegel* Archive



Source: Bruegel. Note: *Der Spiegel* after removal of shorter articles (150,000 articles).

La Stampa

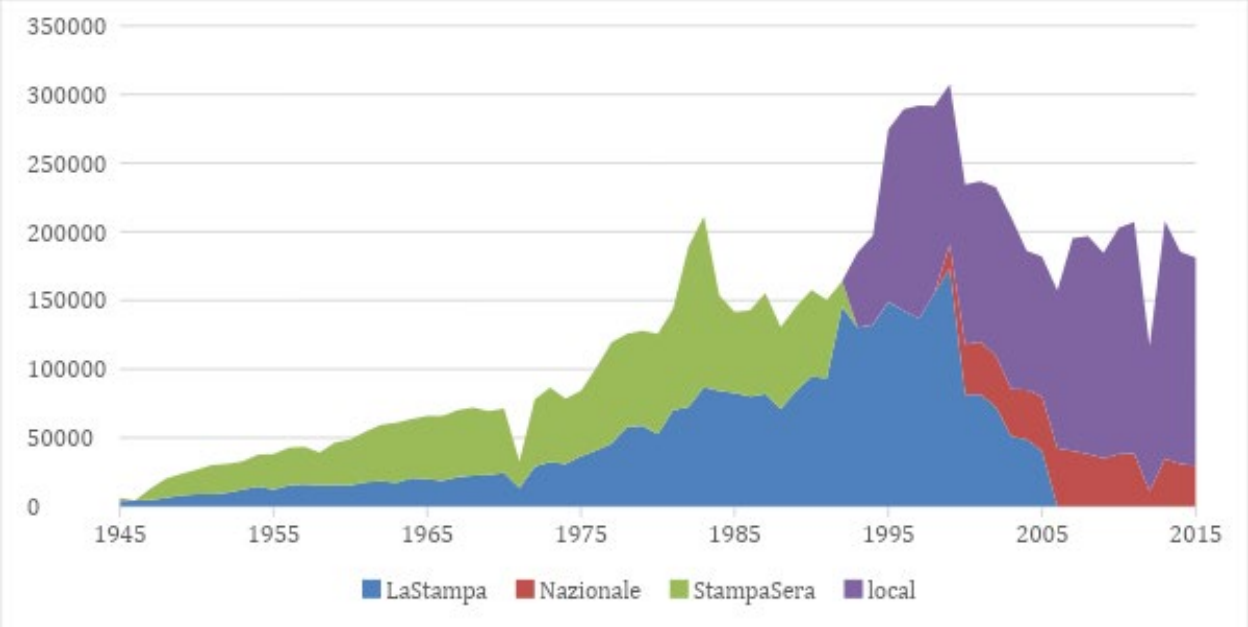
La Stampa's archive was by far the most complex to handle, both in size (9.9 million articles) and because of its composition. *La Stampa*, in fact, had many different editions, and a large number of articles belong to local news divisions, rather than the 'national' news. As mentioned, originally *La*

[7] Available at : https://www.bruegel.org/2019/07/talking-about-europe-die-zeit-and-der-spiegel-1940s-2010s/

Stampa was composed of two archives (websites), overlapping. The composition is summarised in the following table:

	Old Archive: 1945-2006 (archivio.lastampa.it)	New archive: 1992-2019 (archiviolastampa.it)
Local news	Stampa Sera (local and national)	La Stampa (local editions)
National	La Stampa	La Stampa (Nazionale)

Figure A.5: *La Stampa*'s raw archive

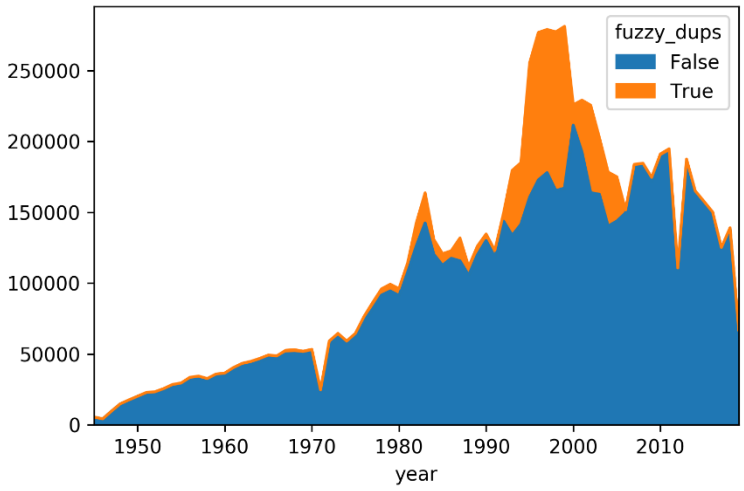


Source: Bruegel.

In order to check for duplicates, we deleted duplicates with identical titles and identical texts. Furthermore, we cleaned near duplicates (hence addressing the near-duplicates OCR errors) by calculating a measure of string similarity across articles. We used a measure of cosine similarity in order to detect near-duplicate documents. Cosine similarity interprets each document as a vector, in a space of dimensionality equal to the number of unique in the whole corpus. Cosine similarity measures the cosine distance between the document vectors. For efficient implementation, we calculated pairwise similarities of texts day by day. By doing so, we avoided creating big, highly sparse and cumbersome matrices, by reducing the dictionary's size, and eliminating above the 0.98

threshold. The logic of the algorithm, therefore, is to calculate the pairwise similarities for articles in the same day, and then keep only one of those that largely 'match' another one.

Figure A.6: Volume of fuzzy duplicates (based on cosine similarity) in *La Stampa*



Source: Bruegel.

In Figure A.6, we illustrate how in the case of *La Stampa's* archive, the detection of near-duplicates focused on the overlapping period (1992-2006), in which expectedly the two databases contained the same articles.

Moreover, the archive is inconsistently composed of different types of news: local and national. This composition is critical for our purposes. The mix of local editions (such as *La Stampa Torino*, *La Stampa Vercelli*, etc) is clearly distinguishable only in the new archive. The absence of classification for the previous part of the archive, between 1945 and 1992, remains problematic.

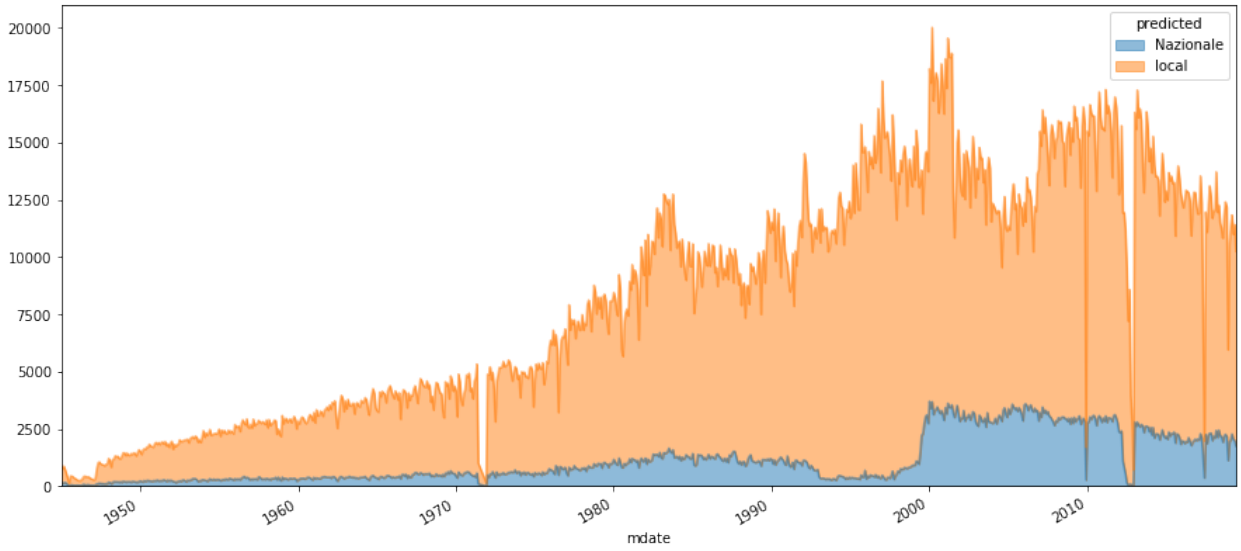
Whereas in the new archive the amount of local news, on average, represents 88 percent of the total, in the older archive, this information is absent. Unfortunately, other information that we scraped, potentially useful to solve this problem, proved to be too unreliable to be leveraged. For example, the page number of the articles in the newspaper was highly inconsistent and showed a number of errors probably due to the digitisation process. Hence, in order to classify the old archive (composed of *La Stampa* and *Stampa Sera*), we trained a machine learning model on the new archive, able to replicate the same division across the old archive.

We implemented a Linear Support Vector Classifier (SVC) based on the word frequency. Intuitively, the model classifies text into two labels ('local' or 'national'), based on the distribution of words in it. The predictor vectorises every term using a TFIDF (Term Frequency Inverse Document Frequency) model. We divided the 1992-2019 archive into two random sets, one for training and one for testing the model. Respectively, they represented 66.6 percent and 33.3 percent of the whole sample. In this fashion, we could implement a traditional training algorithm, relying on Python's *sklearn* library. The model yielded a test accuracy of 0.95.

We predicted the two labels by applying the trained SVC to the 1945-2006 archive. In Figure A.7, we present the archive's volume consistently divided into 'local' and 'national' news. Intuitively, this process simply allows us to exploit the partition of the new archive, to obtain the same in the old one.

This classification suffers from obvious limitations. It is time independent, and hence 'blind' to changes in language and writing style, and the labelling is only available for recent times. Despite this, to our knowledge, this was the most efficient method in dealing with highly unstructured text data, and therefore increasing the sample consistency across time.

Figure A.7: La Stampa's *archive consistently divided between "local" and "national" news via SVC classifier*



Source: Bruegel.

Annex 2 : Keywords

French

Communautés européennes, Communauté européenne du charbon et de l'acier, Communauté économique européenne, Communauté européenne de l'énergie atomique, Communauté européenne, Union européenne, Haute autorité, Commission des Communautés européennes, Commission européenne, Conseil de l'Union européenne, Conseil UE, Conseil européen, Conseil U.E., Assemblée parlementaire européenne, Parlement européen, Cour de justice européenne, Cour de justice de l'Union européenne, Cour des comptes européenne, Coopération politique européenne, Politique étrangère et de sécurité commune, Justice et affaires intérieures, Coopération policière et judiciaire en matière pénale, EUROPOL, Unité de coopération judiciaire de l'Union européenne, Union économique et monétaire, Union économique et monétaire européenne, Union monétaire européenne, Mécanisme européen de change, Système monétaire européen, euro, Monnaie unique, Monnaie commune, Zone euro, eurozone, Eurocratie, Institut monétaire européen, Banque central européenne, Système européen des banques centrales, Serpent monétaire européen, Service européen pour l'action extérieure, Politique étrangère et de sécurité commune, Politique de sécurité et de défense commune, Coopération politique européenne, Schengen, Traités de Rome, Traité de Maastricht, Acte unique européen, Acte unique, Traité de Lisbonne, CE, CECA, CEE, EURATOM, CEE, UE, UEM, ECU, IME, BCE, SEAE, PESC, PSDC, CPE

German

Europäische Gemeinschaften, Europäische Gemeinschaft für Kohle und Stahl, Montan-Union, Montanunion, Europäische Wirtschaftsgemeinschaft, Europäische Atomgemeinschaft, Europäische Gemeinschaft, Europäische Union, Hohe Behörde, Kommission der Europäischen Gemeinschaften, Europäische Kommission, Besonderer Ministerrat, Rat der Europäischen Union, Europäischer Rat, Gemeinsame Versammlung, Europäisches Parlament, Europäischer Gerichtshof, Gerichtshof der Europäischen Union, Gerichtshof der EU, Europäischer Rechnungshof Europäische Politische Zusammenarbeit, Gemeinsame Außen- und Sicherheitspolitik, Justiz und Inneres, Polizeiliche und justizielle Zusammenarbeit in Strafsachen, Europäisches Polizeiamt, Einheit für justizielle Zusammenarbeit der Europäischen Union, Wirtschafts- und Währungsunion, Europäische Wirtschafts- und Währungsunion, Europäische Währungsunion, Europäischer Wechselkursverbund, Europäisches Währungssystem, Europäische Währungseinheit, Euro, Einheitswährung, Gemeinschaftswährung, Eurozone, Eurokratie, Euroland, Euroraum, Europäisches Währungsinstitut, Europäische Zentralbank,

Europäisches System der Zentralbanken, Europäische Währungsschlange, European External Action Service, Common Foreign and Security Policy, Common Security and Defence Policy, European Political Cooperation, Schengener, Römische Verträge, Vertrag von Maastricht, Einheitliche Europäische Akte, Einheitliche Akte, Vertrag von Lissabon, EG, EGKS, EWG, EURATOM, EG, EU, EPZ, GASP, PJZS, Europol, Eurojust, WWUEWI, EZB, EEAS, CFSP, CSDP, EPC

Italian

Comunità europea, Comunità europea del carbone e dell'acciaio, Comunità economica europea, Comunità europea dell'energia atomica, Comunità europea, Unione europea, Unione europea, Alta Autorità, Commissione delle Comunità europee, Commissione europea - Consiglio dell'Unione europea, Consiglio UE, Consiglio europeo, Consiglio UE, Assemblea parlamentare europea, Parlamento europeo, Corte di giustizia europea, Corte di giustizia dell'Unione europea, Corte dei conti europea, Cooperazione politica europea, politica estera e di sicurezza comune, giustizia e affari interni, cooperazione di polizia e giudiziaria in materia penale, EUROPOL, Unità di cooperazione giudiziaria dell'Unione Europea, Unione Economica e Monetaria, Unione Economica e Monetaria Europea, Unione Monetaria Europea, Meccanismo di cambio europeo, Sistema Monetario Europeo, Moneta Unica, Moneta Comune, Eurozona, Eurocrazia, Istituto Monetario Europeo, Banca Centrale Europea, Sistema Europeo delle Banche Centrali, Serpente monetario europeo, Servizio europeo per l'azione esterna, Politica estera e di sicurezza comune, Politica di sicurezza e di difesa comune, Cooperazione politica europea, Schengen, Trattati di Roma, Trattato di Maastricht, Atto unico europeo, Atto unico, Trattato di Lisbona, Trattato CE, CECA, CEE, EURATOM, UE, UEM, ECU, IME, BCE, SEAE, PESC, PSDC, PESD, CPE

Annex 3: Topics extracted with LDA procedure

Der Spiegel

Label	Topic ID	Number of articles	Average keywords per article	Keywords
European	130	480	7.16	ewg, brüssel, europäisch, gemeinschaft, brüsseler, europa, land, prozent, frankreich, gemeinsam
European	121	421	6.80	prozent, wirtschaft, staat, land, regierung, unternehmen, deutschland, wachstum, investition, hoch
European	136	716	6.68	merkel, schäuble, kanzlerin, angela_merkel, politik, seehofer, koalition, union, deutschland, berlin
International politics	144	1328	6.27	england, britisch, brite, london, großbritannien, britische, londoner, englisch, engländer, europa
International politics	143	1595	6.12	russland, usa, westen, land, europa, nato, eu, moskau, welt, obama
International politics	111	65	5.72	frankreich, französisch, franzose, paris, französische, de_gaulle, pariser, europa, gaulle, de
International politics	119	71	5.19	italien, italienisch, italiener, rom, italienische, berlusconi, mailand, mafia, neapel, mailänder
Uncertain	108	234	5.17	bleiben, lassen, gelten, eher, stark, problem, folge, deutlich, zeigen, entwicklung
Uncertain	117	508	5.02	mal, sehen, leute, sagen, leben, einfach, denken, eigentlich, finden, welt
Uncertain	138	2015	4.89	alt, geschichte, scheinen, welt, lassen, mann, bleiben, gelten, vergangenheit, längst
Uncertain	114	60	4.59	polizei, terrorist, anschlag, polizist, täter, mann, ermittler, gruppe, waffe, beamte
Not included	120	154	4.52	stadt, bauen, architekt, haus, bau, bürgermeister, gebäude, projekt, entstehen, alt
Uncertain	133	338	4.45	geld, stiftung, steuer, vermögen, finanzamt, spende, fiskus, luxemburg, liechtenstein, firma
Not included	131	90	4.38	lafontaine, köhler, biedenkopf, rau, saar, saarbrücken, saarländisch, oskar_lafontaine, saarland, saarländer
International politics	145	3020	4.32	usa, amerikanisch, amerika, amerikaner, amerikanische, dollar, new_york, welt, europa, million_dollar
International politics	148	3425	4.27	japan, firma, japaner, unternehmen, japanisch, markt, industrie, japanische, produkt, konzern
Uncertain	139	2236	4.18	alt, jung, arbeiten, job, generation, junge, inzwischen, wichtig, schaffen, nennen
Uncertain	141	1708	3.97	unternehmen, konzern, manager, vorstand, mitarbeiter, aufsichtsrat, aktionär, siemens, übernehmen, übernahme
Uncertain	132	320	3.92	gewerkschaft, arbeiter, betrieb, arbeit, arbeiten, arbeitnehmer, unternehmen, streik, stunde, mitarbeiter
Uncertain	147	5104	3.90	krieg, general, armee, soldat, mann, offizier, lassen, truppe, beginnen, führer
Uncertain	129	372	3.86	gesellschaft, revolution, linke, marx, intellektuelle, politisch, revolutionär, philosoph, link, kapitalismus
International politics	125	84	3.80	österreich, wien, österreichisch, wiener, österreicher, wulff, kreisky, haider, hartmann, lucke
Uncertain	112	119	3.73	regierung, partei, parlament, wahl, premier, politisch, land, politiker, abgeordnete, opposition
Not included	128	255	3.70	buch, schreiben, leben, roman, autor, schriftsteller, vater, tod, frau, geschichte
Uncertain	137	375	3.68	strom, deutschland, energie, kraftwerk, prozent, kosten, rwe, anlage, unternehmen, deutsch

Not included	123	402	3.65	stadt, sterben, straße, opfer, tod, leben, gewalt, tote, krieg, töten
Not included	118	79	3.46	fahrer, straße, autobahn, fahren, unfall, lkw, auto, leber, autofahrer, organ
Uncertain	142	1108	3.43	ddr, osten, westen, sed, dresden, leipzig, ostdeutsch, ostdeutsche, sachsen, ost-berlin
Uncertain	127	82	3.38	telekom, post, sommer, deutschland, kunde, sender, zumwinkel, deutsch, sehen, wettbewerb
International politics	135	125	3.38	iran, teheran, land, saudi-arabien, iranisch, schah, scheich, dubai, golf, iranische
Not included	101	104	3.29	woche, lassen, treffen, stehen, halten, gespräch, müssen, monat, erklären, chef
Not included	140	538	3.20	könig, prinz, sohn, herzog, wagner, königin, schloß, vater, graf, de
Not included	134	64	3.16	roth, neumann, schäfer, bastian, meier, kelly, club, petra, marco, charlotte
Not included	124	161	3.10	tier, vogel, wolf, hund, art, jäger, leben, million, wald, natur
Not included	149	493	2.92	pferd, reiter, winkler, simon, reiten, gewinnen, rennen, tim, robert, aachen
Not included	146	936	2.80	fernsehen, zdf, ard, sender, programm, sendung, zuschauer, farbe, senden, photo

Le Monde

Label	Topic ID	Number of articles	Average keywords per article	Keywords
European	198	22986	5.28	président, angela_merkel, nicolas_sarkozy, dirigeant, chef_etat, bruxelles, juncker, vouloir, françois_hollande, monsieur_sarkozy
European	189	4021	4.98	commercial, accord, etat_unis, commerce, négociation, pays, omc, gatt, libre_échange, marché
European	193	27963	4.78	système, exemple, cas, principe, forme, technique, méthode, existe, formule, pratique
European	168	2505	4.52	ue, pays, union_européen, union, membre, etat_membre, adhésion, élargissement, bruxelles, candidat
European	192	19786	4.1	conseil, comité, conférence, membre, réunion, travail, ministre, réunir, organisation, représentant
International Politics	172	622	4.03	allemand, allemagne, bonn, français, gouvernement, france, chancelier, accord, traité, république_fédéral
Uncertain	171	2834	3.98	pouvoir, mesure, cas, contrôle, prévoir, système, condition, devoir, permettre, disposition
Uncertain	180	9247	3.92	grand, nouveau, année, important, devoir, pouvoir, permettre, partie, venir, également
International Politics	185	2328	3.89	serbe, kosovo, yougoslavie, belgrade, serbie, bosnie, yougoslave, international, guerre, albanais
Uncertain	170	2366	3.88	banque, financier, crédit, bancaire, établissement, prêt, marché, dette, risque, milliard_euro
Uncertain	190	20398	3.84	année, hausse, baisse, trimestre, croissance, rapport, augmenter, progression, progresser, atteindre
Uncertain	187	11240	3.81	doute, savoir, point, voir, faire, raison, agir, donner, ici, cas
International Politics	199	25548	3.77	américain, washington, état_unis, président, monsieur, canada, bush, congrès, administration, canadien
European	167	1298	3.77	politique, action, programme, objectif, plan, moyen, effort, domaine, développement, nécessaire

Uncertain	196	52332	3.71	français, france, paris, étranger, hexagone, anglais, partenaire, côté, national, position
Uncertain	174	4089	3.64	faire, temps, jour, devoir, fois, mettre, prendre, pouvoir, aller, commencer
Uncertain	175	2751	3.6	emploi, travail, chômage, entreprise, salaire, social, heure, jeune, salarié, temps
Uncertain	188	6936	3.52	ministre, français, estaing, france, monsieur giscard, monsieur, paris, visite, président république, entretien
Uncertain	186	5379	3.5	guerre, peuple, juif, histoire, résistance, paix, ennemi, hitler, victoire, mort
European	182	1510	3.44	état unis, américain, occidental, europe, atlantique, militaire, conférence, u.r.s.s., défense, washington
European	177	1671	3.36	pétrole, prix, production, acier, pétrolier, haute autorité, million tonne, charbon, sidérurgie, producteur
Uncertain	194	12252	3.33	droit homme, prison, justice, peine, autorité, condamner, arrêter, libye, prisonnier, police
Not included	195	8769	3.26	religieux, catholique, chrétien, religion, église, eglise, pape, islam, musulman, ii
Uncertain	183	2261	3.18	satellite, programme, spatial, espace, lancer, américain, lancement, mission, lanceur, devoir
European	165	941	3.17	aide, fonds, financier, prêt, programme, accorder, aider, financement, soutien, financer
European	178	4556	3.16	ministre, gouvernement, etat, ministère, annoncer, budget, économie, bercy, finance, dossier
Uncertain	179	8840	3.14	mars, avril, mai, février, janvier, mois, juin, fin, premier, annoncer
European	166	1350	3.13	accord, négociation, discussion, accepter, obtenir, négocier, compromis, engager, point, signer
European	176	2070	3.12	liste, ps, rpr, udf, socialiste, président, député, opposition, verts, campagne
Uncertain	191	5780	3.12	produit, environnement, france, ogm, utiliser, industriel, culture, consommateur, interdire, déchet
Not included	197	10616	3.06	groupe, média, publicité, communication, publicitaire, canal+, agence, presse, français, société
Not included	151	139	3.05	france, homme, grand, général gaulle, français, gaulle, général, faire, savoir, vouloir
International Politics	164	888	3.03	allemand, allemagne, berlin, bonn, europe, rhin, ouest, république fédéral, fédéral, rfa
Not included	150	160	3.03	monde, homme, vivre, vie, jamais, savoir, mort, peur, temps, croire
Uncertain	161	238	3.02	socialiste, parti, communiste, gauche, politique, parti communiste, parti socialiste, socialisme, dirigeant, mouvement
Not included	162	574	3.01	national, administration, pouvoir, état, organisation, service, organisme, mission, institution, gestion
Not included	135	184	3.01	rapport, étude, résultat, enquête, note, publier, analyse, chiffre, expert, point
Uncertain	163	932	2.98	congrès, organiser, débat, mouvement, réunir, représentant, réunion, participer, thème, tenir
Uncertain	158	456	2.96	société, groupe, capital, banque, participation, général, français, actionnaire, financier, assurance
Uncertain	181	6157	2.94	petit, ville, ici, jour, venir, rue, heure, maison, village, habitant
Not included	184	2686	2.91	vin, grand, année, marque, bouteille, qualité, bordeaux, alcool, bon, domaine
Not included	132	221	2.89	euro, action, titre, groupe, valeur, résultat, hausse, annoncer, million euro, cours
Not included	173	1530	2.87	jeu, modèle, machine, ordinateur, appareil, microsoft, logiciel, système, apple, image

Uncertain	159	178	2.86	parti, politique, gouvernement, libéral, coalition, conservateur, droite, formation, monsieur, ministre
Not included	154	472	2.86	jour, fois, faire, mettre, déjà, bataille, prendre, semaine, face, tenter
Not included	169	1462	2.83	justice, affaire, juge, tribunal, avocat, cour, condamner, procès, judiciaire, procédure
Not included	139	446	2.76	déclarer, mardi, lundi, mercredi, jeudi, indiquer, vendredi, annoncer, estimer, affirmer
Not included	160	590	2.76	art, exposition, musée, artiste, grand, oeuvre, œuvre, tableau, galerie, objet
Uncertain	136	118	2.75	vouloir, faire, pouvoir, entendre, devoir, savoir, demander, accepter, prêt, demande

La Stampa

Label	Topic ID	Number of articles	Average keywords per article	Keywords
European	194	59782	5.02	europeo, unione, europa, paesi, comune, parlamento, strasburgo, comunità, nazionale, integrazione
European	158	331	3.98	presidente, europeo, vertice, unione, ue, ciampi, europa, paesi, presidenza, prodi
European	145	141	3.62	italia, deficit, economico, economia, pil, ciampi, crescita, politico, tesoro, finanziario
European	132	113	3.52	cee, l', comunità, comunitario, commissione, europeo, bruxelles, paesi, italia, lira
European	176	10286	3.41	accordare, trattativa, volere, soluzione, chiedere, compromettere, proporre, incontrare, accettare, negoziare
European	192	22556	3.4	politico, europa, potere, nazione, guerra, potenza, proprio, america, né, occidente
Uncertain	187	14783	3.28	prezzo, aumentare, costare, aumento, caro, lira, costo, tariffa, medio, benzina
Uncertain	199	68968	3.25	rispettare, aumentare, registrare, crescita, calere, crescere, dato, incrementare, periodare, positivo
European	181	16136	3.22	accordare, inteso, firmare, accordo, raggiungere, trattare, patto, entrare, prevedere, impegnare
Not included	175	4556	3.15	incontrare, presidente, visitare, roma, colloquio, esteri, colloquiare, ambasciatore, problema, rapporto
International politics	180	10391	3.06	vienna, annunciare, austriaco, austria, decisione, venire, chiedere, invitare, appellare, dovere
Uncertain	151	931	3.06	prestito, debito, credito, pagare, banca, pagamento, contare, interesse, finanziario, mutuo
Uncertain	156	445	3.05	finanziaria, riformare, tremonti, manovrare, pensione, economia, intervento, roma, misura, ministero
Uncertain	184	7153	3.02	produrre, qualità, produzione, consumatore, produttore, mercato, alimentare, fruttare, consumere, latta
Uncertain	167	1330	3	prodi, ulivo, politico, d'_alema, presidente, sinistro, partire, leader, ds, romano prodi
European	165	2588	2.98	proporre, comitato, proposta, riunione, progettare, commissione, presentare, rappresentare, iniziativa, documentare
Uncertain	152	255	2.93	lavoratore, sindacato, dipendere, azienda, sindacare, scioperare, sindacale, contrattare, occupazione, lavorare
European	120	469	2.93	europa, l', europeo, italia, continente, unito, paesi, unico, l', vecchio_continente

Uncertain	173	4614	2.92	sviluppare, ricercare, nuovo, settore, sistemare, qualità, mercato, investimento, tecnologia, innovazione
European	140	129	2.82	crisi, finanziario, mercato, economia, debito, riforma, grecia, politico, bce, greco
Uncertain	170	6030	2.8	potere, dovere, possibile, possibilità, trovare, difficile, venire, altro, attuale, almeno
Uncertain	198	26753	2.79	malattia, salute, farmaco, caso, curare, usare, malato, medico, rischiare, colpire
Uncertain	191	23734	2.79	lingua, nome, parlare, valore, parola, venire, usare, diverso, scrivere, italiano
Not included	183	7888	2.76	de, politico, partire, pei, socialisto, partito, craxi, comunista, segretario, psi
European	155	893	2.76	sanzione, controllo, dovere, multare, ricorrere, controllare, provvedimento, rispettare, multa, pagare
European	147	132	2.7	politico, partire, sinistro, socialisto, destro, elezione, leader, votare, partito, popolare
Uncertain	159	1724	2.68	importare, ottenere, buono, risultato, perchè, risultare, commentare, potere, spiegare, positivo
International politics	148	124	2.62	politico, partire, comunista, congresso, socialisto, democratico, discorrere, unità, partito, libertà
European	150	662	2.62	italia, francia, germania, olanda, europeo, spagna, gran_bretagna, inghilterra, europa, svezia
Not included	154	457	2.61	né, volere, sapere, credere, leggere, morale, sentire, proprio, parola, vero
Not included	142	194	2.61	fiscale, tassa, impostare, reddito, tassare, iva, pagare, contribuire, imposta, fisco
Uncertain	179	6409	2.6	azienda, produzione, settore, produrre, industriale, stabilimento, industriare, società, mercato, fatturare
Uncertain	160	1132	2.57	provincia, comuni, territorio, regione, provinciale, presidente, locale, ente, sindaco, comune
International politics	164	1495	2.57	spagnolo, spagna, madrid, europeo, barcellona, argentina, de, portoghese, europa, lisbona
Not included	182	11823	2.56	corso, correre, formazione, sede, giovane, attività, professionale, centro, tecnico, informazione
Not included	162	2978	2.54	terzo, punto, °, secondo, quarto, classificare, primo, ultimo, posizione, precedere
Uncertain	193	17134	2.5	chiesa, papa, cattolico, religioso, cristiano, vaticano, dio, vescovo, don, ii
Uncertain	134	161	2.49	progettare, area, strutturare, realizzare, recuperare, edificio, centrare, comune, sede, ospitare
European	143	150	2.48	impresa, azienda, piccolo, industriale, imprenditore, imprendere, settore, presidente, economico, nuovo
Not included	146	445	2.44	arrivare, potere, venire, mille, lavorare, dieci, dovere, mettere, spiegare, dare
Not included	163	1436	2.44	associazione, iniziativa, progettare, fondazione, attività, sociale, aiutare, progetto, promuovere, contributo
Uncertain	188	19258	2.44	bianco, colorire, rosso, nero, terra, lucere, occhio, vedere, vecchio, antico
Uncertain	186	16439	2.42	strada, quartiere, venire, abitare, poco, cittadino, centrare, locale, piccolo, gente
Not included	196	20835	2.41	sanremo, imperia, ventimiglia, fiore, nizza, sanremese, bordighera, riviera, mare, comune
Not included	177	4601	2.41	tv, rai, programmare, televisivo, canale, televisione, trasmissione, dirigere, trasmettere, rete
Uncertain	139	127	2.36	inchiesta, indagine, società, guardia_finanza, imprenditore, procurare, denunciare, scoprire, pm, vicenda
Not included	168	1985	2.36	polizia, arrestare, drogare, carcere, agire, venire, organizzazione, mafia, arrestato, indagine

Not included	197	23638	2.32	teatro, spettacolo, tel, scena, compagnia, regio, sala, attore, l', cinema
Not included	135	123	2.28	gestione, società, servizio, comune, gestire, rifiuto, servizo, consorzio, costo, impiantire

Bruegel, Rue de la Charité 33, B-1210 Brussels
(+32) 2 227 4210
info@bruegel.org
www.bruegel.org