The RTBF Corpus : a dataset of 750,000 Belgian French news articles published between 2008 and 2021

Louis Escouflaire [1,2], Jérémie Bogaert [3], Antonin Descampe [1] et Cédrick Fairon [2]
[1] CENTAL - UCLouvain, [2] ORM - UCLouvain, [3] ICTEAM - UCLouvain
(louis.escouflaire/jeremie.bogaert/antonin.descampe/cedrick.fairon)@uclouvain.be
International Conference on Corpus Linguistics 2023 (JLC23)

Introduction

News corpora provide a large and diverse representation of written language that is essential for various fields of linguistics, such as lexicology, discourse analysis, sociolinguistics, and for training language models in NLP (Tognini-Bonelli, 2021). Their diachronic aspect allows for an investigation of language evolution over time and of its back and forth influence on society, making them valuable resources for linguistic research and teaching (Hilpert & Gries, 2016). Likewise, in media and journalism studies, news corpora are important both for qualitative and quantitative research (Conboy, 2007).

In this paper, we introduce the RTBF Corpus, a large diachronic corpus of 767,204 Belgian French news articles published between 2008 and 2021 by the Belgian public service media RTBF. We present the contents and structure of the corpus, along with the different layers of metadata available for each text. We also describe the three different versions of the articles available in the corpus (depending on the cleaning and preprocessing steps applied to the text). The RTBF corpus is freely available online in CSV format [1], for research and teaching purposes only.

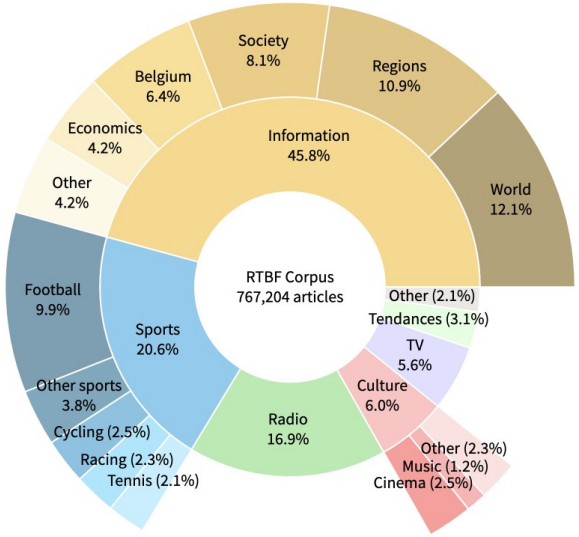

FIGURE 1 – Article distribution per feed (in percentages).

The RTBF Corpus

There exist few open-source French news corpora of considerable size and interest. The *Est Républicain* Corpus, released in 2009, is to our knowledge the largest freely available French news corpus with around 149 million words (Gaiffe & Nehbi, 2009). The articles it contains were published between 1999 and 2003 and between 2006 and 2011, and cover mostly local news of the Eastern part of France (Seddah et al., 2012), which may restrict the variety of topics mentioned in the corpus.

The RTBF Corpus is a freely available Belgian French news corpus, like the *Est Républicain* corpus. However, the RTBF Corpus contains more than 214 million words, which makes it 1.4 times larger. RTBF is a national media which covers all kinds of topics, among which international and Belgian

1. https ://dataverse.uclouvain.be/dataset.xhtml ?persistentId=doi :10.14428/DVN/PEVSSI

news, sports and culture. The corpus is chronologically continous (unlike the *Est Republicain* corpus) : the articles were published between 2008 and 2021, allowing for longitudinal investigations over the whole 2010 decade and more, and for analyses on news coverage of recent events (Escouflaire et al., 2022 ; Escouflaire et al., 2023). The corpus can also be a useful resource for experiments involving machine learning or requiring large amounts of data, such as news genre classification. Such large-scale resources are limited for the French language, making this corpus a useful asset.

The news director of RTBF officially handed us the rights over this dynamic corpus, which will be updated regularly with new articles published on the RTBF website. The corpus will be available online for free downloading, if the user agrees to follow the terms of use which are attached to the data. In short, the corpus may exclusively be used for research or teaching purposes (no commercial usage is permitted), it must be exploited in an ethical manner, and its user must inform RTBF of any planned publications related to the results and conclusions obtained from the corpus, and the media has the ability to request relevant comments and observations made by them to be included.

Corpus description

RTBF (*Radio-télévision belge de la Communauté Française*) is the public service broadcasting organization for the French-speaking community of Belgium. As a public service media, it is directly funded by the Belgian government and has three main missions : inform, educate, and entertain a public as large as possible in the French-speaking Belgian community. Besides operating television channels and radio stations, RTBF also operates a news website since 2008, on which web-only press articles are published daily. Through scientific collaboration with RTBF, we received access to the entirety of the web articles published on their website from 2008 to 2021. As shown in Figure 2, the publication rate has changed over the years, likely due to editorial movements inside the media. As of January 2023, between 1,000 and 1,500 articles are uploaded every week on their website.

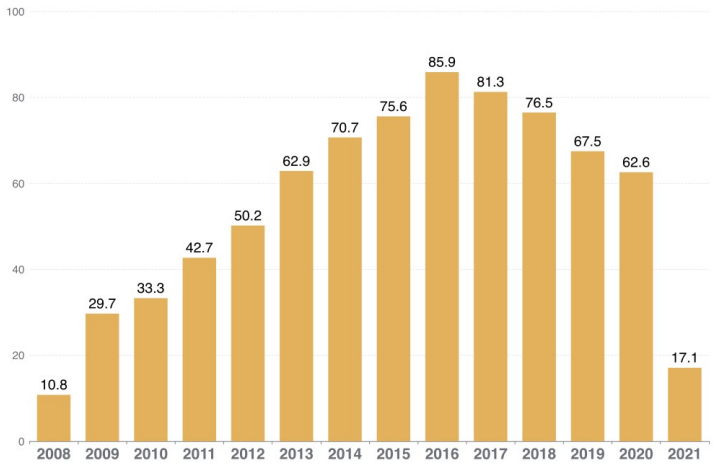

FIGURE 2 – Article distribution per year (in thousands of articles).

The corpus contains 767,204 articles, for a total of 214,072,620 words. The mean amount of words per article is 279. The full text of each article is in the *text* column of the corpus, along with seven columns consisting of different pieces of metadata attached to the article : *ID*, *title*, *publication date*, *signature*, *feed*, s*category* and *keyword*. We present the contents of all columns in more detail :

- ***ID*** : The article's internal ID, which is a number containing between 3 and 8 digits generated by the media source. This ID can be exploited to easily find its original page and layout on the RTBF website. A straight Google search with the format "*RTBF [ID]*" will return the article's link as the first result.
- ***title*** : The article's title, with an average length of 10 words. The lead paragraph and subtitles of the article are part of the main *text* column.

- ***publication date*** : The day on which the first version of the article was published on the RTBF website. The content of some articles may have been updated after their publication, but this information is not present in the corpus.

- ***signature*** : The name of the journalist who wrote the article, the radio or TV program it was originally derived from (cfr. *feed) further*, or the original source which the information is attributed to. About 43% of all articles in the corpus are signed solely or in collaboration with *Belga* or *AFP* (Agence France-Presse), two renowned press agencies delivering straight news. A great number of press agency dispatches are published daily on the RTBF website, with or without enrichments made by RTBF journalists. 1% of all articles were signed by a person or an entity who signed only one article in the corpus.

- ***feed*** : The structural and topical section of the RTBF website in which the article was published. The entire corpus is divided into 7 feeds : *Information*, *Sports*, *Radio*, *Culture*, *TV*, *Tendances* and *Other*. Their percentage distribution in represented in the inner circle of Figure 1. The *Information* feed, which is also the default feed on the RTBF website, contains mostly news belonging to the *information* genre of journalism (Grosse, 2001), i.e. content that is usually considered objective and factual by the source and by readers. Most press agency dispatches are found in this feed. However, the *Information* feed also contains 5,000 articles belonging to the *opinion* genre, i.e. op-eds or columns (those articles are marked with the *chronicles* or *opinions* category tag). The *Sports* and *Culture* feeds also include some amount of opinionated articles (*chronicles)*, representing respectively 416 (0.2%) and 2,800 (6%) of their articles. Then, the *Radio* and *TV* feeds contain articles which were derived from programs broadcasted on RTBF radio or television channels. *Tendances* contains lifestyle articles (e.g. fashion, food, technology, health, travel). The *Other* feed groups articles that were not classified into the 6 other feeds, among which many press releases and weather reports.

- ***category*** : Further classification related to the article's topics. Categories can help to group articles into more specific themes, as shown on the outer circle of Figure 1. Categories with small amounts of articles (e.g. *Basketball*, *Gaming*) were not represented on Figure 1, but were instead grouped into the *Other* category of the feed (*e.g. Other sports*). For the *Radio* and *TV* feeds, categories (not shown on Figure 1) are related to the specific TV channel or radio station from which the broadcast-related articles were derived (e.g. "Matin Première").

- ***keyword*** : Word or phrase attributed to some articles by the journalist or source who wrote them to further annotate their content (e.g. *Europe*, *environment*, *Internet*). About 28,5% articles in the corpus were assigned a keyword.

Corpus cleaning and preprocessing

Three different versions of the article's full text are available in the corpus. Each version contains the text at different stages of data cleaning and preprocessing :

- *HTML* text : The raw version of the text directly received from the RTBF, containing HTML tags. This version of the text can be used for research dealing with metatextual information (e.g. words highlighted in bold or italic, headings).

- *cleaned* text : The version of the text after multiple steps of data cleaning. To get this version, we removed HTML tags and fixed bugs related to text encoding of special characters. Some metatextual elements are still included in this version, such as signatures added at the end of some texts and links referring to other articles of the website.

- *preprocessed* text : The version obtained after multiple preprocessing steps. Most signatures at the end of articles (e.g. *with Belga*) and links to other articles (e.g. *Read also : ...*) were systematically removed from the articles. Numbers and URLs were replaced by placeholder tokens, <NUM> and <URL>. This version should be used carefully, as some elements that

may be relevant for some experiments could be missing from the text. Those preprocessing steps were initially derived and designed for the creation of the InfOpinion corpus (*Bogaert* et al., *2023*), a subcorpus of the RTBF corpus fit for a text classification task.

Notes

The RTBF Corpus was initially created in the context of a PhD program carried out at UCLouvain, in partnership with RTBF. The research is funded by FRS-FNRS (Belgian National Fund for Scientific Research) and conducted by PhD student Louis Escouflaire at ILC (Institute for Language and Communication) under the guidance of Antonin Descampe (Observatory for Research on Media and Journalism) and Cédrick Fairon (Center for Natural Language Processing).

References

Bogaert, J., Escouflaire, L., de Marneffe, M.-C., Descampe, A., Fairon, C. & Standaert, F.-X. (2023). TIPECS : A corpus cleaning method using machine learning and qualitative analysis. *International Conference on Corpus Linguistics 2023 (JLC23).*

Conboy, M. (2007). *The language of the news.* London : Routledge.

Escouflaire, L., Descampe, A., Fairon, C. (2022). L'évolution de la subjectivité linguistique dans le journalisme web du XXIe siècle : analyse d'un corpus belge francophone d'articles de 2010 à 2021. *JADT 2022 : 16th International Conference on Statistical Analysis of Textual Data.* Naples, Italy.

Escouflaire, L., Descampe, A., Lits, G., Fairon, C. (2023). Analyzing the semantic evolution of bias in French news articles using word embeddings. *Digital Humanities Benelux 2023.* Brussels, Belgium.

Gaiffe, B. & Nehbi, K. (2009). Le corpus de l'Est Républicain. *Technical report, Atilf.* http ://www.cnrtl.fr/corpus/estrepublicain/.

Hilpert, M., & Gries, S. T. (2016). Quantitative approaches to diachronic corpus linguistics. The Cambridge handbook of English historical linguistics, 36-53.

Küppers, A., & Ho-Dac, L. M. (2011). Un corpus de presse francophone pour l'étude de l'impact d'Internet sur les pratiques langagières. *CJC Praxiling : Corpus, données, modèles : approches qualitatives et quantitatives.*

Seddah, D., Candito, M., Crabbé, B., & Anguiano, E. H. (2012). Ubiquitous usage of a broad coverage French corpus : Processing the Est Républicain corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3249-3254.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at work.* Amsterdam/Philadelphia : John Benjamins.