A pilot corpus study of stance and compression in English-language digital news headlines

Nigora Satibaldiyeva Alimqul qizi
n.satibaldiyeva@uzswlu.uz
Teacher,
Uzbek State World Languages University

Annotation *This article examines how a small corpus can still reveal stable editorial preferences in digital news headlines. The study uses a pilot corpus of 42 English-language headlines collected from the Reuters Technology page and the Kun.uz English homepage on April 6-7, 2026. The analysis combines simple corpus counts with close qualitative reading. Four markers were coded: headline length, numerals, attribution verbs, and explicit geographical references. The results show that the two datasets are close in average length, but they package public reality differently. Reuters headlines are strongly corporate and market-centered, while Kun.uz foregrounds state institutions, public regulation, and national framing. Numerals occur more often in the Kun.uz sample, whereas Reuters relies more heavily on company names, deal vocabulary, and high-speed risk framing. Because the corpus is small and opportunistic, the findings are interpreted cautiously. Even so, the pilot design demonstrates that corpus methods help describe news style with evidence rather than intuition.*

Keywords *Corpus linguistics, media texts, digital news, headlines, discourse analysis, stance, compression*

Ingliz tilidagi raqamli yangilik sarlavhalarida baholash va kompressiyaning pilot korpus tadqiqi

Satibaldiyeva Nigora Alimqul qizi
n.satibaldiyeva@uzswlu.uz
O‘qituvchi,
O‘zbekiston davlat jahon tillari universiteti

Annotatsiya *Ushbu maqola kichik hajmdagi korpus ham raqamli yangilik sarlavhalaridagi barqaror tahririy afzalliklarni ko‘rsatib bera olishini tahlil qiladi. Tadqiqot materiali sifatida 2026-yil 6-7-aprel kunlari Reuters Technology sahifasi va Kun.uz English bosh sahifasidan to‘plangan 42 ta ingliz tilidagi sarlavhadan iborat pilot korpus olindi. Tahlilda sodda korpus hisoblari va yaqin matniy o‘qish birlashtirildi. To‘rtta ko‘rsatkich kodlandi: sarlavha uzunligi, sonli markerlar, atributiv fe’llar va aniq geografik ishoralar. Natijalar ikki to‘plamning o‘rtacha uzunligi bir-biriga yaqinligini, biroq ijtimoiy voqelikni paketlash usuli farqlanishini ko‘rsatdi. Reuters korporativ subyektlar va bozor dinamikasini markazga qo‘ysa, Kun.uz davlat institutlari, jamoat tartibga solinishi va milliy ramkani ko‘proq oldinga chiqaradi. Sonli birliklar Kun.uz namunalarida ko‘proq uchraydi, Reuters esa kompaniya nomlari, bitim leksikasi va tezkor risk freymlariga ko‘proq tayanadi. Korpus kichik va imkon qadar yig‘ilgani sababli xulosalar ehtiyotkor talqin qilinadi.*

Kalit so‘zlar *Korpus lingvistikasi, mediamatnlar, raqamli yangiliklar, sarlavhalar, diskurs tahlili, baholash, kompressiya*

The Lingua Spectrum
Cogito, ergo sum

The Lingua Spectrum
Volume 4, April 2026

Пилотное корпусное исследование оценки и компрессии в англоязычных цифровых новостных заголовках

Сатибалдиева Нигора Алимқул қизи
n.satibaldiyeva@uzswlu.uz
Преподаватель,
Узбекский государственный университет мировых языков

Аннотация *В статье рассматривается, каким образом даже небольшой корпус позволяет выявить устойчивые редакционные предпочтения в цифровых новостных заголовках. Материалом послужил пилотный корпус из 42 англоязычных заголовков, собранных со страницы Reuters Technology и англоязычной главной страницы Kun.uz 6-7 апреля 2026 года. Анализ сочетает простые корпусные подсчёты и качественное близкое чтение. Были закодированы четыре показателя: длина заголовка, наличие числовых маркеров, глаголы атрибуции и явные географические ссылки. Результаты показывают, что средняя длина заголовков в двух массивах близка, однако способы упаковки социальной реальности различаются. Reuters строит заголовки вокруг корпоративных акторов и рыночной динамики, тогда как Kun.uz выдвигает на первый план государственные институты, общественное регулирование и национальное рамкирование. Числовые элементы чаще встречаются у Kun.uz, а Reuters активнее использует названия компаний, лексику сделки и ускоренное риск-ориентированное фреймирование. Из-за малого объёма корпуса выводы интерпретируются осторожно.*

Ключевые слова *Корпусная лингвистика, медиатексты, цифровые новости, заголовки, дискурс-анализ, оценка, компрессия*

Introduction

Corpus linguistics changed media research in one important way: it made repeated textual behaviour visible. Instead of selecting two or three convenient examples and then building a large conclusion around them, the researcher can begin with observable frequency, distribution, and recurrence. This shift has been central both to corpus-based discourse analysis and to corpus-informed translation and media studies (Baker, 1993; Biber et al., 1998). In news discourse, such a shift matters because headlines are highly compressed sites of meaning. They do not simply summarize events. They rank actors, select angles, activate values, and signal what should be treated as urgent, risky, profitable, or nationally relevant (Bednarek, 2006; Bednarek & Caple, 2014, 2017).

At the same time, large corpora are not the only valid route. A small pilot corpus cannot produce universal claims, but it can reveal stable editorial habits if the dataset is handled carefully and the limitations are named openly. This article therefore asks a modest question: what can a small corpus tell us about stance and compression in contemporary English-language digital news headlines? To answer it, I compare two publicly accessible headline sets collected on April 6-7, 2026 from the Reuters Technology page and the Kun.uz English homepage. The aim is not to rank one outlet as better or worse. The aim is to show how corpus reasoning can describe

different editorial preferences with explicit evidence.

Theoretical background

Work in corpus-assisted discourse studies has repeatedly shown that quantitative signals become most meaningful when they are interpreted together with discourse context. Baker (1993) argued early that corpus methods would reshape language-oriented disciplines because they expose recurrent linguistic patterns that are difficult to detect intuitively. Later work in CADS refined this argument by treating counts not as an end in themselves, but as a guide for close reading and ideological interpretation (Gabrielatos & Baker, 2008; Gillings et al., 2023; Taylor & Marchi, 2018).

News discourse offers a particularly productive field for this approach. Bednarek and Caple (2014) describe news values as discursively constructed rather than naturally given. In that view, newsworthiness is assembled through lexical choice, attribution, naming practices, emotional charge, and multimodal design. Bednarek (2016) adds that attribution is not a decorative appendage to the news story. It helps position voices, distribute responsibility, and calibrate evidential weight. For headline analysis, this means that a short clause like *company cautions on geopolitics* or *government to allocate over 1 trillion UZS* does more than transmit information. It also builds an implicit map of agency, urgency, and legitimacy.

Headline compression is equally relevant. Digital headlines often omit articles, reduce subordinate structure, and pack several semantic relations into a narrow syntactic space. Such compression supports speed and scanability, but it also sharpens framing. An actor placed in initial position receives immediate prominence; a numeral can function as a credibility cue; a reporting verb such as *says*, *warns*, or *confirms* can lower institutional risk while preserving the headline's force. These are exactly the kinds of choices that corpus methods can capture even in a small dataset.

Materials and method

The corpus for this pilot study consists of 42 English-language headlines. Twenty headlines were taken from the Reuters Technology page, and twenty-two were taken from the Kun.uz English homepage and its visible latest-news block. All items were collected from publicly accessible pages on April 6-7, 2026 (Kun.uz, 2026; Reuters, 2026). Because the dataset was gathered from live pages rather than from an archived balanced corpus, it should be treated as an opportunistic pilot sample, not as a statistically representative model of either outlet.

Outlet	Section / page	Headlines	Dominant topical zone	Collection date
Reuters	Technology	20	Corporate technology, markets, platform risk	7 April 2026
Kun.uz English	Homepage + latest news	22	Public policy, economy, technology, society	6–7 April 2026

Table 1. *Composition of the pilot corpus*

Four markers were coded manually: average headline length in words, presence of numerals, presence of attribution verbs, and presence of explicit geographical reference. These markers were selected for practical reasons. Length gives a rough indicator of compression; numerals often package value, scale, or urgency; attribution verbs indicate evidential positioning; geographical references signal spatial framing. After the count-based stage, the headlines were read qualitatively to identify recurrent actor choices, domain preferences, and evaluative tendencies. This mixed procedure follows the logic of corpus-

assisted discourse studies in which counting narrows the field of interpretation but does not replace contextual reading (Gillings et al., 2023; Taylor & Marchi, 2018).

Results and Discussion

The first result is a useful one precisely because it is not dramatic. The two headline sets are close in average length. Reuters headlines average 11.4 words, while the Kun.uz sample averages 11.1 words. In other words, the difference between the outlets does not lie in simple brevity. Both write for fast digital scanning. The divergence appears in what they choose to compress and foreground.

Outlet	Mean words	With numerals	With attribution verbs	With geographical reference
Reuters	11.4	3 (15.0%)	7 (35.0%)	10 (50.0%)
Kun.uz English	11.1	8 (36.4%)	7 (31.8%)	12 (54.5%)

Table 2. *Selected headline markers in the pilot corpus*

Numerals occur more than twice as often in the Kun.uz sample. This is not accidental. The outlet repeatedly packages public issues through measurable quantities: *over 1 trillion UZS*, *7.1% in March*, *30% cut in customs fees*, *$500 million*, *first launch in 2028*. Such numerical packaging gives the headline an administrative and policy-oriented texture. It presents events as governable, countable, and institutionally anchored. Reuters also uses numerals, but less frequently in this sample. When it does so, the number usually serves a market or transaction frame, as in *$65 billion* or *80% of package volume*.

Attribution verbs appear in both datasets at roughly similar levels, yet their discourse function differs. Reuters headlines often use attribution to manage legal or market uncertainty: *Nikkei Asia reports*, *sources say*, *government says*, *court rules*. The wording keeps the headline moving quickly while distributing epistemic responsibility outward. Kun.uz also uses attribution, but more often in a public-information mode: *confirms cyberattack*, *warns of potential mudslides*, *reports sexual harassment*, *comments on Middle East crisis*. Here attribution supports service journalism and public accountability more than market risk management.

Actor choice offers the clearest contrast. Reuters headlines are strongly corporate and firm-centered. Initial position is commonly occupied by companies, platforms, investors, or named economic actors: *OpenAI urges California, Delaware to investigate Musk's anti-competitive behavior*, *Samsung flags eightfold jump in quarterly profit as AI chip demand pumps prices*, *Broadcom signs long-term deal to develop Google's custom AI chips*, and *Foxconn first-quarter revenue jumps, company cautions on geopolitics*. Such openings immediately naturalize a world in which corporations are the main engines of relevance.

Kun.uz English, by contrast, repeatedly foregrounds the state, the nation, or public institutions: *President Mirziyoyev sets new mandates for officials to support youth startups and education*, *Government to allocate over 1 trillion UZS for Ramadan charity initiatives*, *Uzbekistan to develop sovereign satellite system with first launch in 2028*, and *Uzbekistan enforces new safety regulations for temporary public installations*. In this sample, the lexical item *Uzbekistan* appears nine times, which indicates a strong national framing strategy.

A different pattern emerges in evaluative loading. Neither outlet depends on overtly emotional adjectives in every line, but both use small lexical triggers that quietly shape stance. Reuters leans toward risk, competition, and disruption through items such as *worry*, *snags*, *subdued*, *probe*, and *turmoil*. Kun.uz uses a

more civic and administrative set of evaluative cues, for example *shadow network*, *cyberattack*, *safety concerns*, *fraud risks*, and *fabricated reports*. The point is not that one outlet is more objective than the other. Rather, each outlet routinizes a different register of seriousness.

From a corpus perspective, these differences matter because they are patterned rather than anecdotal. A single headline can mislead. A small group of headlines, when counted and then interpreted, begins to show the editorial logic beneath surface variation. Reuters in this sample organizes relevance around firms, deals, platform competition, and transnational economic tension. Kun.uz English organizes relevance more often around public administration, national institutions, measurable reform, and citizen-oriented information. The average length stays similar, but the discourse priorities diverge.

A limitation should be stated directly. The corpus is small, live, and time-bound. It captures a snapshot rather than a season, and it compares a specialized technology page with a broader national homepage. For that reason, the results should not be stretched into a general theory of digital journalism. Still, the pilot remains useful. It demonstrates a workable research design for larger studies and shows how corpus methods can make headline analysis more disciplined, more transparent, and less dependent on impressionistic selection.

Conclusion

This article set out to test what a small corpus can reveal about digital news style. The answer is modest but clear. Even a pilot dataset can uncover stable tendencies when the method is explicit and the interpretation remains cautious. In the 42-headline corpus examined here, Reuters and Kun.uz English show similar levels of formal compression, yet they differ in how they package relevance. Reuters foregrounds firms, products, deals, and corporate risk. Kun.uz English foregrounds the state, public regulation, national institutions, and quantified policy information.

For media-text research, this matters in two ways. First, corpus work helps replace abstract impressions with evidence that can be checked, revised, and expanded. Second, headline analysis should not be treated as a minor supplement to full-text study. In digital journalism, the headline is often the first framing event and sometimes the only part many readers see. A larger follow-up study could compare archived datasets across longer time spans, separate topic categories more strictly, and extend the design to Russian- and Uzbek-language headline corpora. The present pilot does not finish that work. It opens it.

References:

1. Baker, M. (1993). *Corpus linguistics and translation studies: Implications and applications.* In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), Text and technology: In honour of John Sinclair. 233–250. John Benjamins. https://doi.org/10.1075/z.64.15bak
2. Bednarek, M. (2006). *Evaluation in media discourse: Analysis of a newspaper corpus.* Continuum.
3. Bednarek, M. (2016). Voices and values in the news: News media talk, news values and attribution. *Discourse, Context & Media, 11,* 27–37. https://doi.org/10.1016/j.dcm.2015.11.004
4. Bednarek, M., & Caple, H. (2014). Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse & Society, 25*(2), 135–158. https://doi.org/10.1177/0957926513516041

5. 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
6. Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804489
7. Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996–2005. *Journal of English Linguistics, 36*(1), 5–38. https://doi.org/10.1177/0075424207311247
8. Gillings, M., Mautner, G., & Baker, P. (2023). *Corpus-assisted discourse studies*. Cambridge University Press. https://doi.org/10.1017/9781009168144
9. Kun.uz. (2026, April 7). *News of Uzbekistan and the world*. https://kun.uz/en/
10. Reuters. (2026, April 7). *Technology*. https://www.reuters.com/technology/
11. Taylor, C., & Marchi, A. (Eds.). (2018). *Corpus approaches to discourse: A critical review*. Routledge. https://doi.org/10.4324/9781315179346