Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-04514-7 OPEN

The evolution of online news headlines

Pietro Nickl [1,2,3]✉, Mehdi Moussaïd[1,4] & Philipp Lorenz-Spreen [1,5]

As the written word has moved online, new technological affordances and pressures – such as accelerated cycles of production and consumption – have changed how news headlines are produced and selected. Previous literature has linked certain strategies (e.g., clickbait) and linguistic features (e.g., length, negativity) to the success of text online (e.g., clicks). We tracked changes in the prevalence of those features in a sample of ca. 40 million news headlines across the last two decades from English-language outlets worldwide, focusing on the period in which the headline format adapted to the online context. We drew from a broad set of lexical, syntactic and semantic features from the literature to find the signature of the transition to online formats in the journalistic output of the last two decades. Many – but not all – of these features have become more prevalent over time, such as length and negativity. This systematic shift appeared across news outlets from different countries, political leanings, and of different journalistic quality. This may indicate an adaptation to the new affordances and pressures of the digital, online environment, and raises questions for the design of online environments in the future.

[1] Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. [2] Humboldt-Universität zu Berlin, Faculty of Life Sciences, 10099 Berlin, Germany. [3] Max Planck School of Cognition, Leipzig, Germany. [4] School of Collective Intelligence, Mohammed VI Polytechnic University, Rabat, Morocco. [5] Center Synergy of Systems and Center for Scalable Data Analytics and Artificial Intelligence, TUD Dresden University of Technology, Dresden 01069, Germany. ✉email: nickl@mpib-berlin.mpg.de

Introduction

In a crowded marketplace where everybody is shouting, what do vendors call out to attract people's attention? How might they adapt their strategies as this marketplace gets more competitive over time? To answer these questions, we could use an archive that records all the market cries, to study how their language has changed since the first market stalls were set up: A coherent pattern would indicate an adaptation to the changing system. In the present study, the marketplace in question is the online digital environment, the market cries are news headlines attracting attention to news articles, the vendors are content producers (journalists, copy-writers, and editors), the consumers are readers, and the currency is their limited attention.

In the digital age, large amounts of content can be produced quickly and inexpensively, and transmitted and stored with high fidelity (Acerbi 2019). From an economic perspective this represents a contestable market (Munger 2020): because up-front costs are low, entry to this market is potentially open to many more content producers. An abundance of content means increased competition for limited attention, fiercer algorithmic, social and personal selection, and a higher turnover of content (Lorenz-Spreen et al. 2019), a situation aptly described as an *attention economy* (Simon 1971). With an overabundance of content, only a small fraction of content eventually reaches the user, even less is ultimately consumed and engaged with (see González-Bailón et al. 2023, for the funnel of engagement). Increasingly, content is readily created to fit demand and further optimized through audience feedback, exploiting human psychological tendencies – e.g. towards more negative, social, or clickbaity content. As a result, this type of content should become more prevalent over time (Hills 2019).

Previous literature has focused on linguistic features driving consumption (J. M. Stubbersfield 2022; Gligorić et al. 2023; Robertson et al. 2023), sharing and engagement (Goldenberg and Gross, 2020; J. M. Stubbersfield, 2022), the features favoured in production, and the features that catch readers' attention (Kuiken et al. 2017; D. Molina et al. 2021; Scott 2021). If features such as length (Kuiken et al. 2017; Robertson et al. 2023; Gligorić et al. 2023) or negative sentiment (Gligorić et al. 2023; Robertson et al. 2023) are favoured by either producers or consumers, they may become more prevalent over time. For an overview of the findings from the previous literature, see Supplementary Table 1; see Supplementary Fig. 2 for a complete overview of features that we examine in the present study.

In this study, we quantitatively investigate the evolution of online headlines over the last two decades. Though the format of news headlines was established long before the internet, the pressures on headlines have increased dramatically since the medium moved online with 1) single articles, rather than entire newspaper issues, now competing for reader attention, also called the "unbundling of journalism" (Bisceglia 2023), 2) the unprecedented ability to track and quantify success, 3) A/B-testing to optimize wording (Hagar and Diakopoulos 2019) and 4) pay-per-click journalism (Levitte and Davidovitz 2010; Staff 2023).

In contrast to fuzzy cues about popularity and quality of content in offline transmission, online content often comes with explicit metrics (Acerbi 2019). If content is produced to maximize a particular metric, it may be very successful, until the algorithm changes. At a time when the Facebook newsfeed algorithm preferentially recommended content that received a lot of clicks, the soft-news outlet Upworthy coined a particular kind of clickbait style (Munger 2020): long, conversational headlines that strategically induce a curiosity gap (Scott 2021) that readers can only bridge by clicking to open the article (see Table 1). As a result, Upworthy content flooded Facebook newsfeeds in 2013. When Facebook began prioritizing other forms of engagement (e.g., likes, shares, comments, attention minutes etc.) over click-through rate, Upworthy content became less visible and the business faltered (Munger 2020). This example illustrates how fickle online success can be, as artificial success metrics are subject to redesign. Although click-through rate may no longer be the golden metric of online success on social media platforms, inviting clicks may still be a desirable feature for online content. Let us therefore take a closer look at the phenomenon of clickbait.

Besides being designed to attract clicks, clickbait style does not have one defining necessary or sufficient feature. Instead, the literature identifies a range of features typically employed by clickbait outlets and that may make a headline more clickable. From here on, we refer to these features as *clickbait features*, or collectively *as clickbait style*. As Table 1 illustrates, clickbait tends to be characterized by long, conversational-style headlines, which stand in stark contrast to the succinct style of traditional headlines (Chakraborty et al. 2016). Traditional headlines read more like a telegram, with information condensed in as few content words as possible, often lacking the functional words (e.g., pronouns, articles) characteristic of natural speech. They often lack a (main) verb (first and third examples, Table 1), or take the format of a noun phrase (second example, notice that the whole construct functions syntactically as a noun). By clearly presenting the most crucial information, traditional headlines enable the reader to make an informed decision about which content to consume. Clickbait headlines, in contrast, pique readers' interest with a vague promise of surprising or interesting content, but may not even indicate what content to expect, thus creating an information gap (Loewenstein 1994).

One linguistic device that triggers a curiosity gap is *forward reference* (Blom and Hansen 2015; Kuiken et al. 2017; Scott 2021), in which a reference is made to something that will only be specified later in the text. Demonstratives ("You Shouldn't Be Able To Pay For **This**"), pronouns without antecedent or clear referent ("**She** Didn't Believe **It** When **She** Saw **It**"), adverbs, and wh-words ("**Here's Why** …") can constitute a forward reference, thereby introducing a curiosity gap that readers can only close by clicking the headline (Scott 2021). A comparative corpus analysis of clickbait outlets (amongst which, Upworthy) and high-quality journalistic outlets found that demonstratives (these, this, that, those) and personal pronouns (I, you, he, she, it etc.) were much more common in the clickbait headlines (Scott 2021). Analyzing

Table 1 traditional headlines compared to typical clickbait-style headlines.

Outlet	Example headline
The Guardian	*Shell's future in Nigeria in doubt*
The New York Times	*Second Thoughts on Lighter Sentences for Drug Smugglers*
The Times of India	*Ahmedabad firm to run ILO pilot project*
Upworthy	*I'm No Supreme Court Expert, But I Kinda Think You Shouldn't Be Able To Pay For This?*
	A Group Of Women Had A Hypothesis About Sexism. Then They Made A Bunch Of Pies To Prove It.
	A Librarian Wouldn't Let This Kid Check Out Books Because Of His Skin Color. That Backfired For Her.

the Upworthy Research Archive, a dataset of the outlet's A/B tests, Gligorić et al. found links between consumption (click-through rate) and the use of personal pronouns, negativity, and headline length (Gligorić et al. 2023). Using the same Upworthy corpus, Robertson et al. found that negative sentiment and length increased click-through rate, while linguistic complexity and positive sentiment decreased it (Robertson et al. 2023). For a list of linguistic features studied in the previous literature, see Supplementary Table 1. Based on this literature, we extracted a broad range of linguistic features, including formal features (e.g., length), lexical and syntactic features, and sentiment. We operationalized the phenomenon of forward reference by looking at pronouns, determinatives, wh-words and specific substrings (e.g. "here's why") in individual headlines.

The present study tracks the prevalence of these features in online news headlines over time. We investigate which features become more frequent in the production output as news outlets increasingly move online. To conceptualize this process, we use the framework of cultural evolution – with the linguistic features corresponding to more or less adaptive cultural traits (Boyd and Richerson 1988; O'Brien et al. 2010; Mesoudi 2011; Acerbi 2019). The popularity in production of a linguistic feature and its performance in relevant consumption metrics are, presumably, linked: Journalists and copy-writers may consciously establish a link between a certain feature and its online traction (e.g., by tracking performance in A/B tests), and headlines that receive more online visibility are also more likely to become salient models for future headlines. This process of cultural selection requires no insight into its mechanisms on behalf of the people involved: Even if neither producers or consumers understand which features are effective and why, an adaptive feature may become prevalent in the population of traits (Henrich 2016). We suspect that features that bestow a selection advantage at any stage of production, diffusion or consumption, will become more prevalent in the corpus over time, replacing less adaptive variants. Therefore, our main research question is: How does the frequency of linguistic features linked to clicks change over time, across outlets?

To answer this question, we made use of several large corpora of digital headlines from news outlets from around the world. We analyzed four datasets spanning the last two decades: *The New York Times* (U.S.), *The Guardian* (U.K.), *The Times of India*, and headlines from the broadcaster *ABC News Australia* (AU), which we here refer to collectively as BIG4. We also analyzed the News on the Web (NOW) corpus, which contains ca. 30 million headlines from many more outlets and countries. We used a range of Natural Language Processing methods, including dictionary-based methods, sentiment analysis, and constituency parsing, to measure linguistic features at scale across all ca. 40 Million headlines in our sample. Our results provide a comprehensive documentation of the temporal developments of those features, showing a universal trend of increasing prevalence of several linguistic features that are associated with clickbait style or higher click-through rates, such as length, negativity, the occurrence of certain word types (e.g., wh-words, pronouns, verbs), and a shift away from traditional headline formats, indicating an overall stylistic shift.

Methods

Corpora. Several datasets contribute to our collection of headlines: BIG4 combines headlines from four representative news outlets across two decades (thus covering the early years of digital journalism), and the News on the Web corpus (NOW) contains a broader range of news websites, but starting later, in 2010. We also included two reference corpora for comparison: one containing clickbait headlines and another containing scientific preprint titles.

The BIG4 corpus covers a longer range that captures the transition from print to hybrid/online journalism. It includes headlines from *The New York Times* and *The Guardian* (both collected using the official API), and two freely available datasets with headlines from *The Times of India* and *ABC News Australia*, summing to ca. 9 million headlines. The *New York Times* API contains both print and digital-only content, and includes publications from as far back as 1851. In this study, we focus on the time range from 2000 onwards in order to cover the transition of the medium to online environments. While there is little documentation for the Times of India and ABC Australia datasets, we verified that searching for early article headlines brings up a digital version of the article, meaning that these articles also appeared online. The ABC Australia corpus raises some questions: For early years (e.g., 2003), some headlines could not be found online, and for the later years (e.g., 2020), the headline corresponded to a URL, but with a different headline for the same article. For years in between (e.g., 2010, 2015), the headlines match the URL and the current titles of that article. This could reflect editorial curation practices, such as selecting a better-performing headline later on, but also raises doubts as to the quality of this particular dataset. As the dataset seems to be preprocessed (lacking punctuation and being all lower-cased), it is possible that the headlines in this corpus were actually extracted directly from the URL, which would be problematic if URLs do not accurately reflect the headlines. With these caveats in mind, and presenting the disaggregated descriptive results, we decided to include this corpus in our analyses.

We also purchased the Corpus of News on the Web (NOW) – a curated collection of English-language news text ranging from 2010 to the present, sourced from news websites across 20 countries where English is used (Davies 2016) and summing to 30 million headlines. Daily, new URLs are retrieved via Google News, and the respective articles are added to the corpus. This sampling strategy is supposed to capture the current discourse across English-speaking media. In contrast to the corpora obtained through the Guardian and *New York Times* APIs, the NOW corpus does not offer a complete record of the production of any specific outlet. Perhaps due to the sampling strategy, the corpus composition changes over time: Some outlets are heavily sampled from in a specific year, and not at all in another. There is also an issue with the data quality, as for some articles, only part of the headline was scraped. These cases are obvious, as these headlines end in '…'. We therefore carried out our analysis on the entire, unfiltered NOW corpus, as well as on various subsets of the NOW corpus in order to ensure robustness of our findings. The figures in this paper report a cleaned version of the NOW corpus with two filtering steps: Incomplete headlines were removed, then the four outlets for which we already had separate corpora were removed. Note that *The Times of India* is the most frequent source in the NOW corpus and seems to be relatively oversampled. For a robustness check, we include figures in the Supplementary Information for the unfiltered NOW corpus (Supplementary Fig. 1). We also applied an even stricter criterion, including only outlets that have consistently contributed at least 500 headlines per year since 2010, to ensure continuity in our data and to counteract artifacts from the sampling strategy (see Supplementary Table 6 for a list of these outlets, and Supplementary Table 2 for more information about the full set and two subsets).

Besides serving as robustness checks, the full set and two subsets represent different angles to the question of how headlines have changed: Every day, new articles are added to the NOW corpus based on the Google News API. In its entirety,

the NOW corpus may paint a representative picture of what the news sound like at any given point in time. On the other hand, this aggregate view obscures the fact that different outlets contribute to this picture. Filtering the NOW down to specific outlets that consistently contributed to this data, we can get a sense of how particular outlets have employed different language over time, providing a view analogous to the BIG4 dataset.

For benchmark comparisons, we included a clickbait-style corpus (Matias et al. 2021) and a corpus of 2,276,611 scientific preprint titles in STEM fields (arXiv dataset). For more information about these corpora, see Supplementary Table 2. The Upworthy dataset is an excellent benchmark for the clickbait features, as Upworthy is commonly considered the prime example of clickbait style (Chakraborty et al. 2016; Munger 2020; Scott 2021). Since this relatively small corpus spans a short time frame (2013–2015), we use it as a static benchmark, without the time dimension.

Natural language processing. We used a Python pipeline to clean, tokenize, part-of-speech-tag, and analyze the headlines in terms of our selected features (see Supplementary Fig. 7 for an overview of the pipeline). For sentiment analysis, we used the Flair package (Akbik et al. 2018) to classify headlines as positive or negative. Flair returns a label (negative or positive) along with a score for each headline. We set a relatively high threshold (0.9), for choosing the label suggested by flair; we labeled headlines with scores of 0.9 or lower as "neutral".

We also captured the syntactic structure of the headline as an important aspect of style. For this, we used a constituency parser using Python libraries spaCy (Honnibal et al. 2020) and benepar (Kitaev and Klein 2018; Kitaev et al. 2019), which constructs a hierarchical representation of the sentence with labeled parts (constituents). We focused on the top-most label, which captures whether the entire headline is a sentence (S), noun phrase (NP), etc. For a visualization of constituency structures, see Supplementary Fig. 5; for a flowchart detailing the preprocessing and analysis steps, see Supplementary Fig. 7.

Statistical analysis. To quantify the relationship between the individual features across time, we performed linear regressions for the continuous features (e.g., number of words), and logistic regressions for the binary features (e.g., occurrence of a wh-word) with year as the only predictor variable. Due to the prohibitive running times, we abandoned a Bayesian approach, instead running regression models using the lme4 package in R (Bates et al. 2015). While chosen for feasibility, we think a frequentist approach is appropriate given the several million data points at our disposal.

The rise of headline length over time was among the most robust and salient trends we observed. At the same time, we found a clear link between length of the headline and other linguistic features (Fig. 5). This raised the question of whether the increase in other features was just a byproduct of increased headline length. This would be the case if length was the driver of these other features. If so, rather than just regressing features on year as a sole predictor, controlling for the effect of headline length would disentangle their contributions. This reasoning suggests a mediator analysis of the effect of year on linguistic variables mediated by length. Although this is a useful and popular analysis tool, it would be misleading in this context because it would imply a specific causal structure behind the data-generating process. This is problematic because we know this causal structure does not apply and our variables do not lend themselves to causal inference.

Firstly, the variable year is useful to describe a trend over time, but time itself is not a cause. Instead, it contains several causal factors, such as new technologies, sinking production costs and accelerating dynamics of online attention. This already hinders causal inference with the variables at our disposal. Secondly, a mediator analysis would assume that year increases length, which in turn causes certain linguistic features to appear with some (generally higher) frequency (total effect of year on features), while there may be a (probably attenuated) direct effect of year on feature. For causal inference to be possible, the flow of causation would have to be unidirectional between these variables. However, the causal structure in question is more complex than the mediator diagram suggests. For instance, there is mutual causation between length and a linguistic feature: While each additional word provides an opportunity for the inclusion of a specific word or feature (length causes feature), each feature in turn affects the length (feature causes length). Unobserved variables, such as style may also be at work: Journalists and editors do not produce one word after another, which sometimes result in full sentences or noun phrases, clickbaity or dry headlines. Instead, they make holistic decisions on the headline level with consequences for the word count and features used. Because holistic features such as style are difficult to capture, we limited ourselves to tracking the frequency of measurable linguistic features over time.

With these assumptions about the underlying process, we concluded that length would be a bad control, resulting in an uninterpretable and misleading regression (Achen 2005; Rohrer 2018); see Cinelli et al. (2022) for an introduction to Directed Acyclic Graphs. Instead, we used regressions as a means to quantify how features changed over time, without implying causality. To address the relationship between length and features, we include descriptive plots of feature occurrence as a function of number of words in Fig. 5.

Classifying outlets by political leaning and journalistic quality. To classify outlets by political leaning, we relied on the AllSides Media Bias Chart (AllSides Media Bias Chart 2019). Based on online, political U.S. content, this chart categorizes outlets into five categories (left, lean-left, centre, lean-right, right), which we collapsed into the three categories left-leaning, centre, and right-leaning. According to this classification, The New York Times is left-leaning, The Washington Times is right-leaning, and BBC news is a centre outlet.

To classify outlets by journalistic quality, we relied on the Ad Fontes Media Bias Chart (Application Version 2.7.2) (Interactive Media Bias Chart 2024). This chart has two axes: political leaning (left to right on the X axis, in numeric values from −42 to 42, without discrete categories) and journalistic quality (from less reliable to more reliable, on the Y axis, in numeric values from 0 to 64, with demarcations into red, orange, yellow, green). To avoid confounding with political leaning, we focused on the narrow middle section of this dimension (−10 and 10). For journalistic quality, the Ad Fontes Media Bias Chart offers four sections: red (0–16), orange (16–24), yellow (24–40) and green (40–64). We use these cutoffs in our analysis, taking all outlets that fall within the green area as quality journalism, and everything in the yellow area as of lower journalistic quality. According to this partition, the former category includes The Guardian, The New York Times, The LA Times, Forbes, The Wall Street Journal, and Al Jazeera, while the second category includes outlets such as The Mirror, The Daily Mail, Upworthy and The New York Post (for the full list, see Supplementary Table 3). Note that we did not include outlets from the orange and red sections of lowest journalistic quality.

We make all the code for the above analyses available, and provide information about the datasets used in the Data Availability statement.

Results

Overall rise in linguistic features associated with clickbait style or higher click-through rate. For descriptive results, we measured the prevalence of all linguistic features (as listed in Supplementary Fig. 2) across the BIG4 and NOW corpora. For an exemplary overview of the most important categories of linguistic features, we selected eight features: the number of words (length), the occurrence of at least one verb, occurrence of a pronoun, occurrence of a wh-word, the syntactic property of whether the headline constitutes a noun phrase or a full sentence (as these are the most frequent syntactic formats (see Supplementary Table 5), and overall negative or positive sentiment.

Figure 1 shows average values and relative frequencies (for binary occurrence measures) of this selected set of features over time: In each panel, the left side shows the trends for each outlet in the BIG4 corpus as well as their overall average, and the right side shows the average over almost 20 million headlines of the filtered NOW corpus (for the same analyses on the whole NOW corpus and a more strictly filtered subset, see Methods section and Supplementary Fig. 1). The orange and black dotted lines serve as contrasting benchmarks, representing the frequency of features in a dataset of Upworthy headlines (as typical clickbait corpus; orange) and of features in a dataset of titles of scientific preprints in STEM fields on arXiv (as a contrasting format; black). While similar in form and function, the upworthy and arXiv benchmarks illustrate starkly contrasting characteristics, marking extreme points on specific features. For example, the occurrence of pronouns or wh-words is characteristic of clickbait headlines, while these features are virtually absent in the arXiv corpus. On many features, the news headlines shift from being more similar to the arXiv corpus to being more similar to the upworthy corpus. Taken together, these trends are indicative of an overall stylistic shift: There is a clear rising trend in the frequency of multiple features that are linked to clickbait style or a higher click-through rate – many of which (e.g., length, verbs, pronouns, and wh-words; Fig. 1a–d), are especially prevalent in the clickbait corpus we use as a benchmark. Some of these features, such as the occurrence of a pronoun or a wh-word, are virtually absent from the arXiv benchmark, while they have increased in news headlines. These trends are consistently present in both the BIG4 and NOW corpora.

There are also changes in the preferred syntactic format. The noun-phrase headline is the dominant format of scientific preprint titles (arXiv corpus), while it is an uncommon format for clickbait headlines (upworthy corpus). While relatively common as a headline format in earlier years (in 2000, noun phrases made up roughly half of all headlines by The New York Times), it became generally less popular in both the BIG4 and NOW corpora, over time (Fig. 1e). In some outlets, the noun-phrase headline format may have been replaced in part by the full-sentence headline format (Fig. 1f). In the NOW corpus, the use of full-sentence headlines increased until 2018, then decreased. This pattern could be an artefact induced by the sampling strategy of the NOW corpus, reflecting a change in its composition over time, rather than a trend in outlets changing their formats twice in this regard. If we include only outlets that were consistently included in the NOW corpus, we again see a clear rising trend for full-sentence headlines (see Supplementary Fig. 1).

In the Big4 and NOW corpora, news headlines seem to have consistently gotten more negative and less positive over time, whereas the arXiv corpus exhibits the opposite pattern, indicating a news-specific development. For a comprehensive overview of the trends for all features we measured, see Supplementary Fig. 2. In sum, we found converging evidence across different corpora that the language of news headlines has changed consistently over time.

Figure 2 summarizes the results of linear and logistic regression models that we employed to further quantify the relationships between publication year and the occurrence of linguistic features in a given headline. We performed a standalone regression for each variable on the predictor publication year of the format: linguistic_feature ~ 1 + publication year (centred). Since we could not assume unilateral causation between variables, we could not include meaningful controls (see Methods section, "Statistical analysis," for a discussion of this problem). Using only publication year as a predictor variable allowed us to quantify the relationship between publication year and each feature without making causal assumptions. The results in Fig. 2 show that in both BIG4 and NOW, the number of words is consistently positively associated with later years of publication, thereby confirming the trend observed in Fig. 1 of headlines growing in length. The regression results on the binary measures of linguistic features also confirm the trends shown in Fig. 1: The occurrence of wh-words, verbs and pronouns are all positively associated with later years of publication.

The noun-phrase headline is negatively associated with publication year across the BIG4 and NOW corpora. The full-sentence headline is positively associated with later publication years in the BIG4 corpus, but not the NOW corpus. This may reflect a change in NOW's composition over time due to its sampling strategy. Focusing only on outlets that are sampled consistently in the NOW corpus, we found a clear increase in full-sentence headlines (Supplementary Fig. 1). Headlines containing a full sentence have become more frequent in both BIG4 and NOW, and in the NOW corpus, especially the noun-phrase-plus-full-sentence headline (Supplementary Fig. 2). The regression confirmed a positive association between headline negativity and publication year, and conversely a negative association between positivity and publication year consistently across both corpora. These regressions serve to quantify the trends already outlined above. For a discussion of how the data limited causal investigation and interpretation, see the Methods section "Statistical analysis."

After establishing that various features associated with clickbait style or a higher click-through rate have increased over time across the two large corpora, we made use of further aspects of those datasets to explore whether these developments only hold for specific outlets or types of content. We also explored the relationship between the length of a headline and the occurrence of linguistic features.

Overall trends hold across different types of outlets. One advantage of the NOW corpus is that it comprises a plethora of news outlets, allowing us to explore whether the trends we observed were driven by outlets of one specific political leaning or by journalistic quality (e.g., a sensationalist tabloid versus an outlet known for more sober, fact-based reporting). Figure 3 compares specific subsets of outlets from within the NOW corpus, which we split along the dimensions political leaning and journalistic quality. We used the AllSides Media Bias Chart (Version 9.2) to classify the outlet's political leaning (AllSides Media Bias Chart 2019). To classify outlets according to journalistic quality, we relied on the Ad Fontes Media Bias chart (Application Version 2.7.2) (Interactive Media Bias Chart 2024). Overall, the trends were similar between outlets with different

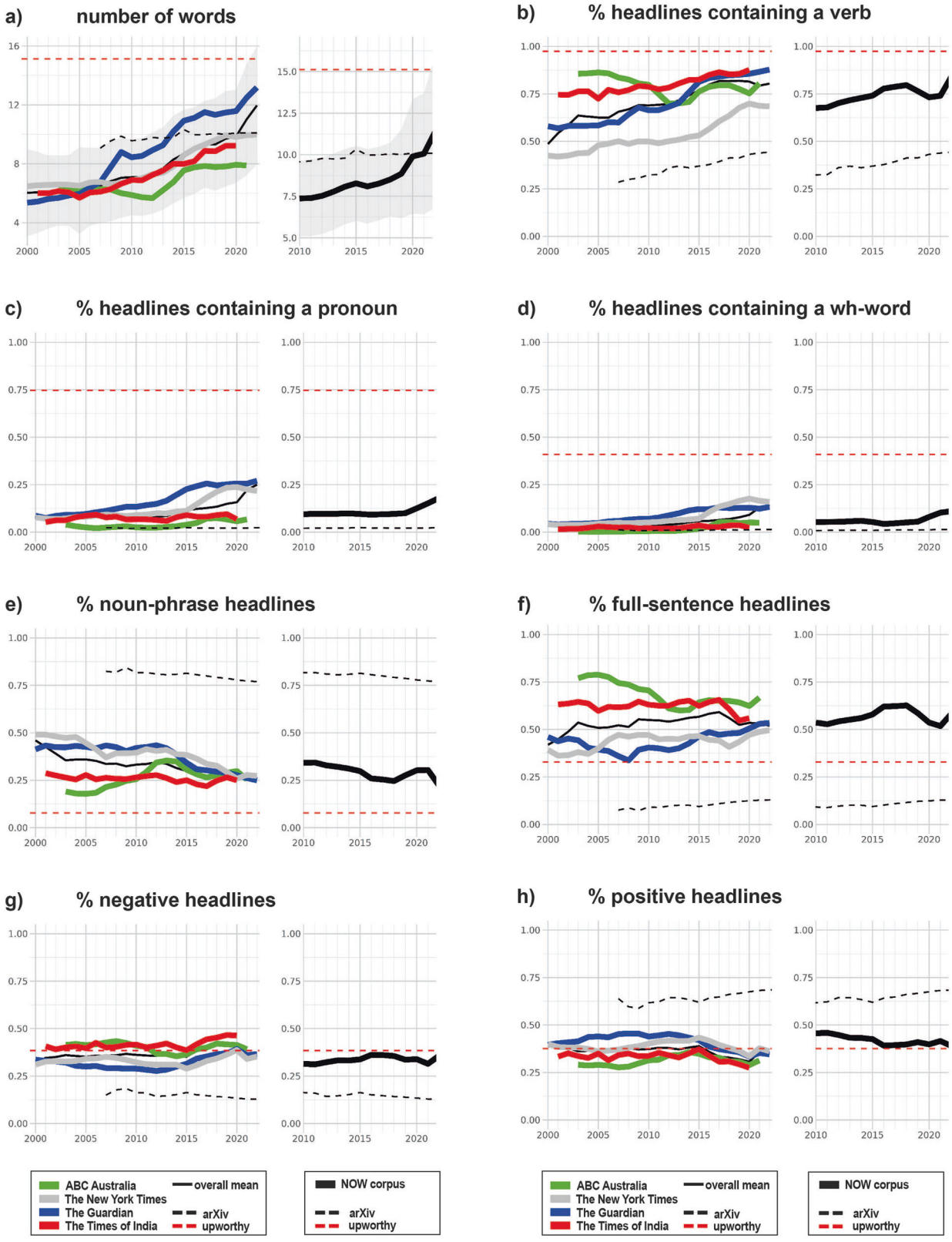

Fig. 1 Changing frequency of linguistic features over time in four major news outlets (*The New York Times, The Guardian, The Times of India*, and *ABC Australia*) over the last two decades, and in the News On the Web (NOW) corpus, starting in 2010. a Mean number of words per headline. Shaded area indicates 1 SD from the mean across outlets (thin black line). **b**–**h** Proportion of headlines exhibiting a specific feature. Thin dashed lines represent two benchmark corpora: clickbait headlines (orange) and academic preprint titles (black).

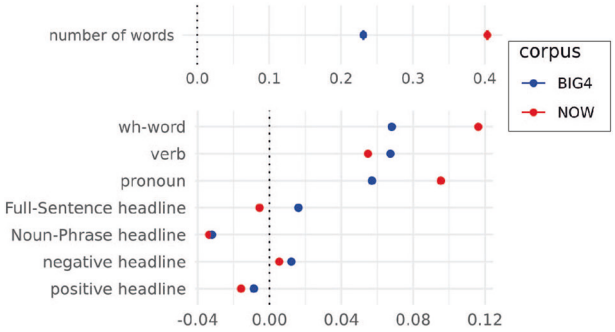

Fig. 2 Estimated coefficients of the predictor variable "publication year" (centred) for different linguistic features. These are the result of a linear regression model for the variable "number of words" and of logistic regression models for the remaining, binary variables. For each year, headlines become longer (by approximately one fourth of a word); for the occurrence of a wh-word, verb, pronoun, and negative headlines, the coefficients are consistently positive across corpora, whereas the coefficients are negative for noun-phrase headlines and positive headlines. For full-sentence headlines, the coefficient signs differ between data sets, although this may be due to a sampling strategy for the NOW corpus. For the overview of all features with their coefficients see Supplementary Fig. 2.

political leanings and of different journalistic quality, and mirrored the trends outlined in Fig. 1. However, we also observed some differential trends: For example, full-sentence headlines and negative headlines became more frequent in right-leaning outlets, and less so in centre and left-leaning outlets. Perhaps surprisingly, there was little difference between news sites of high and lower journalistic quality, pointing to a general development that may be specifically picked up by right-leaning outlets rather than by high or lower quality journalism.

Overall trends hold across different topics. To investigate the possibility that the trends we observed can be attributed to a shift in topics across time within instead of between outlets, we examined the information about individual sections (e.g., world news, U.S. news, sports) provided by the *New York Times* corpus (see Supplementary Table 4 for the full list and frequencies of these sections). We again found evidence for the same general trends shared across different topics (Fig. 4): A rise in headline length, an increase in the occurrence of a verb, pronoun, or wh-word in headlines, a steep decline in the use of the noun-phrase format in favour of a full-sentence headline, and an overall increase in negativity.

This corpus also marks the type of material for each headline, distinguishing, for example, news from op-eds and paid death notices. "News" is by far the most frequent type of material (Supplementary Table 4). It is thus unsurprising that this category follows the overall trends outlined above (Supplementary Fig. 3), although results are less clear for noun-phrase versus full-sentence headlines.

The *New York Times* corpus also specifies the length of the headline in relation to the length of the entire article, thereby allowing for an exploratory analysis. The number of words in both headlines and articles increased over time (Supplementary Fig. 4). While longer articles may be surprising considering the increased competition for limited attention, this trend is consistent with sunk production costs changing journalistic practices online by allowing more words to be put both in headlines and articles. Longer articles also provide more space for ads, a financial consideration that may counteract the pressure of consumer preferences.

There is no trivial link between headline length and the linguistic features studied. Given the observation of increasingly long headlines, we aimed to better understand the relationship between the length of headlines and the linguistic features we studied. There is a positive correlation between the relative frequency of occurrence of many of our features and the number of words in a headline. This relationship, however, is moderated dramatically by genre, as shown in Fig. 5: In an Upworthy headline, even very short headlines are likely to contain a verb, or to a lesser extent, a pronoun or wh-word, while the title of a scientific preprint in a STEM field can get quite long without containing any of these features (Fig. 5a–c). In the nonlinear relationship between length and these clickbait features, the news headline corpora more closely resemble the Upworthy corpus.

Since most of our features record the occurrence of a word (type), their probability increases almost cumulatively with length. However, the relationship is more complex than a direct increase of feature occurrence according to length, and it is not at all clear that length is the causal driver of such an increase. The relationship could easily be reversed: The use of linguistic features may drive the length of the headline (see Methods section for a description of the data-generating process).

The relationship between syntactic formats (Fig. 5d, e) is more nuanced: For example, the more words a headline or title consists of, the less likely it is to be a noun phrase. This may indicate that overall, noun phrases become less readable with increasing length. At the same time, the arXiv dataset illustrates that long noun phrases are a theoretically viable title format (Fig. 5d). Full-sentence headlines tend to be longer than noun-phrase headlines, but they do not become more likely as length increases. Again, there might be an optimum length for the full-sentence format, before a journalist decides to split what might be a single-sentence headline into a two-sentence or compound headline. Such decisions will in turn have consequences for the language, and features, used. As Fig. 5 shows, length is clearly intimately connected with other features, but importantly, this relationship does not hold universally for all English texts: Clickbait headlines and scientific preprint titles have different profiles, with news headlines falling somewhere in between, but closer to the clickbait headlines. Sentiment (Fig. 5f, g), however, does not seem to depend on headline length. Overall, the trends we observed cannot be explained by a general or linear dependence of linguistic features on number of words used.

Discussion

Starting with a list of features that are associated with clickbait style or higher click-through rate, we tracked the prevalence of these features in the production output of online news outlets across the last two decades. On many features, the news headlines shift to being more similar to clickbait headlines (and less like the titles found in scientific preprints). Taken together, these trends are indicative of an overall stylistic shift. Our findings point to a general increase of features in online news headlines that are associated with clickbait style or a higher click-through rate: longer headlines, a shift away from the dry noun phrase – possibly giving way to a variety of syntactic alternatives – and growing negative sentiment. This systematic shift at different levels of analysis – lexical, syntactic, and semantic – applies across outlets based in different countries, with different political leanings and of different journalistic quality.

Our observational data records the frequency of linguistic features in concurrently produced headlines over time. Several mechanisms interact to produce these data, involving the producers and consumers of headlines, as well as their environment with its unique pressures and affordances. We can only speculate

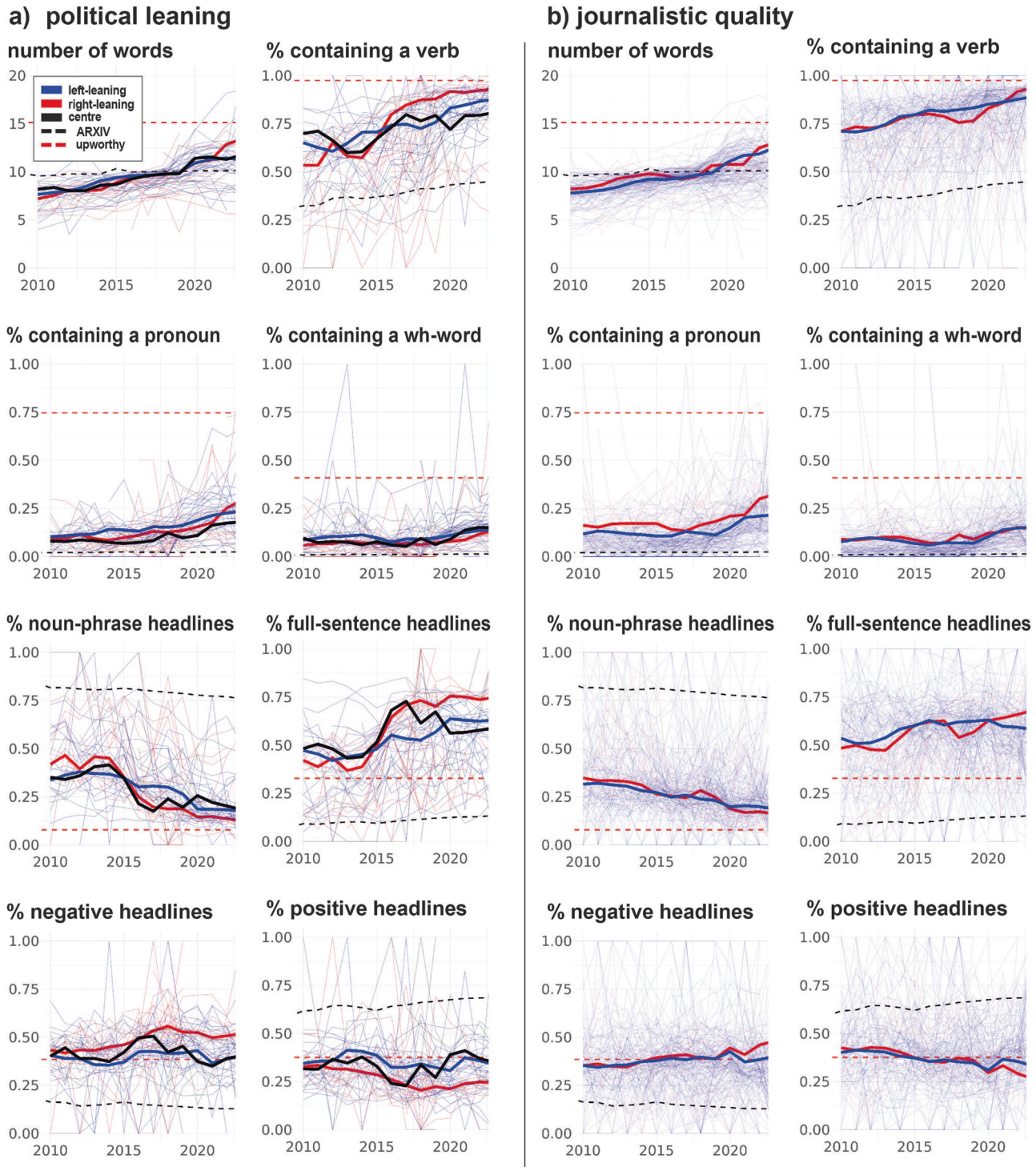

Fig. 3 Linguistic trends in outlets of different political leanings and of different journalistic quality. The development mirrors the general trends observed in Fig. 1, across outlets of different political leaning (**a**) and different journalistic quality (**b**). The dashed lines represent two benchmark corpora: clickbait headlines (orange) and academic preprint titles (black).

about the underlying cognitive, psychological and technological processes.

In the framework of cultural evolution, a rise in the frequency of any cultural trait may be due to cultural drift – i.e., random copying – alone. But given that we investigated features that are thought to make a headline more successful online, and that we observe a consistent pattern of increase in these features during a period of broad adoption of online platforms for news consumption, it is reasonable to assume that our observations track adaptations to online environments. These changes may be driven by cognitive, social, or technological processes: the producers, consumers, affordances and pressures of the digital online environment, and their interaction.

An intuitive explanation of this trend would be that production follows demand (Munger 2024): Although readers may not be aware of their preferences, they may reliably choose to click or engage with

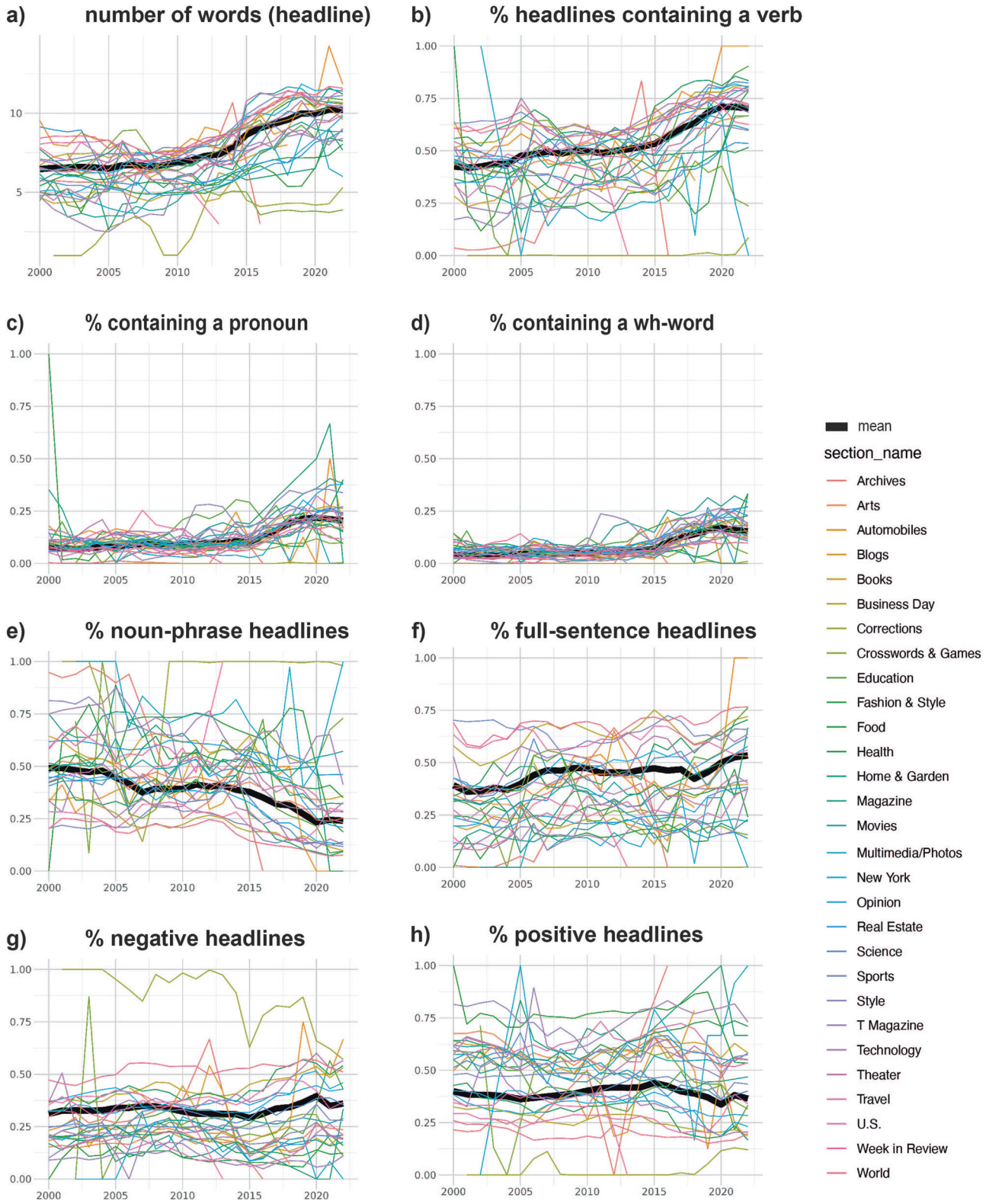

Fig. 4 Headlines in the *New York Times* corpus according to their section (e.g., U.S. news, world news, sports). The bold black line indicates the overall mean. While longer headlines, the occurrence of a verb, pronoun, wh-word, the full-sentence and negative headlines became more prevalent over time (**a**–**d**, **f**, **g**), noun-phrase headlines and positive headlines seem to have become less popular (**e**, **h**).

content that grabs their attention, piques their curiosity, and promises somehow attractive (e.g., negative) information. Such *cognitive* tendencies interact with elements of the online environment, such as recommendation algorithms. Algorithms that track different success metrics may result in different content being successful, depending, for instance, on whether it optimizes clicks or engagement – prime examples of the impact that machines have on the evolution of culture (Brinkmann et al. 2023).

The producers of the headlines have an economic incentive to attract readers' attention – whether they are freelance journalists

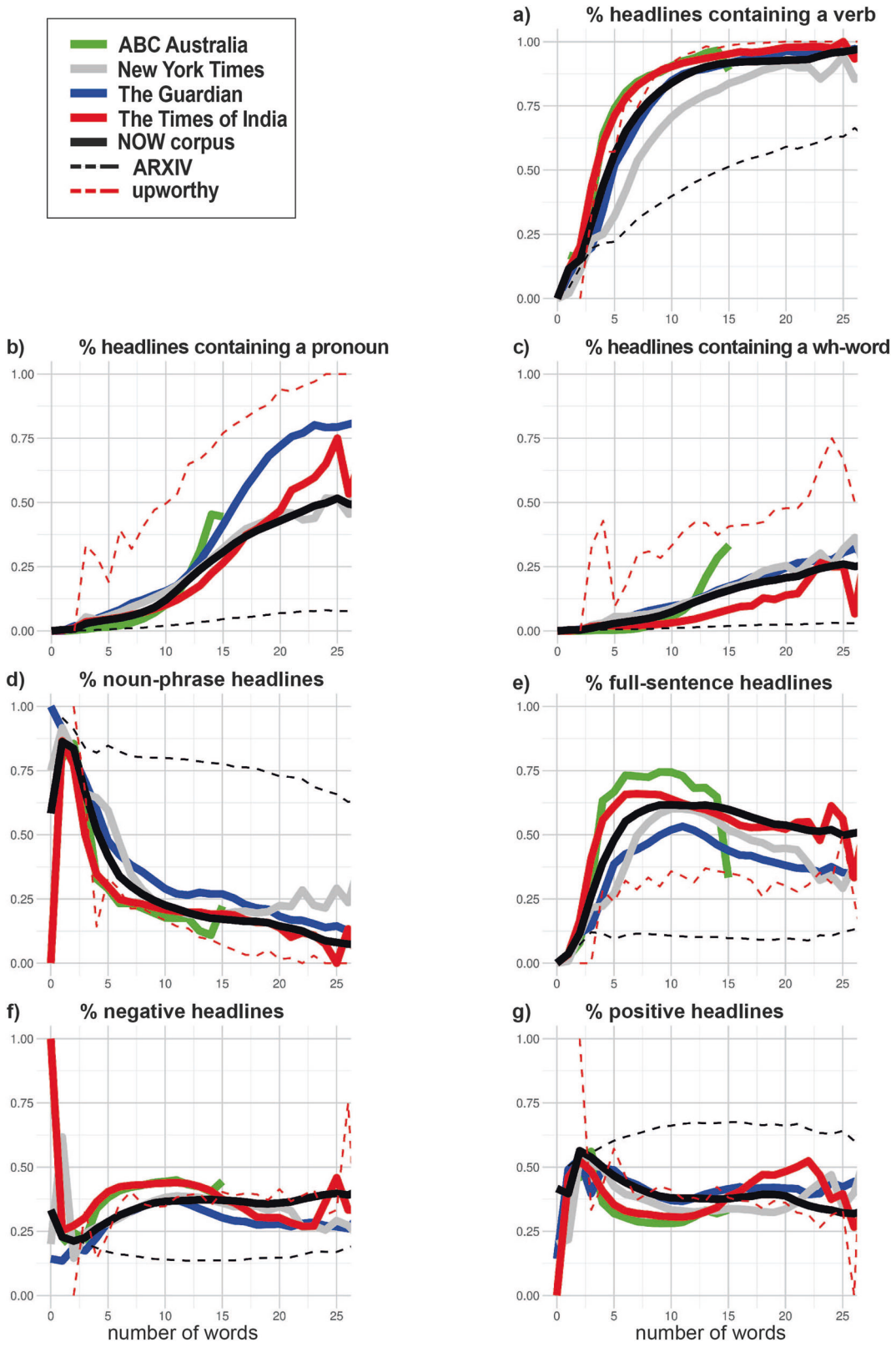

Fig. 5 Select linguistic features as a function of length (number of words). While the probability of several features increases as a function of length, this relationship is strongly modulated by corpus type: The increase is generally steeper for the clickbait corpus, and more attenuated for the scientific titles, with the news falling somewhere in between. This indicates that including certain words (such as verbs, pronouns, wh-words; (**a**–**c**), or choosing specific syntactic formats (**d**, **e**) is more a stylistic choice, rather than a linguistic necessity given a certain word count.

getting paid per click, or copy-writers and established editors trying to secure traffic, and therefore ad revenue, to their website. Techniques like A/B testing allow producers to try several variants of a headline to identify the most successful (e.g., most clicked) one. This technology is used by Upworthy (Matias et al. 2021) and *The New York Times* alike (Bulik 2016). Also in this scenario, it is ultimately the users who choose the most appealing headline for an article.

Headline producers nevertheless bring their own limitations, preferences, and biases to the production process. One study found that journalists tasked with rewriting headlines for an online audience increased headline length, even though it did not affect click-through rate (Kuiken et al. 2017). Furthermore, a journalist with ethical scruples may be more limited in their choice of headline than one who is less concerned about the truthfulness of content (J. Stubbersfield et al. 2018), possibly leading them to actively resist consumer preferences. At the same time, headline writers are social learners who are exposed to, influenced by, and competing with the productions of other writers, potentially giving rise to more complex *social* dynamics.

Finally, the digital medium itself has drastically changed the production process: Content can now be produced quickly, at low cost, and with little technical training (Acerbi 2019; Munger 2020). If the high costs of print production formerly imposed a pressure on journalists to express themselves succinctly, the digital format may have lifted this constraint. People write succinctly when messages are expensive (e.g., telegrams) or cumbersome to type (e.g., SMS texting), but may turn to more natural language when such costs are lifted (e.g., instant messaging). Perhaps new digital technologies, and the era of the Web 2.0, also known as the social Web, have facilitated the rise of an informal written language (McCulloch 2019). On this account, the patterns we see reflect new *technological* affordances of the digital medium: A two-sentence clickbait headline may just not have been feasible in print media.

The consistent finding that headlines became longer with the move online is, however, not trivial. Nor was it expected, as brevity has clear advantages: In language evolution, shorter words are more likely to survive (Li et al. 2024). In fact, an inverse trend has been described for book titles: titles of English novels became shorter throughout the 18th and 19th centuries (Moretti 2009). "The market expands, and titles contract", (p. 141): With more and more novels being written, the titles shifted from long descriptions to intriguingly short titles – often just a proper name (e.g., Austen's "Emma"). Instead of providing a plot summary, titles began to serve as catchy ads. This discrepancy between book titles and online headlines may be explained in terms of the importance of memory in offline evolution: A book title may need to be memorable, for instance, to be brought up in conversation. Online headlines, on the other hand, may optimize for short-term attention rather than long-term memory.

Our descriptive analysis has limitations. While we map the trajectory of individual features across time, there are probably functional dependencies between individual features. We specifically investigated the relationship of headline length to the other linguistic features. On one hand, headline length stands in an enabling relationship to many of these: More words per headline may simply offer more opportunities for pronouns or wh-words to appear. On the other hand, a stylistic choice (e.g., formulating a full-sentence headline) carries simultaneous choices about which words to include (e.g., a verb) and the length of a headline. We clearly saw that there is no trivial link between headline length and our features, as many of them are virtually absent in the arXiv titles, no matter their length. News headlines became more like the clickbait outlet and less like the scientific titles, although these also changed following an attenuated pattern (for changes in scientific paper titles see also Hyland and Zou 2022; Jiang and Hyland 2023). Our stylistic analysis could fruitfully be augmented using large language models: Features that would require manual annotation, such as detecting forward reference or telegraphic speech could be delegated to large language models. We ran a very preliminary analysis that yielded poor inter-rater reliability, but as the technology and methods mature and best practices are established, they will likely be better able to answer more nuanced questions about linguistic strategies and styles at scale.

We lack the measurements to empirically confirm possible mechanisms that explain our results. For example, the increase in length may be due to cognitive, social, or technological pressures: Previous literature has linked length to clickbait style, and descriptively to click success, targeting or reflecting consumer *cognition* respectively. At the same time, it may constitute a *social* trend among content producers, perhaps reflecting an overall style shift in written language. Lastly, it may reflect a *technological* affordance of the digital format. While these different explanations are not mutually exclusive, their contributions could be investigated with more granular data, causal models, and controlled experiments. For example, our predictor variable "publication year" is not a mechanistically meaningful variable like competition pressure, or production cost. Experiments could shed light on the mechanistic structure of the content-generating process by controlling aspects of the environment, such as varying economic incentives for participants to produce successful headlines or imposing certain constraints on title length. Knowing more about specific platforms and outlets and their business models, as well as the incentives of producers and readers, may inform causal models (e.g., agent-based models) of the process.

Covering the transition of the news format to the relatively new online environment, this study aimed to find its signature in the journalistic output of the last two decades. If the trends we observe are undesirable (e.g., due to detrimental effects on autonomy or on fostering an informed citizenry), it would be worth questioning the environment, and the pressures its architecture imposes. Unlike possibly stable psychological tendencies, this highly artificial environment can be redesigned. Although abrupt changes in online success metrics may prevent content from continuously adapting to the online environment (cf. Mesoudi and Thornton 2018, for cumulative cultural evolution), the arbitrariness of online success metrics presents an opportunity to design sustainable, meaningful online metrics that benefit society and the individual: e.g., the diversity of readership may indicate higher quality (Bhadani et al. 2022), or algorithms may act as bridging systems (Ovadya and Thorburn 2023). *The Guardian* now displays a "Deeply read" list next to its "Most viewed" list on its homepage, offering readers an additional criterion for deciding where to spend their time online (What is the 'deeply read' list? 2024). Going a step further, users themselves could be empowered to select the metrics that matter to them (Lorenz-Spreen et al. 2020).

The present analysis also has implications for the style-based detection of misinformation: Previous research has linked many features that we examined to mis- or disinformation, such as increased negativity, length, use of pronouns and differential use of punctuation, with length appearing to be the most important cue (Lebernegg et al. 2024). If such linguistic "fingerprints of misinformation" (Carrasco-Farré 2022) exist (in this case, sentiment, morality and readability; for the trends of readability in the present corpora, see Supplementary Fig. 2), they would present a very resource-effective guide to flag content even before it would need to be laboriously fact-checked. Our findings show that many of these features have become more prevalent in news headlines overall, also in established, traditional news outlets, meaning that

their diagnostic value may have diminished with the adaptation to online pressures. Going forward, the validity of such linguistic cues may need to be re-assessed against overall systematic temporal trends and monitored continuously.

In the age of low-cost production of digital content, content can be engineered to exploit consumer preferences with data-driven precision, while social platforms have rewritten the laws of its distribution. This study quantifies the impacts of these technological developments on the language being used in English-language headlines worldwide over the last 20 years. With this study, we hope to contribute to the understanding of how cognitive, social, and technological factors interact in the process of content production and the wider online information environment, as insights into these mechanisms are crucial for the democratic co-design of online environments.

Data availability

The code for the preprocessing, linguistic and statistical analysis, and individual subplots is available at https://github.com/pietronickl/EvolutionOfNewsHeadlines. The license we purchased for the NOW corpus cannot be shared, but researchers can purchase a license at https://www.english-corpora.org/now/ *ABC News Australia*: This dataset is freely available for download on Kaggle at https://www.kaggle.com/datasets/therohk/million-headlines/data *The Times of India*: This dataset is freely available for download on the Harvard Dataverse at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DPQMQH The official APIs for *The Guardian* (https://open-platform.theguardian.com/documentation/) and *The New York Times* (https://developer.nytimes.com/apis) can be used to collect the headlines for these outlets. Upworthy headlines: The exploratory package can be downloaded from OSF at https://osf.io/3vqmp arXiv: The arXiv dataset can be downloaded from Hugging Face at https://huggingface.co/datasets/arxiv_dataset.

Received: 25 July 2024; Accepted: 29 January 2025;
Published online: 13 March 2025

References

Acerbi A (2019) Cultural Evolution in the Digital Age. 1st ed. Oxford University PressOxford. https://doi.org/10.1093/oso/9780198835943.001.0001

Achen CH (2005) Let's put garbage-can regressions and garbage-can probits where they belong. Confl Manag Peace Sci 22(4):327–339. https://doi.org/10.1080/07388940500339167

Akbik A, Blythe D, Vollgraf R (2018) Contextual string embeddings for sequence labeling. In: Bender, EM, Derczynski, L, Isabelle, P (eds.) Proc. 27th Int. Conf. Comput. Linguist. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 1638–1649

AllSides media bias chart (2019) AllSides. February 21

Bates D, Mächler M, Bolker B et al. (2015) Fitting linear mixed-effects models using **lme4**. J. Stat Softw 67(1). https://doi.org/10.18637/jss.v067.i01

Bhadani S, Yamaya S, Flammini A et al. (2022) Political audience diversity and news reliability in algorithmic ranking. Nat Hum Behav 6(4):495–505. https://doi.org/10.1038/s41562-021-01276-5

Bisceglia M (2023) The unbundling of journalism. Eur Econ Rev 158:104532. https://doi.org/10.1016/j.euroecorev.2023.104532

Blom JN, Hansen KR (2015) Click bait: forward-reference as lure in online news headlines. J Pragmat 76:87–100. https://doi.org/10.1016/j.pragma.2014.11.010

Boyd R, Richerson PJ (1988) Culture and the evolutionary process. Paperback ed. University of Chicago Press, Chicago

Brinkmann L, Baumann F, Bonnefon J-F et al. (2023) Machine culture. Nat Hum Behav 7(11):1855–1868. https://doi.org/10.1038/s41562-023-01742-2

Bulik M (2016) Which headlines attract most readers? N. Y. Times., June 13, sec. Times Insider

Carrasco-Farré C (2022) The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. Humanit Soc Sci Commun 9(1):162. https://doi.org/10.1057/s41599-022-01174-9

Chakraborty A, Paranjape B, Kakarla S et al. (2016) Stop clickbait: detecting and preventing clickbaits in online news media. In: 2016 Ieeeacm Int. Conf. Adv. Soc. Netw. Anal. Min. Asonam. IEEE pp. 9–16

Cinelli C, Forney A, Pearl J (2022) A crash course in good and bad controls. Sociol Methods Res 004912412210995. https://doi.org/10.1177/00491241221099552

D. Molina M, Sundar SS, Rony MMU et al. (2021) Does clickbait actually attract more clicks? three clickbait studies you must read. In: Proc. 2021 Chi Conf Hum Factors Comput. Syst. ACM, Yokohama Japan, pp. 1–19. https://doi.org/10.1145/3411764.3445753

Davies (2016) Corpus of news on the web (now)

Gligorić K, Lifchits G, West R et al. (2023) Linguistic effects on news headline success: evidence from thousands of online field experiments (registered report). Edited by K Sasahara. Sasahara, K (ed.) PLOS One. 18(3):e0281682. https://doi.org/10.1371/journal.pone.0281682

Goldenberg A, Gross JJ (2020) Digital emotion contagion. Trends Cogn Sci 24(4):316–328. https://doi.org/10.1016/j.tics.2020.01.009

González-Bailón S, Lazer D, Barberá P et al. (2023) Asymmetric ideological segregation in exposure to political news on facebook. Science 381(6656):392–398. https://doi.org/10.1126/science.ade7138

Hagar N, Diakopoulos N (2019) Optimizing content with a/b headline testing: changing newsroom practices. Media Commun 7(1):117–127. https://doi.org/10.17645/mac.v7i1.1801

Henrich J (2016) The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter. Princeton University Press, Princeton. https://doi.org/10.1515/9781400873296

Hills TT (2019) The dark side of information proliferation. Perspect Psychol Sci 14(3):323–330. https://doi.org/10.1177/1745691618803647

Honnibal M, Montani I, Van Landeghem S et al. (2020) SpaCy: industrial-strength natural language processing in python. https://doi.org/10.5281/zenodo.1212303

Hyland K, Zou HJ (2022) Titles in research articles. J Engl Acad Purp 56:101094. https://doi.org/10.1016/j.jeap.2022.101094

Interactive media bias chart (2024). Ad Fontes Media. https://adfontesmedia.com/interactive-media-bias-chart/. Accessed November 21

Jiang FK, Hyland K (2023) Titles in research articles: changes across time and discipline. Learn Publ 36(2):239–248. https://doi.org/10.1002/leap.1498

Kitaev N, Cao S, Klein D (2019) Multilingual constituency parsing with self-attention and pre-training. In: Proc. 57th Annu. Meet. Assoc. Comput. Linguist. Association for Computational Linguistics, Florence, Italy, pp. 3499–3505. https://doi.org/10.18653/v1/P19-1340

Kitaev N, Klein D (2018) Constituency parsing with a self-attentive encoder. In: Proc. 56th Annu. Meet. Assoc. Comput. Linguist. Vol. 1 Long Pap. Association for Computational Linguistics, Melbourne, Australia, pp. 2676–2686. https://doi.org/10.18653/v1/P18-1249

Kuiken J, Schuth A, Spitters M et al. (2017) Effective headlines of newspaper articles in a digital environment. Digit J 5(10):1300–1314. https://doi.org/10.1080/21670811.2017.1279978

Lebernegg N, Eberl J-M, Tolochko P et al. (2024) Do you speak disinformation? computational detection of deceptive news-like content using linguistic and stylistic features. Digit. Journal.:1–24. https://doi.org/10.1080/21670811.2024.2305792

Levitte M, Davidovitz G (2010) Pay-per-click is the new online paradigm. The guardian., May 31, sec. Media

Li Y, Breithaupt F, Hills T et al. (2024) How cognitive selection affects language change. Proc Natl Acad Sci 121(1):e2220898120. https://doi.org/10.1073/pnas.2220898120

Loewenstein G (1994) The psychology of curiosity: a review and reinterpretation. Psychol Bull 116(1):75–98. https://doi.org/10.1037/0033-2909.116.1.75

Lorenz-Spreen P, Lewandowsky S, Sunstein CR et al. (2020) How behavioural sciences can promote truth, autonomy and democratic discourse online. Nat Hum Behav 4(11):1102–1109. https://doi.org/10.1038/s41562-020-0889-7

Lorenz-Spreen P, Mønsted BM, Hövel P et al. (2019) Accelerating dynamics of collective attention. Nat Commun 10(1):1759. https://doi.org/10.1038/s41467-019-09311-w

Matias JN, Munger K, Le Quere MA et al. (2021) The upworthy research archive, a time series of 32,487 experiments in U.S. media. Sci Data 8(1):195. https://doi.org/10.1038/s41597-021-00934-7

McCulloch G (2019) Because internet: understanding the new rules of language. Riverhead Books, New York

Mesoudi A (2011) Cultural evolution: how Darwinian theory can explain human culture and synthesize the social sciences. University of Chicago Press, Chicago, Ill

Mesoudi A, Thornton A (2018) What is cumulative cultural evolution? Proc R Soc B Biol Sci 285(1880):20180712. https://doi.org/10.1098/rspb.2018.0712

Moretti F (2009) Style, inc. reflections on seven thousand titles (British novels, 1740–1850). Crit Inq 36(1):134–158. https://doi.org/10.1086/606125

Munger K (2020) All the news that's fit to click: the economics of clickbait media. Polit Commun 37(3):376–397. https://doi.org/10.1080/10584609.2019.1687626

Munger K (2024) The YouTube Apparatus. 1st ed. Cambridge University Press. https://doi.org/10.1017/9781009359795

O'Brien MJ, Lyman RL, Mesoudi A et al. (2010) Cultural traits as units of analysis. Philos Trans R Soc B Biol Sci 365(1559):3797–3806. https://doi.org/10.1098/rstb.2010.0012

Ovadya A, Thorburn L (2023) Bridging systems: open problems for countering destructive divisiveness across ranking, recommenders, and governance (version 3). arXiv. https://doi.org/10.48550/ARXIV.2301.09976
Robertson CE, Pröllochs N, Schwarzenegger K et al. (2023) Negativity drives online news consumption. Nat Hum Behav 7(5):812–822. https://doi.org/10.1038/s41562-023-01538-4
Rohrer JM (2018) Thinking clearly about correlations and causation: graphical causal models for observational data. Adv Methods Pract Psychol Sci 1(1):27–42. https://doi.org/10.1177/2515245917745629
Scott K (2021) You won't believe what's in this paper! clickbait, relevance and the curiosity gap. J Pragmat 175:53–66. https://doi.org/10.1016/j.pragma.2020.12.023
Simon H (1971) Designing organizations for an information-rich world. In Comput Commun Public Interest Johns Hopkins Press, Baltimore
Staff (2023) Twitter to let publishers charge users per article read, says Elon Musk. The guardian., April 30, sec. Technology
Stubbersfield J, Tehrani J, Flynn E (2018) Faking the news: intentional guided variation reflects cognitive biases in transmission chains without recall. Cult Sci J 10(1):54–65. https://doi.org/10.5334/csci.109
Stubbersfield JM (2022) Content biases in three phases of cultural transmission: a review. Cult Evol 19(1):41–60. https://doi.org/10.1556/2055.2022.00024
What is the 'deeply read' list? (2024). The guardian., February 28, sec. Info

Acknowledgements

We thank Deborah Ain for her help in editing this manuscript.

Author contributions

Conceptualization: P.N., M.M., P.L-S.; Methodology: P.N., P.L-S.; Data acquisition, linguistic and statistical analysis, writing—original draft: P.N.; Writing—review and editing: P.N., M.M., P.L-S.; Supervision: P.L-S.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Ethical approval

This research was conducted using publicly available news corpora collected from online sources, for which the authors did not perform studies involving human participants.

Informed consent

This research was conducted using publicly available news corpora collected from online sources, for which the authors did not perform studies involving human participants.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-04514-7.

Correspondence and requests for materials should be addressed to Pietro Nickl.