Discovering and Visualising Stories in News

Marieke van Erp, Gleb Satyukov, Piek Vossen and Marit Nijssen

The Network Institute
VU University Amsterdam
marieke.van.erp,gleb.satyukov,piek.vossen@vu.nl,m.w.nijsen@student.vu.nl

Abstract

Daily news streams often revolve around topics that span over a longer period of time such as the global financial crisis or the healthcare debate in the US. The length and depth of these stories can be such that they become difficult to track for information specialists who need to reconstruct exactly what happened for policy makers and companies. We present a framework to model stories from news: we describe the characteristics that make up interesting stories, how these translate to filters on our data and we present a first use case in which we detail the steps to visualising story lines extracted from news articles about the global automotive industry.

Keywords: narratives, storylines, visualisation

1. Introduction and Motivation

News often revolves around topics that span over a longer time period e.g. the global financial crisis[1], the affordable health care act in the US[2] or the Intercity Direct train in the Netherlands[3]. In particular for policy and decision makers, it is important to be able to reconstruct the story around these topics. Currently, information specialists at governments and companies alike spend a fair part of their time querying search engines, sorting out duplicate articles, finding complementing information, etc. to put together these stories.

Within the NewsReader project[4], we aim to aid these users by automatically generating stories around news topics. To this end, our systems ingest large streams of daily news, extract events, their participants, locations and times and store these in a knowledge base. The number of facts in this knowledge base, will quickly grow to a size that while structured, is too large to search effectively by humans. We are therefore modelling and visualising storylines from these facts, to aid users in their search. Storylines not only summarise the essential developments and changes in a compact way compared to document or trend representations of news, they also have more explanatory power with respect to these changes. Storylines provide insights in the motivation and causes of changes, and can show the possible strategies and intentions.

In this paper, we present our framework to model stories from news: we describe the characteristics that make up interesting stories, how these translate to filters on our data and we present our first visualisation of such a story.

The work on defining ‘stories’, ‘plots’ and ‘narratives’ goes back to Aristotle and each of these concepts have slightly different meanings in different research fields. For our purpose, we take the following definitions:

story: sequence of events that together form a reconstruction of what took place around a certain topic

plot: the causal and logical structure which connects events

narrative: account of a sequence of events

The main distinction we make between story and narrative is that we aim to steer away from any stylistic markup of the account, that is, we aim to present ‘dry’ statements such as who did what, when and where. Naturally, any choices in word use to refer to events or focus of an utterance influence the narrative and may indicate the speaker’s opinion on the event. For now we assume that by aggregating the accounts of many sources on the same event, differences in perspectives are smoothed out.

In this paper, we present the following contributions:

- a theoretical framework for stories (Section 2.)
- a translation of the theoretical framework into a visualisation (Section 3.)
- a concrete use case around the global automotive industry (Section 4.)

2. What makes a story

There is a fair body of work in literary science devoted to explaining and modelling story, plot and narrative structure (cf. (Bremond, 1980; Fludernik, 1996; Bortolussi and Dixon, 2003)). It is generally accepted that stories have a beginning, a process and an outcome. What makes a story a story rather than a description is that the begin state can ameliorate or deteriorate through various stages and actions (Bremond, 1980). However, there is no consensus yet as to what exactly the characteristics are that determine which elements are relevant.

2.1. Event Structure

According to Bortolussi and Dixon (2003), three main structures to understand the relationship between story elements are defined:

[1] http://www.worldbank.org/financialcrisis/ Retrieved 10 October 2013

[2] http://www.theatlanticwire.com/topics/affordable-care-act/ Retrieved 10 October 2013

[3] http://www.railjournal.com/index.php/tag/Fyra.html?channel=00 Retrieved 11 October 2013, formerly known as “Fyra”.

[4] http://www.newsreader-project.eu

Stereotypic Experience: commonly encountered events in the world can be captured in generalised structures, or 'scripts', e.g. in the case of a company takeover, one company is usually bigger or financially stronger than the other, companies negotiate, the takeover is finalised and a press release takes place (Shank and Abelson, 1977)

Character Plans: actors have goals, and by understanding actions in relation to such goals one can interpret the narrative plot, e.g. in case of a company takeover, the companies involved have the goal to become financially stronger, discussing takeover options is a plan for achieving the goal and finally a complete or partial takeover can ensue (Shank and Abelson, 1977)

Causal Chains: events are classified as a cause or enabling event of a second event if they are a prerequisite for the second event to happen e.g. before a takeover takes place, negotiations need to take place (Mackie, 1980)

In order to model news stories, we choose the causal chain as the most viable theory to work with. The reason for this is that it is currently not possible to model all possible scripts that can make up a story, which would be required for the stereotypical experience model. Alternatively, we could define certain generic scenarios for particular domains at a higher level, but we still want to stay flexible in the stories that can be generated from the data in order to be able to find the 'needles in the haystack'. The variation in news topics and their accompanying stories is so great that modelling fine-grained scripts for this is prohibitive.

The difficulty with character plans for the NewsReader domain is that the character plans theory was developed for the literary domain, in which the character and her development are central notions. While one can interpret participants described in the news as characters and the changes resulting from the event influences as their development, intentions and plans are more difficult to detect and visualize. For now, we rather expect that users will be able to guess such intentions and plans from the visualisation of the changes as such.

Bortolussi and Dixon (2003) discuss not being able to count on the information on chains being present in the text of the narrative as an essential difficulty. For our purpose, this is less of a difficulty, as the primary target of news organisations is to provide this information. Naturally, there are gaps in what a single news source can and does publish, but the fact that we mine multiple sources over time should leverage this problem. No news item tells the full story but over time many different sources may.

2.2. Event selection

One issue that is not resolved by literary story theories is the issue of magnitude in the news domain. For most literary science research regarding stories the working unit is the narrative as presented in a novel, movie or play. Even though there is great variation in the length or duration of these narratives, these sizes cannot be compared to the thousands of events that are related to 'bigger' news topics such as the global financial crisis. In order to select the relevant story elements there, one cannot simply present them all on a time line or some graphical representation; the relevant events need to be filtered first. An additional problem is that many different stories are told in the news and we need to know how to separate one story of another. The situation in the news can be compared with a whole library of novels many of which appear to be reworks of the same story.

The latter problem is solved by clustering news published on the same day and applying cross-document event coreference to all news in the same cluster to find out which items report on the same events and stories. We will not discuss this problem in this paper and assume that such a grouping of coreferential event-coreference has been done. For a more detailed explanation of this, the reader is referred to (Cybulska and Vossen, 2014; Vossen et al., 2014)).

For the former problem, we are experimenting with various manners of filtering events by relevance for a particular story. In the news domain, topics are identified both by content and by their 'trendiness'. The content of topics is largely dependent on the use case and the information seeker's particular interests. 'Trendiness' of a topic is influenced by how popular the topic is, for which frequency of mentions and spread over different sources is an approximation.

For further defining relevance we use the notion of a climax or dramatic turn in a story. We expect that every story has a point at which a dramatic turn takes place. Typically, this is something that is reported many times over in the news or a topic for which there is a strong sentiment, i.e. a trending topic. An example of such a climax could be Google suing Facebook. According to classical storyline theories, such a climax is the result of a process that builds up to it. This means there are other events that precede the climax but have some causal relation to it. These events may not have been trending (i.e. frequently reported or loaded with strong sentiments) in the news but now become relevant because they build up to the climax. Similarly, all events happening after the event are relevant because they follow from it, even if not trending. Note that the climax can also be an event that happens despite the build-up of previous events. An exampe is Volkswagen taking over Porsche while everybody expected the opposite.[5]

We thus can come to a definition of criteria for relevance of events:

- trending events: occur frequently and have strong sentiments.
- events preceding or following trending events (not necessarily trending themselves) that involve the same participants or have explicit causal relations.
- events of trendy persons: persons that occur frequently and about which people have strong sentiments (anything Obama does is relevant).

[5] `http://www.theguardian.com/business/2012/jul/05/volkswagen-buys-porsche` Retrieved March 22 2014

- events of non-trending people that participed in trending events before (victims of earthquakes become trending people whose future events become more trending).
- the number of participants effected by an event.
- the culturally dependent impact of an event on their participant.

Most of these criteria can directly be represented by numerical values. Any visualisation should take the above criteria into account for interfacing and selecting of the complete data generated. For now, we focus first on trendiness. In future work, we would also like to be able to detect less trending events related to trendy events, but this is out of the scope of this paper.

Apart from the intrinsic criteria to filter or select certain events instead of others, there are also user-driven selections. For this, we take inspiration from (Van den Akker et al., 2011) in which narrative chains are built by the users. In this approach, users can choose to filter by actor, location or topic. This allows one to, for example, 'follow' an actor (e.g. a person or an organisation) through time. A storyline focusing on a location or topic could be the earthquake and aftermath of the Indian Ocean earthquake that hit on 26 December 2004 or the Fukushima Daiichi nuclear disaster. Here the distinction focuses on whether the narrative only contains events happening at a particular location or whether related events at other locations are also taken into account (such as the international debates about tsunami alerts or nuclear power). However by presenting only one actor, location or topic, the story line often does not give a complete reconstruction of the relevant story. Therefore clusters of relevant actors, locations or (sub)topics need to be taken into account, where relevance is determined by whether actors or locations interacted or are mentioned in conjunction with the main actor or location (potentially filtered by frequency). Determining relevant related topics depends on whether they are mentioned together, but also through general scripts such as those discussed in Subsection 2.1.

2.3. Event extraction

We have a linguistic processing pipeline that takes in news articles on which it performs tokenisation, part-of-speech tagging, chunking, shallow parsing, named entity recognition (including locations and time), event-detection, and semantic role labelling for English (Agerri et al., 2014). We aggregate events across documents using a lemma-based clustering technique (Cybulska and Vossen, 2014). Aggregating mentions to instances allows us to combine information about the same event from different sources, for example if not all information is present in one source, or to update the event information as more information becomes available.[6]

3. Story line visualisation

Visualisations have been used along with data to aid the user in understanding for over 150 years. Often mentioned as the first infographic is Minard's graphic of Napoleon's march to Russia depicting the decreasing size of his army, drawn in 1861[7]. Simple graphs and diagrams are examples of such tools used within larger text to augment the story (Tufte, 1983). The original text was still leading and the visualisations were there only to convey additional detailed information. With the advent of big data and the improvement in computer graphics, visualisations are now used to aggregate large amounts of data in order to make sense of it.

However, many visualisations (also called infographics) that aim to tell a story are still largely the result of manual labour (McCandless, 2009). An analysis of different elements present in narrative visualisations is presented in (Segel and Heer, 2010). They identify 7 different genres of visualisations i.e. magazine style, annotated chart, partitioned poster, flow chart, comic strip, slide show and film/video/animation. For our domain and task at hand (i.e. aiding information specialists in reconstructing a story) an *annotated chart* seems the most suitable data representation as the core of our visualisation. Furthermore, as we are attempting to visualise big data it is impossible to present all information in one view, therefore it is important that the user can interact with the visualisation in order to filter and drill down on a particular selection. Therefore the visualisation should be *reader-driven* according to Segel and Heer (2010)'s classification of design dimensions.

The most common way to visualise stories revolves around some sort of time line (Yau, 2011). On a time line, events can be ordered chronologically, allowing the user to follow the course of the narrative. However, in our use case, we would like to visualise multiple storylines, that can also express interactions between different actors in our domain. There is a large body of work on visualising interactions between participants through so-called graph visualisations in which participants are represented by nodes and their interactions by edges, such as provided by Gephi (Bastian et al., 2009). However, here the temporal dimension is not immediately apparent, which is an important element for narratives.

Since we adopted the Causal Chains approach for modeling stories our visualisation should focus on the temporal and causal sequencing of events, where causal relations can be boiled down to ineractions between actors. To represent both the temporal dimension and interaction between actors, we base our visualisation on ideas from (Shahaf et al., 2012) who use a metro map metaphor to represent story lines through time that intersect. Each line in the visualisation represents a storyline around a particular participant, for example in Figure 1, the red line could represent the storyline of Volkswagen and the blue line the storyline of Ford. As can be seen, the storylines are mostly independent, but in some cases they intersect, which could represent an event in which Volkswagen interacts with Ford, such as for example in 2008 when Volkswagen surpassed Ford as the third

[6] Cybulska and Vossen (2010) found that as sources report on events that happened further back in time (more distant in time), a more generalised 'zoomed out' picture is presented whereas reports of events after they just happened tend to provide accounts on smaller details.

[7] http://www.edwardtufte.com/tufte/minard

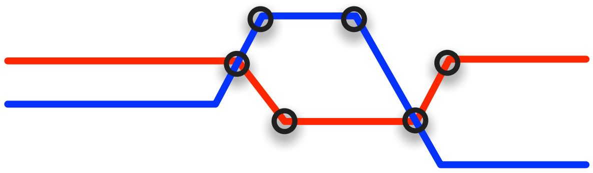


Figure 1: Metro line example of storylines around two actors which intersect at different points in time.

largest automaker.[8]

However, presenting a user with only intersecting storylines does not give him/her the context of the stories. We therefore also include two additional panels in our visualisation that provide information about the participants involved as taken from the participant's DBpedia page and a zoom function on the selected storyline which displays the events of only that storyline in a separate pane.

We have chosen to use paper.js[9] as the implementation to draw our metro-line visualisation, in which separate time lines of various actors intersect on collective events. For the time line visualization a separate module was used built by Almende B.V.[10]. Some integration was needed to make the distinct libraries work together and provide seamless interactivity for the user. For the detail view we use DBpedia services to pull up detailed information regarding the selected actor and present it to the user via a Wikipedia "infobox"-inspired format. The visualisation is served as an application built using Angular MV on top of a Node.js backend. The most challenging parts of the application currently lies in the volume of data. At the moment, we are importing static data selections from the greater NewsReader knowledge base, but we are investigating options to query the knowledge base realtime to generate the visualisations. As far as integration of the various panels is concerned, we have kept everything modular, making it possible to "plug-and-play" additional functionality as is required. Although the three distinct panels seem separated, the key feature resides in the interaction aspects of the visualisation, which allow the user to play around to investigate and review the data in a different perspective.

4. The Global Automotive Industry

In this section, we present a reconstruction of the stories unfolding in the global automotive industry between 2003 and 2013. This is an interesting period as the interplay between the more affordable and exclusive car makers changes over the course of the global financial crisis.

For this use case, we have processed 64,540 news articles about cars using the pipeline mentioned in Subsection 2.3. and loaded the obtained instances into an RDF triplestore. Our base dataset of extracted information about the global automotive industry is then enriched with background knowledge from various sources such as DBpedia.[11]

This setup allows us to easily query the extracted data and visualise the results. To scope our example in this section, we focus on the 10 most mentioned car makers in our corpus (see Table 1).

From the second columns in this table, it can be seen that the number of events is enormous. In order to visualise all of these, one would have to zoom in to a very small time period to be able to see any of the details. Also, the number of events the Ford Motor Company is involved in is 12 times as large as the number of events the Nissan Motor Company is involved in. We therefore further filter by looking at the key persons involved per company (through the *dbpedia-owl:keyPerson* of relation). This brings the number of events down to the numbers shown in the third column of Table 1.

The metro lines in the main pane of the visualisation shown in 2 intersect at events in which different companies are involved simultaneously. The event bubbles show the event id of the event in the knowledge base and the event label to give an indication of the type of event that took place. Further event information can be brought up by clicking on the bubble such as when the event took place and where. This information can also be brough up via the event boxes on the timeline in the bottom pane. The demo can be found at `http://www.newsreader-project.eu/results/demos/`.

5. Challenges

The challenges that make it difficult to reconstruct story lines lie both in the preprocessing of the data and the visual representation.

First of all, we know that the NLP pipeline is not perfect. The performance of the current NLP modules lies between 50% and 90%, depending on the module. By ingesting large amounts of news articles in which some repetition occurs, we can mitigate the consequences for the recall of the modules. In the visualisations presented, we perform some additional filtering through background knowledge on the named entity recognition classes to remove too outrageous items. When we for example query the knowledge base for the most commonly occurring actors in our domain, we add the restriction that these should be of type `http://schema.org/Organization` or `http://schema.org/Person`. However, we hereby lose instances in which the NLP pipeline classified a country mentioned as an actor, as this would normally be of type `http://schema.org/Place` in our background knowledge. Oftentimes countries are locations, but in common discourse often a country name is mentioned when an act of its government is meant. This is a near-identity and coreference problem that we aim to address in future work. Related to this problem is that the actors in our domain are not always stable or have clear boundaries. Chrysler for

[8] `http://www.leftlanenews.com/vw-overtakes-ford-as-3rd-largest-automaker.html` Retrieved 19 March 2014.

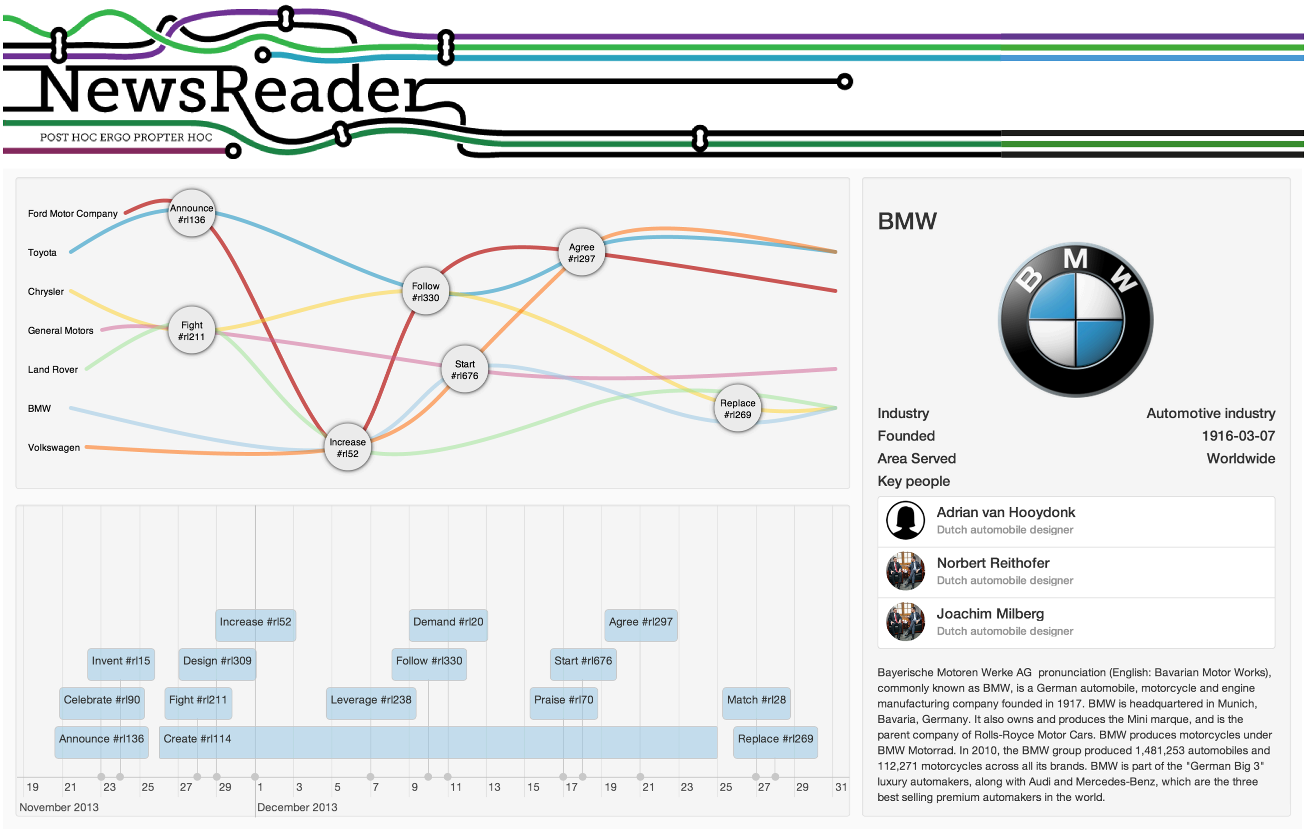
[9] `http://paperjs.org/`

[10] `http://visjs.org/docs/timeline.html`

[11] Due to copyright restrictions we cannot make the full dataset available, however, the instances derived from the news articles can be found at `http://www.newsreader-project.eu/results/data/`.

actor	Number of Events	Events after filtering
`http://dbpedia.org/resource/Ford_Motor_Company`	122,965	2086
`http://dbpedia.org/resource/Toyota`	33,953	151
`http://dbpedia.org/resource/Chrysler`	31,680	457
`http://dbpedia.org/resource/General_Motors`	31,348	19
`http://dbpedia.org/resource/Land_Rover`	27,532	105
`http://dbpedia.org/resource/BMW`	26,855	77
`http://dbpedia.org/resource/Volkswagen`	22,394	511
`http://dbpedia.org/resource/Fiat`	16,656	466
`http://dbpedia.org/resource/Honda`	14,383	0
`http://dbpedia.org/resource/Nissan_Motor_Company`	10,609	336

Table 1: Most commonly occurring actors in global automotive industry use case dataset



BMW

Industry	Automotive industry
Founded	1916-03-07
Area Served	Worldwide

Key people

Adrian van Hooydonk
Dutch automobile designer

Norbert Reithofer
Dutch automobile designer

Joachim Milberg
Dutch automobile designer

Bayerische Motoren Werke AG pronunciation (English: Bavarian Motor Works), commonly known as BMW, is a German automobile, motorcycle and engine manufacturing company founded in 1917. BMW is headquartered in Munich, Bavaria, Germany. It also owns and produces the Mini marque, and is the parent company of Rolls-Royce Motor Cars. BMW produces motorcycles under BMW Motorrad. In 2010, the BMW group produced 1,481,253 automobiles and 112,271 motorcycles across all its brands. BMW is part of the "German Big 3" luxury automakers, along with Audi and Mercedes-Benz, which are the three best selling premium automakers in the world.

Figure 2: Story lines showing the reconstruction of the Volkswagen Vorst story

example manufactures Fiat models in the US. As of 2014, Fiat owns 100% of the Chrysler Group LLC. On a more fine grained level, we see for example that a car plant in Belgium near Brussels, used to be called Volkswagen Vorst, but since 2007 the plant is called Audi Brussels as its focus became manufacturing Audi cars.[12] Again this is a type of near-identity and coreference problem that needs to be addressed. The difference between this challenge and the previous one is that this is a challenge at the semantic level, between instances, whereas the previous challenge is to be resolved at the text level, between mentions. Within the semantic web community, this problem has been described and solutions have been proposed, but they have not yet been implemented widely (Halpin and Hayes, 2010).

The main challenge for the visual representation lies in the choices that we made for filtering the data and its different elements. Currently, we have chosen to filter the number of events in order to make the visualisation more manageable, but we also acknowledge that frequency-based visualisations of datasets can be a very useful manner of providing users insights into the data. In this work, we have chosen to drill down to a smaller selection of data in order to be able to visualise a narrative structure, and ideally, a user is also presented 'the bigger picture'. It is inevitable that multiple visualisations are to be designed to visualise every possible relevant view of a dataset, as there is not a one-size-fits-all solution. In future work, we aim to investigate which story elements are most important to users to find out for example whether we need to focus more on showing frequencies or geographic distributions.

[12] Audi was already part of the Volkswagen Group since 1965.

6. Discussion

We presented a framework for formalising stories based on prior work in literature science, a visual translation of this framework and a use case. Although our visualisation provides more detail than currently used aggregate visualisations such as trend lines, there are of course trade-offs and still some open issues.

The first issue is that actors and events can have a hierarchical structure. Sometimes in our global automotive use case, Volkswagen the company as a whole is a participant in an event, sometimes only a particular part of the company or even a particular person is. We can represent such hierarchies formally, but they do not translate well to visualisations. We now choose to visualise all actors individually, an extra layer or even a separate visualisation is needed to show such relationships. Alternatively, we can let the user choose to group lines according to a hierarchical relation or to differentiate in the reverse direction. Resolving the identity/coreference challenges mentioned in the previous section play a key role here.

The second issue is that more filtering is needed in order to make the visualisation still intelligible when covering a larger topic. For this we aim to include sliders that enable users to zoom in and out to toggle between for example more fine-grained story lines or more or fewer participants. In future work, we will include a view that selects between different perspectives on a story, in the Volkswagen case this could be the story as told by the union or the story told by the company.

Finally, we want to distinguish between actual factual changes and speculated changes in the news. Many new items speculate about jobs that may be lost as a result of changes in the production or speculate about the production itself. Parallel to the actual storyline, we can have prediction points representing the speculated projections.

Acknowledgments

The authors would like to thank Ben Peperkamp and Dick Schram for their helpful advice and discussions on narrative theory. This research was funded by the European Union's 7th Framework Programme via the NewsReader (ICT-316404) project.

7. References

Agerri, Rodrigo, Bermudez, Josu, and Rigau, German. (2014). Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31.

Bastian, Mathieu, Heymann, Sebastien, and Jacomy, Mathieu. (2009). Gephi: An open source software for exploring and manipulating networks. In *e Third International ICWSM Conference (2009)*.

Bortolussi, Marisa and Dixon, Peter. (2003). Events and plot. In *Psychonarratology: Foundations for the Emperical Study of Literary Response*. Cambridge University Press.

Bremond, Claude. (1980). The logic of narrative possibilities. *New Literary History*, 11(3):387–411.

Cybulska, Agata and Vossen, Piek. (2010). Event models for historical perspectives: Determining relations between high and low level events in text, based on the classification of time, location and participants. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valetta, Malta, May 19-21.

Cybulska, Agata and Vossen, Piek. (2014). Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31.

Fludernik, Monika. (1996). *Towards a 'Natural' Narratology*. Routledge.

Halpin, Harry and Hayes, Patrick J. (2010). When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In *Proceedings of LDOW2010*, Raleigh, NC, USA, April 27.

Mackie, J. L. (1980). *The cement of the universe: A study of causation*. Clarendon, Oxford, England, UK.

McCandless, David. (2009). *Information is Beautiful*. HarperCollins.

Segel, Edward and Heer, Jeffrey. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*.

Shahaf, Dafna, Guestrin, Carlos, and Horvitz, Eric. (2012). Trains of thought: Generating information maps. In *Proceedings of WWW 2012*, Lyon, France, April 16-20.

Shank, R. C. and Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Erlbaum, Hillsdale, NJ.

Tufte, Edward. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Conn, USA.

Van den Akker, Chiel, Legêne, Susan, Van Erp, Marieke, Aroyo, Lora, Segers, Roxane, Van der Meij, Lourens, Van Ossenbruggen, Jacco, Schreiber, Guus, Wielinga, Bob, Oomen, Johan, and Jacobs, Geertje. (2011). Digital hermeneutics: Agora and the online understanding of cultural heritag. In *Proceedings of the Third ACM WebSci Conference (WebSci'11)*, Koblenz, Germany, June 14–17.

Vossen, Piek, Rigau, German, Serafini, Luciano, Stouten, Pim, Irving, Francis, and Hage, Willem Robert Van. (2014). Newsreader: recording history from daily news streams. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31.

Yau, Nathan. (2011). *Visualize this: The FlowingData Guide to Design, Visualization and Statistics*. John Wiley & Sons.