FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Visualizing News Stories from Annotated Text

Catarina Justo dos Santos Fernandes

U. PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado em Engenharia Informática e Computação

Supervisor: Sérgio Nunes

Second Supervisor: Ricardo Campos

October 26, 2023

Visualizing News Stories from Annotated Text

Catarina Justo dos Santos Fernandes

Mestrado em Engenharia Informática e Computação

October 26, 2023

Abstract

As the world is producing more information than ever before, the amount of news and journalistic information produced has also increased. A key factor contributing to this is the digital transition that has been happening in the last decades. This increase in the amount of content available and the number of channels where it is possible to come in contact with it has led to some challenges for the average reader. Namely, these challenges include filtering available news content in search of relevant information and interpreting the content of the chosen news piece itself. Such tasks demand a certain degree of time and effort that people might not have, which can make the general public feel overwhelmed, as they are constantly exposed to information they cannot process.

Over the years, several solutions have been proposed to address these problems. In this research work, we build upon the NewsLines visualization tool (developed in the past in our research lab) to explore and expand further avenues within the narrative visualization domain. This tool presents the user with a storyline that allows for the visualization of events, participants, and how they are related in a concise way, along with temporal and spatial information. Contrary to Newslines, which only allowed for visualization of manually fabricated news narratives, we aim to explore the tool by integrating it with real annotated data from the news domain. The annotated news dataset we utilized was collected as part of the Text2Story project, which is also the project under which this research was conducted.

To achieve these goals, we developed a pipeline to convert the essential information in the BRAT annotation files of the project to JSON, a versatile and easy-to-understand file format compatible with NewsLines. Additionally, the NewsLines tool was updated to accommodate the new input schema and better adapt to real data from the news domain.

In order to evaluate the extended tool, we conducted interviews with eight researchers with diverse backgrounds ranging from computer science to linguistics from the Text2Story team. The interviews focused on better understanding the key elements that should be represented in a news narrative visualization, giving feedback on the developed tool, and brainstorming other possible news visualizations. A survey was also sent out to five faculties of the University of Porto in order to gather feedback on the usefulness of the tool from the general public. In the end, we gathered valuable new ideas for future visualizations and improvements to the tool.

Keywords: Visualization of narratives, Text visualization, Human-computer interaction

Resumo

Como o mundo está a produzir mais informação do que nunca, a quantidade de notícias e de informação jornalística produzida também aumentou. Um dos principais factores que contribui para esse crescimento é a transição digital que tem vindo a ocorrer nas últimas décadas. Este aumento da quantidade de conteúdos disponíveis e do número de canais onde é possível entrar em contacto com os mesmos tem levado a alguns desafios para o leitor comum. Nomeadamente, estes desafios incluem a filtragem das notícias disponíveis em busca de informação relevante e a interpretação do conteúdo da própria notícia escolhida. Estas tarefas exigem um certo grau de tempo e esforço que as pessoas podem não ter, o que pode levar a que o público se sinta sobrecarregado, uma vez que está constantemente exposto a informações que não consegue processar.

Ao longo dos anos, foram propostas várias soluções para resolver estes problemas. Neste trabalho de investigação, baseamo-nos na ferramenta de visualização NewsLines (desenvolvida no passado no nosso laboratório de investigação) para explorar e expandir outras vias no domínio da visualização de narrativas. Esta ferramenta apresenta ao utilizador uma *storyline* que permite a visualização de eventos, participantes e das relações entre si de forma concisa, juntamente com informação temporal e espacial. Ao contrário do Newslines, que apenas permitia a visualização de narrativas fabricadas manualmente, o nosso objetivo é explorar a ferramenta integrando-a com dados reais anotados do domínio das notícias. O conjunto de notícias anotadas que utilizámos foi recolhido como parte do projeto Text2Story, que é também o projeto no âmbito do qual esta investigação foi realizada.

Para atingir estes objetivos, desenvolvemos uma *pipeline* para converter a informação essencial dos ficheiros de anotação BRAT do projeto para JSON, um formato versátil, fácil de compreender e compatível com o NewsLines. Além disso, a ferramenta NewsLines também foi atualizada para suportar o novo *schema* de dados de entrada e para se adaptar melhor a dados reais do domínio das notícias.

Para avaliar a ferramenta extendida, realizámos entrevistas com oito investigadores da equipa do Text2Story, com formação diversificada, desde a informática à linguística. As entrevistas centraram-se em compreender melhor quais os elementos-chave que devem ser representados numa visualização de uma narrativa de notícias, recolher feedback sobre a ferramenta desenvolvida e fazer um brainstorming de outras possíveis visualizações de notícias. Foi também enviado um inquérito a cinco faculdades da Universidade do Porto, a fim de recolher a opinião do público em geral sobre a utilidade da ferramenta. No final, conseguimos reunir novas ideias valiosas para futuras visualizações e melhorias da ferramenta.

Keywords: Visualização de narrativas, Visualização de texto, Interação humano-computador

Acknowledgements

First of all, I would like to thank my supervisors Sérgio Nunes and Ricardo Campos for all the guidance throughout the last year. I would also like to thank the Text2Story[1] team for welcoming me into the project and for the helpful participantion in the interviews. Additionally, I would also like to express my gratitude for everyone who helped the evaluation process by filling out the survey.

Thank you to my family, this journey would not have been possible without your unwavering love and support,

To my friends, thank you for your support, laughs, company and motivation when things were not looking their best. I'm happy I got to cross paths with so many good people in the last five years, wherever and however it was.

I could not help but mention TUNAFE, and how much I grew with all of you. Thank you for making my journey in FEUP so memorable and for teaching me the true meaning of *Amizade na Faculdade de Engenharia*.

Lastly, a special thanks to Alex for putting up with my frequent rants and for being the power-house of love and support that made things a little easier.

Catarina Fernandes

"How lucky I am to have something that makes saying goodbye so hard."

Winnie the Pooh

Contents

List of Figures

List of Tables

Abbreviations and Symbols

API   Application Programming Interface
BRAT   Brat Rapid Annotation Tool
DRS   Discourse Representation Structure
DRT   Discourse Representation Theory
JSON   JavaScript Object Notation
MVC   Model-View-Controller
MSC   Message Sequence Chart
NER   Named Entity Recognition
NLP   Natural Language Processing
OCR   Optical Character Recognition
PDF   Portable Document Format
SRL   Semantic Role Labeling

Chapter 1

Introduction

Stories and narrative structures have been an important way of sharing knowledge and information ever since humankind existed. Oftentimes, those stories would be accompanied by an image, as seen in cave paintings in pre-historical art. The format of narratives has evolved, but they remain culturally significant for sharing information about how we perceive the world.

In this chapter, we provide contextual background, introduce the goals of our research work, and present the structure of this document.

1.1 Motivation

It is known that the world is producing more information than ever before. According to DOMO's Data Never Sleeps 10.0 Infographic [23], dated from 2022, every minute 5.9M Google searches were conducted, 500 hours of video were uploaded to YouTube, and 347.2K tweets were shared on Twitter. With figures rising in social media, it would be virtually impossible for any human to process such an amount of information.

Likewise, the amount of news and journalistic information has increased over the last decade [74]. A major factor contributing to this is the digital transition that resulted from the internet's growing popularity. Consequently, news publishers have had to incorporate a sustainable digital strategy in their business models to not depend on the print press. The current digital landscape also includes a rich offer of advertising-funded websites and recently popular forms of media such as podcasts and short videos, TikTok's format of choice, for example. These new forms of media have expanded the reach of digital content and provided new opportunities for engagement and interaction.

This increase in the amount of content available and the number of channels where it is possible to come in contact with it has led to some challenges for the average reader [63]. The first is filtering the available news content in search of relevant information, which refers to choosing the most relevant news piece to be further explored. This choice can be influenced by the degree of knowledge of the reader, as more informed readers may be looking for specific information on a particular subject, while less informed readers may be searching for introductory material with

more context. Secondly, interpreting the content of the chosen news piece itself can prove to be a difficult task. If the content is a large and dense corpus of text, exploring the news piece and finding relevant information can be particularly challenging. In short, the general public is expected to manually filter the content they consume through all the information pieces at their disposal and in the document itself to extract the relevant content.

Such tasks demand a certain degree of time and effort that people might not have. This can mean the general public is susceptible to feeling overwhelmed, as they are constantly exposed to information they cannot process. Consequently, one may be faced with readers who have lost motivation to stay informed.

To fight this phenomenon, manual efforts are sometimes made to reduce the mental load put on readers. Visualizations and video essays are ways to summarize what is known up until a certain point in a way that provides the reader with the most relevant information regarding a particular matter in a centralized way. One example of such summarization is Público's timeline of World War I[1].

However, this approach has evident disadvantages: it is a heavily manual process, so it can only be made available for some stories, being reserved for selected ones. Therefore, articles about niche or lesser-known topics cannot take advantage of this approach.

All things considered, having a visualization tool that semi-automatically produces content based on the input of a news article could be of great advantage to the public and journalists.

1.2 Context

The research work presented in this document is built on top of the work of Mariana Costa [22], *Narrative Visualization of News Stories*. Her dissertation proposed the NewsLines tool[2] as a visualization method to help manually analyze news articles in a temporally ordered sequence that succinctly displays the actors, events, and locations of a news story narrative.

In this research work, we aim to extend the research done by Mariana Costa by exploring her tool with real news input, and automating the input production.

The primary objective of this thesis is to explore Mariana Costa's NewsLines [22] visualization solution in narratives. Since the NewsLines tool has only been tried out in illustrative cases, using manually created input, it is now our intention to explore it and integrate it with the annotated news from the Text2Story project.

This thesis was carried out as part of the Text2Story[3] P2020 project which is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within the project UIDB/50014/2020 and LA/P/0063/2020. Overall, the project aims to develop a conceptual framework and operational pipeline to allow narrative extraction from

---

[1] https://acervo.publico.pt/primeira-grande-guerra/cronologia#1
[2] https://github.com/marianafcosta/news-story-viz
[3] https://text2story.inesctec.pt/

textual sources and visually represent them in an automatic or semi-automatic way, focusing on journalistic text written in Portuguese.

1.3   Goals

The main goal of this thesis is to extend the NewsLines [22] tool by automating the production of JSON input and extending the visualization tool itself. This way, we developed a pipeline that converts BRAT annotated data from the Text2Story project to a more versatile file format. This would allow easier integration of that data with visual representation tools, thereby removing one barrier in automatically representing a news text visually.

In summary, the goal of this thesis is based on the hypothesis presented below, which will be studied throughout the document:

> *Automating the conversion of BRAT to JSON fosters the automatic creation of visual-*
> *ization storylines, thus empowering large-scale news story understanding.*

1.4   Contributions

With the completion of this work, we obtained a pipeline that integrates data annotated in BRAT format with the NewsLines tool, removing one barrier of visually representing a news text automatically. The contributions of this research work include:

1. A pipeline to convert the BRAT annotated data of the project to JSON, the input format of NewsLines, using sentences as units for event representation.

2. An extension of such a pipeline by using the semantic role links in the data to extract the relevant events and using them as units for event representation.

3. Improvements of the NewsLines tool, in particular: adjusting it so it was able to take a JSON file with a revised schema as input, extending the interface to allow for the representation of an extended storyline, and adding an option for viewing the BRAT annotations of the selected source file.

4. Gathered knowledge on what narrative elements are deemed important to be in a news visualization and brainstormed visual representation ideas through interviews and sketches with the Text2Story team.

5. Extension of the Text2Story package[4] by adding the option to convert the BRAT annotations to JSON.

---

[4]https://pypi.org/project/text2story

1.5   Thesis Structure

The document divides itself into seven chapters. Chapter 2 presents the literature review of this thesis's relevant fields of study. After, Chapter 3 introduces the Text2Story pipeline and provides contextual information regarding BRAT annotated data and discourse representation structures. Chapter 4 explores the NewsLines tool in detail and offers the necessary information for its extension. Afterwards, Chapter 5 goes over the automatic generation of NewsLines input problem, and the techniques used to generate it from annotations. Chapter 6 demonstrates our system by going over two use case examples and discussing the advantages and limitations of the tool. Finally, Chapter 7 reviews the results of the interviews and survey to evaluate the prototype, and lastly, in Chapter 8, we present the final remarks and ideas for future work.

Chapter 2

Related work

To assist people in managing the growing number of text documents, researchers from the text analytics and information visualization communities have developed tools to facilitate text analysis and make it easier for people to handle information overload [54].

This chapter will explore the current state-of-the-art in information visualization, with a focus on the sub-field of narrative visualization. Before delving into the chosen works, the concepts of visualization and some of its sub-fields will be covered to provide context for the specific task of visualizing news.

2.1 Information Visualization

One of the most widely adopted definitions of information visualization is that of Card et al. [17]: "the use of computer-supported, interactive visual representations of abstract data to amplify cognition". This definition emphasizes the role of visualization in data analysis. In other words, the computer processes the data and displays one or several visual representations. At the same time, the end users analyze the data by interacting with the representations output by the process. [16]

Research in this field is driven by the study of perceptions in psychology. Science has proven that our brains can effectively process large amounts of information and signals when presented visually [68]. We can think of the saying "a picture is worth a thousand words" for illustration. Therefore, information visualization techniques can convert large and complex data (such as demographic data, social networks, and document corpora) into visual representations, enabling users to uncover patterns within the data or verify analysis results. [16]

Analyzing unfamiliar documents can be difficult due to the extensive text corpus, dynamic changes in data, and information from multiple sources. This is where text visualization comes in [30]. There are four categories of visualizations designed to aid in the analysis of these documents [16]:

1. Showing similarity, i.e., illustrating content-wised similarities of different documents

2. Showing content, i.e., illustrating different aspects of the text data content

3. Showing opinions and emotions, i.e., summarizing the sentiment or emotional profiles of individuals based on their text data

4. Exploring the corpus, i.e., developed to help analysts or end-users to explore the text data efficiently

Text analysis encompasses several tasks, such as summarization [4], word-level analysis [46], topic visualization [36, 11], event visualization [49], and storyline visualization [65, 81].

2.2 Narrative Visualization and Storytelling

This section presents an exploration of the narrative visualization design space in Subsection 2.2.1 and an overview of a narrative extraction pipeline based on the survey by Santana et al. [75] in Subsection 2.2.2.

2.2.1 Exploring the Design Space

Segel et al. [76] explored and analysed the design space of narrative visualization. Their organization of the design space divides features into three groups: genre, visual narrative tactics, and narrative structure tactics. A summary is presented in Table 2.1.

The examples found were categorized into seven genres: magazine style, annotated chart, partitioned poster, flow chart, comic strip, slide show, and film/video/animation. The genres mostly vary in terms of the number of frames and the ordering of their visual elements, though these characteristics are not mutually exclusive. Deciding on the appropriate genre for a certain narrative representation depends on the complexity of various factors, such as the data and the story, the target audience, and the chosen medium for communication.

The visual narrative tactics can be subdivided into three topics: visual structuring, highlighting, and transition guidance. Visual structuring refers to how the narrative structure is communicated to the reader. It helps orient the reader early on (e.g., via an establishing shot, checklist, or consistent visual platform) and allows users to track their progress through the visualization (e.g., via a progress bar or a timeline slider). Highlighting refers to the visual techniques that help guide the viewer's attention toward particular elements in the visualization (e.g., through the use of color, motion, framing, size, audio). Transition guidance refers to the strategies to move within or between visual scenes without confusing the user.

Likewise, narrative structure tactics are also subdivided into three topics: ordering, interactivity, and messaging. Ordering refers to the ways of arranging the sequence in which viewers navigate the visualization. This sequence can either be linear (prescribed by the author), random access (no path at all), or user-directed (the user must select a path among multiple alternatives). Interactivity regards how a viewer can manipulate the visualization (e.g., filtering, selecting, searching,

Topics and subtopics		Examples
Genre		magazine style, annotated chart, partitioned poster, flow chart, comic strip, slide show, and film/video/animation
Visual narrative tactics	Visual structuring	establishing shot, checklist, consistent visual platform, progress bar, timeline slider
	Highlighting	use of color, motion, framing, size, audio
	Transition guidance	continuity editing, animated transitions, object continuity, camera motion
Narrative structure tactics	Ordering	linear, random access, user-directed
	Interactivity	filtering, selecting, searching, navigating, explicit instruction, tacit tutorial, initial configuration
	Messaging	labels, captions, headlines, annotations, articles, introductions, summaries

Table 2.1: Design space analysis of narrative visualization [76].

or navigating) and how they learn those methods (e.g., explicit instruction, tacit tutorial, or initial configuration). Messaging refers to the way observations and commentary in a visualization are communicated to the user. It can be done through short text fields (such as labels, captions, headlines, and annotations) or more substantial descriptions (such as articles, introductions, and summaries).

The authors also place narrative visualizations in a spectrum of approaches, from author-driven to reader-driven. The author-driven approach has a strictly linear progression that leaves no room for interactivity. Film and static slide shows are typical examples of this approach, as seen in cinema, educational videos, and business presentations. In contrast to the author-driven approach, the reader-driven approach does not have a predetermined sequence of images and encourages interactivity. This technique allows for the diagnosis of data, the discovery of patterns, and the formation of hypotheses. An essential attribute of narrative visualization is its flexibility in balancing both approaches. It is vital to provide space for limited interactivity within a more structured narrative.

A few hybrid models were found to be the most common by the authors, as seen in Figure 2.1. The Martini Glass structure starts by following the author's intended narrative, opening up to a reader-driven stage once it is finished, where the user is free to interact and explore the data. The Interactive Slideshow balances both the reader-driven and the author-driven approaches. Its structure is made up of a regular slideshow that, within the confines of each slide, incorporates interaction mid-narrative, allowing the user to explore specific topics. The Drill-Down Story approach presents itself in the reader-driven end of the spectrum. It presents a broad theme and allows the audience to choose which particular topic instance to explore. Despite being a reader-driven approach, it still requires considerable amounts of authoring to define which stories to include.
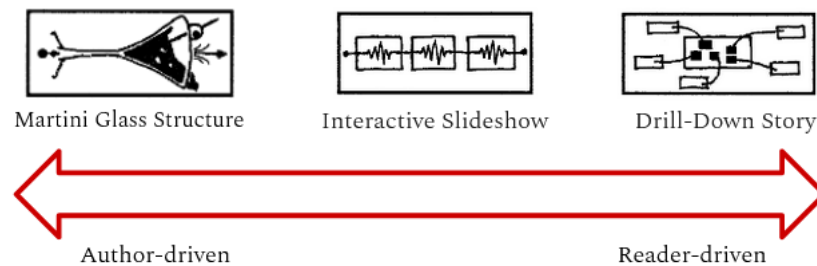
Figure 2.1: Common hybrid models positioned in the author-driven to reader-driven spectrum of narrative visualization approached [76].

Ware [85] states that data display for analysis can often be split into *exploration* and *explanation*. The first concept focuses more on allowing individuals to make discoveries within the data and uncover patterns, while the second is linked with making someone understand a particular topic or prove that a certain interpretation is correct.

To understand the difference, the author states that it is key to consider who or what controls the cognitive sequence. In data exploration, the cognitive processes of the individual who is analyzing are in control. For explanation, on the contrary, the author is in charge of the cognitive flow. The aim is that the audience can absorb a presenter-controlled sequence of visual elements and lines of thought.

This definition can be linked to the one of Segel and Heer [76]. Data exploration shares the same characteristics as a reader-driven approach in visualization, given that the user controls the story flow. Alternatively, data explanation follows an author-driven approach by focusing on the lines of thought that the author wants to transmit.

Pérez-Montoro [71] proposed a narrative visualization taxonomy based on the one by Segel et al. [76] for digital news media. Besides including interaction driver as a topic, ranging from author-driven to reader-driven, his taxonomy includes the ratio visualization-story as a topic. This topic is described as the weight of the information visualization concerning the whole story and is measured using a scale of quartiles: Q1, Q2, Q3, and Q4. Stories in the first quartile (Q1) have a ratio higher than 75%, which can be seen in news pieces like Los Angeles Times' *Don't waste your time at Disneyland. Here's how to avoid the lines*[1]. Those in the second one have a ratio between 50% and 75%. In the third, the ratio must be between 25% and 49%. In the last quartile, the ratio must be below 25%., as exemplified in Quartz's article *England says Oliver is the most popular boys' name, but it's actually Muhammad*[2].

---

[1]https://www.latimes.com/projects/la-fi-disneyland-ride-wait-time/
[2]https://qz.com/1082778/popular-baby-names-muhammad-is-actually-the-most-boys-name-in-england-and-wales-not-oliver

PRE-PROCESSING & PARSING

Sentence Segmentation
Text Tokenization
Data Cleaning
Normalization
Stopwords Removal
Stemming
Lemmatization
Part-of-Speech Tagging
Chunking
Dependency Parsing
Coreference Resolution

IDENTIFICATION & EXTRACTION OF COMPONENTS

Events
Participants
Time
Space

LINKING COMPONENTS

Temporal and Event Linking
Relation Extraction
Semantic Role Labeling
Discourse Relation Parsing
Entity Linking

REPRESENTATION

Formal Representation
Temporal Representation
Visual Representation

EVALUATION

Evaluation Metrics
User Studies

Figure 2.2: The narrative extraction pipeline [75].

2.2.2 Narrative Visualization Pipeline

According to Santana et al. [75], in their survey on narrative extraction from text, the first step in the narrative extraction task is data acquisition and annotation. Data acquisition is a simple process, but it may raise problems due to the lack of available datasets [38, 37] and copyright issues [88]. Data annotation requires data preparation processes such as stripping the data of unimportant information and choosing the most adequate data format while keeping in mind the original format of the data, the tools used in the following phases, and how the annotated corpus will be made accessible. The annotation scheme chosen has to consider the target language but, at the same time, be comprehensive enough to be applied to other datasets. The Ontologies of Linguistic Annotation (OLiA) [18], and ISO 24617—Language resource management—Semantic annotation framework are examples of annotation standards that have gained widespread acceptance and use. In the case of narrative annotation, the annotation framework must include many intertwined semantic layers to enable temporal, referential, thematic, and spatial annotations, since they provide information about important narrative elements such as participants, events, time, space, and the relationships between them. Some annotation tools that include such characteristics are: MMAX2 [60], MAE and MAI [80], BRAT [78], or ANNIS [48].

The pipeline can be divided into five main stages: pre-processing and parsing, identification and extraction of narrative components, linking components, representation of narratives, and evaluation. Figure 2.2 summarizes these stages.

The first stage, pre-processing and parsing, is composed of lexical and syntactical tasks. Lexical tasks split the text into sentences and tokens [66], basic text units, and normalize them into different forms [44, 9]. Syntactical tasks identify the grammar class of those tokens [12] and the dependencies between them [59].

The aim of the identification and extraction of narrative components stage is to extract events, participants, time, and space information. Event detection looks at event structure and its four subtasks to ensure detection. Following ACE terminology [52], the subtasks are event mention, event trigger, event argument, and argument role. This terminology defined eight event types and thirty-three subtypes based on previously defined subtasks. Participants are usually detected using

Named Entity Recognition (NER) approaches. The most common named entities are person, organization, and location, although other types can be considered, such as numeric expressions (e.g., time, date, money, and percent expressions) [62]. Time information detection can be split into two subtasks: extraction, to identify temporal expressions, and normalization, to normalize the found temporal expressions [79]. Identifying places in narratives is a challenging task and has had approaches with entity recognition and toponym resolution techniques [14, 7], which have made possible the recognition of named locations, facilities, and geopolitical entities.

The linking components stage is essential to understand the meaning of the narrative elements. To do this, the relations between temporal expressions and events must be identified, and timelines must be constructed. Strategies for linking events to temporal information have been found without the need for constructing timelines [58]. Other strategies were outlined for stories in specific [73]. Understanding the relations between pairs of identified entities is also relevant to understanding the narrative. These relations also include objectal relations, that is, relations that link referentially related entities. This task is called relation extraction task [72]. Another important task is to link identified entities with the corresponding information in a knowledge base. For example, entity linking would take the sentence *Bush was a former president of the USA* and link *Bush* to the American politician *George H. Bush*, instead of the rock band *Bush*. To understand *who* did *what* to *whom* [34], and *when*, *where*, *why*, and *how* [41], we use Semantic Role Labeling (SRL). This analysis is made on a semantic level, and it captures predicate-argument relations. In a narrative-focused approach, it is also relevant to establish relations of meaning, such as cause and result, e.g., between the different parts of the text. That task is named discourse relation parsing [51].

The next stage of the narrative extraction pipeline is the representation stage. This can be done at the conceptual and the visual levels. At the conceptual level, the most common approaches are to use narrative ontologies [19], frameworks that allow schemes to be built for complex concept representation, and formal semantic representations, usually structures with a recursive formal meaning that have a model-theoretic interpretation such as logic-based representations and abstract meaning representations [45, 8]. At the visual level, Segel and Heer's [76] seven genres of narrative visualizations are mentioned to provide a broad starting point. Additionally, the survey emphasizes visualizations that show the extracted narrative as close as possible to its original form [57, 6]. Another visualization type mentioned was representations of participants and events over time, projecting these elements in two-dimensional layouts [61, 47]. Lastly, knowledge graphs were also discussed [25, 5]. The challenges found in the narrative visualization field were also discussed [29, 82].

2.3 Visualization in News Narratives

2.3.1 Current State of the News Landscape

According to the Digital News Report 2023 from the Reuters Institute for the Study of Journalism [63], which gathers information from six continents and 46 markets, the journalism industry

is experiencing hardship.

Despite specific challenges varying from country to country, publishers face common challenges: weak audience engagement and low trust in an age of abundant digital and social media.

Even though we are living in a climate of political and economic challenges, consumption patterns have shifted, and the general public is becoming less interested in news consumption. Fewer than half of the participants said they were very interested in the news, while at the same time, the number of people avoiding news, often or sometimes, reached record numbers at around thirty percent. This avoidance can be categorized into two groups, periodically limiting all news sources and purposefully restricting news exposure during particular times or for certain topics.

Bearing that in mind, it is not surprising that TV and print continue to fall in most markets, and online and social consumption do not bridge the gap. Additionally, shocks such as the Covid-19 pandemic and the Ukraine War accelerated the shift to digital, mobile, and platform-dominated media environments, with further implications for the business models and formats of journalism.

Despite this change, people in most countries still say they prefer to read the news rather than watch or listen to them, as reading gives them more control on how they access the information. Still, video consumption and podcasting have been growing at a steady pace, the latter still a minority but growing in popularity with educated and younger audiences.

Brand loyalty patterns are also changing, as only around a fifth of respondents prefer to look up news on a brand's website or app, a number that has decreased since 2018 (32%). Young people especially prefer to get their news through side-door routes such as social media, search, or mobile aggregators.

One of the advantages of said side-door sources is that they are often free. Due to financial constraints households face, such as cost of living or the high price of subscriptions, a significant part of the public (83%) says to be satisfied with the news they can get for free. Consequently, the amount of online news subscriptions is decreasing, and so is audience engagement.

However, that increase in access through side-door routes poses a few disadvantages. One notable downside is the growing public disquiet about misinformation and algorithms, reaching near unprecedented levels. This can be explained by the shift from traditional social media such as Facebook and Twitter to youth-focused networks like TikTok and Instagram, which are more strongly reliant on their recommendation algorithms [89, 56]. Regarding news consumption, audiences tend to pay more attention to celebrities and social media personalities than to journalists on networks like TikTok, Instagram, and Snapchat. This clearly contrasts with Facebook and Twitter, that maintain news media and journalists central in the conversation. With less than a third of respondents saying that having stories selected for them based on previous consumption is a good way to get news, it is unsurprising that there is such a wide concern about news and how they are selected.

In addition, despite the idea that making news more accessible through the internet would widen democratic debate, the opposite is happening, with fewer people now participating in online news than in the recent past.

Figure 2.3: Slideshow visualization of a news snippet using the Text2Icons tool [84].

2.3.2 Narrative Visualization Research Works

During the research, many approaches and visualization styles were studied. We gathered literature through a Google Scholar keyword search using keywords related to the information visualization, narrative visualization, and storytelling fields of study. Additionally, we also conducted a recursive exploration through the analyzed papers. In the end, the works were selected to show the wide range of visualization approaches already implemented. We also selected works related to the Text2Story project.

Valente et al. [84, 83] propose a tool for automatic text extraction and conversion into an icons visualization. An example can be seen in Figure 2.3 The tool uses the information regarding extracted narrative elements by the Text2Story pipeline, stored in the DRS files (more about Discourse Representation Structure files in Section 3.3), and produces a graphical representation with little to no human intervention. The narrative elements are obtained from DRS and then translated into English to search for corresponding icons in the icon sources. The Text2Icons visualization is presented through a vertical strip or a slideshow. The tool was evaluated by two surveys. One measured the ability of a set of icons to represent a sentence, and the second measured how well certain icons represented a term.

Gonçalves et al. [32, 31] propose an online platform named Text2Storyline. This platform was based on Time-Matters [15], an online platform that gives users an automatic overview of the key time periods and associated text stories in a short period of time, going past the need to read heavy text documents. Text2Storyline added the feature of allowing queries to be inputted, in addition to the text input and URL input previously available, and the feature of identifying potentially relevant people, events, locations, objects, and concepts to enable users to broaden their knowledge. The interface of this tool, for a given narrative, has five components to display: *Annotated Text*, *Storyline*, *Entities*, *Temporal Clustering*, and *Word Cloud*. Figure 2.4 shows the annotated text and storyline components. The evaluation methods were composed of two surveys, one to understand user preferences regarding information presentation and the other to evaluate the results obtained compared to other systems.

Shixia Liu et al. [54] built an interactive, time-based text analysis tool to aid users looking to analyze large bodies of text, as seen in Figure 2.5. TIARA, or Text Insight via Automated,

(a) Annotatd text.   (b) Storyline.
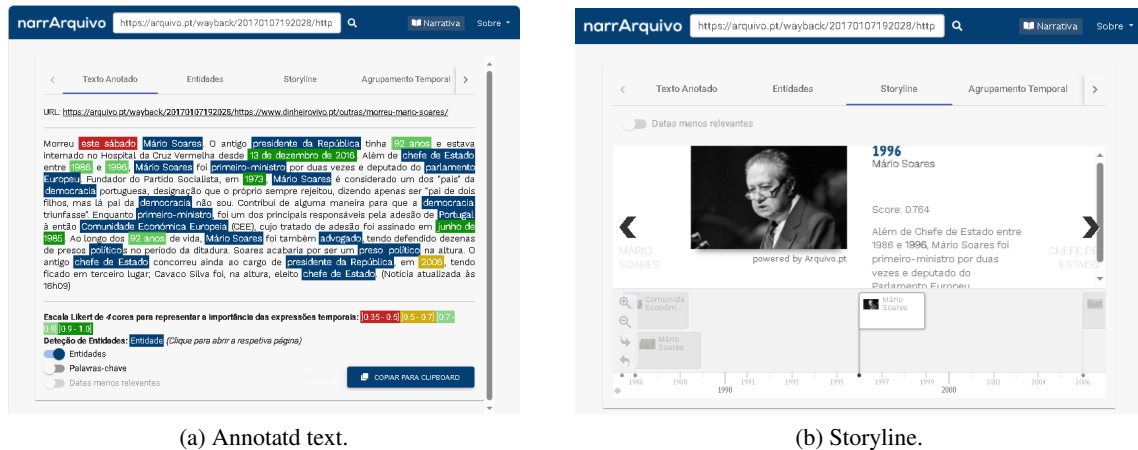
Figure 2.4: Interface of the Text2Storyline tool [32].

Responsive Analysis, uses the Latent Dirichlet Allocation model to automatically extract a set of topics, each depicted by a set of keywords and its probabilistic distributions, from a text collection. That output is then used to summarize complex text and present it visually as a stacked graph so the general user can understand it. As for interaction, a user can zoom in on a section through a magic lens to see more information, focus on a particular keyword and explore its occurrences in the text, and visualize domain-specific metadata, such as email authors and receivers, in a node-link view. The tool was evaluated through user tests.

The work of Dou et al. [24], as seen in Figure 2.6, explores an event-focused perspective of information extraction and visualization, having online news and Twitter feeds as the main data sources. The visualization presents a timeline of the events and the attributes that characterize it, who, what, when, and where. LeadLine incorporates topic modeling through the Latent Dirichlet Annotation (LDA) model, event detection through the Early Event Detection (EED) method, and Named Entity Recognition (NER) techniques to extract crucial information automatically. As for interaction, it allows users to adjust the granularity level of the shown events and explore the event attributes. The tool was evaluated through case studies with eight users.

Michelle Wilkerson et al. [86] propose Story Builder, as seen in Figure 2.7, a plug-in for CODAP, targeted at young students. CODAP, the Common Online Data Analysis Platform, is a free, online, open-source drag-and-drop interactive data analysis system built to be used in schools. It is inserted into the Writing Data Stories project (NSF IIS-1900606), which intends to familiarize students with data and data storytelling. CODAP allows young students to load data as a .csv and add data visualizations. Those visualizations come in the shape of scatter plots and maps and can have common data transformations (e.g., creating new variables, grouping data hierarchically, filtering or sorting their dataset) applied to them. The Story Builder plug-in allowed for a core innovation in CODAP's visualizations: the ability to add a temporal dimension. Consequently, data could be shaped and re-told as a sequence of moments, snapshots of the CODAP document that evolve as students move through the analysis process. As the story progresses and the user moves through the sequence of moments, the visualization windows may decrease, increase, or go
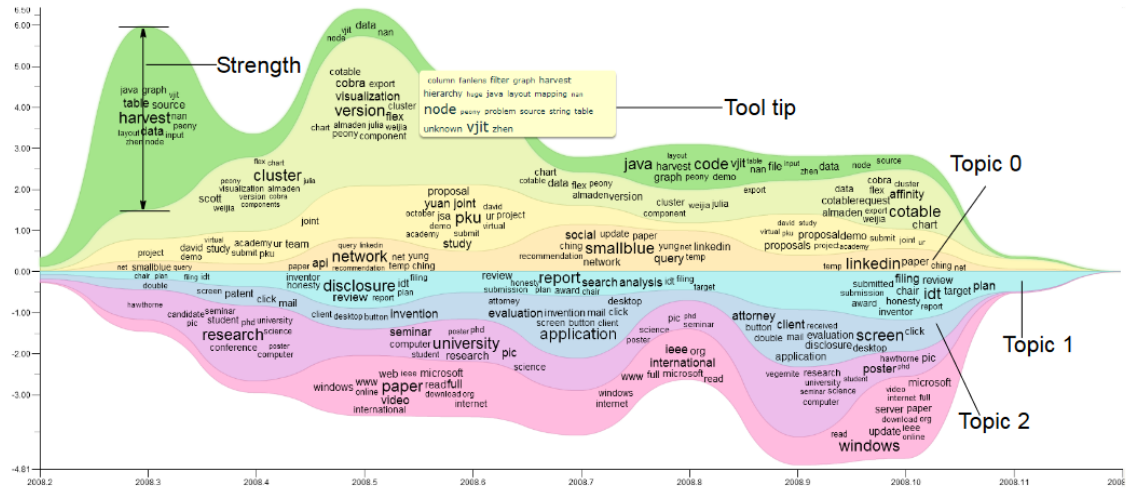
Figure 2.5: Visual interface of TIARA [54].
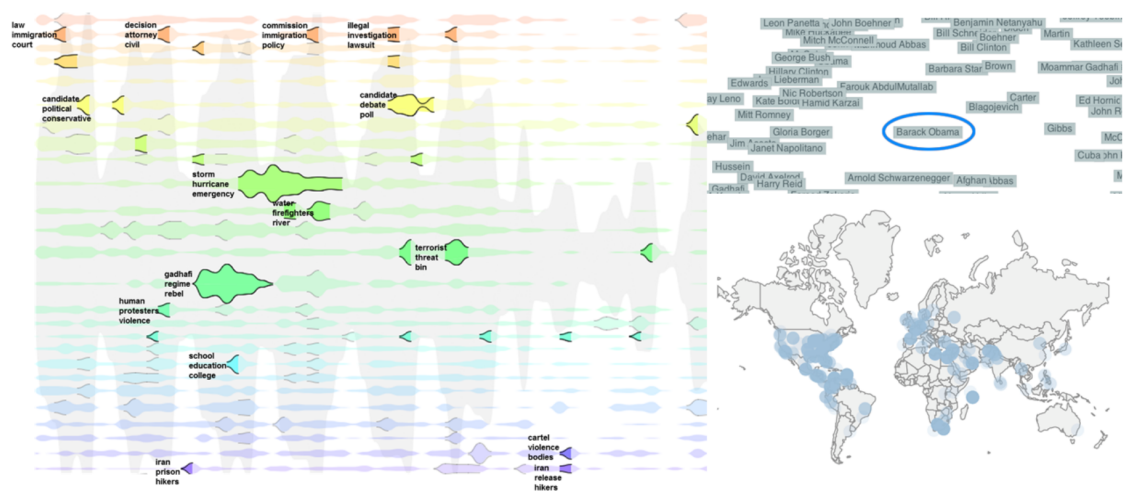
Figure 2.6: The many visual interfaces of LeadLine [24].
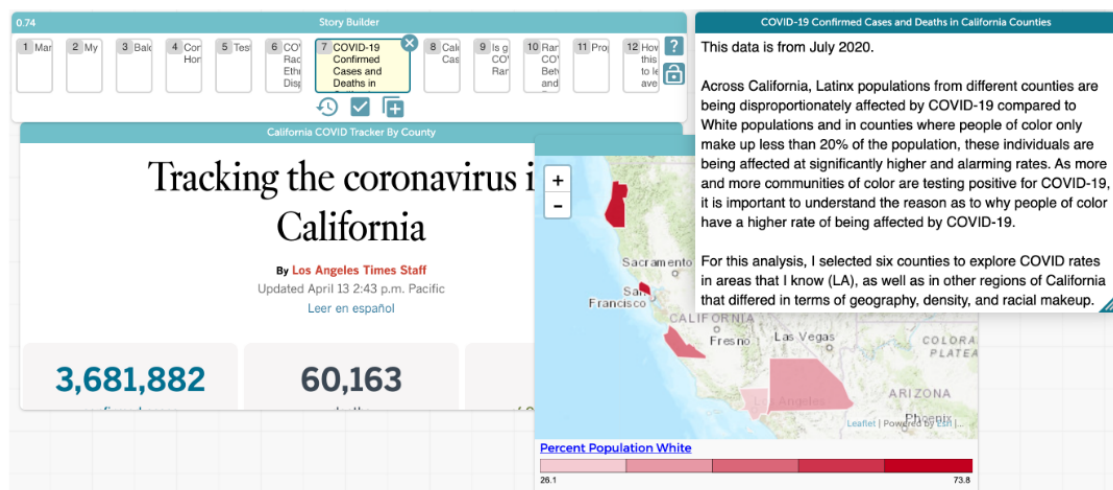
Figure 2.7: Story Builder, a CODAP plug-in for storytelling [86].

to the background or foreground. The tool was being tested at the time of the release of the paper with users of adolescent age.

Shixia Liu et al. [53] propose StoryFlow, an optimized approach to generating the storyline representation. This approach is based on the narrative chart designed by Munroe [61] and it effectively manages hierarchical entity relationships across time, as seen in Figure 2.8. To hierarchically organize the visual elements (e.g., lines), the pipeline creates a dynamic relationship tree, followed by three optimization algorithms that intend to create an aesthetically appealing and compelling layout. Each line represents an entity, and lines side-by-side represent entity interactions. However, this layout method does not account for contextual information, such as the locations of the narrative. The optimization goals mainly focus on one of the most important metrics according to a previous study [70], the number of line crossings. Besides this, it seeks to reduce the wiggle number, distance, and white space. In the end, the work introduces a hierarchy-aware storyline layout to visually display hierarchical relationships among entities, as well as an interactive and progressive rendering method for a large number of storylines with adaptable levels of detail. Interaction-wise, the user can examine the details of an aggregation of lines by expanding the layer (e.g., Bundling), they can add/delete an entity (e.g., Adding/Deleting), they can order the visual elements (e.g., Interactive Ordering), and they can follow a specific entity by straightening its line and bringing focus to it (e.g., Line straightening). To test the capabilities of the tool, a case study was conducted with a PhD student and a professor.

2.4 Discussion

Analyzing our findings from Section 2.3.2, we can categorize the works in terms of the visualization technique they present. Such can be seen in Table 2.2.

What can be concluded from the table is that there is a wide variety of techniques that can be used for information visualization. A total of nine different possible representation types were
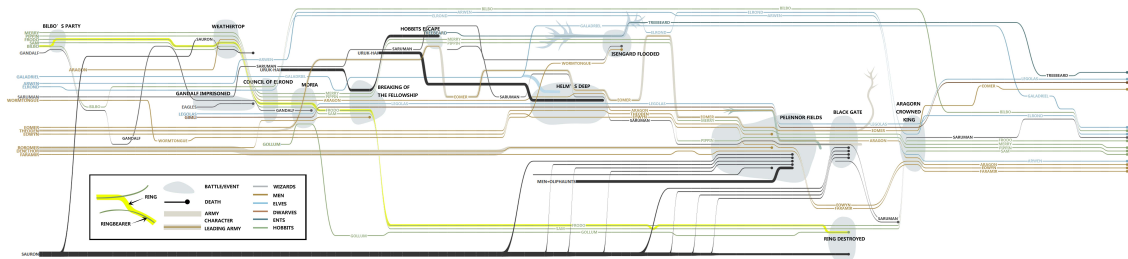
Figure 2.8: A storyline visualization of the movie The Lord of the Rings using the StoryFlow tool [53].

	Node-link	River	Scatter plot	Map	Timeline	Pixel-based/Area	Icons	Interactive Slideshow	Tag Cloud
TIARA	✓	✓							
The CODAP Story Builder			✓	✓				✓	
StoryFlow					✓				
Text2Icons							✓		
Leadline		✓		✓		✓			
Text2Storyline					✓				✓

Table 2.2: Visualization techniques identified in the papers studied.

	Select	Explore	Reconfigure	Encode	Abstract/Elaborate	Filter	Connect
TIARA	✓	✓		✓	✓	✓	✓
The CODAP Story Builder	✓	✓	✓	✓	✓	✓	
StoryFlow	✓	✓	✓		✓		
Text2Icons		✓	✓				
Leadline	✓	✓	✓	✓	✓	✓	✓
Text2Storyline	✓	✓	✓	✓	✓	✓	✓

Table 2.3: Interactivity features identified in the papers studied.

gathered from the articles analyzed.

We also analyzed the interactivity features of the tools studied, as seen in Table 2.3. The taxonomy used for the analysis was that of Yi et al. [87]. In their research, they proposed seven general categories of interaction techniques widely used in Information Visualization: *select* (to choose a subset of data items for further operations), *explore* (to examine a different subset of data items), *reconfigure* (to change the spatial arrangement of data representations), *encode* (to change the visual representation of data items), *abstract/elaborate* (to adjust the level of detail shown in data representations), *filter* (to remove uninteresting data items from view), and *connect* (to show relationships among data items). The most common feature was explore, followed by select, reconfigure, and abstract/elaborate.

Additionally, we analyzed the evaluation methods used in the research papers. These findings are summarized in Table 2.4. We found an even distribution across all methods: user testing, surveys, and case studies.

	User tests	Surveys	Case studies
TIARA	✓		
The CODAP Story Builder	✓		
StoryFlow			✓
Text2Icons		✓	
Leadline			✓
Text2Storyline		✓	

Table 2.4: Evaluation methods identified in the papers studied.

2.5 Summary

In this chapter, we covered the key concepts for understanding the narrative visualization field. We studied approaches already implemented to better understand the wide range of possible visualizations, including their representation type, interactivity features, or evaluation techniques.

Now, with a deeper understanding of the narrative visualization field, we can propose an approach to tackle the domain-specific matter of news visualization.

Chapter 3

Text2Story Pipeline

This chapter thoroughly examines the specific narrative visualization pipeline we will work with in this research work.

It begins with a brief overview of the pipeline in Section 3.1, followed by an in-depth analysis of the annotation process in Section 3.2. Section 3.3 explores the relevance of Discourse Representation Structures (DRS) to the project, while Section 3.4 focuses on the visualizations within the pipeline.

3.1 Overview

The initial stage of this pipeline involves a corpus of text from the news genre, which has been annotated according to a specific annotation scheme. After having the BRAT annotated text, we convert it into an intermediate logic-based representation, a DRS file. From there, a visualization is generated, either a MSC or a Knowledge graph.

3.2 Annotation

The project annotation scheme [77, 50] is composed of an harmonization of 4 different standards, ISO-24617-1/4/7/9 [40, 41, 43, 42]. These standards stem from the *Language resource management-Semantic annotation framework*: Part 1- Time and events [40], Part 4- Semantic roles [41], Part 7- Spatial Information [43] and Part 9- Referential annotation framework [42]. Besides promoting interoperability, this model has allowed for manual annotation of relevant features enabling the generation of visual representations of the narratives in news texts.

The annotation tool used to manually annotate the news corpora was BRAT [78], Brat Rapid Annotation Tool. This web-based tool allows users to annotate existing text documents and facilitates collaboration by providing an online environment for collective annotation.

Silvano et al. [77, 50] have stated that the annotation process has four stages, as seen in Table 3.1. The first stage regards the temporal layer, annotating time expressions, events, and temporal relations between them. The second stage annotates discourse entities, the referring expressions

	Annotation of events
1st stage	Annotation of times
	Annotation of temporal, aspectual, and subordination links
2nd stage	Annotation of participants
	Annotation of objectal relations
3rd stage	Annotation of semantic role links
4th stage	Annotation of spatial relations and measures
	Annotation of spatial and measure links

Table 3.1: Stages of the annotation process.

used to denote them, and the relations between them. For the third stage, we identify thematic relations between predicates and their arguments and modifiers. Lastly, the fourth stage integrates a set of entity and link structures to annotate spatial information.

The annotations resulting from this framework can be mainly categorized into two groups: Entity structures and Link structures. The structures and their attributes can be consulted in Tables 3.2 and 3.3.

Entity structures can be classified as:

- Events, usually used to categorize main verbs, contain attributes that describe their class, type, part of speech, tense, aspect, verb form, mood, modality, and polarity.

- Times, used to categorize temporal expressions that represent time unit classification. Its attributes describe its type, value, temporal function (usually the publication time, to serve as a reference point for other temporal expressions within the document), and anchortime (a reference point for the news story).

- Participants, used to categorize named entities, which may have attributes to describe its individuation domain (whether it is an individual or set, for example), type domain (whether it is a person, organization, facility, and others), its involvement in the event, and its lexical head (noun or pronoun).

- Measures, corresponding to magnitudes of a spatial dimension or a relation (quantitative or non-quantitative), are characterized by their value, unit, and modifier (non-mathematical).

- Spatial Relations, describing textual fragments that indirectly connect to qualitative spatial relations or movement. They can be represented by the topological and path defining attributes.

Meanwhile, Link structures include:

- Temporal links, which establish different temporal relationships between two events, two times, or between an event and a time.

- Aspectual links, which define relations between aspectual events and their event arguments.

Entity Structures		
Events	class	occurrence, state, reporting, perception, aspectual, I-action, I-state
	type	state, process, transition
	pos	verb, noun, adjective, preposition
	tense	present, past, future, imperfect, none
	aspect	progressive, perfective, imperfective, imperfective-progressive, perfective-progressive, none
	vform	none, gerundive, infinitive, participle
	mood	none, subjunctive, conditional, future, imperative
	modality	*dever, poder, ter de, ser capaz de*
	polarity	negative, positive
Time	type	date, time, duration, set
	value	specific value
	anchortime	time ID (select relevant time)
	temporal function	publication_time
Participants	lexical head	noun, pronoun
	involvement	0 ,1, >1, all, undefined
	domain	individuation: set, individual, mass
		types: per, org, loc, obj, nat, fac, veh, path, pl_water, pl_celestial, pl_civil, pl_country, pl_mountain, pl_mount_range, pl_capital, pl_region, pl_state, other
Measures	value	real numbers or nonquantitative measurement expressions
	unit	unit of measurement
	modifier	non-mathematical modifiers
Spatial Relations	topological	dc, ec, po, tpp, itpp, ntpp, intpp, eq, in
	pathDefining	start, end, mids; goalDefining

Table 3.2: Text2Story entity annotation scheme [77].

- Spatial links, which relate the location of the figure (the entity being located) with the location of the ground (where the figure is located). The markable is a spatial relation.

- Subordination links, that represent the different relations between subordinated and subordinating events.

- Objectal links, that describe the referential relation between two entities.

- Semantic role links, which identify how an entity is involved or participates in an event. These links can be established between an event (the predicate) and a participant or another event (the argument).

When the annotation process is finished, the document is exported and viewed as a BRAT annotated file. The BRAT annotation file is a result of the manual annotation process previously described. An example of a news story, Lusa 0 from the *Lusa* collection, can be found in Section A.1, and the BRAT annotations resulting from that file in Section A.2.

Link Structures	
Temporal links	before, after, includes. is_included, during, simultaneous, identity, begins, ends, begun_by, ended_by
Aspectual links	initiates, culminates, terminates, continues, reinitiates
Spatial links	qualitative, movement, measure
Subordination links	intensional, evidential, neg_evidential, factive, counter_factive, condition
Objectal links	objectalIdentity, partof, subset, memberOf, referentialDisjunction
Semantic role links	agent, source, location, path, goal, time, theme, instrument, partner, patient, pivot, cause, beneficiary, result, reason, purpose, manner, medium, means, setting, initialLocation, finalLocation, distance, amount, attribute

Table 3.3: Text2Story link annotation scheme.

3.3 Discourse Representation Structures

Following the annotation step, the annotated file is parsed to build an intermediate logic-based representation, a DRS file. This is done by the Brat2DRS module [5] of the pipeline.

DRS files result from applying the Discourse Representation Theory [45] framework. This approach allows for a straightforward representation of nominal and temporal anaphoric relations and other linguistic phenomena, one sentence at a time.

Building such a file requires creating a dictionary of the linguistic elements in the BRAT annotation and assigning a symbolic variable to each event. After that, for each expression, textual DRS statements that convey the event's properties, such as actors, time expressions, and relations between them, are generated.

This representation was thought to be a high-level dense depiction of the narrative, designed to allow the mapping and representation of all the elements in a narrative.

However, not all the information in the BRAT annotation can be found in the generated DRS file, as seen in Table 3.4. In summary, regarding entities, we can retrieve data about events and participants, and regarding links, there is information about temporal, objectal, and semantic role links. The main consequences of this are:

- While it is possible to determine relative temporal connections between events, such as their sequential order or simultaneity, there is no information in the DRS file about specific time references or absolute temporal expressions.

- DRS files contain little spatial information. Although the semantic roles related to location exist in DRS files, the spatial relations and spatial links are not present.

- Most information about the attributes of the Participants entity structure is lost.

Structures		Found in the DRS file
	Events	✓
	Times	
Entity Structures	Participants	✓
	Measures	
	Spatial Relations	
	Temporal links	✓
	Aspectual links	
Links structures	Spatial links	
	Subordination links	
	Objectal links	✓
	Semantic role links	✓

Table 3.4: Information present in the generated DRS files from BRAT.

3.4 Visualization

The visualization step of the Text2Story pipeline is encapsulated in the DRS2Viz module. This module takes a DRS file as input and deploys a web application with the generated visualization of the chosen file.

The DRS is parsed to extract participants, events, and relationships and stores them in independent data structures. As participants appear frequently in the narrative referred to by different names, the parser marks those different occurrences as referring to the same entity. After the parsing step, the visualization engine generates the representations supported by the project, Message Sequence Charts, and Knowledge Graphs.

3.4.1 Message Sequence Charts

MSCs, or Message Sequence Charts, take a shape known from other domains, such as software engineering and NLP. It is composed of a visual representation that resembles the requirements capture system in the early stages of design in the telecommunications domain [33] and the actors and actor interactions in an NLP scenario [35].

This visualization was implemented using the mscgen_js library [1], which renders message sequence charts from MSC strings. The complete MSC output generated for the Lusa 0 text file can be seen in Figure 3.1.
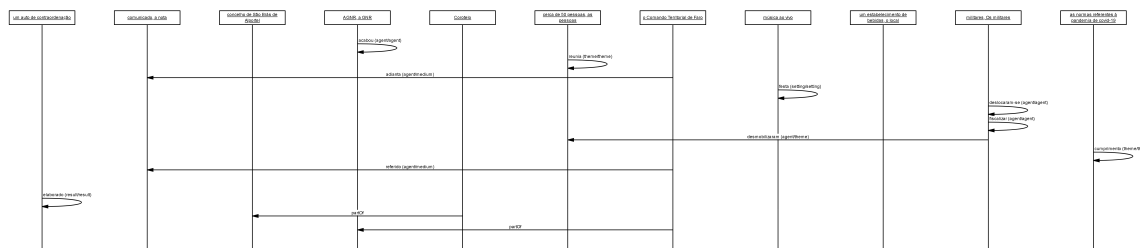
Figure 3.1: Message Sequence Chart of the Lusa 0 text file.

3.4.2 Knowledge Graphs

Ehrlinger et al. [25] define a Knowledge Graph as a tool that collects and incorporates information into an ontology, using a reasoning system to generate new knowledge. Additionally, other possible representations of Knowledge Graphs have been published [27, 67, 69].

Within Text2Story, the knowledge graph was implemented with visjs [2], a browser-based visualization library. The graph displays entities as nodes and links as edges, and users can interact with it by moving, zooming, and repositioning nodes and edges. The beginning of the knowledge graph generated for the Lusa 0 file can be seen in Figure 3.2.
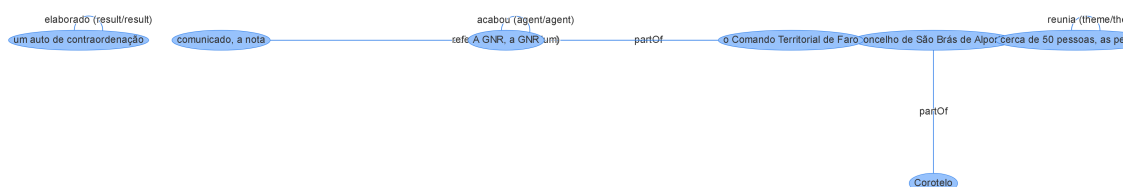
Figure 3.2: Knowledge Graph of the Lusa 0 file.

3.5 Summary

This chapter provided an overview of the Text2Story pipeline, which was designed to extract narratives from text and generate visual representations of such narratives in news texts. The pipeline consists of three main steps: annotation, logic-based representation, and visualization. Each step was described in detail, elaborating on its implementation and relevance for the project.

The annotation step uses a harmonized scheme based on four ISO standards to annotate temporal, referential, semantic, and spatial information in news texts. The logic-based representation step converts the annotated texts into Discourse Representation Structures (DRS), a formal language for natural language semantics. Finally, the visualization step generates two types of visualizations from the DRS: Message Sequence Charts (MSCs) and Knowledge Graphs.

Chapter 4

NewsLines

As previously mentioned in Sections 1.2 and 1.3, this research will be using Costa's work [22] as a foundation, with the main goal of extending it and automating the generation of its input. Therefore, after studying the input source and format in Chapter 3, we will now explore the tool developed by Costa.

This chapter analyzes the NewsLines tool, studying its starting point, interface and main components, interaction features, input data, and architecture.

4.1 Overview

NewsLines [22] was implemented having D3 Narrative Layout as a base, an open-source (licensed under the MIT License) narrative charts layout engine for D3 [26] which aims to mimic the storyline visualizations of Munroe [61]. The engine requires a JSON file of characters and scenes as input and outputs the position of each event in the storyline. The styling and presentation of the output are left to the discretion of the user.

On top of the engine previously mentioned, NewsLines comes as a browser-based prototype that presents a visualization of said participants and events, allowing a user to explore the content of a news story through a narrative-driven lens, with a focus on the participants, events, locations, and dates, and the relationships between them.

The source code can be found on Github [21], licensed under the MIT license, and a live version is also available [1].

4.2 Interface

The interface consists of three main sections: the visualization, the event information module, and the sidebar, as seen in Figure 4.1.

---

[1]https://marianafcosta.github.io/news-story-viz/
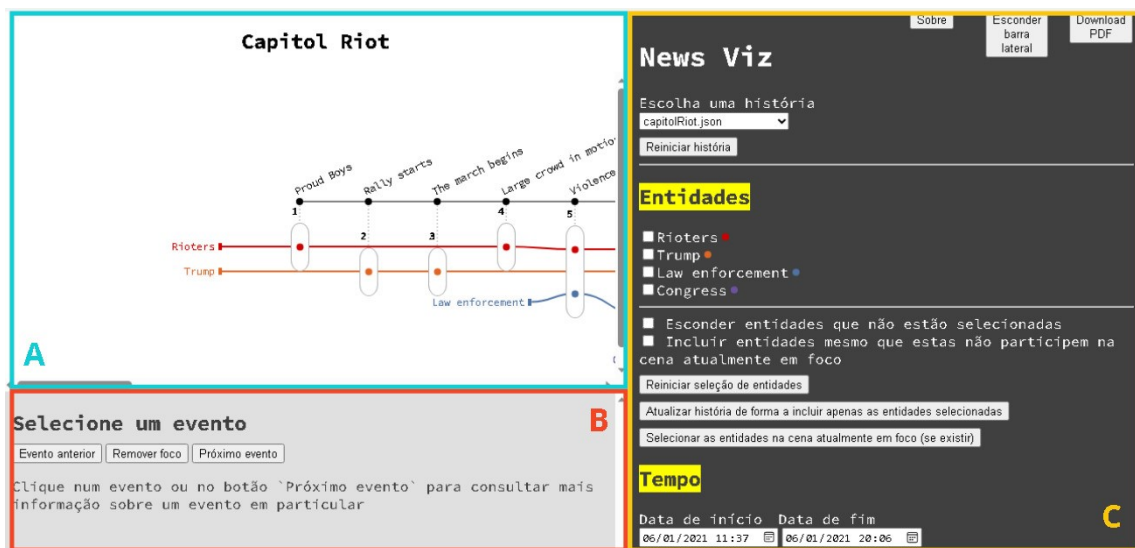
Figure 4.1: The interface of NewsLines. Section A depicts the storyline visualization, Section B depicts the Event information module, which shows the information regarding a selected event, and Section C depicts the sidebar, where most user interaction occurs.

The visualization, section A of Figure 4.1, is the core component of the interface. It is constituted by a storyline where events are central. Events are characterized by a description, title, the entities that participate in them, and, optionally, a date and a location. Visually, events are represented as ellipses, with circles inside representing the appearance of entities in the event. The color of these circles corresponds to the entity participating in the event. Each participant has a colorful line attributed to it and its name on the leftest edge of the line. The colors of the participant's lines can either be defined by the user when providing the input or randomly generated.

A black horizontal timeline is located above the participants' lines, linking the events to their respective descriptions or dates.

The visualization itself, as seen in Figure 4.2, also allows for multiple interactions. The user can hover over an event, which then shows the date and location of that particular event through a tooltip. The tooltip is only visible if the event in question is in focus or if no events are in focus. The user can also drag a number of components of the visualization. Events can be dragged vertically, Event introduction nodes can be dragged either vertically or horizontally, and the nodes in the timeline can be dragged horizontally. Besides this, it is also possible to click on an event to set it as the event on focus, which will update the event information module with the information of that specific event.

The Event information module, section B of Figure 4.1, displays the information relative to the event in focus (or placeholder text in case no event is in focus). A more detailed view can be found in Figure 4.3. The buttons "Next event" or "Previous event" are used to move forward or backward in the narrative sequentially by changing the event in focus. To return to a state where no events are selected, the "Reset focus" button should be pressed.
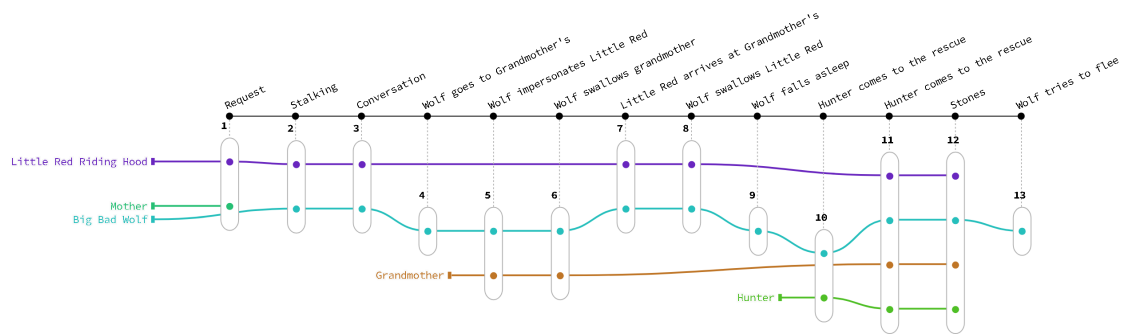
Figure 4.2: The visualization module of NewsLines, displaying the "Little Red Riding Hood" story.

The module displays the Event title at the top, the time and location information below (respectively identified by the clock and red pin icons), and the event description below the interactive buttons. In the description, every time a participant is mentioned, their name or synonyms appear highlighted using that character's color.
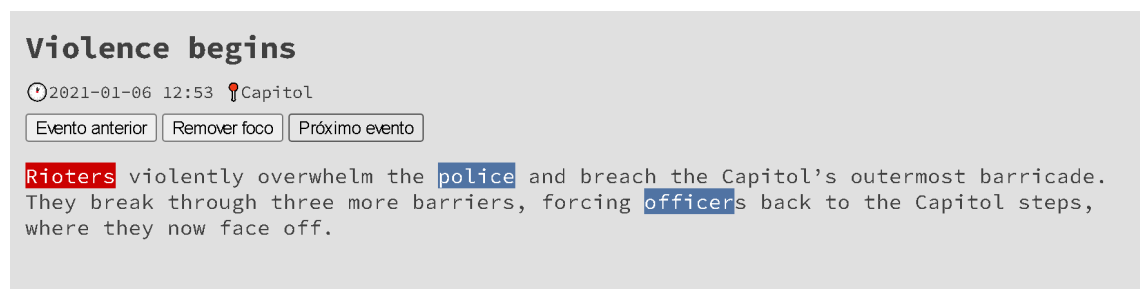
Figure 4.3: The Event information module of NewsLines, focusing on the scene "Violence begins" from the "Capitol Riot" story, with Law enforcers and Rioters as participants.

The collapsible sidebar, section C of Figure 4.1, is the main center of interaction. It is composed of 3 main sections: Entities, Time, and Events. Each section groups the possible ways any user can interact with the tool. Additionally, at the top, there is an option to download the visualization as a PDF (reflecting all current interactions and updates applied to it) and an option to get contextual information through an About page, which provides information on the project, licenses, and links to the source-code and the D3 Narrative Charts implementation. The sidebar is further detailed in Figure 4.4.

The Entities section, section A of Figure 4.4, allows for a user to focus on certain entities of their choice through the entity checkboxes, to hide the entities that are unfocused, to always show the entities in focus even if they do not appear in the current scene in focus, to update the story showing only the focused entities, to select the entities in the scene currently in focus, and to revert to the initial state.

The Time section, section B of Figure 4.4, allows filtering by a temporal interval specified by the user by either hiding the events outside of the interval or updating the visualization to show only the events inside the interval, changing the display of the storyline to show the time information

instead of the event description, and changing the spacing of the event nodes by updating them to be proportional according to their date.

The Event section, section C of Figure 4.4, lists all the events displayed and allows the user to choose which ones to select. It is also tied to the Entity section by giving the option of hiding the events that contain none of the selected entities.

4.3 Input Data

The input data of NewsLines follows a JSON format standard. The file should have two main mandatory sections, a section for characters and another for scenes. Both should be dictionaries containing a number of elements.

Each character in the characters section is uniquely identified by the *id* field. The *name* field is also required. Optionally, the *affiliation* and *synonyms* fields can also be specified. The *affiliation* field refers to the RGB code of the character line in the visualization, and the synonyms field represents the list of other names that refer to the same entity.

In the scene section, each object requires the *characters*, *title*, and *description* fields. The *characters* field has the list of entities that take part in the event, the *title* field regards the short summary of the event to be displayed in the storyline visualization, and the *description* presents the full text of the event, to be displayed in the Event information section of NewsLines. Additionally, each scene can have information regarding its *date* and *location*.

This JSON format is similar to the one presented in the D3 Narrative Layout [26], with a few modifications. The detailed schema and an example can be found in Appendix B.1 and B.2 respectively.

Upon initialization, the system loads a pre-defined narrative. The JSON file is then parsed to ensure it matches the format required by the application, and if so, the narrative is set up and the interface updated. After this initial setup, the interface is ready to respond to user interaction. It is important to note that none of the changes in the interface are reflected in the original input file, as a copy of the original is kept to ensure that resetting the story is possible. When the user selects another storyline in the interface, a new file is loaded, and this process repeats.

4.4 Architecture

The tool was built using the Model-View-Controller architecture, described in Figure 4.5. This structure was chosen to ensure no direct dependencies between models and views.

The models encompass all information relevant to the narrative, including events' descriptions, titles, and their absolute positions within the visualization. The views manage the visual interface and are responsible for tasks such as linking input elements to event listeners and updating the interface to conform to the changes in the model. The controllers serve as a connection and point of contact between the models and views and are responsible for updating the models or views

according to the events triggered. An example of an interaction through the MVC architecture can be observed in Figure 4.6.

4.5 Summary

NewsLines was built based on an open-source engine, D3 Narrative Layout [26], which was then adapted to generate a visualization that better fits the needs of news story representations. With this in mind, two extra modules were added to the interface, an event information module and a sidebar. The visualization depicts a storyline composed of the events in a news story. The event information module presents a detailed view of a selected event. Finally, the sidebar groups most interaction options, presenting a "control panel" to the user. The web-based prototype was implemented with the Model-View-Controller architecture, and the input files were manually fabricated following the schema presented in Appendix B.1. An example of an input file can be found in Appendix B.2.

Figure 4.4: The collapsible sidebar of NewsLines. Section A depicts the Entity interactions, section B depicts the Time interactions, and Section C depicts the Events interactions.
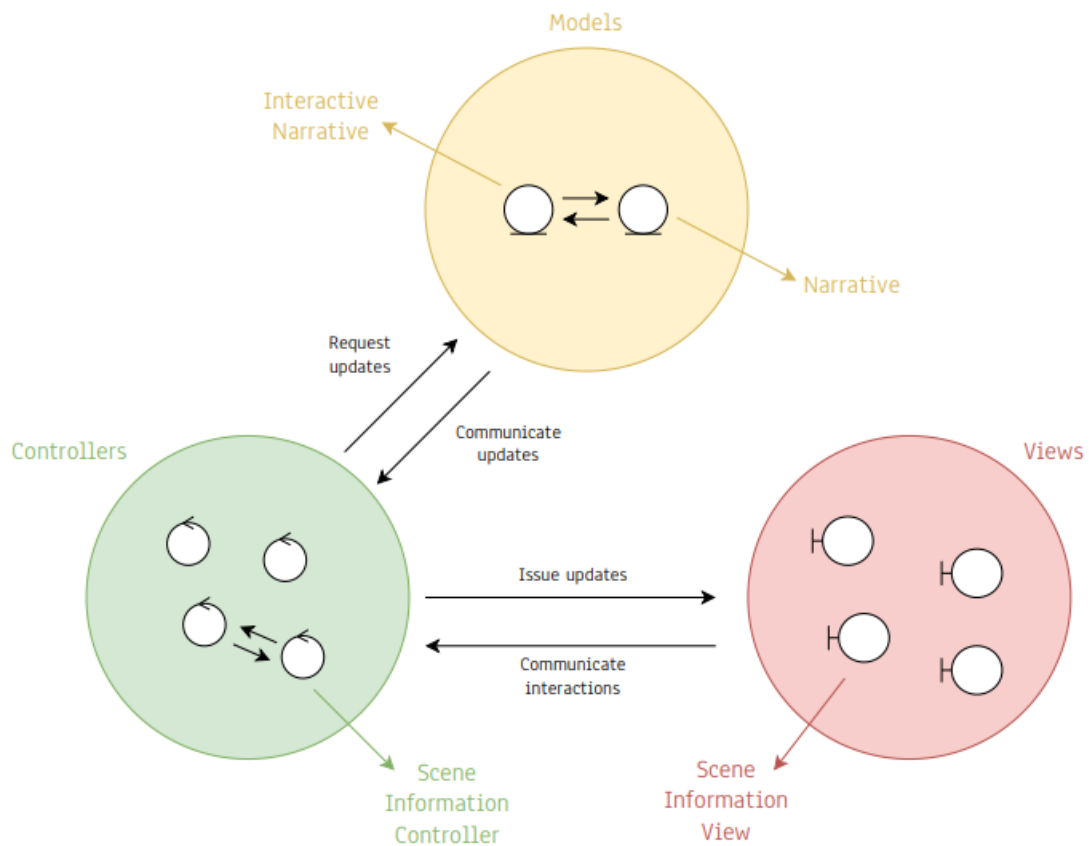
Figure 4.5: A diagram describing the adopted Model-View-Controller approach [22].
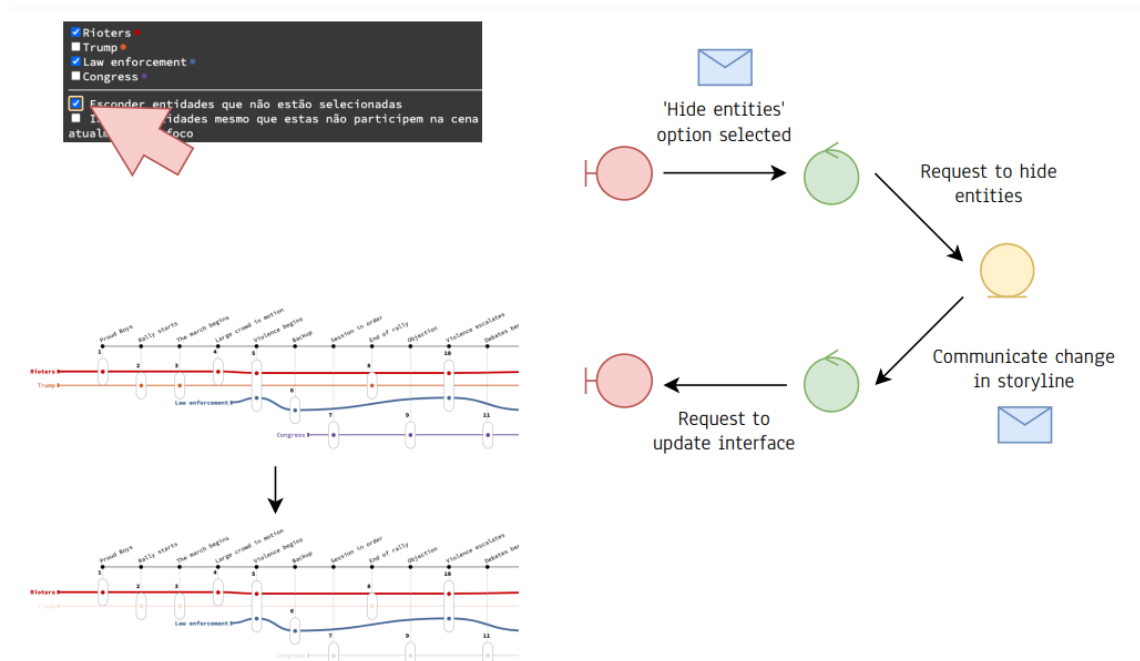
Figure 4.6: An example of an interaction in the application and how the Model-View-Controller architecture processes it [22].

Chapter 5

Methodology

Having explored the NewsLines tool, we will now dive deeper into formulating a problem definition and outline our approach to address it. This chapter identifies the specific challenges that need to be overcome, explains the methods we used to tackle them, and discusses the limitations of the proposed solution.

5.1 Problem Definition

The primary objective of this thesis is to explore Mariana Costa's NewsLines [22] visualization solution in narratives. Since the NewsLines tool has only been tried out in illustrative cases, using manually created input, it is now our intention to explore it and integrate it with the annotated news from the Text2Story project. For that, it is necessary to integrate the project data into the NewsLines tool and evaluate the result with the news narratives. Additionally, we introduced changes in the NewsLines tool to better adapt it to this context and evaluated such changes.

By expanding the tool and adapting it to the news context, we hope to improve its usefulness and make it easier to generate visual representations of news narratives.

5.2 Proposed Solution

The problem at hand required us to analyze which of the file formats made available by the data of the Text2Story project would be fitting as a starting point for our solution. Having analyzed the Text2Story pipeline in Chapter 3, we realized we had two options: the BRAT annotation files and the DRS files. Furthermore, upon exploring DRS files in Section 3.3, we were made aware that these file structures had three limitations that might stall the integration of the data in NewsLines. The first is that, at the time, there are no specific time references or absolute temporal expressions in DRS files, only temporal links, which establish different temporal relationships between events, times, and events and times. However, because most temporal links in the news stories of the dataset are of simultaneous nature, this information alone is not enough to use in the integration. The second is that currently the files present insufficient spatial information, due to the lack of

the qualitative spatial links. Lastly, at the moment, the information about participant attributes is mostly lost, presenting only information of the entity itself. Given these three findings, we decided to use the BRAT annotations as the main source of information.

This integration problem could now be divided into two sub-problems: how to convert the BRAT annotations into a JSON schema that captures the relevant information for narrative visualization and how to extend the NewsLines tool to adapt to the new JSON schema and provide a better user experience. Additionally, we would also look for ways to improve the NewsLines tool outside the data integration part.

For the first sub-problem, we developed a Python pipeline to convert the annotation files into a JSON format, maintaining all the narrative elements and their relations, thereby automating the production of such files. The NewsLines JSON input schema was then updated to better accommodate the information extracted from the annotations, and the tool was modified to enable the integration of the new document schema, providing a solution to the second sub-problem. The modified source code can be found on GitHub [28] and a live version is also available[1]. The code is licensed under the MIT license.

Lastly, we conducted interviews with eight researchers of the Text2Story team to gather collect about NewsLines and gather ideas for other approaches of news visualizations through brainstorming sketches. Additionally, we released a survey to five Faculties of the University of Porto to understand how usefull people find the tool and gather feedback for its improvement.

5.3   Existing Approaches

BRAT standoff to JSON [39] is an open-source tool developed by Astutic Artificial Intelligence. The main focus of the tool is to convert BRAT annotation files to JSONL, according to the annotations configuration file. In short, the tool takes a *.ann* file and a text file as input, and produces a *.jsonl* output file based on the configuration file. JSONL is a text-based format similar to JSON, except it uses newline characters to separate JSON data. This format, also known as JSON Lines, uses the *.jsonl* file extension. As seen in Figure 5.1, the annotation configuration file, usually *annotation.conf*, follows a line-oriented structure similar to other text-based configuration systems. The file is divided into four sections, for *entities*, *relations*, *events*, and *attributes*. Each section follows a simple list format and characterizes a number of tags, explaining the annotation structure through the number and type of arguments, if any, values, and possible hierarchies in the annotation. The output of the tool is a *.jsonl* file that consists of a line entry for each annotation file analyzed. Each line is made up of the source text from the *.txt* file and a list of entities and their beginning and end character numbers in the text.

ScreenJSON [20] is a programmable and cross-platform screenplay data model for the web. It was designed to provide a data model that can be imported and exported across different screenwriting platforms, including web-based platforms like Amazon and WriterDuet, standardizing the format of *screenplays for computers*. Despite the many positive aspects of the data model, the

---

[1]https://catarina03.github.io/news-story-viz-prototype/

```
[entities]
Person
Organization
GPE

[relations]
Located            Arg1:Person, Arg2:GPE
Geographical_part  Arg1:GPE,    Arg2:GPE
Family             Arg1:Person, Arg2:Person

[events]
Report Reporter-Arg:Person|Organization|GPE, Event-Arg:<EVENT>

[attributes]
Individual   Arg:<ENTITY>
Mention      Arg:<ENTITY>, Value:Name|Nominal|Other
Negation     Arg:<EVENT>
Confidence   Arg:<EVENT>, Value:High|Neutral|Low
```

Figure 5.1: Example of an annotation configuration file.

one of interest to us is its JSON format. The file is composed of a *document* container object that describes the document content, such as the language, authors, license, and scenes. Each scene is a container that defines the *Setting & Context*, *Ordering*, *Metadata*, and *Pre-Production Requirements Data*. Participant information is also defined in the *Scene* container object.

Screenplay Parser [10] is a PDF parser that converts screenplay PDFs to JSON. It functions well with standard PDFs, but not with those that use Optical Character Recognition (OCR). The output file structure is an array of dictionaries with information on each scene. Scene structures can be of four different types: *Action*, *Character*, *Dual_Dialogue*, and *Transition*.

Bratpy [3] is a library developed for handling BRAT annotation files. One of the modules *brat.json* converts annotations to a JSON output compatible with the BRAT visualization tool. Given that this library was mostly built for personal reasons, there is a lack of public documentation.

Table 5.1 summarizes the findings of our research. Although ScreenJSON and Screenplay Parser output a JSON file, they are missing the most important characteristic of using a BRAT annotation file as their starting point. Bratpy allows us to produce a JSON file from annotations. However, the output is generated to be compatible with the BRAT visualization tool, which means that the JSON will contain a lot of information that is not useful for the scope of this thesis. BRAT standoff to JSON, on the other hand, does not generate a JSON file with enough information to fit our needs, as the output is limited to the input text and the annotated entities. Information such as relations and attributes is missing from the output.

In conclusion, after conducting our research, we found none of the tools we evaluated were adequate for constructing the proposed solution and changing them for such purpose would be difficult, if possible.

	BRAT standoff to JSON [39]	ScreenJSON [20]	Screenplay Parser [10]	Bratpy [3]
Outputs a JSON file	✓		✓	✓
JSON file contains participant information	✓	✓		✓
JSON file contains event information		✓	✓	✓
BRAT annotation files as input	✓			✓
Needs configuration file	✓			
Documented	✓	✓	✓	
Is a data model		✓		
Is a conversion tool	✓		✓	✓

Table 5.1: Comparison of the existing approaches.

5.4 Data

Before delving further, it is important to provide more context on news production. Journalism covers a wide range of domains, including business, environment, politics, and many others, reporting on events like stock market changes, new temperature records, and policy implementations. Not all of these domains are fit to be represented by a storyline visualization, as they may lack relevant information such as participants, events, locations, or time-related information.

These pieces of information, essential narrative elements, are addressed by the core journalistic questions: *What?*, *Who?*, *Where?*, *When?* and *Why?*. Some also consider *How?* another key question, as it refers to the methods or processes behind a story. These questions are typically answered by journalists in their articles, either explicitly or implicitly. In summary, the selection of the data set of news articles to annotate [64] was made previously, in the context of the Text2Story project, with these questions in mind. The files selected resembled narratives with a storyline and consequently were able to answer these questions.

The data were obtained from 119 news files, carefully curated from the largest news agency in Portugal, Agência Lusa, according to the previously mentioned criteria.

After selecting the news texts, those files were annotated by hand following the scheme in Section 3.2 by the team of the Text2Story project, generating 119 annotation files. Those annotations were then the input of the Text2Story pipeline, generating the corresponding DRS representations and MSC and Knowledge Graph visualizations. However, we will concentrate solely on the files annotated with BRAT, as they were selected for our integration.

5.5 Integrating Text2Story Data in NewsLines

As mentioned in Section 4.3, the NewsLines visualization takes a JSON file as input. In contrast, the files made available by the Text2Story pipeline are the original news text files, BRAT annota-

tion files, and DRS files. Thus, we need a way to convert the information stored in the files into JSON.

For this particular task, we decided to work with the BRAT annotation files for two main reasons. Firstly, the BRAT file format has more documentation available, making it easier to work on. Secondly, compared to the DRS format, the BRAT format contains more information relevant to the integration with NewsLines, such as location data. This additional information makes the BRAT format a better choice for our needs.

As stated in Section 4.3, the original NewsLines schema, depicted in the Listing B.1, has three essential parts: title, participants list, and scenes list. Each participant has a list of synonyms, and each scene has a list of participants appearing in it. Having made this analysis, we identified that the first step was to extract information regarding the Participants, as both the participants' and scenes' lists depend on them. The following step was Event extraction, as they provide the critical information needed for the scenes list section. Lastly, we needed to extract the time and spatial information, which are attributes that provide additional information about each scene.

All these tasks required a determined amount of preprocessing beforehand, which consisted of parsing all the annotated data and gathering it in an adequate data structure for further processing. As previously stated in Section 3.2, annotated data can be of two different types: entity annotations and link annotations. These two types represent different information and have different structures.

```
<EntityID>\t<Tag> <StartingCharacter> <EndingCharacter>\t<Value>
```

Figure 5.2: Entity annotation structure.

As seen in Figure 5.2, entity annotations can be separated into three sections, each separated by a character tabulation \t. The first section regards the Entity ID, a unique identifier for each entity. According to its Entity type, the first character of the identifier can be an *E* for Events or a *T* for Participants, Time, Measures, and Spatial Relations. The second section is made up of the Entity tag (*Event*, *Participant*, *Time*, *Measure*, or *Spatial_Relation*), and the character that corresponds to the beginning and the end of the entity reference in the text. The last section represents the text value of the referred entity. An example can be seen in the first line of Listing 5.1.

```
<RelationID>\t<Tag> <FirstArgument> <SecondArgument>
```

Figure 5.3: Link annotation structure.

Link annotations can be separated into two sections by a character tabulation \t, as seen in Figure 5.3. The composition of these annotations is similar to the entities one, the only difference being that the first character of the Relation identifier is an *R*, and there is no text value, only two arguments that specify the entities linked by the relation. An example can be seen in the last two lines of Listing 5.2.

After gathering all the annotation data in a dictionary, we group the information provided by the *attribute* tags to its corresponding entity, leaving us with a dictionary organized in the following way:

```
{EntityID: EntityAndAttributeInformation}
```

5.5.1   Participants

The Participants tag clusters together a big amount of information, as seen in Table 3.2, such that not all the entities classified as Participants can be considered characters in the visualization. So, a few assumptions were made to extract the relevant characters of the narrative. One of these assumptions pertained to the Individuation domain attribute, which could only be classified as either *Individual* or *Set*:

Individual: *O rapaz entrou na sala / The boy entered the room*

Set: *Os rapazes entraram na sala / The boys entered the room*

The second assumption concerned the participant type domain attribute. To be considered a character, the attribute value had to be either be *Per*, for person, or *Org*, for organization:

Per: *O rapaz leu o livro / The boy read the book*

Org: *A EDP subiu os preços / EDP raised the prices*

The third and last assumption was that the lexical head attribute could only be *Noun* instead of *Pronoun*, as the same pronoun can refer to different characters in a text. An example of an Entity that would be considered a Participant can be seen in Listing 5.1, taken from Text2Story's Lusa 0 example.

```
T24    Participant 108 113    A GNR
A0     Lexical_Head T24 Noun
A1     Individuation_Domain T24 Individual
A2     Participant_Type_Domain T24 Org
A3     Involvement T24 1
```

Listing 5.1: Example annotation of a Participant entity.

Next, synonyms of the participants were detected. For that, we used the objectal link structures, more precisely *ObjectalIdentity* and *PartOf*. *ObjectalIdentity* means two arguments refer to the same entity, while *PartOf* means one argument is part of another through a part/whole relationship. We go through them recursively so that all levels of objectal relations are considered. While going through them, we add them to a dictionary of synonyms of the original entity. Listing 5.2 shows an example of this annotation, taken from Text2Story's Lusa 0 example.

```
T24    Participant 108 113    A GNR
A0     Lexical_Head T24 Noun
A1     Individuation_Domain T24 Individual
A2     Participant_Type_Domain T24 Org
```

```
A3     Involvement T24 1
T29    Participant 274 303    o Comando Territorial de Faro
A0     Lexical_Head T29 Noun
A1     Individuation_Domain T29 Individual
A2     Participant_Type_Domain T29 Org
A3     Involvement T29 1
T30    Participant 305 310    a GNR
A0     Lexical_Head T30 Noun
A1     Individuation_Domain T30 Individual
A2     Participant_Type_Domain T30 Org
A3     Involvement T30 1
R21    OLINK_objIdentity Arg1:T30 Arg2:T24
R22    OLINK_partOf Arg1:T29 Arg2:T30
```

Listing 5.2: Example annotation of Objectal links.

Having both the characters and their synonyms, we can build the character section of the JSON file. The *id* and *name* match those of the entity that first appeared in the text, the *affiliation* is randomly generated, and the *synonyms* are all the entities that have direct or indirect objectal relations with the original entity. An example of a character in the JSON file is provided in Listing 5.3.

```
{
    "id": "T24",
    "name": "A GNR",
    "affiliation": "rgb(144,6,245)",
    "synonyms": [
        "a GNR",
        "o Comando Territorial de Faro"
    ]
}
```

Listing 5.3: Item of the characters dictionary in the JSON file generated from the file Lusa 0.

5.5.2 Events

Regarding events, the selection requires filtering the initial dictionary for entries with the Events tag, which is the only filtering required. Listing 5.4 shows an example of the Events annotation, taken from Text2Story's Lusa 0 example.

```
T1  Event 114 120 acabou
A0  Class T1 Occurrence
A1  Event_Type T1 Transition
A2  Pos T1 Verb
A3  Tense T1 Past
A4  Aspect T1 Perfective
A5  Polarity T1 Pos
```

Listing 5.4: Example annotation of an Event entity.

5.5.3 Spatial Information

To extract spatial information, we have to start, once more, by looking at the *Participants* tag and its attributes. The methodology was similar to the one in the participants' extraction (see Section 5.5.1) with the difference that the participant type domain attribute had a different range of values.

For the same reasons stated in the Participants Section (Section 5.5.1), we kept the Individuation domain attribute to *Individual* or *Set* and the Lexical head attribute to *Noun*. For the entity to be considered a location, its participant type domain value had to be one of the following: *Fac*, for facility, *Path*, *Pl_water*, for bodies of water, *Pl_celestial*, for planets and other cosmic components, *Pl_mountain*, *Pl_civil*, for political regions or administrative areas, *Pl_country*, *Pl_mount_range*, for mountain ranges, *Pl_capital*', *Pl_region*, for non-political or non-administrative regions, or *Pl_state*. Listing 5.5 shows an example of a valid location annotation, taken from Text2Story's Lusa 2 example.

```
T47    Participant 190 196    Sintra
A207   Lexical_Head T47    Noun
A208   Individuation_Domain T47    Individual
A209   Participant_Type_Domain T47    Loc
A210   Involvement T47    1
R38    OLINK_partOf Arg1:T46 Arg2:T47
R39    OLINK_objIdentity Arg1:T47 Arg2:T45
R40    OLINK_partOf Arg1:T47 Arg2:T48
```

Listing 5.5: Example annotation of a location.

This annotation would lead to a dictionary entry such as the one seen in Listing 5.6.

```
{
    "value": "Sintra",
    "synonyms": "T45",
    "Part_Of": "T48",
    "Parts": "T46",
    "id": "T47"
```

```
}
```

Listing 5.6: Item of the locations dictionary in the JSON file generated from the Lusa 2 example.

Based on the spatial information extracted from the participants, we could detect if a given location is present in a scene. For that, we need to define what is considered a scene. As described below, in Section 5.5.5, we tested two approaches: scenes could be a whole sentence or solely an event and its essential information.

For approach one, assuming a scene to be a whole sentence, we checked if a particular location was present in the scene by detecting if the sentence contained the location value, which is the same methodology NewsLines [22] followed for character detection in a scene, as described in Section 4.2. However, throughout the text, different location instances can have the same value. To prevent a scene from having two location instances with the same value, we included an additional step, which consisted of determining the index numbers of both the first and last characters in the sentence relative to the text and confirming that the starting character of the location was between those two numbers.

On the contrary, for approach two, if a scene is made up solely of an event and its information, there are two steps to approach the problem. Firstly, we need to extract the information regarding a specific semantic role link, SRLINK_location. This semantic role link states where an event occurs, or a state is true, or a thing exists. An example of it can be found in Listing 5.7, taken from Text2Story's Lusa 107 example.

```
T3     Event 133 157    atropelada sem gravidade
A5     Class T3 Occurrence
A6     Event_Type T3 Transition
A7     Pos T3 Verb
A8     Tense T3 Past
A9     Aspect T3 Perfective
A10    VForm T3 Participle
A12    Polarity T3 Pos
T36    Participant 170 188    distrito de Aveiro
A238   Lexical_Head T36 Noun
A239   Individuation_Domain T36 Individual
A240   Participant_Type_Domain T36 Loc
A241   Involvement T36 1
R58    SRLINK_location Arg1:T3 Arg2:T36
```

Listing 5.7: Example annotation of the location semantic role link.

The second step involves qualitative spatial links. These links depict the location of the Figure (the entity being located) with the location of the Ground (where the Figure is located), as seen in Listing 5.8, taken from Text2Story's Lusa 0 example. We extracted the qualitative spatial links and filtered them to obtain the ones that link events to locations. After that, we check if that event is present in the scene and, if so, associate the location with the scene.

```
T3     Event 139 151    festa ilegal
A121   Class T3 Occurrence
A122   Polarity T3 Pos
A123   Event_Type T3 Process
A124   Pos T3 Noun
A125   Tense T3 Pres
A126   Aspect T3 Perfective
T28    Spatial_Relation 152 155    num
A96    Type T28 Topological
A97    Topological T28 Disjunction_TTP_NTTP
A98    PathDefining T28 Starts
T42    Participant 156 159    bar
A5     Lexical_Head T42 Noun
A6     Individuation_Domain T42 Individual
A7     Participant_Type_Domain T42 Fac
A8     Involvement T42 1
R71    QSLINK_figure Arg1:T28 Arg2:T3
R62    QSLINK_ground Arg1:T28 Arg2:T42
```

Listing 5.8: Example annotation of qualitative spatial links.

5.5.4   Temporal Information

To obtain information regarding date and time, we had to filter the annotations by *Time* tag and its attributes. This alone gave us enough information to build the dates section in the JSON file. However, to determine if a given time is present in a certain scene, we needed to apply a different methodology.

Once more, the methodology depends on what is considered a scene. Assuming a scene to be a whole sentence, approach one, we check if a particular time is present in the scene by detecting if the sentence contains the value of the *Time* entity. This is the same technique used for spatial information, as described in Section 5.5.3. Throughout the file, different Time instances can have the same value (for example, if the text references the word *hoje*, *today* in English, in more than one sentence). To prevent a scene from having two date instances with the same value, we included an additional step, which consisted of determining the index numbers of both the first and last characters in the sentence relative to the text and confirming that the starting character of the Time entity was between those two numbers.

For approach two, if a scene is made up solely of an Event and its information, it is unlikely that the temporal information will be explicitly present. Therefore, we need to get that information through Temporal links. As mentioned in Section 3.2 and Table 3.3, temporal links, *TLINK* 's in the BRAT annotation files, can establish temporal relationships between two events, two times, and an event and a time. In this case, we are only interested in the relations between events and times. This relation type is usually represented by the *TLINK_isIncluded* tag, as seen by the example in Listing 5.9, taken from Text2Story's Lusa 1 example. In conclusion, this relation type enables us

to identify the temporal information associated with a specific event and, consequently, provide a more complete picture of a given scene.

```
T21    Time 1028 1032    hoje
A0     Time_Type T21 Date
A1     TemporalFunction T21 Publication_Time
A2     value T21 2020-10-15
T22    Event 1033 1068    presente às Autoridades Judiciárias
A0     Class T22 Occurrence
A1     Event_Type T22 Transition
A2     Pos T22 Verb
A3     Tense T22 Present
A4     Polarity T22 Pos
R19    TLINK_isIncluded Arg1:T22 Arg2:T21
```

Listing 5.9: Example annotation of a date.

After identifying all the *Time* entities occurring in a scene, we added the IDs of the time entities to the *date* field in the scenes section of the final JSON file, as seen in Listing 5.10, taken from Text2Story's Lusa 1 example. Additionally, we assembled a dictionary of all the *Time* entities to be stored in the JSON file in a specific section named *dates*. Each entry in the *dates* section is made up of four fields: *id*, a mandatory field identifying the time entity ID, *time*, a mandatory field with the value of the time entity, *TemporalFunction*, the type of temporal anchor that the time entity represents for other temporal expressions in the document (if it is the publication date, for example), and *value*, which represents the date in a valid date format. An example of a valid entry of this section can be seen in Listing 5.11, taken from Text2Story's Lusa 1 example.

```
{
    "characters": [
        "T26",
        "T28"
    ],
    "description": "Segundo a PSP, o detido é hoje presente
    às Autoridades Judiciárias.",
    "title": "Segundo a PSP, o detido é hoje presente às
    Autoridades Judiciárias.",
    "date": [
        "T21"
    ],
    "location": []
}
```

Listing 5.10: Item of the scenes dictionary in the JSON file.

```
{
    "time": "hoje",
    "TemporalFunction": "Publication_Time",
    "value": "2020-10-15",
    "id": "T21"
}
```

Listing 5.11: Item of the dates dictionary in the JSON file.

5.5.5 Scene Generation

For the first approach, we decided to start with the most straightforward method for scene genera-tion, generating scenes whose titles are a full sentence of the text, to ensure a baseline that would test the integration. For that, we took each text file and split it into sentences using NLTK's tok-enize package[2], to get all the sentences in the text. The next step was going through the character list to see if any of the character values or their synonyms were present in the sentence, and if so, their id would be added to the characters list. We followed the same methodology with the location field. An example of a baseline scene in the JSON file can be observed in Figure 5.12, taken from Text2Story's Lusa 0 example.

```
{
    "characters": [
        "T24",
        "T27"
    ],
    "description": "A GNR acabou no sábado com uma festa ilegal
    num bar em Corotelo, concelho de São Brás de Alportel,
    que reunia cerca de 50 pessoas, foi hoje anunciado.",
    "title": "A GNR acabou no sábado com uma festa ilegal
    num bar em Corotelo, concelho de São Brás de Alportel,
    que reunia cerca de 50 pessoas, foi hoje anunciado.",
    "date": "",
    "location": "T25"
}
```

Listing 5.12: Item of the scenes dictionary in the JSON file.

However, splitting the narrative into scenes this way would make for long scenes with possibly more than one event happening in each scene, which could cause some confusion in the visual-ization. So, to fix this problem, we decided to generate scenes that would show only a summary of the narrative events, in line with the original NewsLines design. The summaries should be as short as possible and try to convey only the most important information.

---

[2]https://www.nltk.org/api/nltk.tokenize.html

To do that, we took advantage of the semantic roles link structure. Semantic roles, as explained previously in Section 3.2, identify how an entity is involved in an event and can have a number of values, as seen in Table 3.3.

For the sake of generating a relevant summary of each event, we have decided to make a selection of the relevant semantic role links. Regarding the criteria of what was relevant or not, we have reunited the semantic role links that answer the *Who?*, *What?*, and *Why?* journalistic questions. The questions *When?* and *Where?* have been disregarded as they will be answered in their own specific field of the JSON file for each scene. We have then identified the following links:

- *agent, pivot, patient, partner, beneficiary, cause, source, goal* - all of these either identify the roles of the participants in the scene or establish a connection between an event and a participant. They can either be the central entity initiating an event (intentionally or unintentionally) or simply undergoing a change directly caused by the event or by a participant in it. They answer the *Who?* journalistic question.

- *amount, path, attribute, theme, result* - all of these represent a piece of information that answers the *What?* journalistic question. Without it, the sentence would be deemed incomplete to a user.

- *reason, purpose* - both of these answer the *Why?* journalistic question, as they explain the occurrence of an event or the intention of a participant in performing a certain action.

After identifying which semantic role links would be helpful in scene generation, we analyzed the BRAT annotations, leading us to some key findings. Firstly, we found that many of the annotations were done sequentially, i.e., the annotations regarding the first event in the text file were placed at the top of the file, followed by the annotations of the second event. However, not all the annotations were ordered, so our solution would have to consider those cases. Secondly, because most sentences with relevant information often start with a participant as a subject, the annotations will reflect that and, in a scene, the first semantic role link to be found will be one of type *agent*, *pivot* or *patient*.

It is also important to remember the nature of the selected news files, as discussed in Section 5.4. The news files that were chosen resemble narratives and provide a storyline flow that gives a strong emphasis to not only events but participants too.

With all this in mind, an algorithm was developed with the intention of grouping relations that characterized the same event.

To prepare the data, the script extracted all the relations that had a *Tag* attribute starting with *SRLINK* and then filtered them to make sure their *Tag* matched the chosen semantic role links, i.e., SRLINK_agent. Next, the script parsed the filtered semantic role links one by one, in the same order as they were extracted, and grouped them into clusters of information that later originate scenes.

To group the relations, a number of key rules were set:

1. *beginning points* are semantic role links that start a given scene by linking information about a participant and an event. In this case, *SRLINK_agent*, *SRLINK_pivot* and *SRLINK_patient*. These semantic role links can also link two events, but in these cases, the part of speech of one of the events will be a noun.

2. For a number of relations to be part of the same group, they must have one tag_ref field with the same value between each sequential pair of relations. An example can be seen in Listing 5.13, taken from Text2Story's Lusa 1 example.

```
1  R37     SRLINK_pivot Arg1:T1 Arg2:T26
2  R38     SRLINK_theme Arg1:T1 Arg2:T2
3  R39     SRLINK_theme Arg1:T2 Arg2:T27
4
5  Caption:
6  T1 - "suspeito"
7  T2 - "tráfico"
8  T26 - "Um homem de 20 anos"
9  T27 - "droga"
```

Listing 5.13: Example of three semantic role links that could be grouped together in a scene.

The script first checks if the SRLINK is a beginning breakpoint. If it is, and if it could be part of the current scene being constructed, the script will evaluate the next SRLINK. If not, the script will check all previous scenes to see if the information can be grouped with any other previously constructed scene (in case there is information out of order). If it is part of any scene, we move on to evaluate the next SRLINK. If not, the program flow continues, the current scene is added to the scenes dictionary and a new scene begins.

The next step involves evaluating whether the scene is empty. If it is not, the same process of checking if a relation is part of the current scene or any previous scene will be repeated. If the scene is empty and the SRLINK is a beginning breakpoint, then the SRLINK will be added to the scene as its first relation.

After processing all the SRLINKs in a file, the script will go over all the possible unordered SRLINKs that could not fit into any scene before them. It will then check if they can fit into any of the scenes that have been identified, following the procedure previously described. After all the information is grouped into scenes, the scenes will be evaluated to see if they pass certain exclusion criteria to ensure the quality of the information:

1. If the scene has only one relation, and that semantic role link has one of their arguments of type *Event* and that event is of type *Reporting*, then it will be considered as a case of a report to a news agency, which is information that is not relevant for the storyline itself.

2. There must only be one agent for each scene. If there is a scene with two semantic role links of type agent, we might be mixing two different events in the same scene.

After the clusters of information were grouped together and those that did not meet the criteria for a scene were removed, the scene text was generated. This was done by going through all groups of semantic role links and, for each one, determining the order of the words involved in each semantic role link. If it was the first relation of a group, the relation would be of type *SRLINK_agent*, *SRLINK_pivot*, or *SRLINK_patient*, and one of its two *tag_ref* attributes would be matched to a participant. The participant is written first in the sentence, followed by the event. The order of subsequent relations is determined by comparing them to the previous relation. Since the following relation must share one *tag_ref* with the previous one, the common *tag_ref* would already be written, so the other *tag_ref* is added to the sentence. This process is repeated until there are no more relations to iterate over and all information groups have been processed.

Ultimately, each approach produces a single JSON file, resulting in two JSON files for each annotated narrative. One of the files will be named *lusa_0* (using Lusa 0 as an example) and the other *lusa_0_extended*. The former corresponds to the file with the sentence-based approach while the latter represents the event-based one.

5.6 Improvements to NewsLines

We made several updates to the NewsLines code to improve its functionality. Firstly, we modified the code to support the new JSON input schema, which can be found in Section B.3, with an example in Section B.4. This involved updating the data processing and visualization modules to handle the new data format. The user must still interact with the dropdown menu to select the annotated source file for the visualization. However, a checkbox has been added below the dropdown to enable switching between the sentence-based and event-based approaches. Additionally, we added a new feature that allows users to view the BRAT source annotations by clicking on a button located at the top of the sidebar. This provides users additional context and information about the data used to generate the scene text. Finally, we made changes to the user interface. These changes include updating the location icon to make it more visible and easier to see, updating the *times* title in the sidebar module to *dates*. Additionally, tooltips were also added to the scene summaries in the storyline, so that the full description of the event can also be seen in the visualization itself, instead of only in the event information module.

5.7 Limitations

A few limiting factors were detected during the development process. On the scene generation topic, the second approach, summary generation from the annotation files, might produce summaries that are not clear to the reader. Given that it focuses on putting essential information together, linguistic connectors are usually left out, making the produced scene text less clear to read by a user.

Another limitation is that this information extraction technique was designed for a specific type of news story with characteristics that resemble those of a narrative. Therefore, expanding it

to fit other news genres could be difficult, limiting its use. Additionally, many of the source news files are short and focus the main events of the narrative in the first two sentences. From there, the news body expands on the information already stated. This factor makes it difficult to showcase the full potential of NewsLines, as the storylines generated from these files will be, more often than not, short and with few events and participants.

Lastly, the annotation files themselves were found to be incomplete at times, which can result in incorrect or incomplete scenes being generated.

5.8   Summary

This chapter describes the methodology used in this study to integrate the BRAT annotated data from Text2Story into Mariana Costa's NewsLines tool. The approach involved converting the Text2Story files into a JSON format, maintaining all the narrative elements and their relations, and automating the production of such files. The NewsLines source code was also modified to enable the integration of the JSON document, and improvements were made to the tool.

Chapter 6

System Demonstration

After learning about how Text2Story data is integrated into NewsLines, we will demonstrate the advantages and limitations of the pipeline developed in this research work by analyzing two examples, Lusa 112 and Lusa 120. Section 6.1 gives us an overview of the demonstration. Section 6.2 and Section 6.3 analyse the chosen examples. Finally, Section 6.4 goes over the advantages, limitations, and possible adaptations of the system.

6.1 Overview

To explore how well our pipeline works, we analyzed two files in-depth, Lusa 112 and Lusa 120. These files were chosen because they showcase different positive aspects of the framework, as well as different negative aspects. A live online version is also available[1] for real-time exploring and interacting and the source code [28] is also available on GitHub.

6.2 Use Case - Lusa 112

For the first use case, we analyzed the representation of file Lusa 112. Overall, this news story regards changes in the government. Ricardo Pinheiro, a socialist deputy, will replace José Mendes as Secretary of State for Planning. The President of the Republic of Portugal has accepted the Prime Minister's proposal to dismiss, at his request, José Mendes as Secretary of State for Planning and the nomination of Ricardo Pinheiro for the same role. The text and annotation files can be found in Appendix A.3 and A.4, and the sentence-based and event-based JSON files in Appendix B.5 and Appendix B.6 respectively. This narrative contains six participants, and all of them are active actors in the events: Ricardo Pinheiro (a socialist deputy), José Mendes, the Secretary of State to the Minister for Planning, the President of the Republic, the Prime Minister, and the City Council of Campo Maior. This means that the participants are linked to events through relations such as

---

[1]https://catarina03.github.io/news-story-viz-prototype/

the semantic role link *SRLINK_agent*, *SRLINK_patient*, and *SRLINK_pivot*. As a result, the event-based visualization shows the same number of participants as the sentence-based one. Figure 6.1 shows the sentence-based visualization, and Figure 6.2 shows the event-based visualization.
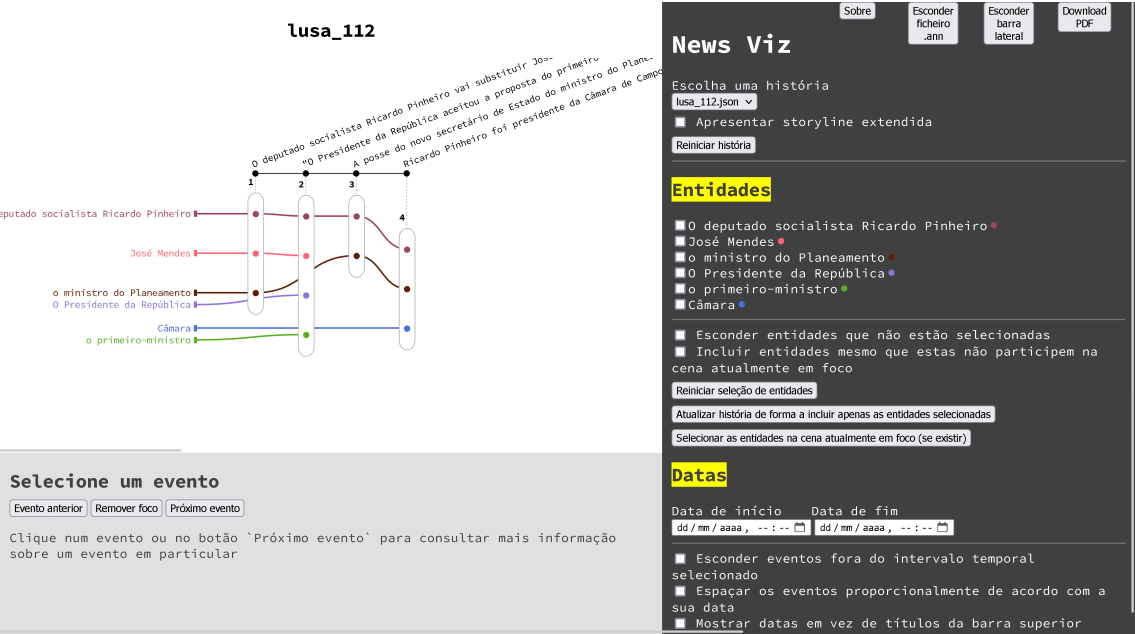
Figure 6.1: The sentence-based visualization of the Lusa 112 file.
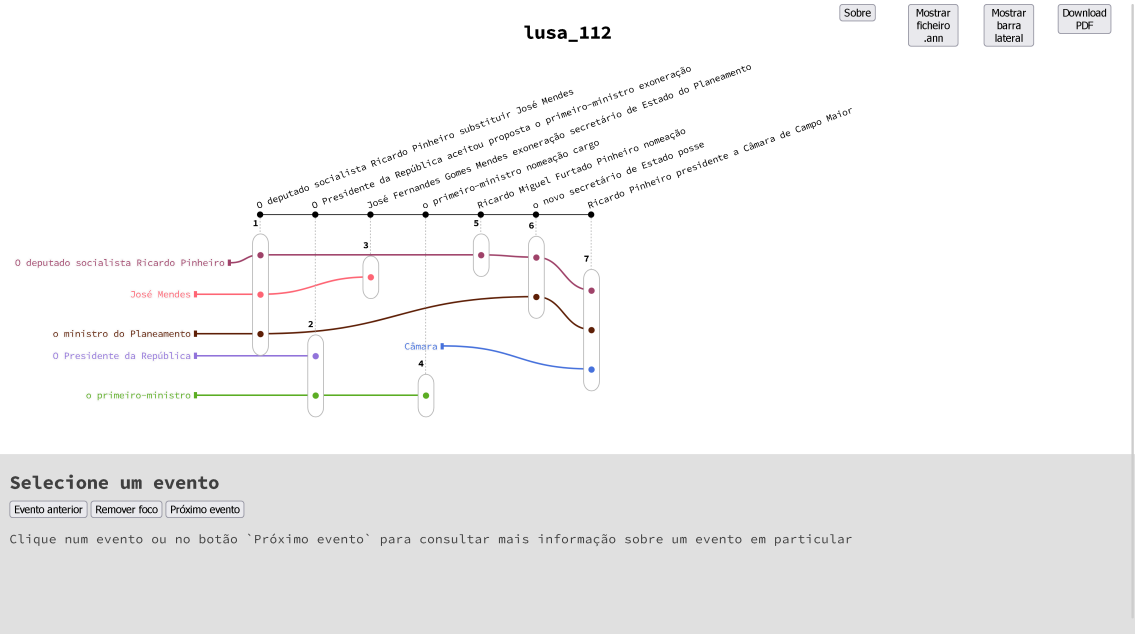
Figure 6.2: The event-based visualization of the Lusa 112 file.

As seen in Figure 6.2, the main points of the narrative are present in the visualization. Firstly, we learn about the main event, the replacement of José Mendes for Ricardo Pinheiro, a socialist deputee. After we learn additional information, we see that the President of the Republic accepted

the exoneration proposal. The next scene gives us more information on the exoneration. We learn that José Mendes was exonerated from the Secretary of State for Planning role. The following scene tells us that the Prime Minister nominated someone for the role, and the scene after clarifies that this nomination from the Prime Minister is for Ricardo Pinheiro. After, we learn that the new Secretary of State to the Minister for Planning will take office. Lastly, we learn about the previous role of Ricardo Pinheiro, who was President of the Campo Maior City Council. From this, we can conclude that, despite the lack of articles in the scene text at times, we can still follow the narrative without much difficulty.
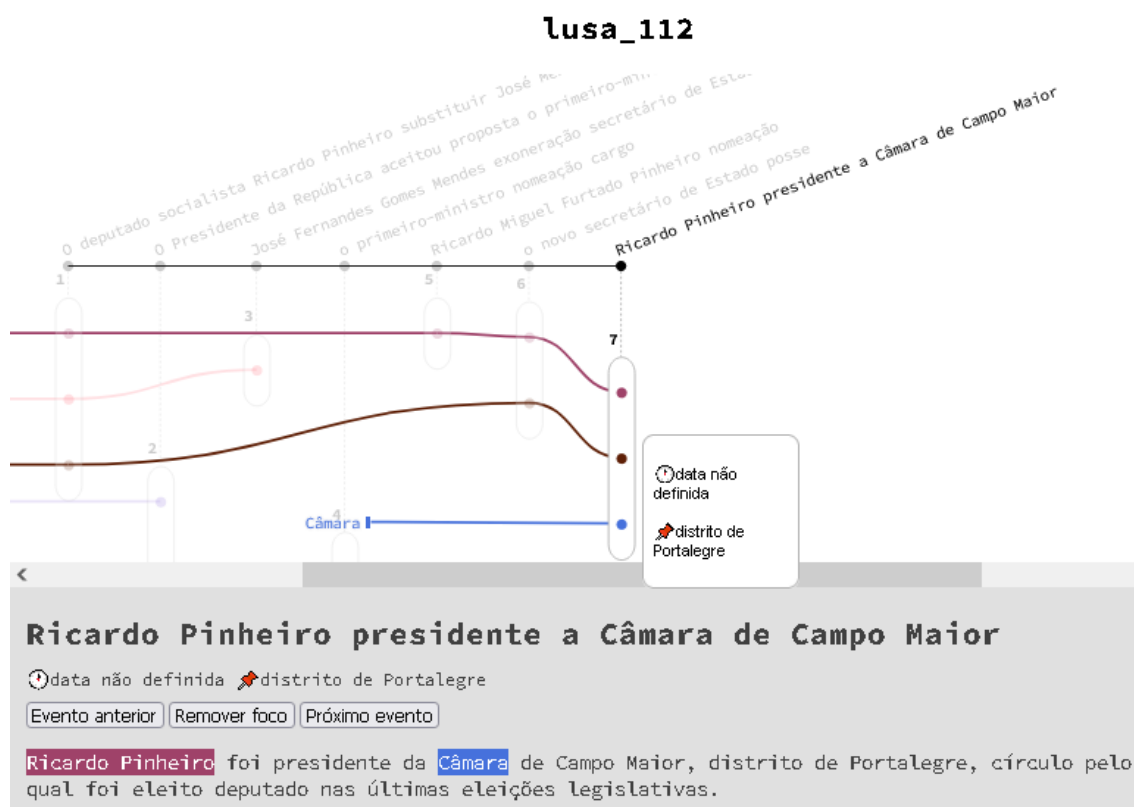
Figure 6.3: The event-based visualization of the Lusa 112 file, having the last scene selected.

Additionally, we also note that there is a lack of spatial and temporal information. Regarding spatial information, there are two mentions in the original text. One is in the last scene, indicating the former role of Ricardo Pinheiro, and can be seen in the tooltip of the event and in the event information module, as seen in Figure 6.3. The other one is in the scene where we learn that the Secretary of State for Planning will take office. The original sentence mentions that the ceremony will happen in Palácio de Belém (the official residence of the President of the Republic). However, this information is not represented in the visualization, as seen in Figure 6.4. The reason for that is that Palácio de Belém, a participant entity with location as its participant tyle domain, is linked through a semantic role link location to the event *terá lugar* (*will take place*, in English), and not to the event represented, *posse* (*take office*, in English). Listing 6.1 shows part of the annotations
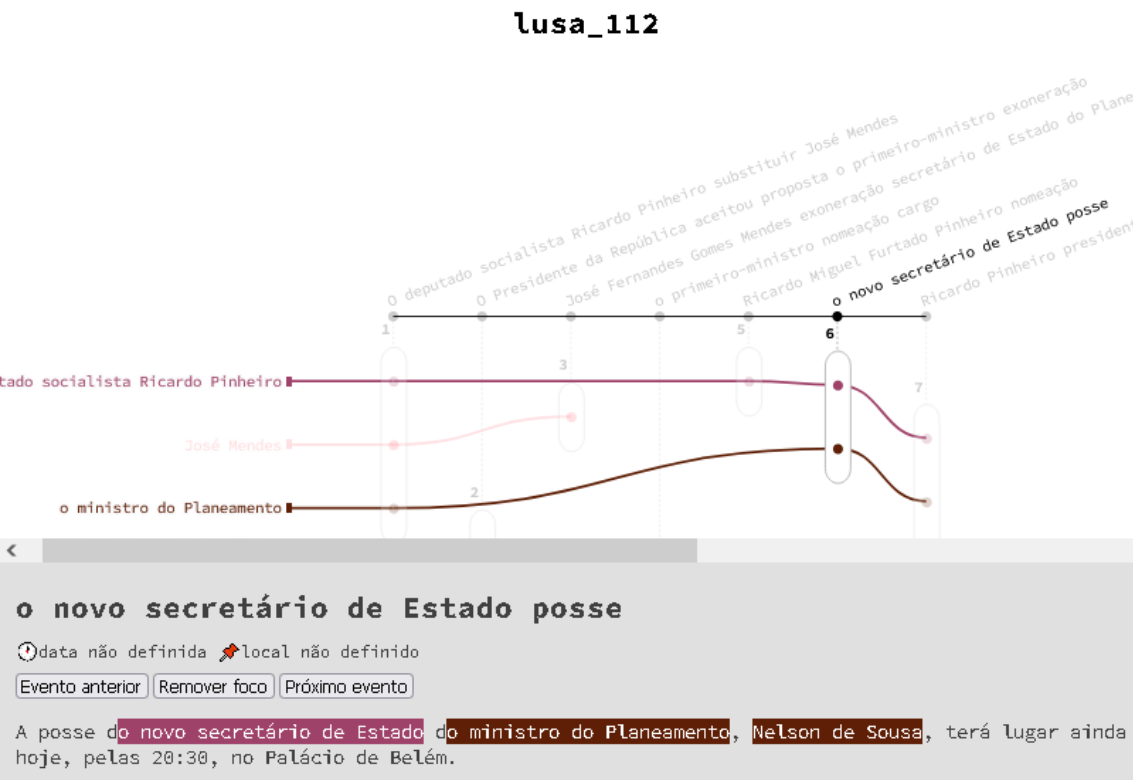
lusa_112

Figure 6.4: The event-based visualization of the Lusa 112 file, showing the lack of spatial information.

that originate this scene.

```
1  T15 Event 593 598 posse
2  T16 Event 673 683 terá lugar
3  T33 Participant 600 627 o novo secretário de Estado
4  T36 Participant 710 729 o Palácio de Belém.
5  R56 SRLINK_patient Arg1:T15 Arg2:T33
6  R58 SRLINK_location Arg1:T16 Arg2:T36
```

Listing 6.1: Annotations for a scene missing location in the file Lusa 112.

Besides this, we can see in Figure 6.4 that the only mention of time is also not displayed. This happens for the same reason as before: the temporal information is linked to the event *terá lugar*, and since that specific event entity is not included in the scene, the time associated with it does not appear in the visualization. The relevant annotations that refer to this issue are shown in Listing 6.2.

```
1  T16 Event 673 683 terá lugar
2  T17 Time 690 694  hoje
3  T18 Time 702 707  20:30
4  R16 TLINK_isIncluded Arg1:T16 Arg2:T17
5  R17 TLINK_isIncluded Arg1:T18 Arg2:T17
```

Listing 6.2: Annotations for a scene missing temporal information in the file Lusa 112.

6.3 Use Case - Lusa 120

For the second use case, we chose to analyze the Lusa 120 file. This news story is about a boy who was in a severe state after being shot near Amadora. The seventeen-year-old boy was shot at dawn in the Casal da Mira neighborhood in Amadora, in the Lisbon district. The Lisbon Metropolitan Command (Cometlis) stated that the boy was severely injured and was taken to Professor Doutor Fernando Fonseca Hospital. Cometlis stated that the causes of the incident are unknown, and no suspects have yet been identified or arrested. The case was handed over to the Judicial Police (PJ). The complete text and annotation files can be found in Appendix A.5 and Appendix A.6, and the sentence-based and event-based JSON files in Appendix B.7 and B.8 respectively.
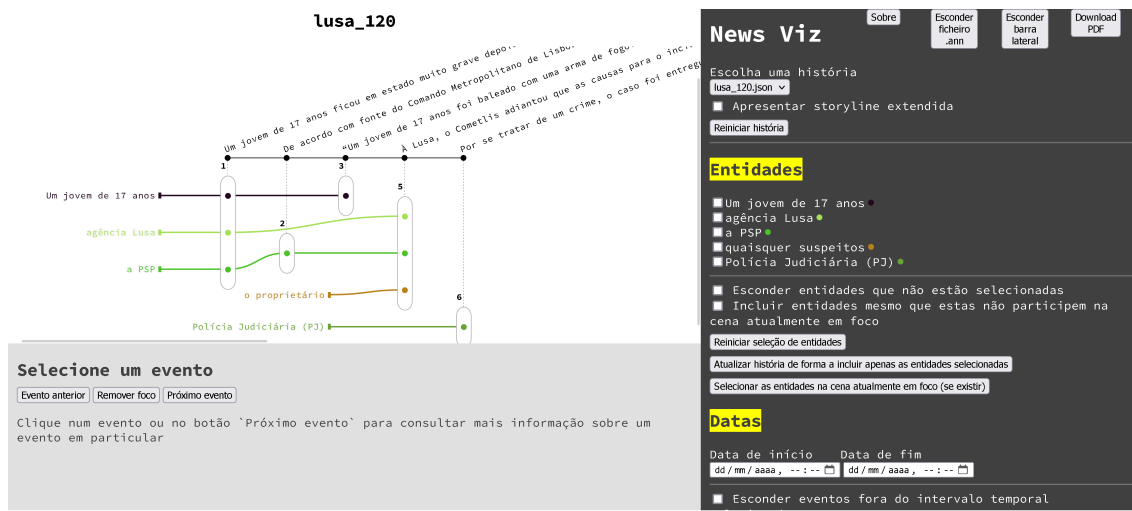
Figure 6.5: The sentence-based visualization of the Lusa 120 file.

In the sentence-based visualization, there are five participating entities (the boy, the Lusa news agency, PSP/Cometlis, the suspects, and the *Polícia Judiciária (PJ)*), while in the event-based visualization, there are only four. In the latter case, the entity of *Polícia Judiciária (PJ)* disappears, as seen in Figure 6.5 and Figure 6.6. This is because of the semantic roles that link *Polícia Judiciária (PJ)* to the event. As seen in Listing 6.3, the only event that includes *Polícia Judiciária (PJ)* does not match the criteria of the scene generation algorithm, discussed in Section 5.5.5. In this case, the first semantic role link (*SRLINK_theme*) does not belong to the list of possible
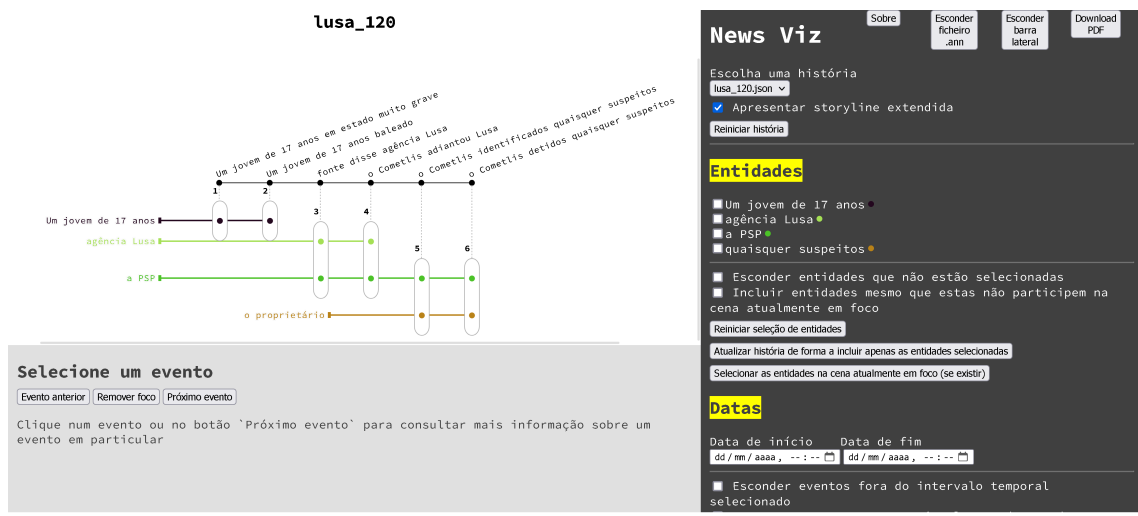
Figure 6.6: The event-based visualization of the Lusa 120 file.

semantic role links for the beginning point of a scene, so the algorithm will try to match it to previous scenes but not begin a new one.

```
1  T22 Event 786 790 caso
2  A102  Pos T22 Noun
3  T23 Event 795 803 entregue
4  T24 Event 835 847 investigação
5  T41 Participant 806 829 Polícia Judiciária (PJ)
6  R72 SRLINK_theme Arg1:T23 Arg2:T22
7  R73 SRLINK_goal Arg1:T23 Arg2:T41
8  R74 SRLINK_purpose Arg1:T23 Arg2:T24
```

Listing 6.3: Annotations of *Polícia Judiciária (PJ)* in the file Lusa 120.

In this case, however, understanding the full news article from the visualization is more challenging. As seen in Figure 6.6, the first scene states that a seventeen-year-old boy is in a severe condition, the second complements this idea by revealing that the boy has been shot, and the third scene states that the source communicated (the information of the previous two scenes) to the Lusa news agency. The next scene says the Lisbon Metropolitan Command (Cometlis) told Lusa (the information in the following scenes). The next scene just puts together the information that the police identified any suspects. While it is possible to understand the meaning through context and the text presented, it is more challenging due to the lack of words that connect the sentence and emphasize that no suspects were identified. The same happens in the last scene, changing only the event from *identified* to *detected*. Additionally, besides the missing sentence fragments making it more challenging to interpret the meaning of some scenes, there is also information not represented in any scenes. One example of this was the one previously discussed, of *Polícia Judiciária (PJ)* not being present in the visualization.

Regarding temporal and spatial information of the sentences represented in the visualization,
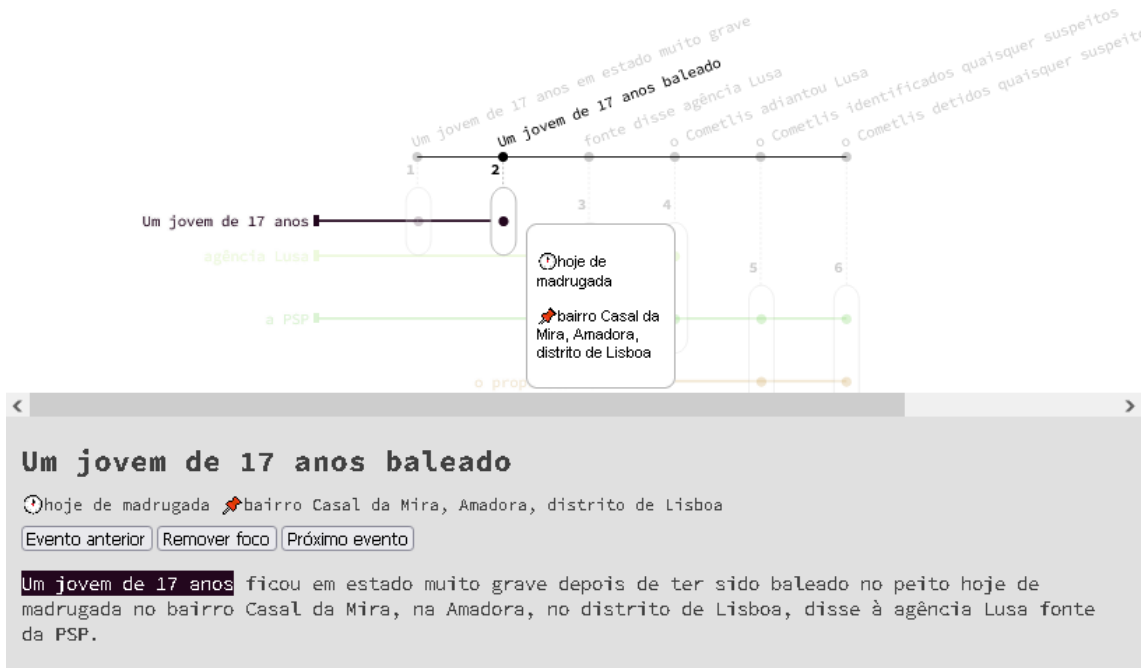
Figure 6.7: The event-based visualization of the Lusa 120 file, having the second scene selected.

all the times and locations mentioned are represented in a scene. In this particular case, the information is represented in the second scene, as it is linked to the event *shot*, as shown in Figure 6.7.

6.4 Discussion

As seen in the first use case, in Section 6.2, this tool works well with participant-focused narratives. If the events of the news narrative are mainly driven by the actions of people or organizations, the algorithm will better detect and represent those events. This approach fits well with NewsLines, as its visual representation also focuses on the participating entities. Besides this, the temporal and spatial information linked to a certain event is represented in an intuitive way in the specific scene that contains that event. Moreover, the BRAT source code sidebar allows a user to see the source code side to side with the final result, as seen in Figure 6.8. While this might not be useful to the general public, it is a relevant feature for someone who works with annotations or for further development and debugging of the tool. Lastly, due to the way the scene generation algorithm reads a BRAT annotated file, files that organize their content sequentially will be displayed by our tool in a way that is better understood.

One of the negative aspects is that, as explained in Section 6.3, it is possible that information does not get displayed because the corresponding annotation does not match the criteria needed by our algorithm. To our knowledge, this frequently happens when the first semantic role link relation in a scene is of type *SRLINK_theme*. This relation is similar to the *SRLINK_patient*. It differs only by the fact that if a participant is referenced in the relation, it is not structurally changed or affected
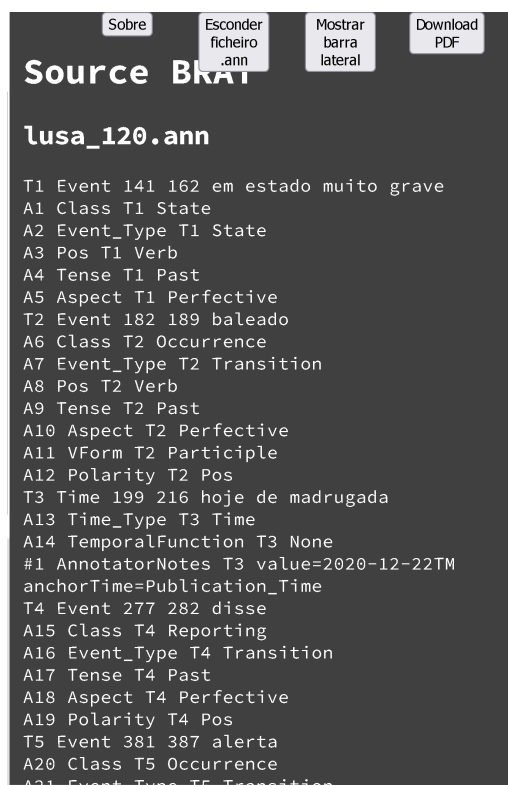
Figure 6.8: The source BRAT sidebar of the Lusa 120 file visualization.

by the event also referenced in the relation. Consequently, this would also be a viable relation to include in the beginning points list. However, given that this relation type appears more frequently throughout the news annotations, we could not find a criteria to distinguish both cases during the development stage.

Another negative aspect is that, given the lack of frequent date information, the date section of the sidebar (seen in Figure 6.9) is often unable to serve its purpose. In this dataset, since the news available are short and mostly talk about one main event, it is difficult to have a file where the filtering by date option is possible.
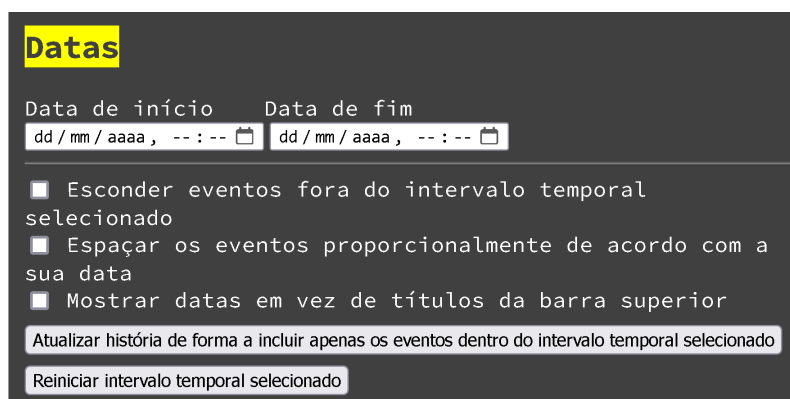
Figure 6.9: Date section of the NewsLines sidebar.

Lastly, given that the news narratives of the dataset are short, the usual format of the news narrative follows a simple structure: firstly, the main event is introduced; secondly, the same event is referred to again, this time in more detail and with additional information, and lastly, the outcome of the event is stated. Having this structure, it is natural that some information is repeated as the storyline goes forward.

In the end, despite the limitations of this pipeline, automating JSON input file generation opens up a world of possibilities in a large-scale scope. It allows for automating the production of news narrative visualizations from the moment we have the news story text file and the BRAT file with the annotated news narrative. While this research work focuses on NewsLines, the visualization can be adapted to represent the narratives in different visual styles. Additionally, if we wished to represent narratives of other genres, that would also be possible. At most, one type of change would have to occur. For a narrative that was not participant-focused, it could make sense to change the algorithm for scene generation (and the visualization itself), as it would cease to be relevant to rely on semantic role links that are heavily connected to the participant entities to decide where scenes begin. It is worth noting that we are relying on the assumption that the BRAT annotation standard will remain the same throughout the different narratives.

6.5 Summary

This chapter demonstrates how our tool functions by reviewing two examples of narrative visualizations (Lusa 112 and Lusa 120) and their corresponding advantages and limitations. Additionally, we discuss how this system can be adapted to other narrative styles and visualizations.

Chapter 7

Evaluation

The evaluation of our proposal was conducted through two methods: a qualitative analysis based on semi-structured interviews with eight researchers involved in the Text2Story project, as seen in Section 7.1, and a primarily quantitative analysis using a survey released to five Faculties of the University of Porto, gathering a total of 65 responses, as described in Section 7.2. The main objectives of the evaluation were to assess the usefulness of the NewsLines prototype with the automatically generated JSON files as input, as well as to gather feedback and suggestions for improvement, or even ideas for new visualizations.

7.1 Interviews

To test and gather feedback to guide further development, user interviews were conducted with the Text2Story team. The interviews consisted of semi-structured talks with eight researchers from the Faculty of Sciences, Arts and Humanities, and Engineering, with varied education levels. The participants pool was comprised of five men and three women. The questions used as guidelines for the interviews can be found in Table 7.1 and the original interview script can be found in Appendix D, Section D.1.

The first section (from Q1 to Q4) was mandatory and asked the participants to name which narrative elements should be present in a visual representation of a news story, how much focus each element should have, and how each element related to the others. Additionally, participants were asked to describe how they imagined the visual representation. The second section (Q5) consisted of sketching a prototype of a visual representation, considering the answers given in the previous section. Finally, the third section (from Q6 to Q24) showed three visualization prototypes: the sentence-based and the event-based adaptations of NewsLines [22], and a sketch of a graph-based visualization. For each prototype, a series of qualitative questions were asked to gather feedback on the positive aspects and areas of improvement.

	Question
	Section 1
Q1	What elements must be present in a visual representation of a news narrative?
Q2	Rate the elements above using the scale *essential, important, useful, neutral, insignificant*, where an essential element should be given more focus and an insignificant element should be given less focus.
Q3	Describe how these elements are related visually using concepts such as *part of*, *composed of*, and *associated with*. For example, a character could be associated with an event, and an event could be made up of several characters.
Q4	Describe how you imagine the visual representation of a news narrative.
	Section 2
Q5	Sketch a prototype, taking into account the elements previously mentioned and how they relate to each other.
	Section 3
Q6, Q12, Q18	What are the positive aspects of this visualization?
Q7, Q13, Q19	What aspects could be improved?
Q8, Q14, Q20	What are the negative aspects?
Q9, Q15, Q21	Are there elements that should be given more focus? Which ones? Why?
Q10, Q16, Q22	Are there elements that should be given less focus? Which ones? Why?
Q11, Q17, Q23	Is there any missing information you feel is important?

Table 7.1: Questions used as guidelines for the semi-structured interviews.

(a) Rating narrative elements by importance, normalized. (b) Rating narrative elements by importance, not normalized.
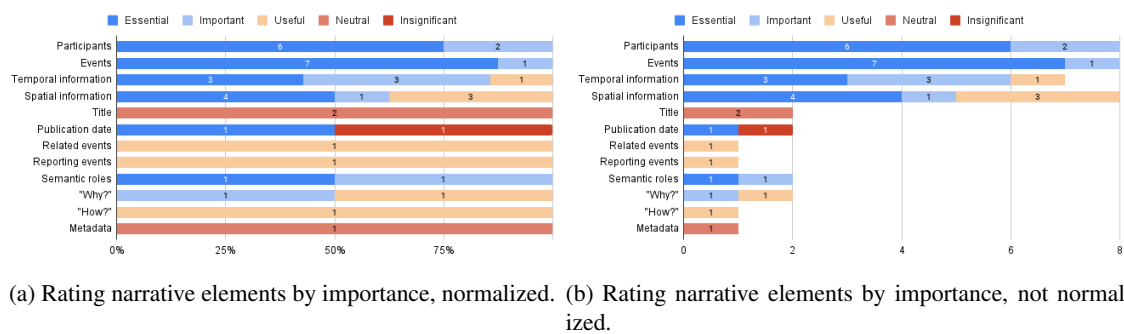
Figure 7.1: Rating narrative elements by importance.

7.1.1 Characterizing the Visualization

The first two sections of the interviews were meant to help us understand the characteristics of a useful visual representation of a news story. Firstly, we needed to understand which elements should be present in the visualization and which ones should have more focus.

In total, as seen in Figures 7.1a and 7.1b, we found that the elements that had the biggest positive rating were, by descending order, Events, Participants, Semantic roles, Temporal information, and Spatial information. All of these elements count with no ratings lesser than useful. Contrary to the others, Semantic Roles were only mentioned twice by the interviewees but obtained a rating of important and essential. The small number of mentions can possibly be attributed to the specificity level of the element, as semantic roles are very prevalent in the annotations but might not come up to people who do not work closely with them. Title and Metadata, when mentioned, are not considered important.

When comparing element classifications between STEM study fields (5 participants) and non-STEM ones (3 participants), as demonstrated by Figure 7.2a and Figure 7.2b, the most striking difference is the feeling towards publication date, which in the STEM group has the lowest classification and in the non-STEM the highest. This can perhaps be due to how the non-STEM respondents are responsible for the annotation process. During the process, the publication date is used as a reference point to calculate other dates. STEM participants were also the only ones mentioning Semantic Roles. However, non-STEM participants mentioned both the questions *Why?* and *How?*, which can be answered by the semantic roles *reason, purpose*, and *instrument, means* respectively. Non-STEM users also gave much emphasis on the importance of answering the *What?*, *Who?*, *When?*, and *Where?* journalistic questions, which might be the reason why the Participants, Events, Temporal information, and Spatial information had a higher percentage of the essential rating.

When asked to describe how elements relate to each other visually (Q3), four users said participants should be associated with events, two said events should be composed of participants, and another said that participants should be related to each other through events. Additionally, one said that the association of participants and events should be made through semantic role links. Regarding spatial information, three users associate events with places, and another three said

(a) STEM ratings of the narrative elements.

(b) Non-STEM ratings of the narrative elements.
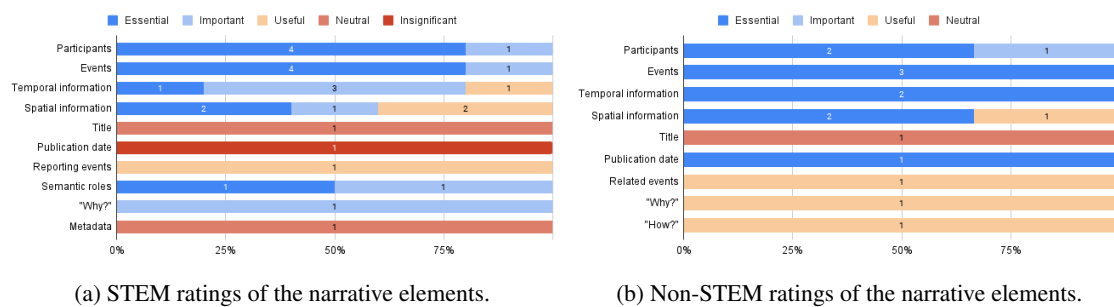
Figure 7.2: Rating narrative elements by importance according to study field.

events should be composed of places. Another said that a participant belonged to a place. One participant said events can be related to other events, and another two spoke of events as being able to contain other events. Temporally, events should be composed of a time, according to three users. Another user stated that many events can be in the same temporal space. Two participants stated that events should be indexed by time, and one suggested the use of TLINKS for that matter. Lastly, one participant mentioned that metadata should be linked to the news story. In conclusion, most respondents agree that events and participants should be linked, the same applies to events and spatial information and events and temporal information. The other pairs of relations were mentioned less. Figure 7.3 summarizes these findings with a top four of the most mentioned relations.
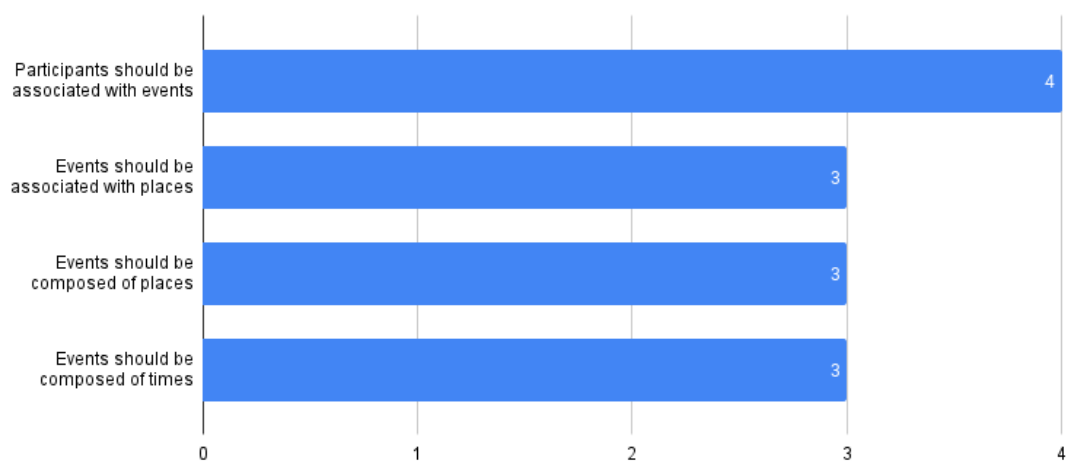
Figure 7.3: Top four of mentioned relations between narrative elements (Q3).

As for the description of the visual representation (Q4), five participants mentioned temporal ordering and timelines, one of which presented two possible orderings: chronological and narrative. Another participant described a representation where events were as boxes in the timeline, and on hover, participant and location information would appear. One participant stated that while a timeline representation works well for events that follow a time sequence, in some news narratives, the same event is repeated throughout the story, and in those cases, a bubble graph
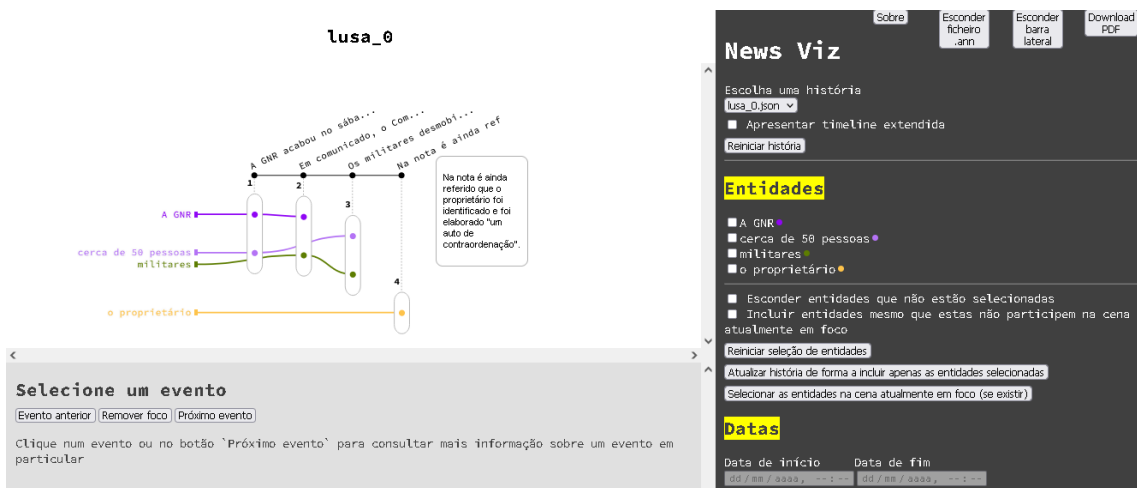
Figure 7.4: The interface of the sentence-based adaptation of NewsLines, with no selected scenes.

representation could better depict the simultaneity relations. Another two mentioned graph representations using the central word of an event as a node. The rest of the information would be linked to it, using semantic roles as edges. One participant mentioned that it would be interesting to have a representation with a set of images related to each other. Another user gave three examples of different visualization types: a 4x4 comic strip with the four key moments, a diagram with participants, event lines and timestamps, and an infographics map. Finally, one user stated the importance of understanding the needs of the reader and accounting for that in the visualization.

The visualization sketches (Q4) can be found in Appendix D.2.

7.1.2 Evaluating Prototypes

The last section showcased three visualization prototypes. The first two are extensions of NewsLines, one displaying the storyline with whole sentences and another with a summary of each event. The last prototype was a sketch of a graph-based visualization. For each visualization, there were six questions, as seen in section three of Table 7.1, guiding the interview that focused on understanding the positive aspects of each visualization, the areas of improvement, if there were elements to be given more or less focus, and if there was information missing.

7.1.2.1 The Sentence-based Adaptation of NewsLines

For this prototype, interviewees were given access to its online version, allowing them to interact with it freely. Figures 7.4 and 7.5 show the interface of the online version, at the time. Before diving into the question, a brief introduction of the prototype's features was made, describing the three modules of the interface and going over the possible interactions. The participants were then asked to evaluate the tool and suggest how to improve it. The interviews were conducted in the development stage when participant, event, and spatial information were already integrated into the prototype, using the news story Lusa 0, introduced in Section 3.2, as an example.
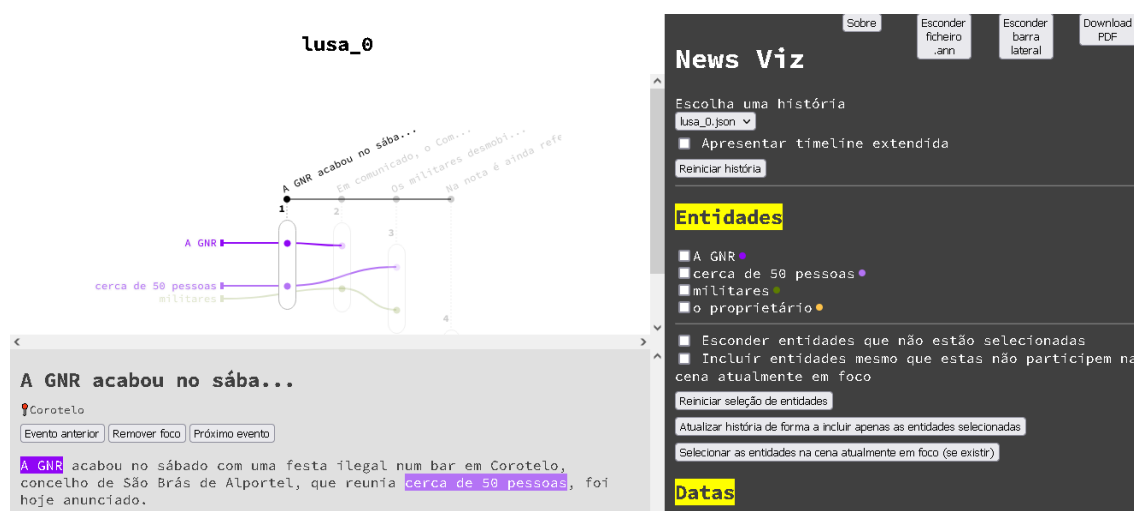
Figure 7.5: The interface of the sentence-based adaptation of NewsLines, with scene one selected.

As for positive feedback, six participants found the visualization an intuitive way of showing the relations between an entity and an event. Three found it clear, and another three found the storyline representation positive. One participant found the different colors of the actors to be a good aspect. Another considered the sentence-by-sentence analysis and being able to read the full news story from there a positive experience. Finally, one participant found the interactive sidebar to have interesting options.

Regarding the negative points, five people noted that the temporal information was missing. Two stated they could not extract the information on the main events from the visualization. Another two mentioned that long sentences get cut from the visualization. Two participants found it confusing when character lines would pass behind the character aggregator. One person said it was confusing having the text above the storyline. Lastly, one participant noted that two levels of the story are mixed up: the main narrative and the reporting events.

Finally, participants made suggestions for improvement. One suggested having temporal indexing so a user can better understand the whole story. Regarding the character aggregator, two participants mentioned that the sentence could appear on hover like the spatial information instead of above the black line, and another suggested changing the aggregator outline thickness based on the duration of the event. One would find it interesting if important participants appeared above. Another person proposed having the most important elements appear in bold in the tooltip of the sentence. Lastly, one participant advised us to focus on the most important information, which is mentioned in the first two sentences of this type of news.

Table 7.2 summarizes the findings through a thematic analysis [55, 13]. In this case, codes are words or short phrases that assign an essence-capturing attribute a section or portion of a participant's response. They are used to partially reduce qualitative data and identify key areas for further analysis. Examples depict spoken responses from participants that fit the theme of the code.

Codes	Examples
Positive Feedback	
Participant/event relations	*It's intuitive to know which events each participant takes part in.*
Clear	*Clean, tidy, and attractive.*
Sequencing in-line	*Shows time sequence, the way people are used to.*
Color-coding	*Different colors to mark different things is a good thing.*
Analysis by sentence	*You can analyze the story by looking at each of the phrases.*
Customizable focus	*When we select it, only what we want to see becomes highlighted.*
Displays whole news story	*It's good to have the news itself.*
Interactivity	*Interesting side buttons.*
Negative feedback	
Confusing	*The text above the timeline points gets confusing.*
Character lines	*The line passing behind the aggregator is not a very explicit way of showing that the character does not participate.*
Cut sentences	*The beginning of sentences alone doesn't give much information.*
Mixed story levels	*The highlighted story and sources, which are not part of the story itself, should be on separate levels.*
Missing temporal information	*You can't see the time information.*
Lacks events	*We can't get that information [events] from the visualization just by looking at it.*
Suggestions	
Temporal indexation	*If we could have temporal indexing, we could understand the whole story.*
Text in aggregator	*It would make more sense for the text to appear when you hover over the character aggregator ball.*
Aggregator thickness	*The little ball surrounding the characters could have varying depending on the duration of the event.*
Participant hierarchy	*Important participants could appear above.*
Bold	*The most important elements could appear in bold in the tooltip of the sentence that appears when you hover over the event.*
Avoid Sentence Analysis	*Do not analyze the news using the sentence criteria.*
Capture the most important	*Capture what is most important, which is in the first two sentences, what follows is less important as it resumes the important information that was said at the beginning.*

Table 7.2: Thematic analysis of the feedback of the sentence-based NewsLines adaptation.
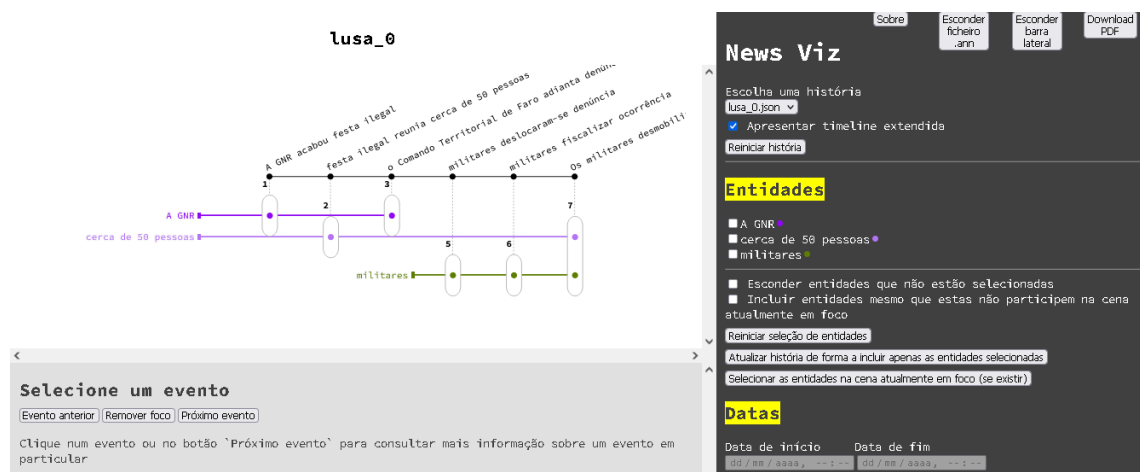
Figure 7.6: The interface of the event-based adaptation of NewsLines, with no selected scenes.

7.1.2.2 The Event-based Adaptation of NewsLines

This prototype was introduced as an extension of the previous one. The majority of the elements stayed the same, the only difference being that the text of each scene above the black line was no longer whole sentences but event summaries generated from the annotations. Participants had to check the *Extend visualization* checkbox to access this prototype. Figures 7.6 and 7.7 show the prototype interface at the time.

Regarding positive feedback, four participants agreed that providing only an event summary for each storyline scene was enough to obtain the information needed. Three participants stated that this visualization was easier to understand. Another two participants found it easy to visualize character relations. One said the event-based approach gave the visualization a cleaner look. Another mentioned that the linearity of the story was better understood. Finally, we got the same feedback regarding interactivity, customizable focus, and displaying the whole news story as the previous visualization.

As for negative points, some feedback was repeated, namely the one regarding the lack of temporal information, mixed story levels, character lines, and cut sentences. Additionally, two participants mentioned the need to include semantic roles, as the event information can become confusing without it. Another mentioned that the location information needed to be presented more consistently. One user found the *Extend visualization* checkbox difficult to find. Lastly, one participant noticed one of the visualization participants was not being displayed.

Finally, some suggestions were made. Four participants advised us to highlight the important elements in the text, two of them suggesting putting them in bold. One participant mentioned it would be interesting to be able to show the elements in chronological order. Lastly, another participant suggested swapping the scene numbers with the event verb.

Table 7.3 summarizes these findings in a thematic analysis.
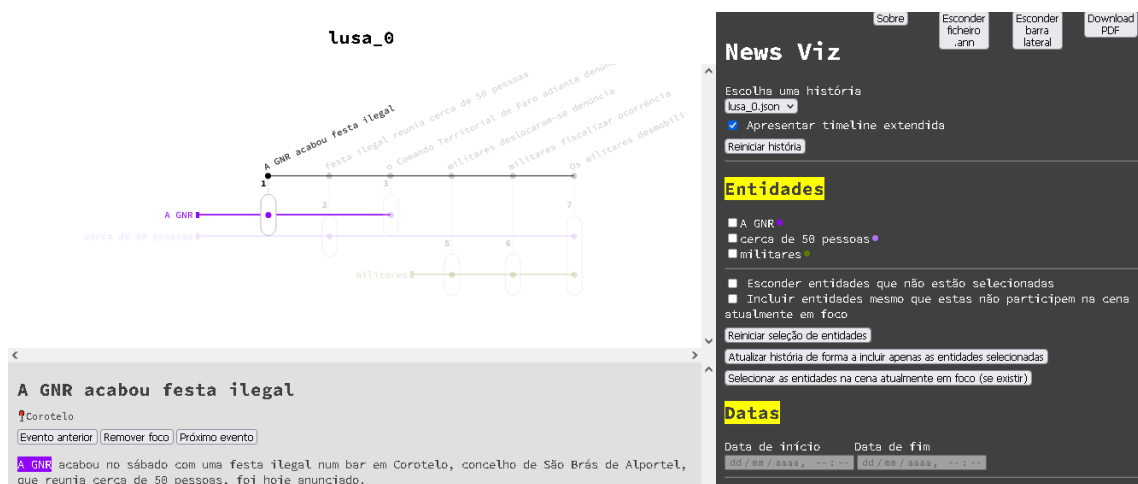
Figure 7.7: The interface of the events-based adaptation of NewsLines, with scene one selected.

Codes	Examples
Positive Feedback	
Easy to visualize	*It's easy to visualize the characters and allows you to see the relationships well.*
Clean	*It gives a cleaner look.*
Linearity	*The linearity of things is better understood.*
Easier to understand	*This [visualization] helps to better understand the news.*
Displays whole news story	*It's good to have the news itself.*
Customizable focus	*When we select it, only what we want to see becomes highlighted.*
Summary	*What is in black is enough to obtain information.*
Interactivity	*Interesting side buttons.*
Negative feedback	
Character lines	*The character lines passing behind are confusing.*
Semantic roles	*Visual information becomes more confusing than the other because the semantic relationships are missing.*
Inconsistent location	*The locations must be uniform at all points.*
Cut sentences	*Words should not be cut off.*
Temporal indexation	*Not temporally indexed.*
Mixed story levels	*The two levels continue to appear.*
Missing temporal information	*Time information is missing.*
Checkbox	*The checkbox took a while to find.*
Missing participants	*One of the participants disappeared in this visualization.*
Suggestions	
Sorting	*It would be interesting to sort chronologically.*
Highlight important elements	*The most important elements could appear in bold in the tooltip of the sentence that appears when you hover over the event.*
Numbers	*Swap scene numbers by the event verb.*

Table 7.3: Thematic analysis of the feedback of the event-based NewsLines adaptation.

(a) Graph-based visualization, no event in focus selected.

(b) Graph-based visualization, scene one selected.

Figure 7.8: Graph-based visualization interface.
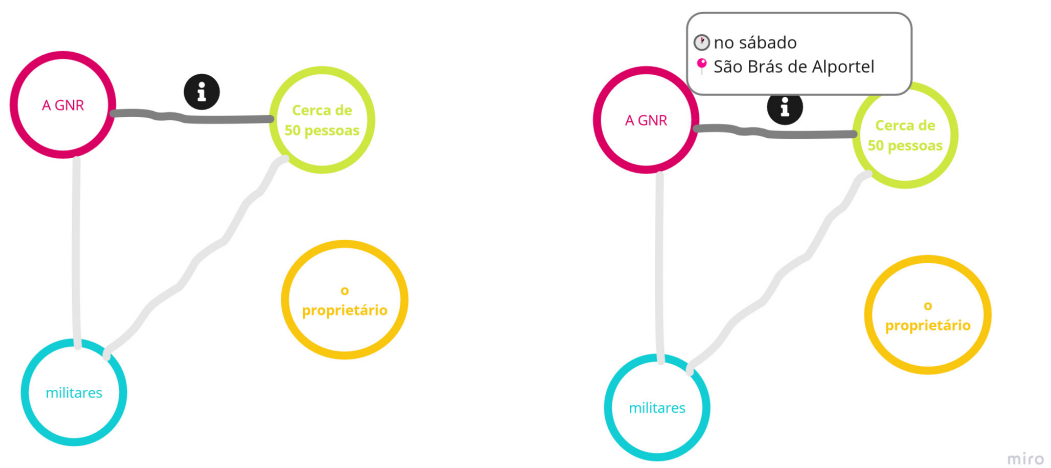
Figure 7.9: Graph-based visualization, scene information on hover.

7.1.2.3   The Graph-based Visualization

The third prototype consisted of three sketches of a graph visualization focused on the participants, as seen in Figures 7.8a, 7.8b, and 7.9. Each sketch represented a different state in the interaction: no events in focus, one event in focus, and one event in focus and hovering on the edge.

Regarding positive feedback, three participants stated that the relations shown in the graph were very clear. Another participant considered that having different colors on the scheme was positive. Another user felt there was no information overload, as additional information would only appear on hover. Another user liked that we can see if there are no explicit relations between one participant and the rest. One user found the visualization compact. Finally, one participant pointed out that highlighting the participants and their relations with each other was positive.

As for the negative aspects, two participants found the visualization insufficient to understand the news story thoroughly. Another found it strange that entities could appear alone. Additionally, one user considered the grey lines in the visualization irrelevant.

Lastly, the participants gave four suggestions. Six participants advised us to add the events to clarify the relations. Two participants proposed that the temporal information be added, one of which suggested we use the publication date as a reference. One participant added that using a bolder font would make the visualization more understandable. Lastly, another user suggested we swap the participants with the events to make it an event-centric visualization.

Table 7.4 contains a brief thematic analysis of the responses gathered.

7.2   Survey

After the semi-structured interviews provided us with deeply explored and discussed responses, we also decided to release a survey to evaluate the final prototype. In order to reach as many people as possible, we designed the survey to be attractive to fill out. This meant making it short and mainly having close-ended questions. The result was a survey with fifteen close-ended questions and one open-ended question.

The survey was sent by email to the students of five faculties of the University of Porto: Faculty of Engineering, Faculty of Nutrition and Food Sciences, Faculty of Medicine, and Faculty of Sciences and School of Medicine and Biomedical Sciences. These faculties were chosen because they were the ones that allowed access to their email system. The email system only allowed the emails to be sent to students. Only those enrolled in the 2022/23 school year were contacted. In total, 65 responses were obtained out of the 26642 emails sent.

The survey questions divide themselves into three categories: demographic information (Q1 to Q3), online news reading habits (Q4 to Q9), and prototype evaluation (Q10 to Q16). The list of survey questions can be found in Table 7.5 and the complete form in Appendix E.

The first question group, demographic information, allowed us to trace a participant profile based on age, affiliation, and occupation. The second question group, online news reading habits,

Codes	Examples
Positive Feedback	
Color-coding	*The scheme itself having different colors is good.*
Clear relations	*Graphs show relations more perceptively.*
No information overflow	*Information on hover so there's no information overflow.*
Easier to understand	*It's easier and more intuitive to understand the relationships between participants.*
Shows lack of relations	*Shows if there is no explicit relation between one participant and others.*
Compact	*It's compact, closer to a knowledge graph.*
Highlights participants	*Highlights the participants and how they relate to each other.*
Negative feedback	
Entities alone	*It's strange for an entity to appear alone.*
Insufficient	*You don't know what happened just by looking at the visualization.*
Line relevancy	*Light gray lines not very relevant.*
Suggestions	
Swap participants with events	*This visualization for events would be interesting.*
Bold	*Put everything in a bolder font.*
Add events	*The relation itself is not made clear, the event should be added on the lines or in the bubble.*
Temporal information	*Provide temporal information with the date of publication as a reference.*

Table 7.4: Thematic analysis of the feedback of the graph-based visualization prototype.

	Question
Q1	Age group
Q2	Affiliation
Q3	Occupation
Q4	How frequently do you read online news?
Q5	How overwhelmed do you feel with the amount of news you are exposed to daily?
Q6	How frequently do you encounter visual elements (e.g., charts, timelines) in online news?
Q7	When a news article has visual elements (e.g. charts, timelines), how useful are they to the comprehension of that news article?
Q8	Which of the following expressions better describes your news reading style?
Q9	Which elements should be present in a visual representation of a news narrative?
Q10	Given the elements previously mentioned, classify them into the following categories: essential, important, useful, neutral, insignificant. knowing that an essential element would have the most emphasis and an insignificant element less emphasis.
Q11	How would you rate the usefulness of the first tool, automatic generation of a storyline from the sentences in a news item?
Q12	How do you think the first tool contributed to a better understanding of the news item presented earlier?
Q13	How would you rate the usefulness of the second tool, automatic generation of a storyline from the events of a news story?
Q14	How do you think the second tool contributed to a better understanding of the news item presented earlier?
Q15	Would you be interested in using these visualizations to complement news articles?
Q16	Comments

Table 7.5: Questions asked in the survey.

gathered information on reading profile and how much the visual elements are valued in a news article. Finally, the third section, prototype evaluation, first obtained information on which elements were deemed relevant to be visually represented in a news text and then asked the participants to evaluate the tool.

7.2.1 Demographic Information

The majority (66.2%) of the participants are between 18 and 24 years old, but there is also a significant portion (23.1%) between the ages of 25 and 39. The other age groups have smaller representation, 7.7% for between the ages of 40 and 59 and 3.1% for over 59. Figure 7.10 contains the age distribution.
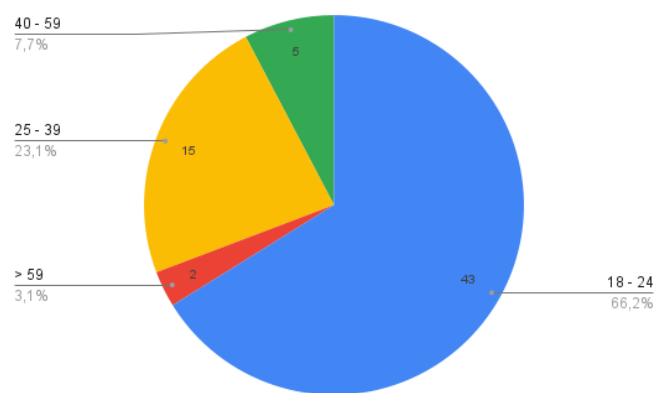
Figure 7.10: Age group distribution in survey responses.

Regarding affiliation, almost half of the responses are from the Faculty of Engineering/FEUP (49.9%), followed by the Faculty of Sciences/FCUP (15.4%), the Faculty of Medicine/FMUP (13.8%), and the School of Medicine and Biomedical Sciences/ICBAS (10.8%). There was also a small percentage of participants from the Faculty of Arts and Humanities/FLUP, a joint degree between FEUP and FLUP, the Faculty of Psychology and Educational Sciences/FPCEUP, and the Faculty of Nutrition and Food Sciences/FCNAUP. Figure 7.11 contains the affiliation distribution.

The overwhelming majority of the participants are students (58 responses corresponding to 89.2% of participants), with five responses from researchers and one response from a student worker and a worker.

The biggest disparity is in the age and affiliation of the participants, with little difference being observed regarding occupation. This corresponds to our expectations, as the survey's distribution medium was each faculty's dynamic email tool.

7.2.2 Online News Reading Habits

As seen in Figure 7.12, although the most voted option was that participants read the news between 1 and 3 times a week (32.31%), we find that around half of survey respondents (49.23%) consults
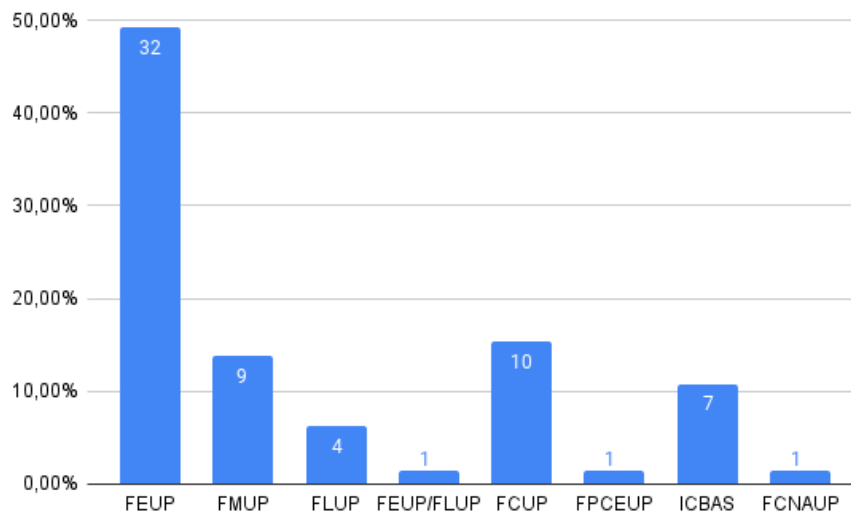
Figure 7.11: Affiliation distribution in survey responses.

online news at least once a day. We can also observe that (9.23%) of respondents are not interested in reading the news often, as they do so less than once a week.

Regarding the feeling of news overload, responses skew slightly more toward the feeling of being overloaded with the amount of online news, although it is not a significant skew. The results of question Q6 point to an even distribution where responses do not skew towards a side. This could indicate that while visualizations might not be uncommon in online news, text remains the primary communication medium. However, while visual elements might not often be present in news articles, when they are, they are considered to be extremely useful by more than a quarter of survey participants (26.15%) and overall very useful for the comprehension of the news story by 75.38% of respondents. Figure 7.13 illustrates this reader profile.

When asked about their reading method, if a panoramic view of the news is preferred over reading chosen news articles in a detailed way, the vast majority of participants (67.69%) followed the panoramic view approach, and 16.92% of the enquired do not have a preference, as seen in Figure 7.14.

7.2.3 Prototype Evaluation

Similarly to the interview script, we started this section by asking which narrative elements should be present in the visual representation of a news story. To our surprise, there were a few differences when compared to the semi-structured interviews with Text2Story team members. Both participants and publication date had the worst results, with 66.15% and 50.77% correspondingly. Title gathered 78.46% of votes, similar to spatial information and events, showing that survey participants find it more useful as a visual element than the interviewees. The data on which elements are deemed useful to be represented can be observed in Figure 7.15.
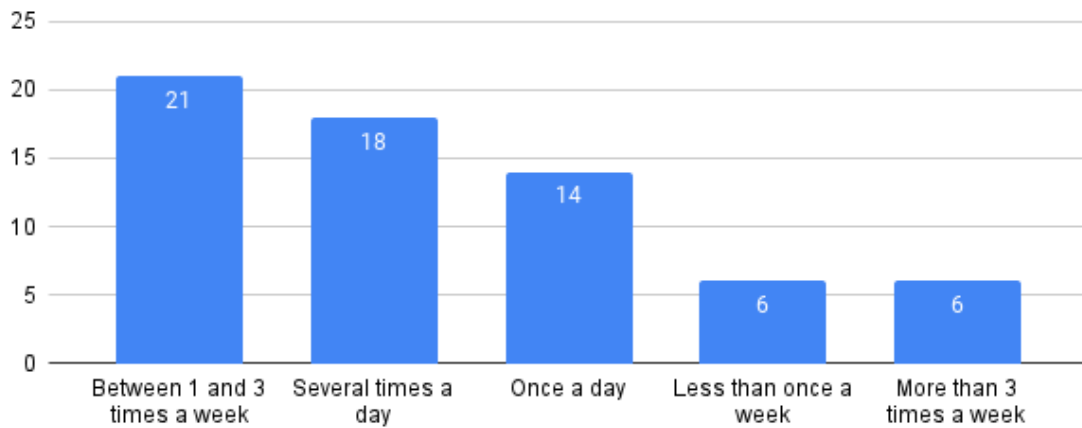
Figure 7.12: How frequently participants read online news (Q4).
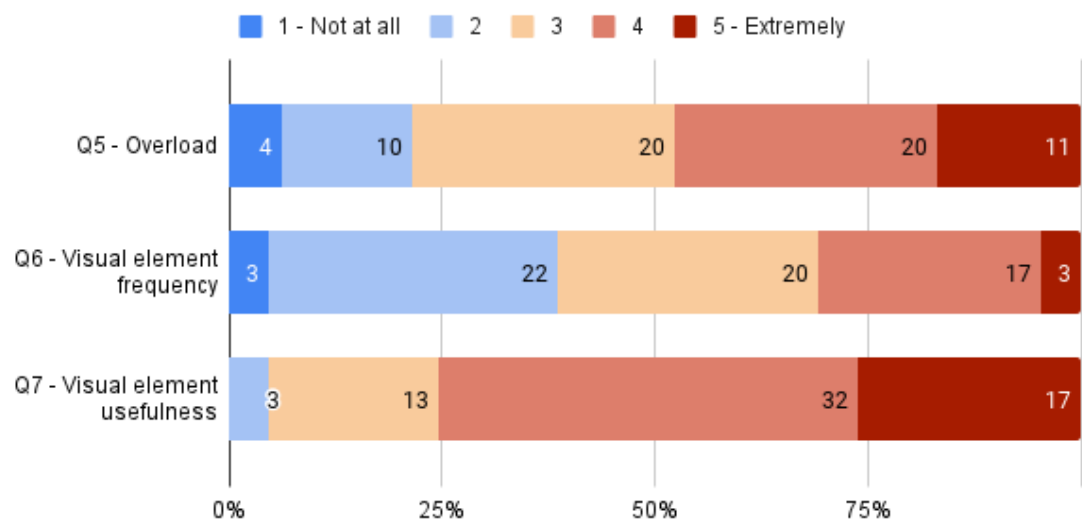
Figure 7.13: Distribution of responses regarding Q5, Q6 and Q7.
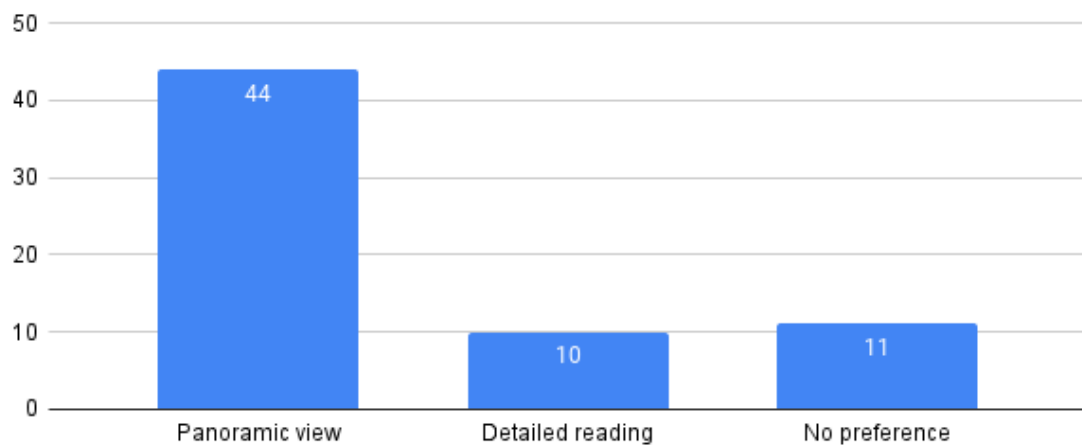
Figure 7.14: Distribution of responses regarding preferred reading method (Q8).
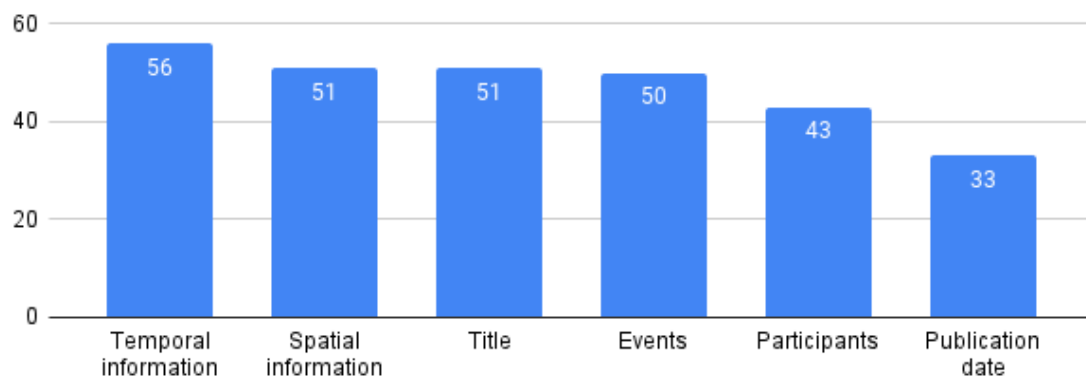
Figure 7.15: Elements that should be present in the visual representation of a news narrative (Q9).

In line with the previous question, we asked participants to rate the elements mentioned on how important it would be that they are present in the visualization on a scale from essential to insignificant. As seen in Figure 7.16, Title is deemed essential by the majority (58.46%), followed by the temporal information and events. Spatial information is also deemed important, with 70.70% of participants classifying it as important or essential. Participants and publication date obtained the worst results, as expected after the outcome of question Q9, with little over half of the answers (55.38% and 53.85% respectively) rating them as important or essential.

Two visualizations were presented to the survey participants, one displaying the sentences from the news piece and one displaying an event summary for each event of the news narrative. Both approaches are described in Section 5.5.5. Figure 7.17 and Figure 7.18 show how useful or how much the visualization contributes to a better understanding of the news piece. The second approach, displaying a summary of each event, shows a slight improvement in usefulness (15% more participants rated it with a 4 or 5) and better understanding (8% more participants rated it with a 4 or 5).

Figure 7.19 shows the interest level of using the tool visualizations to complement news stories, with 61.54% answers showing an interest level of 4 and 5.

7.2.4 Open-ended Question

A thematic analysis of the answers was conducted for the one open-ended question of the survey, which can be found in Table 7.6. A total of five answers of varied length were gathered. Four participants showed interest in the tool, and two of those noted its usefulness for narratives with longer and more complex timelines. One participant highlighted the ability to extract only the relevant events of an article. Another noted that it could be an interesting addition when spreading news on social media. Two out of the four people interested, despite seeing the potential of the tool, emphasized that the interface needs improvement. One participant suggested testing the tool in a real context by gathering feedback from non-STEM users. Another questioned the usefulness of the tool in its current state. Finally, the participant who did not show interest in the tool noted
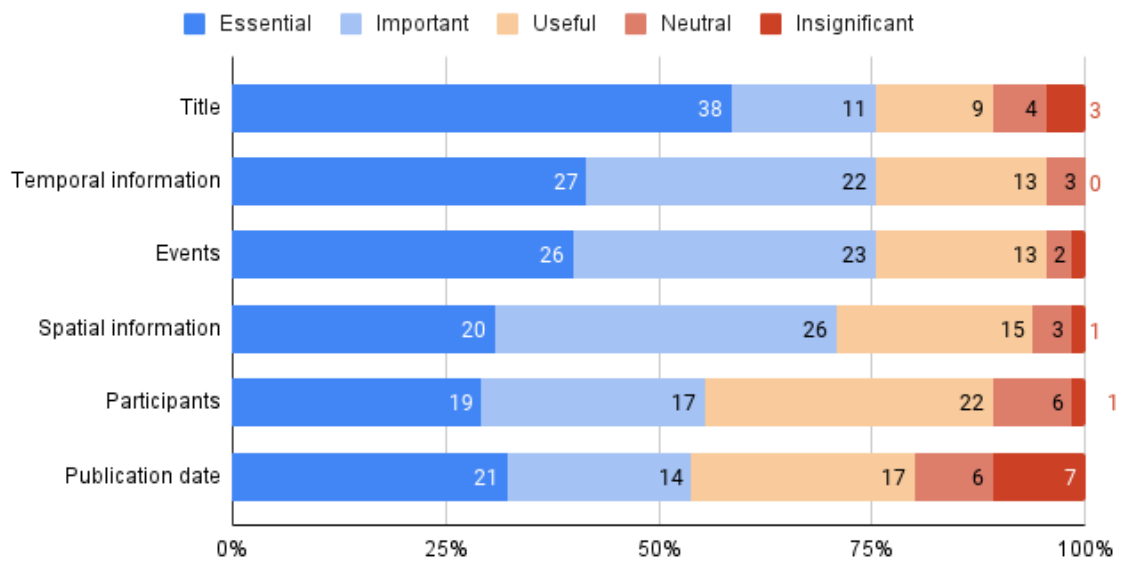
Figure 7.16: Rating each element on how important it would be in a visual representation of a news narrative (Q10).
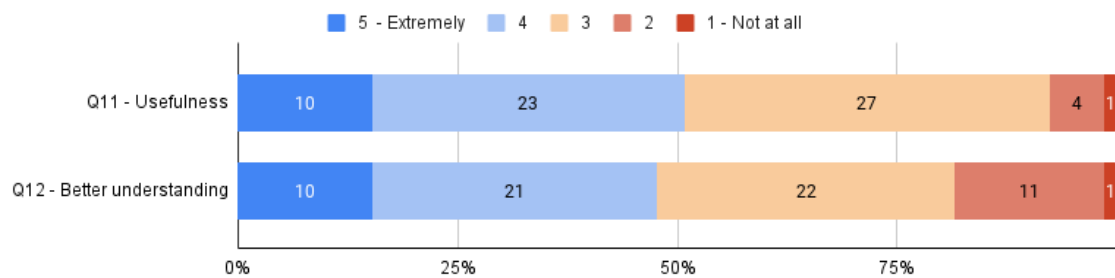
Figure 7.17: Rating the usefulness (Q11) of the news story sentence-based visualization and its contribution to a better understanding of the source news article (Q12).
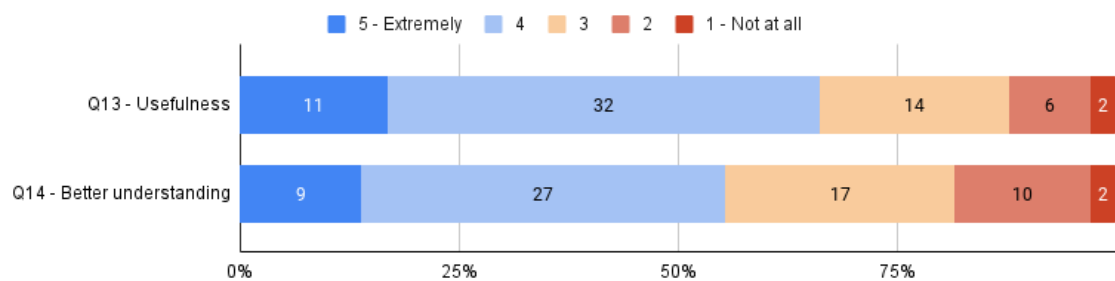
Figure 7.18: Rating the usefulness (Q13) of the news story event-based visualization and its contribution to a better understanding of the source news article (Q14).
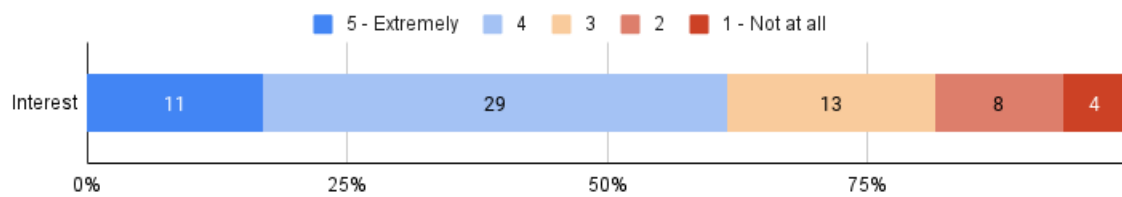
Figure 7.19: Rating the level of interest shown by the survey participants (Q15) on the visualization tool.

that the current layout is difficult to process and that too much emphasis is being given to the narrative participants.

Overall, while most of the survey participants found the tool interesting, many denoted that it is not intuitive and that the layout is confusing, emphasizing the need for improvements at the design level.

7.2.5 Discussion

After examining the outcomes of each question, we conducted an analysis to determine the relationships between the data from different questions. We found it relevant to analyze if demographic data or news reading habits had any influence on the feedback of the tool. Figure 7.22a and Figure 7.22b show that there has been an increase in how useful the visualization is deemed to be when presented with the example displaying the events, especially on the two youngest demographics. In the 18 to 24 demographic, the number of responses rating the visualization as 5, extremely useful, rose from zero to seven, and the number of people rating it as 4 also increased from eighteen to twenty-one. In the 25 - 39 demographic, the difference resides mainly in the number of people rating it as 4, which increased from four to eight, and in the number of people rating it a 3, which decreased from eight to three, showing little change in the number of negative ratings. However, results show little variance between the two visualizations in terms of improving the understanding. When comparing the responses regarding usefulness, there is no difference between the two oldest age groups, while the two youngest age groups show a polarization in their results.

Additionally, we decided to analyse the relationship between how overwhelmed participants feel with the amount of news they are exposed to and the level of interest they have in the tool. We cannot draw conclusions regarding the people who describe themselves as feeling overwhelmed with the amount of news they are exposed to and the people who remain neutral on that topic as results are very similar. Participants who are neutral have a higher overall classification, with an average interest level of 3.70 on a scale from one to five. However, their interest level distribution is more polarized percentage-wise compared to participants who feel overwhelmed. The latter group has an average interest level of 3.52. Regarding people who don't feel overwhelmed, they show less interest in the tool than the other tool groups, showing an average interest level of 3.36 and a

Codes	Examples
Positive feedback	
Highlights the relevant points	*The tool is promising in that it makes it possible to highlight the relevant points of the news.*
Social media	*The idea is very good, especially for spreading news on social media*
Useful for news with a longer time span	*I think the tool can be very useful in a number of situations, especially for news stories with a longer time span.*
Negative feedback	
Interface design	*The design - the way it is presented to the user - could be improved to be less confusing and to present the facts more clearly.*
Doubts in usefulness	*I see potential in the tool but, in its current state, I'm not entirely convinced of its usefulness.*
Too much participant focus	*The current layout is a little difficult to digest in its entirety and, in my opinion, gives too much focus to the actors in the story, who take up as much as 3/4 of the visualization space.*
Suggestions	
Non-STEM user tests	*To study the usefulness of use in a real context, I suggest collecting opinions from users outside the STEM area.*

Table 7.6: Written feedback on the open question of the survey.

(a) Normalized.   (b) Not normalized.
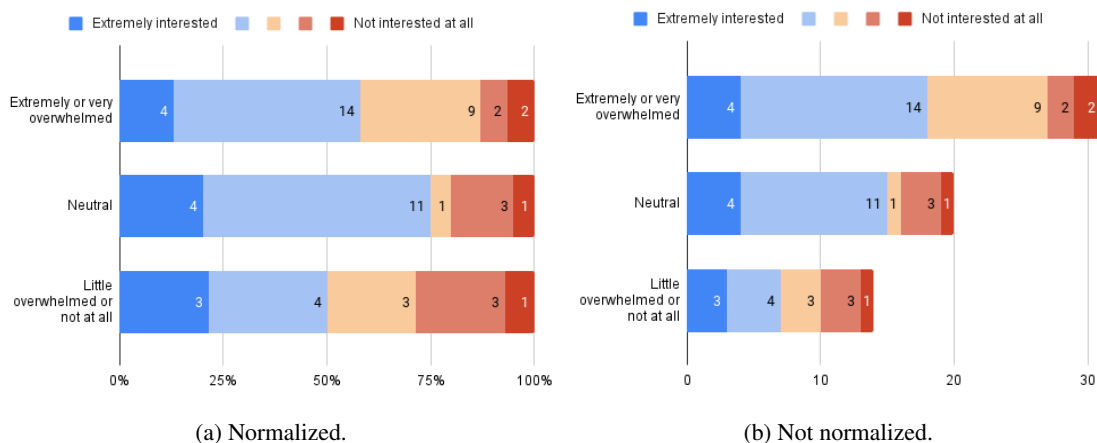
Figure 7.20: Cross-referencing the level of overwhelmed participants feel with the level of interest in the tool.

total of 28.57% of participants stating that they have little to no interest in the tool. Figure 7.20 summarizes these findings.

Lastly, we analysed how the preferred reading method can be related to the level of interest in the tool. Figure 7.21 displays the results. Our analysis did not yield any conclusive results, as the distribution is similar between all the reading methods. The biggest difference is seen on the participants who prefer to read news pieces in a more detailed way. The participants demonstrate a slightly higher level of interest in the tool and a slightly less negative attitude towards it.

7.3 Results Comparison

Survey respondents had different preferences when compared to the interviewees. Especially comparing which elements were found relevant in a visualization, survey participants considered title the most essential element to be present in a visualization, as seen in Figure 7.16. On the contrary, interview participants rated title as *neutral* when mentioned, having been mentioned only 20% of the time, as seen in Figure 7.1a. Additionally, besides the title, survey participants considered events and temporal information as the next most important elements, while interviewed researchers considered events and participants as such.

Despite these differences, both the surveys and the interviews gathered positive feedback on the sentence-based and event-based visualizations. The event-based approach had better results than its sentence-based counterpart, showing more positive feedback in the interviews and a better evaluation regarding its usefulness and contribution to a better understanding of the news piece in the survey.

(a) Normalized.
(b) Not normalized.
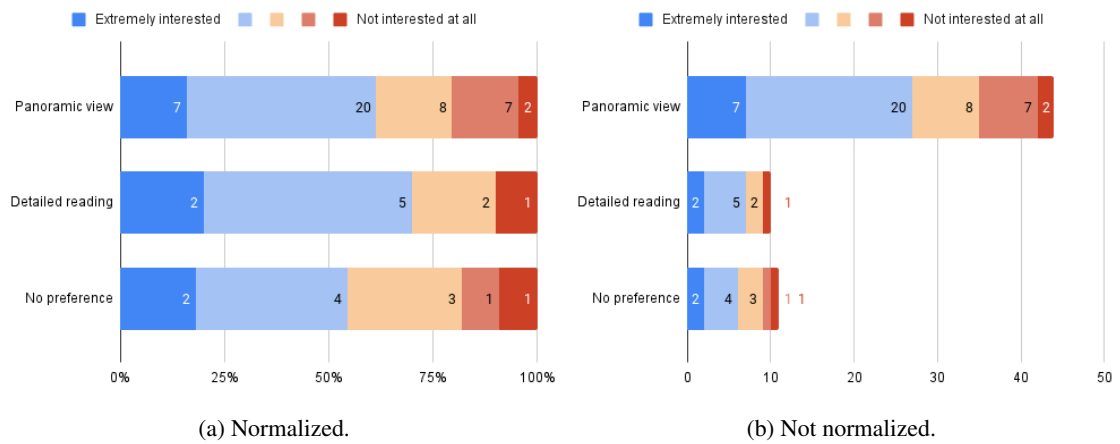
Figure 7.21: Cross-referencing the preferred reading method with the level of interest in the tool.

(a) Usefulness of the news story sentence visualization (Q11).
(b) Usefulness of the news story event visualization (Q13).
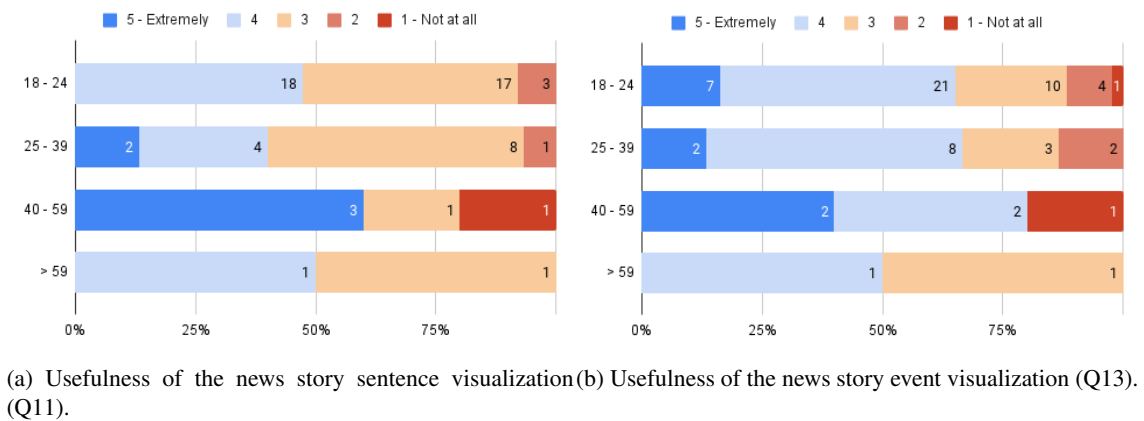
Figure 7.22: Comparison between the usefulness of each visualization according to the group.

(a) Better understanding of the news story sentence visualization (Q12).
(b) Better understanding of the news story event visualization (Q14).
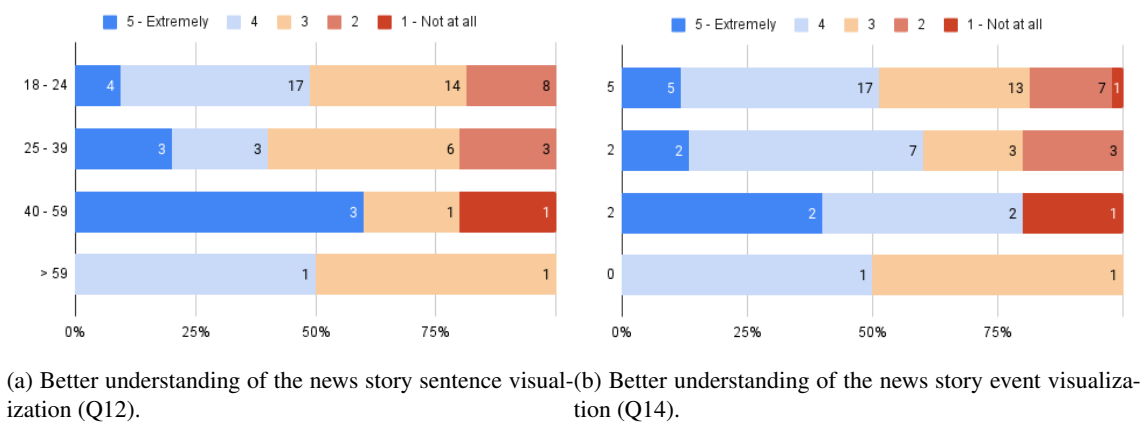
Figure 7.23: Comparison between the better understanding of each visualization according to age group.
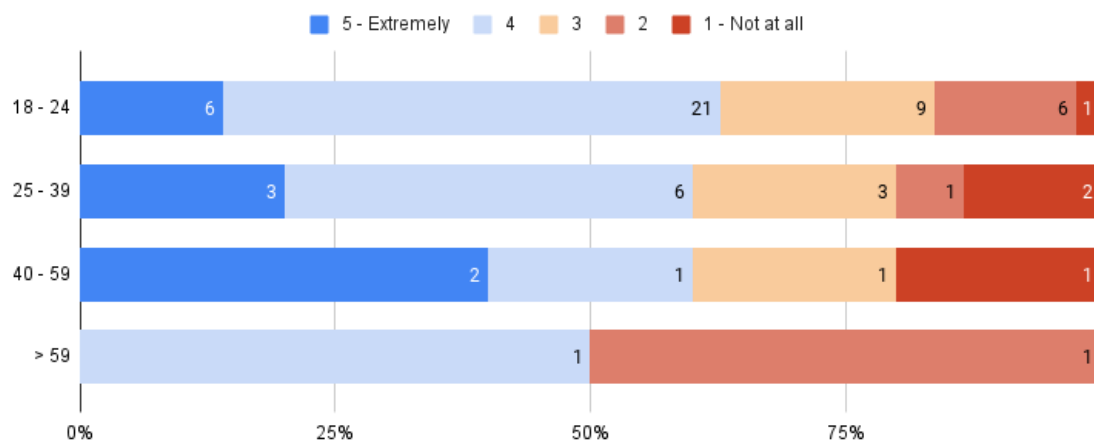
Figure 7.24: Comparison between the level of interest (Q15) in the tool.

7.4 Limitations

In terms of semi-structured interviews, there are a few concerns regarding the validity of the results obtained, including the number of participants and the characteristics of the sample. Our primary objective was to gather ideas for other visualizations, improve NewsLines, and assess the quality of the changes that have already been made. The participant pool consisted of eight researchers: four from the Faculty of Sciences, three from the Faculty of Arts and Humanities, and one from Engineering. All of them had previous experience with the text2story project. Given their expertise in the field, their responses were influenced by their knowledge of the project's developments and available resources, which may have resulted in a less diverse range of answers. It could be worthwhile to conduct interviews with individuals who are not familiar with the project in order to gain a fresh perspective.

Regarding the survey, three main topics might concern the validity of results: the question types, the comprehensiveness of the examples, and the sample characteristics.

The survey questions were designed to be simple and easy to answer to attract as many participants as possible. The downside is that we get only a small picture of the participants' reasoning and do not get to explore their thought process. Additionally, we do not know exactly what type of news story participants think of when asked to describe what visual elements should be present in its representation, information that would be relevant for the analysis.

Before delving into the final part of the questionnaire and gathering input about the prototype, a captioned figure was shown to the participant, indicating the meaning of each visual element. Additionally, a minute-long video was shown before the questions about the two visualizations. The visualization examples that illustrate the working tool focused on showing the features that have suffered changes since Costa's [22] implementation and that are relevant to the general public. What has suffered the most changes is the textual content of the scenes in the storyline. Therefore, features such as editing the visualization and filtering it by entity, date, or events were not shown, as they had been previously evaluated in Costa's work. However, the consequence of this choice

might have been that survey participants found it less interesting than if those features had been showcased.

The sample characteristics of the results could also raise validity concerns due to the low number of results, especially considering the vast majority of respondents are students and the low number of participants older than forty not from the Faculties of Engineering, Science, Medicine, or Biomedical faculties.

7.5 Summary

This chapter evaluates the interface of NewsLines, its improvements, and how it contributes to a better understanding of a news story, using real news story data. The evaluation method includes a qualitative analysis based on semi-structured interviews with eight researchers involved in the Text2Story project, and a quantitative analysis based on a survey with 65 participants.

The semi-structured interviews aim to gather feedback and suggestions for improvement from the researchers. The analysis covers three prototypes: a sentence-based NewsLines adaptation, an event-based NewsLines adaptation, and a graph-based visualization prototype. The survey aims to measure the usefulness, usability, and user satisfaction of the sentence-based and event-based NewsLines adaptations.

The analysis reveals the strengths and weaknesses of each prototype, as well as the challenges and opportunities for future work.

Chapter 8

Conclusions and Future Work

In this thesis, we explored Mariana Costa's NewsLines [22] tool for narrative visualization. Because the tool had only been tried out in illustrative cases, we integrated it with the annotated news data from the Text2Story project and gathered feedback from the public and the Text2Story team in order to assess and improve the tool.

For that, we have proposed converting BRAT annotated news story files to JSON and integrating those JSON files into the NewsLines tool. Additionally, we proposed modifications to the NewsLines tool to enable it to adapt to the updated JSON schema and enhance it as a whole.

We started by analyzing the current state of the news landscape and techniques for information and narrative visualization, as well as storytelling. Furthermore, we analyzed five works in the narrative visualization domain to find a wide variety of visualization types being adopted, with *Map* and *River* being most commonly adopted in the tools studied.

In Chapter 3, we examined the Text2Story pipeline, the two file types that are part of it, BRAT annotations and DRS files, and how they are produced. Consequently, we analyzed the annotation process and how that information is translated from one format to the other. Moreover, we looked into the visualizations generated through this pipeline, message sequence charts, and knowledge graphs.

Chapter 4 focuses on studying the components of Mariana Costa's NewsLines [22]. This open-source web prototype aims to display news story visualizations in a similar way to the storyline ones by Munroe [61]. We went over the NewsLines interface, input data, and architecture in detail to understand the inner workings of the tool.

Chapter 5 introduces the problem definition and our approach to tackling it. We started by reviewing existing approaches that aim to convert files to a JSON format. Then, we explained our specific methodology to integrate the annotated data into Costa's visualization tool and the improvements made in NewsLines.

Chapter 6 showcases the framework developed in this research work. For this demonstration, two examples were analysed, illustrating the advantages and limitations of the tool.

The evaluation of our solution, in Chapter 7, lies in two evaluation methods, interviews with researchers of the project, and a survey released to the members of five faculties of the University of Porto. The main goals of the interviews were to gather feedback on the implemented solution and to get ideas, not only on how to improve NewsLines but also on new types of visualizations to develop in the future. The survey aimed to understand the online news reading habits of the general population and see how they have changed since Costa's survey. Besides this, it aimed to identify which news elements people considered most valuable in a visualization and assess the usefulness of the sentence-based and event-based integration approaches to the final user. We obtained vastly different results in the interviews and the survey regarding what elements were relevant to add to a visualization. However, most people in both evaluation methods showed interest and saw potential in the prototype. This way, our hypothesis *Automating the conversion of BRAT to JSON fosters the automatic creation of visualization storylines, thus empowering large-scale news story understanding* can be deemed valid. Not only did participants from both the interviews and the survey showed interest and found our system useful, but also our framework allows for large-scale automation of news narrative visualizations production from real data.

Based on our results and feedback, we identified some possible directions for future work. One of them would be focused on the text annotations. It would be interesting to further analyze the *theme* semantic role, as it encompasses two different use cases, one of which is similar to the one of the *patient* semantic role. The only difference between the two is that the *theme* role is not structurally changed by the event or in the state in which it occurs. Given that this role also can link a participant to an event, it would be relevant to expand its role in the event-based scene generation technique. In addition to that, it would be relevant to explore the feedback gathered in the interviews regarding changes in the storyline visualization, especially the comments involving the addition of semantic roles to the storyline, as it could make the scene summaries easier to understand. Lastly, another potential direction for future work could involve the development of a new visualization, taking into account the sketch prototypes created during the interviews.

References

[1] mscgenjs library. Available at: https://www.npmjs.com/package/mscgenjs. Online: accessed September 2023.

[2] vis.js. Available at: https://visjs.org/. Online: accessed September 2023.

[3] bratpy [source-code]. Available at: https://github.com/amadanmath/bratpy, 2022. Online: accessed September 2023.

[4] Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys J. Kochut. Text summarization techniques: A brief survey. *CoRR*, abs/1707.02268, 2017.

[5] Evelin Amorim, Alexandre Ribeiro, Brenda Salenave Santana, Inês Cantante, Alípio Jorge, Sérgio Nunes, Purificação Silvano, Antonio Leal, and Ricardo Campos. Brat2viz: a tool and pipeline for visualizing narratives from annotated texts. In Ricardo Campos, Alípio Mário Jorge, Adam Jatowt, Sumit Bhatia, and Mark A. Finlayson, editors, *Proceedings of Text2Story - Fourth Workshop on Narrative Extraction From Texts held in conjunction with the 43rd European Conference on Information Retrieval (ECIR 2021), Lucca, Italy, April 1, 2021 (online event due to Covid-19 outbreak)*, volume 2860 of *CEUR Workshop Proceedings*, pages 49–56. CEUR-WS.org, 2021.

[6] Alok Baikadi, Julius Goth, Christopher Mitchell, Eun Ha, Bradford Mott, and James Lester. Towards a computational model of narrative visualization. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 7, pages 2–9, 2011.

[7] David Bamman, Sejal Popat, and Sheng Shen. An annotated dataset of literary entities. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2138–2144. Association for Computational Linguistics, 2019.

[8] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186, 2013.

[9] Toms Bergmanis and Sharon Goldwater. Context sensitive neural lemmatization with lematus. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA,*

June 1-6, 2018, Volume 1 (Long Papers), pages 1391–1400. Association for Computational Linguistics, 2018.

[10] Ean Bisma. Screenplay Parser [source-code]. Available at: https://github.com/SMASH-CUT/screenplay-pdf-to-json, 2020. Online: accessed September 2023.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[12] Bernd Bohnet, Ryan T. McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. Morphosyntactic tagging with a meta-bilstm model over context sensitive token encodings. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2642–2652. Association for Computational Linguistics, 2018.

[13] Virginia Braun and Victoria Clarke. Thematic analysis. In Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf, and Kenneth J. Sher, editors, *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, pages 57–71. American Psychological Association.

[14] Julian Brooke, Adam Hammond, and Timothy Baldwin. Bootstrapped text-level named entity recognition for literature. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics, 2016.

[15] Ricardo Campos, Jorge Duque, Tiago Cândido, Jorge Mendes, Gaël Dias, Alípio Jorge, and Célia Nunes. Time-matters: Temporal unfolding of texts. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, pages 492–497. Springer, 2021.

[16] Nan Cao and Weiwei Cui. *Introduction to Text Visualization*, volume 1 of *Atlantis Briefs in Artificial Intelligence*. Atlantis Press, 2016.

[17] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in information visualization - using vision to think*. Academic Press, 1999.

[18] Christian Chiarcos. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 4569–4577. European Language Resources Association (ELRA), 2014.

[19] Fabio Ciotti. Toward a formal ontology for narrative. *MATLIT: Materialidades da Literatura*, 4(1):29–44, 2016.

[20] Alex Coppen. Screenplay JSON [source-code]. Available at: https://github.com/alexc-hollywood/screen-json, 2017. Online: accessed September 2023.

[21] Mariana Costa. NewsLines (version 1.0.0) [source-code]. Available at: `https://github.com/marianafcosta/news-story-viz`, 2021. Online: accessed July 2023.

[22] Mariana Filipa da Costa. Narrative visualization of news stories. Master's thesis, Faculty of Engineering of the University of Porto, Porto, July 2021.

[23] Domo. Data never sleeps 10.0. Available at: `https://www.domo.com/data-never-sleeps`. Online: accessed September 2023.

[24] Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *7th IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2012, Seattle, WA, USA, October 14-19, 2012*, pages 93–102. IEEE Computer Society, 2012.

[25] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. In Michael Martin, Martí Cuquet, and Erwin Folmer, editors, *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016*, volume 1695 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[26] Simon Elvery. Narrative charts [source-code]. Available at: `https://github.com/abcnews/d3-layout-narrative`, 2016. Online: accessed August 2023.

[27] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of dbpedia, freebase, opencyc, wikidata, and YAGO. *Semantic Web*, 9(1):77–129, 2018.

[28] Catarina Fernandes. NewsLines (version 2.0.0) [source-code]. Available at: `https://github.com/catarina03/news-story-viz`, 2023. Online: accessed August 2023.

[29] Ana Raquel de Ponte Figueiras. How to tell stories using visualization: strategies towards narrative visualization. 2016.

[30] Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. Visualization of text document corpus. *Informatica (Slovenia)*, 29(4):497–504, 2005.

[31] Francisco Gonçalves, Ricardo Campos, and Alípio Jorge. Text2storyline: Generating enriched storylines from text. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III*, volume 13982 of *Lecture Notes in Computer Science*, pages 248–254. Springer, 2023.

[32] Francisco Manuel Pires Gonçalves. Text2storyline: Generating enriched storylines from text. Accepted: 2023-02-01T00:10:13Z.

[33] David Harel and PS Thiagarajan. Message sequence charts. *UML for real: design of embedded real-time systems*, pages 77–105, 2003.

[34] Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. Jointly predicting predicates and arguments in neural semantic role labeling. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 364–369. Association for Computational Linguistics, 2018.

[35] Swapnil Hingmire, Nitin Ramrakhiyani, Avinash Kumar Singh, Sangameshwar Patil, Girish Palshikar, Pushpak Bhattacharyya, and Vasudeva Varma. Extracting message sequence charts from Hindi narrative text. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 87–96, Online, July 2020. Association for Computational Linguistics.

[36] Thomas Hofmann. Probabilistic latent semantic indexing. *SIGIR Forum*, 51(2):211–218, 2017.

[37] Nancy Ide. Case study: the manually annotated sub-corpus. *Handbook of Linguistic Annotation*, pages 497–519, 2017.

[38] Nancy Ide, Randi Reppen, and Keith Suderman. The american national corpus: More than the web can provide. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*. European Language Resources Association, 2002.

[39] Astutic Artificial Intelligence. BRAT standoff to JSON [source-code]. Available at: https://github.com/astutic/brat-standoff-to-json, 2022. Online: accessed September 2023.

[40] ISO Central Secretary. Language resource management — Semantic annotation framework (SemAF) — Part 1: Time and events (SemAF-Time, ISO-TimeML). Standard ISO 24617-1:2012, International Organization for Standardization, Geneva, CH, 2012.

[41] ISO Central Secretary. Language resource management — Semantic annotation framework (SemAF) — Part 4: Semantic roles (SemAF-SR). Standard ISO 24617-4:2014, International Organization for Standardization, Geneva, CH, 2014.

[42] ISO Central Secretary. Language resource management — Semantic annotation framework — Part 9: Reference annotation framework (RAF). Standard ISO 24617-9:2019, International Organization for Standardization, Geneva, CH, 2019.

[43] ISO Central Secretary. Language resource management — Semantic annotation framework — Part 7: Spatial information. Standard ISO 24617-7:2020, International Organization for Standardization, Geneva, CH, 2020.

[44] Abdul Jabbar, Sajid Iqbal, Manzoor Ilahi, Shafiq Hussain, and Adnan Akhunzada. Empirical evaluation and study of text stemming algorithms. *Artif. Intell. Rev.*, 53(8):5559–5588, 2020.

[45] Hans Kamp and Uwe Reyle. *From Discourse to Logic - Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42 of *Studies in linguistics and philosophy*. Springer, 1993.

[46] Owen Kaser and Daniel Lemire. Tag-cloud drawing: Algorithms for cloud visualization. *CoRR*, abs/cs/0703109, 2007.

[47] Nam Wook Kim, Benjamin Bach, Hyejin Im, Sasha Schriber, Markus Gross, and Hanspeter Pfister. Visualizing nonlinear narratives with story curves. *IEEE transactions on visualization and computer graphics*, 24(1):595–604, 2017.

[48] Thomas Krause and Amir Zeldes. ANNIS3: A new architecture for generic corpus query and visualization. *Digit. Scholarsh. Humanit.*, 31(1):118–139, 2016.

[49] Milos Krstajic, Enrico Bertini, and Daniel A. Keim. Cloudlines: Compact display of event episodes in multiple time-series. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2432–2439, 2011.

[50] António Leal, Purificação Silvano, Evelin Amorim, Inês Cantante, Fátima Silva, Alípio Mario Jorge, and Ricardo Campos. The place of ISO-space in Text2Story multilayer annotation scheme. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 61–70, Marseille, France, June 2022. European Language Resources Association.

[51] Jiaqi Li, Ming Liu, Bing Qin, and Ting Liu. A survey of discourse parsing. *Frontiers of Computer Science*, 16(5):165329, 2022.

[52] Linguistic Data Consortium. ACE (automatic content extraction) English Annotation Guidelines for Events. Standard, 2005.

[53] Shixia Liu, Yingcai Wu, Enxun Wei, Mengchen Liu, and Yang Liu. Storyflow: Tracking the evolution of stories. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2436–2445, 2013.

[54] Shixia Liu, Michelle X. Zhou, Shimei Pan, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. Interactive, topic-based visual text summarization and analysis. In David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy Lin, editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 543–552. ACM, 2009.

[55] Chad R Lochmiller. Conducting thematic analysis with qualitative data. *The Qualitative Report*, 26(6):2029–2044, 2021.

[56] Frank Martin Mehlhose, Michael Petrifke, and Christoph Lindemann. Evaluation of graph-based algorithms for guessing user recommendations of the social network instagram. In *15th IEEE International Conference on Semantic Computing, ICSC 2021, Laguna Hills, CA, USA, January 27-29, 2021*, pages 409–414. IEEE, 2021.

[57] Daniele Metilli, Valentina Bartalesi, and Carlo Meghini. Steps towards a system to extract formal narratives from text. In *Text2Story@ ECIR*, pages 53–61, 2019.

[58] Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, 2016.

[59] Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. Rethinking self-attention: Towards interpretability in neural parsing. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 731–742. Association for Computational Linguistics, 2020.

[60] Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with mmax2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214, 2006.

[61] Randall Munroe. Movie narrative charts. Available at: https://xkcd.com/657/. Online: accessed June 2023.

[62] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[63] Nic Newman, Richard Fletcher, Kirsten Eddy, Craig T Robertson, and Rasmus Kleis Nielsen. Digital news report 2023. 2023.

[64] Sérgio Nunes, Alípio Jorge, António Leal, Evelin Amorim, Hugo Sousa, Inês Cantante, Purificação Silvano, and Ricardo Campos. Text2Story Lusa [data set]. *INESC TEC*, 2023.

[65] Michael Ogawa and Kwan-Liu Ma. Software evolution storylines. In Alexandru C. Telea, Carsten Görg, and Steven P. Reiss, editors, *Proceedings of the ACM 2010 Symposium on Software Visualization, Salt Lake City, UT, USA, October 25-26, 2010*, pages 35–42. ACM, 2010.

[66] David D Palmer. Tokenisation and sentence segmentation. *Handbook of natural language processing*, pages 11–35, 2000.

[67] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508, 2017.

[68] Mary C. Potter, Brad Wyble, Carl Erick Hagmann, and Emily S. McCourt. Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics*, 76(2):270–279.

[69] Jay Pujara, Hui Miao, Lise Getoor, and William W. Cohen. Knowledge graph identification. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha F. Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, volume 8218 of *Lecture Notes in Computer Science*, pages 542–557. Springer, 2013.

[70] Helen C. Purchase. Which aesthetic has the greatest effect on human understanding? In Giuseppe Di Battista, editor, *Graph Drawing, 5th International Symposium, GD '97, Rome, Italy, September 18-20, 1997, Proceedings*, volume 1353 of *Lecture Notes in Computer Science*, pages 248–261. Springer, 1997.

[71] Mario Pérez-Montoro and Xaquín Veira-González. Information visualization in digital news media. In Mario Pérez-Montoro, editor, *Interaction in Digital News Media: From Principles to Practice*, pages 33–53. Springer International Publishing.

[72] Yongbin Qin, Weizhe Yang, Kai Wang, Ruizhang Huang, Feng Tian, Shaolin Ao, and Yanping Chen. Entity relation extraction based on entity indicators. *Symmetry*, 13(4):539, 2021.

[73] Melissa Roemmele and Andrew Gordon. An encoder-decoder approach to predicting causal relations in stories. In *Proceedings of the First Workshop on Storytelling*, pages 50–59, 2018.

[74] Paul Sagan and Tom Leighton. The Internet & the future of news. *Daedalus*, 139(2):119–125, 04 2010.

[75] Brenda Salenave Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. A survey on narrative extraction from textual data. *Artif. Intell. Rev.*, 56(8):8393–8435, 2023.

[76] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1139–1148, 2010.

[77] Purificação Silvano, António Leal, Fátima Silva, Inês Cantante, Fatima Oliveira, and Alípio Mario Jorge. Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 1–13, Groningen, The Netherlands (online), June 2021. Association for Computational Linguistics.

[78] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April 2012. Association for Computational Linguistics.

[79] Jannik Strötgen and Michael Gertz. Multilingual and cross-domain temporal tagging. *Lang. Resour. Evaluation*, 47(2):269–298, 2013.

[80] Amber Stubbs. MAE and MAI: lightweight annotation and adjudication tools. In *Proceedings of the Fifth Linguistic Annotation Workshop, LAW 2011, June 23-24, 2011, Portland, Oregon, USA*, pages 129–133. The Association for Computer Linguistics, 2011.

[81] Yuzuru Tanahashi and Kwan-Liu Ma. Design considerations for optimizing storyline visualizations. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2679–2688, 2012.

[82] Chao Tong, Richard C. Roberts, Rita Borgo, Sean P. Walton, Robert S. Laramee, Kodzo Wegba, Aidong Lu, Yun Wang, Huamin Qu, Qiong Luo, and Xiaojuan Ma. Storytelling and visualization: An extended survey. *Inf.*, 9(3):65, 2018.

[83] Joana Valente, Alípio Jorge, and Sérgio Nunes. Text2icons: linking icons to narrative participants (position paper). In Ricardo Campos, Alípio Mário Jorge, Adam Jatowt, Sumit Bhatia, and Marina Litvak, editors, *Proceedings of Text2Story - Fifth Workshop on Narrative Extraction From Texts held in conjunction with the 44th European Conference on Information Retrieval (ECIR 2022), Stavanger, Norway, April 10, 2022*, volume 3117 of *CEUR Workshop Proceedings*, pages 111–116. CEUR-WS.org, 2022.

[84] Joana Maria Lima Valente. Text2icons: using AI to tell a story with icons. Master's thesis, Porto, November 2021. Accepted: 2022-09-11T08:45:51Z.

[85] Colin Ware. Chapter nine - images, narrative, and gestures for explanation. In Colin Ware, editor, *Information Visualization (Third Edition)*, Interactive Technologies, pages 325–343. Morgan Kaufmann, third edition edition.

[86] Michelle Hoda Wilkerson, William Finzer, Tim Erickson, and Damaris Hernandez. Reflective data storytelling for youth: The CODAP story builder. In *IDC '21: Interaction Design and Children, Athens, Greece, 24-30 June, 2021*, pages 503–507. ACM, 2021.

[87] Ji Soo Yi, Youn ah Kang, John T. Stasko, and Julie A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1224–1231, 2007.

[88] Amir Zeldes. *Multilayer corpus studies*. Routledge, 2018.

[89] Yuhang Zhao. Analysis of TikTok's success based on its algorithm mechanism. In *2020 International Conference on Big Data and Social Sciences (ICBDSS)*, pages 19–23. IEEE.

Appendix A

Annotations

This appendix will present an example of a text file from the Lusa dataset and the corresponding BRAT annotation. Section A.1 and Section A.2 show the text and annotation file correspondingly. Section A.3 and Section A.4 present the Lusa 112 file, and Section A.5 and Section A.6 the Lusa file 120 from the Text2Story dataset.

A.1 Text File Example

Listing A.1 shows the Lusa 0 text file, from the Text2Story Lusa dataset.

```
1 Redação, 11 out 2020
2 VAM // JH
3 Covid-19: GNR acabou com festa ilegal com 50 pessoas em São Brás de Alportel
4 A GNR acabou no sábado com uma festa ilegal num bar em Corotelo, concelho de São Br
      ás de Alportel, que reunia cerca de 50 pessoas, foi hoje anunciado.
5 Em comunicado, o Comando Territorial de Faro da GNR adianta que, na sequência de
      uma denúncia a informar sobre a existência de "uma festa com música ao vivo num
       estabelecimento de bebidas", militares do Posto Territorial de São de Alportel
       deslocaram-se ao local para fiscalizar a ocorrência.
6 Os militares desmobilizaram as pessoas para "cumprimento das normas referentes à
      pandemia de covid-19".
7 Na nota é ainda referido que o proprietário foi identificado e foi elaborado "um
      auto de contraordenação".
```

Listing A.1: Example of a text file from the Text2Story Lusa dataset.

A.2 Annotation Example

Listing A.2 shows the Lusa 0 annotated file, resulting from the Lusa 0 text file.

```
1 T24 Participant 108 113 A GNR
```

```
 2 A1  Lexical_Head T24 Noun
 3 A2  Individuation_Domain T24 Individual
 4 A3  Participant_Type_Domain T24 Org
 5 A4  Involvement T24 1
 6 T42 Participant 156 159 bar
 7 A5  Lexical_Head T42 Noun
 8 A6  Individuation_Domain T42 Individual
 9 A7  Participant_Type_Domain T42 Fac
10 A8  Involvement T42 1
11 T25 Participant 163 171 Corotelo
12 A9  Lexical_Head T25 Noun
13 A10 Individuation_Domain T25 Individual
14 A11 Participant_Type_Domain T25 Pl_civil
15 A12 Involvement T25 1
16 T23 Participant 173 205 concelho de São Brás de Alportel
17 A13 Lexical_Head T23 Noun
18 A14 Individuation_Domain T23 Individual
19 A15 Participant_Type_Domain T23 Pl_civil
20 A16 Involvement T23 1
21 T27 Participant 218 237 cerca de 50 pessoas
22 A17 Lexical_Head T27 Noun
23 A18 Individuation_Domain T27 Individual
24 A19 Participant_Type_Domain T27 Per
25 A20 Involvement T27 1
26 T21 Participant 262 272 comunicado
27 A21 Lexical_Head T21 Noun
28 A22 Individuation_Domain T21 Individual
29 A23 Participant_Type_Domain T21 Obj
30 A24 Involvement T21 1
31 T29 Participant 274 295 o Comando Territorial
32 A25 Lexical_Head T29 Noun
33 A26 Individuation_Domain T29 Individual
34 A27 Participant_Type_Domain T29 Org
35 A28 Involvement T29 1
36 T46 Participant 299 303 Faro
37 A33 Lexical_Head T46 Noun
38 A34 Individuation_Domain T46 Individual
39 A35 Participant_Type_Domain T46 Pl_civil
40 A36 Involvement T46 1
41 T30 Participant 307 310 GNR
42 A37 Lexical_Head T30 Noun
43 A38 Individuation_Domain T30 Individual
44 A39 Participant_Type_Domain T30 Org
45 A40 Involvement T30 1
46 T31 Participant 401 415 música ao vivo
47 A41 Lexical_Head T31 Noun
48 A42 Individuation_Domain T31 Individual
49 A43 Participant_Type_Domain T31 Other
50 A44 Involvement T31 1
```

```
51  T32 Participant 420 446 estabelecimento de bebidas
52  A45 Lexical_Head T32 Noun
53  A46 Individuation_Domain T32 Individual
54  A47 Participant_Type_Domain T32 Fac
55  A48 Involvement T32 1
56  T33 Participant 449 458 militares
57  A49 Lexical_Head T33 Noun
58  A50 Individuation_Domain T33 Individual
59  A51 Participant_Type_Domain T33 Per
60  A52 Involvement T33 1
61  T34 Participant 462 498 Posto Territorial de São de Alportel
62  A53 Lexical_Head T34 Noun
63  A54 Individuation_Domain T34 Individual
64  A55 Participant_Type_Domain T34 Fac
65  A56 Involvement T34 1
66  T48 Participant 462 479 Posto Territorial
67  A57 Lexical_Head T48 Noun
68  A58 Individuation_Domain T48 Individual
69  A59 Participant_Type_Domain T48 Fac
70  A60 Involvement T48 1
71  T49 Participant 483 498 São de Alportel
72  A61 Lexical_Head T49 Noun
73  A62 Individuation_Domain T49 Individual
74  A63 Participant_Type_Domain T49 Pl_civil
75  A64 Involvement T49 1
76  T35 Participant 516 521 local
77  A65 Lexical_Head T35 Noun
78  A66 Individuation_Domain T35 Individual
79  A67 Participant_Type_Domain T35 Loc
80  A68 Involvement T35 1
81  T36 Participant 552 564 Os militares
82  A69 Lexical_Head T36 Noun
83  A70 Individuation_Domain T36 Individual
84  A71 Participant_Type_Domain T36 Per
85  A72 Involvement T36 1
86  T37 Participant 580 590 as pessoas
87  A73 Lexical_Head T37 Noun
88  A74 Individuation_Domain T37 Individual
89  A75 Participant_Type_Domain T37 Per
90  A76 Involvement T37 1
91  T38 Participant 613 653 normas referentes à pandemia de covid-19
92  A77 Lexical_Head T38 Noun
93  A78 Individuation_Domain T38 Individual
94  A79 Participant_Type_Domain T38 Other
95  A80 Involvement T38 1
96  T39 Participant 659 663 nota
97  A81 Lexical_Head T39 Pronoun
98  A82 Individuation_Domain T39 Individual
99  A83 Participant_Type_Domain T39 Obj
```

```
100  A84 Involvement T39 1
101  T40 Participant 685 699 o proprietário
102  A85 Lexical_Head T40 Noun
103  A86 Individuation_Domain T40 Individual
104  A87 Participant_Type_Domain T40 Per
105  A88 Involvement T40 1
106  T17 Participant 734 760 um auto de contraordenação
107  A89 Lexical_Head T17 Noun
108  A90 Individuation_Domain T17 Individual
109  A91 Participant_Type_Domain T17 Obj
110  A92 Involvement T17 1
111  T43 Time 9 20 11 out 2020
112  A93 TemporalFunction T43 Publication_Time
113  T2   Time 124 130  sábado
114  A94 TemporalFunction T2 Publication_Time
115  T5   Time 243 247  hoje
116  A95 TemporalFunction T5 Publication_Time
117  T28 Spatial_Relation 152 155  num
118  A96 Type T28 Topological
119  A97 Topological T28 Disjunction_TTP_NTTP
120  A98 PathDefining T28 Starts
121  T22 Spatial_Relation 160 162  em
122  A99 Type T22 Topological
123  A100  Topological T22 Disjunction_TTP_NTTP
124  A101  PathDefining T22 Starts
125  T47 Spatial_Relation 296 298  de
126  A102  Type T47 Topological
127  A103  Topological T47 Disjunction_TTP_NTTP
128  A104  PathDefining T47 Starts
129  T41 Spatial_Relation 416 419  num
130  A105  Type T41 Topological
131  A106  Topological T41 Disjunction_TTP_NTTP
132  A107  PathDefining T41 Starts
133  T50 Spatial_Relation 480 482  de
134  A108  Type T50 Topological
135  A109  Topological T50 Disjunction_TTP_NTTP
136  A110  PathDefining T50 Starts
137  T44 Spatial_Relation 513 515  ao
138  A111  Type T44 Topological
139  A112  Topological T44 Equal
140  A113  PathDefining T44 End
141  T1   Event 114 120 acabou
142  A114  Class T1 Occurrence
143  A115  Polarity T1 Pos
144  A116  Event_Type T1 Transition
145  A117  Pos T1 Verb
146  A118  Tense T1 Past
147  A119 Aspect T1 Perfective
148  T3   Event 139 151 festa ilegal
```

```
149  A121  Class T3 Occurrence
150  A122  Polarity T3 Pos
151  A123  Event_Type T3 Process
152  A124  Pos T3 Noun
153  A125  Tense T3 Pres
154  A126  Aspect T3 Perfective
155  T26 Event 207 210 que
156  A128  Class T26 Occurrence
157  A129  Polarity T26 Pos
158  A130  Event_Type T26 Process
159  A131  Pos T26 Noun
160  A132  Tense T26 Pres
161  A133  Aspect T26 Perfective
162  T4   Event 211 217 reunia
163  A135  Class T4 State
164  A136  Polarity T4 Pos
165  A137  Event_Type T4 State
166  A138  Pos T4 Verb
167  A139  Tense T4 Past
168  A140  Aspect T4 Progressive
169  T6   Event 248 257 anunciado
170  A142  Class T6 Reporting
171  A143  Polarity T6 Pos
172  A144  Event_Type T6 Transition
173  A145  Pos T6 Verb
174  A146  Tense T6 Past
175  A147  Aspect T6 Perfective
176  T7   Event 311 318 adianta
177  A149  Class T7 Reporting
178  A150  Polarity T7 Pos
179  A151  Event_Type T7 Transition
180  A152  Pos T7 Verb
181  A153  Tense T7 Present
182  A154  Aspect T7 Perfective
183  T8   Event 344 352 denúncia
184  A156  Class T8 Reporting
185  A157  Polarity T8 Pos
186  A158  Event_Type T8 Transition
187  A159  Pos T8 Noun
188  A160  Tense T8 Pres
189  A161  Aspect T8 Perfective
190  T9   Event 355 363 informar
191  A163  Class T9 Reporting
192  A164  Polarity T9 Pos
193  A165  Event_Type T9 Transition
194  A166  Pos T9 Verb
195  A167  Tense T9 Pres
196  A168  Aspect T9 Perfective
197  T10 Event 372 382 existência
```

```
198  A170  Class T10 State
199  A171  Polarity T10 Pos
200  A172  Event_Type T10 State
201  A173  Pos T10 Noun
202  A174  Tense T10 Pres
203  A175  Aspect T10 Perfective
204  T11 Event 391 396 festa
205  A177  Class T11 Occurrence
206  A178  Polarity T11 Pos
207  A179  Event_Type T11 Process
208  A180  Pos T11 Noun
209  A181  Tense T11 Pres
210  A182  Aspect T11 Perfective
211  T12 Event 499 512 deslocaram-se
212  A184  Class T12 I_Action
213  A185  Polarity T12 Pos
214  A186  Event_Type T12 Transition
215  A187  Pos T12 Verb
216  A188  Tense T12 Past
217  A189  Aspect T12 Perfective
218  T13 Event 527 537 fiscalizar
219  A191  Class T13 Occurrence
220  A192  Polarity T13 Pos
221  A193  Event_Type T13 Process
222  A194  Pos T13 Verb
223  A195  Tense T13 Pres
224  A196  Aspect T13 Perfective
225  T14 Event 540 550 ocorrência
226  A198  Class T14 Occurrence
227  A199  Polarity T14 Pos
228  A200  Event_Type T14 Process
229  A201  Pos T14 Noun
230  A202  Tense T14 Pres
231  A203  Aspect T14 Perfective
232  T15 Event 565 579 desmobilizaram
233  A205  Class T15 I_Action
234  A206  Polarity T15 Pos
235  A207  Event_Type T15 Transition
236  A208  Pos T15 Verb
237  A209  Tense T15 Past
238  A210  Aspect T15 Perfective
239  T16 Event 597 608 cumprimento
240  A212  Class T16 Occurrence
241  A213  Polarity T16 Pos
242  A214  Event_Type T16 Process
243  A215  Pos T16 Noun
244  A216  Tense T16 Pres
245  A217  Aspect T16 Perfective
246  T18 Event 672 680 referido
```

```
247  A219  Class T18 Reporting
248  A220  Polarity T18 Pos
249  A221  Event_Type T18 State
250  A222  Pos T18 Verb
251  A223  Tense T18 Present
252  A224  Aspect T18 Perfective
253  T19 Event 704 716 identificado
254  A226  Class T19 Occurrence
255  A227  Polarity T19 Pos
256  A228  Event_Type T19 Transition
257  A229  Pos T19 Verb
258  A230  Tense T19 Past
259  A231  Aspect T19 Perfective
260  T20 Event 723 732 elaborado
261  A233  Class T20 Occurrence
262  A234  Polarity T20 Pos
263  A235  Event_Type T20 Transition
264  A236  Pos T20 Verb
265  A237  Tense T20 Past
266  A238  Aspect T20 Perfective
267  R68 TLINK_identity Arg1:T5 Arg2:T43
268  R1   TLINK_isIncluded Arg1:T1 Arg2:T2
269  R2   TLINK_before Arg1:T4 Arg2:T1
270  R69 TLINK_before Arg1:T3 Arg2:T1
271  R8   TLINK_identity Arg1:T11 Arg2:T3
272  R19 TLINK_identity Arg1:T3 Arg2:T26
273  R75 TLINK_simultaneous Arg1:T4 Arg2:T26
274  R4   TLINK_after Arg1:T6 Arg2:T4
275  R89 TLINK_simultaneous Arg1:T8 Arg2:T4
276  R3   TLINK_isIncluded Arg1:T6 Arg2:T5
277  R31 TLINK_identity Arg1:T7 Arg2:T6
278  R7   TLINK_before Arg1:T8 Arg2:T7
279  R32 TLINK_identity Arg1:T18 Arg2:T7
280  R6   TLINK_isIncluded Arg1:T9 Arg2:T8
281  R48 TLINK_after Arg1:T12 Arg2:T8
282  R70 TLINK_simultaneous Arg1:T10 Arg2:T9
283  R9   TLINK_simultaneous Arg1:T11 Arg2:T10
284  R12 TLINK_identity Arg1:T14 Arg2:T11
285  R11 TLINK_after Arg1:T13 Arg2:T12
286  R13 TLINK_after Arg1:T15 Arg2:T12
287  R46 TLINK_simultaneous Arg1:T14 Arg2:T13
288  R14 TLINK_after Arg1:T16 Arg2:T15
289  R55 TLINK_after Arg1:T19 Arg2:T15
290  R16 TLINK_before Arg1:T19 Arg2:T18
291  R17 TLINK_after Arg1:T20 Arg2:T19
292  R21 OLINK_objIdentity Arg1:T30 Arg2:T24
293  R20 OLINK_partOf Arg1:T25 Arg2:T23
294  R27 OLINK_objIdentity Arg1:T37 Arg2:T27
295  R18 OLINK_objIdentity Arg1:T39 Arg2:T21
```

```
296  R22 OLINK_partOf Arg1:T29 Arg2:T30
297  R24 OLINK_objIdentity Arg1:T35 Arg2:T32
298  R25 OLINK_objIdentity Arg1:T36 Arg2:T33
299  R88 OLINK_partOf Arg1:T34 Arg2:T49
300  R28 SRLINK_agent Arg1:T1 Arg2:T24
301  R29 SRLINK_patient Arg1:T1 Arg2:T3
302  R35 SRLINK_theme Arg1:T6 Arg2:T1
303  R33 SRLINK_pivot Arg1:T4 Arg2:T3
304  R76 SRLINK_theme Arg1:T6 Arg2:T3
305  R23 SRLINK_cause Arg1:T4 Arg2:T26
306  R34 SRLINK_theme Arg1:T4 Arg2:T27
307  R5  SRLINK_medium Arg1:T7 Arg2:T21
308  R37 SRLINK_agent Arg1:T7 Arg2:T29
309  R59 SRLINK_agent Arg1:T18 Arg2:T29
310  R38 SRLINK_theme Arg1:T7 Arg2:T12
311  R77 SRLINK_theme Arg1:T7 Arg2:T8
312  R78 SRLINK_theme Arg1:T7 Arg2:T9
313  R79 SRLINK_theme Arg1:T7 Arg2:T10
314  R80 SRLINK_theme Arg1:T7 Arg2:T11
315  R81 SRLINK_theme Arg1:T7 Arg2:T13
316  R82 SRLINK_theme Arg1:T7 Arg2:T14
317  R10 SRLINK_reason Arg1:T12 Arg2:T8
318  R40 SRLINK_theme Arg1:T9 Arg2:T10
319  R41 SRLINK_pivot Arg1:T10 Arg2:T11
320  R42 SRLINK_setting Arg1:T11 Arg2:T31
321  R44 SRLINK_agent Arg1:T12 Arg2:T33
322  R50 SRLINK_agent Arg1:T13 Arg2:T33
323  R47 SRLINK_purpose Arg1:T12 Arg2:T13
324  R49 SRLINK_theme Arg1:T13 Arg2:T14
325  R51 SRLINK_agent Arg1:T15 Arg2:T36
326  R52 SRLINK_theme Arg1:T15 Arg2:T37
327  R53 SRLINK_purpose Arg1:T15 Arg2:T16
328  R54 SRLINK_theme Arg1:T16 Arg2:T38
329  R15 SRLINK_medium Arg1:T18 Arg2:T39
330  R58 SRLINK_theme Arg1:T18 Arg2:T19
331  R43 SRLINK_theme Arg1:T18 Arg2:T20
332  R83 SRLINK_theme Arg1:T18 Arg2:T20
333  R60 SRLINK_patient Arg1:T19 Arg2:T40
334  R56 SRLINK_result Arg1:T20 Arg2:T17
335  R36 SLINK_evidential Arg1:T6 Arg2:T1
336  R39 SLINK_evidential Arg1:T7 Arg2:T12
337  R64 SLINK_evidential Arg1:T9 Arg2:T10
338  R45 SLINK_intensional Arg1:T12 Arg2:T13
339  R63 SLINK_intensional Arg1:T15 Arg2:T16
340  R73 SLINK_evidential Arg1:T18 Arg2:T19
341  R74 SLINK_evidential Arg1:T18 Arg2:T20
342  R71 QSLINK_figure Arg1:T28 Arg2:T3
343  R62 QSLINK_ground Arg1:T28 Arg2:T42
344  R26 QSLINK_figure Arg1:T22 Arg2:T42
```

```
345  R72 QSLINK_figure Arg1:T22 Arg2:T42
346  R57 QSLINK_ground Arg1:T22 Arg2:T25
347  R85 QSLINK_figure Arg1:T47 Arg2:T29
348  R84 QSLINK_ground Arg1:T47 Arg2:T46
349  R30 QSLINK_figure Arg1:T41 Arg2:T11
350  R61 QSLINK_ground Arg1:T41 Arg2:T32
351  R87 QSLINK_figure Arg1:T50 Arg2:T48
352  R86 QSLINK_ground Arg1:T50 Arg2:T49
353  R65 MOVELINK_figure Arg1:T12 Arg2:T33
354  R66 MOVELINK_spatialRelation Arg1:T12 Arg2:T44
355  R67 MOVELINK_targetSpatialRelation Arg1:T12 Arg2:T35
356  R90 OLINK_memberOf Arg1:T33 Arg2:T30
```

Listing A.2: Example of a BRAT annotated file from the Text2Story Lusa dataset.

A.3 Lusa 112 Text File

Listing A.3 shows the Lusa 112 text file, from the Text2Story Lusa dataset.

```
1  Lisboa, 06 nov 2020
2  PMF // JPS
3  Governo: Ricardo Pinheiro substitui José Mendes como secretário de Estado do
      Planeamento
4  O deputado socialista Ricardo Pinheiro vai substituir José Mendes no cargo de
      secretário de Estado do Planeamento, mudança que foi hoje anunciada através de
      uma nota divulgada no portal da Presidência da República.
5  "O Presidente da República aceitou a proposta do primeiro-ministro de exoneração, a
       seu pedido, de José Fernandes Gomes Mendes, como secretário de Estado do
      Planeamento, e a nomeação, para o mesmo cargo, de Ricardo Miguel Furtado
      Pinheiro", lê-se na nota.
6  A posse do novo secretário de Estado do ministro do Planeamento, Nelson de Sousa,
      terá lugar ainda hoje, pelas 20:30, no Palácio de Belém.
7  Ricardo Pinheiro foi presidente da Câmara de Campo Maior, distrito de Portalegre, c
      írculo pelo qual foi eleito deputado nas últimas eleições legislativas.
```

Listing A.3: Lusa 112 text file from the Text2Story Lusa dataset.

A.4 Lusa 112 Annotation File

Listing A.4 shows the Lusa 112 annotated file, resulting from the Lusa 112 text file.

```
1  T1  Event 163 173 substituir
2  A1  Class T1 Occurrence
3  A2  Pos T1 Verb
```

```
 4  A3  Tense T1 Future
 5  A4  Aspect T1 Imperfective
 6  A5  VForm T1 Infinitive
 7  A6  Mood T1 Future
 8  A7  Polarity T1 Pos
 9  T2  Event 235 242 mudança
10  A8  Class T2 Occurrence
11  A9  Event_Type T2 Transition
12  A10 Pos T2 Noun
13  A11 Polarity T2 Pos
14  T3  Time 251 255  hoje
15  A12 Time_Type T3 Date
16  A13 TemporalFunction T3 Publication_Time
17  #1  AnnotatorNotes T3 value=2020-11-06
18  T5  Event 256 265 anunciada
19  A20 Class T5 Occurrence
20  A21 Event_Type T5 Transition
21  A22 Pos T5 Verb
22  A23 Tense T5 Past
23  A24 Aspect T5 Perfective
24  A25 VForm T5 Participle
25  A26 Polarity T5 Pos
26  T6  Event 286 295 divulgada
27  A27 Class T6 Occurrence
28  A28 Event_Type T6 Transition
29  A29 Pos T6 Verb
30  A30 Tense T6 Past
31  A31 Aspect T6 Perfective
32  A32 VForm T6 Participle
33  A33 Polarity T6 Pos
34  T7  Event 362 369 aceitou
35  A34 Class T7 I_Action
36  A35 Event_Type T7 Transition
37  A36 Pos T7 Verb
38  A37 Tense T7 Past
39  A38 Aspect T7 Perfective
40  A39 Polarity T7 Pos
41  T8  Event 372 380 proposta
42  A40 Class T8 Occurrence
43  A41 Event_Type T8 Transition
44  A42 Pos T8 Noun
45  A43 Polarity T8 Pos
46  T9  Event 405 415 exoneração
47  A44 Class T9 Occurrence
48  A45 Event_Type T9 Transition
49  A46 Pos T9 Noun
50  A47 Polarity T9 Pos
51  T10 Event 423 429 pedido
52  A48 Class T10 I_Action
```

```
53   A49 Event_Type T10 Transition
54   A50 Pos T10 Noun
55   A51 Polarity T10 Pos
56   T11 Event 509 517 nomeação
57   A52 Class T11 I_Action
58   A53 Event_Type T11 Transition
59   A54 Pos T11 Noun
60   A55 Polarity T11 Pos
61   T12 Event 576 581 lê-se
62   A56 Class T12 Occurrence
63   A57 Event_Type T12 Process
64   A58 Pos T12 Verb
65   A59 Tense T12 Present
66   A60 Aspect T12 Progressive
67   A61 Polarity T12 Pos
68   T15 Event 593 598 posse
69   A70 Class T15 State
70   A71 Event_Type T15 State
71   A72 Pos T15 Noun
72   A73 Polarity T15 Pos
73   T16 Event 673 683 terá lugar
74   A74 Class T16 Occurrence
75   A75 Event_Type T16 Transition
76   A76 Pos T16 Verb
77   A77 Tense T16 Future
78   A78 Aspect T16 Imperfective
79   A79 Mood T16 Future
80   A80 Polarity T16 Pos
81   T17 Time 690 694  hoje
82   A81 Time_Type T17 Date
83   A82 TemporalFunction T17 Publication_Time
84   #2  AnnotatorNotes T17  value=2020-11-06 anchorTime=Publication_Time
85   T18 Time 702 707  20:30
86   A83 Time_Type T18 Time
87   #3  AnnotatorNotes T18  value=2020-11-06T20:30 anchorTime=Publication_Time
88   T19 Event 834 840 eleito
89   A85 Class T19 I_Action
90   A86 Event_Type T19 Transition
91   A87 Pos T19 Verb
92   A88 Tense T19 Past
93   A89 Aspect T19 Perfective
94   A90 VForm T19 Participle
95   A91 Polarity T19 Pos
96   T20 Event 841 849 deputado
97   A92 Class T20 State
98   A93 Event_Type T20 State
99   A94 Pos T20 Noun
100  A95 Polarity T20 Pos
101  T21 Event 751 761 presidente
```

```
102  A96 Class T21 State
103  A97 Event_Type T21 State
104  A98 Pos T21 Adjective
105  A99 Polarity T21 Pos
106  T22 Event 189 194 cargo
107  A100  Class T22 State
108  A101  Event_Type T22 State
109  A102  Pos T22 Noun
110  A103  Polarity T22 Pos
111  T23 Event 532 537 cargo
112  A104  Class T23 State
113  A105  Event_Type T23 State
114  A106  Pos T23 Noun
115  A107  Polarity T23 Pos
116  R2  TLINK_identity Arg1:T2 Arg2:T1
117  R3  TLINK_isIncluded Arg1:T5 Arg2:T3
118  R4  TLINK_before Arg1:T5 Arg2:T2
119  R7  TLINK_before Arg1:T8 Arg2:T7
120  R8  TLINK_before Arg1:T7 Arg2:T6
121  R9  TLINK_identity Arg1:T23 Arg2:T22
122  R10 TLINK_after Arg1:T9 Arg2:T8
123  R11 TLINK_simultaneous Arg1:T11 Arg2:T9
124  R12 TLINK_after Arg1:T12 Arg2:T11
125  R14 TLINK_before Arg1:T10 Arg2:T8
126  R15 TLINK_identity Arg1:T16 Arg2:T15
127  R16 TLINK_isIncluded Arg1:T16 Arg2:T17
128  R17 TLINK_isIncluded Arg1:T18 Arg2:T17
129  R18 TLINK_before Arg1:T21 Arg2:T16
130  R19 TLINK_before Arg1:T19 Arg2:T21
131  R21 SLINK_evidential Arg1:T5 Arg2:T2
132  R22 SLINK_factive Arg1:T7 Arg2:T8
133  R23 SLINK_intensional Arg1:T10 Arg2:T9
134  R24 SLINK_evidential Arg1:T12 Arg2:T7
135  R25 SLINK_intensional Arg1:T19 Arg2:T20
136  T4  Participant 120 158 O deputado socialista Ricardo Pinheiro
137  A14 Lexical_Head T4 Noun
138  A15 Individuation_Domain T4 Individual
139  A16 Participant_Type_Domain T4 Per
140  A17 Involvement T4 1
141  T24 Participant 174 185 José Mendes
142  A18 Lexical_Head T24 Noun
143  A19 Individuation_Domain T24 Individual
144  A108  Participant_Type_Domain T24 Per
145  A109  Involvement T24 1
146  T25 Participant 198 233 secretário de Estado do Planeamento
147  A110  Lexical_Head T25 Noun
148  A111  Individuation_Domain T25 Individual
149  A112  Participant_Type_Domain T25 Other
150  A113  Involvement T25 1
```

```
151  T26 Participant 297 333 o portal da Presidência da República
152  A114  Lexical_Head T26 Noun
153  A115  Individuation_Domain T26 Individual
154  A116  Participant_Type_Domain T26 Other
155  A117  Involvement T26 1
156  T27 Participant 336 361 O Presidente da República
157  A118  Lexical_Head T27 Noun
158  A119  Individuation_Domain T27 Individual
159  A120  Participant_Type_Domain T27 Per
160  A121  Involvement T27 1
161  T28 Participant 382 401 o primeiro-ministro
162  A122  Lexical_Head T28 Noun
163  A123  Individuation_Domain T28 Individual
164  A124  Participant_Type_Domain T28 Per
165  A125  Involvement T28 1
166  T29 Participant 419 422 seu
167  A126  Lexical_Head T29 Pronoun
168  A127  Individuation_Domain T29 Individual
169  A128  Participant_Type_Domain T29 Per
170  A129  Involvement T29 1
171  T30 Participant 434 461 José Fernandes Gomes Mendes
172  A130  Lexical_Head T30 Noun
173  A131  Individuation_Domain T30 Individual
174  A132  Participant_Type_Domain T30 Per
175  A133  Involvement T30 1
176  R26 OLINK_objIdentity Arg1:T29 Arg2:T30
177  R27 OLINK_objIdentity Arg1:T30 Arg2:T24
178  T31 Participant 468 503 secretário de Estado do Planeamento
179  A134  Lexical_Head T31 Noun
180  A135  Individuation_Domain T31 Individual
181  A136  Participant_Type_Domain T31 Per
182  A137  Involvement T31 1
183  R28 OLINK_objIdentity Arg1:T31 Arg2:T25
184  T32 Participant 542 573 Ricardo Miguel Furtado Pinheiro
185  A138  Lexical_Head T32 Noun
186  A139  Individuation_Domain T32 Individual
187  A140  Participant_Type_Domain T32 Per
188  A141  Involvement T32 1
189  R29 OLINK_objIdentity Arg1:T32 Arg2:T4
190  T33 Participant 600 627 o novo secretário de Estado
191  A142  Lexical_Head T33 Noun
192  A143  Individuation_Domain T33 Individual
193  A144  Participant_Type_Domain T33 Per
194  A145  Involvement T33 1
195  T34 Participant 629 654 o ministro do Planeamento
196  A146  Lexical_Head T34 Noun
197  A147  Individuation_Domain T34 Individual
198  A148  Participant_Type_Domain T34 Per
199  A149  Involvement T34 1
```

```
200  T35 Participant 656 671 Nelson de Sousa
201  A150  Lexical_Head T35 Noun
202  A151  Individuation_Domain T35 Individual
203  A152  Participant_Type_Domain T35 Per
204  A153  Involvement T35 1
205  R30 OLINK_objIdentity Arg1:T35 Arg2:T34
206  R31 OLINK_objIdentity Arg1:T33 Arg2:T32
207  R32 OLINK_partOf Arg1:T33 Arg2:T34
208  T36 Participant 710 729 o Palácio de Belém.
209  A154  Lexical_Head T36 Noun
210  A155  Individuation_Domain T36 Individual
211  A156  Participant_Type_Domain T36 Loc
212  A157  Involvement T36 1
213  T37 Participant 730 746 Ricardo Pinheiro
214  A158  Lexical_Head T37 Noun
215  A159  Individuation_Domain T37 Individual
216  A160  Participant_Type_Domain T37 Per
217  A161  Involvement T37 1
218  R33 OLINK_objIdentity Arg1:T37 Arg2:T33
219  T38 Participant 763 786 a Câmara de Campo Maior
220  A162  Lexical_Head T38 Noun
221  A163  Individuation_Domain T38 Individual
222  A164  Participant_Type_Domain T38 Org
223  A165  Involvement T38 1
224  T40 Participant 788 810 distrito de Portalegre
225  A170  Lexical_Head T40 Noun
226  A171  Individuation_Domain T40 Individual
227  A172  Participant_Type_Domain T40 Pl_civil
228  A173  Involvement T40 1
229  T41 Participant 812 819 círculo
230  A174  Lexical_Head T41 Noun
231  A175  Individuation_Domain T41 Set
232  A176  Participant_Type_Domain T41 Per
233  T42 Participant 825 829 qual
234  A178  Lexical_Head T42 Pronoun
235  A179  Individuation_Domain T42 Set
236  A180  Participant_Type_Domain T42 Per
237  A181  Involvement T42 1
238  A177  Involvement T41 1
239  R36 OLINK_objIdentity Arg1:T42 Arg2:T41
240  T43 Event 862 883 eleições legislativas
241  A182  Class T43 Occurrence
242  A183  Event_Type T43 Transition
243  A184  Pos T43 Noun
244  R37 TLINK_isIncluded Arg1:T43 Arg2:T19
245  R38 SRLINK_agent Arg1:T1 Arg2:T4
246  A185  Event_Type T1 Process
247  R39 SRLINK_patient Arg1:T1 Arg2:T24
248  R40 SRLINK_setting Arg1:T1 Arg2:T22
```

```
249  R41 SRLINK_theme Arg1:T22 Arg2:T25
250  R42 SRLINK_theme Arg1:T5 Arg2:T2
251  R45 SRLINK_location Arg1:T6 Arg2:T26
252  R46 SRLINK_agent Arg1:T7 Arg2:T27
253  R47 SRLINK_theme Arg1:T7 Arg2:T8
254  R48 SRLINK_source Arg1:T8 Arg2:T28
255  R49 SRLINK_theme Arg1:T8 Arg2:T9
256  R50 SRLINK_patient Arg1:T9 Arg2:T30
257  R51 SRLINK_theme Arg1:T9 Arg2:T31
258  R52 SRLINK_agent Arg1:T11 Arg2:T28
259  R53 SRLINK_purpose Arg1:T11 Arg2:T23
260  R54 SRLINK_patient Arg1:T11 Arg2:T32
261  R56 SRLINK_patient Arg1:T15 Arg2:T33
262  R57 SRLINK_theme Arg1:T16 Arg2:T15
263  R58 SRLINK_location Arg1:T16 Arg2:T36
264  R59 SRLINK_pivot Arg1:T21 Arg2:T37
265  R60 SRLINK_theme Arg1:T21 Arg2:T38
266  R62 SRLINK_location Arg1:T21 Arg2:T40
267  R63 OLINK_objIdentity Arg1:T41 Arg2:T40
268  R64 SRLINK_source Arg1:T19 Arg2:T41
269  R65 SRLINK_theme Arg1:T19 Arg2:T20
270  R66 SRLINK_patient Arg1:T19 Arg2:T37
271  R67 SRLINK_setting Arg1:T19 Arg2:T43
272  T44 Time 8 19 06 nov 2020
273  A186  Time_Type T44 Date
274  A187  TemporalFunction T44 Publication_Time
275  #4  AnnotatorNotes T44  value=2020-11-06
276  R68 TLINK_identity Arg1:T3 Arg2:T44
277  R34 OLINK_partOf Arg1:T38 Arg2:T40
278  T39 Spatial_Relation 296 297  n
279  A84 Topological T39 Disjunction_TTP_NTTP
280  R35 QSLINK_ground Arg1:T39 Arg2:T26
281  T45 Spatial_Relation 709 710  n
282  A166  Topological T45 Disjunction_TTP_NTTP
283  R61 QSLINK_ground Arg1:T45 Arg2:T36
284  T46 Participant 775 786 Campo Maior
285  A167  Lexical_Head T46 Noun
286  A168  Individuation_Domain T46 Individual
287  A169  Participant_Type_Domain T46 Pl_civil
288  A188  Involvement T46 1
289  T47 Participant 765 771 Câmara
290  A189  Lexical_Head T47 Noun
291  A190  Individuation_Domain T47 Individual
292  A191  Participant_Type_Domain T47 Org
293  A192  Involvement T47 1
294  T48 Spatial_Relation 772 774  de
295  A193  Topological T48 Disjunction_TTP_NTTP
296  R69 QSLINK_figure Arg1:T48 Arg2:T47
297  R70 QSLINK_ground Arg1:T48 Arg2:T46
```

```
298  T13 Participant 583 589 a nota
299  A62 Lexical_Head T13 Noun
300  A63 Individuation_Domain T13 Individual
301  A64 Participant_Type_Domain T13 Obj
302  A65 Involvement T13 1
303  R1  SRLINK_medium Arg1:T12 Arg2:T13
304  T14 Participant 277 285 uma nota
305  A66 Lexical_Head T14 Noun
306  A67 Individuation_Domain T14 Individual
307  A68 Participant_Type_Domain T14 Obj
308  A69 Involvement T14 1
309  R5  SRLINK_medium Arg1:T6 Arg2:T14
310  R6  OLINK_objIdentity Arg1:T13 Arg2:T14
```

Listing A.4: Lusa 112 BRAT annotated file from the Text2Story Lusa dataset.

A.5   Lusa 120 Text File

Listing A.5 shows the Lusa 120 text file, from the Text2Story Lusa dataset.

```
1  Amadora, Lisboa, 22 dez 2020
2  JML // ROC
3  Jovem fica em estado muito grave depois de ser baleado no peito na Amadora
4  Um jovem de 17 anos ficou em estado muito grave depois de ter sido baleado no peito
       hoje de madrugada no bairro Casal da Mira, na Amadora, no distrito de Lisboa,
     disse à agência Lusa fonte da PSP.
5  De acordo com fonte do Comando Metropolitano de Lisboa (Cometlis), o alerta para a
       ocorrência foi dado às 00:53.
6  "Um jovem de 17 anos foi baleado com uma arma de fogo. Ficou ferido em estado muito
        grave e foi transportado para o Hospital Professor Doutor Fernando Fonseca [
     Amadora-Sintra]", referiu.
7  À Lusa, o Cometlis adiantou que as causas para o incidente são desconhecidas e que
       ainda não foram identificados ou detidos quaisquer suspeitos.
8  Por se tratar de um crime, o caso foi entregue à Polícia Judiciária (PJ) para
       investigação.
```

Listing A.5: Lusa 120 text file from the Text2Story Lusa dataset.

A.6   Lusa 120 Annotation File

Listing A.6 shows the Lusa 120 annotated file, resulting from the Lusa 120 text file.

```
1  T1  Event 141 162 em estado muito grave
2  A1  Class T1 State
```

```
 3  A2   Event_Type T1 State
 4  A3   Pos T1 Verb
 5  A4   Tense T1 Past
 6  A5   Aspect T1 Perfective
 7  T2   Event 182 189 baleado
 8  A6   Class T2 Occurrence
 9  A7   Event_Type T2 Transition
10  A8   Pos T2 Verb
11  A9   Tense T2 Past
12  A10  Aspect T2 Perfective
13  A11  VForm T2 Participle
14  A12  Polarity T2 Pos
15  T3   Time 199 216  hoje de madrugada
16  A13  Time_Type T3 Time
17  A14  TemporalFunction T3 None
18  #1   AnnotatorNotes T3 value=2020-12-22TM anchorTime=Publication_Time
19  T4   Event 277 282 disse
20  A15  Class T4 Reporting
21  A16  Event_Type T4 Transition
22  A17  Tense T4 Past
23  A18  Aspect T4 Perfective
24  A19  Polarity T4 Pos
25  T5   Event 381 387 alerta
26  A20  Class T5 Occurrence
27  A21  Event_Type T5 Transition
28  A22  Pos T5 Noun
29  A23  Polarity T5 Pos
30  T6   Event 395 405 ocorrência
31  A24  Class T6 Occurrence
32  A25  Event_Type T6 Transition
33  A26  Pos T6 Noun
34  A27  Polarity T6 Pos
35  T7   Event 410 414 dado
36  A28  Class T7 Occurrence
37  A29  Event_Type T7 Transition
38  A30  Pos T7 Verb
39  A31  Tense T7 Past
40  A32  Aspect T7 Perfective
41  A33  Polarity T7 Pos
42  T8   Time 418 423  00:53
43  A34  Time_Type T8 Time
44  A35  TemporalFunction T8 None
45  #2   AnnotatorNotes T8 value=2020-12-22T00:53
46  T9   Event 450 457 baleado
47  A36  Event_Type T9 Transition
48  A37  Pos T9 Verb
49  A38  Tense T9 Past
50  A39  Aspect T9 Perfective
51  A40  Polarity T9 Pos
```

```
52  T10 Event 486 492 ferido
53  A41 Class T10 Occurrence
54  A42 Event_Type T10 Transition
55  A43 Pos T10 Verb
56  A44 Tense T10 Past
57  A45 Aspect T10 Perfective
58  A46 Polarity T10 Pos
59  T11 Event 496 514 estado muito grave
60  A47 Class T11 State
61  A48 Event_Type T11 State
62  A49 Pos T11 Noun
63  A50 Polarity T11 Pos
64  T12 Event 521 533 transportado
65  A51 Event_Type T12 Transition
66  A52 Pos T12 Verb
67  A53 Tense T12 Past
68  A54 Aspect T12 Perfective
69  A55 Polarity T12 Pos
70  T13 Event 603 610 referiu
71  A56 Event_Type T13 Transition
72  A57 Pos T13 Verb
73  A58 Tense T13 Past
74  A59 Aspect T13 Perfective
75  A60 Polarity T13 Pos
76  T14 Event 631 639 adiantou
77  A61 Class T14 Reporting
78  A62 Event_Type T14 Transition
79  A63 Pos T14 Verb
80  A64 Tense T14 Past
81  A65 Aspect T14 Perfective
82  A66 Polarity T14 Pos
83  T15 Event 647 653 causas
84  A67 Class T15 State
85  A68 Event_Type T15 State
86  A69 Pos T15 Noun
87  A70 Polarity T15 Pos
88  T16 Event 661 670 incidente
89  A71 Class T16 Occurrence
90  A72 Event_Type T16 Transition
91  A73 Pos T16 Noun
92  A74 Polarity T16 Pos
93  T17 Event 675 688 desconhecidas
94  A75 Class T17 State
95  A76 Event_Type T17 State
96  A77 Pos T17 Verb
97  A78 Tense T17 Present
98  A79 Aspect T17 Imperfective
99  A80 Polarity T17 Pos
100 T18 Event 711 724 identificados
```

```
101  A81 Event_Type T18 Transition
102  A82 Pos T18 Verb
103  A83 Tense T18 Past
104  A84 Aspect T18 Perfective
105  A85 Polarity T18 Neg
106  A86 Class T18 Occurrence
107  T19 Event 728 735 detidos
108  A87 Class T19 Occurrence
109  A88 Tense T19 Past
110  A89 Aspect T19 Perfective
111  A90 Polarity T19 Neg
112  T20 Event 761 770 se tratar
113  A91 Class T20 State
114  A92 Event_Type T20 State
115  A93 Pos T20 Verb
116  A94 VForm T20 Infinitive
117  A95 Polarity T20 Pos
118  T21 Event 777 782 crime
119  A96 Class T21 State
120  A97 Event_Type T21 State
121  A98 Pos T21 Noun
122  A99 Polarity T21 Pos
123  T22 Event 786 790 caso
124  A100  Class T22 Occurrence
125  A101  Event_Type T22 Transition
126  A102  Pos T22 Noun
127  A103  Polarity T22 Pos
128  T23 Event 795 803 entregue
129  A104  Class T23 Occurrence
130  A105  Event_Type T23 Transition
131  A106  Pos T23 Verb
132  A107  Tense T23 Past
133  A108  Aspect T23 Perfective
134  A109  Polarity T23 Pos
135  T24 Event 835 847 investigação
136  A110  Class T24 Occurrence
137  A111  Event_Type T24 Process
138  A112  Pos T24 Noun
139  A113  Polarity T24 Pos
140  A114  Class T9 Occurrence
141  R1  SLINK_evidential Arg1:T14 Arg2:T17
142  R2  SRLINK_theme Arg1:T14 Arg2:T17
143  R3  SLINK_evidential Arg1:T13 Arg2:T10
144  R4  SRLINK_theme Arg1:T13 Arg2:T10
145  R5  SLINK_evidential Arg1:T4 Arg2:T1
146  R6  SRLINK_theme Arg1:T4 Arg2:T1
147  R7  TLINK_before Arg1:T2 Arg2:T1
148  R8  TLINK_isIncluded Arg1:T2 Arg2:T3
149  R9  TLINK_after Arg1:T4 Arg2:T2
```

```
150  R10 TLINK_before Arg1:T6 Arg2:T5
151  R11 TLINK_after Arg1:T7 Arg2:T6
152  R12 TLINK_isIncluded Arg1:T7 Arg2:T8
153  R13 TLINK_identity Arg1:T7 Arg2:T5
154  R14 TLINK_before Arg1:T5 Arg2:T4
155  R15 TLINK_identity Arg1:T9 Arg2:T2
156  R16 TLINK_after Arg1:T10 Arg2:T9
157  R17 TLINK_simultaneous Arg1:T11 Arg2:T10
158  R18 TLINK_after Arg1:T12 Arg2:T10
159  R19 TLINK_after Arg1:T13 Arg2:T12
160  R21 TLINK_before Arg1:T9 Arg2:T7
161  R22 TLINK_isIncluded Arg1:T14 Arg2:T13
162  R23 TLINK_identity Arg1:T16 Arg2:T9
163  R24 TLINK_includes Arg1:T17 Arg2:T14
164  R25 TLINK_simultaneous Arg1:T18 Arg2:T17
165  R26 TLINK_simultaneous Arg1:T19 Arg2:T18
166  R27 TLINK_identity Arg1:T22 Arg2:T16
167  R28 TLINK_isIncluded Arg1:T24 Arg2:T23
168  R29 TLINK_before Arg1:T16 Arg2:T14
169  R30 TLINK_identity Arg1:T21 Arg2:T20
170  R31 TLINK_identity Arg1:T22 Arg2:T21
171  R32 TLINK_simultaneous Arg1:T20 Arg2:T19
172  T25 Participant 115 134 Um jovem de 17 anos
173  A115  Lexical_Head T25 Noun
174  A116  Individuation_Domain T25 Individual
175  A117  Participant_Type_Domain T25 Per
176  A118  Involvement T25 1
177  T26 Participant 193 198 peito
178  A119  Lexical_Head T26 Noun
179  A120  Individuation_Domain T26 Individual
180  A121  Participant_Type_Domain T26 Other
181  A122  Involvement T26 1
182  R33 OLINK_partOf Arg1:T26 Arg2:T25
183  T27 Participant 220 240 bairro Casal da Mira
184  A123  Lexical_Head T27 Noun
185  A124  Individuation_Domain T27 Individual
186  A125  Participant_Type_Domain T27 Pl_civil
187  A126  Involvement T27 1
188  T28 Participant 245 252 Amadora
189  A127  Lexical_Head T28 Noun
190  A128  Individuation_Domain T28 Individual
191  A129  Participant_Type_Domain T28 Pl_civil
192  A130  Involvement T28 1
193  T29 Participant 257 275 distrito de Lisboa
194  A131  Lexical_Head T29 Noun
195  A132  Individuation_Domain T29 Individual
196  A133  Participant_Type_Domain T29 Pl_civil
197  A134  Involvement T29 1
198  R34 OLINK_partOf Arg1:T27 Arg2:T28
```

```
199  R35 OLINK_partOf Arg1:T28 Arg2:T29
200  T30 Participant 285 297 agência Lusa
201  A135  Lexical_Head T30 Noun
202  A136  Individuation_Domain T30 Individual
203  A137  Participant_Type_Domain T30 Org
204  A138  Involvement T30 1
205  T31 Participant 298 303 fonte
206  A139  Lexical_Head T31 Noun
207  A140  Individuation_Domain T31 Individual
208  A141  Participant_Type_Domain T31 Per
209  A142  Involvement T31 1
210  T32 Participant 305 310 a PSP
211  A143  Lexical_Head T32 Noun
212  A144  Individuation_Domain T32 Individual
213  A145  Participant_Type_Domain T32 Org
214  A146  Involvement T32 1
215  R36 OLINK_partOf Arg1:T31 Arg2:T32
216  T33 Participant 326 331 fonte
217  A147  Lexical_Head T33 Noun
218  A148  Individuation_Domain T33 Individual
219  A149  Participant_Type_Domain T33 Per
220  A150  Involvement T33 1
221  T34 Participant 333 377 o Comando Metropolitano de Lisboa (Cometlis)
222  A151  Lexical_Head T34 Noun
223  A152  Individuation_Domain T34 Individual
224  A153  Participant_Type_Domain T34 Org
225  A154  Involvement T34 1
226  R37 OLINK_partOf Arg1:T33 Arg2:T34
227  R38 OLINK_objIdentity Arg1:T33 Arg2:T31
228  R39 OLINK_partOf Arg1:T34 Arg2:T32
229  T35 Participant 426 445 Um jovem de 17 anos
230  A155  Lexical_Head T35 Noun
231  A156  Individuation_Domain T35 Individual
232  A157  Participant_Type_Domain T35 Per
233  A158  Involvement T35 1
234  R40 OLINK_objIdentity Arg1:T35 Arg2:T25
235  T36 Participant 462 478 uma arma de fogo
236  A159  Lexical_Head T36 Noun
237  A160  Individuation_Domain T36 Individual
238  A161  Participant_Type_Domain T36 Obj
239  A162  Involvement T36 1
240  T37 Participant 539 600 o Hospital Professor Doutor Fernando Fonseca [Amadora-
         Sintra]
241  A163  Lexical_Head T37 Noun
242  A164  Individuation_Domain T37 Individual
243  A165  Participant_Type_Domain T37 Loc
244  A166  Involvement T37 1
245  T38 Participant 614 618 Lusa
246  A167  Lexical_Head T38 Noun
```

```
247  A168   Individuation_Domain T38 Individual
248  A169   Participant_Type_Domain T38 Org
249  A170   Involvement T38 1
250  R41 OLINK_objIdentity Arg1:T38 Arg2:T30
251  T39 Participant 620 630 o Cometlis
252  A171   Lexical_Head T39 Noun
253  A172   Individuation_Domain T39 Individual
254  A173   Participant_Type_Domain T39 Org
255  A174   Involvement T39 1
256  R42 OLINK_objIdentity Arg1:T39 Arg2:T34
257  T40 Participant 736 755 quaisquer suspeitos
258  A175   Lexical_Head T40 Noun
259  A176   Individuation_Domain T40 Individual
260  A177   Participant_Type_Domain T40 Per
261  A178   Involvement T40 0
262  T41 Participant 806 829 Polícia Judiciária (PJ)
263  A179   Lexical_Head T41 Noun
264  A180   Individuation_Domain T41 Individual
265  A181   Participant_Type_Domain T41 Org
266  A182   Involvement T41 1
267  R43 OLINK_refDisjunct Arg1:T41 Arg2:T39
268  R44 SRLINK_pivot Arg1:T1 Arg2:T25
269  R45 SRLINK_reason Arg1:T1 Arg2:T2
270  R46 SRLINK_patient Arg1:T2 Arg2:T25
271  R47 SRLINK_location Arg1:T2 Arg2:T26
272  R48 SRLINK_location Arg1:T2 Arg2:T27
273  R49 SRLINK_location Arg1:T2 Arg2:T28
274  R50 SRLINK_location Arg1:T2 Arg2:T29
275  R51 SRLINK_goal Arg1:T4 Arg2:T30
276  R52 SRLINK_agent Arg1:T4 Arg2:T31
277  R53 SRLINK_theme Arg1:T5 Arg2:T6
278  R54 SRLINK_theme Arg1:T7 Arg2:T5
279  R55 SRLINK_patient Arg1:T9 Arg2:T35
280  R56 SRLINK_instrument Arg1:T9 Arg2:T36
281  R57 SRLINK_patient Arg1:T10 Arg2:T35
282  R58 SRLINK_manner Arg1:T10 Arg2:T11
283  R59 SRLINK_patient Arg1:T12 Arg2:T35
284  R60 SRLINK_finalLocation Arg1:T12 Arg2:T37
285  R61 SRLINK_agent Arg1:T13 Arg2:T33
286  R62 SRLINK_goal Arg1:T14 Arg2:T38
287  R63 SRLINK_agent Arg1:T14 Arg2:T39
288  R64 SRLINK_theme Arg1:T17 Arg2:T15
289  R65 SRLINK_theme Arg1:T15 Arg2:T16
290  R66 SRLINK_agent Arg1:T18 Arg2:T39
291  R67 SRLINK_theme Arg1:T18 Arg2:T40
292  R68 SRLINK_theme Arg1:T19 Arg2:T40
293  R69 SRLINK_agent Arg1:T19 Arg2:T39
294  R70 SRLINK_reason Arg1:T22 Arg2:T20
295  R71 SRLINK_theme Arg1:T20 Arg2:T21
```

```
296  R72 SRLINK_theme Arg1:T23 Arg2:T22
297  R73 SRLINK_goal Arg1:T23 Arg2:T41
298  R74 SRLINK_purpose Arg1:T23 Arg2:T24
299  T42 Spatial_Relation 190 192   no
300  A183  Topological T42 Disjunction_TTP_NTTP
301  R75 QSLINK_figure Arg1:T42 Arg2:T2
302  R76 QSLINK_ground Arg1:T42 Arg2:T26
303  T43 Spatial_Relation 242 244   na
304  A184  Topological T43 Disjunction_TTP_NTTP
305  R77 QSLINK_figure Arg1:T43 Arg2:T27
306  R78 QSLINK_ground Arg1:T43 Arg2:T28
307  T44 Spatial_Relation 254 256   no
308  A185  Topological T44 Disjunction_TTP_NTTP
309  R79 QSLINK_ground Arg1:T44 Arg2:T29
310  R80 QSLINK_figure Arg1:T44 Arg2:T28
311  T45 Participant 360 366 Lisboa
312  A186  Lexical_Head T45 Noun
313  A187  Individuation_Domain T45 Individual
314  A188  Participant_Type_Domain T45 Pl_civil
315  A189  Involvement T45 1
316  T46 Participant 333 356 o Comando Metropolitano
317  A190  Lexical_Head T46 Noun
318  A191  Individuation_Domain T46 Individual
319  A192  Participant_Type_Domain T46 Org
320  A193  Involvement T46 1
321  T47 Spatial_Relation 357 359   de
322  A194  Topological T47 Disjunction_TTP_NTTP
323  R81 QSLINK_figure Arg1:T47 Arg2:T46
324  R82 QSLINK_ground Arg1:T47 Arg2:T45
325  T48 Spatial_Relation 534 538   para
326  A195  PathDefining T48 End
327  R83 MOVELINK_spatialRelation Arg1:T12 Arg2:T48
328  R84 MOVELINK_targetSpatialRelation Arg1:T12 Arg2:T37
329  T49 Event 315 321 acordo
330  A196  Class T49 Reporting
331  A197  Event_Type T49 Transition
332  A198  Pos T49 Noun
333  A199  Polarity T49 Pos
334  R20 SRLINK_agent Arg1:T49 Arg2:T33
335  R85 TLINK_identity Arg1:T49 Arg2:T4
336  R86 TLINK_identity Arg1:T13 Arg2:T49
337  R87 TLINK_identity Arg1:T14 Arg2:T13
```

Listing A.6: Lusa 120 BRAT annotated file from the Text2Story Lusa dataset.

Appendix B

Input Data

This appendix will showcase the JSON schemas that we had to work with during the development phase. In Section B.1 we show the original NewsLines JSON Schema, in Section B.1 we provide an example of a JSON file with the previous schema, in Section B.3 we show the modified JSON schema, and in Section B.4 we provide an example of a file following that schema.

Additionally, Section B.5 and Section B.6 showcase the sentence-based and the event-based JSON representations of the Lusa 112 file. Lastly, Section B.7 and Section B.8 show the sentence-based and the event-based JSON representations of the Lusa 120 file.

B.1   NewsLines JSON Schema

Listing B.1 shows the original NewsLines JSON schema.

```
1  {
2    "$schema": "https://json-schema.org/draft/2020-12/schema",
3    "title": "News story",
4    "description": "Main story elements necessary to the creation of a storyline
          visualization",
5    "type": "object",
6    "properties": {
7      "title": {
8        "description": "The title of the story",
9        "type": "string"
10     },
11     "characters": {
12       "description": "Entities who participate in the story",
13       "type": "array",
14       "items": {"$ref" : "#/$defs/character"}
15     },
16     "scenes": {
17       "description": "Events in a story",
18       "type": "array",
19       "items": {"$ref" : "#/$defs/scene"}
```

114

```
20        }
21      },
22      "$defs": {
23        "character": {
24          "type": "object",
25          "required": [ "id", "name" ],
26          "properties": {
27            "id": {
28              "type": "string",
29              "description": "The ID of the entity"
30            },
31            "name": {
32              "type": "string",
33              "description": "The name of the entity"
34            },
35            "affiliation": {
36              "type": "string",
37              "description": "RGB value of the entity line in the visualization in the
                    format 'rbg(R, G, B)'"
38            },
39            "synonyms": {
40              "type": "array",
41              "descriptions": "Other names by which this entity can be referred to in
                    the description texts",
42              "items": {"type" : "string"}
43            }
44          }
45        },
46        "scene": {
47          "type": "object",
48          "required": [ "characters", "description", "title" ],
49          "properties": {
50            "characters": {
51              "type": "array",
52              "description": "ID's of the entities that participate in this event",
53              "items": { "type": "string"}
54            },
55            "description": {
56              "type": "string",
57              "description": "Description of this event"
58            },
59            "title": {
60              "type": "string",
61              "description": "Title of this event"
62            },
63            "date": {
64              "type": "string",
65              "descriptions": "Date of this event. Must be in a valid date format"
66            },
```

```
67        "location": {
68          "type": "string",
69          "descriptions": "Location of this event"
70        }
71      }
72    }
73  },
74  "required": [ "title", "characters", "scenes" ]
75 }
```

Listing B.1: Original NewsLines JSON schema

B.2 NewsLines JSON File Example

Listing B.2 depicts an example of a JSON file using the original NewsLines [22] schema.

```
1  {
2    "title": "Capitol Riot",
3    "characters": [
4      {
5        "id": "rioters",
6        "name": "Rioters",
7        "affiliation": "rgb(204, 0, 0)",
8        "synonyms": ["proud boys", "large crowd", "mob", "ashli babbit"]
9      },
10     {
11       "id": "trump",
12       "name": "Trump",
13       "affiliation": "rgb(222, 103, 38)"
14     },
15     {
16       "id": "congress",
17       "name": "Congress",
18       "affiliation": "rgb(110, 81, 160)",
19       "synonyms": ["nancy pelosi", "house republicans", "ted cruz", "house", "
             senate", "mike pence", "james lankford", "pelosi", "representatives"]
20     },
21     {
22       "id": "law-enforcement",
23       "name": "Law enforcement",
24       "affiliation": "rgb(80, 115, 163)",
25       "synonyms": ["capitol police", "police", "national guard", "plainclothes
             officers", "officer", "sergeant-at-arms", "f.b.i.", "swat"]
26     }
27   ],
28   "scenes": [
29     {
```

```
30       "characters": ["rioters"],
31       "description": "Members of the Proud Boys, a far-right group, join protesters
             , who had been assembled on the Capitol lawn since 10 a.m.",
32       "title": "Proud Boys",
33       "date": "2021-01-06 11:37:00",
34       "location": "Capitol"
35     },
36     {
37       "characters": ["trump"],
38       "description": "President Trump begins speaking at a rally near the White
             House, about a mile away from the Capitol.",
39       "title": "Rally starts",
40       "date": "2021-01-06 12:03:00",
41       "location": "Rally"
42     },
43     {
44       "characters": ["trump"],
45       "description": "Mr. Trump tells rallygoers to walk down to the Capitol. \"
             After this, w e re  going to walk down, and I'll be there with you.\"",
46       "title": "The march begins",
47       "date": "2021-01-06 12:04:00",
48       "location": "Rally"
49     },
50     {
51       "characters": ["rioters"],
52       "description": "A video shows a large crowd walking from the rally down
             Constitution Avenue to the Capitol.",
53       "title": "Large crowd in motion",
54       "date": "2021-01-06 12:29:00",
55       "location": "Constitution Avenue"
56     },
57     {
58       "characters": ["rioters", "law-enforcement"],
59       "description": "Rioters violently overwhelm the police and breach the
             C a p i t o l s outermost barricade. They break through three more barriers,
             forcing officers back to the Capitol steps, where they now face off.",
60       "title": "Violence begins",
61       "date": "2021-01-06 12:53:00",
62       "location": "Capitol"
63     },
64     {
65       "characters": ["law-enforcement"],
66       "description": "Capitol Police in riot gear are seen arriving to reinforce
             the line of officers on the Capitol steps.",
67       "title": "Backup",
68       "date": "2021-01-06 12:58:00",
69       "location": "Capitol"
70     },
71     {
```

```
72      "characters": ["congress"],
73      "description": "Speaker Nancy Pelosi begins the joint session of Congress,
            where both chambers meet to certify the presidential vote.",
74      "title": "Session in order",
75      "date": "2021-01-06 13:03:00",
76      "location": "Capitol"
77    },
78    {
79      "characters": ["trump"],
80      "description": "Trump ends his rally speech.",
81      "title": "End of rally",
82      "date": "2021-01-06 13:12:00"
83    },
84    {
85      "characters": ["congress"],
86      "description": "Several House Republicans, backed by Senator Ted Cruz of
            Texas, object to certifying A r i z o n a s vote, sending the House and
            Senate to debate in separate chambers.",
87      "title": "Objection",
88      "date": "2021-01-06 13:13:00"
89    },
90    {
91      "characters": ["rioters", "law-enforcement"],
92      "description": "By the Capitol steps, rioters continue to clash with the
            police. By now, reinforcements from local police have arrived to help.
            Both sides spray chemical agents.",
93      "title": "Violence escalates",
94      "date": "2021-01-06 13:15:00"
95    },
96    {
97      "characters": ["congress"],
98      "description": "The Senate and House debates begin in their respective
            chambers.",
99      "title": "Debates begin",
100     "date": "2021-01-06 13:30:00"
101   },
102   {
103     "characters": ["law-enforcement"],
104     "description": "Capitol Police requests National Guard assistance as rioters
            outside tear through scaffolding.",
105     "title": "Help necessary",
106     "date": "2021-01-06 13:49:00"
107   },
108   {
109     "characters": ["rioters"],
110     "description": "Rioters make it past two barriers on the east side of the
            Capitol and can now approach the doors of the building.",
111     "title": "Closer to the building",
112     "date": "2021-01-06 13:58:00"
```

```
113      },
114      {
115        "characters": ["rioters"],
116        "description": "Another mob breaches the final barricade on the buildings
              west side and approaches an entrance near the Senate chamber.",
117        "title": "Final barricade",
118        "date": "2021-01-06 14:10:00"
119      },
120      {
121        "characters": ["rioters"],
122        "description": "The first rioters make it inside the Capitol building.",
123        "title": "Break-in",
124        "date": "2021-01-06 14:11:00"
125      },
126      {
127        "characters": ["congress"],
128        "description": "As Senators continue to debate, Vice President Mike Pence is
              ushered off the Senate floor.",
129        "title": "Mike Pence is ushered off",
130        "date": "2021-01-06 14:12:00"
131      },
132      {
133        "characters": ["congress"],
134        "description": "Senator James Lankford, Republican of Oklahoma, is
              interrupted on the floor of the Senate, which is called into recess. A
              minute later, a mob arrives steps from a door to the Senate chamber.",
135        "title": "Mob arrives",
136        "date": "2021-01-06 14:13:00"
137      },
138      {
139        "characters": ["congress"],
140        "description": "In the House Chamber on the south side of the building, Ms.
              Pelosi is ushered off the floor. The debate in the House continues.",
141        "title": "Nancy Pelosi is ushered off",
142        "date": "2021-01-06 14:15:00"
143      },
144      {
145        "characters": ["congress"],
146        "description": "The House goes into recess, and representatives remain in the
               chamber.",
147        "title": "Recess",
148        "date": "2021-01-06 14:18:00"
149      },
150      {
151        "characters": ["trump"],
152        "description": "Mr. Trump criticizes Mr. Pence on Twitter.\"Mike Pence
              didnt have the courage to do what should have been done to protect our
              Country and our Constitution, giving States a chance to certify a
```

```
                  corrected set of facts, not the fraudulent or inaccurate ones which they
                  were asked to previously certify. USA demands the truth!\"",
153       "title": "Trump criticizes Pence",
154       "date": "2021-01-06 14:24:00"
155     },
156     {
157       "characters": ["rioters"],
158       "description": "Rioters breach a second entry into the building on the east
                  side. Mr. Trump calls Senator Mike Lee, Republican of Utah, though he
                  means to call Senator Tommy Tuberville, Republican of Alabama. Mr. Lee
                  hands the phone to Mr. Tuberville, who tells the president that Mr. Pence
                   had been escorted out of the chamber. Some time in the next 30 minutes,
                  Representative Kevin McCarthy, the minority leader, calls Mr. Trump
                  asking him to call off the riot, says another House representative Mr.
                  McCarthy tells about the call.",
159       "title": "Rioters breach second entry",
160       "date": "2021-01-06 14:26:00"
161     },
162     {
163       "characters": ["trump", "congress"],
164       "description": "Mr. Trump calls Senator Mike Lee, Republican of Utah, though
                  he means to call Senator Tommy Tuberville, Republican of Alabama. Mr. Lee
                   hands the phone to Mr. Tuberville, who tells the president that Mr.
                  Pence had been escorted out of the chamber. Some time in the next 30
                  minutes, Representative Kevin McCarthy, the minority leader, calls Mr.
                  Trump asking him to call off the riot, says another House representative
                  Mr. McCarthy tells about the call.",
165       "title": "Call",
166       "date": "2021-01-06 14:27:00"
167     },
168     {
169       "characters": ["congress"],
170       "description": "The House goes back into session, even with the mob in the
                  building.",
171       "title": "Back into session",
172       "date": "2021-01-06 14:27:00"
173     },
174     {
175       "characters": ["congress"],
176       "description": "The House is called into a final recess. Someone yells, \"Sit
                  down!\"",
177       "title": "\"Sit down!\"",
178       "date": "2021-01-06 14:30:00"
179     },
180     {
181       "characters": ["congress"],
182       "description": "Senators are evacuated from the Senate chamber as House
                  members remain in theirs.",
183       "title": "Senate evacuation",
```

```
184          "date": "2021-01-06 14:31:00"
185        },
186        {
187          "characters": ["rioters", "law-enforcement"],
188          "description": "The first rioters reach a rear door of the House chamber,
                 where members remain inside. Plainclothes officers train their guns on
                 the door.",
189          "title": "Closer to the House",
190          "date": "2021-01-06 14:35:00"
191        },
192        {
193          "characters": ["congress"],
194          "description": "Representatives begin evacuating the House chamber.",
195          "title": "House evacuation",
196          "date": "2021-01-06 14:39:00"
197        },
198        {
199          "characters": ["rioters"],
200          "description": "Rioters reach the doors of the S p e a k e r s Lobby.",
201          "title": "Speaker's Lobby",
202          "date": "2021-01-06 14:42:00"
203        },
204        {
205          "characters": ["rioters", "law-enforcement"],
206          "description": "Ashli Babbitt is shot and killed by an officer as she
                 attempts to break into the S p e a k e r s Lobby. The last representatives
                 are still being evacuated.",
207          "title": "Rioter killed",
208          "date": "2021-01-06 14:44:00"
209        },
210        {
211          "characters": ["rioters"],
212          "description": "Rioters make it to the floor of the Senate chamber, where
                 they are seen sitting at the dais and hanging from the balcony.",
213          "title": "In the Senate chamber",
214          "date": "2021-01-06 14:47:00"
215        },
216        {
217          "characters": ["law-enforcement"],
218          "description": "The D.C. National Guard is activated, but its members would
                 not arrive at the Capitol for more than two hours.",
219          "title": "National Guard activated",
220          "date": "2021-01-06 15:04:00"
221        },
222        {
223          "characters": ["rioters"],
224          "description": "Rioters now breach a third door at the b u i l d i n g s southeast
                 side.",
225          "title": "Third door breached",
```

```
226        "date": "2021-01-06 15:16:00"
227      },
228      {
229        "characters": ["law-enforcement"],
230        "description": "An F.B.I. SWAT team is seen at the building.",
231        "title": "F.B.I",
232        "date": "2021-01-06 16:14:00"
233      },
234      {
235        "characters": ["trump"],
236        "description": "Mr. Trump tweets a video in which he tells his supporters at
               the Capitol that he loves them.\"It was a landslide election. And
               everyone knows it. Especially the other side. But you have to go home.
                   T h e r e s  never been a time like this when such a thing happened when
                they could take it away from all of us. From me, from you, from our
               country. This was a fraudulent election.    Go home. We love you. You're
                very special.\"",
237        "title": "F.B.I",
238        "date": "2021-01-06 16:17:00"
239      },
240      {
241        "characters": ["law-enforcement"],
242        "description": "The sergeant-at-arms, who is responsible for the security of
               Congress, tells lawmakers and reporters that the Capitol is now secure.
               Meanwhile, members of Congress continue to shelter in largely undisclosed
                locations.",
243        "title": "Capitol secure",
244        "date": "2021-01-06 17:34:00"
245      },
246      {
247        "characters": ["law-enforcement"],
248        "description": "The D.C. National Guard arrives at the Capitol, swears in
               with the Capitol Police and begins support operations.",
249        "title": "National Guard arrives",
250        "date": "2021-01-06 17:40:00"
251      },
252      {
253        "characters": ["law-enforcement"],
254        "description": "Curfew begins in Washington, D.C. Officers arrest many who
               defy the order.",
255        "title": "Curfew",
256        "date": "2021-01-06 18:00:00"
257      },
258      {
259        "characters": ["trump"],
260        "description": "In a tweet, Mr. Trump once again says that the election was
               stolen.\"These are the things and events that happen when a sacred
               landslide election victory is so unceremoniously & viciously stripped
```

```
              away from great patriots who have been badly & unfairly treated for so
              long. Go home with love & in peace. Remember this day forever!\"",
261        "title": "More tweets",
262        "date": "2021-01-06 18:01:00"
263      },
264      {
265        "characters": ["law-enforcement"],
266        "description": "Capitol Police, local police and the National Guard establish
              a perimeter on the west side of the Capitol.",
267        "title": "Perimeter established",
268        "date": "2021-01-06 18:14:00"
269      },
270      {
271        "characters": ["congress"],
272        "description": "The Senate resumes debate over the certification of
              A r i z o n a s vote in the Senate chamber. The House would reconvene about
              an hour later.",
273        "title": "Debates resumed",
274        "date": "2021-01-06 20:06:00"
275      }
276    ]
277  }
```

Listing B.2: CapitolRiot.json - a JSON file using the original NewsLines [22] JSON schema

B.3   Modified JSON Schema

Listing B.3 shows the modified JSON schema used by our implementation.

```
1  {
2      "$schema": "https://json-schema.org/draft/2020-12/schema",
3      "title": "News story",
4      "description": "Main story elements necessary to the creation of a storyline
           visualization",
5      "type": "object",
6      "properties": {
7        "title": {
8          "description": "The title of the story",
9          "type": "string"
10       },
11       "characters": {
12         "description": "Entities who participate in the story",
13         "type": "array",
14              "items": {"$ref" : "#/$defs/character"}
15       },
16       "locations": {
17         "description": "Locations which appear in the story",
```

```
18        "type": "array",
19            "items": {"$ref" : "#/$defs/location"}
20      },
21      "dates": {
22        "description": "Dates which appear in the story",
23        "type": "array",
24            "items": {"$ref" : "#/$defs/date"}
25      },
26      "scenes": {
27        "description": "Events in a story",
28        "type": "array",
29            "items": {"$ref" : "#/$defs/scene"}
30      }
31    },
32      "$defs": {
33      "character": {
34        "type": "object",
35        "required": [ "id", "name" ],
36        "properties": {
37            "id": {
38                "type": "string",
39                "description": "The ID of the entity"
40            },
41            "name": {
42                "type": "string",
43                "description": "The name of the entity"
44            },
45            "affiliation": {
46                "type": "string",
47                "description": "RGB value of the entity line in the visualization
                      in the format 'rbg(R, G, B)'"
48            },
49            "synonyms": {
50                "type": "array",
51                "descriptions": "Other names by which this entity can be referred
                      to in the description texts",
52                "items": {"type" : "string"}
53            }
54        }
55      },
56      "location": {
57        "type": "object",
58        "required": [ "id", "value" ],
59        "properties": {
60            "id": {
61                "type": "string",
62                "description": "The ID of the entity"
63            },
64            "value": {
```

```
65                  "type": "string",
66                  "description": "The name of the entity"
67              },
68          "Part_Of": {
69                  "type": "string",
70                  "description": "The ID of the entity this location is a part of,
                        through a OLINK_PartOf relation"
71              },
72          "Parts": {
73                  "type": "string",
74                  "description": "The ID of the entity that is a direct part of this
                        location, through a OLINK_PartOf relation"
75              },
76          "synonyms": {
77                  "type": "string",
78                  "descriptions": "The ID of the entity that is a direct synonym of
                        this location, through a OLINK_Identity relation",
79              }
80          }
81      },
82      "date": {
83        "type": "object",
84        "required": [ "id", "time" ],
85        "properties": {
86            "id": {
87                  "type": "string",
88                  "description": "The ID of the entity"
89              },
90          "time": {
91                  "type": "string",
92                  "description": "The value of the entity."
93              },
94          "TemporalFunction": {
95                  "type": "string",
96                  "description": "The type of temporal anchor the entity represents
                        for other temporal expressions in the document"
97              },
98          "value": {
99                  "type": "string",
100                 "descriptions": "The date in a valid date format"
101             }
102         }
103     },
104       "scene": {
105       "type": "object",
106       "required": [ "characters", "description", "title" ],
107       "properties": {
108           "characters": {
109                 "type": "array",
```

```
110              "description": "ID's of the entities that participate in this event
                    ",
111              "items": { "type": "string"}
112          },
113          "description": {
114              "type": "string",
115              "description": "Description of this event"
116          },
117          "title": {
118              "type": "string",
119              "description": "Title of this event"
120          },
121          "date": {
122              "type": "array",
123              "descriptions": "IDs of the dates of this event.",
124              "items": { "type": "string"}
125          },
126          "location": {
127              "type": "array",
128              "descriptions": "IDs of the locations of this event",
129              "items": { "type": "string"}
130          }
131        }
132      }
133    },
134    "required": [ "title", "characters", "scenes" ]
135  }
```

Listing B.3: Modified NewsLines [22] JSON schema

B.4   JSON File Example Following the Modified Schema

Listing B.4 depicts an example of a JSON file using the modified schema from our implementation.

```
1  {
2      "title": "lusa_0",
3      "characters": [
4          {
5              "id": "T24",
6              "name": "A GNR",
7              "affiliation": "rgb(127,63,142)",
8              "synonyms": [
9                  "GNR",
10                 "o Comando Territorial",
11                 "militares",
12                 "Os militares"
13             ]
```

```
14            },
15            {
16                "id": "T27",
17                "name": "cerca de 50 pessoas",
18                "affiliation": "rgb(142,27,56)",
19                "synonyms": [
20                    "as pessoas"
21                ]
22            },
23            {
24                "id": "T40",
25                "name": "o proprietário",
26                "affiliation": "rgb(0,101,193)"
27            }
28        ],
29        "locations": [
30            {
31                "value": "bar",
32                "id": "T42"
33            },
34            {
35                "value": "Corotelo",
36                "Part_Of": "T23",
37                "id": "T25"
38            },
39            {
40                "value": "concelho de São Brás de Alportel",
41                "id": "T23"
42            },
43            {
44                "value": "Faro",
45                "id": "T46"
46            },
47            {
48                "value": "estabelecimento de bebidas",
49                "id": "T32"
50            },
51            {
52                "value": "Posto Territorial de São de Alportel",
53                "Part_Of": "T49",
54                "id": "T34"
55            },
56            {
57                "value": "São de Alportel",
58                "id": "T49"
59            },
60            {
61                "value": "local",
62                "synonyms": "T32",
```

```
63              "id": "T35"
64          }
65      ],
66      "dates": [
67          {
68              "time": "11 out 2020",
69              "TemporalFunction": "Publication_Time",
70              "id": "T43"
71          },
72          {
73              "time": "sábado",
74              "TemporalFunction": "Publication_Time",
75              "id": "T2"
76          },
77          {
78              "time": "hoje",
79              "TemporalFunction": "Publication_Time",
80              "id": "T5"
81          }
82      ],
83      "scenes": [
84          {
85              "characters": [
86                  "T24"
87              ],
88              "description": "A GNR acabou no sábado com uma festa ilegal num bar em
                        Corotelo, concelho de São Brás de Alportel, que reunia cerca de 50
                        pessoas, foi hoje anunciado.",
89              "title": "A GNR acabou festa ilegal",
90              "location": [
91                  "T42",
92                  "T25"
93              ],
94              "date": [
95                  "T2"
96              ]
97          },
98          {
99              "characters": [
100                  "T27"
101              ],
102              "description": "A GNR acabou no sábado com uma festa ilegal num bar em
                        Corotelo, concelho de São Brás de Alportel, que reunia cerca de 50
                        pessoas, foi hoje anunciado.",
103              "title": "festa ilegal reunia que cerca de 50 pessoas",
104              "location": [
105                  "T42",
106                  "T25"
107              ],
```

```
108             "date": []
109         },
110         {
111             "characters": [
112                 "T24"
113             ],
114             "description": "Em comunicado, o Comando Territorial de Faro da GNR
                    adianta que, na sequência de uma denúncia a informar sobre a existê
                    ncia de 'uma festa com música ao vivo num estabelecimento de
                    bebidas', militares do Posto Territorial de São de Alportel
                    deslocaram-se ao local para fiscalizar a ocorrência.",
115             "title": "o Comando Territorial adianta denúncia existência festa ocorr
                    ência",
116             "location": [
117                 "T46",
118                 "T32"
119             ],
120             "date": []
121         },
122         {
123             "characters": [],
124             "description": "Em comunicado, o Comando Territorial de Faro da GNR
                    adianta que, na sequência de uma denúncia a informar sobre a existê
                    ncia de 'uma festa com música ao vivo num estabelecimento de
                    bebidas', militares do Posto Territorial de São de Alportel
                    deslocaram-se ao local para fiscalizar a ocorrência.",
125             "title": "festa existência",
126             "location": [
127                 "T32"
128             ],
129             "date": []
130         },
131         {
132             "characters": [
133                 "T24"
134             ],
135             "description": "Em comunicado, o Comando Territorial de Faro da GNR
                    adianta que, na sequência de uma denúncia a informar sobre a existê
                    ncia de 'uma festa com música ao vivo num estabelecimento de
                    bebidas', militares do Posto Territorial de São de Alportel
                    deslocaram-se ao local para fiscalizar a ocorrência.",
136             "title": "militares deslocaram-se denúncia",
137             "location": [],
138             "date": []
139         },
140         {
141             "characters": [
142                 "T24"
143             ],
```

```
144            "description": "Em comunicado, o Comando Territorial de Faro da GNR
                   adianta que, na sequência de uma denúncia a informar sobre a existê
                   ncia de 'uma festa com música ao vivo num estabelecimento de
                   bebidas', militares do Posto Territorial de São de Alportel
                   deslocaram-se ao local para fiscalizar a ocorrência.",
145            "title": "militares fiscalizar ocorrência",
146            "location": [],
147            "date": []
148        },
149        {
150            "characters": [
151                "T24",
152                "T27"
153            ],
154            "description": "Os militares desmobilizaram as pessoas para '
                   cumprimento das normas referentes à pandemia de covid-19'.",
155            "title": "Os militares desmobilizaram as pessoas cumprimento normas
                   referentes à pandemia de covid-19",
156            "location": [],
157            "date": []
158        },
159        {
160            "characters": [
161                "T40"
162            ],
163            "description": "Na nota é ainda referido que o proprietário foi
                   identificado e foi elaborado 'um auto de contraordenação'.",
164            "title": "o proprietário identificado",
165            "location": [],
166            "date": []
167        }
168    ]
169 }
```

Listing B.4: Lusa 0

B.5 Lusa 112 Sentence-based JSON File

Listing B.5 depicts the Lusa 112 sentence-based JSON file.

```
1 {
2     "title": "lusa_112",
3     "characters": [
4         {
5             "id": "T4",
6             "name": "O deputado socialista Ricardo Pinheiro",
7             "affiliation": "rgb(96,203,27)",
```

```
 8              "synonyms": [
 9                  "o novo secretário de Estado",
10                  "Ricardo Pinheiro",
11                  "Ricardo Miguel Furtado Pinheiro"
12              ]
13          },
14          {
15              "id": "T24",
16              "name": "José Mendes",
17              "affiliation": "rgb(194,29,121)",
18              "synonyms": [
19                  "José Fernandes Gomes Mendes"
20              ]
21          },
22          {
23              "id": "T27",
24              "name": "O Presidente da República",
25              "affiliation": "rgb(52,142,64)"
26          },
27          {
28              "id": "T28",
29              "name": "o primeiro-ministro",
30              "affiliation": "rgb(125,40,164)"
31          },
32          {
33              "id": "T34",
34              "name": "o ministro do Planeamento",
35              "affiliation": "rgb(62,24,74)",
36              "synonyms": [
37                  "Nelson de Sousa",
38                  "o novo secretário de Estado",
39                  "Ricardo Pinheiro"
40              ]
41          },
42          {
43              "id": "T47",
44              "name": "Câmara",
45              "affiliation": "rgb(171,105,138)"
46          }
47      ],
48      "locations": [
49          {
50              "value": "o Palácio de Belém.",
51              "id": "T36"
52          },
53          {
54              "value": "distrito de Portalegre",
55              "id": "T40"
56          },
```

```
57          {
58              "value": "Campo Maior",
59              "id": "T46"
60          }
61      ],
62      "dates": [
63          {
64              "time": "hoje",
65              "TemporalFunction": "Publication_Time",
66              "id": "T3"
67          },
68          {
69              "time": "hoje",
70              "TemporalFunction": "Publication_Time",
71              "id": "T17"
72          },
73          {
74              "time": "20:30",
75              "id": "T18"
76          },
77          {
78              "time": "06 nov 2020",
79              "TemporalFunction": "Publication_Time",
80              "id": "T44"
81          }
82      ],
83      "scenes": [
84          {
85              "characters": [
86                  "T4",
87                  "T24",
88                  "T34"
89              ],
90              "description": "O deputado socialista Ricardo Pinheiro vai substituir
                    José Mendes no cargo de secretário de Estado do Planeamento, mudanç
                    a que foi hoje anunciada através de uma nota divulgada no portal da
                     Presidência da República.",
91              "title": "O deputado socialista Ricardo Pinheiro vai substituir José
                    Mendes no cargo de secretário de Estado do Planeamento, mudança que
                     foi hoje anunciada através de uma nota divulgada no portal da
                    Presidência da República.",
92              "date": [
93                  "T3"
94              ],
95              "location": []
96          },
97          {
98              "characters": [
99                  "T4",
```

```
100                    "T24",
101                    "T27",
102                    "T28"
103                ],
104            "description": "'O Presidente da República aceitou a proposta do
                      primeiro-ministro de exoneração, a seu pedido, de José Fernandes
                      Gomes Mendes, como secretário de Estado do Planeamento, e a nomeaçã
                      o, para o mesmo cargo, de Ricardo Miguel Furtado Pinheiro', lê-se
                      na nota.",
105            "title": "'O Presidente da República aceitou a proposta do primeiro-
                      ministro de exoneração, a seu pedido, de José Fernandes Gomes
                      Mendes, como secretário de Estado do Planeamento, e a nomeação,
                      para o mesmo cargo, de Ricardo Miguel Furtado Pinheiro', lê-se na
                      nota.",
106            "date": [],
107            "location": []
108        },
109        {
110            "characters": [
111                    "T4",
112                    "T34"
113                ],
114            "description": "A posse do novo secretário de Estado do ministro do
                      Planeamento, Nelson de Sousa, terá lugar ainda hoje, pelas 20:30,
                      no Palácio de Belém.",
115            "title": "A posse do novo secretário de Estado do ministro do
                      Planeamento, Nelson de Sousa, terá lugar ainda hoje, pelas 20:30,
                      no Palácio de Belém.",
116            "date": [
117                    "T17",
118                    "T18"
119                ],
120            "location": [
121                    "T36"
122                ]
123        },
124        {
125            "characters": [
126                    "T4",
127                    "T34",
128                    "T47"
129                ],
130            "description": "Ricardo Pinheiro foi presidente da Câmara de Campo
                      Maior, distrito de Portalegre, círculo pelo qual foi eleito
                      deputado nas últimas eleições legislativas.",
131            "title": "Ricardo Pinheiro foi presidente da Câmara de Campo Maior,
                      distrito de Portalegre, círculo pelo qual foi eleito deputado nas ú
                      ltimas eleições legislativas.",
132            "date": [],
```

```
133          "location": [
134              "T40",
135              "T46"
136          ]
137      }
138    ]
139 }
```

Listing B.5: Lusa 112 sentence-based JSON file.

B.6   Lusa 112 Event-based JSON File

Listing B.6 depicts the Lusa 112 event-based JSON file.

```
1  {
2      "title": "lusa_112",
3      "characters": [
4          {
5              "id": "T4",
6              "name": "O deputado socialista Ricardo Pinheiro",
7              "affiliation": "rgb(96,203,27)",
8              "synonyms": [
9                  "o novo secretário de Estado",
10                 "Ricardo Pinheiro",
11                 "Ricardo Miguel Furtado Pinheiro"
12             ]
13         },
14         {
15             "id": "T24",
16             "name": "José Mendes",
17             "affiliation": "rgb(194,29,121)",
18             "synonyms": [
19                 "José Fernandes Gomes Mendes"
20             ]
21         },
22         {
23             "id": "T27",
24             "name": "O Presidente da República",
25             "affiliation": "rgb(52,142,64)"
26         },
27         {
28             "id": "T28",
29             "name": "o primeiro-ministro",
30             "affiliation": "rgb(125,40,164)"
31         },
32         {
33             "id": "T34",
```

```
34              "name": "o ministro do Planeamento",
35              "affiliation": "rgb(62,24,74)",
36              "synonyms": [
37                  "Nelson de Sousa",
38                  "o novo secretário de Estado",
39                  "Ricardo Pinheiro"
40              ]
41          },
42          {
43              "id": "T47",
44              "name": "Câmara",
45              "affiliation": "rgb(171,105,138)"
46          }
47      ],
48      "locations": [
49          {
50              "value": "o Palácio de Belém.",
51              "id": "T36"
52          },
53          {
54              "value": "distrito de Portalegre",
55              "id": "T40"
56          },
57          {
58              "value": "Campo Maior",
59              "id": "T46"
60          }
61      ],
62      "dates": [
63          {
64              "time": "hoje",
65              "TemporalFunction": "Publication_Time",
66              "id": "T3"
67          },
68          {
69              "time": "hoje",
70              "TemporalFunction": "Publication_Time",
71              "id": "T17"
72          },
73          {
74              "time": "20:30",
75              "id": "T18"
76          },
77          {
78              "time": "06 nov 2020",
79              "TemporalFunction": "Publication_Time",
80              "id": "T44"
81          }
82      ],
```

```
83        "scenes": [
84            {
85                "characters": [
86                    "T4",
87                    "T24",
88                    "T34"
89                ],
90                "description": "O deputado socialista Ricardo Pinheiro vai substituir
                        José Mendes no cargo de secretário de Estado do Planeamento, mudanç
                        a que foi hoje anunciada através de uma nota divulgada no portal da
                         Presidência da República.",
91                "title": "O deputado socialista Ricardo Pinheiro substituir José Mendes
                        ",
92                "location": [],
93                "date": []
94            },
95            {
96                "characters": [
97                    "T27",
98                    "T28"
99                ],
100                "description": "'O Presidente da República aceitou a proposta do
                        primeiro-ministro de exoneração, a seu pedido, de José Fernandes
                        Gomes Mendes, como secretário de Estado do Planeamento, e a nomeaçã
                        o, para o mesmo cargo, de Ricardo Miguel Furtado Pinheiro', lê-se
                        na nota.",
101                "title": "O Presidente da República aceitou proposta o primeiro-
                        ministro exoneração",
102                "location": [],
103                "date": []
104            },
105            {
106                "characters": [
107                    "T24"
108                ],
109                "description": "'O Presidente da República aceitou a proposta do
                        primeiro-ministro de exoneração, a seu pedido, de José Fernandes
                        Gomes Mendes, como secretário de Estado do Planeamento, e a nomeaçã
                        o, para o mesmo cargo, de Ricardo Miguel Furtado Pinheiro', lê-se
                        na nota.",
110                "title": "José Fernandes Gomes Mendes exoneração secretário de Estado
                        do Planeamento",
111                "location": [],
112                "date": []
113            },
114            {
115                "characters": [
116                    "T28"
117                ],
```

```
118            "description": "'O Presidente da República aceitou a proposta do
                   primeiro-ministro de exoneração, a seu pedido, de José Fernandes
                   Gomes Mendes, como secretário de Estado do Planeamento, e a nomeaçã
                   o, para o mesmo cargo, de Ricardo Miguel Furtado Pinheiro', lê-se
                   na nota.",
119            "title": "o primeiro-ministro nomeação cargo",
120            "location": [],
121            "date": []
122        },
123        {
124            "characters": [
125                "T4"
126            ],
127            "description": "'O Presidente da República aceitou a proposta do
                   primeiro-ministro de exoneração, a seu pedido, de José Fernandes
                   Gomes Mendes, como secretário de Estado do Planeamento, e a nomeaçã
                   o, para o mesmo cargo, de Ricardo Miguel Furtado Pinheiro', lê-se
                   na nota.",
128            "title": "Ricardo Miguel Furtado Pinheiro nomeação",
129            "location": [],
130            "date": []
131        },
132        {
133            "characters": [
134                "T4",
135                "T34"
136            ],
137            "description": "A posse do novo secretário de Estado do ministro do
                   Planeamento, Nelson de Sousa, terá lugar ainda hoje, pelas 20:30,
                   no Palácio de Belém.",
138            "title": "o novo secretário de Estado posse",
139            "location": [],
140            "date": []
141        },
142        {
143            "characters": [
144                "T4",
145                "T34",
146                "T47"
147            ],
148            "description": "Ricardo Pinheiro foi presidente da Câmara de Campo
                   Maior, distrito de Portalegre, círculo pelo qual foi eleito
                   deputado nas últimas eleições legislativas.",
149            "title": "Ricardo Pinheiro presidente a Câmara de Campo Maior",
150            "location": [
151                "T40"
152            ],
153            "date": []
154        }
```

```
155     ]
156 }
```

Listing B.6: Lusa 112 event-based JSON file.

B.7 Lusa 120 Sentence-based JSON File

Listing B.7 depicts the Lusa 120 sentence-based JSON file.

```
1  {
2      "title": "lusa_120",
3      "characters": [
4          {
5              "id": "T25",
6              "name": "Um jovem de 17 anos",
7              "affiliation": "rgb(226,231,216)"
8          },
9          {
10             "id": "T30",
11             "name": "agência Lusa",
12             "affiliation": "rgb(250,128,245)",
13             "synonyms": [
14                 "Lusa"
15             ]
16         },
17         {
18             "id": "T32",
19             "name": "a PSP",
20             "affiliation": "rgb(45,167,83)",
21             "synonyms": [
22                 "fonte",
23                 "o Comando Metropolitano de Lisboa (Cometlis)",
24                 "o Cometlis"
25             ]
26         },
27         {
28             "id": "T40",
29             "name": "quaisquer suspeitos",
30             "affiliation": "rgb(65,116,84)"
31         },
32         {
33             "id": "T41",
34             "name": "Polícia Judiciária (PJ)",
35             "affiliation": "rgb(41,60,71)"
36         }
37     ],
38     "locations": [
```

```
39              {
40                  "value": "bairro Casal da Mira",
41                  "Part_Of": "T28",
42                  "id": "T27"
43              },
44              {
45                  "value": "Amadora",
46                  "Part_Of": "T29",
47                  "id": "T28"
48              },
49              {
50                  "value": "distrito de Lisboa",
51                  "id": "T29"
52              },
53              {
54                  "value": "o Hospital Professor Doutor Fernando Fonseca [Amadora-Sintra
                        ]",
55                  "id": "T37"
56              },
57              {
58                  "value": "Lisboa",
59                  "id": "T45"
60              }
61          ],
62          "dates": [
63              {
64                  "time": "hoje de madrugada",
65                  "TemporalFunction": "None",
66                  "id": "T3"
67              },
68              {
69                  "time": "00:53",
70                  "TemporalFunction": "None",
71                  "id": "T8"
72              }
73          ],
74          "scenes": [
75              {
76                  "characters": [
77                      "T25",
78                      "T30",
79                      "T32"
80                  ],
81                  "description": "Um jovem de 17 anos ficou em estado muito grave depois
                        de ter sido baleado no peito hoje de madrugada no bairro Casal da
                        Mira, na Amadora, no distrito de Lisboa, disse à agência Lusa fonte
                         da PSP.",
82                  "title": "Um jovem de 17 anos ficou em estado muito grave depois de ter
                        sido baleado no peito hoje de madrugada no bairro Casal da Mira,
```

```
                        na Amadora, no distrito de Lisboa, disse à agência Lusa fonte da
                        PSP.",
 83            "date": [
 84                "T3"
 85            ],
 86            "location": [
 87                "T27",
 88                "T28",
 89                "T29"
 90            ]
 91        },
 92        {
 93            "characters": [
 94                "T32"
 95            ],
 96            "description": "De acordo com fonte do Comando Metropolitano de Lisboa
                        (Cometlis), o alerta para a ocorrência foi dado às 00:53.",
 97            "title": "De acordo com fonte do Comando Metropolitano de Lisboa (
                        Cometlis), o alerta para a ocorrência foi dado às 00:53.",
 98            "date": [
 99                "T8"
100            ],
101            "location": [
102                "T45"
103            ]
104        },
105        {
106            "characters": [
107                "T25"
108            ],
109            "description": "'Um jovem de 17 anos foi baleado com uma arma de fogo
                        .",
110            "title": "'Um jovem de 17 anos foi baleado com uma arma de fogo.",
111            "date": [],
112            "location": []
113        },
114        {
115            "characters": [],
116            "description": "Ficou ferido em estado muito grave e foi transportado
                        para o Hospital Professor Doutor Fernando Fonseca [Amadora-Sintra
                        ]', referiu.",
117            "title": "Ficou ferido em estado muito grave e foi transportado para o
                        Hospital Professor Doutor Fernando Fonseca [Amadora-Sintra]',
                        referiu.",
118            "date": [],
119            "location": [
120                "T37"
121            ]
122        },
```

```
123            {
124                "characters": [
125                    "T30",
126                    "T32",
127                    "T40"
128                ],
129                "description": "À Lusa, o Cometlis adiantou que as causas para o
                       incidente são desconhecidas e que ainda não foram identificados ou
                       detidos quaisquer suspeitos.",
130                "title": "À Lusa, o Cometlis adiantou que as causas para o incidente sã
                       o desconhecidas e que ainda não foram identificados ou detidos
                       quaisquer suspeitos.",
131                "date": [],
132                "location": []
133            },
134            {
135                "characters": [
136                    "T41"
137                ],
138                "description": "Por se tratar de um crime, o caso foi entregue à Polí
                       cia Judiciária (PJ) para investigação.",
139                "title": "Por se tratar de um crime, o caso foi entregue à Polícia
                       Judiciária (PJ) para investigação.",
140                "date": [],
141                "location": []
142            }
143        ]
144 }
```

Listing B.7: Lusa 120 sentence-based JSON file.

B.8 Lusa 120 Event-based JSON File

Listing B.8 depicts the Lusa 120 event-based JSON file.

```
1  {
2      "title": "lusa_120",
3      "characters": [
4          {
5              "id": "T25",
6              "name": "Um jovem de 17 anos",
7              "affiliation": "rgb(226,231,216)"
8          },
9          {
10             "id": "T30",
11             "name": "agência Lusa",
12             "affiliation": "rgb(250,128,245)",
```

```
13              "synonyms": [
14                  "Lusa"
15              ]
16          },
17          {
18              "id": "T32",
19              "name": "a PSP",
20              "affiliation": "rgb(45,167,83)",
21              "synonyms": [
22                  "fonte",
23                  "o Comando Metropolitano de Lisboa (Cometlis)",
24                  "o Cometlis"
25              ]
26          },
27          {
28              "id": "T40",
29              "name": "quaisquer suspeitos",
30              "affiliation": "rgb(65,116,84)"
31          },
32          {
33              "id": "T41",
34              "name": "Polícia Judiciária (PJ)",
35              "affiliation": "rgb(41,60,71)"
36          }
37      ],
38      "locations": [
39          {
40              "value": "bairro Casal da Mira",
41              "Part_Of": "T28",
42              "id": "T27"
43          },
44          {
45              "value": "Amadora",
46              "Part_Of": "T29",
47              "id": "T28"
48          },
49          {
50              "value": "distrito de Lisboa",
51              "id": "T29"
52          },
53          {
54              "value": "o Hospital Professor Doutor Fernando Fonseca [Amadora-Sintra
                    ]",
55              "id": "T37"
56          },
57          {
58              "value": "Lisboa",
59              "id": "T45"
60          }
```

```
61        ],
62        "dates": [
63            {
64                "time": "hoje de madrugada",
65                "TemporalFunction": "None",
66                "id": "T3"
67            },
68            {
69                "time": "00:53",
70                "TemporalFunction": "None",
71                "id": "T8"
72            }
73        ],
74        "scenes": [
75            {
76                "characters": [
77                    "T25"
78                ],
79                "description": "Um jovem de 17 anos ficou em estado muito grave depois
                        de ter sido baleado no peito hoje de madrugada no bairro Casal da
                        Mira, na Amadora, no distrito de Lisboa, disse à agência Lusa fonte
                         da PSP.",
80                "title": "Um jovem de 17 anos em estado muito grave",
81                "location": [],
82                "date": []
83            },
84            {
85                "characters": [
86                    "T25"
87                ],
88                "description": "Um jovem de 17 anos ficou em estado muito grave depois
                        de ter sido baleado no peito hoje de madrugada no bairro Casal da
                        Mira, na Amadora, no distrito de Lisboa, disse à agência Lusa fonte
                         da PSP.",
89                "title": "Um jovem de 17 anos baleado",
90                "location": [
91                    "T27",
92                    "T28",
93                    "T29"
94                ],
95                "date": [
96                    "T3"
97                ]
98            },
99            {
100               "characters": [
101                   "T30",
102                   "T32"
103               ],
```

```
104              "description": "Um jovem de 17 anos ficou em estado muito grave depois
                      de ter sido baleado no peito hoje de madrugada no bairro Casal da
                      Mira, na Amadora, no distrito de Lisboa, disse à agência Lusa fonte
                       da PSP.",
105              "title": "fonte disse agência Lusa",
106              "location": [],
107              "date": []
108          },
109          {
110              "characters": [
111                  "T30",
112                  "T32"
113              ],
114              "description": "À Lusa, o Cometlis adiantou que as causas para o
                      incidente são desconhecidas e que ainda não foram identificados ou
                      detidos quaisquer suspeitos.",
115              "title": "o Cometlis adiantou Lusa",
116              "location": [],
117              "date": []
118          },
119          {
120              "characters": [
121                  "T32",
122                  "T40"
123              ],
124              "description": "À Lusa, o Cometlis adiantou que as causas para o
                      incidente são desconhecidas e que ainda não foram identificados ou
                      detidos quaisquer suspeitos.",
125              "title": "o Cometlis identificados quaisquer suspeitos",
126              "location": [],
127              "date": []
128          },
129          {
130              "characters": [
131                  "T32",
132                  "T40"
133              ],
134              "description": "À Lusa, o Cometlis adiantou que as causas para o
                      incidente são desconhecidas e que ainda não foram identificados ou
                      detidos quaisquer suspeitos.",
135              "title": "o Cometlis detidos quaisquer suspeitos",
136              "location": [],
137              "date": []
138          }
139      ]
140 }
```

Listing B.8: Lusa 120 event-based JSON file.

Appendix C

Installation

This appendix will present the installation instruction for the NewsLines [22] tool.

C.1 Instruction Manual

Listing C.1 show the local installation instructions for the tool. Additionally, there is also a live version available[1].

```
1  ## Local installation
2
3  Clone the repository:
4  ```
5  git clone https://github.com/catarina03/news-story-viz
6  ```
7
8  Use Python's `http.server` or any equivalent service to start a local server.
9  Run the following command in the root folder:
10  ```
11  python3 -m http.server
12  ```
```

Listing C.1: Instructions for local installation

---

[1]https://catarina03.github.io/news-story-viz-prototype/

Appendix D

Interviews

Section D.1 of this appendix presents the original script of the semi-structure interviews. Section D.2 showcases the visualization descriptions and sketches, Q3 and Q4 of the interviews, as seen in Table 7.1. Of the eight people interview, seven agreed to draw a visualization they envisioned.

D.1 Interviews Script

Figures D.1, D.2, D.3, D.4, D.5, and D.6 show the original interview script.

D.2 Interview Sketches

Figures D.7, D.8, D.9, D.10, D.11, D.12, and D.13 depict the interview sketches of the Text2Story team.

[Introdução]

Esta entrevista está a ser realizada no contexto da dissertação de mestrado "Visualization of Narrative Structures", que se engloba no projeto Text2Story.

O projeto tem como objetivo a criação de visualizações, de forma automática ou semi-automática, que enriqueçam e complementem os elementos narrativos encontrados no texto original.

Nesta entrevista pretendo recolher ideias para novas visualizações e sugestões de melhorias para visualizações já existentes.

[Contexto]

Primeiro vamos começar por definir o que é uma narrativa. Uma narrativa, ou uma história, é definida pelo Oxford English Dictionary como "um relato de uma série de acontecimentos, factos, etc., apresentados por ordem e com o estabelecimento de ligações entre eles".

No entanto, os dados disponíveis para este projeto são narrativas do domínio das notícias. Sendo assim, as narrativas consideradas serão apenas narrativas noticiosas. Segue abaixo um exemplo:

lusa_0.txt:

https://drive.google.com/file/d/1inFOvq1JSDT2-w9sfIZhpbqQzXc7_obr/view?usp=share_link

Redação, 11 out 2020
VAM // JH
Covid-19: GNR acabou com festa ilegal com 50 pessoas em São Brás de Alportel
A GNR acabou no sábado com uma festa ilegal num bar em Corotelo, concelho de São Brás de Alportel, que reunia cerca de 50 pessoas, foi hoje anunciado.
Em comunicado, o Comando Territorial de Faro da GNR adianta que, na sequência de uma denúncia a informar sobre a existência de "uma festa com música ao vivo num estabelecimento de bebidas", militares do Posto Territorial de São de Alportel deslocaram-se ao local para fiscalizar a ocorrência.
Os militares desmobilizaram as pessoas para "cumprimento das normas referentes à pandemia de covid-19".
Na nota é ainda referido que o proprietário foi identificado e foi elaborado "um auto de contraordenação".

[Questões iniciais]

Pensar nestas questões para o caso de uma notícia genérica, não apenas para o exemplo dado.

1. Quais os elementos que devem estar presentes numa representação visual de uma narrativa noticiosa?
   - personagens/intervenientes
   - eventos
   - contextualização temporal
   - contextualização espacial
   - título
   - data de publicação
   - outros…

Tendo isto em conta, naturalmente haverá elementos que será mais importante destacar.

Figure D.1: Introduction, context, and Q1 of the interview script.

2. Tendo em conta os elementos referidos anteriormente, classifique-os usando a seguinte escala, sendo essencial um elemento a dar mais destaque e insignificante um elemento a dar menos destaque:
   essencial > importante > útil > neutro > insignificante

3. Descreva como se relacionam estes elementos visualmente usando conceitos como "parte de", "composto por", "associado a", etc. Por exemplo, uma personagem está associada a um evento, um evento é composto por várias personagens. ?

4. Descreva a forma como imagina a representação visual de uma narrativa noticiosa?

5. A próxima fase da entrevista será opcional, consiste em observar alguns exemplos de protótipos de visualizações de narrativas. Antes disso, para não condicionar o pensamento, peço que esquematize um protótipo, tendo em conta os elementos referidos e a forma como se relacionam.

[Dar espaço para esquematizar algum protótipo que possa surgir em mente]

Whiteboard colaborativo: https://webwhiteboard.com/

Figure D.2: Q2 to Q5 of the interviews script.

[Exemplos de visualização]

Estes exemplos encontram-se a usar os dados do ficheiro lusa_0, da coleção de notícias da lusa.

Exemplo de visualização event-based, storyline com frases:

Demo interativa: https://catarina03.github.io/news-story-viz-prototype/
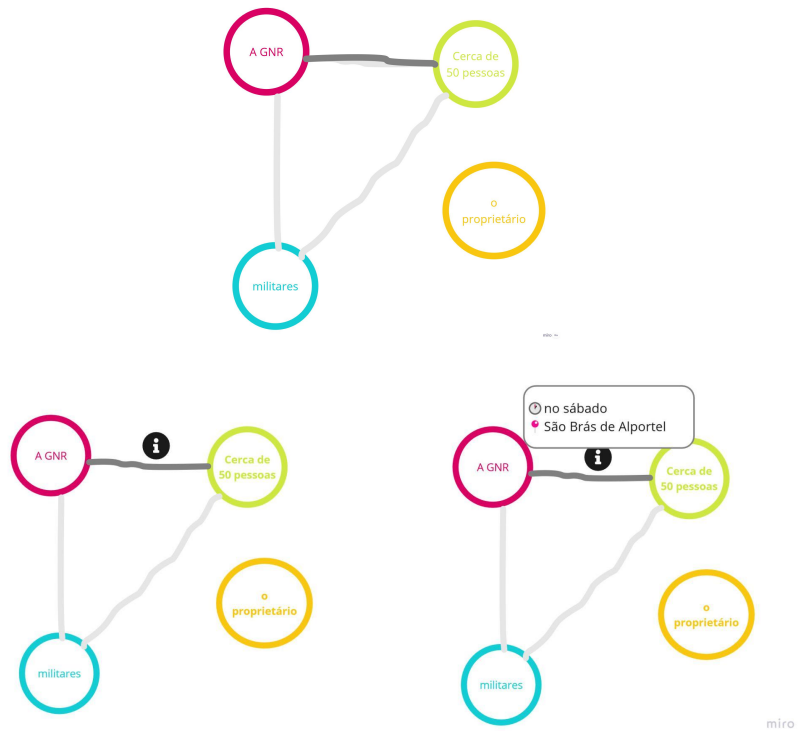
Nenhuma cena selecionada:
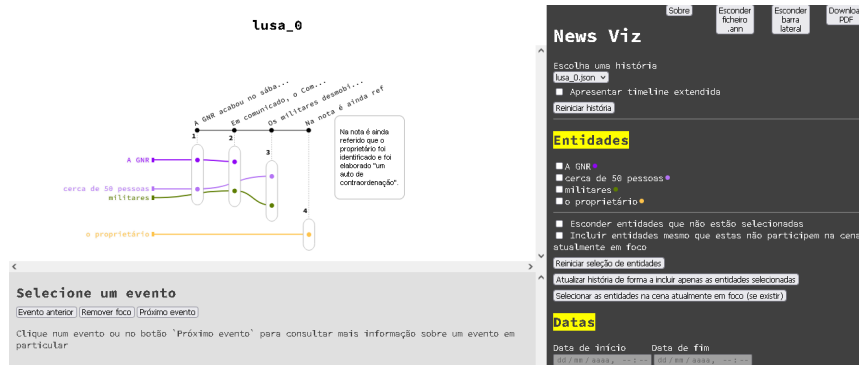
Cena 1 selecionada:
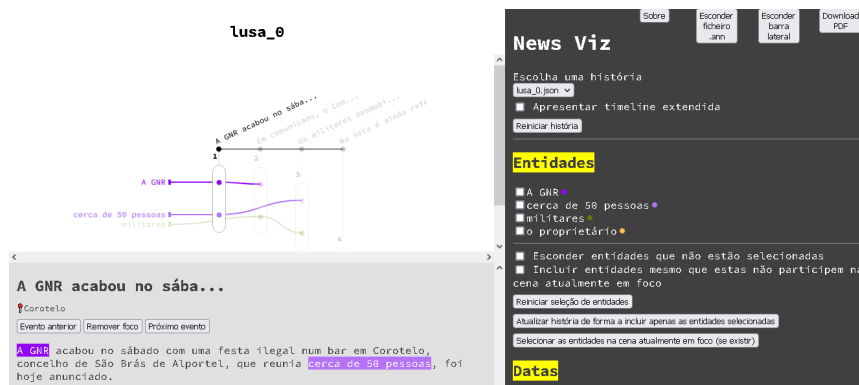
6. Quais são os pontos positivos desta visualização?
7. Que aspetos poderiam ser melhorados?
8. Quais são os pontos negativos?
9. Há elementos que deveriam ter mais destaque? Quais? Porquê?
10. Há elementos que deveriam ter menos destaque? Quais? Porquê?
11. Falta algum tipo de informação que sente ser importante?

Figure D.3: Q6 to Q11 of the interviews script.

Exemplo de visualização event-based, storyline com frases:

Demo interativa [storyline extendida]: https://catarina03.github.io/news-story-viz-prototype/

Nenhuma cena selecionada:
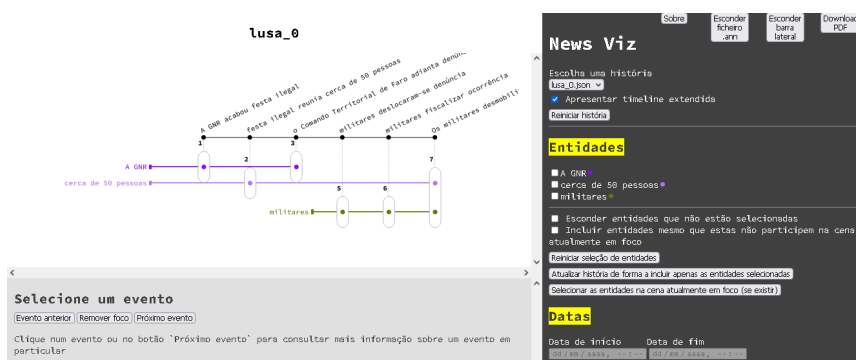
Cena 1 selecionada:
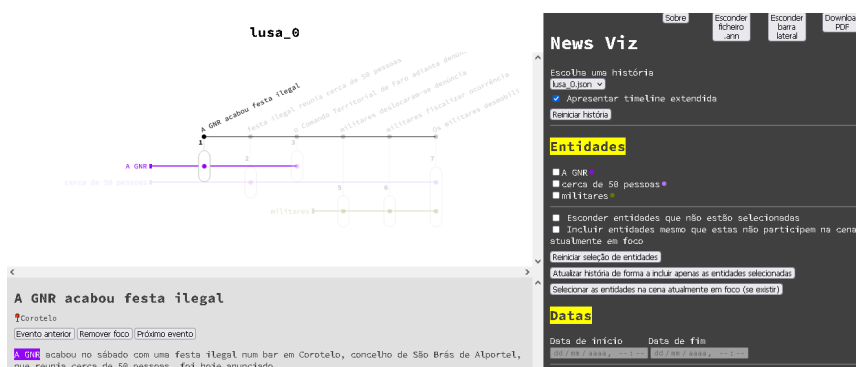
12. Quais são os pontos positivos desta visualização?
13. Que aspetos poderiam ser melhorados?
14. Quais são os pontos negativos?
15. Há elementos que deveriam ter mais destaque? Quais? Porquê?
16. Há elementos que deveriam ter menos destaque? Quais? Porquê?
17. Falta algum tipo de informação que sente ser importante?

Figure D.4: Q12 to Q17 of the interviews script.

Exemplo de um protótipo de visualização character-based, grafo:

Um exemplo de um protótipo de representações diferentes da cena 1 do ficheiro lusa_0 (caso um conjunto de personagens estejam presentes na mesma cena, a ligação entre elas aparece assinalada numa cor mais escura):

Um exemplo de um protótipo de representações da narrativa do ficheiro lusa_0, sem nenhuma cena selecionada:
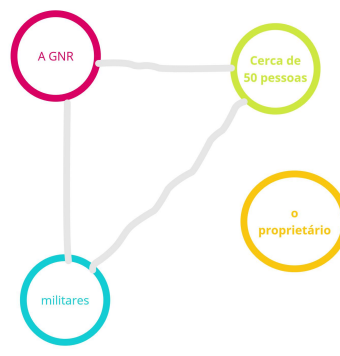
Figure D.5: Introducing the graph-based visualization prototype.

18. Quais são os pontos positivos desta visualização?
19. Que aspetos poderiam ser melhorados?
20. Quais são os pontos negativos?
21. Há elementos que deveriam ter mais destaque? Quais? Porquê?
22. Há elementos que deveriam ter menos destaque? Quais? Porquê?
23. Falta algum tipo de informação que sente ser importante?

Links para as imagens originais:

Storyline view com frases, nenhuma cena selecionada:
https://drive.google.com/file/d/1ISN1Q66lfnjVqqUtzmhxBSMoaDMX_jAc/view?usp=sharing
Storyline view com frases, cena 1 selecionada:
https://drive.google.com/file/d/1C6dJb7MvFpbAKJEhu4mioVPYIJFaTUvQ/view?usp=drive_link
Storyline view com eventos, nenhuma cena selecionada:
https://drive.google.com/file/d/1iDivpmyp45aaMwRSvX0F_Vj8-wwDFUFo/view?usp=sharing
Storyline view com eventos, cena 1 selecionada:
https://drive.google.com/file/d/1-1hBRbDJ9CrxUw6xqeeQo7HDyW5avY_6/view?usp=drive_link
Character view., nenhuma cena selecionada:
https://drive.google.com/file/d/1UNEOEiDaI3JkbfPG1IUgKFtV0SEvcsxq/view?usp=share_link
Character view, cena 1 selecionada:
https://drive.google.com/file/d/1Yzb_yRmG4_CancKulxGUOzTBUB51GyNK/view?usp=share_link
Character view, cena 1 selecionada, com informação:
https://drive.google.com/file/d/1Ab5yxJWhoMJFcupU-Mui77GsGj9_UMso/view?usp=share_link

Figure D.6: Q18 to Q23 of the interviews and the image links for the visualizations presented.
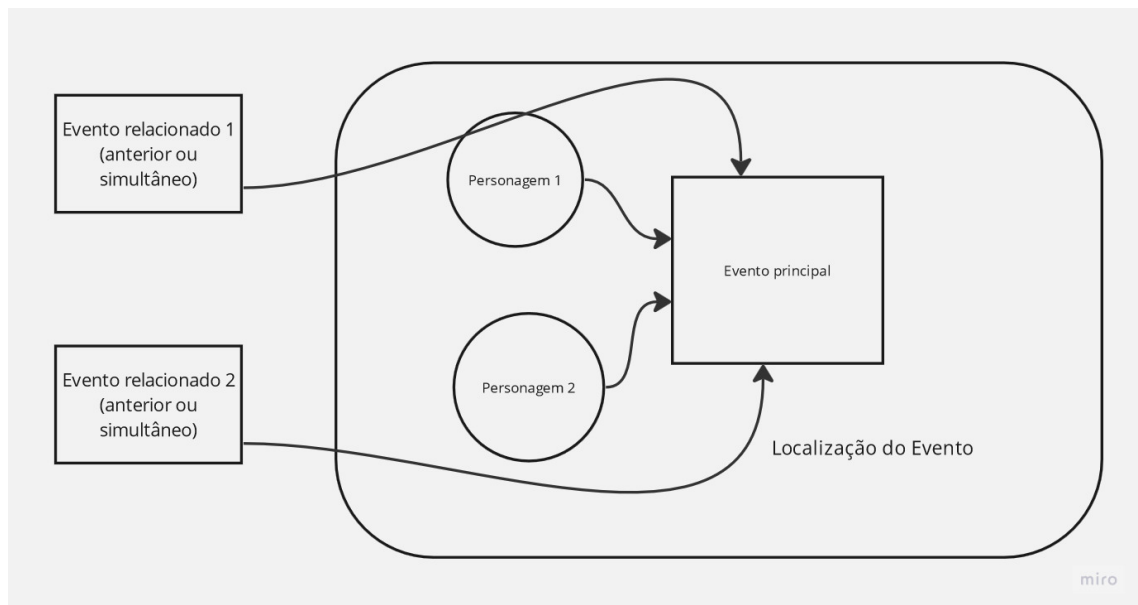
Figure D.7: Interview sketch of the participant that envisioned the visualization as *a set of images related to each other*.
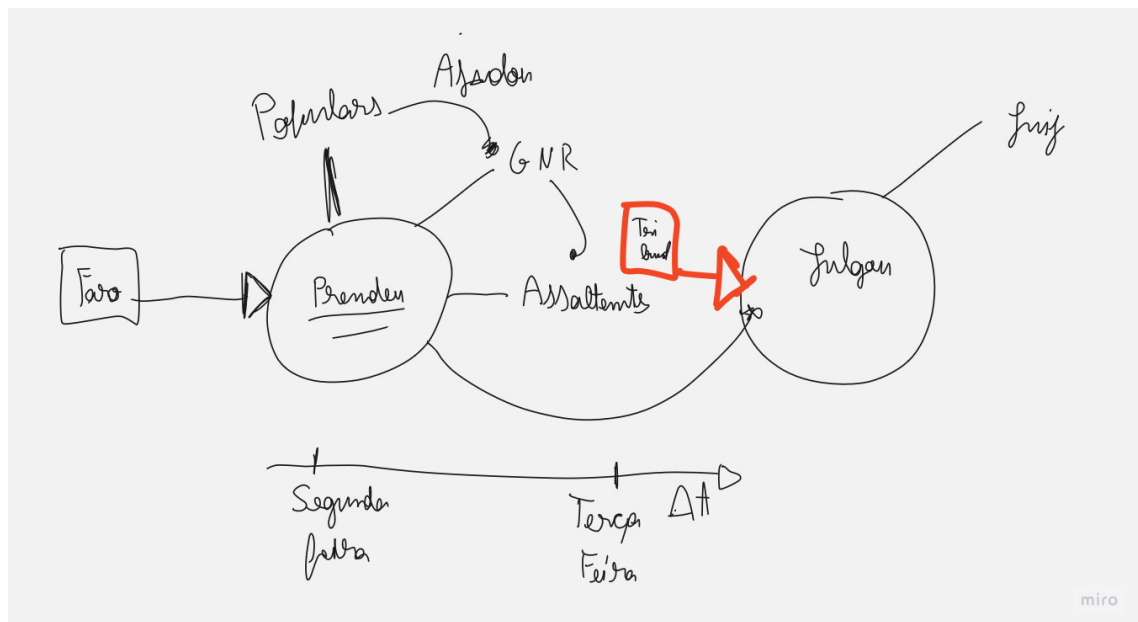

Figure D.8: Interview sketch of the participant that envisioned the visualization as *temporally ordered things, giving events more focus and linking them to participants*.
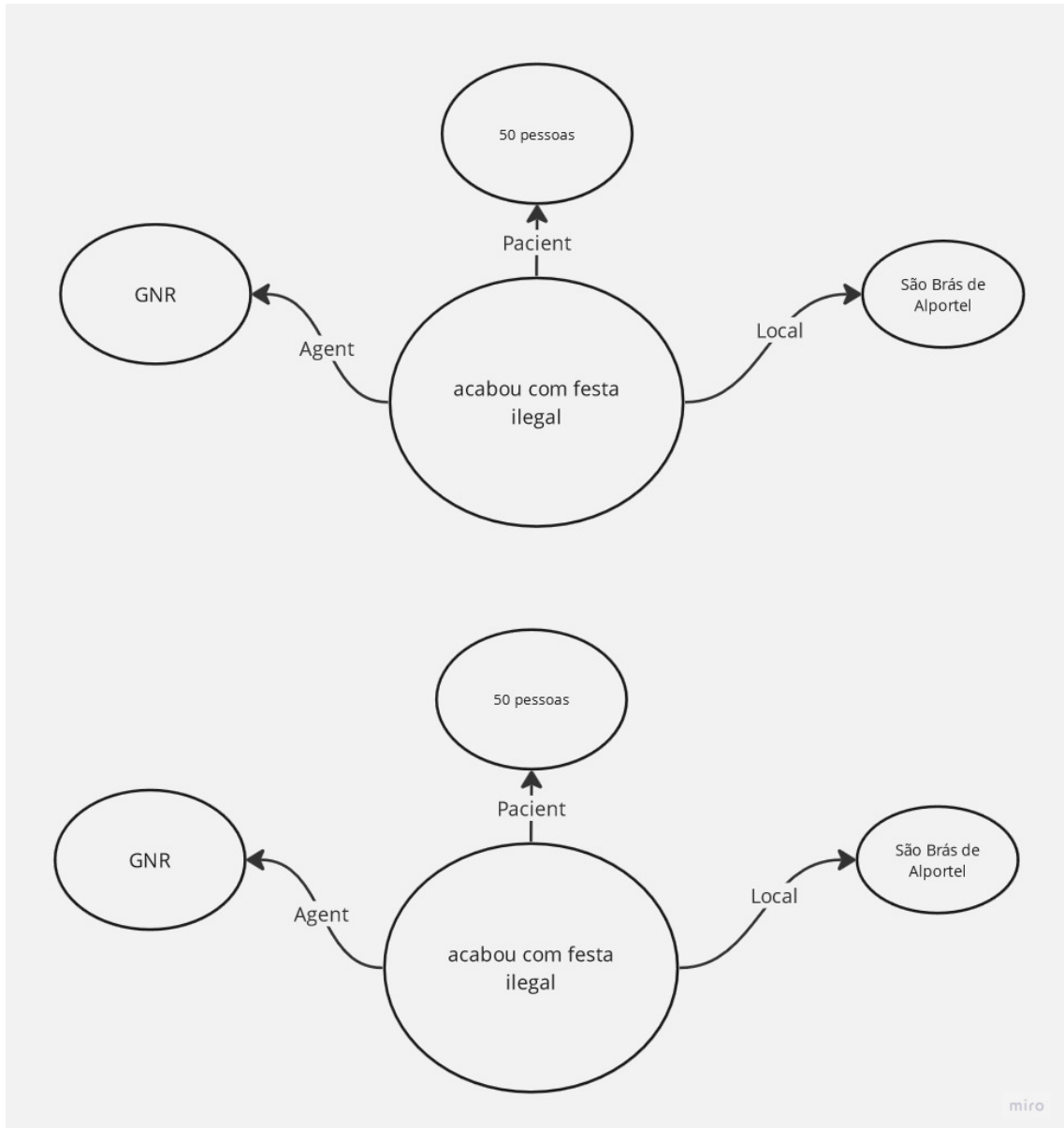
Figure D.9: Interview sketch of the participant that envisioned the visualization as *graph ontology, the event is the main node, and the rest of the information is linked to it, using semantic roles as edge, ordered chronologically or in the order of the narrative*.
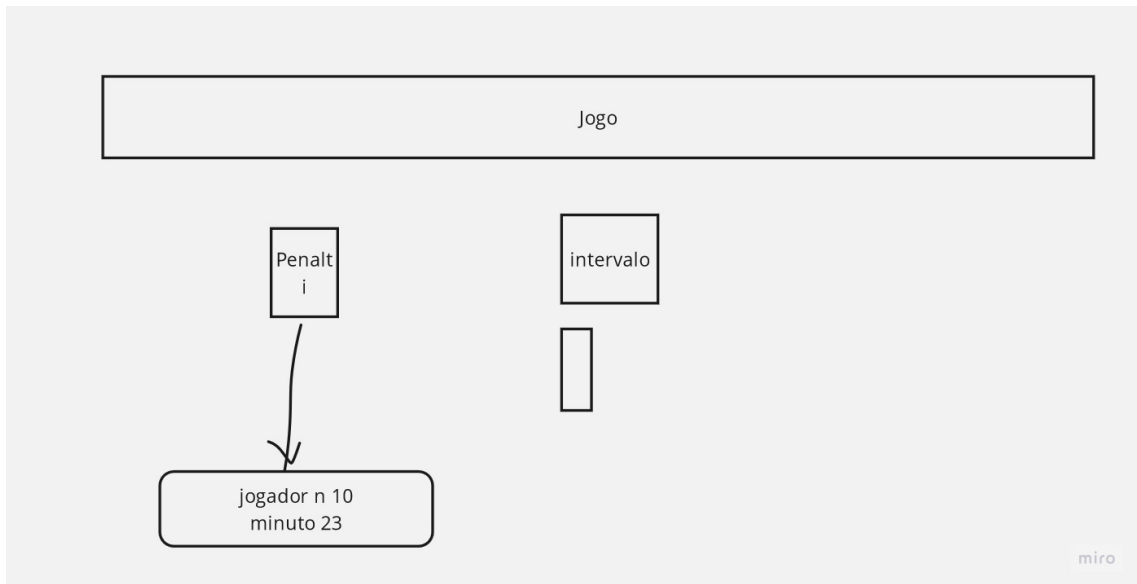
Figure D.10: Interview sketch of the participant that envisioned the visualization as *a timeline of events in boxes that on hover over the boxes show the participants and location, and above, a larger or smaller box showing the duration of the event (i.e. a soccer game, when someone scores a goal it's a smaller box that shows who scored the goal and when).*
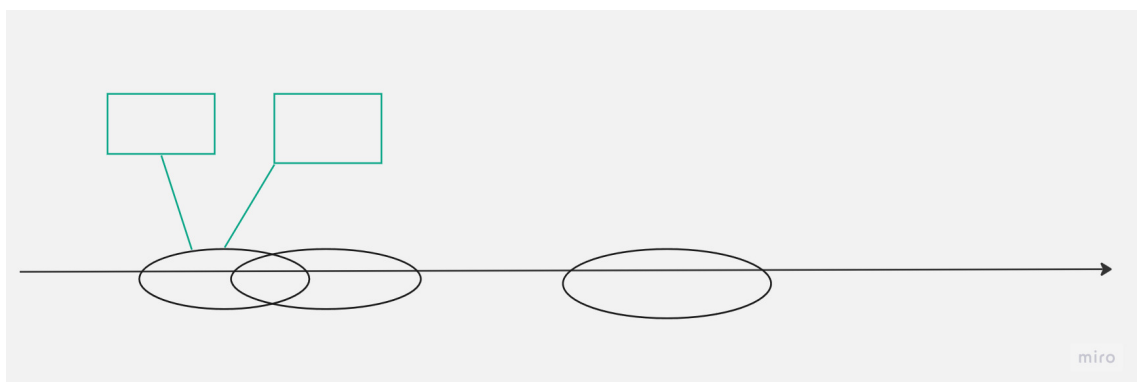

Figure D.11: Interview sketch of the participant that envisioned the visualization as *if there was temporal succession, a timeline, if there is temporal simultaneity, a bubbles representation.*
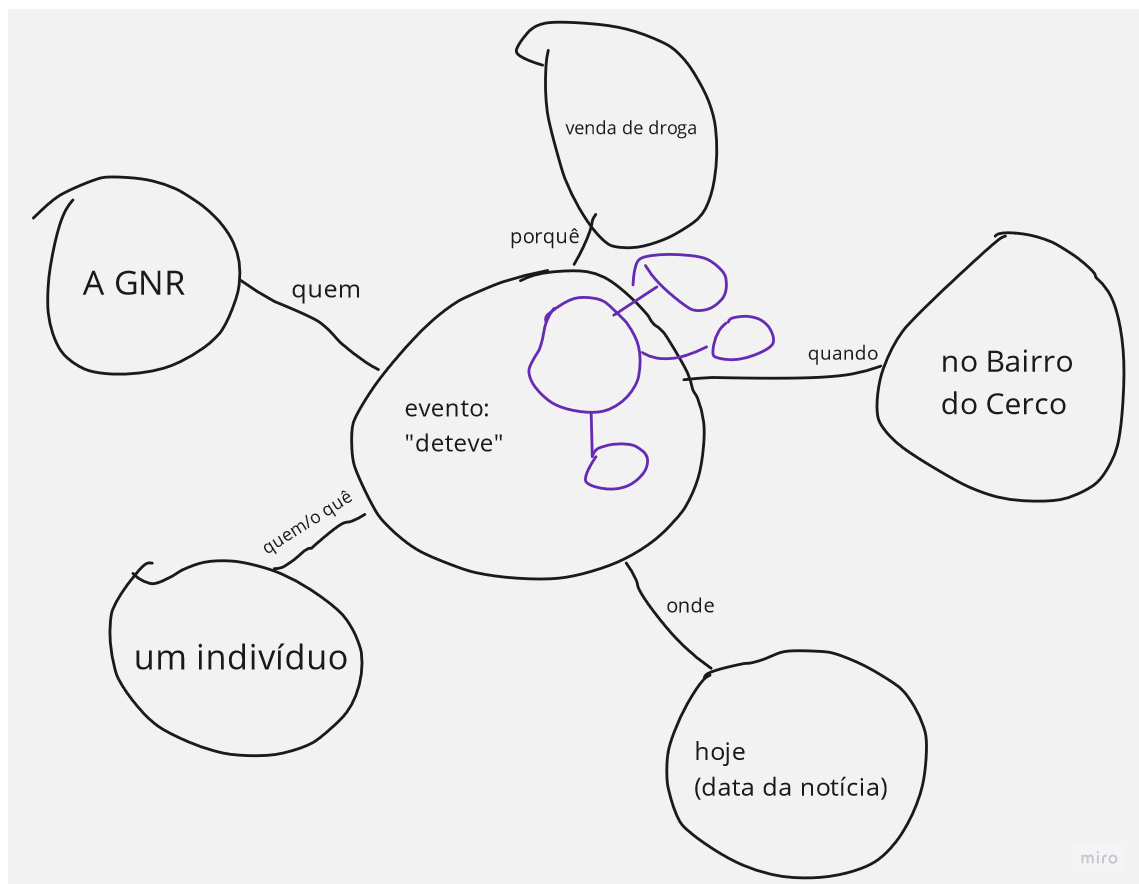
Figure D.12: Interview sketch of the participant that envisioned the visualization as *a central ball (event described by the verb), and balls with the other elements linked to it, and in the line the name of the relationship.*
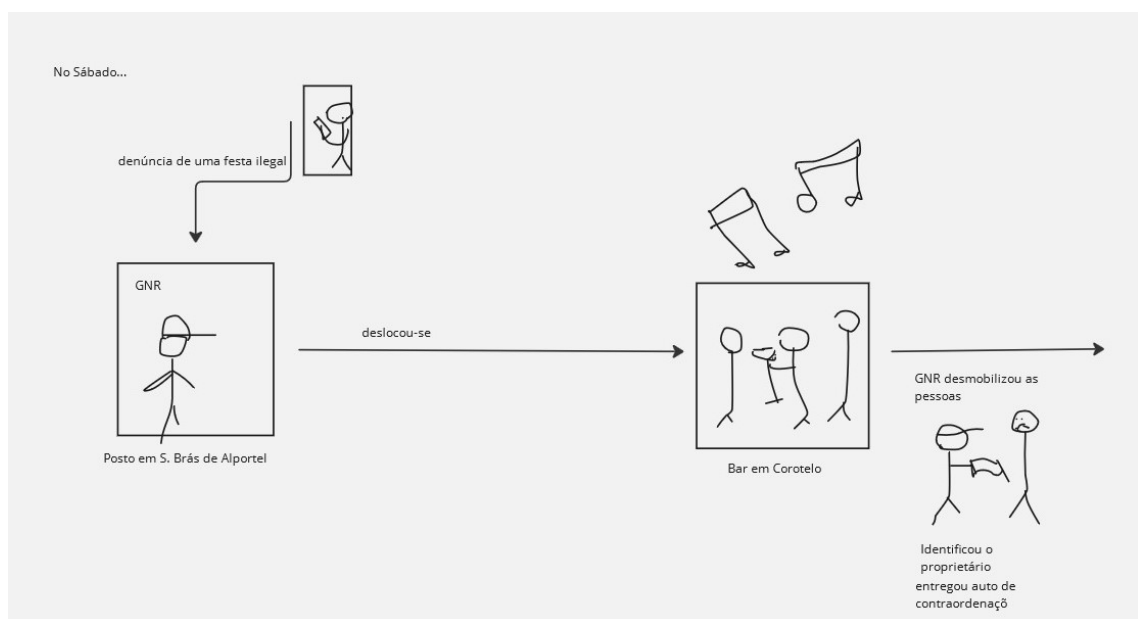
No Sábado...

denúncia de uma festa ilegal

GNR

deslocou-se

Posto em S. Brás de Alportel

Bar em Corotelo

GNR desmobilizou as pessoas

Identificou o proprietário entregou auto de contraordenaçõ

Figure D.13: Interview sketch of the participant that envisioned the visualization as *a line, where we have a sequence of events (if it were a comic book it would be a 4x4 with the 4 key moments), or in a more diagrammatic way with participants, lines to show the events and timestamps, another way would be a map like in infographics.*

Appendix E

Survey

This appendix will present the questions of the survey sent to the University of Porto.

E.1   Survey Questions

Figures E.1, E.2, E.3, E.4, and E.5 show the questions of the survey.

NewsLines

Este questionário tem como principal objetivo avaliar a utilidade de duas visualizações no que diz respeito à sua capacidade de auxílio à compreensão de uma notícia. Esta ferramenta foi desenvolvida no âmbito da minha tese de mestrado "Visualization of Narrative Structures" e consiste numa ferramenta de geração automática de visualizações de notícias, a partir de anotações BRAT.
A primeira secção recolhe dados demográficos, a segunda faz um levantamento sobre hábitos de leitura de notícias online e, por fim, a terceira secção avalia dois protótipos de visualizações.

Informação demográfica

Idade *

○ < 18

○ 18 - 24

○ 25 - 39

○ 40 - 59

○ > 59

Afiliação (sigla da faculdade, por exemplo, FEUP, FLUP) *

A sua resposta

Ocupação *

○ Estudante

○ Docente

○ Técnico

○ Investigador

○ Outra:

Figure E.1: Section one of the survey, demographic information.

Hábitos de leitura de notícias online

Com que frequência lê notícias online? *

◯ Menos de uma vez por semana

◯ Entre 1 a 3 vezes por semana

◯ Mais de 3 vezes por semana

◯ Uma vez ao dia

◯ Várias vezes ao dia

Quão sobrecarregado se sente com a quantidade de notícias a que é exposto diariamente?   *

	1	2	3	4	5	
Nada sobrecarregado	◯	◯	◯	◯	◯	Muito sobrecarregado

Com que frequência se depara com elementos visuais (gráficos, cronologias, etc) em notícias online?   *

	1	2	3	4	5	
Nunca	◯	◯	◯	◯	◯	Sempre

Quando uma notícia inclui elementos visuais (gráficos, cronologias, etc), quão úteis são para a compreensão da notícia?   *

	1	2	3	4	5	
Nada útil	◯	◯	◯	◯	◯	Muito útil

Qual das seguintes expressões descreve melhor o seu método de leitura das notícias?   *

◯ Prefiro ter uma visão panorâmica das notícias, mesmo que não consiga ler todos os artigos detalhadamente

◯ Prefiro ler com atenção os artigos com que me deparo, analisando cada um detalhadamente

◯ Não tenho preferência por nenhum dos métodos

Figure E.2: Section two of the survey, online news reading habits.

Ferramenta de geração de visualização de notícias a partir de anotações de texto

Quais os elementos que devem estar presentes numa representação visual de   *
uma narrativa noticiosa?

☐ Participantes

☐ Eventos

☐ Informação temporal

☐ Informação espacial

☐ Título

☐ Data de publicação

☐ Outra: _____

Tendo em conta os elementos referidos anteriormente, classifique-os usando a   *
seguinte escala, sendo essencial um elemento a dar mais destaque e
insignificante um elemento a dar menos destaque:

	Essencial	Importante	Útil	Neutro	Insignificante
Participantes	○	○	○	○	○
Eventos	○	○	○	○	○
Informação temporal	○	○	○	○	○
Informação espacial	○	○	○	○	○
Título	○	○	○	○	○
Data de publicação	○	○	○	○	○

Figure E.3: First part of section three of the survey, prototype evaluation.

Protótipo

Abaixo será apresentada uma ferramenta de visualização de notícias. Esta ferramenta procura enriquecer e complementar os elementos narrativos encontrados no texto original, salientando principalmente as ligações entre eventos e participantes.

O primeiro vídeo mostra uma visualização storyline, considerando que cada frase do texto original é um evento. O segundo vídeo mostra o mesmo formato de visualização, extraindo os principais eventos do texto e representando-os.

Esta visualização foi gerada de forma automática através de anotações BRAT feitas no texto original.

O seguinte texto é usado como exemplo:
Redação, 11 out 2020
VAM // JH
Covid-19: GNR acabou com festa ilegal com 50 pessoas em São Brás de Alportel
A GNR acabou no sábado com uma festa ilegal num bar em Corotelo, concelho de São Brás de Alportel, que reunia cerca de 50 pessoas, foi hoje anunciado.
Em comunicado, o Comando Territorial de Faro da GNR adianta que, na sequência de uma denúncia a informar sobre a existência de "uma festa com música ao vivo num estabelecimento de bebidas", militares do Posto Territorial de São de Alportel deslocaram-se ao local para fiscalizar a ocorrência.
Os militares desmobilizaram as pessoas para "cumprimento das normas referentes à pandemia de covid-19".
Na nota é ainda referido que o proprietário foi identificado e foi elaborado "um auto de contraordenação".
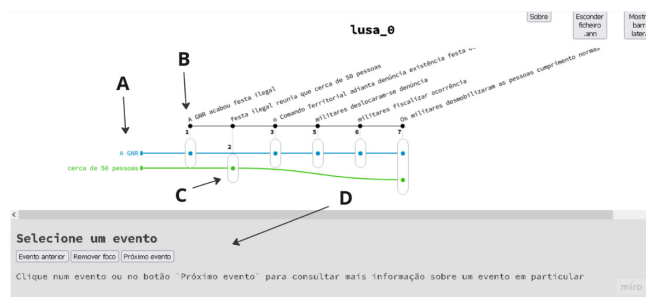
Segue abaixo a legenda de uma visualização:
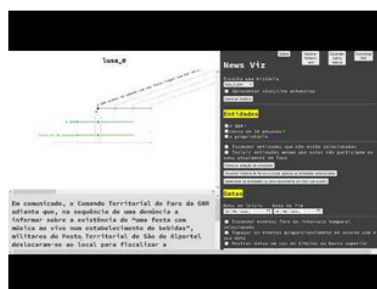A - Participantes/Entidades
B - Evento
C - Participação de uma entidade num evento
D - Descrição de um evento em foco, se existir

Visualização legendada

Geração de uma storyline a partir das frases de uma notícia

Link: https://www.youtube.com/watch?v=gxmnuZrvUUk

Figure E.4: Second part of section three of the survey, prototype evaluation.

Como classificaria a utilidade da primeira ferramenta, geração automática de *
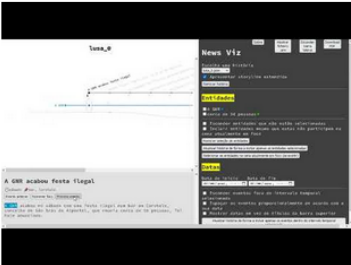uma storyline a partir das frases de uma notícia?

	1	2	3	4	5	
Nada útil	○	○	○	○	○	Muito útil

Como considera que a primeira ferramenta contribuiu para um melhor *
entendimento da notícia apresentada anteriormente?

	1	2	3	4	5	
Não contribuiu em nada	○	○	○	○	○	Contribuiu muito

Geração de uma storyline a partir dos eventos de uma notícia

Link: https://www.youtube.com/watch?v=lGi20NfzN1Y

Como classificaria a utilidade da segunda ferramenta, geração automática de *
uma storyline a partir dos eventos de uma notícia?

	1	2	3	4	5	
Nada útil	○	○	○	○	○	Muito útil

Como considera que a segunda ferramenta contribuiu para um melhor *
entendimento da notícia apresentada anteriormente?

	1	2	3	4	5	
Não contribuiu em nada	○	○	○	○	○	Contribuiu muito

Estaria interessado em usar estas visualizações como complemento a artigos de *
notícias?

	1	2	3	4	5	
Nada interessado	○	○	○	○	○	Muito interessado

Observações:

A sua resposta

Figure E.5: Third part of section three of the survey, prototype evaluation.