

RESEARCH ARTICLE OPEN ACCESS

Building A Digital Corpus Architecture for War Terminology Research: Extracting Opinion and Action Narratives from Ukrainian News Media

Aiko Nakamura

Department of Linguistics and Cultural Studies, Kyoto International Institute of Humanities, Japan

Haruto Fujisawa

Faculty of Letters and Philological Research, Tokyo Metropolitan University of Humanities, Japan

Received: 09 June 2026 Accepted: 05 July 2026 Published: 01 August 2026

ABSTRACT

The rapid expansion of digital news ecosystems has transformed the study of conflict discourse by creating unprecedented volumes of textual data that reflect political positions, public attitudes, and strategic narratives. This study develops a digital corpus architecture for investigating war-related terminology and narrative structures in Ukrainian news media, focusing on the extraction and analysis of opinion and action narratives. The research integrates corpus linguistics principles, web-based data collection approaches, and computational text analysis techniques to construct a systematic framework for identifying lexical patterns associated with conflict representation. Drawing upon corpus-driven methodologies, the study examines how war terminology functions as a linguistic mechanism for framing military events, political decisions, humanitarian concerns, and social responses. The proposed architecture combines automated data acquisition, corpus annotation, concordance analysis, collocation measurement, and multidimensional interpretation to analyze large-scale news discourse. The theoretical foundation is informed by corpus-based discourse analysis, where context and pragmatic functions determine the meaning of linguistic choices (Adolphs, 2008). The findings demonstrate that a structured digital corpus enables researchers to identify recurring war lexicons, ideological positioning, and narrative transformations across news sources. The research contributes a methodological framework for digital humanities, media linguistics, and conflict discourse studies by demonstrating how computational corpus approaches can reveal patterns of meaning construction within contemporary war reporting.

Keywords: Digital corpus architecture; War terminology; Corpus linguistics; Ukrainian news media; Web scraping; Discourse analysis; Conflict narratives; Lexical analysis; Computational linguistics.

1. INTRODUCTION**1.1 Background**

The relationship between language and conflict has become increasingly significant in contemporary communication environments, where wars are not only

military events but also discursive phenomena shaped through media representation. News platforms continuously produce textual narratives that classify actors, define actions, construct moral positions, and influence public interpretation. In this context, war-related

vocabulary becomes a central analytical object because lexical choices determine how events are categorized and understood. Terms associated with military operations, political responses, humanitarian consequences, and international cooperation do not merely describe reality; they actively participate in the construction of conflict narratives.

Corpus linguistics provides an effective methodological foundation for examining such linguistic patterns because it enables researchers to analyze extensive collections of authentic texts through systematic computational methods. Corpus-based approaches allow the investigation of frequency, contextual variation, collocation behavior, and semantic relationships within large datasets. Adolphs (2008) emphasizes that corpus analysis must consider the relationship between linguistic forms and contextual functions, demonstrating that meaning emerges through patterns of usage rather than isolated words. This perspective is particularly relevant for war discourse research, where identical lexical items may carry different ideological meanings depending on their journalistic context.

The emergence of digital journalism has created new possibilities for building specialized corpora focused on contemporary conflicts. Online news articles contain rapidly changing terminology reflecting battlefield developments, diplomatic negotiations, international reactions, and public opinion. However, the scale and complexity of digital information require structured approaches for collecting, organizing, and analyzing textual materials. Digital corpus architecture addresses this challenge by integrating automated data collection, linguistic annotation, and computational analysis into a unified research framework.

The Russian-Ukrainian conflict provides a significant case for studying war terminology because its media representation involves multiple narrative dimensions. News reports discuss military strategies, territorial developments, international assistance, political leadership, migration, and humanitarian consequences. Sources such as El Mundo, El País, Deutsche Welle, and Ukrainian news platforms demonstrate how different media environments construct narratives through specific lexical selections (Andrés & Velasco, 2025; Bassets, 2025; Kinkartz, 2025b; Editorial, 2025). A digital corpus approach allows these linguistic choices to be examined systematically rather than through limited qualitative

observations.

1.2 Problem Statement

Despite increasing availability of online news data, the systematic study of war terminology remains challenged by several methodological limitations. Traditional discourse analysis often relies on manually selected texts, which may restrict the ability to identify large-scale linguistic tendencies. Furthermore, rapidly evolving conflicts generate new terminology and changing narrative structures that require flexible analytical systems.

Existing corpus tools provide valuable capabilities for linguistic investigation, but researchers require specialized architectures adapted to conflict discourse. General-purpose corpus systems may identify frequency patterns and concordances, yet they may not sufficiently address the semantic complexity of war narratives, where terms frequently contain political, emotional, and ideological dimensions.

Another significant challenge involves connecting extracted lexical patterns with broader narrative functions. The appearance of terms such as “military support,” “territorial defense,” “refugees,” or “weapons assistance” cannot be interpreted only through frequency statistics. These expressions operate within larger discourse structures that represent responsibility, legitimacy, threat, and action. Therefore, a research framework is needed that combines computational extraction with contextual interpretation.

1.3 Research Relevance

The development of a digital corpus architecture for war terminology research is relevant because contemporary conflicts increasingly unfold within digital information environments. News media serve as major platforms where political meanings are negotiated and public understanding is shaped. By examining language patterns systematically, researchers can better understand how conflicts are represented and how narratives evolve over time.

Corpus-based approaches also contribute to methodological advancement in digital humanities. The integration of web scraping, corpus management systems, and linguistic analysis tools enables researchers to process large-scale textual resources efficiently. Automated collection methods provide opportunities for constructing

dynamic corpora capable of capturing ongoing discourse changes (Automated data collection: Web scraping example, part 1, 2017; Gupta, 2024).

The research is further supported by developments in corpus analysis software. Tools such as AntConc and Sketch Engine demonstrate the importance of computational environments for examining lexical relationships, word frequencies, and contextual patterns (Anthony, 2005; Kilgarriff et al., 2008). These technologies create opportunities for developing specialized analytical models for conflict-related discourse.

1.4 Research Objectives

This study aims to:

1. Develop a digital corpus architecture for extracting and analyzing war-related terminology from Ukrainian news media.
2. Examine lexical patterns associated with opinion and action narratives within conflict reporting.
3. Establish a methodological framework combining web scraping, corpus construction, annotation, and computational analysis.
4. Investigate how war vocabulary contributes to narrative framing and interpretation.
5. Identify the theoretical and practical implications of corpus-based approaches for conflict discourse research.

1.5 Scope and Significance

The study focuses on digital news discourse related to the Russian-Ukrainian conflict, emphasizing terminology associated with military actions, political responses, humanitarian issues, and international relations. The research does not attempt to evaluate political positions of individual media organizations; instead, it investigates linguistic structures through corpus-based methods.

The significance of this research lies in developing an analytical bridge between computational linguistics and discourse studies. By creating a structured corpus architecture, the study provides researchers with a reproducible approach for examining how language constructs representations of war. The framework can

support future investigations in media studies, computational humanities, and linguistic analysis of geopolitical events.

2. LITERATURE REVIEW

2.1 Corpus Linguistics as a Framework for Discourse Investigation

Corpus linguistics has developed as a major methodological approach for examining language through systematically collected textual datasets. Rather than relying exclusively on researcher intuition, corpus methods allow linguistic patterns to emerge from empirical evidence. Baker (2006) highlights that corpus-based discourse analysis provides tools for identifying recurring linguistic structures and exploring how language reflects social meanings.

A fundamental principle of corpus research is the relationship between linguistic patterns and contextual interpretation. Adolphs (2008) argues that corpus investigation requires attention to pragmatic functions because words acquire meaning through their communicative environments. This theoretical position is particularly important for war terminology, where vocabulary frequently carries implicit evaluations and ideological assumptions. A term describing military activity may represent defense, aggression, resistance, or intervention depending on the surrounding discourse.

Biber (2014) further demonstrates the importance of multidimensional analysis in corpus studies by emphasizing that linguistic variation cannot be understood through single features alone. War narratives contain multiple dimensions, including emotional expression, institutional language, political argumentation, and descriptive reporting. Therefore, a digital corpus architecture must incorporate multiple analytical layers rather than focusing only on lexical frequency.

2.2 Digital Corpus Construction and Data Acquisition

The construction of specialized corpora requires systematic methods for collecting, cleaning, and organizing textual data. Digital news environments provide extensive resources, but automated collection methods are necessary because manual compilation cannot efficiently manage large-scale information. Web scraping techniques enable researchers to extract publicly available

textual materials and organize them into structured datasets (Gupta, 2024; Johnson, 2024).

Davies and Parodi (2022) emphasize the importance of developing growing corpora capable of incorporating continuously expanding linguistic data. This concept is highly relevant to war discourse because conflict terminology changes rapidly as events develop. A dynamic corpus architecture allows researchers to monitor lexical shifts and identify emerging narrative patterns.

Software platforms such as Sketch Engine demonstrate how corpus systems can support advanced linguistic exploration through keyword analysis, concordances, and collocation measurements (Kilgarriff et al., 2008; Kilgarriff et al., 2014). Similarly, AntConc provides accessible computational methods for analyzing textual collections and identifying linguistic relationships (Anthony, 2005). These tools establish the technical foundation for constructing specialized war terminology corpora.

2.3 Corpus Annotation and Contextual Interpretation

Annotation represents a crucial stage in corpus development because raw textual data requires linguistic organization before meaningful analysis can occur. Leech (1997) explains that corpus annotation introduces structured linguistic information into computer-readable texts, allowing researchers to examine grammatical, semantic, and discourse-related features.

For war terminology research, annotation may include categories such as military actions, political actors, humanitarian concepts, and evaluative expressions. Such classification enables researchers to distinguish between different narrative functions. For example, the same word related to “support” may refer to military assistance, humanitarian aid, or political solidarity depending on its annotated context.

Adolphs (2008) reinforces the importance of contextual analysis by demonstrating that corpus evidence must be interpreted through pragmatic relationships. In war discourse, isolated vocabulary analysis may overlook deeper narrative meanings. Therefore, annotation and contextual examination are essential components of a reliable digital corpus architecture.

2.4 Research Gap and Theoretical Positioning

Although existing corpus methodologies provide strong analytical foundations, limited attention has been given to specialized digital architectures designed specifically for war terminology extraction. Previous research has demonstrated the effectiveness of corpus tools, but additional frameworks are required to integrate automated collection, conflict-specific annotation, and narrative interpretation.

The reviewed literature reveals three major research gaps. First, existing corpus approaches often focus on general language analysis rather than rapidly evolving conflict discourse. Second, technical corpus construction and interpretive discourse analysis are frequently treated separately. Third, there is a need for frameworks capable of connecting computational findings with narrative meanings.

This study addresses these gaps by proposing a digital corpus architecture that combines technical data processing with discourse-oriented interpretation. Building upon corpus theory, computational tools, and contextual analysis principles, the research positions war terminology as both a linguistic dataset and a social meaning system.

3. METHODOLOGY

3.1 Research Design and Methodological Framework

This research adopts a corpus-driven computational discourse analysis approach to develop a digital corpus architecture for examining war terminology and narrative structures in Ukrainian news media. The methodology integrates principles from corpus linguistics, digital text collection, computational analysis, and discourse interpretation. Unlike traditional qualitative approaches that examine limited textual samples, the proposed framework enables systematic investigation of large-scale media datasets by combining automated extraction techniques with linguistic interpretation.

The methodological design is based on the assumption that war-related language is not merely a collection of isolated terms but a structured system of meanings produced through repeated media practices. Therefore, the research architecture is designed to identify lexical frequency, contextual relationships, collocation patterns, and narrative functions. The approach follows the corpus linguistic principle that linguistic evidence should be derived from authentic textual materials rather than predetermined

assumptions (Baker, 2006).

The proposed methodology consists of five interconnected layers:

1. Digital news data acquisition layer
2. Corpus processing and organization layer
3. Linguistic annotation layer
4. Computational analysis layer
5. Narrative interpretation layer

These layers collectively form a digital corpus architecture capable of transforming unstructured news materials into analyzable linguistic resources.

3.2 Digital Corpus Architecture Model

The proposed architecture represents a structured computational environment for collecting and analyzing war-related news discourse. The framework is designed around a continuous pipeline where textual information moves from online sources to analytical outputs.

3.2.1 Data Acquisition Layer

The first component of the architecture involves collecting digital news articles from online media platforms. Since contemporary conflict reporting is continuously updated, automated acquisition methods are required to maintain corpus relevance. Web scraping techniques provide mechanisms for extracting article content, metadata, publication dates, and source information from online platforms (Automated data collection: Web scraping example, part 1, 2017).

The data acquisition layer includes:

- Source identification
- Automated webpage retrieval
- Article text extraction
- Metadata preservation
- Duplicate content removal

For example, news reports discussing Ukrainian military requirements, international weapons assistance, or migration consequences can be collected from different media environments, including Ukrainian and European news platforms (Editorial, 2025; Bassets, 2025; Kinkartz, 2025b). These sources provide diverse representations of the conflict, allowing comparative linguistic analysis.

The main methodological challenge at this stage is maintaining data consistency. Online news articles vary in structure, formatting, and accessibility. Therefore, preprocessing mechanisms are necessary to transform heterogeneous materials into standardized corpus documents.

3.3 Corpus Construction and Data Processing Framework

3.3.1 Corpus Development Process

Corpus construction involves transforming collected news materials into a structured linguistic database. The process requires several stages:

Data Cleaning

Raw digital texts frequently contain advertisements, navigation elements, hyperlinks, and irrelevant content. Data cleaning removes these elements while preserving meaningful linguistic information.

Text Normalization

Normalization includes:

- Converting different encoding formats
- Removing unnecessary symbols
- Standardizing punctuation
- Preparing texts for computational analysis

Corpus Storage

The processed documents are stored with metadata including:

- Source publication
- Date of publication

- Article category Humanitarian Vocabulary
- Language characteristics Examples:
- Conflict-related thematic classification
 - refugees
 - civilians
 - displacement
 - assistance

The development of growing corpora is particularly relevant for conflict studies because terminology changes according to political and military developments. Davies and Parodi (2022) emphasize the importance of expanding corpora that can represent changing linguistic environments.

3.4 Linguistic Annotation Framework

Annotation provides the connection between computational processing and linguistic interpretation. The proposed framework applies multi-level annotation to classify war-related expressions according to their semantic and discourse functions.

3.4.1 Lexical Annotation

Lexical annotation identifies important war-related terminology. Categories may include:

Military Action Vocabulary

Examples:

- attack
- defense
- weapons
- battlefield
- operation

Political and Diplomatic Vocabulary

Examples:

- negotiations
- sanctions
- support
- alliance

Evaluative Vocabulary

Examples:

- threat
- victory
- aggression
- resistance

This classification enables researchers to analyze not only which words appear frequently but also what social meanings they represent.

3.4.2 Contextual Annotation

A central feature of the architecture is contextual annotation. According to Adolphs (2008), corpus analysis must examine pragmatic functions because linguistic forms acquire significance through their communicative contexts. Therefore, war terminology is analyzed according to surrounding linguistic structures.

For example, the term “support” may appear in different narrative contexts:

- Military support provided by allied countries
- Humanitarian support for displaced populations
- Political support for national sovereignty

Although the lexical form remains identical, its discourse function changes according to context.

The contextual annotation layer therefore records:

- Nearby words
- Sentence-level meaning
- Speaker or institutional perspective
- Narrative orientation

3.5 Computational Analysis Framework

3.5.1 Frequency and Keyword Analysis

Frequency analysis identifies dominant vocabulary patterns within the corpus. High-frequency terms provide initial indicators of major themes represented in news discourse.

However, frequency alone cannot explain meaning. A frequently occurring term may represent different narratives depending on its contextual distribution. Therefore, frequency analysis is combined with qualitative interpretation.

Corpus tools such as AntConc enable researchers to examine word frequency, concordances, and keyword patterns within textual collections (Anthony, 2005). These capabilities support systematic identification of important war-related expressions.

3.5.2 Collocation and Co-occurrence Analysis

Collocation analysis examines relationships between words that frequently occur together. Evert (2005) demonstrates that statistical analysis of word co-occurrences provides insights into semantic relationships within corpora.

Within the proposed architecture, collocation analysis identifies patterns such as:

- military + assistance
- territorial + defense
- humanitarian + crisis
- international + support

These relationships reveal how media narratives organize concepts around specific conflict themes.

For example, if “support” frequently appears with “weapons,” the narrative emphasis may focus on military assistance. If it appears with “humanitarian,” the discourse may shift toward civilian consequences.

3.6 Corpus Analysis Tools and Technical Environment

The methodology incorporates established corpus analysis platforms to support computational investigation.

3.6.1 AntConc-Based Analysis

AntConc provides accessible tools for examining textual patterns, including:

- Concordance generation
- Keyword identification
- Cluster analysis
- Frequency calculation

Anthony (2005) highlights that AntConc was designed as a practical corpus analysis toolkit capable of supporting detailed linguistic investigations.

3.6.2 Sketch Engine-Based Analysis

Sketch Engine contributes advanced corpus processing capabilities, including:

- Word sketches
- Collocation analysis
- Corpus comparison
- Lexical profiling

Kilgarriff et al. (2008) describe Sketch Engine as a platform designed for practical lexicographic analysis, while Kilgarriff et al. (2014) demonstrate its continued development as a corpus research environment.

3.6.3 LancsBox Integration

LancsBox provides additional visualization and analytical possibilities for corpus exploration. Its application supports network-based examination of lexical relationships and discourse patterns (LancsBox X, 2025).

The combination of these tools creates a flexible analytical environment where computational results can be interpreted through discourse theory.

3.7 War Narrative Extraction Model

The central analytical component of the proposed architecture is the War Narrative Extraction Model (WNEM). This model identifies how vocabulary constructs opinion and action narratives.

The model contains three analytical dimensions:

Dimension 1: Action Representation

This dimension examines vocabulary describing events, decisions, and operations.

Examples:

- military movements
- strategic actions
- government decisions
- international responses

The objective is to identify how news discourse represents activities and responsibilities.

Dimension 2: Opinion Construction

This dimension analyzes evaluative language reflecting perspectives and interpretations.

Examples:

- legitimacy claims
- threat descriptions
- moral evaluations
- political judgments

Opinion narratives are important because media texts frequently combine factual reporting with interpretive framing.

Dimension 3: Actor Positioning

This dimension investigates how linguistic choices represent different participants.

Actors may include:

- governments
- military organizations
- international institutions
- civilian populations

The analysis examines how vocabulary assigns roles such as defender, aggressor, victim, or supporter.

3.8 Analytical Procedure

The complete research procedure follows these sequential stages:

Stage 1: Digital Data Collection

Online news articles related to the Russian-Ukrainian conflict are collected through automated extraction methods.

Stage 2: Corpus Preparation

Collected materials are cleaned, normalized, and stored with metadata information.

Stage 3: Annotation

Texts are categorized according to war terminology, narrative function, and contextual meaning.

Stage 4: Computational Analysis

Frequency analysis, keyword extraction, concordance analysis, and collocation measurement are conducted.

Stage 5: Interpretive Analysis

Computational findings are examined through corpus discourse principles to identify narrative structures.

This integrated approach ensures that statistical results are connected with meaningful discourse interpretation.

3.9 Methodological Reliability and Limitations

The proposed architecture improves reliability by combining multiple analytical techniques rather than depending on a single measurement method. Computational analysis identifies large-scale patterns, while contextual interpretation explains their significance.

However, several limitations exist. Automated data collection may encounter technical restrictions, website structure changes, or incomplete information extraction. Furthermore, corpus analysis cannot completely eliminate researcher interpretation because decisions regarding annotation categories and analytical focus involve theoretical choices.

Another limitation concerns media diversity. News sources represent particular editorial environments, meaning that corpus findings should be interpreted as representations of discourse patterns rather than absolute descriptions of political reality.

Despite these limitations, the digital corpus architecture provides a systematic and reproducible methodology for studying war terminology and media narratives.

4. RESULTS

The implementation of the proposed digital corpus architecture reveals several important patterns regarding the organization and representation of war terminology within Ukrainian news media discourse. The findings demonstrate that computational corpus analysis can identify not only frequently occurring lexical items but also broader narrative structures associated with military actions, political positioning, and social consequences of conflict.

4.1 Emergence of Structured War Lexicons

The corpus analysis indicates that war-related vocabulary forms interconnected lexical networks rather than isolated word groups. Terms associated with military operations, international assistance, territorial issues, and humanitarian concerns appear through repeated contextual relationships. This finding supports the corpus linguistic perspective that meaning is produced through patterns of usage and contextual association rather than individual lexical units (Adolphs, 2008).

The constructed corpus architecture enables the identification of dominant semantic fields within conflict

reporting. Military terminology generally clusters around concepts of defense, weapons, operations, and strategic responses, while political terminology connects with diplomacy, international cooperation, and governmental decisions. Humanitarian vocabulary demonstrates another narrative layer by emphasizing civilian experiences, displacement, and social consequences.

These patterns confirm that war reporting operates through multiple interconnected discourse domains. The same conflict event can simultaneously be represented as a military issue, political challenge, humanitarian crisis, or international security concern.

4.2 Identification of Opinion and Action Narratives

A major finding of the study is the distinction between action-oriented narratives and opinion-oriented narratives. Action narratives primarily describe events, decisions, and operational developments. They contain verbs and expressions associated with movement, intervention, assistance, and strategic responses.

Opinion narratives, in contrast, contain evaluative expressions that construct interpretations of events. These narratives frequently involve judgments regarding responsibility, legitimacy, urgency, and threat. The corpus framework demonstrates that opinion formation is embedded within linguistic choices rather than appearing only in explicit commentary sections.

This finding reflects Adolphs' (2008) argument that pragmatic meaning emerges through contextual relationships. War terminology cannot be fully understood without examining how lexical items function within broader communicative environments.

For example, terms related to military assistance may represent cooperation and protection in one context, while representing escalation and confrontation in another. Therefore, contextual analysis remains essential for accurate interpretation.

4.3 Computational Identification of Lexical Relationships

The application of corpus analysis techniques reveals significant relationships between frequently co-occurring terms. Collocation analysis demonstrates that war vocabulary operates through predictable semantic combinations.

The analysis identifies several major lexical associations:

- military terms connected with strategic actions,
- political terms connected with international responses,
- humanitarian terms connected with civilian impact,
- evaluative terms connected with conflict interpretation.

These relationships demonstrate that corpus architecture provides deeper analytical capabilities than simple frequency counting. Through computational measurement of word associations, researchers can identify conceptual structures embedded within news discourse.

The effectiveness of such analysis corresponds with the contributions of corpus tools such as AntConc and Sketch Engine, which support systematic examination of lexical relationships and contextual patterns (Anthony, 2005; Kilgarriff et al., 2008).

4.4 Digital Corpus Architecture as a Research Infrastructure

The findings demonstrate that the proposed architecture functions as more than a data collection mechanism. It represents a research infrastructure capable of supporting continuous analysis of changing conflict discourse.

Because war terminology evolves rapidly, static datasets may fail to capture emerging expressions and changing narratives. The architecture's capacity for continuous data integration provides researchers with opportunities to monitor linguistic transformation over time. This aligns with the concept of expanding corpora discussed by Davies and Parodi (2022), where growing datasets allow researchers to examine language change.

The framework also demonstrates that computational methods can complement traditional discourse analysis. Automated processing increases analytical scale, while linguistic interpretation preserves contextual understanding.

4.5 Findings Summary

The major findings of the study can be summarized as

follows:

First, war terminology in Ukrainian news media forms complex lexical networks representing military, political, and humanitarian dimensions of conflict.

Second, opinion and action narratives can be systematically distinguished through corpus-based annotation and contextual analysis.

Third, computational tools provide effective mechanisms for identifying lexical patterns, but interpretation requires attention to pragmatic context.

Fourth, a digital corpus architecture enables continuous and reproducible investigation of evolving war discourse.

Overall, the findings demonstrate that corpus-based approaches provide a powerful framework for understanding how contemporary conflicts are linguistically constructed and communicated.

5. DISCUSSION

5.1 Theoretical Implications

The findings contribute to corpus-based discourse research by demonstrating the importance of combining computational analysis with contextual interpretation. Traditional quantitative corpus approaches often emphasize measurable linguistic patterns, while discourse analysis focuses on social meaning. The proposed architecture connects these perspectives by treating war terminology as both measurable linguistic data and socially meaningful discourse.

The study reinforces the theoretical position that context determines linguistic function. As emphasized by Adolphs (2008), corpus analysis requires examination of how language operates within specific communicative environments. The present research extends this principle into digital conflict studies by showing that war vocabulary must be analyzed through both computational patterns and narrative functions.

The results also support Baker's (2006) argument that corpus methods can reveal relationships between language and social representation. News discourse does not simply report conflict events; it organizes them through recurring lexical choices and narrative structures.

5.2 Practical Implications

The proposed digital corpus architecture has several practical applications.

For media researchers, the framework provides a systematic method for examining how different news environments construct conflict narratives. Researchers can compare terminology across sources, identify shifts in representation, and analyze changes in public discourse.

For computational linguists, the architecture demonstrates how corpus technologies can be adapted for specialized domains. The integration of scraping methods, annotation systems, and corpus tools provides a foundation for developing domain-specific linguistic models.

For educators and digital humanities researchers, the framework offers opportunities for teaching corpus-based analysis and demonstrating how language reflects geopolitical processes.

5.3 Relationship with Previous Research

The findings are consistent with previous corpus research emphasizing the importance of structured textual analysis. Anthony (2005) demonstrated the value of accessible corpus tools for linguistic investigation, while Kilgarrieff et al. (2014) showed the continued development of computational corpus environments.

However, this research extends previous approaches by focusing specifically on war terminology and narrative extraction. Existing corpus platforms provide technical capabilities, but they do not automatically explain the ideological and pragmatic dimensions of conflict language.

The integration of corpus architecture with narrative analysis represents an important advancement because it addresses the limitation of purely statistical approaches.

5.4 Methodological Trade-offs and Challenges

Although the proposed framework provides significant analytical advantages, several trade-offs must be considered.

The first challenge concerns automation versus interpretation. Automated methods allow analysis of large datasets but may overlook subtle meanings, irony, or politically sensitive expressions. Human interpretation

remains necessary to explain complex discourse patterns.

The second challenge involves corpus representativeness. Online news sources reflect specific editorial practices and audience expectations. Therefore, findings should be understood as patterns within selected media environments rather than complete representations of public opinion.

The third challenge concerns technological sustainability. Digital platforms frequently change structures, requiring continuous adaptation of data collection methods.

5.5 Limitations and Future Research Directions

The study has several limitations. The corpus architecture focuses primarily on textual news materials and does not incorporate multimedia elements such as images, videos, or social media discussions. Future research could expand the framework toward multimodal conflict analysis.

Additionally, further studies could incorporate multilingual comparison between Ukrainian, Russian, English, and European news sources. Such comparative research could reveal how different linguistic communities construct competing interpretations of the same events.

Future developments may also integrate artificial intelligence-based semantic models to enhance automated identification of narrative structures while maintaining human-centered interpretive approaches.

6. CONCLUSION

This study developed a digital corpus architecture for investigating war terminology and extracting opinion and action narratives from Ukrainian news media. By integrating corpus linguistics, automated data collection, annotation frameworks, and computational analysis methods, the research demonstrates how digital approaches can enhance understanding of contemporary conflict discourse.

The findings show that war-related vocabulary functions as a complex system through which media organizations construct interpretations of military events, political decisions, and humanitarian consequences. Rather than representing conflict through neutral terminology alone, news discourse organizes meanings through recurring lexical patterns, contextual relationships, and narrative structures.

The proposed framework contributes to digital humanities and corpus-based discourse research by providing a methodological model capable of analyzing large-scale conflict-related textual data. It demonstrates that computational techniques are most effective when combined with contextual linguistic interpretation.

The research also highlights the continuing importance of corpus architecture development in rapidly changing digital environments. As conflicts increasingly unfold within information ecosystems, systematic analysis of language becomes essential for understanding how narratives are produced, circulated, and transformed.

Future research should expand this framework through multilingual corpora, multimodal datasets, and advanced computational models capable of examining increasingly complex forms of digital conflict communication.

REFERENCES

1. Adolphs, S. (2008). *Corpus and context: Investigating pragmatic functions in spoken discourse*. John Benjamins Publishing.
2. Andrés, H., & Velasco, M. (2025, May 29). Guerra Ucrania –Rusia, últimas noticias. *El Mundo*.
3. Anthony, L. (2005). AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. *Proceedings of the IEEE International Professional Communication Conference*, 729–737.
4. AntConc. (2021). SAGE Publications, Ltd.
5. Automated data collection: Web scraping example, part 1. (2017).
6. Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.
7. Bassets, M. (2025, May 28). Alemania amplía la ayuda a Ucrania para fabricar armas que alcancen territorio ruso. *Ediciones EL PAÍS S.L.*
8. Biber, D. (2014). Opening. *Multi-Dimensional analysis*. In *Studies in Corpus Linguistics* (pp. xxix–xxxviii). John Benjamins Publishing Company.
9. Carter, R. (2004). *Language and creativity: The art of common talk*. Routledge.
10. Collins, P., & Yao, X. (2016). Douglas Biber and Randi Reppen (eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press, 2015. Pp. 639. ISBN 978-1-107-03738-0 (hardback). *English Language and Linguistics*, 21(1), 161–168.
11. Cook, G. (2004). *Genetically modified language: The discourse of arguments for GM crops and food*. Routledge.
12. Davies, M. (2004). Student use of large, annotated corpora to analyze syntactic variation. In *Studies in Corpus Linguistics* (pp. 259–269). John Benjamins Publishing Company.
13. Davies, M., & Parodi, G. (2022). Constitución de corpus crecientes del español. In *Lingüística de corpus en español* (pp. 13–32). Routledge.
14. Editorial, T. (2025, May 28). Zelenskyy says what the Ukrainian military needs most. *TCH*.
15. Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations* (Doctoral dissertation, University of Stuttgart).
16. Fendel, V. B. (2024). The i.sicily Sketch engine corpus. *Journal of Open Humanities Data*, 10.
17. Gupta, S. (2024). *Getting started with Web Scraping. In Web Scraping with Python*. Apress.
18. Hanks, P. (2016, April 5). *Create and search a text corpus*. Sketch Engine.
19. Johnson, M. (2024). *Web-scraping*. Litres.
20. Kilgarrieff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36.
21. Kilgarrieff, A., Rychly, P., Smrž, P., & Tugwell, D. (2008). The Sketch Engine. In *Practical Lexicography* (pp. 297–306). Oxford University PressOxford.
22. Kinkartz, S. (2025b, May 28). *Migration: Deutschland*

will es Flüchtlingen schwerer machen. Deutsche Welle.

23. LancsBox X. (2025). Retrieved June 3, 2025, from <https://lancsbox.lancs.ac.uk/>
24. Laurence Anthony's Website. (2025). Retrieved June 3, 2025, from <https://www.laurenceanthony.net/software/antconc/>
25. Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech, & T. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 1–18). Longman.