2 Building a multilingual multi-genre corpus for interdisciplinary research

Data selection, collection techniques, and ethical considerations

Barbara De Cock, Eva Rolin, Ferdinand Teuber, and Romane Werner

1 Rationale behind the data selection[1]

In this section, we detail the rationale behind the data selection for the TrUMPo project.

The project aims to analyse the use of the term *populis** in society in order to understand in which contexts and situations this notion is used, which meaning it conveys in actual discursive practices, and how it circulates within the public debate (see Aulit *et al.*, Chapter 6 in this volume, for a more detailed description).

As we aim to understand how discourse about populism plays a crucial role in the mediatised political debate in European democracies, we compare discourses in the national public sphere of three countries – Belgium, France, and Spain – offering four cases of investigation as Belgium can in some respects be considered to have a separate Dutch-speaking and French-speaking public sphere. The choice of these case studies is justified by the fact that they allow for comparing countries presenting different views on the way the State represents the people or the nations/nationalities which are part of the people (Niessen, 2022). Furthermore, in these countries, actors and actions attached to different ideologies have regularly been qualified as "populist" (see De Cock, 2026).

Given our interest in analysing the meanings of *populism*, we have searched texts including *populism* or *populisms* in the languages concerned (namely, Dutch, French, and Spanish), as well as derived adjectives and nouns (see Table 2.1). We do not include data in the third national language of Belgium, namely, German, nor in the co-official languages of Spain. This choice is motivated by their very limited presence

DOI: 10.4324/9781003610489-3

Table 2.1 Different forms of *populist* and *populism* in the languages included in this study

	Abstract noun	*Person-referring noun*	*Adjective*
Belgium Dutch-speaking	*populisme* *populismes* *populismen*	*populist* *populisten*	*populistisch* *populistische*
Belgium French-speaking and France	*populisme(s)*	*populiste(s)*	*populiste(s)*
Spain	*populismo(s)*	*populista(s)*	*populista(s)*

(for Belgium) or total absence (for Spain) in the national parliament. A further issue for German data would have been the difficulty of finding German data concerning Belgium in the vast amount of German language tweets (see Section 2.1).

In these four political contexts, we focus on the year 2019 because of the multiple elections that have occurred in all three countries that year, including EU elections, national elections, and elections at the substate level. The focus on an electoral year is justified by this being the moment where political competition in a democracy is at its peak and when discursive production is especially intense.

To account for the variety of uses, it is evident that various loci of production (so-called arenas of public discourse (Badouard *et al.*, 2016)) are to be included. The choice was made to include parliamentary debates, media, and Twitter (now X). Parliamentary debates are included because they are the central discussion space of the democracies involved in this study. Even though the access to this arena is also limited by various sociopolitical factors, it is in principle a public arena, whose proceedings are made public. While not necessarily directly reproduced, the discourses produced there play a crucial role in public life. From the perspective of political sciences, it is paramount to not only have the transcripts themselves, but also information concerning the speaker and their political party, which is relevant for subsequent analyses concerning the use of *populis** by certain political orientations and/or by government vs. opposition (see annotation system described in Hambye *et al.*, Chapter 3 in this volume). Being parliamentary regimes, all three chambers share fundamental legislative, representative, oversight, and communicative functions. To fulfil these functions, they have at their disposal various instruments and structures, such as plenary debates, legislative proposals, and parliamentary committees. So, while their levels of activity and institutional strength may differ, the three chambers are sufficiently similar to allow for comparisons on this basis.

Table 2.2 Media sources included in the data collection

	Public-owned media	*Elite media*	*Serious-popular media*	*Free daily and traffic distributed newspaper*
Belgium – Dutch-speaking	vrt.be/vrtnws	De Standaard	Het Laatste Nieuws	Metro
Belgium – French-speaking	rtbf.be/info	Le Soir	La DH	Metro
France	France Television / Radio France	Le Monde	Aujourd'hui en France/Le Parisien	20 minutes
Spain	www.rtve.es/noticias	El País	elespanol.com/elconfidencial.com	20 minutes

From the perspective of communication studies, the choice of mass media to be analysed is of the utmost importance. As shown in Table 2.2, our mass media corpus comprises four media per country. Concretely, have selected (1) public-owned media, that is, public broadcasting service which "cannot retreat from public information" (Coleman, 2002: 89), (2) elite media, that is, "established media [that] typically operate within the consensus, displaying a clear tendency to adopt the hegemonic-official position on issues" (Wiesslitz & Ashuri, 2011), (3) "serious-popular" media (Sparks, 2000: 15) which include newspapers "which have a strong stress upon visual design and contain a large dose of scandal, sports and entertainment but that still demonstrate all, or a significant part, of the same inventory of news values as their more serious cousins" (Sparks, 2000: 15), and (4) free daily and traffic distributed newspapers. Free dailies, "whose most well-known titles are *20 minutes* and *Metro*" (Lamour, 2019: 1167), are given "free of charge to readers" (Serazio, 2009: 649). They concentrate on news and current affairs (Bakker, 2013). They are "short, illustrated and easily consumed content, distributed in large urban areas during the morning or evening peak hours" (Lamour, 2017: 1). Lamour describes them as "popular mass medium" (Lamour, 2019: 1167), the "only print medium able to capture a mass audience equivalent to prime-time TV news bulletins" (Lamour, 2019: 1169). Upon deciding on the extent of our data selection, we have come to realise that these journals are, due to their free distribution in public spaces and public transport, widely read and therefore an important source of information. In 2022, Belgian *Metro* has ceased its activities due to too limited income from publicity (as a consequence of the

pandemic). In 2024, French *20 Minutes* has ceased to publish a paper version and has become an entirely online media. These media outlets have then come to occupy a different place in news consumption and, for the Belgian case, downright lost their relevance. Spanish *20 Minutos,* by contrast, still exists in a paper version in major cities. The different current situation of these free media makes that they are no longer comparable, as they were in 2019.

The choice of Twitter is in the first place justified by the fact that Twitter plays a key role in the cycle of political information (Chadwick, 2013) and political discussion since it is a professional space, used by politicians but also by stakeholders, media, etc. (Jacobs & Spierings, 2018). Especially in the case of public figures, we know that their tweets may also be written by members of their communication team, and some politicians explicitly acknowledge this in their profile. However, the public figure remains the principal of the message in Goffman's terms (Goffman, 1981), that is, the person publicly responsible for the content. Twitter is used by journalists (Vis, 2013) and therefore contributes to the coverage of the political sphere by media professionals. Even though some media or journalists have left Twitter/X after Elon Musk bought the platform and following the 2024 U.S. presidential elections, many journalists remain on this platform, although some only in an observatory capacity. Furthermore, Twitter is used by political actors themselves (Roginsky & De Cock, 2015) but also by citizens who have a political interest (Roginsky & Huys, 2015). It is worth bearing in mind that only a minority of people are indeed Twitter subscribers. If numbers should be used with caution due to the difficulties to find official statistics provided by the platform, statistics provided by marketing websites indicate that, in 2019, 5.9% of the Spanish population, 6.4% of the Belgian population, and 6.1% of the French population were Twitter users. A second reason to use Twitter is that it contributes to "hybridising" and revisiting the "traditional cycle of political information" (Chadwick, 2013). Politicians can use Twitter to spread information themselves, without using press conferences or, indeed, media outlets at all. They often also retransmit news or comment on it.

The parliamentary debates and the mass media could easily be linked to one of the (sub)national public spheres we want to analyse; however, this is less straightforward with tweets. Indeed, while one can enable geolocalisation of one's tweets and/or mention one's location in the Twitter biography, it is not compulsory to do so, making tweets much less clearly ascribed to a specific geographical area. In addition, the geographical indications available bear some questions. The localisation mentioned in one's biography is not necessarily univocal. Indeed, users can choose to mention various locations, representing various aspects of their life (e.g., place of

origin vs. current residence or residence vs. workplace). Moreover, one can tweet while being for a longer or shorter time in a location different than the one indicated in the biography. Finally, some locations are downright fictitious, ranging from toponyms expressing a sociopolitical view (e.g., *Banana republic*) to toponyms that reveal personal hobbies or interests (e.g., *Hogwarts*, the school from the Harry Potter series) (see also Graham *et al*., 2014).

For both practical and content-related reasons, we then decided to first gather all the 2019 tweets with *populis** and then apply various filtering techniques to them.

2 Collection and data treatment techniques

2.1 Tweets

The tweet collection was conducted via the company Tweet Binder, through which we obtained a corpus of tweets in Excel format. Each file corresponds to a specific form of *populis** (e.g., *populisme, populiste*) in 2019 and includes, for each occurrence of *populis**, various fields such as the date, the tweet content, the language used, the username, the number of likes, and the location.

A first round of filtering was carried out using OpenRefine, focusing on the "language" variable to keep only tweets written in our target languages: Dutch, French, and Spanish (whenever applicable). Indeed, some of the forms of *populis** also exist in languages outside the scope of this study, such as *populisten* in Swedish or *populisme* in Catalan, making a language filtering a necessary step. Concerning tweets that contained geolocalisation information (a minority of the tweets), we then refined this filtering further by using geolocation, keeping only tweets in Spanish from Spain, in French from Belgium and France, and in Dutch from Belgium.

The second filtering step was applied to tweets that contained a reference to a localisation other than geolocalisation, for example, mentioning a hometown in one's Twitter biography. This step relied on a manually curated list of cities and towns in Spain, France, and Belgium, compiled from various online sources. Most tweets originated from larger cities rather than smaller villages. This step was implemented in Python using the Pandas library, which is specifically designed for data analysis and manipulation. Pandas offers powerful data structures and functions for efficiently handling structured data, such as tables. During this process, we also noticed that location names were often written in various ways (e.g., *Belgium, Belgïe, Belgique,* and *Belgiiiique*). To account for this variability, we incorporated these noisy variations into our list to improve the accuracy of location matching.

Once each occurrence had been divided into three files (one per language), we applied regular expressions (regex) to transform certain variables, such as converting date formats (e.g., "Sat 05 Jan 2019 13:39:44 GMT") into a simpler day/month format. Thanks to these new variables, we were able to generate visualisations showing minimum, maximum, and average tweet volumes for a given period of time, helping us identify spikes, that is, specific days or months when tweet activity increased. These spikes could then be linked to events identified within the tweet corpus.

Finally, in order to create a specific subcorpus for more qualitative analyses, a random sampling of 1,000 tweets was realised, complemented with one further criterion, namely, the relevance of the tweets to the circulation of discourses on populism in Belgium (both the French-speaking and Dutch-speaking parts), France, and Spain. This criterion is motivated by the ecological approach (see Treré, 2019) we adopt to the way Twitter users can receive tweets. We then aim not to select only tweets published in Spain, France, and Belgium, but rather to select tweets that would be read and considered a relevant part of the reading experience of users from those countries. Indeed, social networks do not function at a specific geographical level (although they can be more popular in some regions than others), and users read messages posted from different countries. Thus, Spanish Twitter users can also frequently read messages from/concerning Latin-American countries. Moreover, such an ecological approach allows for the consideration of tweets that can be published by users who are currently not living in what they consider their homeland. This approach also enables us to take into account the considerable circulation between France and the French-speaking part of Belgium due to the huge interest for French media and politics in the French-speaking part of Belgium. Thus, following a discussion among at least two project members on each specific tweet in this 1,000 tweet dataset, we discard tweets that would be considered entirely irrelevant or not understandable at all (in terms of sociopolitical context) from the Spanish, French, or Belgian perspectives, for example, a tweet published in French concerning a leader described as populist in a French-speaking country with which Belgium has little contact. This leads us to establishing four subdatasets of 666 tweets each.

2.2 Media outlets

To collect both the media and parliamentary data, web scraping techniques have been used. These are adapted to the specificities of the respective websites.

The primary objective of this project is to analyse the usage of the concept of *populis** within public discourse in Belgium, Spain, and France.

To this end, a large-scale corpus of media discourse has been compiled and systematically analysed.

We first present the techniques used for article extraction and data preprocessing. The collection of media articles is carried out through automated web scraping from 16 selected newspapers, representing diverse political orientations and linguistic contexts across the three studied countries.

This process relies on Perl and Python scripts specifically designed to extract structured data from news websites. Given the heterogeneity of web page structures across different media outlets, a tailored approach is adopted, using a combination of HTML parsing techniques, regular expressions, and XPath/CSS selectors to locate and extract relevant textual content.

The extraction workflow follows these key steps:

1 URL Identification: Scripts such as *get_all_urls_le_monde.py* and *get_all_urls_el_pais_1.py* retrieved article URLs from sitemap indexes or dynamically generated news feeds.
2 HTML Parsing and Content Extraction: The BeautifulSoup library (Python) and regular expressions (Perl) help to navigate the HTML DOM,[2] isolating article titles, publication dates, and main body text while filtering out irrelevant sections (e.g., advertisements, navigation bars, and multimedia content).
3 Normalisation and Cleaning: Extracted texts are processed to remove non-linguistic elements, standardise the encoding (UTF-8), and convert data into structured formats (.txt,.csv,.json) for further analysis.

The collected articles are subsequently categorised into two datasets:

- A complete dataset, encompassing all extracted files.
- A filtered dataset, containing only articles in which at least one occurrence of the studied terms is detected.

The created scripts used in this process include *extraction_le_monde.pl*, *extraction_rtve.pl*, *extraction_rtbf.pl*, among others. These scripts ensure consistency across multiple sources while adapting to site-specific HTML structures.

To analyse the linguistic and discursive contexts in which *populism* is used, concordances are generated using the TextStat library. A concordance is a structured representation of a keyword's occurrences within a corpus, showing its left and right co-text to capture contextual patterns.

The concordance-building process follows these steps:

1 Tokenisation and Lemmatisation: Texts are processed to segment words while preserving multi-word expressions (see Table 2.3).
2 Keyword in Context (KWIC) Extraction: The TextStat library generates KWIC concordances, providing a window of context around each occurrence of the target lexeme.
3 Data Structuring: The output files are formatted as spreadsheets (.xlsx), including the following columns (see Figure 2.1):
   a Left Context (preceding words)
   b Lexeme (target word)
   c Right Context (following words)
   d Source File (original document name)
   e Date (when available)

This structured representation allows for both qualitative and quantitative analyses of discursive patterns related to *populism*.

A quantitative analysis of the occurrences of *populism* is conducted through descriptive statistical measures to assess frequency distributions and temporal trends in the data.

Table 2.3 Example of the tokenisation and lemmatisation of a sentence

A	*B*	*C*
Token	*Lemma*	*Grammatical_category*
les	le	DET (determiner)
discours	discours	NOM (noun)
des	de	PREP+DET
politiciens	politicien	NOM
populistes	populiste	ADJ (adjective)
influencent	influencer	VER (verb)
souvent	souvent	ADV (adverb)
l’	le	DET (elided determiner)
opinion	opinion	NOM
publique	public	ADJ
.	.	PONCT

	A	B	C	D	E	F	G
1	sentence	left_context	lexeme	right_context	source_file	url	date
2	"Le MR semble se diriger plus vers la droite et il y a le style de Georges-Louis Bouchez qui est un peu populiste."	qui est un pet	populiste	".	rtlinfo	https://www.	26/10/23

Figure 2.1 Example of the result of data structuring

2.3 Parliamentary debates

In order to compile a list of mentions of the term *populis** in the parliamentary arena of the three selected national contexts, we collect data from the websites of the most important parliamentary chambers of each country (Belgium: www.lachambre.be; France: www.assemblee-nationale.fr; Spain: www.congreso.es). Our approach relies on automated web scraping implemented through Python scripts that are tailored to the specific institutional contexts of each country, the selected document types, and the specific structure of each website. As our goal is to analyse comparable types of parliamentary texts across countries, we include the following categories in our data collection:

1 Transcripts of plenary debates;
2 Transcripts of parliamentary committee debates;
3 Written questions submitted by members of parliament to the government[3];
4 Legislative proposals and other documents debated or adopted by parliament (e.g., resolutions).

In practice, our workflow includes the following common steps:

1 URL Collection
  For all three parliamentary chambers, the first step involves retrieving lists of parliamentary documents from their respective websites. Depending on the website, we rely on Selenium (namely, for the Assemblée nationale), BeautifulSoup[4], and/or regular expressions in order to identify and extract links to the specific documents, for example, plenary debates or legislative bills.
2 Document Download and Storage
  Once URLs have been collected, the scripts download the corresponding documents, typically in PDF or HTML format. To download the documents, we develop custom Python scripts using tools like wget (a command-line program for downloading files) and the requests library (which provides more control when accessing websites). These scripts are tailored to the structure of each parliamentary website and the specific types of documents we have been collecting, such as plenary debates or legislative proposals. Once downloaded, the files are saved in a folder structure that reflects the three national contexts and the various document types. File names follow consistent conventions, typically including the country, document type, and the date of the debate or publication.

3 Content Filtering and Keyword Search
To identify documents relevant to the study of *populis**, a filtering step based on keyword detection has been implemented. Similar procedures to those for the media data are applied to collect the relevant data in datasheets.

As a result of these steps, our dataset for the parliamentary arena consists in two parts: (1) a unique data file with metadata (e.g., date, document type) and keywords in context for the search term, and (2) a complete collection of full text documents, for example, legislative proposals or plenary debate transcripts, in which *populis** occurs.

Given the small size of this corpus, no descriptive statistics have been developed.

3 Ethical issues related to collecting, archiving, and reusing the data

All the data collected for this research have been produced in public spaces and are publicly available. A considerable amount of the discourse producers are public figures (politicians, journalists, researchers, etc.) communicating in their professional and public capacity.

In our study, when considering ethical issues, we adopt a double distinction depending on whether the method was large or small scale, and on whether persons involved were public or private persons.

For more large-scale studies, such as collocation analyses, data are treated in an anonymised way in that the discourse producers' identity is not revealed. We have not removed all proper nouns in the corpora since those that appear in large-scale studies are typically highly frequent names of public persons (e.g., Orbán, Salvini). For more small-scale studies (such as argumentative studies or metaphor analyses), we pseudonymise the data of the discourse producers who are not public figures, as well as mentions of non-public persons in other authors' discourses in the sense that we attribute a label "non-public person" to those references (see also Hambye *et al*., Chapter 3 in this volume). In practice, this concerned especially the Twitter data since the speakers and references to concrete persons in parliamentary and media data only concerned public figures. (When non-public figures are included in the media, they are typically already represented in an anonymised manner.) In presentations and publications, such references within the tweets are replaced by @X.

Despite all our data being public, we have chosen not to retain the references to non-public figures since some of the utterances may be considered sensitive (e.g., calling someone populist). Indeed, discussions concerning *populis** can include strong expressions of political opinions.

Our decisions regarding these data have been inspired by the guidelines of the Association of Internet Researchers (Franzke *et al.*, 2020), as also discussed in Andone (2025).

While analysing the data, we have seen that some of the tweets are no longer visible at the time of analysing or publishing output. The author may have decided to remove a specific tweet or removed their account altogether. Other authors may have become blocked.

The data have been stored on university servers with restricted access, limited to project members. Given the complex and evolving intellectual property rights issues related to the Twitter and media parts of the dataset, only the parliamentary data can be made available beyond the project. All data will be destroyed after 15 years, considering that this is a reasonable time to create all the research outputs and possible follow-up research.

4 Insights for future projects

As it has become clear throughout the project, and as it will become clear in the following chapters, creating this international corpus comprising multiple arenas has greatly enriched our research on *populis**. However, creating a corpus across countries and arenas also implies challenges and requires sufficient resources. We discuss some insights for future projects in what follows.

The possibility to collect tweets in a massive way has varied greatly over the years, which makes it difficult to collect longitudinal corpora (see De Cock & Greco, 2025). When we started the project in 2020, purchasing the tweets via Tweet Binder was the only viable solution to obtain huge quantities of previously produced tweets. Sometime later, academics could use the "Tweet downloader" tool, which required little technical skills and allowed researchers to obtain considerable amounts of tweets for research purposes (see Filardo-Llamas *et al.*, 2025). However, this tool disappeared in February 2023, leaving at the time of writing in practice only paid options to obtain tweets, which raises both economical and ethical questions. (An academic access to X exists but has so far not proven to be a useful or usable route for this kind of research.)

In addition to the difficulties to obtain the tweets, we have also encountered challenges related to tweets that are in the data collection but are no longer available online, for example, because the original poster erased the tweet or closed their account altogether. This also holds for tweets to which a tweet in our corpus is an answer or a retweet, making the interpretation of such tweets inherently complicated. We have made the choice to acknowledge these limitations regarding possible interpretations, since they are intrinsic to the way the platform works.

As pointed out above, the fact that Twitter does not follow geographical boundaries is a challenge for research focusing on a specific geographic entity. However, we have also considered it an opportunity to analyse Twitter precisely as a platform that does not follow classical geographical boundaries and to fully acknowledge in our research design that discourses on/relevant for a specific country can originate outside its geographical boundaries.

When comparing different media, it is paramount to take into account that different newspapers and newspaper archives have different search tools. Some allow for searching the full content, whereas other search systems are rather based on keywords defined by the journal itself. While some overarching (often subscription-based) systems give access to various media in a similar way, such as Lexis or Europresse, none of these included the combination of countries and types of newspapers we aim for. These tools can of course be very helpful if they match your research interests, but we would like to draw attention to the blind spots they create in terms of analysed media and collected data. This has led us to develop our own collection techniques described in this chapter.

The creation of the parliamentary corpus requires more than any other corpus a profound knowledge of the political system under study, particularly in order to consider the relevance and function of different types of subcommissions that constitute spaces of discussion. The added value of working with an interdisciplinary research team is crucial in this respect. The inclusion or exclusion of certain specific spaces in the parliamentary arena ultimately depends on the research aims. In our case, the public nature of discourses at the moment of their production is a relevant factor. A comparative functional approach where one can compare the position spaces take up in parliamentary debate is paramount to assess this. On a more technical note, while parliamentary data are the most easily available in the sense that they do not require paid subscriptions from a perspective of democratic transparency, the underlying technical tools do not always allow for easy massive downloading of the data. Finally, the available transcripts are not transcripts created for research aims. Thus, they tend to not include elements that transcripts for the linguistic or interactional analysis of spoken data would require, such as pauses, hesitations, and repetitions or, if they include them, they do so with a degree of accuracy that is insufficient for linguistic research. Rather, the parliamentary transcripts are edited versions of the text as pronounced in the public arena. This editing can be limited to removing specific features of spoken language, such as pauses or repetitions, but can go further in some national contexts (Gelabert, 2004: 78–79). In the context of this study, this edited version has been useful but a study focusing on spoken language features such as prosodic features and repetitions would need a more detailed transcript.

In this chapter, we have outlined the way in which we have established a corpus involving various languages, various types of sources with their technical specificities, in the context of an interdisciplinary project. We are aware that the decisions we have made and procedures we have followed are not necessarily the ideal solutions for other researchers or other projects. Indeed, these decisions are inspired by the project aims and taken in a specific timeframe. As discussed above, some technical aspects have already evolved in the meantime, particularly access to Twitter/X data. Moreover, as also discussed in Andone (2025), standards and practices concerning ethical aspects of research are in constant evolution.

However, we do hope that sharing our meta-reflection on this data collection can provide useful suggestions and ideas for colleagues who wish to collect data of a somewhat similar nature.

Notes

1 This research is part of the ARC project 20/25–109 "Discourse, populism and democracy – Tracking the uses of 'populism' in media and political discourse" (TrUMPo), funded by the *Fédération Wallonie-Bruxelles*.
2 The HTML DOM (Document Object Model) is a tree-like representation of a web page's structure. It allows programming languages like JavaScript to access, modify, and interact with the elements of a webpage dynamically.
3 We included written questions in our data collection since they are a classic instrument in the hands of members of parliament. Yet, it turned out that in the Belgian case we could not find a single occurrence of *populis** in over 1,700 written questions submitted to the federal government in 2019. In the case of Spain, we excluded written questions from the analysis because they were only available in OCR (image-based) format, leading us to exclude these from the data collection.
4 BeautifulSoup is a Python library used for web scraping and parsing HTML and XML documents. It helps extract data from web pages in a structured way by providing tools to navigate, search, and modify the parsed HTML/XML tree.

References

Andone, Corina. (2025). The ethics of qualitative research in argumentation. In Corina Andone, Marianne Doury, Sara Greco, Kati Hannken-Illjes, & Menno Reijven (Eds.), *Qualitative research methods in argumentation studies* (pp. 199–210). Routledge. https://doi.org/10.4324/9781003502296

Badouard, Romain, Mabi, Clément, & Monnoyer-Smith, Laurence. (2016). Le débat et ses arènes. *Questions de communication*, *30*, 7–23. https://doi.org/10.4000/questionsdecommunication.10700

Bakker, Piet. (2013). The life cycle of a free newspaper business model in newspaper-rich markets. *Journalistica, 1*, 33–51. https://doi.org/10.7146/JOURNALISTICA.V7I1.15802

Bednarek, Monika, Schweinberger, Martin, & Lee, Kelvin K. H. (2024). Corpus-based discourse analysis: From meta-reflection to accountability. *Corpus Linguistics and Linguistic Theory*, *20*(3), 539–566. https://doi.org/10.1515/cllt-2023-0104

Chadwick, Andrew. (2013). *The hybrid media system: Politics and power.* Oxford University Press.

Coleman, Stephen. (2002). From service to commons: Re-inventing a space for public communication. In Jamie Cowling & Damian Tambini (Eds.), *From public service broadcasting to public service communications* (pp. 88–98). IPPR.

De Cock, Barbara. (2026). Do we think populists actually do something? A discursive analysis of populists as agents or patients in tweets in French, Spanish and Dutch. In Noelia Ramón Pérez, & María Pérez Blanco (Eds.), *Multilingual corpus research. Advances and challenges* (pp. 151–173). John Benjamins.

De Cock, Barbara, & Greco, Sara. (2025). Approaching social media argumentation from a linguistics informed perspective: Methodological aspects In Corina Andone, Marianne Doury, Sara Greco, Kati Hannken-Illjes, & Menno Reijven (Eds.), *Qualitative research methods in argumentation studies* (pp. 111–126). Routledge.

Filardo-Llamas, Laura, De Cock, Barbara, Hambye, Philippe, & Shchinova, Nadezda. (2025). "I am not populist". Mechanisms for the re-negotiation of category membership on Twitter. *Pragmatics and Society, 16*(5), 653–675. https://doi.org/10.1075/ps.24027.fil

Franzke, Aline Shakti, Bechmann, Anja, Zimmer, Michael, Ess, Charles M. & the Association of Internet Researchers. (2020). *Internet Research: Ethical Guidelines 3.0*. https://aoir.org/reports/ethics3.pdf

Gelabert, Jaime. (2004). *Pronominal and spatio-temporal deixis in contemporary Spanish political discourse: A corpus-based pragmatic analysis*. PhD thesis, Pennsylvania State University. https://etda.libraries.psu.edu/catalog/6383

Goffman, Erwing. (1981). *Forms of talk*. Basil Blackwell.

Graham, Mark, Hale, Scott A., & Gaffney, Devin. (2014). Where in the world are you? Geolocation and language identification in twitter. *The Professional Geographer*, *66*(4), 568–578. https://doi.org/10.1080/00330124.2014.907699

Jacobs, Kristof, & Spierings, Niels. (2018). A populist paradise? Examining populists' Twitter adoption and use. *Information, Communication & Society*, *22*(12), 1681–1696. https://doi.org/10.1080/1369118X.2018.1449883

Lamour, Christian. (2017). New for free in the late-modern metropolis: An exploration of differentiated social worlds. *Global Media and Communication*, 1–15. https://doi.org/10.1177/1742766517734250

Lamour, Christian. (2019). The legitimate peripheral position of a central medium, *Journalism Studies*, *20*(8), 1167–1183. https://doi.org/10.1080/1461670X.2018.1496026

Niessen, Christoph. (2022). Measuring evolving regional autonomy demands and statutes: Introducing the Sub-state Autonomy Scale (SAS). *Regional Studies, 56*(9), 1589–1603. https://doi.org/10.1080/00343404.2022.2056157

Roginsky, Sandrine, & De Cock, Barbara. (2015). Faire campagne sur Twitter. Modalités d'énonciation et mises en récit des candidats à l'élection européenne.

Les Cahiers du Numérique, *11*(4), 119–144. https://shs.cairn.info/revue-les-cahiers-du-numerique-2015-4-page-119?lang=fr

Roginsky, Sandrine, & Huys, Sophie. (2015). À qui parlent les professionnels politiques? La fabrique des publics des députés européens sur les dispositifs Facebook et Twitter. *Communication*, *33*(2). https://journals.openedition.org/communication/6051

Serazio, Michael. (2009). Free newspapers. In Christopher H. Sterling (Ed.), *Encyclopedia of journalism* (pp. 648–650). Sage.

Sparks, Colin. (2000). Introduction. The panic over tabloid news. In Colin Sparks & John Tulloch (Eds.), *Tabloid tales. Global debates over media standards* (pp. 15–40). Rowman & Littlefield Publishers.

Treré, Emiliano. (2019). *Hybrid media activism. Ecologies, imaginaries, algorithms*. Routledge. https://doi.org/10.4324/9781315438177

Vis, Farida. (2013). Twitter as a reporting tool for breaking news. *Digital Journalism*, *1*(1), 27–47. https://doi.org/10.1080/21670811.2012.741316

Wiesslitz, Carmit, & Ashuri, Tamar. (2011). 'Moral journalists': The emergence of new intermediaries of news in an age of digital media. *Journalism*, *12*(8), 1035–1051. https://doi.org/10.1177/1464884910388236