DHQ: Digital Humanities Quarterly
2021
Volume 15 Number 15

Finding Narratives in News Flows: The Temporal Dimension of News Stories

Blanca Calvo Figueras <blancacalvofigueras_at_gmail_dot_com>, University of Gronigen
Tommaso Caselli <t_dot_caselli_at_rug_dot_nl>, University of Groningen
Marcel Broersma <m_dot_j_dot_broersma_at_rug_dot_nl>, Centre for Media and Journalism Studies, University of Groningen

Abstract

Previous studies indicate that the capacity of media to influence the salience of issues in the public realm is strongly dependent on specific attributes that characterize these issues. In this work, we investigate two *internal* aspects of issue types related to the attribute of *duration*. First, we address whether news stories belonging to different issue types can be identified and represented using a set of quantifiable temporal dimensions (i.e. lifespan, intensity, and burstiness). Second, we conduct a qualitative analysis to investigate whether news stories of different issue types have different *narrative patterns*, regardless of their specific topic. We use a corpus of 50,385 political news articles in Spanish from 2018 as a case study, and propose a novel system to aggregate the articles into stories. Our results show that stories belonging to different issue types do have distinguishing behaviours, especially along the intensity dimension. At the same time, the qualitative analysis indicates a tendency to associate narrative patterns to issue types. This analysis shows the potential of using news stories as research units to study framing strategies.

INTRODUCTION

News stories shape the public perception of reality. Agenda-setting theory suggests that readers not only get informed through news media but also get to grasp the salience of a specific issue on the basis of the amount of information that is published about it [McCombs and Shaw 1972] [Scheufele and Tewksbury 2006] [Wanta Ghanem 2006]. Whereas mass media could function as gatekeepers that determined which issues became news, and thus public [Shoemaker and Reese 2013], this is more complicated in the current online decentralized information ecology. 1

The volume of journalistic articles published by print media has been increasing steadily since the emergence of the World Wide Web [Gómez-Rodríguez et al. 2014]. The transition from paper to the digital format has removed the barrier of space limits allowing online media to publish a virtually unlimited amount of documents. News organizations compete with each other but also with other agents about the attention of news users. In this attention economy the speed of the news cycle has tremendously risen. Studying how news circulates in the information ecology and the dynamics of how broader news stories are constructed and put on the public agenda is therefore increasingly important [Broersma 2019]. At the same time, this limitless space can be used by digital media to silence some topics and emphasize others [Holton and Chyi 2012] [Lee et al. 2017]. In the age of digital news, the frequent publication of articles about a story has become a source of power in digital reporting. Thus, it becomes important to regard the *temporality* of news and the *timing* that newspapers use to inform about a story. 2

A relevant terminological distinction is the difference between a news article and a news story. A news article is a *unitary mention* of a potentially larger sequence of events. A news story is a *unitary sequence* of events that can be reconstructed by aggregating multiple unitary mentions (i.e. news articles). For instance, a news article on the arrest of a suspect (i.e., a unitary mention) can be embedded as a component in a news story of a specific crime. Focusing on 3

news stories allows us to understand how news organizations produce meaning by making sense of various news events and interpreting them as parts of larger societal developments and discussions.

4 The lifespan of digital news articles, i.e. unitary mentions, has been previously investigated by observing how long a specific news article remains in a salient position (e.g. the front page) [Karlsson and Strömbäck 2010] [Lee et al. 2014] [Castillo et al. 2014]. However, in the digital realm, news stories have gained more prominence, as articles remain accessible and are connected to each other [Bright 2014]. This new environment points at the need for aggregation strategies that facilitate the study of news reporting [Trilling and van Hoof 2020]. As Bødker and Brügger (2018) observe, digital newspapers have switched from an archival logic to constructing stories based on networks of links. Hereby, they put the focus on the story as an organizing principle, rather than on the journalistic filtering of the front page. Following Bødker and Brügger (2018), we consider stories as our minimal units of analysis (cf. [Widholm 2016]).

5 Focusing on stories rather than on news articles enables us to investigate the temporal dimensions of newspapers' online reporting. We build on Soroka (2002), who developed a typology that focuses on the attributes of the reported issue when evaluating the effects of agenda-setting. This framework argues that the capacity of media to influence the salience of a particular issue depends on its attributes, namely: (1) *obtrusiveness*; (2) *abstractness*; (3) *dramatism*; and (4) *duration*. Using the first three attributes (i.e. obtrusiveness; abstractness; and dramatism) Soroka develops a three-fold issue-typology that distinguishes between *sensational*, *governmental* and *prominent* issues. Duration is not fully included in his framework, conceivably, because issues on their own do not have an attribute of duration. Issues need to be identified as stories in order to have a temporal dimension. In this article, we investigate how duration contributes to agenda-setting through three quantitative dimensions, namely: (1) *lifespan* (number of consecutive days that articles keep being published), (2) *intensity* (amount of articles per day), and (3) *burstiness* (speed of increase/decrease of the number of articles). Our goal is to examine whether the three types of news issues defined by Soroka can be differentiated along these quantitative temporal dimensions.

6 We design a methodology that extracts all the salient stories in digital newspapers over a fixed period of time. Our approach uses *k*-means clustering on the set of articles of each relevant time span of the target period. It then links the clusters through time by looking at the cosine similarities of their content. The stories are later manually assigned to one of the three types: sensational, governmental or prominent, according to a specific decision tree (see Fig. 4). We have used the data of four Spanish newspapers during the entire period of 2018, scraped from their archives on the section about *politics*, and used weeks as relevant time spans to aggregate single news articles. What we present here is a case study on journalism, while the proposed methodology can be applied to other dimensions of variations or types of text. For instance, recent initiatives on digitization and analysis of historical newspapers (e.g., NewsEye, Impresso) could use our methodology to explore the historical evolution of issue types. On the other hand, its application (with minor changes) to literary texts can support computational textual studies in identifying coherent sequences of events and thus contribute to the advancements in the area of computational narratology.

7 The aggregation method we propose works as a proxy to investigate and visualize the eventual narrative patterns of issue types. Following narratology frameworks [Bal 1997] [Kafalenos 2006], we adopt the notion of narrative patterns as fixed, recurring global structures that yield coherence to the stories configuring the arrangements of the events from the beginning to the end.

8 This paper makes three main contributions: (1) it tests a semi-automated approach to aggregate mentions, i.e. news articles, into sequences of events, i.e. stories; (2) it empirically validates Soroka's framework of issue types; (3) it provides a qualitative analysis of the narrative patterns in the different types of stories, focusing on the timings and content of each part of the plot. This renders the interconnectedness of news stories visible, and shows the temporal structure of various types of stories and how this impacts meaning-making. Furthermore, the identification of narrative patterns can support the investigation of variations in writing styles of the plot of a news story at different moments in time.

9 In the upcoming section, we introduce Soroka’s theoretical framework and the quantitative measures we are going to use for this work (Section 2). We then discuss previous approaches to story extraction and present our methodology

and data, as well as the evaluation of the methodology (Section 3). We then present the quantitative results (Section 4) and our qualitative analysis on narrative patterns (Section 5). Finally, we illustrate our conclusions and future directions (Section 6).

EMPIRICAL APPROACH TO SOROKA'S ISSUE-TYPOLOGY

10 The concept of *agenda-setting* was first used in 1922 by Walter Lippmann. However, it was developed into a formal theory by Maxwell McCombs and Donald Shaw in 1972. They described agenda-setting as the ability of news media to influence the relative importance of topics on the public agenda. Their research on the US presidential campaign of 1968 showed that voters tend to share the media's perception of what is important [McCombs and Shaw 1972].

11 Later work suggests that the impact news media have on certain issues varies depending on the attributes of these issues [Ball-Rokeach and DeFleur 1976] [Downs 1972] [Wanta Ghanem 2006] [Zucker 2017]. Relying on this literature, Soroka (2002) identifies four main attributes that can make the media's influence differ:

1. *Obtrusiveness* refers to the direct experience that individuals on a large scale have with an issue. It has been argued that the less direct experience citizens have with an issue, the more they have to rely on the media for information and interpretation.
2. *Abstractness* opposes concreteness, suggesting that the potential for media effects is diminished when an issue is abstract, since it is more difficult for the audience to visualize and relate to.
3. *Dramatism* refers to the existence of a dramatic event playing a significant role in the issue becoming news.
4. *Duration* represents the time a certain issue has been in the news. Some theories suggest that the longer an issue stays in the media, the lower the chances of media having an influence on the public debate, because people tend to have made up their minds or get bored of the issue [Downs 1972] [Zucker 2017].

12 Soroka proposes to combine the attributes of obtrusiveness, abstractness and dramatism to study the media's agenda-setting capacity. The three types of issues he defines using these attributes are (1) sensational, (2) governmental, and (3) prominent issues.

13 *Sensational issues* have little observable impact on the majority of the population (i.e. unobtrusive). They are concrete and, very often, they are the result of a dramatic event that attracted media's attention. They are mostly put on the public agenda by media coverage. A murder or a corruption case are examples of such issues.

14 *Governmental issues* are perceived as unobtrusive (although they might have a relevant long term impact). They are usually led by political actors, such as political institutions and policy-makers. They are mostly abstract issues that are difficult to comprehend without expertise. They rarely offer any dramatic component. Public debt or the national budget are examples of these issues.

15 *Prominent issues* affect a significant amount of people directly; they are obtrusive and concrete. These kinds of issues become news as a result of real-world experiences, and their outcome can result in a perceivable change for many people. The rise of salaries or pensions are examples of these issues.

16 Soroka's typology puts great emphasis on who or what puts news on the public agenda in the first place. In short, sensational issues are usually put on the public agenda by the media, governmental issues are led by political actors, and prominent issues by real-world conditions.

17 Soroka (2002) describes the possible influence of the attribute duration, however, he does not use it in the definition of his typology. In this contribution, we investigate if the attribute of duration is also a distinctive feature between these three types of news stories. We portray duration through three quantitative dimensions:

1. *Lifespan* is defined as the number of days in which articles about a news story keep being published.
2. *Intensity* corresponds to the number of articles published about a news story in one day.
3. *Burstiness* indicates the speed with which stories go from incipient to the climax. It is measured with the Fano Factor, a ratio between the variance and the mean of counts.

Following Downs (1972), Zucker (2017) and Soroka (2002), we expect sensational issues to be put on the public agenda by the media. Consequently, we expect this type of stories to have a short lifespan, a high intensity and a high burstiness. By comparison, we expect prominent and governmental issues to have a longer lifespan and a lower intensity. In the following section, we introduce a semi-automated methodology for the empirical validation of these hypotheses. 18

AGGREGATION OF NEWS STORIES AND ISSUE TYPES IDENTIFICATION

In this section, we describe the previous approaches to automated systems for news stories extraction (Subsection A); we then present our semi-automatic method to identify news stories (Subsection B), the dataset (Subsection C), the guidelines for the manual classification (Subsection D), and the evaluation of the results (Subsection E). 19

Previous approaches to automated systems for news stories extraction

Identifying news stories within a large corpus of news texts is not a trivial task. An ideal automatic system should be able to identify the linguistic mentions of the reported events and of the associated actors in documents published at a time *t*, and then connect them to other mentions in other documents published at a later time, *t+1*, through a cross-document coreference relation [Lee et al. 2012]. Due to the complexity of the task and the low performance of such automatic systems [Beheshti et al. 2017], previous work has mainly employed topic detection and tracking techniques through the use of clustering. Clustering is an unsupervised machine learning technique that automatically organizes a collection of objects into smaller coherent groups. The end goal is to have clusters whose elements are as similar as possible between them (intra-cluster similarity) and as dissimilar as possible from the elements in the other clusters (inter-cluster similarity). This implies that we cannot determine a priori which criteria are going to define the different clusters. 20

Clustering has been mainly employed to divide a set of news according to their topic (e.g. politics, sports, etc.) [Huang 2008]. However, the use of clustering to identify news stories is very challenging. Some approaches have tried to improve the performance of this task by incorporating properties of news stories, such as timing features [Khy et al. 2008] [Yang et al. 1998] [Trilling and van Hoof 2020], named entities [Kumaran and Allan 2004] and burstiness [He et al. 2007]. In our case, timing and burstiness are measures we are trying to obtain from our system and, consequently, they cannot be part of the design. Named entities have also limited value, since the actors or places appearing in different news stories can be common between stories. 21

Later approaches have used Latent Dirichlet Allocation models to deal with news streams [Guo et al. 2013] [Hong et al. 2016] [Rao et al. 2016]. These techniques, however, focus on the extraction of topics rather than stories. 22

Another approach is timeline generation: automatically generated chronological sequences of events. This is done by using the temporal information attached or contained in single news documents. The timelines are usually centred on a target entity, giving rise to so-called entity-centric timelines [Shahaf and Guestrin 2010] [Baraniak 2019] [Minard et al. 2015]. Other work has proposed to go from timelines to storylines by taking into account the timelines of the actors, i.e. entities, co-participating in event mentions [Laparra et al. 2015] [Laban and Hearst 2017]. Storylines are thus represented as intersections (i.e. co-participation) of independent entity-centric timelines. Vossen, Caselli, and Kontzopoulou (2015) used the basic concepts of the narratology framework to bridge the relations between events of different timelines, starting with the most salient event of the storyline, i.e. the climax point. 23

A successful model for story tracking is the European Media Monitor [Gey et al. 2009]. This system runs every 10 minutes, and clusters all the news articles published during the last 4 hours. In every run, the system picks the clusters with the biggest number of articles, and compares them to the clusters that already existed in the last run. If two clusters have at least 10% of their articles in common, they are merged, becoming a single story. This system allows for a flexible time span and is language independent. 24

In this contribution, we use a similar model to the one in Gey, Karlgren, and Kando (2009). However, given that our data is not real-time, we employ a similar approach to Vossen, Caselli, and Kontzopoulou (2015) to account for time. For 25

every pre-defined *time window*, we identify the most salient stories by clustering the corresponding news articles. Once this is completed, we approximate Vossen, Caselli, and Kontzopoulou (2015)'s proposal of *bridging relations* across time windows by means of similarity scores between pairs of story clusters (see Subsection B) to reconstruct the full story timeline.

A system for news stories extraction

Our pipeline for story extraction consists of (1) clustering the news articles of every week (i.e. the time window) and (2) connecting the clusters across relevant time spans, i.e. weeks, using each of the valid clusters as the climax of a story. The method is language independent, as clustering does not require any language-specific resource, such as lexicons or parsing models.[1] 26

Given a set of news articles from a fixed span (one year), we divide the corpus into smaller units (weeks). For each week, *k*-means clustering is applied.[2] The purpose of the clustering process is to detect the most salient stories of each week, spotting both short and long-span stories. We assign the number of clusters in a semi-automated way by manually reviewing the results of each week. 27

The value of the number of clusters per week (*k*) is assigned using the Elbow method. This method computes the clustering for *k* in the range from 1 to 30 (the range has been empirically chosen through a set of preliminary experiments) and stores the sum of squared errors (SSE) for every set of clusters. The optimal number of clusters is the one after which the SSE starts decreasing slower. This number can be easily detected by plotting the SSE in the y-axis and the number of *k* in the x-axis, which creates an elbow shape in the desired value. 28

In a collection of news articles, it is to be expected that not all articles are going to be part of a story. These isolated articles are gathered by the algorithm in one or two general clusters with very low SSE. These general clusters are detected and deleted using Silhouette analysis [Rouseeuw 1987]. This method plots all the documents in a cluster according to the silhouette coefficient. This coefficient compares the average distance from all data points in the same cluster and the average distance from all data points to the closest cluster. The resulting plot shows which clusters have gathered the articles that do not belong to any story. In Figure 1, clusters 8 and 10 are general clusters and will be removed from our data for further analysis. 29

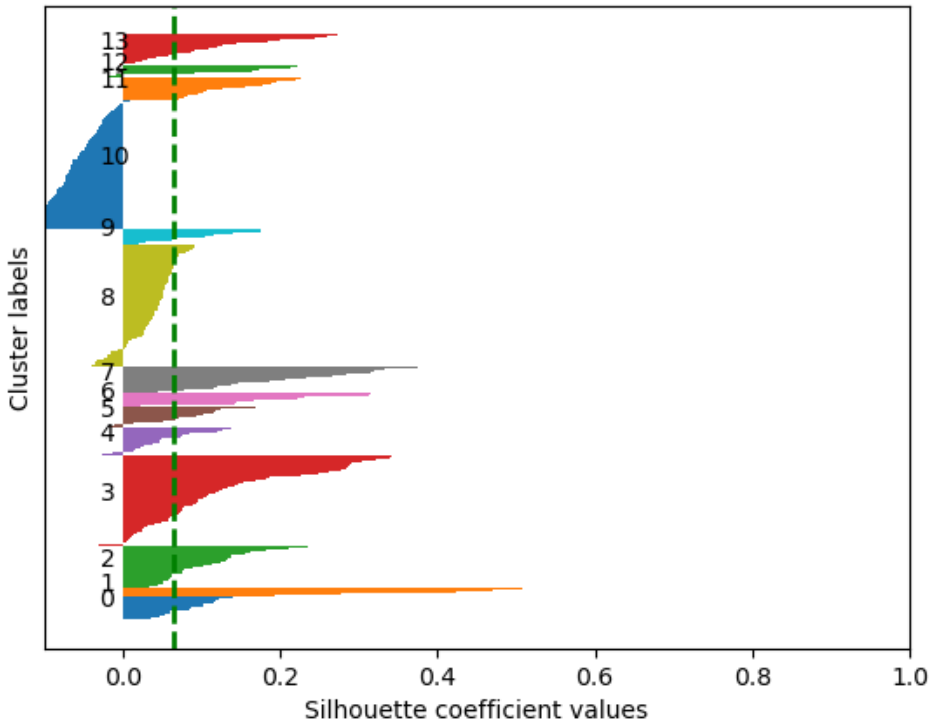

Figure 1. Silhouette coefficients of the clusters of one week

In some cases, the same story is split into more than one cluster. To deal with this issue, we manually merge these clusters by comparing the keywords of their centroids. The centroids are the articles closest to the centre of the cluster and consequently the most representative of its content. In this step, some data cleaning is performed: clusters of international news are excluded (see Subsection C for details), clusters with non-political topics are excluded (e.g. the lottery, Christmas celebrations), and clusters with less than 15 articles are deleted. The process of manual revision takes around 5 minutes per week. 30

The clusters thus extracted are assumed to be the climax of different stories: they have been detected in weeks in which they are salient enough to form a cluster of over 15 articles. Nonetheless, stories are not limited to one week. Hereafter, for each climax (i.e. each extracted cluster), we link articles of the colliding weeks to the climax using cosine similarity. Henceforth, we will refer to the remaining clusters as climaxes (*w*). 31

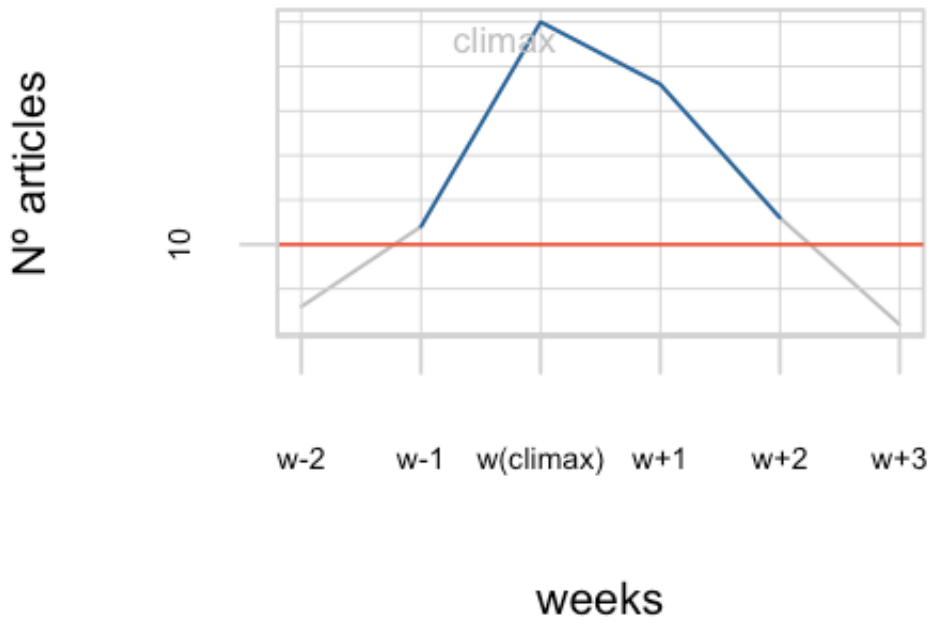

Figure 2. Approach followed to trace the articles belonging to each story

It is often the case that the same story is detected as a climax in more than one week. In these cases, it is reasonable to expect that, after tracing the articles of the colliding weeks, an important proportion of the articles included in these stories are going to be part of both clusters. If more than 33% of the articles contained by a new story are already part of a previous story, and vice versa, the system merges them. The threshold of 0.33 is chosen based on empirical validation on our data. 32

This last step runs for the 52 weeks contained in the study, resulting in stories with very different lifespans and intensities that are the basis of our analysis. The method allows for one article to be part of more than one story. Nonetheless, this happens rarely.[3] 33

Data

To extract the lifespan and intensity of the news stories, it is important to collect an exhaustive dataset of news that gathers all the articles published by one or more media outlets on certain topics during a fixed period. Available news corpora are usually from several media, time periods, and rarely care about exhaustivity. For this reason, we created our own corpus by scraping all articles in 2018 from four major Spanish newspaper outlets: *El Pais*, *El Mundo*, *ABC*, and *El Periódico*. We narrowed the focus to news about politics, collecting for *El Mundo* and *ABC* news from the section *España*, and for *El Pais* and *El Periódico* from the section *Política*. Political news about Spain are included in these categories, with the main difference that *El Pais* and *El Periódico* also include news about international politics in it. International news articles are removed during the cleaning of clusters (see Subsection B). An overview of the resulting corpus is illustrated in Fig. 3. 34

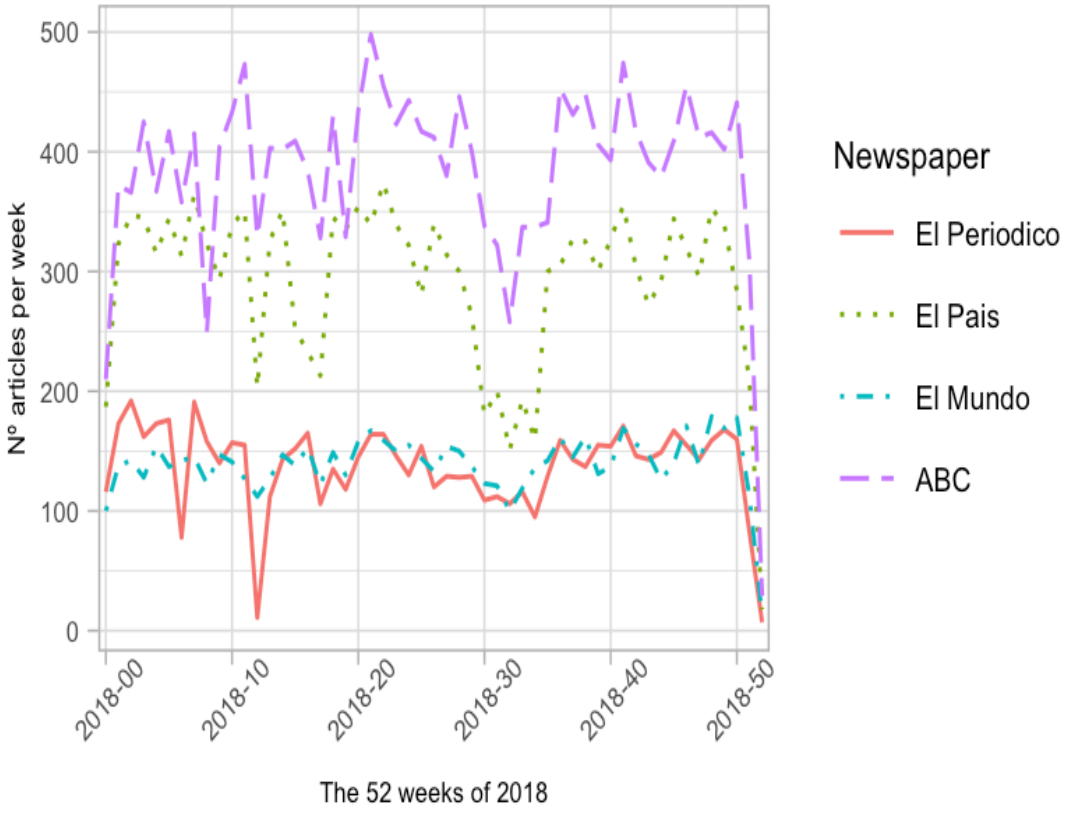

Figure 3. Number of articles per week for each newspaper

Overall, the corpus contains 50,385 news articles: 15,245 from *El Pais*, 7,401 from *El Mundo*, 20,448 from *ABC*, and 7,291 from *El Periódico*. For each news article, the following data are collected: the date of publication, the title, the text, the URL, and the keywords (as specified by each news outlet). In total, an average of 950 articles per week were published. *ABC* is the biggest publisher, with an average of 385 articles per week. *El Pais* follows with an average of 293. *El Periódico* and *El Mundo* published on average 137 and 139 articles a week, respectively.[4] 35

Guidelines for typology labelling

The manual annotation of the news stories was performed following a decision tree based on four questions. The annotation was conducted by one of the authors who is a Spanish native speaker and has extensive knowledge on Spanish politics. Figure 4 graphically illustrates the decision tree used for the manual annotation of the stories into issue types. 36

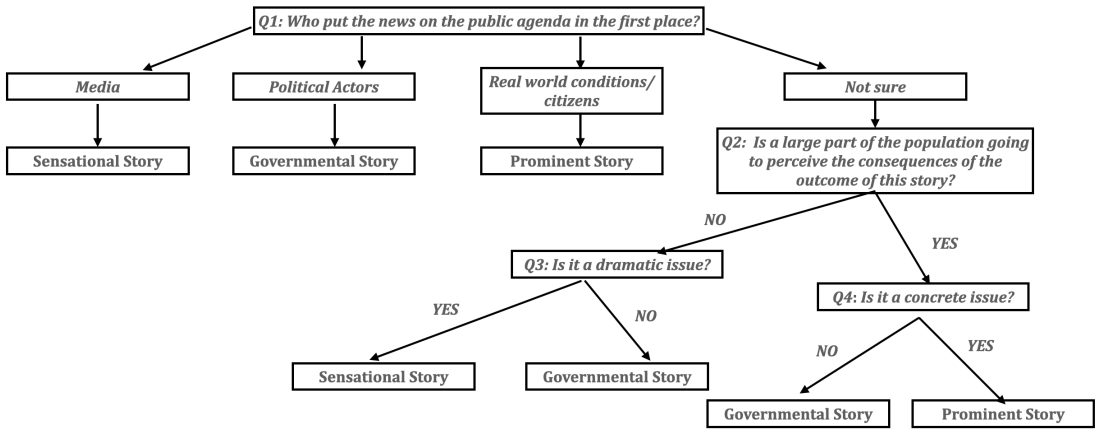

Figure 4. Decision tree used to manually assign news stories to Soroka's (2002) issue types

The first question of the tree distinguishes the actor(s) who put(s) the news in the public agenda (Q1:'"Who put the news on the public agenda in the first place?"'). If the answer to this question is "Media", "Political Actors" or "Real-World Conditions/Citizens", then the story is classified as sensational, governmental or prominent, respectively. In case the answer cannot be clearly identified, the annotator has to move to the second branch of the tree, which checks for the impact of the story in the public by assessing the obtrusiveness of the story (Q2:'"Is a large part of the population going to perceive the consequences of the outcome of this story?"'). A negative answer restricts the choice either to sensational or governmental issues. The exact labelling is triggered by a further question using the dramatism attribute (Q3:'"Is it a dramatic issue?"'). A positive answer leads to a sensational story, otherwise the story is labelled as governmental. In case of a positive answer to Q2 the annotator is posed with another question focused on the concreteness of the story (Q4: '"Is it a concrete issue?"'). In case of a positive answer, the story is labelled as prominent, otherwise it is considered as a governmental one. 37

Evaluation

The performance of the semi-automated system for news stories extraction has been evaluated with a subset of 2,666 manually annotated news articles. The gold standard was created by one of the authors by inspecting the main issue in each article and assigning each of them to a story. The measures used to evaluate are Purity and Normalized Mutual Information (NMI). 38

Purity is a measure of the extent to which clusters are well formed, i.e. contain a single class. To compute Purity, each of the clusters is assigned to its most frequent class, according to the gold standard. Then, the accuracy of this assignment is measured. The Mutual Information score is a complementary measure that assesses the quality of the clustering. In particular, MI provides a measure of the reduction in the entropy of clusters that we compare to the gold labels. This measure is sensitive to the number of clusters. NMI is the normalization of this value between 0 and 1, where 1 means a perfect classification. 39

The evaluation scores resulted in a Purity of 0.896 and NMI of 0.625. To gain a better understanding of the limitations and reliability of our approach we conducted a manual exploration of the aggregated stories. In particular, we used a contingency matrix to see the distribution of errors among the gold and automatically generated clusters of stories. Figure 5 visually illustrates the contingency matrix-based analysis of the errors against the manually aggregated stories. 40

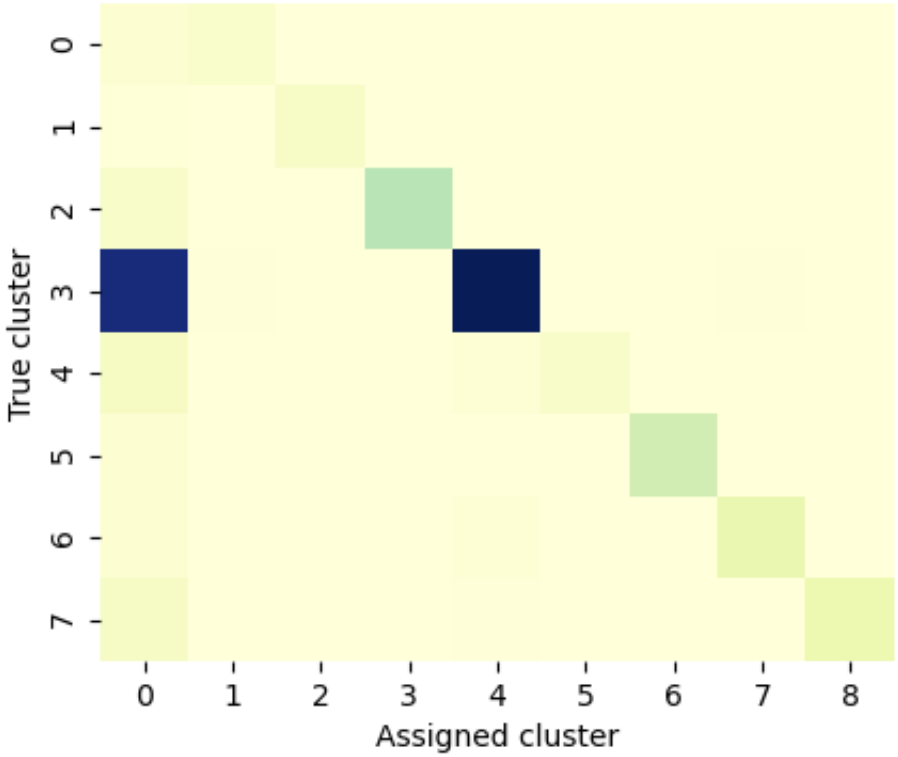

Figure 5. Contingency matrix of the manually annotated subset. 0 includes the articles that were not put in any story by the system but, instead, they were by the annotator.

Overall the proposed system works quite well, with very few errors. However, major classification issues concern the story labelled as Cluster 4. This cluster corresponds to the story about the referendum for the independence of Catalonia. This was the largest story of 2018 in Spain, as the referendum that took place at the end of 2017 had a lot of political consequences in 2018. This story represents a challenge for our methodology, as it spanned over the entire year, with articles appearing almost every day. This results in a very large story with multiple climaxes. As the cluster gets bigger (by iteratively adding new articles from succeeding weeks), it becomes more difficult for the clusters emerging from new climaxes to share 33% of the articles with the existing bigger cluster (see end of Section 3.B). As a result, this story is represented with multiple clusters instead of just one. We considered applying a more fine-grained clustering to distinguish more specific sub-stories from the more general story *Catalan independence referendum*. However, this harms the identification of smaller stories about other topics, making them disappear. A single story being the main story throughout the year is not a common situation. But it can happen in cases of big events like Brexit in the UK in 2018 or Covid-19 in 2020. To obviate these problems and also to be able to better generalize our observations, we decided to exclude this large story from our analysis. 41

When excluding this cluster, we observe an improvement of the NMI (from 0.625 to 0.754). Finally, in Figure 6, we observe that all the differences between the gold data and the output of our semi-automated system are articles that the system ignored (articles in 0 were assigned by the annotator but not by the system). This means that there are no articles assigned to the wrong cluster; the system only missed articles. We therefore conclude that the results are reliable enough to carry on our analysis. 42

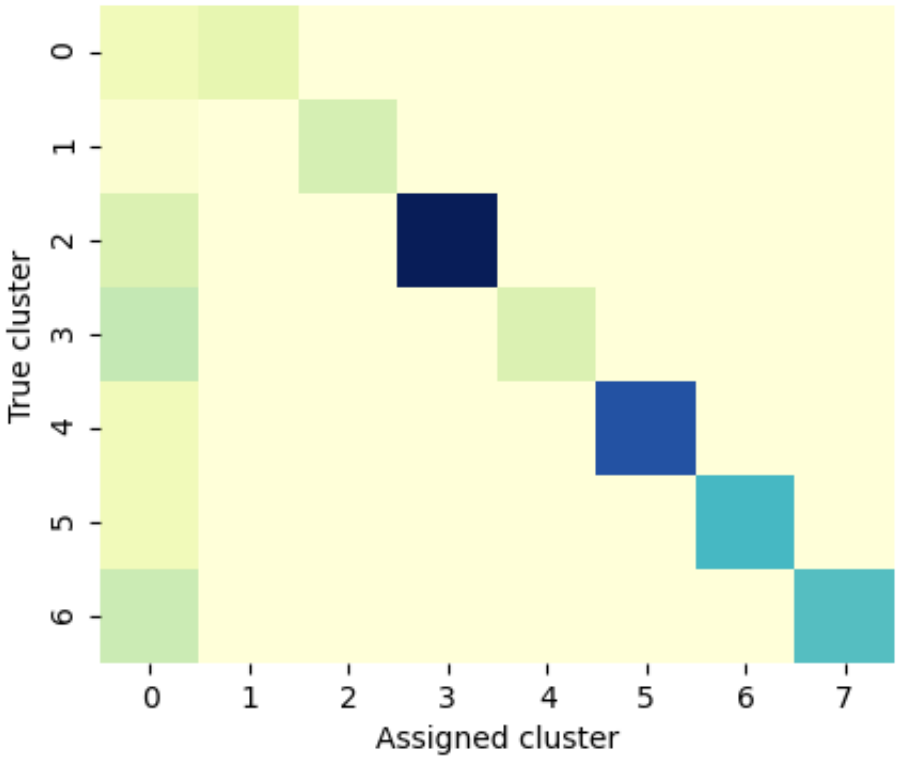

Figure 6. Contingency matrix of the manually annotated subset, when excluding the Catalan independence referendum story. 0 includes the articles that were not put in any story by the system but, instead, they were by the annotator.

QUANTITATIVE RESULTS

Our system, described in Section 3.B, when applied to the corpus described in Section 3.C, results in 80 news stories, from which 40 are labelled as sensational stories, 32 as governmental stories, and 9 as prominent stories. 43

We then calculate the quantitative measures of lifespan, intensity and burstiness for each of these stories. The results have been averaged between all the stories of the same type. Two stories have been removed as they are found to not be representative of the classes.[5] The quantitative measures show that the issue types differentiate only to some extent. 44

	Sensational	Governmental	Prominent
lifespan	11.33	20.76	14.00
intensity	8.07	6.95	5.92
burstiness	6.13	6.38	4.50

Table 1. Averaged quantitative measures for issue type

Stories classified as belonging to sensational issues, i.e. sensational stories, have the highest intensity (8 articles per day on average), the shortest lifespan (11 days on average), and an average burstiness (6.13). These measures show that sensational stories tend to appear out of the blue and have a short lifespan. 45

Governmental stories, i.e. news stories belonging to the governmental issue type, last for long periods (21 days on average) but have a lower intensity (7 articles per day on average). This suggests that governmental stories are planned events which's reporting and development requires a long period of time. 46

Stories belonging to the prominent issue type last for 14 days on average, have a low intensity (6 articles per day) and the lowest burstiness (4.50). 47

We found that only the difference in intensity between types of stories is significant, both between sensational and governmental stories ($0.001849 \leq 0.05$), and between sensational and prominent stories ($0.001998 \leq 0.05$).[6] We can conclude that intensity is the main measurable difference between the different types of stories. It shows that stories that are put on the public agenda by news media themselves, i.e. sensational stories about dramatic events that are triggered by active reporting, get a lot of media attention in a relatively short period of time. Usually, news organizations put a lot of emphasis on such articles, for example by putting them in prominent positions on their website, to attract an audience. News stories that are triggered by political actors or developments in society, such as political events or more structural issues like the environment, are covered with less news articles per day but remain longer on the news agenda. 48

QUALITATIVE ANALYSIS

In this section, we analyse the narrative patterns that emerge from different types of news stories. Newspapers construct stories by turning daily events into narratives [Fulton 2005]. This is achieved by putting in place a series of strategies and linguistic devices. Narratologists have investigated these strategies to identify the presence and recurrence of narrative structures [Propp 1968] [Phelan and Rabinowitz 1994] [Bal 1997]. 49

Narrative structures are what differentiate the fable (the chronological order of the events) from the plot, which underlies the order and manner in which a narrative is presented [Bal 1997] [Kukkonen 2014]. Adopting a post-structuralist perspective, we consider *time* as one of the main features that define the structure of a narrative [Genette 1983] [Abbott 2008]. 50

In all kinds of narratives, the organic temporality of the fable is “subverted by the technical and aesthetic demands of the plot” [Fulton 2005, 240]. When it comes to digital news, the technical settings demand a rapid publication of any relevant novelty about an ongoing story. In this context, the technological and commercial circumstances of production (the narration) “affect the temporal structure of the narrative” [Huisman 2005, 24]. In digital newspapers, the story emerges as a concatenation of single news articles, which compose the daily, if not hourly, development of an event. In this environment, the narrative of a story gets to be expressed by the consecutive publication of articles. 51

Previous work in narratology has attempted to define a general grammar that can describe the construction of narratives. These grammatical systems agree on the general existence of an exposition, predicament, and resolution phase in most narrative texts [Freytag 1900]. News stories are a special kind of narrative that puts its focus on climax events on a routine basis [Tuchman 1973] [Vossen et al. 2015], mostly lacking an exposition or a resolution phase. Besides the climax, the predicament often includes a rising action and a falling action. The rising action starts with an inciting event that brings the attention of the public to the story. The climax is the moment in which the story reaches its peak of tension. In a story about a parliamentary discussion, the climax can be, for instance, the voting day for the approval of a law. A single story can have more than one climax (or sub-climax). After the climax, comes a falling action, which commonly focuses on possible outcomes or consequences of the story. 52

To analyse the narrative structure (i.e. the plot) of the different types of stories, we will use the proxies of lifespan and intensity, but now applied to the different parts of the plot: the climax, the rising action, and the falling action. 53

Sensational stories

Sensational stories are, in general, ephemeral and intense. This can be easily observed in Figure 7, where the smoothed conditional mean of the intensity of each sensational story is illustrated, with the climax in x=0. The rising action of sensational stories is the shortest of the three types. On average, there are 3 days from the first day of reporting to the climax of a sensational story. 54

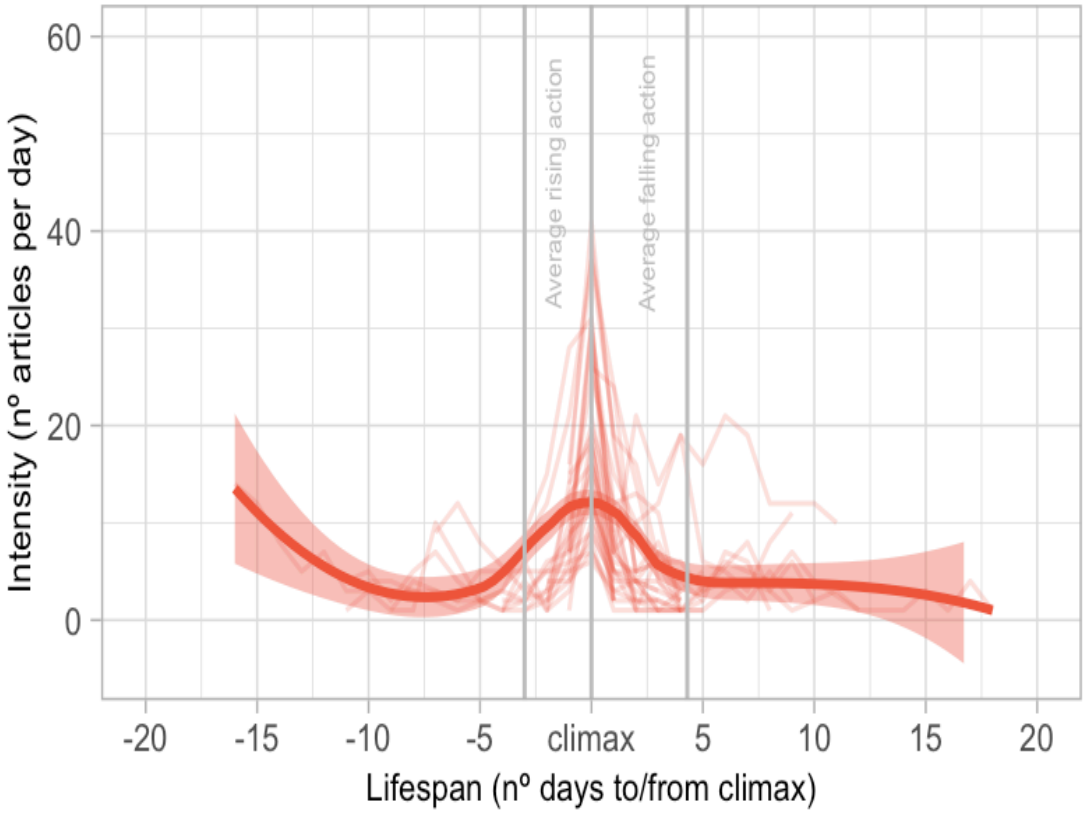

Figure 7. All sensational stories plotted with their climax at x = 0 (thin lines). The smoothed conditional mean is plotted in bold.

The most common narrative pattern for sensational stories is the following: a rising action of 0 to 1 days and a falling action of 3 to 5 days (median). There are 13 out of 40 stories following this structure. 55

In Figure 8, we have labelled the 13 stories by topic and adjusted the size of the shapes by average intensity. The topics that more often recur with this narrative pattern are weather incidents, social injustice, and corruption. These are topics that trigger indignation, but they directly affect very few people. They generally span for one single day. 56

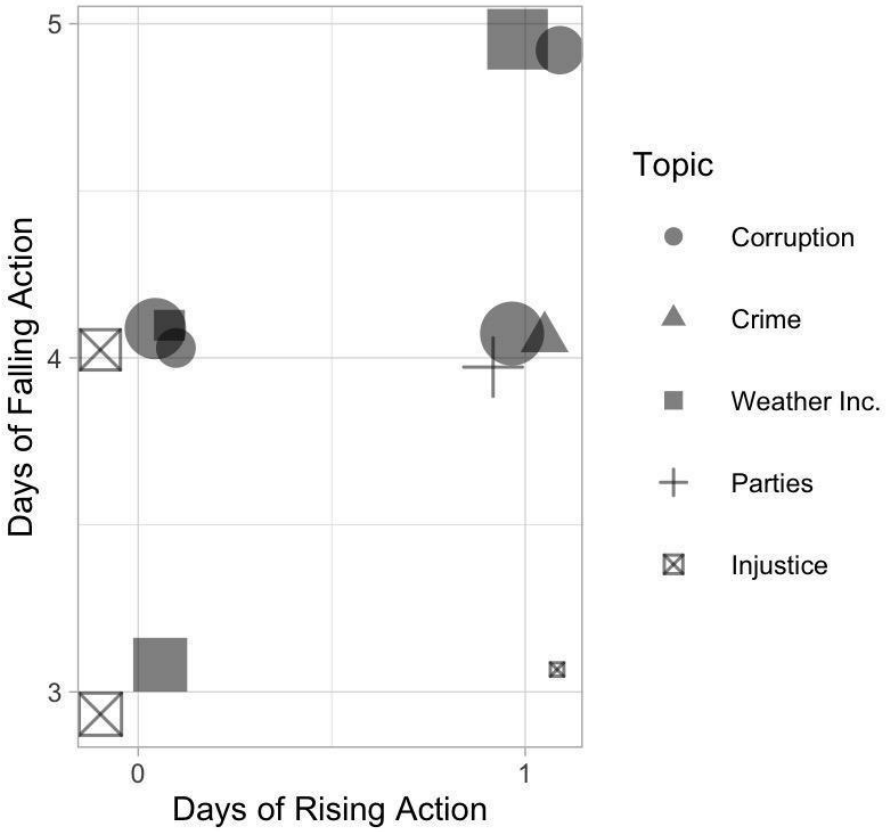

Figure 8. The structure of sensational stories with 0 to 1 days of rising action and 3 to 5 days of falling action. The size of the symbol depends on the average intensity of the story.

We further observe that stories about weather incidents get to its climax the same day that they emerge, having 0 days of rising action.[7] This is not always the case in stories about corruption and social injustice. 57

Three stories are clear instances of social injustice, all concerning court decisions about political freedoms. Two of them are decisions from the European Court of Human Rights (ECHR) that contradict previous decisions of a Spanish court[8], and a third one is the final resolution from a Spanish court, which will later be appealed to ECHR.[9] These three stories follow a similar content structure. They start with an article that explains the court decision, with very similar headlines in all newspapers. Then, some newspapers report about the reactions of the government and the political actors in favour or against the decision. In the case of the two European court decisions (in which the tribunal ruled in favour of political freedom), the reactions are published the same day. In the one coming from a Spanish court (which was accused of ruling against freedom of speech), the reactions are not reported until the day after, making the second day of the story its climax. This shows either less previous attention to the case, or a more unexpected decision compared to the other two court cases, which had already been in the news in the past. The days of the falling action mainly consist of articles about the precedents and the consequences, and opinion pieces. 58

We observe a similar pattern in stories about corruption. On the first day of reporting, multiple articles about the event itself are published. These are followed by reactions from other political actors, and some articles about the context. Finally, in the last days, the articles about the consequences and the opinion pieces are the most common. It can be noticed that two of the stories get to their climax on the second day of reporting. In the first case, the corruption case forces the resignation of the politician involved, making this second event the main climax of the story.[10] This story spans two days, and there are two days of climax. However, in the second story of this kind,[11] reporting starts the day before the court decision. This is an interesting case, as one newspaper published several articles about the background of the corruption case the day before the final sentence. This unusual behaviour seems to indicate a willingness to emphasize this particular story by this media outlet. 59

Governmental stories

Governmental stories are often political discussions or legal debates, which have set timings, often known in advance by journalists. The planned nature of governmental stories gives space for a longer build-up to the story, which results in a longer rising action. In Figure 9, it can be observed that governmental stories tend to have more than one peak (we will also call these sub-climax). For this typology, the curve of the smoothed conditional mean in the main climax is less pronounced and multiple peaks can be observed along the span of the stories (thin lines), mainly in the rising action, but also in the falling action. The rising action of governmental stories takes on average 6.5 days, a difference statistically significant with respect to sensational stories. The average falling action, on the other hand, is similar to sensational stories, lasting around 4 days. 60

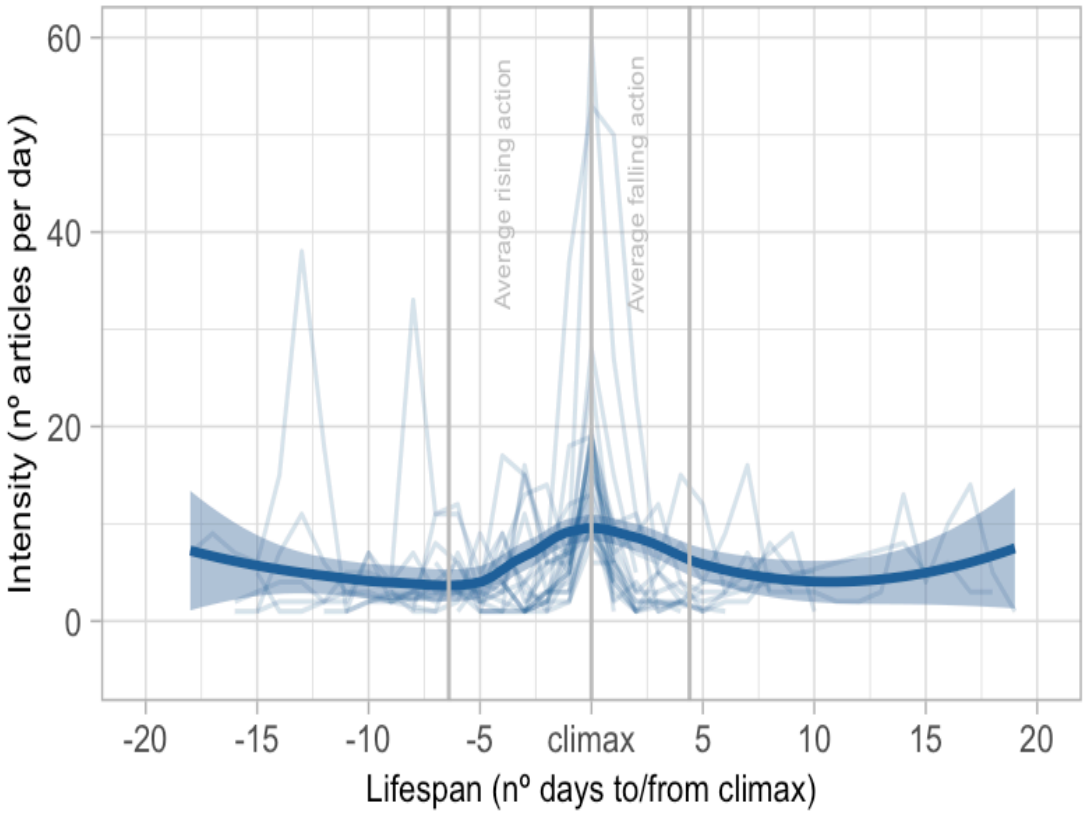

Figure 9. All governmental stories plotted with their climax at x = 0. The smoothed conditional mean is plotted in bold.

At first glance, there are no clear narrative patterns for stories belonging to this issue type. In Figure 10, it can be observed that stories with 3 to 6 days of rising action and 0 to 2 days of falling action are common. These stories have generally one single peak (one climax). However, the big majority of the stories in this typology do not fit in the main pattern represented in Fig. 10. We do observe a common feature between the non-represented stories: most of them have 2 to 4 different peaks (1 climax and 1 to 3 sub-climax). In Figure 10, these stories are identified in dark red. We observe two kinds of narrative patterns in governmental stories, (1) short-spanning stories with one climax, and (2) multiple sub-climax stories. 61

The first kind of stories are generally planned events, with a relatively low media interest, such as the birthday of the King, a vote on the internal statutes of a party, or the visit of the President to Cuba. In the days previous to the climax, these stories start growing by publishing articles about the context, previous similar events and the details on how the central event is going to take place. Also, some opinions of known political actors and some speculation about unknown facts. On the central day of the event, detailed descriptions about how it went are published. In most of the cases, the effects or possible consequences of the central event are published the same day, leaving 0 to 1 days of falling action. If 62

there are opinion pieces, which is not always the case, they are part of the rising action.

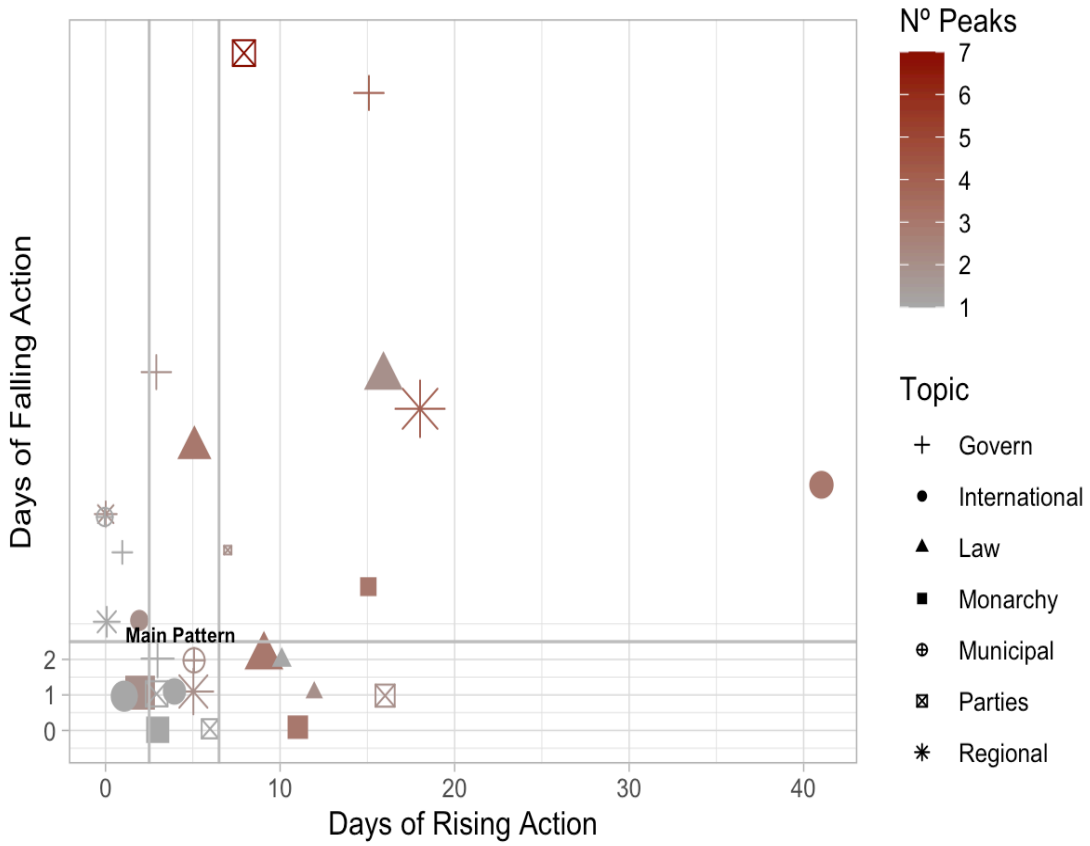

Figure 10. The rising and falling action of all the governmental stories [12], shaped by topic. The size of the symbol indicates the average intensity of the story. The colour shows the number of peaks.

The stories with multiple climaxes are discussions from legislative bodies, ideological struggles between parties, and law approval debates at regional, municipal, or state level. The multiple peaks in this kind of stories are often due to disagreements on the issue by different actors. These stories start like the first kind: they are planned, and the press knows when the involved characters are going to talk about the issue. However, after the first peak, in which a personality has shown its position/decision, new peaks come from other actors disagreeing and asking for an alternative decision. 63

An example of this pattern can be found in the story about the renovation of the General Council of the Judiciary. One first peak emerges when the new members of the council are announced.[13] Though, a larger peak is created when the person who was appointed to be the president of the council announces he will not take the position.[14] 64

Another example of this pattern emerges when the government announces that children in Catalan schools will be able to study in Spanish.[15] The announcement of the government generates the main climax of the story. Some days after, a smaller sub-climax emerges when a court declares a previous measure to facilitate the schooling in Spanish in Catalonia unconstitutional.[16] Finally, a third smaller sub-climax closes the story with the announcement that the Government will take some time to rethink their policies on the regularisation of Spanish in Catalan schools.[17] 65

Prominent stories

Since there are only 9 stories of the type prominent, it is risky to extract many conclusions from them. Figure 11 illustrates the smoothed conditional mean of the intensity of stories belonging to this type, excluding one case: the story of pensions, which is further illustrated in Fig. 12. 66

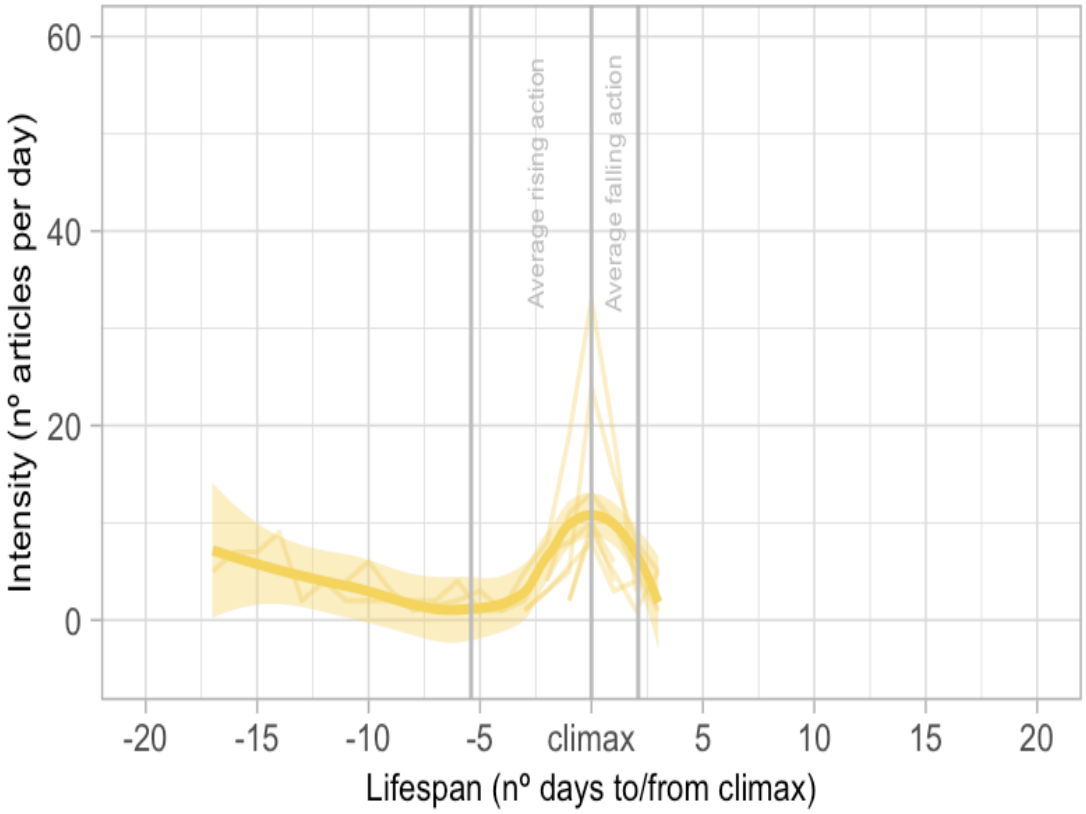

Figure 11. All prominent stories plotted with their climax at x = 0 (thin lines). The smoothed conditional mean is plotted in bold.

It is safe to observe that 6 out of 9 of the prominent stories have a falling action of 1 to 3 days. The rising action, on the other hand, is more varied. These stories are mostly social demands expressed as demonstrations and protests, and the development of policies related to these demands. These observations suggest that media attention to prominent issues is infrequent and fades short after its climax. 67

We have excluded the story about pensions from Fig. 11 because the plot of this story is very different from the other prominent stories. This story is unique in the whole corpus: it has the longest rising action (38 days) and the longest falling action (24 days). It has 8 different peaks, and the highest one has 34 articles published on a single day. The pensions system is a recurrent topic in Spanish politics, as it is public and universal, and the biggest expense of the national budget. 68

In 2018, the yearly uprising of pensions was 0.25%, while prices had risen around 3%. This caused indignation, and the first demonstrations took place in the middle of January. However, *Pensions* does not emerge as a news story until the 8th of February, the day in which the Prime Minister announced some measures to incentivize private retirement funds. From then on, there were demonstrations every week in many cities, with a large one on the 22th of February. After this date, the opposition parties started discussing the topic and the media coverage focused on the politicians, even though the demonstrations kept taking place. On the 14th of March, there was a Parliament session focusing exclusively on the topic of pensions. This day is the highest peak of the story. The Parliament session did not have a clear outcome, and even bigger demonstrations took place after that. By that point, the topic of pensions had become the main point in the political agenda, becoming the trading issue for parties negotiating the national budget (which is a governmental topic). 69

In the story about pensions, we observe that the constant efforts of the social movement brought a prominent issue into the space that is usually taken by governmental issues. This suggests that prominent stories can in fact become very salient, but they require a sustained mobilization. 70

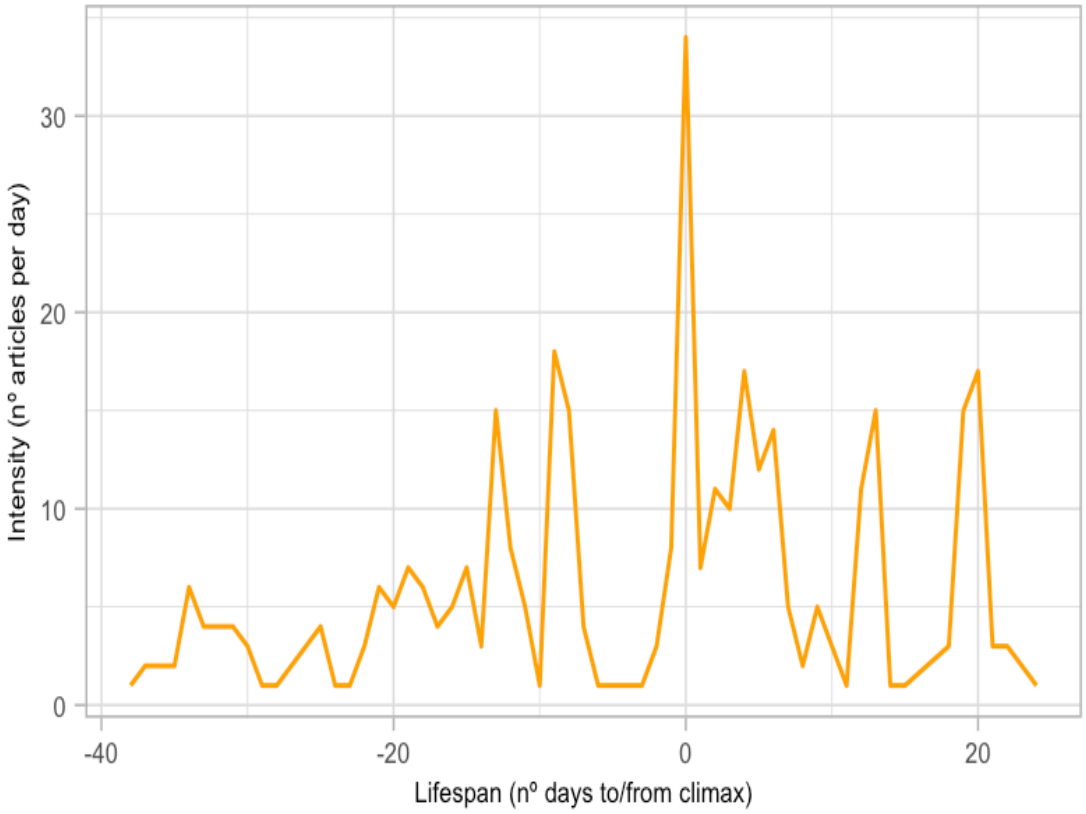

Figure 12. Story of Pensions with climax at x = 0

CONCLUSION

In this work we have defined news stories as *unitary sequences* of events that can be reconstructed by aggregating multiple unitary mentions, i.e. news articles. Our work focused on the development of a semi-automatic method that can be used to facilitate the application of empirical analysis in media studies using large quantities of data. 71

We have shown that clustering methods can be effectively applied to identify news stories, and that news stories are a meaningful unit of analysis. Our approach still requires some manual intervention and background knowledge from the researcher. Rather than a limitation, we consider this a virtuous way to combine quantitative and qualitative analysis to large corpora or collections of datasets. It has proven efficient when applied to an automatically scraped corpus of 50,385 news articles in Spanish, with a purity score of 0.896. A further advantage is represented by its language independent nature: there is no language specific module to pre-process the data, and the story identification is completely unsupervised. Some drawbacks, however, remain unsolved, like the difficulty to deal with very large stories (see Section 4.A). An additional aspect that will be further investigated is the validation of the proposed decision tree for story type classification by means of an extensive inter-annotator agreement study. 72

As a case study, we attempted an empirical validation of Soroka's issue-typology, and especially of the temporal dimension, an aspect that was ignored in the original framework. Time has been quantitatively proxied by three attributes of news stories: lifespan, intensity, and burstiness. We hypothesized for sensational stories to have a short lifespan, a high intensity and a high burstiness and we expected prominent and governmental issues to have a longer lifespan and a lower intensity. The results of the quantitative study using the three temporal attributes showed that, when it comes to time, stories belonging to different issue types differentiate only to some extent. In particular, only *intensity* appears as a significant differentiation attribute, especially for sensational stories. This can be explained by both news values and commercial motivations. Sensational stories such as natural disasters and murders are unexpected, and thus get much attention when they break. Moreover, these stories are able to attract many readers and thus create traffic to news websites. To do justice to their perceived importance and to maximize the effect of such 73

stories, news organizations publish in a relatively short amount of time many articles about them [Harcup and O'Neill 2017].

74 We complemented the quantitative study with a more qualitative analysis inspired by narratology frameworks. In particular, we applied the notion of plot structure to analyse whether there are further distinguishing elements across the types of stories. We thus attempted to measure and visualize the narrative patterns of the types of stories, and in particular, the rising actions, climax, and falling actions. We observed that the days of the rising action are significantly lower in sensational stories compared to governmental stories, while falling actions are similar. A very common trait in governmental stories is having multiple peaks, which correspond to different stages of the political debate. It is interesting to point out that prominent stories have been the least covered stories over the year we analysed. They also had very short falling actions, with the exception of one story.

75 The observed narrative patterns within different issues can be used in the future to identify unusual plots in news reporting, which could indicate an intention of emphasis from a newspaper on a story.

76 The methodology we have illustrated can be used in the future by other scholars in the Humanities interested in conducting computational textual studies of documents that present an evolution of coherent sequences of events in time. In literary studies, our method can help in identifying and aggregating subplots in novels. The code is open and accessible. Required changes are minimal, and in particular they concern the input data and the adaptation of the empirical thresholds for the identification of the elements belonging to a cluster and the merging of the stories over time. At the same time, our methodology can be used to investigate the evolution of issue types in a diachronic perspective, providing an additional level of analysis for digitized collections of historical newspapers.

77 By offering an approach to studying the circulation of news and news stories in the online information ecology, we make a clear contribution to Journalism Studies. In addition, we consider our work useful for other areas in Digital Humanities, especially where the application of computational text analysis methods is involved. Being semi-automatic, our proposed methodology leaves the human in control while the machine is a support to deal with large amounts of data. Once a news story is identified, more fine-grained levels of analysis can be applied. We have applied core concepts from narratology to non-literary texts, opening a new direction in computational narratology. As already stated, having access to plot structures can help investigating variations in writing styles: not only there can be differences across issue types, but it would be interesting to investigate whether there are differences across the components of a plot structure (i.e., rising action, climax, and falling action). In addition to this, having access to stories rather than single unitary mentions can facilitate the investigation of framing and whether this mirrors existing dynamics of power in the society [Haynes 1989] [Entman 1993] [Van Dijk 1995].

78 Further research with this methodology could focus on the linguistic content of these stories, by investigating possible phenomena like language change for different types of news stories, as well as for different moments of the plot. Following previous work on the agenda-setting framework [Boczkowski and Michelstein 2010] [Castillo et al. 2014], the influence of social media interactions on news stories' timing features could be studied, as well as their impact on public opinion. Differences between the plots created by different newspapers about the same story could also be assessed.

Notes

[1] The code of the system can be found in https://github.com/BlancaCalvo/Stories-from-News-Streams .

[2] We use the Sci-kit learn implementation of *k*-means.

[3] 83 out of 20,138 articles are assigned twice (0.4% of the articles).

[4] The code to scrape a similar dataset can be found in https://github.com/BlancaCalvo/Spanish-Newspapers-Scraper .

[5] The first story to be removed is the one about the referendum of independence of Catalonia. As mentioned before, this story spans the whole year and cannot be averaged with the rest. The second deleted story is the parliamentary election in the region of Andalucía. Stories about elections are governmental, as they are led by political actors and do not offer dramatic components. Nonetheless, they are highly

obtrusive, as a huge part of the population is required to take part in them. Because of the ambivalent nature of this story, we exclude it.

[6] The p-value is calculated using the Mann-Whitney U Test.

[7] The incident with 1 rising action day happened over the night.

[8] Estrasburgo dice que quemar fotos del Rey es libertad de expresión and Estrasburgo sentencia que el tribunal que condenó por terrorismo a Otegi no fue imparcial (https://elpais.com/politica/2018/03/13/actualidad/1520933026_224065.html ; https://www.elmundo.es/espana/2018/11/06/5be15f9f468aeb99178b4628.html).

[9] El Supremo ratifica la prisión para el rapero Valtonyc (https://www.elperiodico.com/es/politica/20180220/el-supremo-ratifica-la-prision-para-el-rapero-valtonyc-6637403).

[10] Montón dimite por las irregularidades de su máster pese al apoyo de Sánchez (https://elpais.com/politica/2018/09/11/actualidad/1536664736_500550.html).

[11] Convergència, condenada en el 'caso Palau' por cobrar comisiones ilegales, (http://elpais.com/ccaa/2018/01/15/catalunya/1516001673_733428.html).

[12] Three outlier stories have been removed for a clearer representation.

[13] Manuel Marchena presidirá un Poder Judicial con mayoría progresista (https://elpais.com/politica/2018/11/12/actualidad/1542008392_972068.html).

[14] El juez Marchena renuncia a presidir el Supremo y el Poder Judicial (https://www.elperiodico.com/es/politica/20181120/marchena-supremo-poder-judicial-7156914).

[15] El Gobierno garantiza que se podrá estudiar el 50% en castellano (https://elpais.com/politica/2018/02/16/actualidad/1518780919_449844.html).

[16] El TC tumba los 6.000 euros de Wert para estudiar en castellano en Cataluña (https://www.elperiodico.com/es/politica/20180220/constitucional-declara-inconstitucionales-ayudas-6000-euros-wert-lomce-estudiar-castellano-catalunya-6637168).

[17] Gobierno se tomará 'tiempo' pare regular el castellano en Cataluña (https://www.abc.es/espana/abci-gobierno-tomara-tiempo-para-regular-castellano-cataluna-201802231845_video.html).

Works Cited

Abbott 2008 Abbott, H. P. (2008). *The Cambridge Introduction to Narrative*. Cambridge University Press.

Bal 1997 Bal, M. (1997). *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press.

Ball-Rokeach and DeFleur 1976 Ball-Rokeach, S.J. and DeFleur, M.L. (1976). "A Dependency Model of Mass-Media Effects." *Communication Research*, 3(1), pp. 3–21.

Baraniak 2019 Baraniak, K. and Sydow, M. (2019). "Towards Entity Timeline Analysis in Polish Political News." In *Intelligent Methods and Big Data in Industrial Applications*. Studies in Big Data. Springer International Publishing.

Beheshti et al. 2017 Beheshti, S., Benatallah, B., Venugopal, S. et al. (2017). "A systematic review and comparative analysis of cross-document coreference resolution methods and tools." *Computing* 99, 313–349.

Boczkowski and Michelstein 2010 Boczkowski PJ, Mitchelstein E. (2010) "Is There a Gap between the News Choices of Journalists and Consumers? A Relational and Dynamic Approach." *The International Journal of Press/Politics*. 15(4):420-440.

Bright 2014 Bright, J. and Nicholls, T. (2014). "The Life and Death of Political News: Measuring the Impact of the Audience Agenda Using Online Data." *Social Science Computer Review* 32 (2): 170–81.

Broersma 2019 Broersma, M. (2019). "Epilogue: Situating Journalism in the Digital: A Plea for Studying News Flows, Users, and Materiality." In Scott Eldridge and Bob Franklin (Eds.), *The Routledge Handbook of Developments in Digital Journalism Studies*, 515–526. Abingdon: Routledge.

Bødker and Brügger 2018 Bødker, H. and Brügger, N. (2018). “The Shifting Temporalities of Online News: The Guardian’s Website from 1996 to 2015.” *Journalism* 19 (1): 56–74.

Castillo et al. 2014 Castillo, C., El-Haddad, M., Pfeffer, J. and Stempeck, M. (2014). “Characterizing the Life Cycle of Online News Stories Using Social Media Reactions.” In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* - CSCW ’14, 211–23. Baltimore, Maryland, USA: ACM Press.

Downs 1972 Downs, A. (1972). “Up and down with Ecology — the ‘Issue-Attention Cycle.’” *Public Interest*, 28 (1972: Summer) p.38

Entman 1993 Entman, R. M. (1993). “Framing: Towards clarification of a fractured paradigm.” *Journal of Communication* 43 (4), 51-58.

Freytag 1900 Freytag, G. (1900). “Freytag’s Technique of the Drama: An Exposition of Dramatic Composition and Art.” *An Authorized Translation from the 6th German Ed*. by Elias J. MacEwan. (version 3rd ed. —). 3rd ed. — . Chicago: Scott, Foresman.

Fulton 2005 Fulton, H. (2005). *Narrative and Media.* 1 online resource (xi, 329 pages) : illustrations vols. Cambridge: Cambridge University Press.

Genette 1983 Genette, G. (1983). *Narrative Discourse: An Essay in Method.* Cornell University Press.

Gey et al. 2009 Gey, F., Karlgren, J. and Kando, N. (2009). “Information Access in a Multilingual World: Transitioning from Research to Real-World Applications.” ACM SIGIR Forum 43 (2): 24.

Guo et al. 2013 Guo, X., Xiang, Y., Chen, Q., Huang, Z. and Hao, Y. (2013). “LDA-Based Online Topic Detection Using Tensor Factorization.” *Journal of Information Science* 39 (4): 459–69.

Gómez-Rodríguez et al. 2014 Gómez-Rodríguez, M., Gummadi, K.P. and Scholkopf, B. (2014). *Quantifying Information Overload in Social Media and Its Impact on Social Contagions.* Cornell University Press.

Harcup and O'Neill 2017 Harcup, T. and O’Neill, D. (2017). “What is News? News values revisited (again).” *Journalism Studies* 18(12): 1470-1488.

Haynes 1989 Haynes, J. (1989). *Introducing stylistics*. Allen & Unwin Australia.

He et al. 2007 He, Q., Chang, K. and Lim. E. (2007). “Using Burstiness to Improve Clustering of Topics in News Streams.” In *Seventh IEEE International Conference on Data Mining* (ICDM 2007), 493–98.

Holton and Chyi 2012 Holton, A.E. and Chyi, H.I. (2012). “News and the Overloaded Consumer: Factors Influencing Information Overload Among News Consumers.” *Cyberpsychology, Behavior, and Social Networking* 15 (11): 619–24.

Hong et al. 2016 Hong, L., Yang, W., Resnik, P. and Frias-Martinez, V. (2016). “Uncovering Topic Dynamics of Social Media and News: The Case of Ferguson.” In *Social Informatics, edited by Emma Spiro and Yong-Yeol Ahn, 240–56. Lecture Notes in Computer Science*. Cham: Springer International Publishing.

Huang 2008 Huang, Anna. (2008). “Similarity measures for text document clustering.” *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand.* Vol. 4.

Huisman 2005 Huisman, R. (2005). “Narrative and Media.” In *Narrative and Media*. Cambridge University Press.

Kafalenos 2006 Kafalenos, E. (2006). *Narrative Causalities*. Ohio State UP.

Karlsson and Strömbäck 2010 Karlsson, M. and Strömbäck, J. (2010). “FREEZING THE FLOW OF ONLINE NEWS: Exploring Approaches to the Study of the Liquidity of Online News.” *Journalism Studies* 11 (1): 2–19.

Khy et al. 2008 Khy, S., Ishikawa, Y. and Kitagawa, H. (2008). “A Novelty-Based Clustering Method for On-Line Documents.” *World Wide Web* 11 (1): 1–37.

Kukkonen 2014 Kukkonen, K. (2014). “Plot.” In: Hühn, Peter et al. (eds.): *the living handbook of narratology.* Hamburg: Hamburg University.

Kumaran and Allan 2004 Kumaran, G. and Allan, J. (2004). “Text Classification and Named Entities for New Event Detection.” In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval* - SIGIR ’04, 297. Sheffield, United Kingdom: ACM Press.

Laban and Hearst 2017 Laban, P. and Hearst, M. (2017). “NewsLens: Building and Visualizing Long-Ranging News Stories.” In *Proceedings of the Events and Stories in the News Workshop*, 1–9. Vancouver, Canada: Association for

Computational Linguistics.

Laparra et al. 2015 Laparra, E., Aldabe, I. and Rigau, G. (2015). “From TimeLines to StoryLines: A Preliminary Proposal for Evaluating Narratives.” In *Proceedings of the First Workshop on Computing News Storylines*, 50–55. Beijing, China: Association for Computational Linguistics.

Lee et al. 2012 Lee, H., Recasens, M., Chang, A., Surdeanu, M. and Jurafsky, D. (2012). “Joint entity and event coreference resolution across documents.” In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 489-500).

Lee et al. 2014 Lee, A.M., Lewis, S.C. and Powers, M. (2014). “Audience Clicks and News Placement: A Study of Time-Lagged Influence in Online Journalism.” *Communication Research* 41 (4): 505–30.

Lee et al. 2017 Lee, S.K., Lindsey, N.J. and Kim, K.S. (2017). “The Effects of News Consumption via Social Media and News Information Overload on Perceptions of Journalistic Norms and Practices.” *Computers in Human Behavior* 75 (October): 254–63.

McCombs and Shaw 1972 McCombs, M.E. and Shaw, D.L. (1972). “The Agenda-Setting Function of Mass Media.” *The Public Opinion Quarterly* 36 (2): 176–87.

Minard et al. 2015 Minard, A. L. M., Speranza, M., Agirre, E., Aldabe, I., van Erp, M., Magnini, B., Rigau, G. and Urizar, R. (2015). “Semeval-2015 task 4: Timeline: Cross-document event ordering.” In *9th International Workshop on Semantic Evaluation* (SemEval 2015) (pp. 778-786).

Phelan and Rabinowitz 1994 Phelan, J. and Rabinowitz, P.J. (1994). *Understanding Narrative.* Columbus: Ohio State University Press.

Preiss et al. 2006 Preiss, R.W., Gayle, B.M., Burrell, N., and Allen, M. (2006). *Mass Media Effects Research: Advances Through Meta-Analysis*. Routledge.

Propp 1968 Propp, V. (1968). *Morphology of the Folktale*. University of Texas Press, Austin.

Rao et al. 2016 Rao, Y., Zhong, X. and Lu, S. (2016). “Research on News Topic-Driven Market Fluctuation and Predication.” In 2016 *International Conference on Identification, Information and Knowledge in the Internet of Things* (IIKI), 559–62.

Rouseeuw 1987 Rousseeuw, P. J. (1987). “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.” *Journal of computational and applied mathematics*, 20, 53-65.

Schaeffer 2019 Schaeffer, J.M. (2019). “Fictional vs. Factual Narration.” In: Hühn, Peter et al. (eds.): *the living handbook of narratology*. Hamburg: Hamburg University.

Scheufele and Tewksbury 2006 Scheufele, D. and Tewksbury, D. (2006). “Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models.” *Journal of Communication* 57 (1): 9–20.

Shahaf and Guestrin 2010 Shahaf, D., and Guestrin, C. (2010). “Connecting the Dots between News Articles.” In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* - KDD ’10, 623. Washington, DC, USA: ACM Press.

Shoemaker and Reese 2013 Shoemaker, P., and Reese, S. (2013). “Mediating the Message in the 21st Century.” *A Media Sociology Perspective*. New York: Routledge.

Soroka 2002 Soroka, S.N. (2002). *Agenda-Setting Dynamics in Canada*. UBC Press.

Trilling and van Hoof 2020 Trilling, D. and van Hoof, M. (2020): “Between Article and Topic: News Events as Level of Analysis and Their Computational Identification, Digital Journalism.”

Tuchman 1973 Tuchman, G. (1973). “Making News by Doing Work: Routinizing the Unexpected.” *American Journal of Sociology* 79 (1): 110–31.

Van Dijk 1995 Van Dijk, T. A. (1995). “Discourse analysis as ideology analysis.” In Schäffner, Christina and Wenden, Anita L. (Eds.), *Language and Peace*, 41-58. Amsterdam: Harwood Academic Publishers.

Vossen et al. 2015 Vossen, P., Caselli, T., and Kontzopoulou, Y. (2015). “Storylines for Structuring Massive Streams of News.” In *Proceedings of the First Workshop on Computing News Storylines*, 40–49. Beijing, China: Association for Computational Linguistics.

Wanta Ghanem 2006 Wanta, W. and Ghanem, S. (2006). “Effects of Agenda Setting.” In *Mass Media Effects Research:*

Advances Through Meta-Analysis. Routledge.

Widholm 2016 Widholm, A. (2016). “Tracing Online News in Motion: Time and Duration in the Study of Liquid Journalism”. *Digital Journalism*: Vol 4, No 1. 2016.

Yang et al. 1998 Yang, Y., Pierce, T. and Carbonell, J. (1998). “A Study on Retrospective and On-Line Event Detection.” In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 28-36).

Zucker 2017 Zucker, H.G. (2017). “The Variable Nature of News Media Influence.” *Annals of the International Communication Association*, November.