SC|M
Studies in Communication and Media

FULL PAPER

Linguistic corpora and tools in media and journalism studies

Sprachwissenschaftliche Korpora und Werkzeuge in der Medien- und Journalismusforschung

Valentin Werner & Hendrik Michael

Valentin Werner (PD Dr.), Otto-Friedrich-Universität Bamberg, Institut für Anglistik und Amerikanistik, An der Universität 9, 96045 Bamberg, Germany. Contact: valentin.werner(at)uni-bamberg.de. ORCID: https://orcid.org/0000-0003-2669-3557
Hendrik Michael (Dr.), Julius-Maximilians-Universität Würzburg, Institut Mensch-Computer-Medien im Arbeitsbereich Medien- und Wirtschaftskommunikation, Oswald-Külpe-Weg 82, 97074 Würzburg, Germany. Contact: hendrik.michael@uni-wuerzburg.de. ORCID: https://orcid.org/0000-0002-6818-0533

FULL PAPER

Linguistic corpora and tools in media and journalism studies

Sprachwissenschaftliche Korpora und Werkzeuge in der Medien- und Journalismusforschung

Valentin Werner & Hendrik Michael

Abstract: In the context of methodological discussions surrounding the "interdisciplinary turn" and "computational turn" of communication studies and media and journalism studies in particular, this paper raises awareness for corpus linguistics as a specific form of computer-assisted textual analysis with the potential to contribute to these developments. To date, corpus-based approaches have been embraced in media linguistics but have been underestimated in journalism and media studies. This paper argues that such neglect, especially when compared to other computational approaches towards textual analysis, is unwarranted. It examines relevant work from the areas of media linguistics and media and journalism studies and provides practical examples of how linguistic corpora and tools can inform pertinent research. Eventually, corpus-linguistic approaches are posited as an efficient technique for content analysis. They represent a potentially valuable addition to the (digital) methodological toolbox of media and journalism scholars as a useful "middle-ground" empirical approach that does not require extensive computational skills but allows conducting theory-driven work while maintaining control over the data and research procedures. Additionally, they facilitate the integration of quantitative and qualitative perspectives and open avenues for meaningful triangulation with other methods.

Keywords: Textual analysis, content analysis, interdisciplinarity, triangulation, linguistics.

Zusammenfassung: Im Kontext methodologischer Diskussionen um den „interdisciplinary turn" und den „computational turn" der Kommunikationswissenschaft und insbesondere der Medien- und Journalismusforschung möchte dieser Beitrag auf die Korpuslinguistik als spezifische Form der computergestützten Textanalyse aufmerksam machen, die das Potenzial hat, zu diesen Entwicklungen beizutragen. Bislang wurden korpusbasierte Ansätze, denen vor allem in der Medienlinguistik große Wertschätzung zuteilwurde, in der Medien- und Journalismusforschung weitgehend ausgeklammert. In diesem Beitrag wird argumentiert, dass diese Vernachlässigung, insbesondere im Vergleich zu anderen computergestützten Ansätzen der textuellen Inhaltsanalyse, ungerechtfertigt ist. Entsprechend werden relevante Arbeiten aus den Bereichen der Medienlinguistik sowie der Medien- und Journalismusforschung untersucht und praktische Beispiele dafür geliefert, wie linguistische Korpora und Programme die einschlägige Forschung bereichern können. Schließlich werden korpuslinguistische Ansätze als eine effiziente Technik für die Inhaltsanalyse präsentiert, die eine potenziell wertvolle Ergänzung des (digitalen) methodischen Werkzeugkastens von Medien- und Kommunikationswissenschaftlern darstellen. Dies liegt in ihrer Natur als empirische Herangehensweise, die zwar keine umfassenden Computerkenntnisse erfordert, aber die Durchführung theoriegeleiteter Arbeiten bei gleichzeitiger Kontrolle über die

Daten und Forschungsverfahren ermöglicht. Ferner wird argumentiert, dass korpuslinguistische Verfahren die Integration quantitativer und qualitativer Perspektiven erleichtern und Anknüpfungspunkte für eine zielführende Triangulation mit anderen Methoden bieten.

Schlagwörter: Textanalyse, Inhaltsanalyse, Interdisziplinarität, Triangulation, Linguistik.

1. Introduction

Within the broader field of communication studies and the more specialized area of media and journalism studies, some have acclaimed an interdisciplinary turn (Hase et al., 2022, p. 60) or new interdisciplinarity (Loosen et al., 2022, p. 5) that unlocks as-yet-uncharted territory in terms of both theoretical and methodological approaches used and research questions to be tackled. Fostered by (i) technological advancements, (ii) the availability of tailor-made software tools, and (iii) a dramatic rise in the volume of machine-readable (textual and other) data (van Atteveldt & Peng, 2018, p. 81), this growing interdisciplinarity has been associated with an greater openness towards computational approaches to (re-)asses established concepts and theories (see also Hase et al., 2022, p. 72), hence the regular equation of the interdisciplinary turn with a "computational turn."

What is striking in the view of the fundamental epistemological question of "to what extent are the available methods appropriate to acquire knowledge about new components of journalism and their relationships?" (Sjøvaag & Karlsson, 2016, p. 88) is that the computational approach of corpus linguistics has received little attention to date when it comes to analyzing media texts. Corpus linguistics has been defined as an empirical method for the study of language by way of computer-assisted analysis of machine-readable collections of writing or speech (Meyer, 2023, p. 4), taking heed of the fact that texts are important artifacts that produce and semiotically circulate meaning (see, e.g. Bertrand & Hughes, 2017, p. 224).[1]

Neither are corpus-linguistic methods discussed in the aforementioned recent reviews on interdisciplinarity and methodological innovation nor in methodological textbooks and handbooks (e.g. Bertrand & Hughes, 2017; Riffe et al., 2020; Scharrer & Ramasubramanian, 2021; Oehmer-Pedrazzi et al., 2023), nor do they feature prominently in other articles and chapters with methodical concerns (e.g. Baden et al., 2022; Boumans & Trilling, 2018; Loosen & Schmidt, 2016). Corpus-linguistic tools are sometimes even sweepingly dismissed as "relatively static and inflexible" (Niekler et al., 2023, p. 327).

This also aligns with Lefkowitz (2021, p. 97), who asserts that "the potential of corpus linguistics with regards to journalistic texts has remained largely untapped." Accordingly, one of the overarching goals of this contribution is to bridge a current methodological gap by addressing the persistent disconnect between the larger areas of media linguistics, defined as a "branch of research in linguistic science, focusing on language use in the press, on TV and radio, but also in advertising and

1 On the wider debate whether "text only" analysis is a suitable approach in journalism studies, which cannot be explicated here (but see Section 2.2) due to space constraints, see, e.g., Philo (2007) vs. Fürsich (2009).

online publications" (Lüger, 2017, p. 10), and media and journalism studies (see Bucher & Niemann, 2018, p. 178). Specifically, it aims to demonstrate how corpus linguistics represents a highly versatile approach when it comes to textual analysis and can contribute to the diversification of the methodological toolkit of media and journalism scholars engaging in content analysis.

This is done in the spirit of an emerging (but not yet fully realized) "new interdisciplinarity" that includes methodological, but eventually also theoretical (e.g. reliance on theories outside of one's own research area) and practical (e.g. publications authored by interdisciplinary research teams) facets (Hase et al., 2022, p. 61).[2] To illustrate broader trends and opportunities, the specific focus will lie on the study of journalistic texts as a widespread instantiation of public mediated communication.

To this end, Section 2 provides conceptual background by (i) briefly charting the current disconnect between media and journalism studies and media linguistics and (ii) situating corpus-linguistic approaches as a form of content analysis. Section 3 discusses the opportunities and limitations of established computational approaches to textual analysis, and introduces corpus linguistics as a complement. Section 4 provides a selective review of relevant studies to illustrate the scope of extant work. Section 5 takes a more practical approach and complements Bednarek and Carr (2021), who have established the concept of "computer-assisted digital text analysis." It expands upon their valuable guide to analyzing self-compiled corpora with a commercial tool with regard to using (i) sizable freely available reference corpora and (ii) free tools. Section 6 summarizes and contextualizes the arguments presented.

2. Conceptual background

2.1 Media and journalism studies, media linguistics, and obstacles to interdisciplinarity

When work transcends disciplinary boundaries, one of the common impediments are natural differences in research objectives and terminology use across the respective fields. For instance, technical terms such as "corpus" or "keyword" vary in meaning and usage (see also Section 4.1). Others include forces of inertia impeding methodological innovation, and, importantly, a lack of awareness of scholars of other research within and especially outside of the confines of their discipline (Boumans & Trilling, 2018, pp. 16–17).

An illustration of the latter is the following claim: A consequence of these developments (i.e. the "new interdisciplinarity" mentioned above) is that journalism studies have increasingly opened up towards other disciplines. Simultaneously,

2 Similar developments are observable in the social sciences at large, where to date the "linguistic turn", intended to "make […] room for new types of discourse and conversation analysis" (European Strategy Forum on Research Infrastructures, 2021, p. 108; with media studies notably mentioned as a an "articulate example" of interdisciplinarity in the same report), largely is a "computational turn" (see, e.g., Bucher & Niemann, 2018, p. 180; Windsor, 2021, p. 182), arguably not exploiting the full scope of linguistic approaches.

journalism is "discovered" by other disciplines as a research object (Loosen et al., 2022, p. 5). This supposedly recent "discovery" may seem somewhat grotesque from the perspective of linguists[3] as they have been interested in journalistic language for several decades now (see Bucher, 2014, p. 273), for instance to track linguistic change at large (see, e.g., Bell, 1995; Hundt & Mair, 1999). They further have recognized journalistic language as distinct communicative form (or "register") and have shown a "considerable readiness to adopt insights and impulses from other lines of research […], not least from the science of journalism" (Lüger, 2017, p. 15).

Corpus linguists in particular have acknowledged the significance, prevalence and accessibility of news discourse, even deeming it as "an ideal territory for corpus linguistics" (Marchi, 2022, p. 576). Thus, at best, claims such as the one above highlight the large-scale ignorance of corpus-linguistic techniques on part of media and journalism scholars (see also Lenk, 2013, p. 67).[4] Essentially, this results in a largely one-sided relationship wherein corpus linguists diligently study media texts but media and journalism scholars have overlooked pertinent work and have avoided implementing corpus-linguistic techniques. At the same time, while engagement with journalistic texts has a long tradition in media linguistics, it has to be acknowledged that – *pace* Lüger (2017) – there actually is an undervaluation of concepts and frameworks established in journalism studies (see Section 4 for details). This lack of mutual recognition and the ensuing state that "a fruitful, interdisciplinary contact cannot have won recognition yet" (Lüger, 2017, pp. 25–26) is regrettable in the view of the fact that "linguistic analysis of news discourse can yield valuable corroboration as well as insights for a broader understanding of how journalism operates" (Marchi, 2022, p. 582), and particularly as "both communication research and linguistics have a long tradition of content-analytic techniques" (Mahrt & Scharkow, 2013, p. 27).

2.2 Corpus linguistics as content analysis

To flesh out this common concern and to contextualize corpus linguistics within the extant methodological literature in media and journalism studies, it is beneficial to place it within established taxonomies. These distinguish the three basic methodological strands of interviewing, content analysis, and observation (including

3 Suprisingly, Hase et al. (2022, p. 74) mention information science, engineering and technical science as wider disciplines engaging with journalistic communication, leaving aside linguistics as an apparently (too?) obvious choice.

4 It needs to be highlighted that communication and journalism scholars are by no means the only ones "guilty" of this type of ignorance, as a recent paper discussing the study of language from the perspective of psychology reveals (Jackson et al., 2022). In this contribution, the authors seem to embrace computational approaches for language processing (including techniques such as topic modeling, sentiment analysis, and text classification) but do not discuss the opportunities and limits of corpus linguistics for psychological studies. Further, they misrepresent factual information with claims like "[t]he Oxford English Corpus is the largest corpus of 21st century English, totaling more than 2.1 billion words across multiple English-language cultures" (Jackson et al., 2022, p. 810), while there certainly are larger databases available (e.g. the constantly evolving NOW Corpus with c. 19.3 billion words in July 2024; english-corpora.org/now). They also seem to disregard the longer history of linguistics as a discipline when stating that "[r]esearchers now [sic!] have the tools to analyze and interpret this language" (Jackson et al., 2022, p. 820).

experimental approaches, typically featuring in the study of media reception and media effects), each including "manifold qualitative/non-standardized and quantitative/standardized forms and variants" (Loosen & Schmidt, 2016, p. 563). Corpus linguistics could conventionally be categorized as part of content analysis, viewed as "systematic and replicable examination of symbols of communication [...] to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption" (Riffe et al., 2020, p. 23).

Supporting rationales for deeper engagement with the language of journalistic media texts (aka "news discourse")[5] as symbols of communication stem primarily from their high social impact in constructing and representing "what matters, what makes sense, what time and place we live in, what range of considerations we should take seriously" (Schudson, 1995, p. 14; see also Bell, 1995, p. 23; Mautner, 2008, p. 32). This motivates a functional perspective on (among others) linguistic usage patterns in relevant instantiations to determine their effect on the context, specifically the shaping of public opinion or the representation of social values at a given point in time (Bucher, 2014, p. 271, 275; Meier et al., 2017, p. 357). Concretely, this functional view of language has also been considered in various frameworks prominently featuring in media and journalism studies (see, e.g., the "style dimension" within the concept of "softening the news" according to Reinemann et al., 2012).

Given the aforementioned perspectives, it is not surprising that common definitions of content analysis invite collaboration with theories of language and discourse analysis, since they make the symbolic meaning of media contributions their genuine subject matter. They provide the categories and criteria that are relevant for a content-analytical grasp of media communication as a form of symbolic exchange (Bucher, 2014, p. 288). Specifically, on a methodological level, it has been noted that *corpus*-linguistic approaches in particular may be useful to obtain empirical insights into quantitative and qualitative textual patterns, even if only viewed as a means to triangulate more traditional content-analytic perspectives (Bucher, 2014, p. 289) as well as perspectives from the other methodological strands on the production and reception of relevant communication (Fraas & Pentzold, 2016, p. 229).[6] However, as previously mentioned, corpus-linguistic techniques have rarely been employed in media and journalism research to examine pertinent texts. Instead, content analysis predominantly has employed either more or less extensive manual coding (which can indeed be combined with corpus-linguistic approaches,

5 On the discussion of the use of the terms "discourse" vs. "text" and their variable usage in different research traditions, see, e.g., Bucher (2014) or Fraas and Pentzold (2016). In the present work, we will refer to (news) discourse as a textual manifestation and thus use the terms liberally.

6 Thus, Bucher (2014, p. 272) considers looking at linguistic (and other) symbols of communication even as some kind of cross-sectional "basic science" for media and journalism studies, as it may also become relevant in the study of media effects as well as of the communication between and among producers and recipients.

see Sections 4.2 and 5.2) or automated content analysis aligned with the "computational turn" (see also Section 3).[7]

3. Opportunities and limits of automated content approaches

As suggested above, the "interdisciplinary turn" in media and journalism studies predominantly is a "computational turn" in the sense of relying on computational and information science methods for analyzing very large datasets (aka "big data"). A significant amount of pertinent data is textual (see, e.g., Chen et al., 2023, p. 112) and typically amenable to quantitative techniques used in natural language processing (NLP), including topic modeling, named entity recognition, and sentiment analysis (Sjøvaag & Karlsson, 2016, p. 91). As rightly noted by Boumans and Trilling (2018, pp. 8–9), such automated content analysis, generally defined as a set of approaches "in which the analysis of text is, to some extent, automatically conducted by machines" (Hase, 2023, p. 23), has been embraced by media and journalism scholars as it offers several benefits, such as being able to test hypotheses based on previous qualitative or small-scale quantitative research or to efficiently identify latent patterns in the data that would remain hidden to mere manual assessment.

As the focus of the present contribution is on corpus linguistics rather than NLP, two general issues need to be resolved at this point: (i) corpus linguistics is not to be viewed in complementary opposition to automated content analysis based on NLP, as there are shared interests and a considerable overlap of functionalities in individual implementations. Arguably, corpus linguistics could even be viewed as an "offspring" (in the sense of an implementation) or at least "close relative" of NLP and there have been persistent efforts to "scale up" corpus linguistics through integrating NLP techniques in a meaningful way (see Dunn, 2022, for recent discussion); (ii) we do not intend to present automated content-analytic techniques in an unduly negative light in the following, as they have proven to be a versatile means for putting media and journalism studies on firmer empirical ground in many respects, as already indicated above, and are by no means "blind" to linguistic insights. Yet, such approaches also have various potential limitations on both the theoretical and practical level, as discussed in the methodological literature, and we would like to sketch the potential of how some of these constraints can be addressed using corpus linguistics as a complement (at the same time acknowledging that there are continuous efforts on the part of both the NLP and the media and journalism studies communities to improve algorithms and tools).

A first notable issue at stake is transparency. It is widely recognized that the more sophisticated and algorithmically complex an approach (e.g. towards sentiment analysis) is, the more there is a danger of the tools and programs used rep-

7 (Critical) discourse analysis (CDA), which can be corpus-assisted or not, is occasionally mentioned as a qualitative (and non-standardized) form that contrasts with the idealized systematic quantitative perspective regularly taken in content-analytic work (see Pentzold, 2017, p. 517). For reviews of relevant contributions beyond the ones introduced in Section 4.1, see Bucher (2014, p. 279) and Kelsey (2017).

resenting a "black box" for the regular media and journalism scholar lacking intensive computational training (Pichler & Reiter, 2020, p. 51; Bednarek & Carr, 2021, p. 147), so that a systematic evaluation of relevant tools becomes difficult. This, in turn, may prompt researchers to continue relying on more transparent (non-computational) approaches involving higher levels of manual control (Baden et al., 2022, p. 2). On a more general level, missing transparency may also obstruct replicability in research.

A second recurring theme is the lack of theoretical embedding of big data approaches, so that research is driven by "the bigness of data itself or its processability [rather than] by theoretical questions and substantive problems" (Shahin, 2016, p. 47), a development possibly fostered by the wider availability of relevant commercial software tools, which are (over-)used as ends in themselves rather than as instruments for generating and empirically testing hypotheses grounded in established frameworks (Chen et al., 2023, p. 111). On a related note, it has been observed that reliance on automated analysis carries the risk of taking essential terms such as "frame," "sentiment," or "topic" simply for granted, and that leaving the theoretical status of these constructs implicit can lead to simplistic or flawed operationalizations (Baden et al., 2022, pp. 8–9; Hase et al., 2022, p. 73). This decline of theory-driven research is also linked to broader debates about (i) whether the focus on automated approaches has led to a dominance of purely descriptive work, with a consequent neglect of formal theoretical developments (see, e.g., Waldherr et al., 2021) and (ii) whether the "computational turn" has in fact turned into a "computational overwind," as there is tension between "developers' emphasis on technological and statistical properties" and a "primary concern for operational demands and measurement validity" on the part of researchers (Baden et al., 2022, p. 2).

The latter directly relates to a third limitation, which concerns the reliability and validity of automated analysis. While Sjøvaag and Karlsson (2016, p. 91) had already warned that "computational power can easily substitute scale for depth," Scott (2023) manually tested whether the results of content word searches in widely used databases such as *Nexis* or *Factiva*, which could be considered typical sources for NLP analysis in media and journalism studies, actually retrieved relevant articles. This exercise showed that even the "crude assumption that word forms indicate text meaning" (aka the "bag-of-words" assumption that "the orders of words in a document can be neglected to understand the major content in a document"; Chen et al., 2023, p. 115) is deficient and retrieves a large proportion of irrelevant or only marginally relevant texts, thus introducing considerable "noise" into the data. However, texts collected in this way regularly serve as input for automated analysis of large datasets without the opportunity for additional manual assessment, which can be considered undesirable.

Another aspect operates on a personal and practical level. Despite the increased availability of commercial and freely available software packages mentioned above, media and journalism scholars may face a steep learning curve if they want to competently use big data approaches as they do not have sufficient computer literacy to perform automated analyses. While it is clear that media and journalism scholars have increasingly developed computational skills and have started developing computational methods themselves (i.e. "by the field for the field"), they still may have to rely on interdisciplinary cooperation (Sjøvaag & Karlsson, 2016, p. 91), and it has been argued that progress in this direction is still comparatively slow (see, e.g., Bednarek & Carr, 2021, p. 146). Further, such collaborative efforts can often involve financial burdens (e.g. hiring a programmer or data scientist) that are beyond the resources of both junior researchers and researchers in the Global South (van Atteveldt & Peng, 2018, p. 87). Furthermore, the latter are disadvantaged due to a persistent lack of technical infrastructure and computational power (Madrid-Morales, 2020, p. 74), which may be required to efficiently run sophisticated NLP tools.

The aforementioned limitations point to the larger issue of methodological reflection. A "bigger (data) is better" approach may not always be helpful, as "researchers caught in the 'data rush' seem to have thrown caution to the wind, allowing themselves to be seduced by the appeal of Big Data" (Mahrt & Scharkow, 2013, p. 22) without acknowledging the inherent weaknesses of large-scale automated analysis of (textual) media content. As a potential remedy, researchers in media and journalism studies would be well advised to consider drawing on methods and tools from linguistics, the discipline with arguably the longest tradition in textual research, at least to supplement established analytical techniques (Mahrt & Scharkow, 2013, p. 22).

Therefore, the following section will outline how corpus linguistics in particular can be a viable "middle-ground" approach that provides opportunities for theory-based, quantitative-qualitative analysis while still offering control over the data through tools and procedures that are comparatively transparent and accessible to the regular media and journalism studies scholar and that can be fruitfully combined with other types of (content) analysis (see also Lewis et al., 2013; Madrid-Morales, 2020, p. 79).[8] In other words, corpus linguistics is presented as a digital method for (semi-)automated empirical analysis that still involves human judgement and theoretical relevance and reflection to a greater extent than fully automated computational (and arguably opaque) approaches, so that eventually "scale meets depth" (Shahin, 2016) in a best-case scenario. In addition, from a practical perspective, it is suggested that corpus linguistics as a "complementary way of doing Digital Social Science" (Bednarek & Carr, 2021, p. 132), can still be seen as part of the computational turn outlined in Section 1, albeit without high technical and organizational barriers and extensive time investment on the part of researchers (Bednarek & Carr, 2021, p. 146) but with the possibility of fruitfully combining computational and media studies approaches, as already postulated by Sjøvaag and Karlsson (2016, p. 91), for instance.

8 See also Bednarek and Carr (2021, p. 146) for a general overview of advantages and limitations of corpus-linguistic approaches. On a related note, it must be acknowledged that "methodological transparency" in media and journalism studies commonly is conceptualized as providing actual code (e.g. in R or Python) usable for the replication of automated content analysis procedures. The understanding of transparency here relates to the structure and functionalities of the (corpus-linguistic) tools presented (see Section 5).

4. Corpus-linguistic work on journalistic language

This section presents corpus linguistics as a path towards reflected algorithmic textual analysis, defined as a (best) practice of computer-aided text analysis characterized by its interdisciplinary modularization. These modules are interlinked manual and automatic work steps related to concepts or text phenomena. The allocation of the modules and the interpretation of their results are carried out taking into account domain-related prior knowledge, the operationalizability of the modules and their empirical validation (Pichler & Reiter, 2020, pp. 57–58).

This definition suggests that such an approach is genuinely interdisciplinary and theory-driven/deductive. In particular, corpus-linguistic software has explicitly been mentioned as useful for such hypothesis-testing projects (Pichler & Reiter, 2020, p. 56), whereas other automated content analysis tools (see Section 3) have more frequently been associated with hypothesis-generating (aka explorative/inductive) work (see, e.g., Hase et al., 2022, p. 74). While this is not assessed formally (i.e. through a user experience survey), it is further suggested that corpus-linguistic tools are comparatively transparent (i.e., users can comprehend what the software does as it relies on algorithmically less complex procedures than many other computational tools, which require a certain amount of programming/coding skills) and accessible (e.g. they come with a graphical user interface). On a general note, it is also worth mentioning that corpus linguistics as a field has engaged in extensive meta-methodological reflection within the "replication crisis," including discussion on the issues of reliability, replicability, and open science practices (see, e.g., Sönning & Werner, 2021; Schweinberger & Haugh, forthcoming).

Another strength of a corpus-linguistic approach is that it typically combines qualitative and quantitative research in a systematic and reliable way, balancing automated and human evaluation (Touri & Koteyko, 2015, p. 605; see also Haider & Hussein, 2020, p. 840). Again, we do not claim that such hybrid approaches are absent from other types of computational content analysis altogether (see, e.g., Nelson, 2020 or Baden et al., 2022) but as yet they are the exception rather than the rule and typically work on much vaster sets of data, naturally coming at the cost of a lower amount of researcher "immersion" in the data, which, however, may be a crucial prerequisite for adequate theory-driven interpretation (Carlsen & Ralund, 2022). As stated above, corpus-linguistic approaches often are mixed-method by default, thus are more data-immersive and arguably allow the identification of subtle linguistic differences in the textual material, which is likely to increase the validity and reliability of the findings.

For a fruitful theory-driven application of corpus linguistics in the field of media and journalism (content) studies specifically it is essential to consider (i) how relevant concepts (such as "framing," "emotionality," "news values," or "objectivity") can be operationalized (i.e. how they can be linked to actual linguistic phenomena at the surface text; see Pichler & Reiter, 2020, p. 47), (ii) how quantitative and qualitative perspectives can be efficiently combined to offer a "unique way of looking at language in the media" (Jaworska, 2018, p. 105; see also Windsor, 2021, p. 183), and (iii) how corpus-linguistic approaches can potentially complement other (manual and computational) content-analytic techniques to arrive at valid results and theoretically meaningful interpretations.

While Bednarek and Carr (2021) summarize several pioneering studies (such as Touri & Koteyko, 2015) and provide a succinct overview of how a specific tool, *WordSmith* (Scott, 2020), can be used to gain insights into aspects such as media bias – in the form of (dis-)preferred language for certain social groups –, sourcing and quoting practices of individual media, as well as framing, it may be useful to review work from the areas of media linguistics and media and journalism studies in order to outline the full potential of corpus-based approaches to journalistic language (see also Lefkowitz, 2021, pp. 98–102). While a comprehensive review would go beyond the scope of the present contribution (and would merit an additional systematic review article in itself), the scope and breadth of extant research are illustrated below to approximate a representative overview. The studies presented were chosen as they either represent a more corpus-linguistic (Section 4.1) or media and journalism studies (Section 4.2) perspective on the data, as determined by the association of the publication outlet or the individual author(s) with one discipline or the other.

To promote mutual awareness, some "missed opportunities" of (i) corpus-linguistic work in terms of embedding it in common theories and frameworks used in media and journalism studies and (ii) media and journalism studies work in terms of applying certain corpus techniques are identified, and, in addition to the practical examples given in Section 5, potential points of (future) interdisciplinary contact are highlighted.

4.1 Linguistic studies

Following Jaworska (2018, pp. 101–102; see also Marchi, 2022, pp. 581–585), corpus-linguistic work on news discourse is based either on "news" parts of larger general (aka "reference") corpora (such as the *British National Corpus*, BNC) or on specialized, often self-compiled news corpora, and can broadly be categorized into the two strands of "language use" (form) and "discursive representations" (content), as reflected in Table 1.[9] In addition, it has been observed that there is a certain bias towards the latter strand (and especially its "critical" variant; see also note 7), as well as a strong focus on written discourse (O'Keeffe, 2012, p. 117), and especially legacy media discourse (Nartey & Mwinlaaru, 2019, p. 217), which does not seem too surprising given the easy (now regularly digital) accessibility of online and digitized written news material.

Analyses of the "content" type take media discourse on socially relevant topics (e.g. politics, racism, nationalism, science, gender, disease, migration) as their starting point and adopt a "critical" perspective as they are "often interested in uncovering the mechanisms of ideological work as legitimized through language use" (Jaworska, 2018, p. 103; see also Mautner, 2008, p. 33). To counter views of

9 Corpus-based work on longer-term diachronic (sociolinguistic and pragmatic) developments (see, Conboy, 2014; Paulsen, 2022 or the contributions to Bös & Kornexl, 2015, for instance) is ignored here for the sake of brevity.

critical discourse analysis (CDA) as overly subjective due to its traditional reliance on manual analysis of smaller sets of texts, it has been argued that CDA has benefitted from corpus-based approaches that produce empirical quantitative evidence (i.e. salient terms/keywords in news discourse, see above) and thus lead to generalizable findings (see, e.g., O'Halloran, 2010, pp. 564–565). However, CDA still relies to a considerable extent on follow-up manual analysis (e.g. of individual concordance lines) in what is commonly referred to as corpus-assisted discourse study (see, e.g., Johnson & Partington, 2018). While CDA is a burgeoning subfield (see, e.g., Jeffries & Walker, 2019 or Balfour, 2023 for recent book-length treatments), it is considered sufficient to present selected studies in more detail for illustrative purposes (see Table 1).

4.2 Media and journalism studies

Despite the widespread neglect or at least underestimation of corpus linguistics in the analysis of journalistic texts (see Section 1), there are a few pioneering studies that have appeared with the comparatively short timeframe of a decade. To highlight the potential of corpus-linguistic approaches, a selective overview is presented in Table 2, again with the aim to illustrate the scope and breadth of extant research rather than drawing a comprehensive picture.[10]

The overview in Table 2 illustrates that corpus-linguistic approaches certainly are not unheard of in media and journalism studies and have repeatedly been used to complement other types of content analysis. At the same time, it is clear that individual studies suffer from issues addressed in corpus linguistics (e.g. balancing of corpora that are compared, small corpus sizes) and may have benefitted from applying more sophisticated corpus-linguistic techniques.

4.3 Summary evaluation

The preceding overview, rather than representing a comprehensive summary of the full body of corpus-based work, was guided by the aim to illustrate the scope and breadth of such analyses and their potential relevance for content-analytic media and journalism studies. For instance, this concerns the opportunities for empirical validation of (theoretical) concepts such as tabloidization or also the (data-driven) exploration of aspects such as topic modeling or agenda setting under the prerequisite that corpus queries intended to provide insights into central concepts of media and journalism studies are operationalized in a meaningful way on the basis of linguistic categories. Concurrently, researchers can retain control over the data and can supplement quantitative and qualitative insights from the corpora with manual annotation (as done in Lefkowitz, 2021, for instance). Given this vast potential for "reflected algorithmic textual analysis" as defined above, the persistent

10 Work analyzing journalistic language with corpus-linguistic approaches in other fields is ignored here, but see, e.g., Meier et al. (2017) for an example from sports sciences that relies on Entman's cascading activation model, Atanasova et al. (2019) from health studies that analyzes framing, or Haider and Hussein (2020) for a study from digital humanities focusing on news values using a parallel Arabic-English corpus.

Table 1. Linguistic studies on the language of journalism

Study	Type	Data	Research focus	Method/operationalization	Relevant journalism studies concepts/theories	Comment on "missed opportunities"
Biber (2003)	Form	*Longman Spoken and Written English Corpus* (c. 20 million words, c. 5.4 million words of British news)	Increasing popularization/colloquialization vs. information-centric/economical nature of newspaper discourse (means of comparison: data from Hundt & Mair, 1999)	Quantitative comparison of changing frequencies (markers of informality vs. heavily pre- and/or post-modified nouns and nominalizations)	Style dimension (personal/emotional reporting) (e.g. Reinemann et al., 2012)	More insights into production dimension through triangulation, e.g. with interviews with journalists/observational work regarding their (changing?) working practices in the twenty-first century (Loosen & Schmidt, 2016) would have allowed insights into reasons for the findings
Duguid (2010)	Form	*SiBol 93* (c. 100 million words from the year 1993); *SiBol 05* (c. 145 million words from the year 2005)[1]	Increasing informalization/tabloidization of newspaper discourse + increased amount of evaluation	Quantitative comparison of keywords between the two corpora[2] with a focus on evaluative adverbs and adjectives	Style dimension (personal/emotional reporting) (e.g. Reinemann et al., 2012); emotionality/intimization of journalism (e.g. Steensen, 2016; Wahl-Jorgensen, 2019)	Hypotheses on reasons of changing newsroom practices not substantiated by additional content analysis (e.g. manual annotation of sourcing practices, a quantification of article types or the number of direct quotes; on the latter cf. Clark, 2013) or practitioner observation and interviews

Scott & Tribble (2006)	Form/ content	*Guardian Weekly* corpus (1995–2001; 14 million words)	Gender-sensitive reporting, major themes and topics per year, typical collocations and their semantic prosody[3]	Quantitative keyword analysis, collocation analysis	Topic modeling, agenda setting, framing	Provides evidence for journalists subtly employing "moral evaluation" (Entman, 1993, p. 52) in the sense of framing news stories as an instance of second-level agenda setting (Wu & Coleman, 2009, p. 775), without labeling it as such.
Jaworska & Krishnamurthy (2012)	Content	Self-compiled corpus of articles with the target term *feminism* in the headline and/or in the lead paragraph (daily and weekly newspapers and magazines, c. 3 million words), created from *Nexis*	Contrastive analysis of expression of (negative and sexualized) attitudes towards feminism in the British and German press	Quantitative-qualitative collocation analysis of the patterns premodifying adjective + *feminism* and *feminism is X*	Agenda setting, framing	Relation to relevant journalism concepts left implicit
Collins & Nerlich (2016)	Content	Self-compiled corpus of British press articles with target term *climate change* (c. 750,000 words), derived from *Nexis*	Perception of the notion of climate change following the climate summits in Copenhagen (2009) and Paris (2015)	Quantitative semantic keyword analysis[4] + manual analysis of the key term *uncertainty*	Topic modeling, priming	No explicit relation to relevant journalism studies concepts/theories

Notes. [1]An updated version of the SiBol corpus, now comprising data up to the year 2021 (c. 850 million words) is available at sketchengine.eu/sibol-corpus. [2]Keywords in a corpus-linguistics sense are identified through comparing the frequency of words occurring in a target corpus (e.g. a newspaper corpus) with their frequencies in a reference corpus (e.g. a general corpus of English such as the *British National Corpus*). If a word occurs more frequently in the target corpus it may be called a keyword with a specific relevance and function in the target corpus, and may be worth closer scrutiny. Corpus software usually offers the opportunity to calculate various "keyness" measurements (see, e.g., Sönning, 2024). The higher a keyness value, the more the item is "overrepresented" in the target corpus. [3]Semantic prosody is defined as "the evaluative meaning of extended lexical units" (Ebeling, 2013; see also Louw, 1993), which is especially prominent in near-synonyms (e.g. the (neutral) *Clinton administration* and the *Blair government* vs. the (negatively connoted) *Milosevic regime*).
[4]Semantic keyword analysis contrasts a target corpus with a general reference corpus with the aim of identifying not only salient individual words, but also broader themes (semantic fields) that appear in the relevant discourse (Rayson, 2008).

Table 2. Media and journalism studies analyses using corpus linguistics

Study	Data	Corpus tool(s) used	Research focus	Method/operationalization	Relevant journalism studies concepts/theories	Comments
Krenmayr (2015)	Journalistic texts (from *Independent*, *Guardian*, *Daily Telegraph*) and other genres from the *British National Corpus* (BNC), c. 200,000 words	Information not provided, assumedly *XAIRA* (natcorp.ox.ac.uk/tools/)	Metaphors as rhetorical devices in British quality newspapers	Manual word-by-word metaphor identification by four raters Followed by quantification	Representation of ideologies in discourse	Limited amount of news data (c. 45,000 words only)
Moon (2016)	Self-compiled corpus (from *Factiva*) of materials from Kenyan paper *Daily Nation* and the British *Times*, c. 20 million words	*WordStat*	Cross-cultural analysis of news styles	Identification of content words/topics	Institutional theory, news values, storytelling	Use of more sophisticated corpus-linguistic techniques (such as keyness analysis) possibly with more straightforward insights, imbalance between the subcorpora contrasted (*Daily Nation* part of c. 750,000 words vs. *The Times* part of c. 19.3 million words)
Hansen (2016)	Self-compiled corpus of four Danish newspapers (1997–2013) from *Infomedia*, c. 1.5 million words	*WordSmith Tools*	Future-oriented news discourse, assertiveness vs. speculativeness in news discourse (short-term diachronic perspective)	Quantitative keyness analysis Qualitative concordance analysis of modals	Gatewatching, sourcing practices, accountability, contextual journalism	Complements traditional (manual) content analysis

Lefkowitz (2018, 2021)	Self-compiled corpus from several databases (*Nexis* and newspaper archives, e.g. from the *Times*, 1968–2016)	*AntConc*	Tabloidization vs. public service function of British newspapers (long-term diachronic perspective)	Automatically computed variables on information orientation (average article length, average sentence length, average word length, Flesch-Kincaid readability) Frequency counts (concordances) of (i) markers of personalization (first/second person pronouns, private verbs, i.e. verbs describing mental states or acts), (ii) markers of sensationalism (superlative general adverbs, degree adverbs, superlative degree adverbs, amplifiers, general adverbs, and general adjectives) Manual count of quotes (as a means to increase personalization) Manual content analysis of (i) story topics (politics, health, crime, etc.); (ii) types of story topics (hard vs. soft. vs. general)	Personalization, sensationalism, theory of the public sphere, news values	Combines NLP measures with corpus-linguistic insights and manual content analysis
Ping (2022)	Several smaller (c. 20,000–90,000 words) self-compiled corpora (from *BBC Chinese*, *New York Times Chinese*, *RN*); *English Web Corpus 2015*, *Chinese Web Corpus 2017* as reference corpora	*SketchEngine*	Themes and political stance in the news discourse surrounding the 2014 “Occupy Central” movement in Hong Kong	Keyness analysis to identify salient themes Concordance searches and manual annotation of framing types	Framing, media bias	Triangulation of quantitative-qualitative corpus linguistics with manual qualitative annotation
Li & Zhang (2022)	Self-compiled corpus from the *New York Times* (2000–2016) created from *Nexis*, c. 1 million words; *Open American National Corpus* as reference corpus	*AntConc*	Coverage of Muslims and Islam	Keyness analysis to identify salient themes Collocation and n-gram analysis to determine contextual use	Framing, media bias	

large-scale ignorance of corpus-based approaches and insights (see Section 1) is deplorable.

At the same time, however, several problematic aspects emerged, which could partly explain the reluctance of media and journalism scholars to rely on corpus linguistics as a tool. First, media-linguistic studies would benefit from considering terminological discussion in media and journalism studies (see, e.g., Otto et al., 2017 on delineating partly related concepts such as "sensationalism" vs. "hard/soft news" vs. "infotainment" vs. "tabloidization"). This would facilitate exchange between the disciplines and would help to avoid the use of different terminology although the same or similar conceptualizations are explored (e.g. "informalization" vs. "tabloidization").

Second, media-linguistic corpus studies could arguably benefit enormously from triangulation with other methodologies (e.g. practitioner interviews and observation or manual content analysis to determine sourcing practices) in order to go beyond mere hypothesizing about the reasons for linguistic findings. On a related note, there may be an occasional disconnect with key theoretical frameworks employed in media and journalism studies (e.g. the style dimension of tabloidization) even though the actual linguistic findings are potentially relevant. This suggest that while corpus-linguistic work has gained limited momentum in media and journalism studies, an "interdisciplinary modularization" as defined in Section 4 is yet to be fully realized.

Another obvious limitation of the studies both from the angles of media linguistics and media and journalism studies is that they often rely on self-compiled corpora (see Santaemilia & Maruenda, 2014, Landert, 2015, or Marín Arrese, 2015 for other, including non-English, examples) that are not publicly available or that were compiled relying on commercial aggregator databases subject to substantial licensing fees (such as *Factiva* or *Nexis*). Clearly, this hinders replicability, may disadvantage scholars without adequate resources or may place some demands on researchers in terms of required programming skills to compile relevant materials (e.g. through extensive web-scraping) if such costs are to be avoided (Marchi, 2022, p. 579). While some studies leave it implicit which corpus tools were actually used (e.g. Krennmayr, 2015), others (e.g. Hansen, 2016; Moon, 2016; Ping, 2022) rely on commercial tools that come with licensing costs. On a different note, it appears that most studies have a synchronic or short-term diachronic scope and the vast majority focuses on English as the target language. The subsequent section aims to address some of the shortcomings from a practical perspective.

5. Applying corpus-linguistic methodology

As noted above, there have been some efforts to highlight the potential of corpus-linguistic approaches for media and journalism studies (notably, Bednarek & Carr, 2021), as well as to apply various databases and tools (see Section 4.2). The current section aims to encourage media and journalism scholars to seriously consider corpus-linguistic approaches by lowering the barriers for using them (either on their own or in collaboration with corpus linguists). This is attempted by (i) illustrating the use of large reference corpora that are freely available online for academic purposes and (ii) introducing *AntConc*, a free and extensively documented tool that can also be used for historical press research based on publicly available news archives, an area that has not received much attention to date.

By necessity, selective examples are given in order to familiarize scholars with relevant resources and tools and to briefly illustrate cases of theory-driven corpus-based research possible without expert scripting/programming skills. Obviously, due to space limitations, the full "workflow" for each tool and search and the customization of individual settings (e.g. for the statistical measures to calculate keyness; see note 2 of Table 1) cannot be reproduced here; rather, the aim is to illustrate the range of possible applications of corpus-linguistic approaches in media and journalism studies.

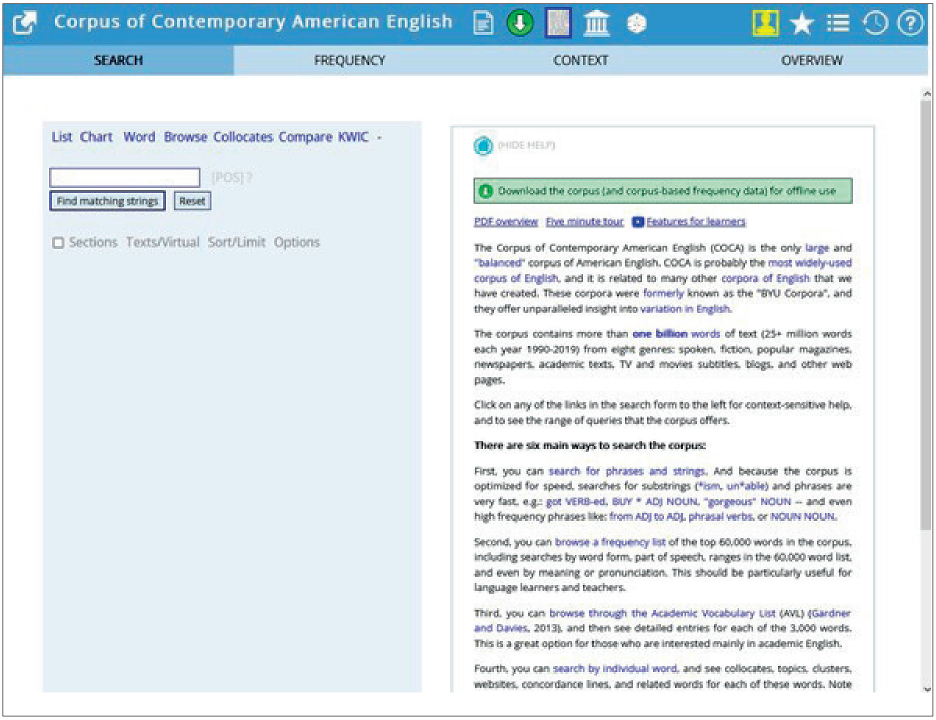
5.1 Reference corpora

Due to its wide availability and social relevance, linguists have for a long time relied on news discourse as study material for analyses of linguistic structure and change in general. Accordingly, carefully annotated (e.g. including part-of-speech tagging, genre categorization or speaker assignment) general reference corpora for various languages (such as COCA for English, english-corpora.org/coca; *Deutsches Referenzkorpus* for German, DeReKo, https://www.ids-mannheim.de/digspra/kl/projekte/korpora/; the *Corpus del Español* for Spanish, corpusdelespanol.org) that are freely available for academic purposes regularly contain substantial news portions (including data from magazines). In addition, there exist large linguistic news corpora that are also freely available (e.g. *News on the Web*; NOW; available for English and Spanish, for instance; Davies, 2017; see also the resources listed at clarin.eu/resource-families/newspaper-corpora) as well as specialized corpora, for instance on headlines (e.g. Kulkarni, 2021). Surprisingly, apart from a few exceptions (e.g. Krennmayr, 2015; see Section 4.2), these data have remained largely untouched in media and journalism studies. This is unfortunate for a number of reasons.

Above all, reference corpora are not just unstructured "blobs" (Davies, 2018, p. 19), which data in aggregators, for instance, arguably are, but present material in a "ready-to-use" format. On the one hand, this means that neither scripting for downloading materials nor preprocessing (e.g. removal of duplicates, meta-information and boilerplate material; lemmatization; part-of-speech tagging) is required. On the other hand, researchers can access the data through online interfaces that are comparatively easy to use and for which there is ample documentation and support material (e.g. in the form of corpus manuals or video tutorials), allowing a smooth start for the uninitiated. At the same time, it is clear that working with such "ready-to-use" material has its limits when it comes to addressing specific research questions that pertain to very recent data or data that are typically not part of such reference corpora (e.g. from social media, where the collection of one's own dataset, e.g. through APIs, may be the preferred approach, see Section 5.2).

As an illustration, consider the one-billion-word COCA, which currently covers the period 1990–2019 and consists of c. 250 million words of newspaper and magazine content (Davies, 2010). While the online interface allows for multiple types of searches (concordances, collocations, diachronic developments) and provides related information on general lexicographical measures, for instance (see Figure 1), it also facilitates contrastive searches (e.g. between news and other genres, between news outlets with different editorial stances, between news discourse from different decades or news articles vs. editorials, national vs. international news, etc.).

Figure 1. Browser interface of COCA (landing page)

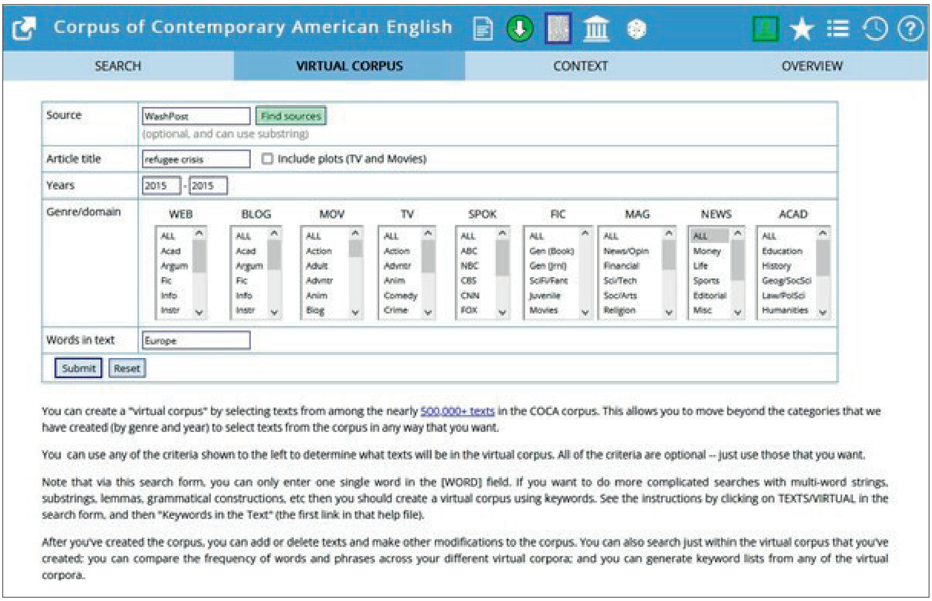


COCA could be used to add a quantitative dimension to analyses of news bias, such as Papacharissi and de Fatima Oliveira (2008), for instance. Concretely, this study of the portrayal of terrorism in US and British news could be complemented by a quick search that determines the semantic prosody of the salient term *terrorist* and possibly some (near-)synonyms such as *rebel* (or *freedom fighter*). Embedded in Entman's (1993) framing approach, an examination of the collocating adjectives to the left of the key term could be carried out in the "newspaper" section of the corpus to track "moral evaluation" in the text as one important instance of the framing process. Notwithstanding general methodological discussions regarding the application of framing as a concept and its operationalization on the textual level (see, e.g., Pentzold, 2017, pp. 521–523), such an analysis using the "Compare" function of the corpus interface reveals, among other things, that while *Islamic terrorist* and *Syrian rebel* are common collocations, the reverse combinations cannot be found in the US press (apart from one example from the *New York Times* in which *Islamic rebel government* is used as a paraphrasing translation of *mujahedeen*).

Using a different search strategy in the "Compare" display, collocations of *terrorist/rebel* and nouns appearing to the right could be identified, returning *terrorist organization(s)*, *terrorist networks*, and *terrorist cell(s)* as highly salient items while *rebel* co-occurs with arguably more "legitimate" (and positively connoted) items such as *force(s)*, *army*, or *movement*. This information on the semantic prosody (see also Section 4.1) provides some empirical insight into the conventional definitions, associations and usages of these terms as used in Western press discourse (which in turn can be contrasted with usages in other sociocultural settings) and could be combined with a manual content analysis of the framing types (episodic vs. thematic), as done in Papacharissi and de Fatima Oliveira (2008), for instance, or with a complementary analysis of multimodal framing (Pentzold, 2017, p. 516).

If a research question requires zooming in on a particular subset of the corpus data, this is also possible using the "virtual corpus" option, in which several filters can be applied individually or in combination depending on the researcher's goal. Criteria here include the selection of a source, a timeframe, a specific word in an article title, as well as specific words in a text (see Figure 2). This facilitates searches within a precisely defined dataset (e.g. in all articles from the *Washington Post* from the year 2015 that have the terms "refugee crisis" and in the title and "Europe" in the text) and is approximating the advanced search capabilities of aggregators such as *Nexis*, but without the need to pay licensing fees or preprocess the data before they can actually be queried along the lines illustrated.

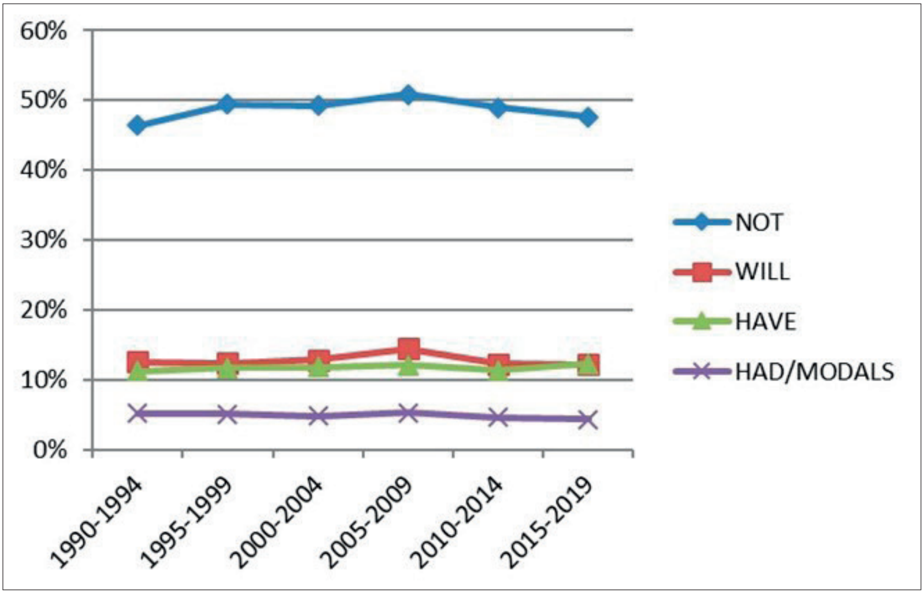
Figure 2. Creating a virtual corpus in COCA



Note that similar functionalities are also available in DeReKo (for German) and NOW, an evolving monitor corpus that uses the same interface as described above and further allows comparison between news from different English- and Spanish-speaking countries. While it cannot be denied that the bias towards English as the most widely studied (and thus most resourced) language in the world gives an advantage to researchers working with this language (see also Madrid-Morales, 2020, p. 76), there are a variety of other (linguistic and media-linguistic) corpora available that could be exploited in a similar way (see, e.g., Schrader-Kniffki et al., 2017 for an overview on Romance languages, including "minor" ones such as Catalan or Galician, or Raharjo et al., 2020 for an outline of plans to develop general corpora for Bahasa Indonesia). This raises the hope that the current Anglo-centrism might at least be levelled out in the future.

A short-term diachronic look at the news sections of reference corpora can further provide evidence on the style dimension of softening the news (or tabloidization) (Reinemann et al., 2012), if the data are searched for markers associated with emotional (e.g. evaluative adjectives and adverbs), colloquial (e.g. contractions and non-standard features), and subjective (e.g. combinations of first-person pronouns and private verbs) language, and their changing frequencies are assessed. Figure 3 synthesizes several searches in the newspaper section of COCA, showing contraction ratios for selected items (*not* vs. *n't*, *will* vs. *'ll*, *have* vs. *'ve*, *had*/modals vs. *'d*). It implies that in this domain, the number of "colloquial" variants apparently has levelled off, a finding that contrasts with results for the period 1960–1990, in which a marked increase of contractions, and thus evidence for tabloidization of newspaper style was found (Hundt & Mair, 1999).

Figure 3. Contraction ratios (in percent) of selected items in COCA (newspaper section)



Arguably, such an approach provides a more objective basis for generalizations than manual content annotation operating with ordinal scales (e.g. individual articles being coded on the "degree of colloquial language […] 3-point-scale, ranging from 0 (not colloquial at all) to 2 (very colloquial)"; see Steiner, 2021, pp. 4–5), which is additionally subject to certain practical limitations as regards personnel resources and interrater reliability, for instance. Nevertheless, to obtain a comprehensive picture, the corpus-linguistic analysis could be complemented by other types of content analyses on the topic and focus dimensions, where manual rating (still) reigns superior.

5.2 Free corpus tool

While carefully compiled and curated reference corpora offer several opportunities for theory-driven work with a broader scope (see Section 5.1), one of their main limitations is that they work only with a predefined set of materials (see Marchi, 2022, pp. 577–578). In addition, they may be limited in terms of historical coverage (but cf. the *Corpus of Historical American English*, for instance, which goes back to the 1820s, contains a 150-million-word portion of newspapers and magazines, and works through the same interface as COCA; english-corpora.org/coha) and as regards their availability for smaller languages. These constraints can be overcome by using standalone corpus tools.

Bednarek and Carr (2021, p. 133; see also Jaworska, 2018, pp. 93–100) list and illustrate several essential functions of such tools (some of which were already mentioned in passing in Sections 4 and 5.1), summarized in Table 3.

Table 3. Essential functions of corpus-linguistic tools

Function	Description
Concordance/Key Word in Context (KWIC)	Inspection of the context of a specific node (word or a combination of words) for qualitative analysis (e.g. as regards connotations)
Wordlist	Frequency list of all words in a corpus
Collocation	List of items that commonly occur within a specific span to the left or right of a target node (word or combination of words)
N-grams/chunks	Repeatedly occurring combinations of *n* words (e.g. trigrams) in the data
Keyness	Identification of items with high keyness/salience, i.e. high relative frequency of words (and potentially also lemmas and n-grams) in one corpus vs. another (reference) corpus
Clusters	Repeatedly occurring combinations (of varying size) of a specific node (word or word combination)
Wordcloud	Display of word frequencies in a corpus based on a wordlist, intuitively visualizing pervasiveness of individual words (commonly: more frequent words appearing larger)

Note. Several of the functions are also available in NLP toolkits (see note to Table 4) commonly used in media and journalism studies. However, informal screening of pertinent publications suggests that these are rarely applied in research.

Table 4. Common corpus-linguistic tools

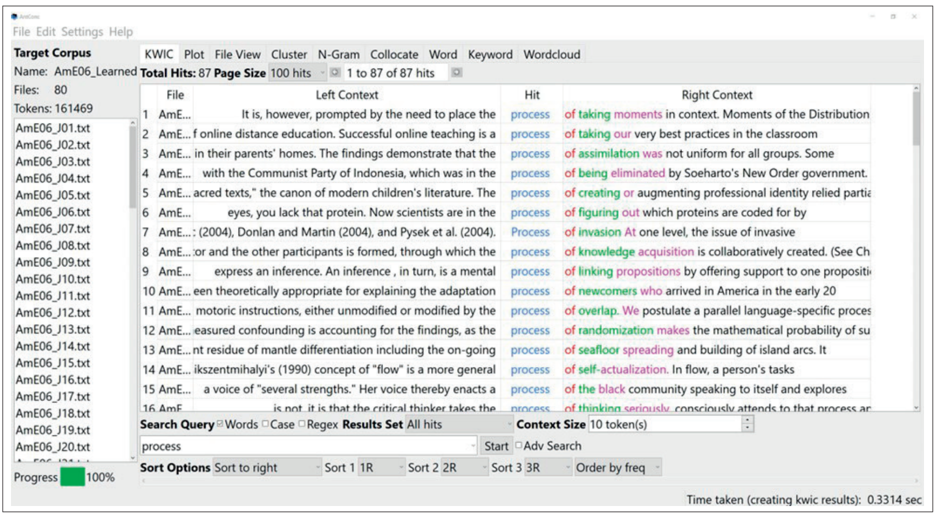
Name	Type	Licensing fee	Editable/ open source	Link	Updated	Notes
AntConc	Standalone desktop app	No	No	laurenceanthony.net/software/antconc/	Yes	Additional corpus analysis (e.g. spelling variant analysis, vocabulary level profiles) and processing (e.g. part-of-speech-tagging) software available (laurenceanthony.net/software.html)
IMS Open Corpus Workbench	Server-based	No	Yes	cwb.sourceforge.io/index.php	Yes	
LancsBox	Standalone desktop app	No	No	corpora.lancs.ac.uk/lancsbox	Yes	Integrated visualization features and statistical analysis
SketchEngine	Browser-based	Yes	No	app.sketchengine.eu	Yes	Open source version *NoSketchEngine* (nlp.fi.muni.cz/trac/noske) available
WordSmith Tools	Standalone desktop app	Yes	No	lexically.net/wordsmith/	Yes	
Wmatrix	Browser-based	Yes	No	ucrel.lancs.ac.uk/wmatrix	Yes	Semantic tagging and keyword analysis

Note. There are text mining tools (like *Voyant Tools*; voyant-tools.org) that offer partly comparable functionality to corpus-linguistic tools as well as "toolkits" for textual analysis implemented in programming languages like R or Python, which are customary in NLP, but ignored here due to the focus on corpus-linguistic tools. Examples are *quanteda* (quanteda.io), *NLTK* (nltk.org) or *ConvoKit* (convokit.cornell.edu), with the last example being specifically designed for the analysis of conversations. Further, there is a broad range of other corpus tools that are no longer updated or only work on specific operating systems (e.g. *CorpKit*, a standalone open source desktop app for Mac and Linux, last updated in 2016; github.com/interrogator/corpkit).

As a proviso, laments should be noted that "[w]hile such tools are typically easy to use, they also impose limitations on researchers, confining their study designs to the predefined boundaries of the software" (Niekler et al., 2023, p. 327). At the same time, it is evident that corpus-linguistic tools are still helpful because of their transparent structure, easy availability and usability, and the fact that they can work flexibly without sophisticated technical infrastructure and independently of the languages studied. By contrast, more sophisticated (NLP) text-analytic tools rely on complex, possibly server-based technology, often require familiarity with bulky user manuals and rely on language models that may not be available for smaller languages (see also Section 3), which means that they come with their own inherent boundaries. In particular, the dominance of Western languages, and especially English, in text-analytic NLP tools has been identified as an area of concern (see Baden et al., 2022, p. 2), as it hinders the participation of other researcher communities and impedes comparative research on an international scale. A list of available tools commonly used in linguistics is shown in Table 4.

Subsequently, *AntConc* is introduced (see Figure 4). It represents the tool most widely used by linguists and has also been extensively applied within the broader sphere of digital humanities and social sciences (see, e.g., Froehlich, 2015; Smith, 2021; Li & Zhang, 2022; Hui, 2023). It also is the tool arguably best documented (see, e.g., the "User Support" section on laurenceanthony.net/software/antconc, and the numerous instructional videos available at youtube.com/results?search_query=antconc) and – unlike other commonly used commercial tools such as *WordSmith* or the browser-based *SketchEngine* – it is available for free. Along the lines of "reflected algorithmic textual analysis" as presented in Section 4, the central question again is one of proper operationalization of concepts that will facilitate relevant theory-driven research.

Figure 4. ***AntConc*** **GUI with KWIC display shown**



Note. Extracted from laurenceanthony.net/software/antconc/screenshots/kwic.png.

To illustrate, consider an analysis to explore objectivity as a core journalistic value (see Donsbach & Klett, 1993) in emerging online news practice in the form of political live blogging (see Michael & Werner, 2023). The Wordlist and N-gram functions of *AntConc* were used to explore a topical corpus, overall establishing fact-based discourse as well as an emphasis on the liveness of the event. Other salient items that emerged from these analyses, such as the frequent combinations *X said* and *I think* or could be subjected to closer scrutiny through Concordance analysis, determining a clear linguistic delineation of reporting and evaluation/ expression of personal stance as well as a partly informal tone. Additional linguistic markers associated with "objective" and "subjective" discourse, such as markers of colloquiality (vs. their more formal variants) could be identified in the literature and traced and quantified in the data, also using the Concordance function. To take advantage of a combined quantitative-qualitative corpus analysis with other approaches, these findings could be complemented by a manual content analysis of sourcing practices to determine whether and how political live blogging strives for transparent and objective reporting through accounting for sources and linking information.

A corpus-based analysis with *AntConc* could also be triangulated with experimental work, such as Burgers and de Graaf (2013). This study starts from the hypothesis that "the degree to which a newspaper article refrains from using neutral language may be an indicator of sensationalistic news" (Burgers & de Graaf, 2013, p. 169). Formally, sensationalism in print news thus is characterized through heightened "language intensity", which is specifically promoted through the use of emotional language (see also Lefkowitz, 2021). In the experiment, participant assessments on the perceived intensity and credibility of a restricted set of news items were elicited under various conditions (intensifiers vs. no intensifiers; positive vs. negative framing). As intensification can operate on various levels (see Burgers & de Graaf, 2013, p. 170), using the Concordance function, a larger corpus of news articles could be searched based on a list of relevant items, such as adjectives (*massive*, *gigantic*), adverbs (*even higher*, *much more intense*), quantifiers (*billions*), nouns (*spectacle, event*), verbs (*gorge* vs. *eat*), exaggerations (*They had to wait forever*) and repetition (*really, really awful*) and results on the presence of such elements could be quantified to be able to assess the amount of sensationalistic reporting across different news categories and across various news outlets. A complementary strategy would be to use the Keyness function, which could be applied to determine (i) whether relevant items are actually salient in "tabloid" news outlets, which are generally associated with more sensationalistic reporting, if material from them is contrasted with "quality" press data and (ii) whether there have been any changes in the amount of sensationalistic reporting over time. Obviously, such analyses can therefore also be related to the style dimension of tabloidization and could be further complemented by a manual content analysis of broader journalistic strategies such as person-centered reporting (Otto et al., 2017, pp. 141–142).

In addition to studying purpose-built self-compiled corpora, *AntConc* can also be used for the linguistic processing of materials from commercial aggregators (see Section 4.2 for examples), reference corpora of various languages that do not have an online interface but are freely available in textual format (see, e.g., the *Helsinki Corpus of Swahili*; korp.csc.fi/download/HCS/a-v2),[11] as well as for analyzing material from digital press archives, such as *Chronicling America* (English; chroniclingamerica.loc.gov), *ANNO* (German; anno.onb.ac.at), or *RetroNews* (French; retronews.fr). While these archives vary considerably in terms of their temporal coverage, level of textual processing and accuracy, search functionalities and filters, as well as data output types (see Koenen, 2021; Michael & Werner, in press), they commonly at least allow the download of textual data, which then (possibly after accuracy checks and pre-processing) can be loaded into the corpus software and queried along the lines presented above. Given the progress in the development of these archives over the past decade, it is hoped that future efforts in digital humanities and social sciences will expand access to and increase the quality of such materials (Marchi, 2022, p. 579), facilitating historical and diachronic research on news discourse. Again, while *AntConc* is "blind" to the language of the processed material, and also allows the analysis of right-to-left languages, for instance, a certain bias in the availability of these archives for larger and Western languages cannot be denied, and greater inclusivity remains a desideratum.

6. Conclusion

The present paper can be seen as a contribution towards the methodological discussion in media and journalism studies (as well as communication studies at large), arguing for an interdisciplinary cross-fertilization between the fields of corpus linguistics and media and journalism studies. Overall, it argued that corpus linguistics, a specific form of computer-assisted textual analysis, has the potential to become part of the "interdisciplinary turn" or "new interdisciplinarity" of media and journalism studies (see Section 1). An ensuing goal was to (re-)raise awareness about corpus linguistics, which to date has suffered from large-scale neglect and has not become part of the canon of computational approaches. It was proposed that this undervaluation (especially compared to NLP approaches) is unjustified, as corpus-linguistic analysis has much to contribute to content analysis, relating to central theories and concepts traceable at the level of the surface text, such as framing, agenda-setting, media bias, or emotionalization (Hase et al., 2022, p. 71).

It is also clear that several of the functionalities typically featuring in corpus linguistics (e.g. collocation or keyness analysis) are also already available in other NLP tools more widely used in media and journalism studies but arguably have only rarely been applied in research for various reasons. Therefore, it was suggested that corpus-linguistic approaches can be a potentially helpful addition to the (digital) methodological toolbox as an efficient "middle-ground" empirical approach that allows researchers of both contemporary and historical news discourse to conduct theory-driven work and to maintain a fair degree of control over the data and research procedures, while it facilitates the integration of quantitative and qualitative perspectives and can be meaningfully triangulated with other methods.

11 Note that COCA (Section 5.1) can also be downloaded and searched in tools such as *AntConc*.

This cross-validation through hybrid research designs, which to date arguably still is in a fledgling state, has even occasionally been labelled the "gold standard" in empirical media and journalism research as the only viable way to adequately describe and analyze socially complex phenomena (Loosen & Schmidt, 2016, p. 563; see also DeCoster & Lichtenstein, 2007). *Mutatis mutandis*, (corpus-based) media linguistics could also benefit from going beyond textual analysis to be in a position to better motivate results, as variously shown in Section 4.1.

Corpus linguistics can also be viewed as part of an effort to fill (or at least circumvent) some of the major gaps identified in current automated text analysis practice (Baden et al., 2022), such as its ability to facilitate theory-driven work and its relative flexibility when it comes to integrating and processing various types of textual data from different languages. In this sense, if properly applied, corpus-linguistic approaches, even though they have existed for a longer time, could be viewed as methodological innovation that serves as a driver for media and journalism studies at large (Sjøvaag & Karlsson, 2016, p. 87).

At the same time, it is clear that corpus linguistics is not a silver bullet and that, depending on the data assessed and individual research questions being tackled, relying on automated content-analytic approaches (or indeed different methodologies altogether) may be superior. Very much in the spirit of Bednarek and Carr (2021, p. 147), corpus linguistics in its various forms therefore was presented as *one* possible addition to the media and journalism scholar's toolkit when it comes to analyzing relevant texts, and the concrete choice of one method or the other in a specific study may be influenced by various theoretical and practical considerations.

Furthermore, while corpus-linguistic tools were presented as easily accessible and usable for media and journalism scholars, continuing collaborative efforts between them and corpus linguists will certainly bear fruit in the sense of aiming for a holistic picture of news discourse, conducting contextualized "reflected algorithmic textual analysis" (Pichler & Reiter, 2020, p. 58; see also Marchi, 2010, p. 165, 2022, p. 585) as a genuinely interdisciplinary endeavor. While there are already a few examples where interdisciplinary collaboration between corpus linguists and media and journalism scholars has successfully been realized (i.e. in terms of publications in leading outlets) on a practical (and also theoretical and methodological) level (see, e.g., Bednarek et al., 2021; Michael & Werner, 2023), this position is also supported by the fact that corpus linguistics is an evolving field in its own right, with other tools and resources than the ones showcased above being available (see Table 3) and being developed by the field for the field itself. Some of them include additional functionalities, such as integrated visualization and subsequent statistical analysis (e.g. *LancsBox*), semantic tagging and keyword analysis (also for languages such as Mandarin Chinese, Indonesian or Dutch), which may be helpful for linguistically informed topic modelling (e.g. *Wmatrix*), as well as the possibility of highly sophisticated queries based on individualized corpus annotation (e.g. *IMS Open Corpus Workbench*). This dynamism similarly applies to methodological reflection and development, for instance as regards corpus query techniques (see, e.g., Bednarek & Bray, 2023) and concrete (statistical) measurements geared to linguistic data (see, e.g., McEnery et al., 2019).

While the current contribution has highlighted the advantages of corpus-linguistic approaches, there are certainly some general limitations that have to be acknowledged. These include the very fact that to date researchers in media and journalism studies may be more familiar with environments such as R or Python, which can also be used to process language data. However, the fact that they require programming skills on the part of the researcher, whereas most corpus-linguistic tools do not (Bednarek & Carr, 2021, p. 146) and come with accessible interfaces (see Section 5), can be considered a strong argument especially for those venturing into empirical textual analysis for the first time. On this note, the practical experience of the authors has shown that the basics of corpus linguistics along the lines developed above (introduction to working with online reference corpora and with *AntConc*) can be taught within a longer session (six contact hours) of an undergraduate interdisciplinary media studies/media linguistics course.

Other potential concerns include the risk of corpus tools or online environments becoming obsolete and the concentration of corpus-based work on textual data. While the former could be considered a natural process as resources and methods evolve, the latter is salient as current news practice, especially in online environments, increasingly relies on visual elements, which may require an extension to multimodal perspectives to be able to account for the interaction between different modes (Bucher, 2014, p. 277; Jaworska, 2018, pp. 103–104). While more work accounting for the interaction between different modes is needed in the future, there are already innovative attempts, for instance on the multimodal construction of newsworthiness (Bednarek & Caple, 2012, 2017).

The focus of the present paper was on news discourse with an emphasis on legacy (print and online) media. However, corpus resources and relevant work are also available on other media, such as blogs and social media (see, e.g., Lutzky & Kehoe, 2022) as well as on transcribed media speech (e.g. the spoken section of COCA or the spoken news commentary section of the BNC) and on entertainment media (see, e.g., the *Sydney Corpus of Television*, syd-tv.com or the *TV Corpus*; english-corpora.org/tv). This means that also work in these domains, which to date appears to have remained ignorant of corpus approaches as well (e.g. Nitsch, 2023), could be conveniently complemented by corpus approaches to increase the validity of results of pertinent content analyses.

References

Atanasova, D., Koteyko, N., Brown, B., & Crawford, P. (2019). Representations of mental health and arts participation in the national and local British press, 2007–2015. *Health*, *23*(1), 3–20. https://doi.org/10.1177/1363459317708823

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures*, *14*(3), 165–183. https://doi.org/10.1080/19312458.2020.1803247

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communi-*

cation Methods and Measures, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Balfour, J. (2023). *Representing schizophrenia in the media: A corpus-based approach to UK press coverage*. Routledge. https://doi.org/10.4324/9781003096054

Bednarek, M, & Bray, C. (2023). Trialing corpus search techniques for identifying person-first and identity-first language. *Applied Corpus Linguistics*, *3*(1), 100046. https://doi.org/10.1016/j.acorp.2023.100046

Bednarek, M., & Caple, H. (2012). *News discourse*. Continuum.

Bednarek, M., & Caple, H. (2017). *The discourse of news values: How news organizations create newsworthiness*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190653934.001.0001

Bednarek, M., & Carr, G. (2021). Computer-assisted digital text analysis for journalism and communications research: Introducing corpus linguistic techniques that do not require programming. *Media International Australia*, *181*(1), 131–151. https://doi.org/10.1177/1329878X20947124

Bednarek, M., Caple, H., & Huan, C. (2021). Computer-based analysis of news values: A case study on national day reporting. *Journalism Studies*, *22*(6), 702–722. https://doi.org/10.1080/1461670X.2020.1807393

Bell, A. (1995). Language and the media. *Annual Review of Applied Linguistics*, *15*, 23–41. https://doi.org/10.1017/S0267190500002592

Bertrand, I., & Hughes, P. (2017). *Media research methods: Audiences, institutions, texts*. Bloomsbury.

Biber, D. (2003). Compressed noun-phrase structures in newspaper discourse: The competing demands of popularization. In J. Aitchison & D. M. Lewis (Eds.), *New media language* (pp. 169–181). Routledge.

Bös, B., & Kornexl, L. (Eds.). (2015). *Changing genre conventions in historical English news discourse*. Benjamins. https://doi.org/10.1075/ahs.5

Boumans, J. W., & Trilling, D. (2018). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. In M. Karlsson & H. Sjøvaag (Eds.), *Rethinking research methods in an age of digital journalism* (pp. 8–23). Routledge.

Bucher, H.-J. (2014). Sprach- und Diskursanalyse in der Medienforschung [Language and discourse analysis in media research]. In M. Karmasin, M. Rath, & B. Thomaß (Eds.), *Kommunikationswissenschaft als Integrationsdisziplin* (pp. 271–298). Springer. https://doi.org/10.1007/978-3-531-19016-7_14

Bucher, H-J. (2017). Massenmedien als Handlungsfeld: Printmedien [Mass media as field of action: Print media]. In K. S. Roth, M. Wengeler, & A. Ziem (Eds.), *Handbuch Sprache in Politik und Gesellschaft* (pp. 298–333). Mouton de Gruyter. https://doi.org/10.1515/9783110296310-014

Bucher, H.-J., & Niemann, P. (2018). Infrastrukturen zur Erforschung medienspezifischer Sprachverwendung [Infrastructures for researching media-specific language use]. In H. Lobin, R. Schneider, & A. Witt (Eds.), *Digitale Infrastrukturen für die germanistische Forschung* (pp. 177–196). Mouton de Gruyter. https://doi.org/10.1515/9783110538663-009

Burgers, C., & de Graaf, A. (2013). Language intensity as a sensationalistic news feature: The influence of style on sensationalism perceptions and effects. *Communications*, *38*(2), 167–188. https://doi.org/10.1515/commun-2013-0010

Carlsen, H. B., & Ralund, S. (2022). Computational grounded theory revisited: From computer-led to computer-assisted text analysis. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517221080146

Chen, Y., Peng, Z., Kim, S. H., & Choi, C. W. (2023). What we can do and cannot do with topic modeling: A systematic review. *Communication Methods and Measures*, *17*(2), 111–130. https://doi.org/10.1080/19312458.2023.2167965

Clark, C. (2013). It's always the same old news! A diachronic analysis of shifting newspaper language style, 1993–2005. In J. Bamford, S. Cavalieri, & G. Diani (Eds.), *Variation and change in spoken and written discourse: Perspectives from corpus linguistics* (pp. 269–282). Benjamins. https://doi.org/10.1075/ds.21.25cla

Collins, L. C., & Nerlich, B. (2016). Uncertainty discourses in the context of climate change: A corpus-assisted analysis of UK national newspaper articles. *Communications*, *41*(3), 291–313. https://doi.org/10.1515/commun-2016-0009

Conboy, M. (2014). Exploring the language of the popular in American and British newspapers 1833–1988. *Journal of Historical Pragmatics*, *15*(2), 159–164. https://doi.org/10.1075/jhp.15.2.01con

Davies, M. (2010). The *Corpus of Contemporary American English* as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

Davies, M. (2017). *The new 4.3 billion word NOW corpus, with 4–5 million words of data added every day*. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2017/general/paper250.pdf

Davies, M. (2018). Using large online corpora to examine lexical, semantic, and cultural variation in different dialects and time periods. In E. Friginal (Ed.), *Studies in corpus-based sociolinguistics* (pp. 19–82). Routledge.

DeCoster, J., & Lichtenstein, B. (2007). Integrating quantitative and qualitative methods in communication research. *Communication Methods and Measures*, *1*(4), 227–242. https://doi.org/10.1080/19312450701636599

Donsbach, W., & Klett, B. (1993). Subjective objectivity: How journalists in four countries define a key term of their profession. *Gazette*, *51*, 53–83. https://doi.org/10.1177/001654929305100104

Duguid, A. (2010). Newspaper discourse informalisation: A diachronic comparison from keywords. *Corpora*, *5*(2), 109–138. https://doi.org/10.3366/E1749503210000419

Dunn, J. (2022). *Natural language processing for corpus linguistics*. Cambridge University Press. https://doi.org/10.1017/9781009070447

Ebeling, S. O. (2013). Semantic prosody in a cross-linguistic perspective. In M. Huber & J. Mukherjee (Eds.), *Corpus linguistics and variation in English: Focus on non-native Englishes*. Varieng. https://varieng.helsinki.fi/series/volumes/13/ebeling/

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*, 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

European Strategy Forum on Research Infrastructures. (2021). *Roadmap 2021: Strategy report on research infrastructures*. ESFRI. https://roadmap2021.esfri.eu/media/1295/esfri-roadmap-2021.pdf

Fraas, C., & Pentzold, C. (2016). Diskursanalyse in der Kommunikationswissenschaft [Discourse analysis in communication studies]. In S. Averbeck-Lietz & M. Meyen (Eds.), *Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft* (pp. 227–240). Springer. https://doi.org/10.1007/978-3-658-01656-2_14

Froehlich, H. (2015). Corpus analysis with *AntConc*. *Programming Historian*, *4*. https://doi.org/10.46430/phen0043

Fürsich, E. (2009). In defense of textual analysis: Restoring a challenged method for journalism and media studies. *Journalism Studies*, *10*(2), 238–252. https://doi.org/10.1080/14616700802374050

Haider, A. S., & Hussein, R. F. (2020). Analysing headlines as a way of downsizing news corpora: Evidence from an Arabic-English comparable corpus of newspaper articles. *Digital Scholarship in the Humanities*, *35*(4), 826–844. https://doi.org/10.1093/llc/fqz074

Hansen, K. R. (2016). News from the future: A corpus linguistic analysis of future-oriented, unreal and counterfactual news discourse. *Discourse & Communication*, *10*(2), 115–136. https://doi.org/10.1177/1750481315611240

Hase, V. (2023). Automated content analysis. In F. Oehmer-Pedrazzi, S. H. Kessler, E. Humprecht, K. Sommer, & L. Castro (Eds.), *Standardized content analysis in communication research: A handbook* (pp. 23–36). Springer. https://doi.org/10.1007/978-3-658-36179-2_23https://doi.org/10.1007/978-3-658-36179-2_3

Hase, V., Mahl, D., & Schäfer, M. S. (2022). Der „Computational Turn“: ein „interdisziplinärer Turn“? Ein systematischer Überblick zur Nutzung der automatisierten Inhaltsanalyse in der Journalismusforschung [The “computational turn”: An “interdisciplinary turn“? A systematic overview of automated content analysis in journalism studies]. *Medien & Kommunikationswissenschaft*, *70*(1-2), 60–78. https://doi.org/10.5771/1615-634X-2022-1-2-60

Hui, H. (2023). What can digital humanities do for literary adaptation studies: Distant reading of children’s editions of *Robinson Crusoe*. *Digital Scholarship in the Humanities*, *38*(4), 1564–1576. https://doi.org/10.1093/llc/fqad059

Hundt, M., & Mair, C. (1999). “Agile” and “uptight” genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics*, *4*(2), 221–242. https://doi.org/10.1075/ijcl.4.2.02hun

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Linquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, *17*(3), 805–826. https://doi.org/10.1177/17456916211004899

Jaworska, S. (2018). Corpus approaches: Investigating linguistic patterns and meanings. In C. Cotter & D. Perrin (Eds.), *The Routledge handbook of language and media* (pp. 93–108). Routledge. https://doi.org/10.4324/9781315673134-8

Jaworska, S., & Krishnamurthy, R. (2012). On the F word: A corpus-based analysis of the media representation of feminism in British and German press discourse, 1990–2009. *Discourse & Society*, *23*(4), 401–431. https://doi.org/10.1177/0957926512441113

Jeffries, L., & Walker, B. (2019). *Keywords in the press: The New Labour years*. Bloomsbury.

Johnson, J. H., & Partington, A. (2018). Corpus-assisted discourse study of representations of the “underclass” in the English-language press: Who are they, how do they behave, and who is to blame for them? In E. Friginal (Ed.), *Studies in corpus-based sociolinguistics* (pp. 293–318). Routledge.

Kelsey, D. (2017). Journalism and critical discourse studies. In J. Flowerdew & J. E. Richardson (Eds.), *The Routledge handbook of critical discourse studies* (pp. 510–524). Routledge. https://doi.org/10.4324/9781315739342-35

Koenen, E. (2021). Forschungssoftware für die Kommunikations- und Mediengeschichte: Epistemologische Herausforderungen und Perspektiven [Research software for communication and media history: Epistemological challenges and perspectives]. *Medien & Kommunikationswissenschaft*, *69*(1), 117–135. https://doi.org/10.5771/1615-634X-2021-1-117

Krennmayr, T. (2015). What corpus linguistics can tell us about metaphor use in newspaper texts. *Journalism Studies*, *16*(4), 530–546. https://doi.org/10.1080/1461670X.2014.937155

Kulkarni, R. (2021). A million news headlines. *Kaggle*. https://www.kaggle.com/datasets/therohk/million-headlines

Landert, D. (2015). Reportable facts and a personal touch: The functions of direct quotes in online news. In J. Arendholz, W. Bublitz, & M. Kirner-Ludwig (Eds.), *The pragmatics of quoting now and then* (pp. 29–52). Mouton de Gruyter. https://doi.org/10.1515/9783110427561-003

Lefkowitz, J. (2018). "Tabloidization" or dual-convergence: Quoted speech in tabloid and "quality" British newspapers 1970–2010. *Journalism Studies*, *19*(3), 353–375. https://doi.org/10.1080/1461670X.2016.1190662

Lefkowitz, J. (2021). *"Tabloidization" and the Internet age* [Doctoral dissertation, University of Oxford]. https://ora.ox.ac.uk/objects/uuid:20f23190-ea82-4cd6-ab5b-de8e09e21f77/download_file?file_format=application%2Fpdf&safe_filename=Lefkowitz_2021_Tabloidization_and_the.pdf&type_of_work=Thesis

Lenk, H. E. H. (2013). Sinn und Nutzen textlinguistischer Medienforschung [Purpose and usefulness of textlinguistic media research]. In Z. Berdychowska, Z. Bilut-Homplewicz, & B. Mikołajczyk (Eds.), *Textlinguistik als Querschnittsdisziplin* (pp. 67–83). Lang.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, *57*(1), 34–52. https://doi.org/10.1080/08838151.2012.761702

Li, K., & Zhang, Q. (2022). A corpus-based study of representation of Islam and Muslims in American media: Critical Discourse Analysis approach. *International Communication Gazette*, *84*(2), 157–180. https://doi.org/10.1177/1748048520987440

Loosen, W., Nölleke, D., & Springer, N. (2022). Journalismusforschung: Disziplin durch Entdisziplinierung [Journalism research: Discipline through untangling disciplines]. *Medien & Kommunikationswissenschaft*, *70*(1-2), 3–16. https://doi.org/10.5771/1615-634X-2022-1-2-3

Loosen, W., & Schmidt, J-H. (2016). Multi-method approaches. In T. Witschge, C. W. Anderson, D. Domingo, & A. Hermida (Eds.), *The Sage handbook of digital journalism* (pp. 562–575). Sage.

Louw, W. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157–176). Benjamins. https://doi.org/10.1075/z.64.11lou

Lüger, H.-H. (2017). Media linguistics: Interfaces to media and communication studies. In K. Bedijs & C. Maaß (Eds.), *Manual of Romance languages in the media* (pp. 10–32). Mouton de Gruyter. https://doi.org/10.1515/9783110314755-002

Lutzky, U., & Kehoe, A. (2022). Using corpus linguistics to study online data. In C. Vasquez (Ed.), *Research methods for digital discourse analysis* (pp. 219–236). Bloomsbury.

Madrid-Morales, D. (2020). Using computational text analysis tools to study African online news content. *African Journalism Studies*, *41*(4), 68–82. https://doi.org/10.1080/23743670.2020.1820885

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20–33. https://doi.org/10.1080/08838151.2012.761700

Marchi, A. (2010). "The moral in the story": A diachronic investigation of lexicalised morality in the UK press. *Corpora*, *5*(2), 161–189. https://doi.org/10.3366/E1749503210000432

Marchi, A. (2022). Corpus linguistics in the study of news media. In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 576–588). Routledge. https://doi.org/10.4324/9780367076399-40

Marín Arrese, J. I. (2015). Epistemicity and stance: A cross-linguistic study of epistemic stance strategies in journalistic discourse in English and Spanish. *Discourse Studies*, *17*(2), 210–225. https://doi.org/10.1177/1461445614564523

Mautner, G. (2008). Analyzing newspapers, magazines and other print media. In R. Wodak & M. Krzyżanowski (Eds.), *Qualitative discourse analysis in the social sciences* (pp. 30–53). Palgrave Macmillan.

McEnery, T., Brezina, V., & Baker, H. (2019). Usage fluctuation analysis: A new way of analyzing shifts in historical discourse. *International Journal of Corpus Linguistics*, *24*(4), 413–444. https://doi.org/10.1075/ijcl.18096.mce

Meier, H. E., Rose, A., & Hölzen, M. (2017). Spirals of signification? A corpus linguistic analysis of the German doping discourse. *Communication & Sport*, *5*(3), 352–373. https://doi.org/10.1177/2167479515610151

Meyer, C. F. (2023). *English corpus linguistics*. Cambridge University Press. https://doi.org/10.1017/9781107298026

Michael, H., & Werner, V. (2023). Hybrid news (in the) making: A content and corpus-based discourse analysis of political live blogs on the 2020 US presidential debates. *Journalism Practice*. https://doi.org/10.1080/17512786.2023.2215254

Michael, H., & Werner, V. (in press). Digital press archives for media and communication history research: From "reading rooms" to virtual research environments. *Studies in Communication Sciences*.

Moon, R. (2016). A corpus-linguistic analysis of news coverage in Kenya's *Daily Nation* and *The Times* of London. *International Journal of Communication*, *10*, 2381–2401. http://hdl.handle.net/1773/33131

Nartey, M., & Mwinlaaru, I. N. (2019). Towards a decade of synergising corpus linguistics and critical discourse analysis: A meta-analysis. *Corpora*, *14*(2), 203–235. https://doi.org/10.3366/cor.2019.0169

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Niekler, A., Kahmann, C., Burghardt, M., & Heyer, G. (2023). The interactive Leipzig Corpus Miner: An extensible and adaptable text analysis tool for content analysis. *Publizistik*, *68*, 325–354. https://doi.org/10.1007/s11616-023-00809-4

Nitsch, C. (2023). Content analysis in the research field of fictional entertainment. In F. Oehmer-Pedrazzi, S. H. Kessler, E. Humprecht, K. Sommer, & L. Castro (Eds.), *Standard-*

ized content analysis in communication research: A handbook (pp. 265–275). Springer. https://doi.org/10.1007/978-3-658-36179-2_23

O'Keeffe, A. (2012). Corpora and media studies. In K. Hyland, M. Handford, & C. M. Huat (Eds.), *Corpus applications in applied linguistics* (pp. 117–131). Continuum.

O'Halloran, K. (2010). How to use corpus linguistics in the study of media discourse. In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 563–577). Routledge.

Oehmer-Pedrazzi, F., Kessler, S. H., Humprecht, E., Sommer, K., & Castro, L. (Eds.). (2023). *Standardized content analysis in communication research: A handbook*. Springer. https://doi.org/10.1007/978-3-658-36179-2

Otto, L., Glogger, I., & Boukes, M. (2017). The softening of journalistic political communication: A comprehensive framework model of sensationalism, soft news, infotainment, and tabloidization. *Communication Theory*, *27*(2), 136–155. https://doi.org/10.1111/comt.12102

Papacharissi, Z., & de Fatima Oliveira, M. (2008). News frames terrorism: A comparative analysis of frames employed in terrorism coverage in U.S. and U.K. newspapers. *The International Journal of Press/Politics*, *13*(1), 52–74. https://doi.org/10.1177/1940161207312676

Paulsen, I. (2022). *The emergence of American English as a discursive variety: Tracing enregisterment processes in nineteenth-century U.S. newspapers*. Language Science Press. https://doi.org/10.5281/zenodo.6207627

Pentzold, C. (2017). Kommunikationswissenschaft [Communication studies]. In K. S. Roth, M. Wengeler, & A. Ziem (Eds.), *Handbuch Sprache in Politik und Gesellschaft* (pp. 514–532). Mouton de Gruyter. https://doi.org/10.1515/9783110296310-023

Philo, G. (2007). Can discourse analysis successfully explain the content of media and journalistic practice? *Journalism Studies*, *8*(2), 175–196. https://doi.org/10.1080/14616700601148804

Pichler, A., & Reiter, N. (2020). Reflektierte Textanalyse [Reflected textual analysis]. In N. Reiter, A. Pichler, & J. Kuhn (Eds.), *Reflektierte algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt* (pp. 43–60). Mouton de Gruyter. https://doi.org/10.1515/9783110693973-003

Ping, Y. (2022). Representations of the 2014 Hong Kong protests in journalistic translation: A corpus-based critical framing analysis of Chinese and English news coverage. *Journalism*, *23*(7), 1509–1529. https://doi.org/10.1177/14648849221074556

Raharjo, S., Utami, E., Yusa, M., & Sutanta, E. (2022). Systematic literature review: Corpus linguistics in Indonesia. In C. Stephanidis, M. Antona, & S. Ntoa (Eds.), *HCII 2022: Communications in computer and information science*. Springer. https://doi.org/10.1007/978-3-031-06417-3_50

Rayson. P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549. https://doi.org/10.1075/ijcl.13.4.06ray

Reinemann, C., Stanyer, J., Scherr, S., & Legnante, G. (2012). Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism*, *13*(2), 221–239. https://doi.org/10.1177/1464884911427803

Riffe, D., Lacy, S., Watson, B. R., & Fico, F. (2020). *Analyzing media messages: Using quantitative content analysis in research*. Routledge. https://doi.org/10.4324/9780429464287

Santaemilia, J., & Maruenda, S. (2014). The linguistic representation of gender violence in (written) media discourse: The term "woman" in Spanish contemporary newspapers.

Journal of Language Aggression and Conflict, *2*(2), 249–273. https://doi.org/10.1075/jlac.2.2.04san
Scharrer, E., & Ramasubramanian, S. (2021). *Quantitative research methods in communication: The power of numbers for social justice*. Routledge. https://doi.org/10.4324/9781003091653
Schrader-Kniffki, M., Colominas, C., Bedijs, K., Bouzas, P., Schneider, S., & Kallweit, D. (2017). Tertiary media corpora of the Romance languages. In K. Bedijs & C. Maaß (Eds.), *Manual of Romance languages in the media* (pp. 290–322). Mouton de Gruyter. https://doi.org/10.1515/9783110314755-014
Schudson, M. (1995). *The power of news*. Harvard University Press.
Schweinberger, M., & Haugh, M. (Eds.). (forthcoming). Reproducibility, replication and robustness in corpus linguistics. Special issue of the *International Journal of Corpus Linguistics*.
Scott, M. (2020). *WordSmith Tools* (Version 8). Lexical Analysis Software. https://lexically.net/wordsmith/
Scott, M. (2023, March 30). *News downloads and text coverage: Case studies in relevance*. Edge Hill Corpus Research Group, Aston. https://lexically.net/downloads/workshops/EdgeHill2023/index.html
Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Benjamins. https://doi.org/10.1075/scl.22
Shahin, S. (2016). When scale meets depth: Integrating natural language processing and textual analysis for studying digital corpora. *Communication Methods and Measures*, *10*(1), 28–50. https://doi.org/10.1080/19312458.2015.1118447
Sjøvaag, H., & Karlsson, M. (2016). Rethinking research methods for digital journalism studies. In B. Franklin & S. A. Eldridge (Eds.), *The Routledge companion to digital journalism studies* (pp. 87–95). Routledge. https://doi.org/10.4324/9781315713793-9
Smith, E. L. (2021). Review of AntConc (Version 3.5.8); WordSmith Tools (Version 8). *Renaissance and Reformation*, *44*(1), 200–214. https://www.jstor.org/stable/27085730
Sönning, L. (2024). Evaluation of keyness metrics: Performance and reliability. *Corpus Linguistics and Linguistic Theory*, *20*(2), 263–288. https://doi.org/10.1515/cllt-2022-0116
Sönning, L., & Werner, V. (2021). The replication crisis, scientific revolutions, and linguistics. *Linguistics*, *59*(5), 1179–1206. https://doi.org/10.1515/ling-2019-0045
Steensen, S. (2016). The intimization of journalism. In T. Witschge, C. Anderson, D. Domingo, & A. Hermida (Eds.). *The Sage handbook of digital journalism* (pp. 113–127). Sage.
Steiner, M. (2021). Soft news/tabloidization (journalistic reporting styles). *Database of Variables for Content Analysis*. https://doi.org/10.34778/2t
Touri, M., & Koteyko, N. (2015). Using corpus linguistic software in the extraction of news frames: Towards a dynamic process of frame analysis in journalistic texts. *International Journal of Social Research Methodology*, *18*(6), 601–616. https://doi.org/10.1080/13645579.2014.929878
van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2-3), 81–92. https://doi.org/10.1080/19312458.2018.1458084
Wahl-Jorgensen, K. (2019). *Emotions, media and politics*. Polity Press.
Waldherr, A., Geise, S., Mahrt, M., Katzenbach, C., & Nuernbergk, C. (2021). Toward a stronger theoretical grounding of computational communication science: How macro

frameworks shape our research agendas. *Computational Communication Research*, *3*(2), 152–179. https://doi.org/10.5117/CCr2021.02.002.wald

Windsor, L. C. (2021). Advancing interdisciplinary work in computational communication science. *Political Communication*, *38*(1-2), 182–191. https://doi.org/10.1080/10584609.2020.1765915

Wu, D. H., & Coleman, R. (2009). Advancing agenda-setting theory: The comparative strength and new contingent conditions of the two levels of agenda-setting effects. *Journalism & Mass Communication Quarterly*, *86*(4), 775–789. https://doi.org/10.1177/107769900908600404