The handle http://hdl.handle.net/1887/133504 holds various files of this Leiden University dissertation.

**Author:** Cunha, E. Landulfo Teixeira Paradela  
**Title:** Contributions to the computational processing of diachronic linguistic corpora  
**Issue date:** 2020-03-19
CHAPTER 2

Building a diachronic corpus of comments extracted from news portals and websites

2.1 Introduction

Within the context of corpus linguistics, an imperative task is the development of tools and resources capable of assisting researchers in the process of collecting and organizing material for analysis.

1 This chapter reproduces with minor changes the article “Xereta: A Brazilian corpus of online news comments” (Cunha et al., under review), submitted for publication and presented (under the title “Contribuições para a coleta e a compilação de um corpus de comentários de portais de notícias”) at the XI International Conference of the Brazilian Linguistics Association (ABRALIN 2019), held in Maceió, Brazil, in May 2019. It is an extended and updated version of the paper “A elaboração de um coletor e de um corpus de comentários extraídos de portais de notícias” (Cunha et al., 2017), published in the Proceedings of the X International Conference of the Brazilian Linguistics Association (ABRALIN 2017) and presented at this conference, held in Maceió, Brazil, in March 2017. See Appendix C for more information.
Despite the large amount of corpora of all types available (written and spoken, general and specialized, synchronic and diachronic, historical and contemporary etc.), in certain cases researchers face situations in which it is impossible to perform the desired study due to the lack of adequate material in already collected corpora. From our experience, this seems to be a recurrent issue when considering online content (such as social media posts, blog entries, online forum discussions and the like): even though this kind of data is often and increasingly useful for researchers from various fields of study, those with little or no programming skills are sometimes prevented from obtaining the desired data for their research due to the lack of suitable and accessible tools and resources\(^2\).

Various types of content available on the Internet – from personal conversations in online chat rooms to media texts aimed to attract broad audiences – may be of interest to corpus linguists, since the character of these texts and the language represented electronically are greatly diverse. Kilgarriff (2005), in a controversial statement, argues that “it is the Web that presents the most provocative questions about the nature of the language” (p. 473); Crystal (2004) adds that “there are good grounds for viewing the arrival of the Internet as an event which is as revolutionary in linguistic terms as it has been technologically and socially” (p. 65).

One of the new text genres that have been growing in use as Internet access increases across the world is the genre comment in news portal. This type of text is frequent in webpages that publish news articles, such as online newspapers and magazines, as a means of interaction between readers and media producers. Indeed, the

\(^2\) This is a particularly pertinent situation in resource-limited parts of the world where higher proportions of the population lack advanced digital skills: technical and technological obstacles to working with certain types of data might increase the gap between developed and developing countries (as well as between rich and poor institutions) regarding research results and quality of scientific outcomes.
sections dedicated to readers’ comments in news portals might be roughly understood as modern versions of the traditional readers’ letters sections in print media.

The investigation of these readers’ comments is relevant for researchers interested in linguistic and textual characteristics of Internet genres, since analyses of these texts allow the study of issues related to this genre in the most varied domains, including lexical, morphosyntactic and pragmatic. They might also be useful for the investigation of phenomena such as language variation and change in the online world, and the relationship between language and technology, while discourse analysts may profitably use this kind of text to examine ideology, bias and representation on the Internet, for instance. Professionals from several other fields of study, including social scientists and journalists concerned with the public perception of news and the relationship between media and society, might also make good use of this kind of material.

It is necessary, therefore, to develop tools to assist the collection and organization of such data, as well as to compile corpora that comprise this text genre. In this chapter, we present two useful resources in this regard: (a) a Web scraper of comments from news portals and websites, developed as open source and as being free for use, modification and distribution; and (b) a freely available corpus composed of comments published at UOL, a major Brazilian news portal. The scraper has a simple format, allowing users with no technical background and limited computing knowledge to simultaneously collect all the comments published in a given set of news by simply informing a list of URLs. The corpus brings not only the texts of the comments themselves, but also important metadata such as dates and times of publication, commentators’ user names and numbers of likes received by the comments. Both resources have been given the name Xereta\textsuperscript{3} – a colloquial word

\textsuperscript{3} Pronounced as [ʃeˈreta] – or, roughly, as shea-RE-ta.
meaning “busybody, nosy” in Brazilian Portuguese.

In the next section, we present related work and the main characteristics of texts posted by readers as comments on news portals. Then we discuss results of the efforts for the elaboration of the two resources introduced here, and present the main conceptual and computational challenges and limitations faced during their elaboration. In addition, we provide a general characterization of the compiled corpus. In Section 2.4, we mention some possibilities for research employing both the scraper and the corpus presented here, hoping that these ideas might be of some value to researchers wishing to use these resources. Finally, we conclude the chapter by presenting future steps that should be followed in order to maintain the project and expand its reach.

2.2 Comments in news portals

News portals and websites are currently responsible for much of the volume of the world Internet traffic. As an illustration, as of August 2019, seven (14%) of the top fifty websites ranked by SimilarWeb in the world were classified in the category “News and Media”. When visualizing data regarding the Netherlands, this number drops to six (12%); however, data from Brazil show that, in this country, the traffic of news sites is extremely high, with the impressive number of twelve (24%) webpages among the top fifty ranked.

At the end of the second decade of the 21st century, the online news territory, that “began as the simple provision of news using websites on the Internet”, has become “an environment of multi-

---

4 SimilarWeb is a company that provides Web analytics services. As of the time of writing of this dissertation, its ranking could be accessed on https://www.similarweb.com/top-websites. The ranking provided by SimilarWeb “calculates the number of monthly unique visitors together with the number of page views across desktop and mobile traffic” (SimilarWeb, 2016).
ple digital platforms and products and numerous ways of accessing news content” (Küng et al., 2016, p. 443). Also according to Küng et al. (2016), at least two eras in the development of online news services are identifiable: while the period from 1993 until approximately 1999 may be called the “era of digital publishing”, the following period can be termed the “era of participation and multimedia”. From the 2000s, with the advent of Web 2.0\(^5\) (O’Reilly, 2007; Cunha, 2012), news portals began to incorporate readers’ participation: on the same page in which a certain news is published, readers are able to expose their opinion and make it available to others interested in that specific piece of news – thus creating not only content, but, above all, conversations and communities (Amoris et al., 2012). In the words of Milioni et al. (2012), “[t]he popularization of web 2.0 has signaled a new era in audience participation, one that is interactive and allows users to produce and publish their content online”.

Comments posted by readers of news portals and online newspapers and magazines – such as Yahoo! News and The New York Times; or G1, Terra and UOL, in Brazil; or NU.nl, in the Netherlands\(^6\) – are expressions of a text genre that grows along with the expansion of Internet use in the world. Figure 2.1 illustrates some of the comments posted by readers in a Yahoo! News article. It is interesting to note that popular articles may reach a considerable number of comments, sometimes thousands of them – as is the case shown in the figure, which reached (at least) 6,684 comments. Some portals allow comments to be posted in response to other users’ comments and, in some cases, readers are able to positively

---

\(^5\) “Web 2.0 refers to the social use of the Web which allow people to collaborate, to get actively involved in creating content, to generate knowledge and to share information online” (Grosseck, 2009, p. 478).

or negatively evaluate comments.

Figure 2.1: Comments posted by readers in a Yahoo! News article. In this example, comments are called *reactions* and may receive replies and positive/negative evaluations by other readers.

In the words of Henrich and Holmes (2013), “[c]omments have the potential to increase our understanding of public opinions, how the public makes decisions and how beliefs are formulated” (p. 1). However, according to Potthast (2009), comment boards on news sites “are often flooded with all kinds of junk and spam, which may be a reason why research has widely neglected comments as
a source of information” (p. 724). Indeed, the scholarship on the topic of comments in online news portals is still somewhat limited. Nonetheless, interesting results have already been published, as we show in the next paragraphs. Some of these results are especially relevant for social and media scientists.

One case of recent scholarship is the work of Lee (2012), who surveyed South Korean individuals in order investigate whether user-generated comments on Internet news sites affect other readers’ perception on the bias of the news report itself. Among her results, we highlight the finding that, in some cases, “people might misattribute the opinions expressed in others’ comments to the news article” (p. 32), demonstrating the influence of comment boards on news reception. Related to this, Milioni et al. (2012) explore the assumption that the possibility of publishing their own comments on news pages gives audience a greater power over influencing the very activity of news production – that is, a power of influencing journalists and media outlets. Their findings suggest that, at least in the analyzed (Greek) context, even though commentators often challenge journalistic viewpoints, “this type of audience participation is not likely to render audiences co-producers of news content in significant ways” (p. 21). Erjavec and Kovačič (2012), using a critical discourse analysis framework, investigate hate speech on Slovenian news websites’ comments, finding that most of the hate speech producers “share characteristics of an authoritarian personality” (p. 899) and are mostly motivated by thrill and fun.

There are also a number of studies that focus on the analysis of this type of Web comments in specific contexts, such as in the political sphere (e.g. Maia et al., 2015). Particularly important for us is the investigation carried out by Rossini (2017), the first user of the scraper presented in this chapter. In her extensive analysis of informal political conversation on the Web, the author employs a systematic content analysis approach to investigate 12,797 comments
extracted from news portals and websites, examining incivility and intolerance in this environment. Among the results obtained, we highlight the one suggesting a normalization of incivility – which is usually employed as a rhetorical resource in situations of disagreement – in these online spaces. The outcomes of her study are relevant and contribute to the understanding of how technology can affect the political and social dynamics of contemporary democratic societies.

The previously mentioned studies are based on surveys and/or qualitative research usually conducted through the individual analysis of each comment. Some other studies have used use larger selections of comments extracted from Web portals as objects of analysis, thus taking more quantitative approach. Potthast (2009), for instance, investigates the descriptive nature of this type of comments, revealing that “10 comments suffice to expect a high similarity between the comments and the commented text; 100-500 comments suffice to replace the commented text in a ranking task” (p. 724). Additionally, Reyes et al. (2010) evaluate humorous features on Web comments with the goal of automatically distinguishing between implicit funny comments from not funny ones.

From a computational perspective, Hsu et al. (2009) propose a machine learning approach to rank comments based on quality – which may be useful, for instance, to promote high-quality comments and filter out low-quality ones, including spams –, while Potthast and Becker (2010) present a tool to help summarize and visualize expressed opinions in the form of comments on the Internet. Their motivation is that popular items often receive thousands of comments (see Figure 2.1 as an example), thus making “visitors read (...) only the newest comments and hence get an incomplete and possibly misleading picture of the overall opinion” (p. 668). Moreover, named-entity recognition in news comments on the Web is addressed by Wan et al. (2011), who propose a method for iden-
tifying person, location and organization names in comments.

All of the above-mentioned works were made possible by the availability of real and authentic data collected from news websites. In small-scale studies, in which the amount of data required is more limited, we observe that in most cases the data collection was performed manually (i.e., researchers visited the news pages and manually copied the comments available). In studies using larger numbers of comments (from a few thousands upwards), data collections were certainly accomplished by computational tools. In other cases, already available corpora of news comments were used. A comparison between different corpora of news comments published up to the writing of this dissertation is presented in Section 2.3.2.

In the next section, we introduce two relevant resources for the research on the text genre comment in news portal, hoping that they will be especially useful for researchers who, for any reason, are not able to develop their own scripts for this type of data collection.

2.3 General description of the resources

Here, we introduce the two resources presented in this chapter: (a) a Web scraper of comments from news portals and websites, developed as open source and as being free for use, modification and distribution; and (b) a freely available corpus composed of comments published at UOL, a major Brazilian news portal.

2.3.1 The Web scraper

In Web science, a scraper refers to an automated agent used to extract data from targeted sources (Upadhyay et al., 2017), thus capturing specific content from webpages. The purpose of the scraper presented here is to return a file containing all comments posted on a given list of news articles published in a news portal. One of its
First, it is important to note that this tool was developed as open source and as being free for use, modification and distribution. In short, it means that anyone can read the code, modify it, contribute, and use it in their own projects. Our Web scraper can be downloaded and used free of charge, without any registration, and if programming-skilled users want to, they might modify it to better meet their goals. The scraper Xereta is licensed under the GNU GPL (General Public License), which guarantees the freedom of running the program for any purpose, studying how it works, adapting it to one’s needs, redistributing copies, and ultimately refining it and release improvements that might benefit the whole community. By acting in this way, we are contributing to the community by lowering costs and providing greater flexibility for researchers (Corrado, 2005).

The scraper presented here was developed in Python and is available in two different formats: for download and for online use. The code available for download is the same as the one available for online use. The downloadable version is intended primarily for those who might want to see the code and modify it, while the online version is best suited for general and quick use, especially by less programming-skilled users. The online version of the scraper is available at http://xereta.herokuapp.com/\(^7\). To use it from the website, the user must simply enter in the available box the list of URLs from which comments should be collected\(^8\). The automati-

\(^7\) If at any time this page is down, it is recommended to look for instructions at https://www.dcc.ufmg.br/~evandrocunha/.

\(^8\) The list of URLs to be inserted must, of course, be made up of articles of interest to the user. There are some ways to automate the process of obtaining these URLs. One of them is the automated collection from robots.txt files and sitemaps. This process is described in the next section, where we detail the collection of the corpus Xereta.
cally downloaded file might be opened in a spreadsheet editor (such as LibreOffice Calc or Microsoft Excel).

The main challenge in designing a Web scraping tool for gathering comments from news sites is that page structures of news portals might completely differ from each other. Given the technology available during the execution of this work, the most reliable way to collect comments from different websites (e.g. The New York Times, Yahoo! News etc.) is to develop a specific piece of code explicitly designed for each one of them. In addition, page structures of different sections (e.g. entertainment, sports, politics etc.) on the same website may also differ. This means that the code used to collect comments from one portal cannot be entirely reused to collect comments from another portal: basically, it is necessary to generate a modified module for each news portal to be collected, which means that developers need to visit different websites (and different sections of the same website) to identify their structural characteristics and the extent to which they differ from each other.

Moreover, the meta-information available on each portal also varies. For example, while some news sites allow comments to be rated positively and negatively, others allow only positive ratings (“likes”), while others do not even offer this possibility. With user convenience in mind, we decided to generate a single output file that compiles all comments from the given URLs, regardless of the source portal. Thus, it was necessary to define the set of meta-information fields to be included in the output file. In the version available at the time of this publication, the output file contains the ten columns mentioned in Table 2.1.

At the time of publication of this work, our Web scraping tool processes pages from two Brazilian portals: Folha de S. Paulo and UOL. However, the idea is to expand this project and make it able to gather content from more websites, especially from Brazilian, Dutch and international portals.
Table 2.1: Columns available in the output file of the Web scraper Xereta.

<table>
<thead>
<tr>
<th>Column title (in Portuguese)</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ID</strong></td>
<td>Comment ID, provided by the news portal</td>
</tr>
<tr>
<td><code>Resposta a...</code></td>
<td>“Reply to...”: ID of the comment to which the comment replies (if the comment is a reply to another comment; if it is not a reply, this field is filled with “0”)</td>
</tr>
<tr>
<td><strong>Título da notícia</strong></td>
<td>Title of the news article</td>
</tr>
<tr>
<td><strong>Data da notícia</strong></td>
<td>Publication date of the news article</td>
</tr>
<tr>
<td><strong>Usuário</strong></td>
<td>Username provided by the commentator</td>
</tr>
<tr>
<td><strong>Comentário</strong></td>
<td>The comment itself</td>
</tr>
<tr>
<td><strong>Data do comentário</strong></td>
<td>Publication date of the comment</td>
</tr>
<tr>
<td><strong>Hora do comentário</strong></td>
<td>Publication time (hour and minute) of the comment</td>
</tr>
<tr>
<td><strong>Aval. positivas</strong></td>
<td>Number of positive ratings received (“likes” or “thumbs up”)</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td>URL of the news article</td>
</tr>
</tbody>
</table>

Figure 2.2 displays the operating diagram of the Web scraper Xereta. We observe that the scraper first receives as input a single file with a list of all URLs to be visited and, at the end, it generates as output a single file containing all the comments (and respective meta-information) gathered from all those URLs. Internally, the scraper’s first task is to consume the file containing the URLs and store them in a list. The program then iterates over this list, visiting one URL at a time. As mentioned above, the algorithm
Figure 2.2: Operating diagram of the Web scraper Xereta. The labels *Site A* and *Site B* represent news portals and websites from where the scraper is able to collect comments. Additional modules can be inserted at any time.
for extracting comments and their meta-information is different for each portal, making it necessary to create a specific module for each site. The scraper then checks the URL string itself to identify the source portal and hence which module to use. It is important to note that, despite the need to develop individual modules, our Web scraper is flexible enough to allow the inclusion of additional modules for other portals simply by implementing them using the same format as the already available ones. After comments are extracted from a given URL, they are included in a list where all comments collected are already stored. At the end, when the entire list of URLs is traversed, the program saves the list of comments in the output file.

The developer interested in inserting new modules to collect comments posted on different websites must follow the steps below. First, it is required to inform the scraper which module to use (since a different module is called for each different website). In the version described in this study, this is done by simply analyzing the URL, as follows:

```python
if "folha.uol.com.br" in url:  #Check if URL is from Folha
    comentarios = coleta_folha(url) #Use module "Folha"
elif "uol.com.br" in url:  #Check if URL is from UOL
    comentarios = coleta_uol(url) #Use module "UOL"
else:  #If URL is unknown...
    comentarios = [] #...do not call any module
    print("URL inválida") #Inform that URL is invalid
```

In the case of adding a new module, it is necessary to inform the scraper which module will be called depending on the format of the URL by inserting a piece of code such as:

```python
elif "nu.nl" in url:  #Check if URL is from NU.nl
    comentarios = coleta_nu.nl(url) #Use module "NU.nl"
```
Then it is needed to implement the module itself, which receives as input the string of the URL to be collected and should return a list of comments. In general, a manual analysis of the page structure is required to identify the URL of the page where the comments are stored. The programmer should then look for the fields that correspond to the columns to be output. It is necessary to implement a site-specific data scraping task, which can be performed in a number of ways (e.g. through regular expression, HTML XPath extracting, CSS selector, API request etc.)\(^9\). In the version of the scraper described here, the list of comments returned should contain the information available in Table 2.1, in that exact order. If it is not possible to collect a certain field or if a given information does not apply for the website (e.g. if the website does not provide the feature of positively rating a comment), the module should fill this information with a null value.

As far as we are concerned, this is the first freely available resource specifically built for the collection of comments published in news portals.

### 2.3.2 The corpus

Using the architecture of this Web scraper, we collected a corpus containing comments (plus corresponding meta-information) posted in news articles published in the Brazilian major news portal UOL. The UOL website\(^{10}\), whose acronym stands for *Universo Online*, was chosen because it is one of the most accessed and traditional Brazilian webpages, being active since 1996 (Moreira et al., 2018). In August 2019, it was ranked by SimilarWeb in the seventh position among the most visited websites in Brazil and second in the category “News and Media”, with an average of approximately

---

\(^9\) Examples using regular expressions are available in the open-source code of the Web scraper Xereta itself.

\(^{10}\) [https://www.uol.com.br/](https://www.uol.com.br/).
650 million visits per month\textsuperscript{11}. UOL publishes news on the most varied topics, including politics, economics, entertainment, sports and science, to name a few.

In 2017, we collected and made available a first version of the corpus Xereta. This version contains 23,455 unique comments posted in news articles published in 2014 and can be called our “pilot corpus”. In 2019, we released a second version, now containing 202,541 unique comments – almost ten times more than in the pilot corpus – posted in news articles published from January 2016 until December 2018, thus comprising three full years of data collection. These three years are especially interesting for the analysis of news portals comments since they are among the most heated times in Brazil’s history, particularly due to the impeachment of (or coup d’état against) President Dilma Rousseff in 2016 and to the 2018 extremely turbulent general balloting that resulted in the election of far-right candidate Jair Bolsonaro. In this chapter, we are referring to this second version of the corpus.

As shown in Figure 2.2, to perform the scraping it is first necessary to provide a list of URLs to be used as input. To obtain this list for our corpus compilation, we followed the steps below:

1. We accessed the \texttt{robots.txt} file\textsuperscript{12} available at \url{https://noticias.uol.com.br/robots.txt};

2. From this file, we obtained the URL of the sitemap\textsuperscript{13} available at \url{https://noticias.uol.com.br/sitemap/index.xml};

3. The XML sitemap obtained included a list of additional

\textsuperscript{11} Data obtained from \url{https://www.similarweb.com/website/uol.com.br}.

\textsuperscript{12} A \texttt{robots.txt} file is a text file placed on sites’ root directory giving instructions to search engine robots.

\textsuperscript{13} \textit{Sitemaps} are lists containing the URLs of the pages in a website, usually intended to help search engine bots to explore, crawl and index site’s webpages.
sitemaps, each one corresponding to one month – from January 2016 until the time of the corpus compilation (July 2019);

4. We gathered all the URLs available in the sitemaps corresponding to the period between January 2016 and December 2018, thus obtaining two full years of data.

It is important to notice that each section of the UOL website has its own robots.txt file and, consequently, its own sitemaps. We gathered webpages available in the sitemap of the section notícias (news). Other corpora might be collected from sections such as economia (economics), whose robots.txt file can be found at https://economia.uol.com.br/robots.txt, or carros (cars), whose robots.txt file can be found at https://www.uol.com.br/carros/robots.txt, for instance\textsuperscript{14}. This is one of the tasks that we intend to accomplish in future work to make our corpus more comprehensive.

After having obtained the list of URLs to be collected, we proceeded to the corpus compilation using the scraper described in the previous section. As mentioned above, 202,541 unique comments were collected, contemplating 5,112 news stories. Table 2.2 displays the number of comments per year and per semester in the corpus. It is clear that this version of the corpus is somewhat imbalanced per year, since the number of comments from 2017 (89,269) is considerably higher than those from 2016 (53,147) and 2018 (60,099)\textsuperscript{15}. Also, Figure 2.3 shows the number of comments per month in the corpus, where the temporal imbalance is again evident. However, when partitioning the data into semesters, this imbalance slightly

\textsuperscript{14} These URLs were accessed on August 2019 and might change at any time.

\textsuperscript{15} There are also 26 comments posted in 2019, due to the fact that some news published in the last days of 2018 were still receiving comments in the first days of 2019.
decreases, with only two semesters out of the range of 27,000-37,000 comments. In future releases, we expect to correct these issues and publish more temporally balanced corpora.

Table 2.2: Number of comments per period (year and semester) in the second version of the corpus Xereta.

<table>
<thead>
<tr>
<th>Year</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comments</td>
<td>53,147</td>
<td>89,269</td>
<td>60,099</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semester</th>
<th>1st</th>
<th>2nd</th>
<th>1st</th>
<th>2nd</th>
<th>1st</th>
<th>2nd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comments</td>
<td>36,149</td>
<td>16,998</td>
<td>27,124</td>
<td>62,145</td>
<td>32,861</td>
<td>27,238</td>
</tr>
</tbody>
</table>

Figure 2.3: Number of comments per month in the second version of the corpus Xereta. The dashed line indicates the mean value (5,626.14).

The corpus includes 6,689,806 word tokens, averaging 33.2 words
per comment, and 109,849 word types. The top ten frequent types in the corpus are listed in Table 2.3. As expected, the most frequent types are all short functional words or high frequency verbs, such as “o” (both a definite article and a pronoun), “que” (both a conjunction and a pronoun), “não” (meaning both “no” and “not”) and “é” (present tense, third-person singular form of the verb “ser”, which means “to be”). The high frequency of these words does not say much about the corpus. However, Table 2.3 also shows the most common nouns in the corpus. This list displays several words linked to Brazilian politics, including the name of two former presidents (Lula and Temer) and a political party (PT).

Table 2.3: Most frequent words (total) and nouns in the second version of the corpus Xereta.

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Noun</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>o</td>
<td>229,836</td>
<td>brasil</td>
<td>25,627</td>
</tr>
<tr>
<td>que</td>
<td>227,927</td>
<td>lula</td>
<td>20,618</td>
</tr>
<tr>
<td>de</td>
<td>212,571</td>
<td>páis</td>
<td>19,352</td>
</tr>
<tr>
<td>e</td>
<td>200,940</td>
<td>povo</td>
<td>18,351</td>
</tr>
<tr>
<td>a</td>
<td>188,481</td>
<td>pt</td>
<td>15,451</td>
</tr>
<tr>
<td>não</td>
<td>114,138</td>
<td>governo</td>
<td>11,788</td>
</tr>
<tr>
<td>é</td>
<td>110,055</td>
<td>anos</td>
<td>11,009</td>
</tr>
<tr>
<td>do</td>
<td>98,838</td>
<td>temer</td>
<td>10,821</td>
</tr>
<tr>
<td>para</td>
<td>74,328</td>
<td>dinheiro</td>
<td>10,511</td>
</tr>
<tr>
<td>se</td>
<td>71,322</td>
<td>presidente</td>
<td>10,402</td>
</tr>
</tbody>
</table>

In total, the comments present in the corpus received 407,250 positive evaluations (“likes” or “thumbs up”), averaging a mere 2.01 per comment. Interestingly, the distribution of positive evaluations among comments appears to follow a “rich-get-richer” pattern: the most liked comment received 297 thumbs up, while 109,603 comments did not receive any positive evaluation.
Figure 2.4: Fragment of the corpus Xereta displayed in the spreadsheet editor LibreOffice Calc.
The number of different usernames in the corpus is 47,393, averaging 4.27 comments per username. To comment on UOL news, readers must register in the website with a unique username – that is, it is not possible for more than one reader to share the same username. This ensures that, in principle, all the comments assigned to the same username were posted by the same reader. On the other hand, it is possible that a same user registered herself more than once, using two different accounts. Again, there seems to be a “rich-get-richer” pattern: the most active reader in the corpus is the writer of no less than 948 comments, while 27,221 users interacted only once. Lastly, we highlight the fact that 40,764 (20.13%) of the comments are actually replies to other comments, which suggests a remarkable interaction between UOL’s readers.

As an illustration, Figure 2.4 depicts a fragment of the corpus Xereta displayed in the spreadsheet editor LibreOffice Calc.

Comparison with other corpora of comments from news portals

The corpus Xereta is not the first corpus of news portals comments ever made available. Here we mention and compare other relevant corpora that contain this type of text. We do not cite data collections that were performed only for the purpose of a specific study and were not made available in corpus format to the scientific community.

At the time of publication of this dissertation, the SFU Opinion and Comments Corpus (SOCC)\textsuperscript{16} (Kolhatkar et al., 2019) is probably the larger corpus of comments available. It is focused on English-written Canadian media sources and includes 663,173 comments, comprised in a five-year period (from January 2012 to December 2016). A small portion of the corpus (1,043 comments)

\textsuperscript{16} Available at https://github.com/sfu-discourse-lab/SOCC.
is also annotated for constructiveness, toxicity, negation and appraisal. Differently from the corpus Xereta, however, the articles considered by SOCC are all opinion articles, not hard news articles.

The Yahoo News Annotated Comments Corpus (YNACC)\textsuperscript{17} (Napoles et al., 2017) is another large and well-curated corpus of news readers comments written in English. It contains 521,608 comments posted in response to Yahoo! News articles. Among these, 9,160 are annotated for sentiment, persuasiveness and tone. To illustrate how laborious and expensive the process of annotation of such a corpus is, the authors report that this task was performed by 26 professional trained editors and 495 untrained crowdsourced workers. Despite being a large corpus, the YNACC is not suitable for diachronic analysis, since it only includes comments for articles published in April 2016. Another corpus of comments from news portals in English is the SENSEI Social Media Annotated Corpus (Barker et al., 2016). This corpus is not comparable to the previous ones (including Xereta), since it contains only 1,850 comments – all of them annotated – collected from a small set of 18 articles from the British news portal The Guardian.

Cotterell et al. (2014) present an Algerian Arabic-French code-switched corpus\textsuperscript{18} that contains 339,504 comments, from which 1,000 are annotated for word level language identification. Fišer et al. (2018), as part of the Janes project\textsuperscript{19} – which aims to develop language resources and tools for Slovene user generated content –, describe a corpus containing 299,219 comments from Slovene news portals and encompassing a period of eight years (from March 2007 to January 2015)\textsuperscript{20}. This corpus is entirely tokenised, sentence seg-

\textsuperscript{17} Available at https://github.com/cnap/ynacc.
\textsuperscript{18} Available at https://github.com/ryancotterell/arabic_dialect_annotation.
\textsuperscript{19} Available at http://nl.ijs.si/janes.
\textsuperscript{20} The news comment corpus Janes-News 1.0 is available at https://www.
mented, word normalised, morphosyntactically tagged, lemmatised and annotated with named entities, which probably makes it the largest annotated corpus of comments up to the publication date of this dissertation.

Besides Xereta, at least two corpora of online news comments in Portuguese have already been compiled. SentiCorpus-PT\textsuperscript{21} (Carvalho et al., 2011) is an European Portuguese annotated corpus of comments to political debates. It includes 2,800 comments from a short period of time (ten days in September 2009). ComentCorpus (Pedro, 2018) is composed by 6,185 comments manually collected from 90 news articles related to the impeachment process of Dilma Roussef, in Brazil. It contains comments from between January and July 2016. Since its content is entirely related to politics, it cannot be considered a corpus of general comments. SentiCorpus-PT and ComentCorpus are focused on opinion mining and irony, and, therefore, are annotated with semantic-discursive information.

Table 2.4 summarizes the main characteristics of the above-mentioned corpora. It becomes evident that much remains to be done in this domain. In particular, there is a lack of large general and temporally broad corpora of news comments in Portuguese, and the corpus Xereta comes to (partially) supply this demand. In future work, we intend to follow the steps of most of the other corpora and add annotation to at least a small portion of our corpus.

2.4 Research possibilities

The two resources described above hold potential for use in various fields of knowledge, including linguistics, communication and media studies, sociology, political science and digital humanities. In this section, we present an overview of a few studies that could employ

\textsuperscript{21} Available at https://github.com/davidsbatista/REACTION-resources
Table 2.4: Basic information regarding different corpora of news comments. The corpora are ordered according to the total number of comments.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language</th>
<th>Number of comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFU Opinion and Canadian Comments Corpus</td>
<td>Canadian, English</td>
<td>663,173 (1,043 annotated)</td>
</tr>
<tr>
<td>Yahoo News Annotated Comments Corpus</td>
<td>English</td>
<td>521,608 (9,160 annotated)</td>
</tr>
<tr>
<td>Algerian Arabic-French code-switched corpus</td>
<td>Algerian, Arabic/French</td>
<td>339,504 (1,000 annotated)</td>
</tr>
<tr>
<td>Janes-News 1.0</td>
<td>Slovene</td>
<td>299,219 (completely annotated)</td>
</tr>
<tr>
<td>Xereta</td>
<td>Brazilian, Portuguese</td>
<td>202,541 (not annotated)</td>
</tr>
<tr>
<td>ComentCorpus</td>
<td>Brazilian, Portuguese</td>
<td>6,185 (completely annotated)</td>
</tr>
<tr>
<td>SentiCorpus-PT</td>
<td>European, Portuguese</td>
<td>2,795 (completely annotated)</td>
</tr>
<tr>
<td>SENSEI Social Media Annotated Corpus</td>
<td>British, English</td>
<td>1,850 (completely annotated)</td>
</tr>
</tbody>
</table>

our Web scraper and/or our corpus, hoping that these ideas might turn into actual research projects.

Many of the potential studies mentioned in the next paragraphs can be accomplished using the already existing corpora analysis tools. Figure 2.5 shows concordance lines of the corpus Xereta in AntConc, a freeware tool for carrying out corpus linguistics research.
Figure 2.5: Concordance lines (fragment) of the corpus Xereta at AntConc using the search term Brasil.
Diachronic research

As mentioned earlier, a diachronic corpus “is a collection of texts including information on the time period to which they relate, e.g. the publication date of a document” (Trevisani and Tuzzi, 2018, p. 130). The corpus Xereta is, by definition, a diachronic corpus, since it provides temporal information concerning the publication of both the news article and the comment. Figure 2.6 displays a fragment of our corpus, highlighting the meta-information regarding date and time of publication of each comment. Also, the files output by our Web scraper contain this information as well.

Figure 2.6: Fragment of the corpus Xereta highlighting comments’ temporal (date and time) information.

Being a diachronic corpus, Xereta might be employed in research aiming to analyze various phenomena (whether linguistic or not) from a temporal perspective. In the field of linguistics, it is clear that research on language variation and change can benefit from temporal information, which can be used to study how language evolved over a period of time. Since news portals comments are part of a relatively new text genre, any corpus containing such texts will cover a short time span – at most a few years. This obviously puts restrictions on the kinds of research on language variation and change that can be performed. However, previous research (e.g. Cunha et al., 2011; Danescu-Niculescu-Mizil et al., 2013; Eisenstein et al., 2014) shows that language use quickly evolves in social media, spreading across social network connections. Therefore, the investigation of the phenomenon of language variation and change in an
online environment, such as the one presented here, has proved to be a very fertile field, particularly in the domain of the lexicon. One practical example is the analysis of neologisms on the Internet (e.g. Rumšienė, 2004; Zhang et al., 2013).

Besides that, one interesting potential line of research is related to linguistic style accommodation – that is, the fact that participants in interactions “tend to nonconsciously converge to one another’s communicative behavior”, coordinating “in a variety of dimensions including choice of words, syntax, pausing frequency, pitch and gestures” (Danescu-Niculescu-Mizil et al., 2011, p. 745). The authors of the study just cited confirmed the hypothesis of linguistic style accommodation in a large dataset of Twitter conversations, and similar studies could analyze this phenomenon in Brazilian Portuguese, focusing on the genre comment in news portal as well.

Regarding fields other than linguistics, we might mention the possibility of using our corpus to investigate changes in readers’ behavior through time. Examples are behaviors such as bullying, incivility, harassment and disrespect, typically present in online discussions and comment boards (see Sarmento and Mendonça, 2016): do they exhibit the same characteristics across time? How do changes in external circumstances, like political or economic scenarios, influence the way readers interact with news articles and with other readers? Studies have shown a sharp rise in political polarization in Brazil in the second decade of the 21st century (see Hunter and Power, 2019): is it possible to corroborate this from the analysis of news portals comments over time? These are all issues that can be explored using our resources.

**Positively and negatively rated comments**

In the UOL news portal, readers are able to positively rate a comment by giving it a “like” (or “thumbs up”). Figure 2.7 depicts two comments extracted from an UOL news story, each having been
positively rated by eight readers, while Figure 2.8 highlights positive rating information in a fragment of the corpus. Unlike in other websites (such as Yahoo! News, shown in Figure 2.1), UOL readers, at the time of the writing of this research, do not have the possibility of rating comments negatively, which is the reason why our corpus does not include a column dedicated to the report of “dislikes” (or “thumbs down”). However, a slight change in the code is enough to include a column devoted to this information when collecting from sites that offer this possibility of interaction to their users.

Figure 2.7: Positively rated comments in the UOL news portal. As highlighted in the image, each of these comments received eight “likes” (or “thumbs up”).

Previous studies have investigated linguistic, stylistic, and discursive characteristics of positively and negatively rated replies in social media (e.g. Siersdorfer et al., 2010; Möller et al., 2019) and in question-and-answer platforms such as Stack Overflow (e.g. Calefato et al., 2015) and Yahoo! Answers (e.g. Shah and Pomerantz, 2010), usually employing a computational approach and with the purpose of predicting comment/answer quality. In addition,
more fine-grained qualitative studies could be conducted in this area. Since our corpus provides not only the number of positive evaluations received per comment, but also other meta-information such as the headline and the URL of the corresponding article, it is possible to comparatively analyze the ratings of comments in the same news story and examine which features make them gain more (or less) positive evaluations by other readers. A potential research question is: are the characteristics of more/less positively evaluated comments the same in articles mentioning different entities (such as two different politicians) or concerning different topics (sports, technology, entertainment etc.)? This type of research question could be addressed both through linguistic and social methodological approaches.

**Anonymity**

Usually, news portals readers are free to post comments using freely created usernames, and it is not difficult to observe a great variability among them: while some readers seem to draw upon their own names (in some cases, even their full names), others use eccentric nicknames that entirely prevent them from being identified. In Figure 2.9, which shows a fragment of the corpus highlighting commentators’ usernames, we find cases of readers who apparently use their...
own names, such as “Fabiano Palma” and “Rafa Ferreira S”, and others who, on the contrary, prefer completely anonymous usernames, like “comp”, “o predador” (the predator) and “Povo” (People).

Correa et al. (2015) show that the linguistic differences between anonymous and identified social media content are so significant that automated classifiers can be trained “to distinguish between them with reasonable accuracy” (p. 71). However, as far as we are aware, this phenomenon has not yet been analyzed for texts of the genre comment in news portal – certainly not for texts in Brazilian Portuguese. One hypothesis to be investigated is whether users with less identifiable usernames tend to exhibit more aggressive and hostile behaviors than (apparently) less anonymous commentators.

The onomatological analysis of commentators’ usernames also seems like an interesting topic for research in its own right. According to Hooker (2019), social media usernames may be completely
random, but can also be used to refer to attributes or characteristics of the users, their preferences, significant dates or events, or their actual names. The author adds that previous research shows that “participants displayed their identities through the choice of their username” (Hooker, 2019, p. 80). For instance, several usernames found in our corpus are somehow expressing political preferences.

The question then arises how they relate to the behavior of commentators in the platform.

**Analysis of conversation threads**

Another piece of meta-information present in the Xereta corpus and in the files output from the scraper indicates whether a comment constitutes a reply to a previously posted comment. Sequences of replies to the same comment can trigger true conversations between readers, which can also be analyzed. These can be called *conversation threads* or *conversation sessions*. Figure 2.10 displays a short conversation thread held by three readers of a particular news story at the UOL webpage. De Choudhury et al. (2009) observe that “[p]eople return to a video post that they have already seen and post further comments (say in YouTube) in response to the communication activity, rather than to watch the video again”. This behavior also occurs in the case of online news commentators – and Figure 2.10 itself is an example, since the original commentator reacted to the replies received by her/his own comment.

Figure 2.11 displays how this information is present in the corpus: in the field “Resposta a...” (*reply to*...), the value “0” indicates comments that are not replies to any other comments, while other values refer to the IDs of the replied comments.

Conversation threads pertaining to online social networks have

---

22 A few selected examples: “Anti PT 2014” (*PT* refers to the Workers’ Party, a major social-democratic political party in Brazil); “Mais a esquerda” (*more to the left*); “Apolítico” (*apolitical*).
Figure 2.10: Replies to a comment in the UOL news portal. In this example, users “Marcos A.” and “Gymno” reply to the original comment of “Oráculo Aprendiz” (the top comment). Then, “Oráculo Aprendiz” replies to both “Marcos A.” and “Gymno” (note that replies are displayed in reverse chronological order).

been the topic of previous research (e.g. Gómez et al., 2008; Ferguson et al., 2014). Caetano et al. (2019), for instance, work on the notion of attention cascades in WhatsApp groups, which can also be applied to the case of replies to comments in news portals. According to these authors, “an attention cascade begins when a user makes an assertion about a topic in a message to the group” and continues when “[o]ther users join and establish a conversation
Figure 2.11: Fragment of the corpus Xereta highlighting the field “Resposta a...” *(reply to...)*. The value “0” indicates comments that are not replies to any other comments, while other values refer to the IDs of the replied comments.

thread by explicitly replying to the root message” (Caetano et al., 2019, p. 29, emphasis in original). The use of the reply feature is then “a signal that the user’s attention was caught by the message she is replying to and that the cascade is a (semi-)structured representation modeling emergent patterns of collective attention” (p. 29, emphasis in original). One of the research questions proposed in the previously mentioned study is: “how different are attention cascades in political and non-political groups?” Similar analyses could be performed in the context of news portals comments, since, as far as we know, no studies in this direction have been carried out so far. Also, “[a]s in any real conversation of a group of people, attention may drift to other (possibly weakly related) topics as the conversation goes on” (Caetano et al., 2019, p. 29-30): a possible line of research could be to investigate how far from the news themselves these conversation threads can get.
Cross-corpus research

Cross-corpus research is the comparative use of different corpora to investigate the same (or similar) phenomena. By using this approach, researchers are able to compare different languages, different varieties of the same language or different periods of time, to give some examples. In this case, it seems opportune to employ our corpus to compare news portals comments with texts from other Internet genres or with comments taken from other contexts, such as letters sections in print media. Researchers might be interested, for example, in comparing the use of a specific linguistic item (a word, an expression, a grammatical structure etc.) between genres.

Berber Sardinha (2013) uses multidimensional analysis to compare different Internet and “traditional” (offline) genres, and to understand the variation between these genres and the writing style specific to them. In his work, however, only the genres “email”, “blog post”, “tweet”, “Facebook post” and “webpage” are considered. The availability of the resources presented here now allows the use of the text genre comment in news portal also to enter into this type of analysis. One of the observed characteristics of most Internet genres is their proximity to informal oral language. Researchers could compare comments obtained from our resources with pieces of text gathered from different sources on this regard, trying to observe how far/close from oral language these comments are when compared to other Internet genres.

In the above paragraphs, we offered some potential research ideas that may use the resources presented in this chapter. Of course, our intention is not to exhaust all the research possibilities, but rather to contribute by providing ideas and suggestions that could be followed in the future. Our main intention here is to promote and facilitate research on this new text genre, which seems
so interesting and relevant nowadays.

2.5 Concluding remarks

In this chapter, we present and describe two useful resources for research in the fields of corpus linguistics, studies on Internet communication, and digital humanities in general: (a) a Web scraper of comments from news portals and websites, developed as open source and as being free for use, modification and distribution; and (b) a freely available corpus composed of comments published at UOL, a major Brazilian news portal. Here, we show that our Web scraper is simple to operate and can be used even by individuals with limited computational skills. This is a feature that we consider very important, given our interest in promoting a broader access to digital content for researchers with little knowledge of programming and Internet data collection, especially in developing countries. We also describe a second (and, at the time of this publication, current) version of our corpus containing more than 200,000 comments (and respective meta-information) posted by readers as reactions to UOL’s news articles. This corpus makes it possible to analyze linguistic, textual, and discursive characteristics of the genre of news comments. Lastly, we mention some ideas about projects that could be carried out using the resources presented here, and we expect these ideas to be of interest to academics who wish to promote research on this developing Internet genre.

As far as we are concerned, ours is not only a contribution to corpus approaches in a language other than English, but also the first available large and general corpus of online news comments in Portuguese language. Besides being useful for academic research in linguistics, journalism, social science and other fields, a corpus of online comments may also contribute to the task of Web archival (Day, 2006; Webster, 2017), since it is preserving textual
material that risks being eliminated by companies (due to maintenance costs savings or website redesigns) or simply disappearing together with their news portal hosts in case of website closure or the bankruptcy of a company. In these cases, this corpus could be used, in the future, as a way of accessing human interactions and Web behaviors from past times.

This is of course a long-term and, ideally, an always-ongoing project. In future steps, we plan to keep improving the Web scraper for it to be able to collect comments from additional news portals, be them Brazilian (e.g. G1, Terra), Dutch (e.g. NU.nl) and other (e.g. The New York Times, Yahoo! News). Also important is the maintenance of the presently available scraper, since news portals might eventually modify their page structures, making the current code obsolete\textsuperscript{23}. Accordingly, we also intend to make available new versions of the corpus, with larger numbers of comments collected from different news portals and in different languages, ultimately enabling cross-language analyses of similarly collected data.

Finally, it is important to mention that the work involved in corpus compilation, preparation and maintenance is extensive, which is a reason why we invite other members of the community interested in this task to participate by improving codes, adapting our Web scraping tool so that it can be used to gather comments from more news portals, and building collections of comments extracted from different sources. All of this would be indispensable for the full development of our resources, and the encouragement for the creation of such a community of developers and users is also one of our goals.

\textsuperscript{23} Actually, this very incident occurred during the execution of this work: UOL slightly changed the structure of its news pages and, in 2019, we needed to restructure the module responsible for scraping from this site to get our Web scraper back into operation.
Acknowledgements

We thank Patrícia Rossini, who, while a doctoral student in Social Communication at Universidade Federal de Minas Gerais, motivated us to elaborate the Web scraper of comments and assisted us by testing preliminary versions in her own research, which culminated in her doctoral dissertation on political conversation, incivility and intolerance in digital environments (Rossini, 2017).