

A study on bias analysis of north korea's nuclear threat by country using a self-expanding crawler

Yonhong Jeong*, Gayeon Ha, Sueyeon Lee

Division of Nuclear Nonproliferation Policy, Korea Institute of Nuclear Nonproliferation and Control,
1418 Yuseong-daero, Yuseong-gu, Daejeon 34101, Republic of Korea

*Corresponding author: jyh1404@kinac.re.kr

***Keywords:** North Korea nuclear, OSINT, self-expanding crawling, media framing

1. Introduction

Assessment of North Korea's nuclear program depends heavily on open-source information (OSINT), in which media reporting is both a primary source and an actor shaping international opinion and policy. Comparative media research has long held that such reporting is not plain fact transmission, but is structurally reconstituted by the political and ideological frames of the bloc an outlet belongs to [1]. That liberal-bloc media (South Korea, the United States, Japan) and socialist-bloc media (Russia, China) report the same nuclear event in markedly different vocabulary is widely observed, yet large-scale, quantitative verification of that bias, and computational identification of the mechanism by which it diverges, remain scarce.

This study asks how bloc bias diverges when a self-expanding (snowball) collection, one that extracts new keywords from what it has already gathered and re-queries with them, is operated in isolation for each bloc. Two literatures motivate the design. Framing and securitization theory predict that liberal outlets foreground security threat, while state media foreground political legitimacy [2, 3]. The algorithmic filter bubble predicts that automated re-collection deepens whatever separation is already present [4]. Combining them allows the separation and its amplification to be measured separately.

2. Methods

2.1 Data and bloc isolation

The substrate is GDELT Global Knowledge Graph (GKG) metadata, which supplies language-normalized structured fields for multilingual articles [5]. Every GKG record from 1 January 2018 to 30 June 2026 whose source domain belongs to an outlet in Table I and whose V2Locations field contains "North Korea" (language-independent, since GKG geocodes place names into English) was materialized once in BigQuery, giving a pool of 170,983 records from which every document below was drawn. Three fields were used: AllNames, the multilingual entities extracted per article; V2Themes, an automatic assignment into roughly 300 event categories; and V2Tone, a document-level tone score computed from English-language lexicons.

Blocs were assigned at the outlet level, by source domain, before collection and never revised, so membership is independent of document content. Two criteria define them. The first is the home state's alignment on the North Korean nuclear issue: the liberal bloc holds outlets of the United States, the United Kingdom, South Korea and Japan, whose governments treat the program as a proliferation threat; the socialist bloc holds outlets of Russia and China, North Korea's principal partners, and of North Korea itself. The second is the media system, market-based and editorially autonomous versus state-directed [1]: every socialist-bloc outlet is state-owned, state-funded or a party organ, and no liberal-bloc outlet is. Ownership, not a site's language or domain suffix, decides membership.

Within each bloc, outlets had to satisfy three conditions: (i) indexing by GDELT under a stable domain throughout the window; (ii) national or international reach with sustained North Korea coverage (a wire service, leading daily, public broadcaster or dedicated North Korea outlet); and (iii) jointly spanning the media types of the bloc, so that within-bloc heterogeneity is measurable. The liberal list therefore mixes two international wires, the Korean national wire, two Korean dailies of opposite editorial leaning, a public broadcaster and two North Korea specialists; the socialist list mixes state agencies, state- or party-affiliated international services, a party organ, a pro-government news site and a business daily (Table I). KCNA and Rodong Sinmun fail (i), being absent from GDELT's index, so North Korea's own organs are excluded and the socialist bloc is, in practice, Russian and Chinese. Documents were stored in physically separated per-bloc databases, so no keyword or frame from one bloc can enter the other bloc's collection.

2.2 Self-expanding snowball and sample construction

Both blocs were run with identical parameters against their own partition of the pool: the same English seeds, "North Korea nuclear" and "North Korea", 15 rounds of 20 documents, hence 300 documents per bloc, and no early stopping. A round proceeds in four steps. (i) Each query phrase of the round is matched as a substring of the lowercased metadata (AllNames, V2Organizations, V2Persons and V2Themes) of every record. (ii) Matching records from the bloc's domains not retrieved earlier are read in a fixed pseudo-random order (a URL

hash, reproducible and not concentrated in time) and allocated round-robin across the round's phrases; records with an empty entity field are skipped and duplicates are removed by a content hash. (iii) The entity phrases of the stored documents pass a rule-based noise filter (over-long phrases, repeated tokens, ceremonial venue names) and are registered in the bloc's keyword table with the round number. (iv) Registered phrases not yet used as queries are ranked by a TF-IDF-type score, term frequency times $\log(1 + N/(1 + df))$, with a floor of three documents (relaxed to two when fewer than six qualify, which happened for 6 of 84 socialist phrases in rounds 0 to 2), and the top six become the next round's queries. No large language model takes part, and the procedure is deterministic: identical seeds and parameters reproduce identical samples.

The phrase "North Korea nuclear" matches only 33 liberal and 6 socialist records in the pool, so round 0 consisted of 10 + 10 and 6 + 14 documents from the two seeds; the remaining 280 documents per bloc were retrieved by 84 expansion phrases each (Table I). Both samples follow the time profile of the pool, with 58 % (liberal) and 54 % (socialist) of documents from the summit diplomacy of 2018 to 2019. The results below replace those of the earlier draft, whose run never issued an expansion query because of a defect in the novelty test.

Table I: Outlet classification and sample composition (n = documents retrieved).

Outlet (domain)	Ctry	Type	n
<i>Liberal bloc (US, GB, KR, JP), n = 300</i>			
Reuters (reuters.com)	GB	wire	118
AP (apnews.com)	US	wire	20
NK News (nknews.org)	US	specialist	17
38 North (38north.org)	US	specialist	3
Yonhap (yna.co.kr)	KR	wire	50
Chosun Ilbo (chosun.com)	KR	daily (cons.)	38
Hankyoreh (hani.co.kr)	KR	daily (prog.)	47
NHK (nhk.or.jp)	JP	public TV	7
<i>Socialist bloc (RU, CN), n = 300</i>			
TASS (tass.com)	RU	state agency	34
RIA Novosti (ria.ru)	RU	state agency	27
Sputnik (sputnikglobe.com)	RU	state-affiliated	92
RT (rt.com)	RU	state-funded	33
Lenta.ru (lenta.ru)	RU	pro-government	23
Kommersant (kommersant.ru)	RU	business daily	17
Xinhua (news.cn)	CN	state agency	23
Global Times (globaltimes.cn)	CN	party-affiliated	33
People's Daily (people.com.cn)	CN	party organ	18

2.3 Metrics

Lexical divergence was quantified by the Jaccard distance between the two blocs' cumulative keyword sets after each round, defined as one minus the ratio of intersection to union, and by a bloc-separation alignment score, defined as the proportion of bloc-exclusive keywords. Topic drift was measured as the share of a round's documents whose metadata carries no nuclear-activity anchor (weapon, test, enrichment, reprocessing, plutonium or warhead terms). The tone difference was tested by bootstrap resampling with 5,000 replicates, computed both at document level and with outlet-level clustering, to reflect the non-independence of documents from one outlet.

3. Results

3.1 Lexical divergence is immediate, not amplified

The inter-bloc Jaccard distance was 0.956 at round 0 and 0.953 at round 14, and stayed within 0.951 to 0.961 throughout, with a bloc alignment of 0.953 (Fig. 1). Expansion therefore does not widen the lexical separation: it is present in the first 20 documents, under identical seeds, and 14 rounds of self-expansion leave it unchanged. The separation belongs to the two source networks, not to the collection loop.

3.2 Opposed framing profiles

The expansion trajectories themselves are the clearest framing evidence. Liberal queries moved through the institutions of the US-ROK alliance (State Department, Blue House, Security Council, Pompeo, Biegun, joint chiefs) and, from round 4, through wire-service boilerplate and Reuters bylines, which alone pulled 42 of the 300 documents. Socialist queries moved through the party-state actors of the Russia-China-DPRK triangle (Putin, Xi Jinping, the Russian Federation, the State Council, the Communist Party, comrade Kim) and through KCNA and Global Times themselves. The V2Themes distribution agreed: crisis (129 against 91 documents), fragility and conflict (141 against 119) and policy (121 against 75) themes were more frequent in the liberal sample, whereas leader (144 against 107) and Russia-related (115 against 54) themes were more frequent in the socialist sample. In both blocs the expansion followed the most frequent named actors rather than nuclear vocabulary: on average 75 % of liberal and 89 % of socialist documents per round carried no nuclear-activity anchor.

3.3 Tone gap and within-bloc heterogeneity

Mean tone was -1.487 (SD 2.574) for the liberal bloc, where 73 % of documents were negative, and -0.313 (SD 3.267) for the socialist bloc, where 43 % were positive. The bootstrap difference was 1.175, with a 95 % confidence interval of [0.715, 1.651] and Cohen's d of 0.40; under outlet-cluster bootstrap the interval widens to [0.18, 2.19] but still excludes zero. The gap is 0.85 in the first batch and 1.17 after 15 rounds; the change is within sampling noise.

The favourability is again not uniform (Fig. 2). Chinese party and state media are the positive pole (People's Daily +2.21, Xinhua +1.00), RT and Global Times are near neutral, and the Russian outlets that contribute most documents are negative (Sputnik -1.46, Lenta -0.79, TASS -0.43). On the liberal side the wires are the most critical (AP -2.95, Reuters -1.96) while Hankyoreh (-0.22) and NK News (0.00) are near neutral. The bloc-level gap is therefore, in part, a gap between Chinese party-state media and Western wires; the socialist bloc is not a single actor.

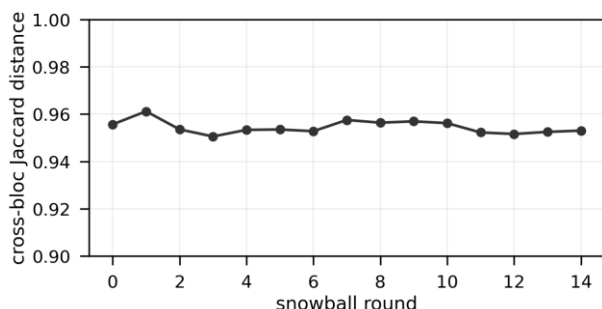


Fig. 1. Cross-bloc keyword Jaccard distance by snowball round.

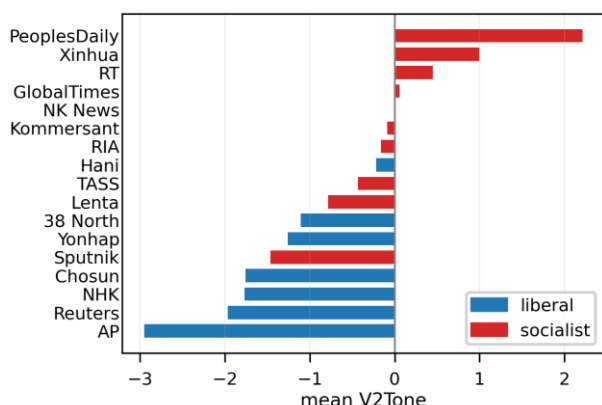


Fig. 2. Per-outlet mean tone; Chinese party and state media are the positive pole, Russian outlets are neutral or negative.

3.4 Robustness

Four seed concepts, the English term and its Korean, Chinese and Russian equivalents, each run with the same expansion for eight rounds, produced inter-bloc distances of 0.94 to 0.97, so the divergence is not an artifact of the English seed. Within-bloc similarity across the four seeds was 0.49 (liberal) and 0.26 (socialist), an order of magnitude above the cross-bloc similarity of 0.03 to 0.06: the seed decides where the walk starts, the bloc decides where it converges. Across ten major North Korea events in a ± 30 day window, the tone gap was event-dependent, peaking at the Russia-North Korea summit. Three language models scoring the metadata of 48 sampled documents as proxy coders agreed with each other (mean Pearson 0.693) more than with V2Tone (0.448), and on metadata alone saw a bloc gap of only 0.29 against 1.14 for V2Tone on the same documents: the V2Tone gap is mainly a body-text lexical signal.

4. Limitations

V2Tone is a dictionary-based score, built on English resources, and its inter-coder reliability is unverified, so ceremonial and party vocabulary can be scored spuriously positive; tone is therefore a supporting signal, and the robust evidence is the lexical separation and the opposed expansion profiles. The 300 documents per bloc are an expansion trajectory, not a random, stratified sample of the 170,983 records: they mirror the pool's concentration in 2018 to 2019, they follow whatever

named entities are most frequent, bylines and GKG transliteration errors included, and most of them lack a nuclear-activity anchor. Generalization is conditional on this mechanism; a year-by-outlet stratified draw with a nuclear-anchored relevance gate is required next. The outlet-cluster interval is wide with only eight or nine outlets per bloc, and major events cluster within the window and were not decomposed from structural bias.

5. Conclusions

Liberal and socialist media reach structurally separated vocabularies and opposed actor networks on the same North Korean events, and that separation is present in the very first batch of seed-matched documents; 14 rounds of self-expansion neither widen nor close it. What expansion does is follow each bloc's institutional source network, alliance institutions and wire bylines on one side, party-state actors and state agencies on the other, so the bias it reproduces is that of source networks and institutional indexing rather than the intent of individual outlets.

The practical implication is that single-bloc OSINT collection is structurally biased. Relying only on liberal sources misses the legitimacy logic that North Korea and its partners articulate, while relying only on socialist sources loses the nonproliferation and verification context. For high-stakes judgments, such as sanctions implementation and capability assessment, balanced, multi-bloc collection with systematic, cross-source verification is an operational necessity rather than an academic preference. All code, schemas, and fixed seeds are published with 227 unit tests for independent reproduction.

ACKNOWLEDGEMENT

This work was supported by the Nuclear Safety Research Program through the Korea Foundation of Nuclear Safety (KoFONS) using the financial resource granted by the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No. 00409134).

REFERENCES

- [1] D. C. Hallin and P. Mancini, Comparing Media Systems: Three Models of Media and Politics, Cambridge University Press, 2004.
- [2] R. M. Entman, Framing: Toward Clarification of a Fractured Paradigm, Journal of Communication, Vol. 43, No. 4, pp. 51-58, 1993.
- [3] T. Balzacq, The Three Faces of Securitization: Political Agency, Audience and Context, European Journal of International Relations, Vol. 11, No. 2, pp. 171-201, 2005.
- [4] E. Pariser, The Filter Bubble: What the Internet Is Hiding from You, Penguin Press, 2011.
- [5] K. Leetaru and P. A. Schrod, GDELT: Global Data on Events, Location, and Tone, 1979-2012, ISA Annual Convention, 2013.