

Nigerian Salafi Discourse Corpus (NSDC): Construction, Methods, and Use

Idris Ayomo Oke

`ioke254746@stu.ui.edu.ng`

University of Ibadan

Temitope Michael Ajayi

University of Ibadan

Research Article

Keywords: corpus construction, Facebook, Salafism, Nigeria, discourse-historical approach, annotation, data cleaning

Posted Date: September 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10966766/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

This paper documents the construction of the Nigerian Salafi Discourse Corpus (NSDC), a purpose-built corpus of 5,791 Facebook posts (2,485,963 words; 2,518,453 tagged tokens) published by Nigerian pages and profiles over a ten-year frame (2015–2025). The NSDC was assembled from eight keyword-filtered exports of the Meta Content Library and processed through a fully documented, machine-verified pipeline executed by Hermes Agent, an agentic artificial-intelligence research assistant. The pipeline performs schema validation, three-pass deduplication, Unicode normalisation and noise removal, SGML-annotated corpus assembly, cross-keyword merging, and calendar-year segmentation. The corpus is annotated in three layers: Penn Treebank part-of-speech and lemma tagging; assignment of tokens to fifteen purpose-built semantic fields; and sentence-level coding of fifteen discourse strategies under the discourse-historical approach. This paper reports aggregate statistics, data-quality challenges and solutions, verification outcomes, a manual reliability audit of the annotation layers, and research applications.

1. Introduction

Corpus-assisted discourse studies (CADS) has matured into a distinctive methodology in which quantitative corpus techniques and qualitative discourse analysis are combined rather than opposed (Baker, 2006; Partington, Duguid and Taylor, 2013). Large, principled corpora now make it possible to identify cumulative and often non-obvious patterns of representation that escape the unaided reader, while close reading keeps the analyst anchored in social context (McEnery and Hardie, 2012; Stubbs, 2001). The analytical value of any such study is contingent on the quality of the underlying corpus, judged against representativeness, cleanliness, machine-readability and transparency of construction (Biber, Conrad and Reppen, 1998; Hunston, 2002).

Notably, social media has become a primary context for religious discourse in Nigeria, where Islam is publicly contested, preached and politicised (Thurston, 2016). Salafism¹ in particular is an object of intense discursive struggle among Sunni Muslims, between Salafi and Sufi constituencies, and in relation to the Nigerian state and the global security discourse. Scholars have analysed this dynamic through ethnography and political science (e.g., Mustapha, 2014; Kendhammer, 2016) but rarely through systematically collected language data. Facebook, one of the most widely used platforms in the country, hosts this contestation at scale, in a dense, orthographically heterogeneous mix of English, Arabic, Roman-script transliterated Arabic and Nigerian languages. Social-media text is now a recognised corpus resource because it combines vernacular authenticity with volume (Gatto, 2014), and this Nigerian Salafi Discourse Corpus (NSDC) was designed to open this context to rigorous corpus-linguistic investigation. To the best of my knowledge, no comparable corpus of Nigerian religious social-media discourse, and none combining part-of-speech, semantic and discourse-strategic annotation, is currently available.

The research agenda that the NSDC therefore serves is defined by a single aim and four objectives. The aim is to examine how Salafism and Salafis are represented in Nigerian Facebook discourse between 2015 and 2025. The objectives are: (i) to identify the linguistic features associated with Salafism and Salafis; (ii) to describe the discursive strategies deployed when discussing them; (iii) to examine the dominant discourses through which they are constructed; and (iv) to trace diachronic shifts across the ten-year frame. In this paper, I report every design decision, from data collection to cleaning rules and annotation categories, each mapped to one or more of these objectives. This paper describes the keyword selection, the Meta approval and download, the instrument, the principle of auditable, verified processing, cleaning, integration, segmentation, annotation, and the research applications the corpus makes possible.

2. Constructing the Corpus

The construction of this corpus operates across a set of levels discussed below: the downloading of the social-media posts, the cleaning and its processes, as well as the annotation of the cleaned data.

2.1 Downloading the data from Facebook: the processes, challenges and result

The construction of the NSDC presupposed access to public Facebook content at scale, a privilege the Meta Content Library extends to researchers only by formal application (Meta, 2026). Applicants must declare an institutional affiliation, specify the aims of the proposed study and accept the platform's terms of use, after which access is granted at Meta's discretion for a defined period. For the research that informed this corpus, I applied for and obtained access as principal researcher, with approval valid until 29 August 2028. The approval confers two further privileges: inviting co-researchers to the project, and access to the Content Library API for programmatic retrieval of the same content at scale.

Data collection proceeded through the Library's search interface, configured identically for each of the eight keywords that neutrally and pejoratively capture discourse around the Salafi movement and ideology (*Salafis*, *Salafism*, *Salafiyyah*, *Ahlus-Sunnah*, *Sunni*, *Wahabism*, *Wahabiyyah*, *Salafool*). Platform, content type and variant were set to Facebook, posts and the downloadable subset; the date range was set to successive June-to-June windows covering 2015 to 2025; and the post surface to pages and profiles with more than 25,000 followers, since selecting events or groups disables the country filter. Nigeria and English were specified; content types covered albums, links and re-shares, story highlights and videos; branded content was included; and both verified and non-verified surfaces were selected.

Three challenges constrained this stage. First, the Library permits only ten downloads per day, so the ten-window frame of a single keyword consumed a full day's quota. On some days, in fact, the limit blocked downloads even when the quota appeared unused. Second, in some exports,

post text appeared scattered in columns other than the text column, and such rows carried no metadata. These rows were therefore excluded, while the four required columns (creation_time, mcl_url, surface.name and text) were copied and arranged manually from the 36-column exports. Third, the Library returned 'No content found' for Ahlus-Sunnah (no posts in 2017–2018), for Wahabiyyah (nothing before 2020) and for Salafool (nothing before 2023). This is only a data-source gaps rather than processing losses. The download stage therefore closed with 7,172 posts and 2,772,120 raw word tokens across the eight exports.

2.1.1 Ethics and legal considerations

Research on publicly available social-media content raises distinct ethical and legal questions (Zimmer, 2010; Townsend and Wallace, 2017; Franzke et al., 2020), and the NSDC's position is set out here. All posts were collected through the Meta Content Library, Meta's research interface, following a formal application that stated the aims of the study and accepted the platform's terms of use. Only public content was retrieved containing posts by pages and profiles with over 25,000 followers. As such, no private groups, messages or non-public data were accessed.

Similarly, the corpus deliberately retains the names of the publishing pages and profiles, together with post URLs. This is done because it is a record of public religious discourse, the actors are public figures and organisations, and attribution is essential for the discourse-historical analysis of voice and authority (Section 6.3). However, no contact details, private communications or data about private individuals beyond what they themselves published are included, and no content was collected from restricted-visibility surfaces. As such, processing of this manifestly public data for research falls within the research exemption of the Nigerian Data Protection Regulation (2019) and the General Data Protection Regulation (EU) 2016/679.

In terms of harm and misuse, this paper notes that religious polemic can be sensitive, and the corpus is thus released for research use under a Creative Commons Attribution 4.0 licence, with the underlying posts remaining the property of their authors, and the cleaning pipeline strips URLs from the text, reducing re-exposure of external content. Users are therefore expected not use the corpus to target, harass or identify private individuals.

2.2 The instrument: Hermes Agent

All cleaning, merging and annotation described in this paper were executed by an agentic artificial-intelligence research assistant: the Hermes Agent, an open-source autonomous agent framework by Nous Research that couples a large language model (LLM) with a tool layer comprising file-system access, a shell, a Python interpreter and document-generation modules (<https://hermes-agent.nousresearch.com>). For this research, the assistant was hosted on a server, operated by the researcher through a messaging interface, and given the operational name *Salih*. The researcher designed the pipeline, the cleaning specification, the annotation scheme and the verification gates; *Salih* executed the implementation.

2.2.1 Architecture

The architecture is deliberately hybrid: the LLM contributes planning, judgment and natural-language reasoning, while every data transformation is executed by deterministic, version-controlled Python scripts. Reproducibility therefore does not depend on the stochastic component: the cleaning pipeline (corpus_pipeline.py), merge module (merge_nsdc.py), annotator (tag_nsdc.py) and verification suites (verify_corpus.py; verify_tagged.py) are ordinary programmes any researcher can re-run identically with the source exports.

2.2.2 Algorithmic capacities

The capacities that matter for the NSDC construction are: (i) schema-robust parsing, which locates headers by content rather than by row number; (ii) three-pass deduplication with reason-tagged drops, so that every excluded row is attributable; (iii) Unicode-aware normalisation, including NFC composition, transliteration-sign mapping and script-block removal defined over code-point ranges; (iv) lexicon-driven semantic coding and rule-based discourse coding, implementing a transparent, inspectable annotation theory; (v) machine verification, in which output invariants (zero noise tokens, count equality, well-formedness) are checked programmatically; and (vi) audit-trail reporting. The full corpus of 2.8 million raw words is thus processed in minutes on a commodity server.

2.2.3 Relation to existing cleaning tools

It is important to state that general-purpose cleaning tools address adjacent but different problems. OpenRefine, for instance, offers a graphical environment for transforming tabular data (Verborgh and De Wilde, 2013); Trifacta/Wrangler in its capacity pioneered interactive visual transformation specification (Kandel et al., 2011); KNIME provides a modular workflow platform (Berthold et al., 2008); and Pandas supplies a scripting substrate for data frames (McKinney, 2010). None of these, however, encodes a corpus-linguistic cleaning specification, with none distinguishes Arabic script from emojis from URLs by Unicode block, normalises Arabic *hamza/ayn* transliteration signs, or preserves sentence boundaries while stripping punctuation. *Hamza* and *ayn* signs (ﺀ, ﺀ) are therefore normalised to the American Standard Code for Information Interchange (ASCII) apostrophe, so that *Qurʾān*, *bidʿah* and *Shiʿa* are retained as Qur'an, bid'ah and Shi'a (cf. Habash, 2010). Conversely, LLM-based wrangling tools (Narayan et al., 2022; cf. Ilyas and Chu, 2019) are flexible but non-deterministic and opaque, incompatible with the auditability corpus construction demands, and spreadsheet cleaning becomes unmanageable at the scale of thousands of documents (Microsoft, 2023). The Hermes

Agent therefore occupies a middle ground in that it is generative where judgment is needed (planning, documentation, anomaly interpretation), and is fully deterministic where correctness is measurable (all transformations and verification).

2.2.4 Capabilities and limitations

The instrument's principal capabilities are end-to-end execution, complete auditability, domain adaptation through custom lexicons, machine-verified delivery and low operational cost. Despite this, it has some limitations too. First, orchestration decisions are made by an LLM and are not formally specified; the mitigation is the fixed-script pipeline and the mandatory verification gate at every stage. Second, annotation is single-annotator by design, and no inter-annotator agreement study has yet been conducted. In Section 6.5, I report a manual reliability audit, and an agreement study with independent coders as future research plan. Third, the statistical tagger underlying layer one (Section 6) is trained on general English and can mis-tag transliterated Arabic and Nigerian English forms. Despite this, the observed proper-noun share of 13.97% is therefore reported transparently, and Section 6.5 quantifies the tagger's behaviour on this register. Fourth, semantic-lexicon coverage is finite, and 78.6% of tokens remain in the unclassified (O) field. Notably, these limitations are normal in corpus building and have been identified in studies like Leech (1993), Garside, Leech and McEnery (1997), and Artstein and Poesio (2008).

3. An auditable process: documentation and verification at every gate

A recurring observation about automated corpus construction is that tools 'drop' the processes they execute by the time the final resource is delivered, leaving intermediate decisions opaque. The NSDC was built on the opposite principle, making the decision record a first-class deliverable rather than an internal log (Tognini-Bonelli, 2001; McEnery and Hardie, 2012). Every stage writes durable artefacts: cleaned workbooks retain a 'raw_text' column and a dedupe-reason column; report spreadsheets carry per-file, per-year and per-post sheets together with excluded-row sheets; statistics files record every count. The process ledger is therefore complete, from source file to final zip.

The processing protocol applies the same skeleton at every stage: load, transform, record and verify. Loading locates the header by scanning the first rows for column names. Transformation executes the stage-specific algorithm (dedupe, cleaning, merging, annotation). Recording writes the structured artefacts described above. Verification then checks invariants programmatically with zero forbidden characters in corpus bodies, zero duplicate bodies, exact equality between spreadsheet word counts and corpus token counts, well-formed SGML documents, and year-field accuracy against creation timestamps. It ensured that a stage is not closed until verification passes, and the verifier is a separate programme from the transformer, so errors cannot be masked by the code that produced them.

Table 1 below summarises the principal data-quality challenges encountered across the eight exports and the algorithmic responses that resolved them. The list is deliberately concrete, as it constitutes the empirical record of what automated cleaning of social-media religious text actually involves.

Table 1
Selected data-quality challenges and their algorithmic responses.

No	Challenge encountered	Algorithmic solution
1	Header not in row 1; empty first column; dozens of empty trailing columns in Facebook exports	Header located by scanning the first three rows for expected column names; empty columns ignored
2	Metadata-less rows (no creation_time or URL; text spilling into the author cell)	Name and text cells merged; row flagged meta_missing and excluded to a recoverable preview sheet
3	Same post exported repeatedly (up to four times; 84 of 111 rows duplicate in one file)	Three-pass deduplication: by post URL, by normalised exact text, by normalised cleaned text; every drop reason-tagged
4	Raw texts differing only by emoji or spacing escape text-level dedupe	Third pass over cleaned text catches noise-level duplicates (76 in Salafiyyah alone)
5	Year-span files such as 2015–2016 covering June 2015 to June 2016	Year always extracted from creation_time, never from the filename
6	Arabic-script Qur'anic quotes, emojis, pictographs and symbols	Script-block and emoji regexes built from integer code-point tuples; NFC normalisation; Arabic hamza/ayn signs mapped to apostrophe
7	URLs, digits and punctuation corrupting token counts	Regex removal; sentence splitting executed before punctuation removal to preserve one sentence per line
8	Meta Content Library returns no data for some keyword-period combinations	Gaps recorded as data-source facts, not treated as processing losses
9	Same post present under different keyword exports	Merge-stage dedupe on the platform post identifier (last URL segment), not the full URL; 237 duplicates removed
10	Off-by-one body extraction adding a phantom closing tag token to every post	Body extracted between the first <text> line and the last </text> line, never by fixed index
11	Clitic-split token inflation (don't to do + n't)	Reported transparently (+ 1.3%); hyphen-split forms re-merged by re-tokenisation
12	Lexicon pitfalls: forbidden lemmatised as forbid; deontic modals inflating mitigation; haram firing inside Boko Haram	Lemma-form lexicon entries; modals reclassified as call-to-action; multiword phrases consume their tokens before single-token coding

4. Data cleaning: from export to corpus

The starting point of the cleaning stage is the set of eight keyword exports (*Salafis*, *Salafism*, *Salafiyyah*, *Ahlus-Sunnah*, *Sunni*, *Wahabism*, *Wahabiyyah*, *Salafool*), each delivered as ten annual Excel files carrying the standard Content Library schema: 'creation_time', 'mcl_url', 'surface.name' and 'text'. Across all exports, the pipeline received 7,172 rows containing 2,772,120 raw word tokens.

The cleaning proceeds in eight steps. At step 1, each export is unzipped and its schema verified by header scan. At step 2, all rows are loaded and fully empty rows are skipped. At step 3, deduplication runs in three passes, marking every drop with its reason: by mcl_url; by normalised exact text; by normalised cleaned text. At step 4, the publication year is extracted from creation_time. At step 5, each text is cleaned against the specification below, with word counts computed before and after. At step 6, the corpus is written as one UTF-8 file per post, with SGML metadata tags and one sentence per line; posts that become empty after cleaning and rows without metadata are excluded and recorded in preview sheets. Step 7 produces a cleaned workbook per source file, preserving original filenames for traceability. The final step (8) packages the corpus in a downloadable zip folder.

The cleaning specification retains English and Roman-script transliterated Arabic (*Salafi*, *Salafiyyah*, *Ahlus-Sunnah*, *Qur'an*, etc.), normalises *hamza/ayn* signs and curly typography to ASCII, and removes Arabic script, emojis, digits, URLs, non-Latin scripts and punctuation, while preserving apostrophes and hyphens within words. Sentence splitting is performed before punctuation removal so that the one-sentence-per-line convention survives cleaning. Raw word counts count whitespace tokens of the original text; clean counts count tokens after cleaning; the two are distinct definitions, since cleaning can legitimately increase token counts when glued emoji fragments split tokens apart. An example is a post with an Arabic Qur'anic quotation, emojis and a URL reduced to its English and Roman-script remainder.

Figure 1 above reports the overall outcome. The within-keyword cleaning removed 1,126 duplicate rows (84 URL-level, 898 exact-text, 144 cleaned-text), 9 metadata-less rows and 9 posts that became empty after cleaning, yielding 6,028 posts and 2,696,412 clean words. Word retention was 97.3%, indicating that the exports were predominantly English-language with modest script noise. Table 2 reports the same outcomes per keyword.

Table 2
Cleaning outcomes by keyword. Totals exclude 9 metadata-less and 9 empty-after-cleaning rows.

Keyword	Raw rows	Final posts	Raw words	Clean words	Words kept %	Duplicates removed
Sunni	3,865	3,394	1,855,794	1,815,449	97.8	466
Salafiyyah	2,381	1,867	481,537	459,685	95.5	512
Salafis	414	390	246,734	236,052	95.7	24
Salafism	312	189	109,496	107,620	98.3	113
Wahabism	97	90	44,035	43,477	98.7	6
Ahlus-Sunnah	51	48	29,208	28,904	99.0	3
Salafool	28	26	709	677	95.5	2
Wahabiyyah	24	24	4,607	4,548	98.7	0
Total	7,172	6,028	2,772,120	2,696,412	97.3	1,126

The exports are markedly skewed towards the *Sunni* keyword, which contributes 3,394 of the 6,028 cleaned posts (56.3%) and 67.3% of clean words. This reflects both the breadth of the term and the dominance of Sunni-facing pages among the surfaces captured. This imbalance is a property of the discourse sampled rather than a processing artefact, and users should segment or weight by keyword for balanced analyses.

Verification at this gate confirmed zero Arabic-script, emoji, digit, URL or non-Latin tokens in corpus bodies. It also confirmed zero duplicate bodies, and exact equality between workbook clean-word counts and corpus token counts.

5. Corpus integration and diachronic segmentation

The eight keyword corpora were merged into a single resource. The critical design decision concerned the deduplication key where the same post can appear under several keyword exports, but the full export URL differs because its dataset identifier segment is keyword-specific. Deduplication therefore operated on the platform post identifier (the final URL path segment), which is stable across exports. Where a duplicate was found, the first occurrence in alphabetical keyword order was kept. The merge removed 237 cross-keyword duplicates while a secondary pass on (normalised body, author) found none further. The merged NSDC thus comprises 5,791 posts and 2,485,963 words, with each document carrying its source keyword in the metadata tag.

For diachronic analysis (objective iv), the corpus is segmented by publication year, producing eleven calendar-year bins (2015 to 2025) within the ten-year frame. Because the annual export windows run from June to June (2015–2016, 2016–2017 and so on), calendar segmentation is the correct procedure for year-level trend analysis, and a single window contributes posts to two adjacent bins. Figure 2 below shows the resulting distribution, which is far from uniform. This shows activity rises steeply from 2015 to a 2017 peak of 753 posts, recedes, then rises again to 715 posts in 2024 before falling to 450 in 2025. Table 3 too pairs post counts with tagged-token counts, and Fig. 3 displays the token series, which peaks in 2016 at 470,713 tokens and declines across the decade to 74,656 in 2025. This divergence between the two series in later years indicates that recent posts are shorter, which is a meaningful diachronic finding about discourse compression.

Table 3
Year-by-year distribution of posts and tagged tokens.

Year	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	Total
Posts	224	635	753	474	439	429	448	534	690	715	450	5,791
Tagged tokens	162,826	470,713	410,501	277,180	226,438	235,991	199,074	178,509	148,208	134,357	74,656	2,518,453

6. Annotation: three layers of linguistic enrichment

Raw corpora are often inert where grammatical, semantic and discourse-pragmatic information is latent in the text until made explicit. For corpus linguistics, annotation therefore converts a word stream into a structured, queryable resource, making patterns measurable and analysis rigorous and replicable (McEnery and Hardie, 2012). The NSDC is therefore annotated on three layers, discussed below in 6.3.

6.1 The decision against external taggers

Although the research design that informed the NSDC envisaged TreeTagger (Schmid, 1994) for part-of-speech tagging and Sketch Engine (Kilgarriff et al., 2014) for lemmatisation, and the corpus format remains natively compatible with both (SGML metadata tags, one sentence per line, plain UTF-8), the annotation was subsequently performed within the same environment. There were four reasons to this decision. First, a single environment

guarantees reproducibility across the entire pipeline. Second, the required layers (semantic-field assignment from a domain lexicon and discourse-strategy coding under the discourse-historical approach (Reisigl and Wodak, 2001)) are provided by no off-the-shelf tagger. Third, the corpus volume is modest enough that local processing takes minutes. Fourth, the vertical output format preserves Sketch Engine compatibility, so the decision costs nothing downstream.

6.2 The problem of tagging religious social-media text

Religious discourse on Nigerian Facebook is a demanding tagging environment. It mixes English with Arabic, Roman-script transliterated Arabic and even Nigerian languages like Yoruba and Hausa, with orthographic variation, dense strings of Arabic names and honorifics, and frequent code-switched quotation. Generic taggers trained on edited prose such as the TreeTagger (Schmid, 1994), CLAWS (Garside, 1987), the Stanford tagger (Toutanova et al., 2003), NLTK's averaged perceptron (Bird, Klein and Loper, 2009) would treat such input as noise or mis-tag it, and none can assign discourse-pragmatic categories. Manual environments such as UAM CorpusTool (O'Donnell, 2008) might also solve the interface problem but would definitely leave the coding and consistency burden on the analyst. Same way, LLM-based annotation is competitive with crowd workers for semantic classification (see, for example, Gilardi, Alizadeh and Kubli, 2023; Tornberg, 2023; Ziemis et al., 2024), but token-level tagging at 2.5 million tokens remains costly and stochastic outputs complicate auditability.

6.3 The annotation architecture

For the annotation, layer one attaches to every token a Penn Treebank part-of-speech tag (Marcus, Santorini and Marcinkiewicz, 1993) and a lemma, using the spaCy statistical model 'en_core_web_md' (Honnibal and Montani, 2017) with the parser and named-entity recogniser disabled. Hyphen-split forms (e.g. Ahlus-Sunnah) are re-merged into single tokens, and this is essential for lemma integrity in transliterated names. Layer two assigns each token to one of fifteen semantic fields (creed, worship practice, religious authority, sect groups, conflict and violence, evaluation, intensification, hedge, and so on) through a purpose-built lexicon; multiword phrases are matched first and their tokens consumed, so that Boko Haram does not trigger a religious-legitimation code on *haram*. Layer three codes each sentence for up to fifteen discourse strategies mapped to the discourse-historical framework: nomination (NOM_MUSLIM, NOM_SALAFI, NOM_SUFI, NOM_SHIA, NOM_SECT), predication (PRED_POS, PRED_NEG), argumentation (ARG_CAUSE, ARG_SCRIPTURE), legitimation (LEGIT_RELIGIOUS, LEGIT_LEGAL), perspectivisation (PERSP_REPORTED), intensification (INTENSIFY), mitigation (MITIGATE) and call-to-action (CALL_ACTION). The mapping from objectives to layers is explicit as POS and lemma serve objectives (i) and (iv); semantic fields serve (i) and (iii); and strategy coding serves (ii) and (iii).

6.4 Results

The tagged corpus comprises 5,791 posts, 149,431 sentences and 2,518,453 tokens, with 57,200 lemma types. Tagged tokens exceed orthographic words by 1.3%, the expected clitic-split inflation of don't to do + n't. Proper nouns account for 13.97% of tokens, a direct measure of the transliterated-Arabic and naming density of the register, subject to the tagger caveats quantified in Section 6.5. Table 4 below summarises the dominant POS categories, Table 5 the semantic fields, Table 6 the strategy codes by post coverage while Figs. 4–6 visualise the same distributions.

Table 4
Dominant part-of-speech categories (Penn Treebank).

POS tag	Count	% of tokens	POS tag	Count	% of tokens
IN (preposition)	352,644	14.0	PRP (pronoun)	109,309	4.3
NNP (proper noun)	351,768	14.0	RB (adverb)	103,234	4.1
NN (noun)	314,152	12.5	CC (conjunction)	99,091	3.9
DT (determiner)	254,471	10.1	VB (verb)	90,213	3.6
JJ (adjective)	159,106	6.3	VBD (past verb)	80,344	3.2
NNS (plural noun)	129,317	5.1	VBZ (3sg verb)	68,524	2.7

Table 5
Semantic-field assignment in the tagged corpus (15 fields; O excluded from Fig. 5).

Semantic field	Tokens	Semantic field	Tokens
PERSON_REFERENCE	167,122	AUTHORITY_SCHOLARS	19,381
AUTHORITY_POLITICS	53,566	CREED_SALAFI	17,685
RELIGION_GENERAL	47,341	MEDIA_COMMUNICATION	17,059
TIME_DIACHRONIC	44,335	SECT_GROUPS	14,478
INTENSIFIER	34,182	EVALUATION_NEG	13,637
DISCOURSE_VERBS	31,928	WORSHIP_PRACTICE	8,164
CONFLICT_VIOLENCE	27,484	HEDGE	22,558
EVALUATION_POS	20,538	O (unclassified)	1,978,995

Table 6
Discourse-strategy codes by post and sentence coverage.

Strategy	Posts	% posts	Sentences	Strategy	Posts	% posts	Sentences
NOM_MUSLIM	4,309	74.4	18,496	LEGIT_LEGAL	1,976	34.1	9,235
CALL_ACTION	3,936	68.0	29,339	NOM_SECT	1,663	28.7	4,237
INTENSIFY	3,688	63.7	28,781	PRED_POS	1,503	26.0	2,758
MITIGATE	3,383	58.4	21,078	PRED_NEG	1,304	22.5	2,459
ARG_CAUSE	3,197	55.2	19,397	ARG_SCRIPTURE	1,019	17.6	3,403
PERSP_REPORTED	3,045	52.6	18,039	NOM_SHIA	955	16.5	3,503
LEGIT_RELIGIOUS	2,737	47.3	12,421	NOM_SUFI	416	7.2	1,194
NOM_SALAFI	2,353	40.6	5,331				

It should be noted that the distributional profile is linguistically plausible as high coverage of nomination and perspectivisation is what one expects of persuasive religious discourse on social media, while the low coverage of scripture argumentation (17.6% of posts) and predication (26.0% and 22.5%) delimits where theological citation actually occurs. The high apparent coverage of call-to-action (68.0%) and intensification (63.7%) should, however, be read cautiously in that the reliability audit in Section 6.5 found these two codes over-generated.

The verification of the tagged corpus passed on all checks with zero malformed tags, zero unbalanced sentences, zero empty cells, zero unknown codes, and exact equality between tagged counts and the merged corpus. The annotated output uses the vertical format of the Sketch Engine and CQP query families, so the tagged layers can be keyed, collocated and compared with standard tools (Gries, 2009; Kilgariff et al., 2014). See sample output below:

```
<doc year="2016" author="Ridwan Lukman Isiaka" url=".../post/1469801934449319/" keyword="Sunni">
```

```
<s>
```

```
Sufism NN Sufism SECT_GROUPS
```

```
Purifying VBG purify CREED_SALAFI
```

```
the DT the O
```

```
Soul NNP soul RELIGION_GENERAL
```

```
</s>
```

```
<strategy NOM_SUFI/>
```

```
</doc>
```

6.5 Reliability and validation

Because the annotation layers are produced by deterministic programmes rather than human coders, their reliability was assessed through a manual audit of stratified random samples from the tagged corpus, following the definitions of Section 6.3.

First, for the semantic-field layer, all 40 inspected non-O assignments (100%) were judged appropriate for the word and its context (e.g., security → AUTHORITY_POLITICS, lying → EVALUATION_NEG), showing the lexicon-driven coding is accurate for the tokens it classifies.

Second, for the part-of-speech layer, 50 NNP-tagged tokens were inspected. This found that 32 (64%) were genuine proper names or name-like religious terms and honorifics (e.g., Usman, El-Zakzaky, Iran, al-Fawzān, SWT); the remainder were titles (Major, Officer, Sheikh), common nouns and Arabic forms mis-tagged by the general-English model. This quantifies, at sample level, the tagger-bias limitation of Section 2.2.4, and is why the NNP share of 13.97% is best read as an upper bound of genuine naming density.

Third, for the discourse-strategy layer, 55 sentences (five per calendar year) were re-coded independently by the researcher and compared with the rule-based output. The automatic codes matched the manual codes exactly in 20 of 55 sentences (36.4%), with a precision of 24.6% and a recall of 27.3% at the level of individual codes. This disagreement is of course systematic as the automatic codes over-generate CALL_ACTION, INTENSIFY and MITIGATE, often on short fragments and boilerplate (e.g., 'Read More' attracts CALL_ACTION automatically but none under manual reading), and they under-detect NOM_MUSLIM, NOM_SHIA, PERSP_REPORTED and ARG_SCRIPTURE which are the codes triggered by reporting verbs ('said', 'explained'), by explicit group nouns ('Shiites'/'Shias' are missed, and the geopolitical referent 'the north' is wrongly tagged NOM_MUSLIM) and by scriptural citation. Two codes, NOM_SALAFI and PRED_NEG, showed high precision (100%) but low recall. Part of the disagreement reflects genuinely borderline judgments, but the patterns are consistent.

As such, these findings have three consequences. First, the semantic-field and lemma layers can support quantitative queries as released. Second, the discourse-strategy layer should currently serve qualitative sampling and hypothesis generation rather than unsupervised quantitative inference. The strategy percentages in Table 6 above are accordingly indicative of this. Third, a refined coding scheme (version 1.1) that tightens the triggers for CALL_ACTION, INTENSIFY and MITIGATE and extends the nomination and reporting lexicons, an inter-annotator agreement study with independent coders, and a gold-standard evaluation are planned.

7. Research applications of the NSDC

The NSDC was built for the four objectives of Section 1, but its design makes it a general-purpose resource for corpus-assisted discourse research on religion, politics and society in Nigeria. All suggestions below are executable with the released artefacts (the cleaned corpus, the tagged vertical corpus and the year segments). Also, because every document retains its metadata (year, author, source keyword), each focus can be restricted to sub-corpora such as a single denomination, year or page cohort (Cf. Hunston, 2002). The discourse-strategy codes should be used with the reliability caveats of Section 6.5 in mind in which at present they are best suited to qualitative sampling and hypothesis generation.

7.1 Stereotypes and representation

The representation of Muslims and of Islam in media and public discourse is a well-established domain of CADS (Baker, Gabrielatos and McEnery, 2013; Richardson, 2004; Ahmed and Matthes, 2017). The NSDC inverts the usual lens by capturing how Muslims represent one another rather than how the majority media represent them. A researcher can extract the concordances of Wahhabi, Salafi and Sufi, compute their collocates with Sketch Engine or CQP, and ask whether Wahhabi functions as a pejorative stereotype in Nigerian usage or as a neutral label. The EVALUATION_NEG and EVALUATION_POS fields provide a ready-made sentiment dimension, letting stereotyping be quantified as the co-occurrence of out-group nominals (NOM_SUFI, NOM_SECT) with negative evaluation across the decade.

7.2 Hegemony and ideological struggle

Gramscian hegemony denotes the winning of consent through discourse rather than coercion (Gramsci, 1971; Laclau and Mouffe, 1985), and van Dijk (1998) has shown how ideologies are reproduced through positive self-presentation and negative other-presentation. The NSDC operationalises this directly where the nomination and predication codes instantiate in-group and out-group construction, and the legitimisation codes (LEGIT_RELIGIOUS, LEGIT_LEGAL) reveal the grounds on which authority is claimed. Therefore, a study could model which strategy legitimises Salafi claims relative to Sufi or Shia counter-claims, or examine how Ahlus-Sunnah is deployed as a hegemonic self-designation annexing the centre of Sunni identity. In another dimension, co-occurrence analysis of NOM_SALAFI with LEGIT_* and NOM_SECT with PRED_NEG would quantify the discursive balance of each constituency.

7.3 Religion and theology in discourse

For researchers in language and religion, the corpus offers a window onto how theological concepts circulate in vernacular discourse. The CREED_SALAFI and WORSHIP_PRACTICE fields delimit creedal and ritual vocabulary (bid'ah, tawhid, mawlid, innovation), and ARG_SCRIPTURE flags sentences in which scripture is invoked as evidence. And because its coverage is modest (17.6% of posts), its distribution is itself a potential finding. A researcher could also examine the concordance of the lemma 'innovation' to study how *bid'ah* is explained, warned against and ridiculed, or trace the *mawlid* debate (the annual controversy over the Prophet's birthday) across years. Useful approaches to these are the Fairclough's three-dimensional model (Fairclough, 2013) and the discourse-historical approach (Wodak and Meyer, 2016).

7.4 Security, conflict and securitisation

Northern Nigeria's religious context has been profoundly shaped by the Boko Haram insurgency, and Salafi discourse is routinely entangled with security talk (Mustapha, 2014; Thurston, 2016). The CONFLICT_VIOLENCE field allows securitisation to be studied. Following Graham, Keenan and Dowd (2004), a researcher can identify how religious actors are lexically and argumentatively linked to, or dissociated from, violence. The multiword handling of Boko Haram in the NSDC treated as a proper-name unit so that haram does not fire a religious-legitimation code particularly keeps security discourse distinct from theology in the annotation.

7.5 Communication and persuasion

Similarly, the corpus is rich in persuasive and rhetorical posts where CALL_ACTION appears in 68.0% of posts and INTENSIFY in 63.7%, although both figures are inflated by the over-generation documented in Section 6.5. It remains a strong resource for digital religious communication, *da'wah* (preaching) rhetoric and online persuasion (see Charteris-Black, 2004; Partington, Duguid and Taylor, 2013). A researcher can analyse how intensifiers (VERILY, TRULY), deontic modals (must, should) and direct address (you, O believers) build exhortation. Another can compare mitigation and intensification across topics or ask whether CALL_ACTION shifted from theological exhortation to political mobilisation over the decade.

7.6 Naming, identity and group boundaries

Naming is where social cognition becomes language (Suleiman, 2011), and the NOM_* codes make naming directly queryable. Same way, social identity theory predicts that in-groups are named richly and positively while out-groups are named generically or pejoratively (Tajfel and Turner, 1979). On this note, a researcher can examine which denominations are referred to by self-chosen names (Ahlus-Sunnah, Salafi) versus ascribed ones (Wahhabi, Shia, Sufi), measure the density of NOM_MUSLIM (74.4% of posts) as the default religious identity, and study the grammar of exclusion; that is which verbs and adjectives co-occur with NOM_SECT, and whether the sect label collocates with terms for deviance, innovation or disbelief. In fact, the shifting frequency of *Wahhabi* versus *Salafi* across 2015–2025 could track the evolution of a pejorative into a contested label, which is of direct relevance to researchers interested in sociolinguistics and onomastics.

7.7 Loyalty, authority and allegiance

Finally, the corpus speaks to loyalty and authority, two themes central to Salafi discourse through *al-wala' wal-bara'* (loyalty and disavowal) and institutionalised deference to scholars. The AUTHORITY_SCHOLARS and AUTHORITY_POLITICS fields, together with PERSP_REPORTED (52.6% of posts), enable the study of reported speech as an authority device. One can see whose words are quoted, how quotative frames attribute or distance, and how loyalty to scholars versus the state is managed (see Wodak, 2015; Kendhammer, 2016). A study could again investigate quotative verbs such as 'said' collocating with scholar names, or trace allegiance shifts in the post-2020 discourse around elections and the state, where LEGIT_LEGAL coverage (34.1% of posts) marks the intrusion of legal-rational legitimation into a primarily religious discourse, asking whether PERSP_REPORTED rises in election years.

8. Conclusion

The NSDC demonstrates an end-to-end, auditable and machine-verified method for constructing a research corpus from social-media exports of religious discourse. Its substantive contribution, importantly, is a decade-long, 2.4-million-word and 2.5-million-token record of how Salafism and Salafis are talked about in Nigerian Facebook discourse, annotated at the levels of grammar, semantics and discourse strategy in a single coherent scheme. Methodologically, it shows that an agentic AI instrument can execute corpus construction to standards of transparency and verifiability comparable to those of conventional pipelines, provided that transformations are deterministic, processes are recorded and verification gates are mandatory. The reliability audit (Section 6.5) indicates that the lexical layers are fit for quantitative use, while the discourse-strategy layer requires refinement as future work comprises an inter-annotator agreement study, a gold-standard evaluation of the strategy codes, an expanded semantic lexicon and a refined coding scheme (version 1.1). The NSDC is offered as a resource for corpus-assisted discourse studies of religion, identity and politics in Africa.

Declarations

Author Contribution

While the authorship of the NSDC is solely for the first author, Idris Ayomo Oke, the overall development of this manuscript belongs to the two authors.

Acknowledgement

We would like to thank Uthman Zakariyah, Esq. whose technical skills and guide helped in building the NSDC reported here. Thanks so much for not getting tired of the several calls and messages.

Data Availability

The NSDC (cleaned corpus, tagged vertical corpus, year segments and pipeline scripts) is available at <https://doi.org/10.5281/zenodo.22165298> (CC BY 4.0).

References

1. Ahmed, S., & Matthes, J. (2017). Media representation of Muslims and Islam from 2000 to 2015: A meta-analysis. *International Communication Gazette*, 79(3), 219–244.
2. Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596.
3. Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.
4. Baker, P., Gabrielatos, C., & McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge University Press.
5. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kotter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, T. (2008). KNIME: The Konstanz Information Miner. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data analysis, machine learning and applications* (pp. 319–326). Springer.
6. Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
7. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.
8. Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis*. Palgrave Macmillan.
9. Fairclough, N. (2013). *Critical discourse analysis: The critical study of language* (2nd ed.). Routledge.
10. Franzke, A. S., Bechmann, A., Zimmer, M., Ess, C., & Association of Internet Researchers. (2020). *Internet research: Ethical guidelines 3.0*. Association of Internet Researchers. <https://aoir.org/reports/ethics3.pdf>
11. Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech, & G. Sampson (Eds.), *The computational analysis of English* (pp. 30–41). Longman.
12. Garside, R., Leech, G., & McEnery, A. (Eds.). (1997). *Corpus annotation: Linguistic information from computer text corpora*. Longman.
13. Gatto, M. (2014). *The web as corpus: Theory and practice*. Bloomsbury.
14. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120.
15. Graham, P., Keenan, T., & Dowd, A. M. (2004). A call to arms at the end of history: A discourse-historical analysis of George W. Bush's declaration of war on terror. *Discourse & Society*, 15(2–3), 199–221.
16. Gramsci, A. (1971). *Selections from the prison notebooks*. Lawrence & Wishart.
17. Gries, S. T. (2009). *Quantitative corpus linguistics with R: A practical introduction*. Routledge.
18. Habash, N. Y. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool.
19. Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. <https://spacy.io>
20. Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
21. Ilyas, I. F., & Chu, X. (2019). *Data cleaning*. ACM Books.
22. Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011). Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)* (pp. 3363–3372). ACM.
23. Kendhammer, B. (2016). *Muslims talking politics: Framing Islam, democracy, and law in northern Nigeria*. University of Chicago Press.
24. Kilgarriff, A., Baisa, V., Busta, J., Jakubicek, M., Kovar, V., Michelfeit, J., Rychly, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36.
25. Laclau, E., & Mouffe, C. (1985). *Hegemony and socialist strategy: Towards a radical democratic politics*. Verso.
26. Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing*, 8(4), 275–281.
27. Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313–330.
28. McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
29. McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 56–61).
30. Meta (2026). *Meta Content Library*. <https://transparency.meta.com/researchtools/>
31. Microsoft (2023). *Microsoft Excel*. <https://www.microsoft.com/microsoft-365/excel>
32. Mustapha, A. R. (Ed.). (2014). *Sects and social disorder: Muslim identities and conflict in northern Nigeria*. James Currey.
33. Narayan, A., Chami, I., Orr, L., & Re, C. (2022). Can foundation models wrangle your data? *Proceedings of the VLDB Endowment*, 16(4), 738–746.

34. O'Donnell, M. (2008). Demonstration of the UAM CorpusTool for text and image annotation. In *Proceedings of the ACL-08: HLT demo session* (pp. 13–16). Association for Computational Linguistics.
35. Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies*. John Benjamins.
36. Reisigl, M., & Wodak, R. (2001). *Discourse and discrimination: Rhetorics of racism and antisemitism*. Routledge.
37. Richardson, J. E. (2004). *(Mis)representing Islam: The racism and rhetoric of British broadsheet newspapers*. John Benjamins.
38. Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing* (pp. 44–49).
39. Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell.
40. Suleiman, Y. (2011). *Arabic, self and identity: A study in conflict and displacement*. Oxford University Press.
41. Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin, & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole.
42. Thurston, A. (2016). *Salafism in Nigeria: Islam, preaching, and politics*. Cambridge University Press.
43. Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.
44. Tornberg, P. (2023). ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. *arXiv*. <https://arxiv.org/abs/2304.06588>
45. Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 252–259).
46. Townsend, L., & Wallace, C. (2017). The ethics of using social media data in research: A new framework. In K. Woodfield (Ed.), *The ethics of online research* (Vol. 2, pp. 189–207). Emerald Publishing. (Advances in Research Ethics and Integrity).
47. van Dijk, T. A. (1998). *Ideology: A multidisciplinary approach*. Sage.
48. Verborgh, R., & De Wilde, M. (2013). *Using OpenRefine*. Packt Publishing.
49. Wodak, R. (2015). *The politics of fear: What right-wing populist discourses mean*. Sage.
50. Wodak, R., & Meyer, M. (Eds.). (2016). *Methods of critical discourse studies* (3rd ed.). Sage.
51. Ziems, C., Held, W., Yang, O., Dhamala, J., Gupta, R., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1), 237–291.
52. Zimmer, M. (2010). But the data is already public: On the ethics of research in Facebook. *Ethics and Information Technology*, 12(4), 313–325.

Footnotes

1. Salafism (from the Arabic *Salafiyyah*, that is, adherence to the *salaf*, the pious predecessors) is a Sunni Islamic ideology that strictly upholds, defends and advocates those Islamic practices which correlate with the way of the Prophet Muhammad (peace be upon him), his companions and the two generations that succeeded them.

Figures

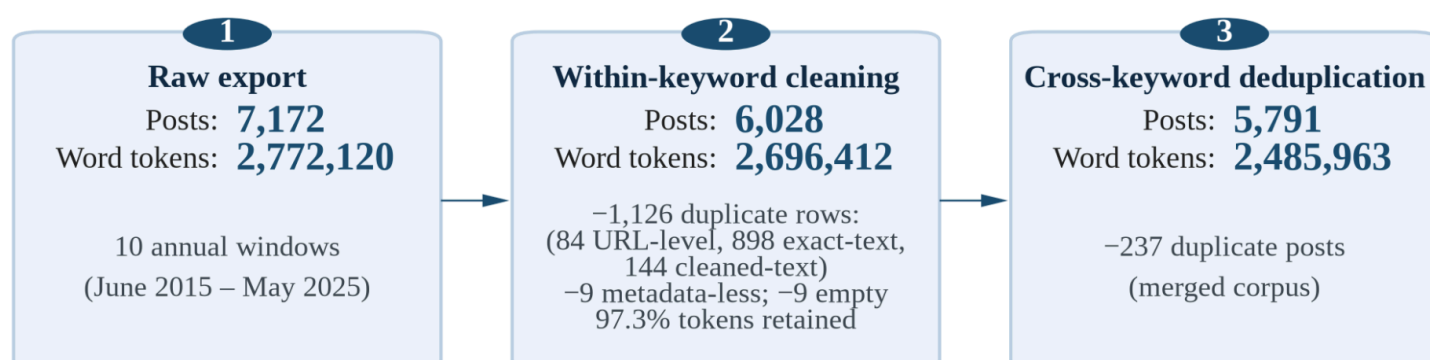


Figure 1

Posts and word tokens retained through the cleaning funnel (7,172 → 5,791 posts; 2,772,120 → 2,485,963 tokens).

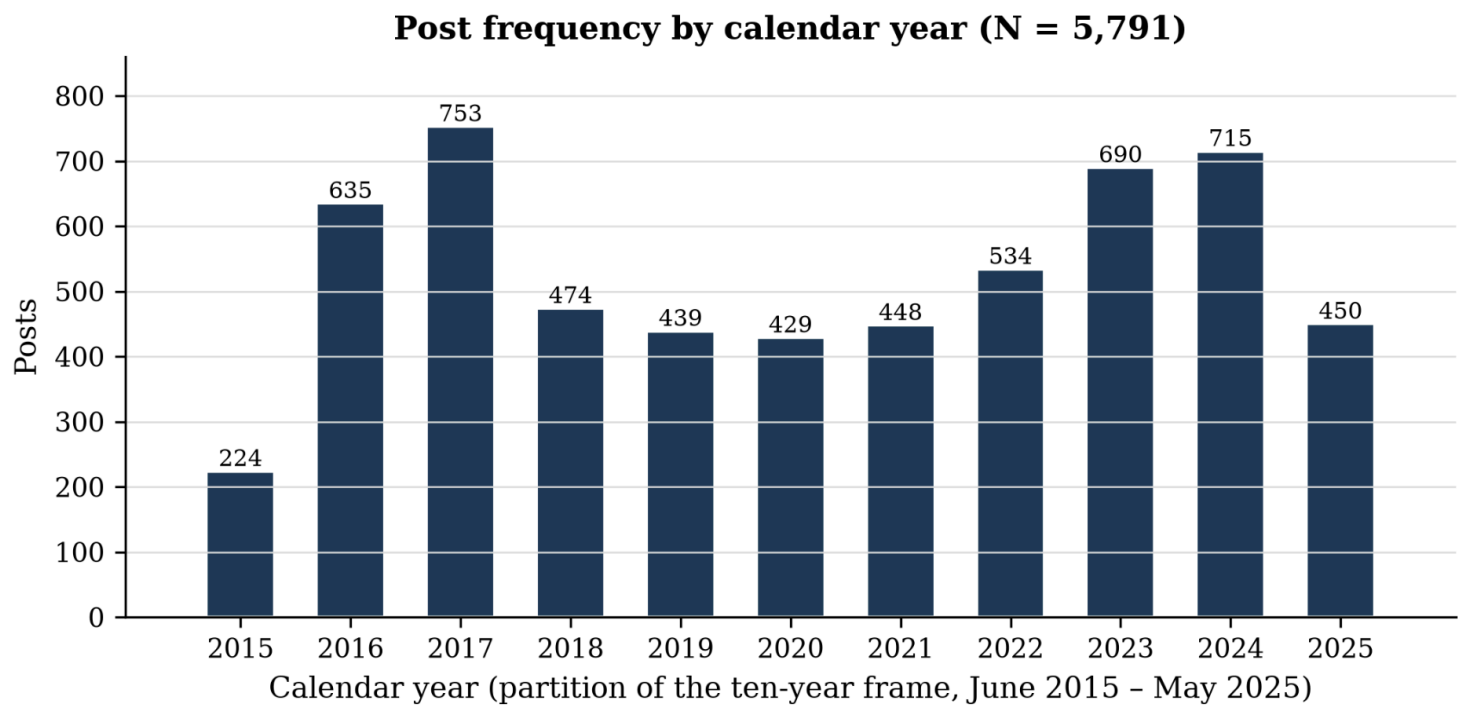


Figure 2

Posts per calendar year within the ten-year frame (N = 5,791).

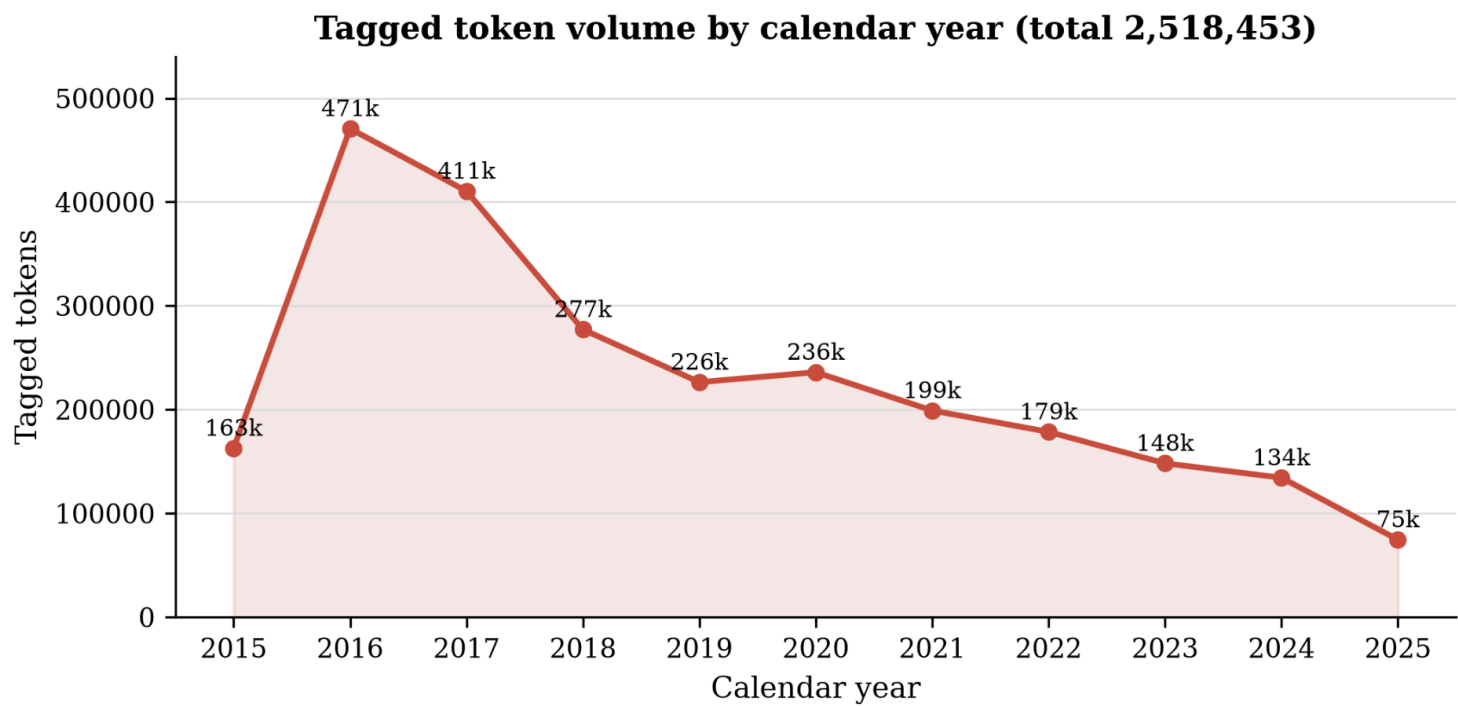


Figure 3

Tagged-token volume per calendar year.

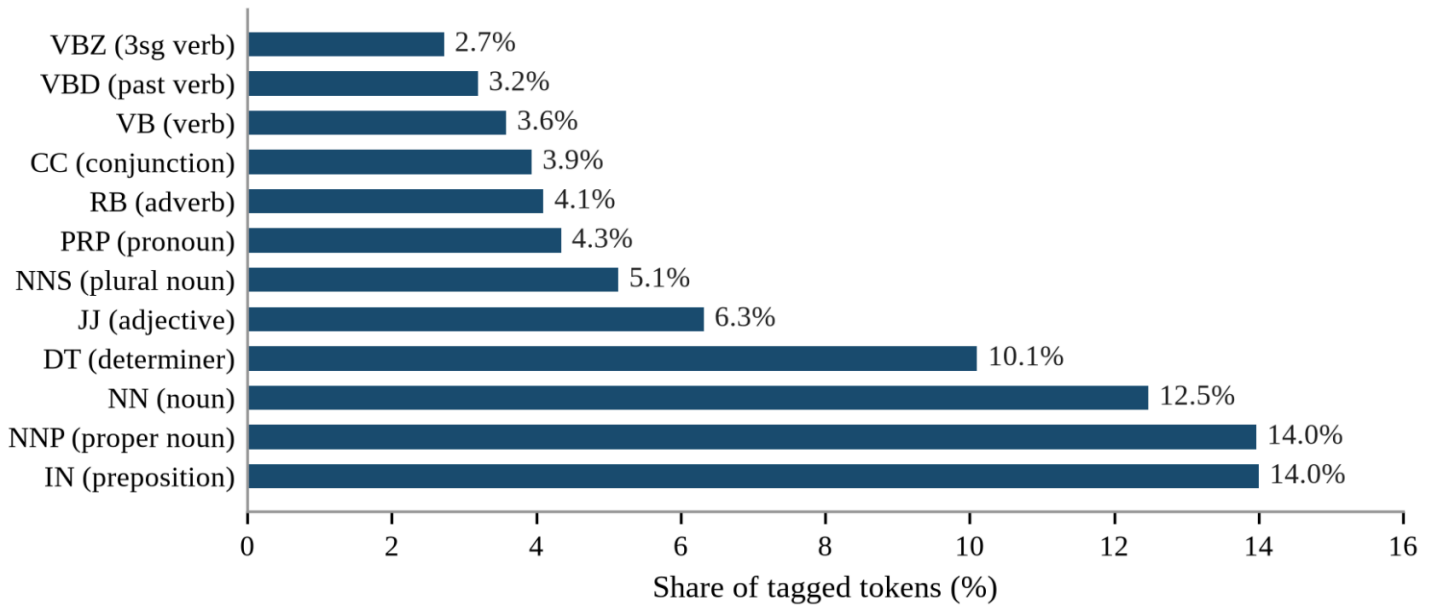


Figure 4

Dominant part-of-speech categories and their shares of tagged tokens.

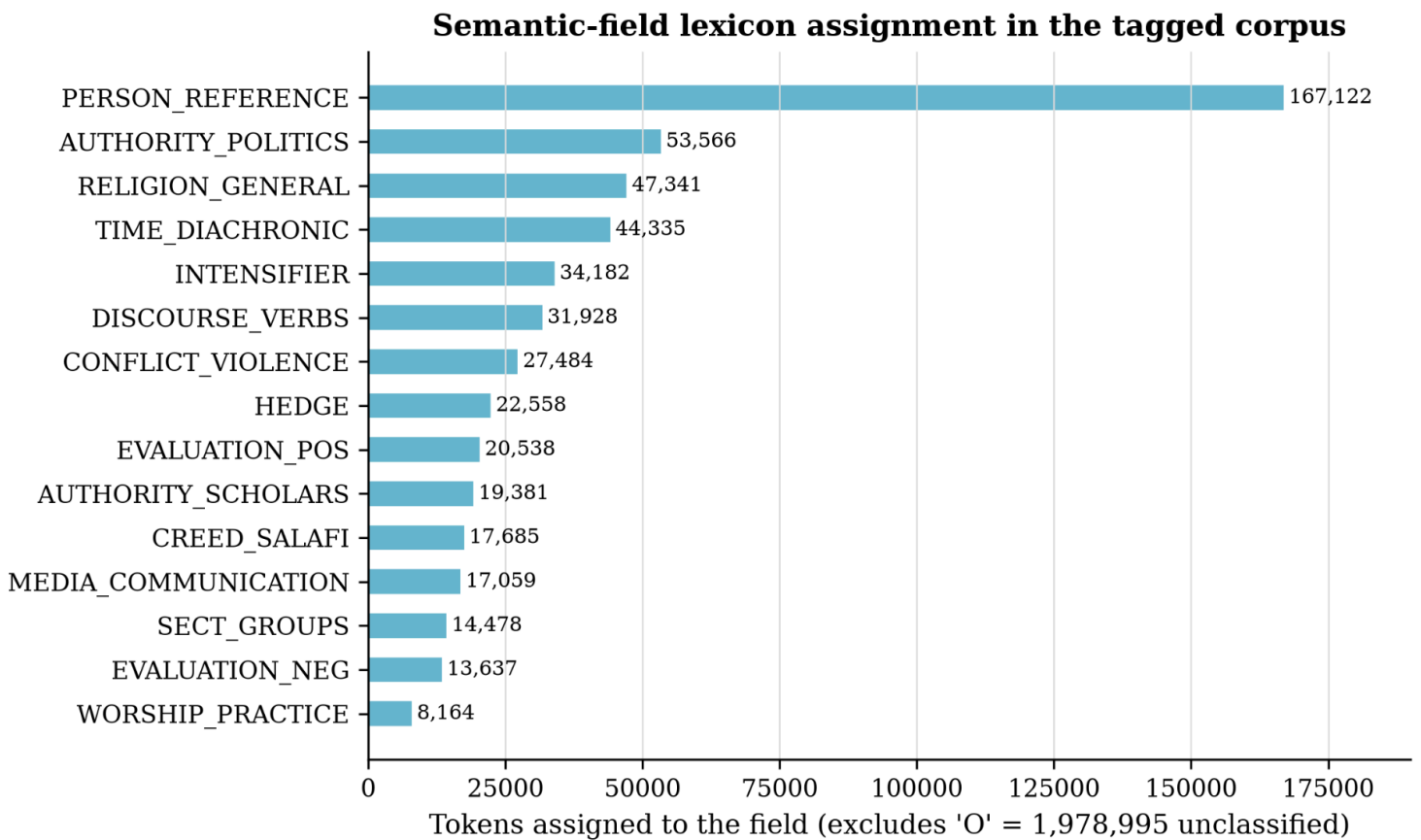


Figure 5

Semantic-field lexicon assignment, top fields by token count.

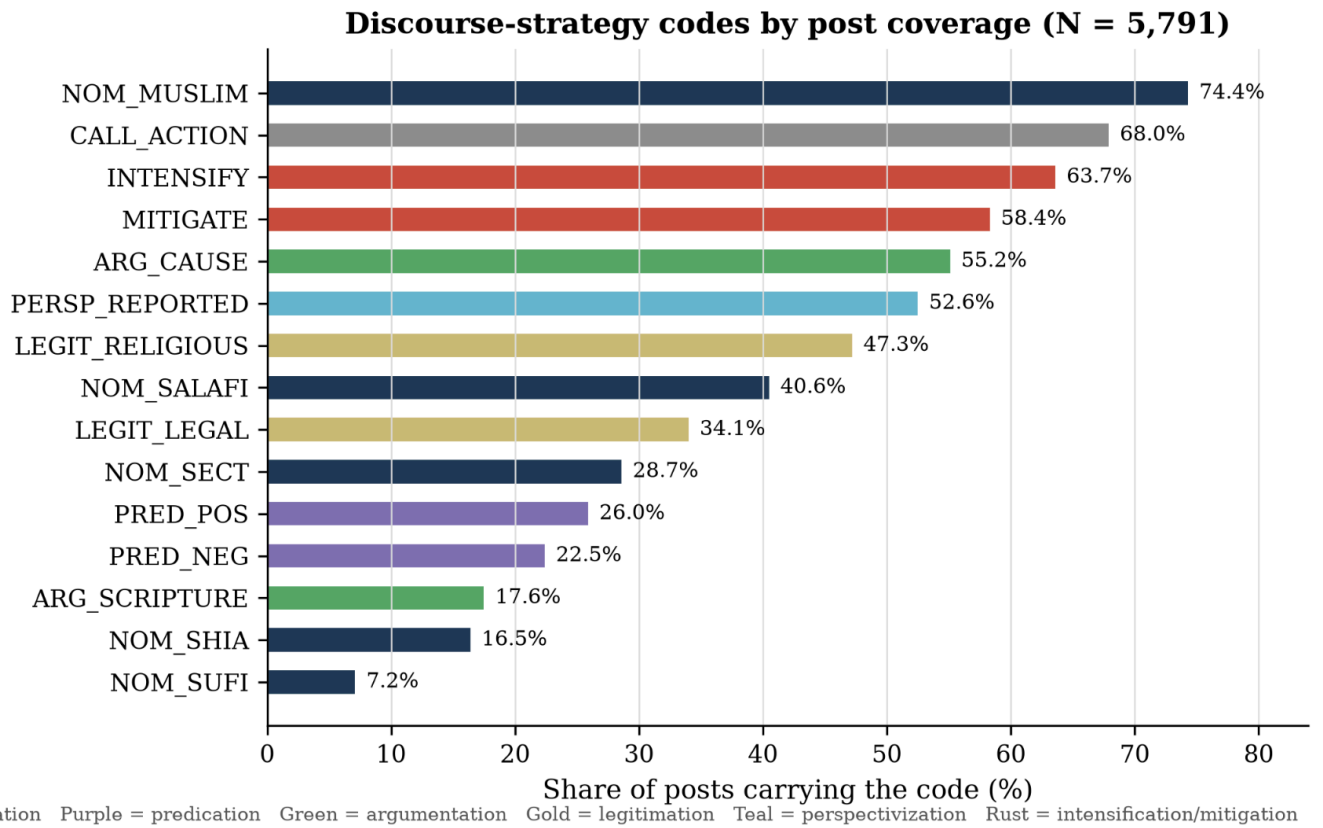


Figure 6

Discourse-strategy codes by share of posts.