［招待：研究論文］

Text-as-data Approach for China Studies

New Opportunities for Familiar Problems

中国研究におけるテキストデータ分析

既存の研究課題に対する新たなアプローチ

Vida Macikenaite

Assistant Professor, Graduate School of International Relations, International University of Japan

マチケナイテ・ヴィダ
国際大学国際関係学研究科講師

Correspondence to: vida@iuj.ac.jp

Abstract: This paper advocates the use of the text-as-data approach – automated or computer-assisted text analysis – for the study of Chinese politics and foreign relations. It reviews how computational text analysis has been applied in China studies and demonstrates its analytical potential through a pilot semantic network analysis of Chinese Ministry of Foreign Affairs spokespersons' remarks on Russia's invasion of Ukraine. The findings reveal the centrality of the United States in China's narrative, the moral framing of China's position, and a subtle discursive shift from “justice” to “dialogue” over time. Nevertheless, the paper concludes that computational text analysis enhances, rather than replaces, human interpretation, opening new avenues for rigorous, data-driven China research validated by human expertise.

本稿は、中国政治および対外関係研究において、テキストデータ分析（text-as-data）アプローチ、すなわち計算機支援型テキスト分析（computer-assisted text analysis）の活用を提唱する。まず、中国研究分野における計算的テキスト分析の適用状況を概観し、続いて、ロシアによるウクライナ侵攻をめぐる中国外交部報道官発言を対象とした予備的な意味ネットワーク分析を通じて、その分析手法としての価値を示す。分析の結果、中国の言説において米国が中心的な参照点として位置づけられていること、中国の立場が規範的・道徳的フレーミングによって構築されていること、さらに時間の経過とともに言説の重点が「正義」から「対話」へと漸進的に移行していることが明らかとなった。他方で本稿は、計算的テキスト分析は研究者の解釈を代替するものではなく、研究者による解釈と検証を前提としてそれを補完・強化する方法論であると結論づける。すなわち、理論的関心と研究者の専門的判断に支えられた、厳密かつデータ駆動型の中国研究に新たな研究展開をもたらすものである。

Keywords: text-as-data, China studies methodology, semantic network analysis, China-Russia relations
テキストデータ分析、中国研究手法、意味ネットワーク分析、中露関係

1. Introduction

China studies traditionally have relied on a careful reading of original government documents, state media articles, and interviews with insiders, alongside archival research. Nevertheless, events in the past decade have increasingly limited scholarly reliance on these sources. Under the Xi Jinping administration, the policy process in China has become even more obscure, and access to the field in China has narrowed, as documented in Japan (Lim et al., 2025) and abroad (The Economist, 2023). Then, COVID-19 pandemic closed the doors to many still available data sources for a significant period of time. Nevertheless, these developments have coincided with other trends in social sciences. There has been a remarkable proliferation of digital data over the last decade, which was also accompanied by the development of new computational analysis methods for social sciences. As a result, social sciences, especially since 2017, have witnessed a remarkable rise of machine learning in academic social sciences (Rahal et al., 2024). Computational analysis and machine learning have provided increasingly accessible tools to analyze social science data. In the field of China studies, where texts have long been a vital research resource, text-as-data approaches have unlocked untapped potential to enhance our understanding of China further – to analyze and explain Chinese politics and foreign policies, often with a higher degree of accuracy and potential to discover new patterns and emerging trends.

This paper seeks to make the case for the use of text-as-data – automated or computer-assisted text analysis – methods to study Chinese politics and foreign relations. It argues that, considering the importance of text and discourse in the Chinese context, text-as-data approaches present great potential for a comprehensive analysis of a wide range of topics, even with limited access to the field.

For this purpose, this paper is divided into two parts. The first part discusses the use of text-as-data methods in the field of China studies to date, highlighting their advantages through the review of existing literature. The second part is based on the original research on China-Russia relations in the context of Russia's invasion of Ukraine. Using an original dataset, which comprises the remarks of the spokesperson of the Ministry of Foreign Affairs of China (MFA), the paper illustrates how semantic network analysis, one of the text-as-data methods, can be used to capture dominant narratives, specific concepts and nuances in official statements. In this way, the paper presents text-as-data analysis methods as the approach that is likely to take a central position in the studies of Chinese domestic and external politics for the upcoming decade to reexamine the questions that have puzzled China scholars to date – China's relations with major powers, its intentions toward the international system and the existing world order, as well as its domestic policy intentions and relations with the society.

The purpose of this paper is confined within a limited scope. It aims to argue that automated text analysis is one of the approaches with the highest potential for China studies in the near future. Nevertheless, it is beyond the scope of this paper to explain specific methods in text-as-data approaches or provide a tutorial on their application. The first would be impossible due to the large number of methodological tools available, while the latter one does not exist. As Grimmer and Stewart (2013) summarize, "there is no globally best method for automated text analysis."

2. Text analysis: the past and future of China studies

Text analysis has been an essential tool in the study of China, and it remains as such today. For example, in the study of elite politics, in-depth field research is difficult; thus, even now the scholars have to rely on careful scrutiny of official speeches, public documents, and the publications by the regime-controlled media to make inferences about Chinese leaders decisions and their intentions. Despite unprecedented access to information, the inner workings of China under Xi Jinping remain opaque, so "analysts have thus revived the Cold War-era discipline of "Pekingology" – the meticulous decoding of subtle signals from Beijing's corridors of power" (Messingschlager, 2025). While China may have been one of the most widely studied subjects in recent years, some China experts call for a careful reconsideration of the ways we think about China. Charles Parton, a British diplomat who spent 22 years working with China, Hong Kong and Taiwan, cautions that "Conventional wisdom should be shunned. <…> We need to look behind 'conventional wisdom'; rather than accept it, we should question its assumptions by going back to what the sources say" (Parton, 2022). He emphasizes the necessity in Xi Jinping's 'New Era' to collect, sort and analyze big open-source datasets.

In support of these arguments, a recent analysis of over 12,000 *People's Daily* articles and a substantial set of Xi Jinping speeches challenges the prevailing assumptions about China's strategic ambitions. Through automated text analysis and a follow-up content analysis, Kang et al. (2025) conclude that China is an inward-looking *status quo* power more concerned with regime stability and sovereignty than global dominance, a significantly different observation to the dominant narrative in the literature recently.

This highlights the key tasks that lie ahead of China scholars in the coming years. As the future direction of China under the centralized leadership of Xi Jinping is debated and the international system is undergoing tremendous changes, researchers carry the responsibility to accurately define the direction China is taking – its relations with major powers and smaller states or regions, its view of the international system, and also domestically, including the stability of the CCP regime. These are all familiar problems, and an extensive number of papers attempt to provide answers. Nevertheless, the proliferation of available data has unlocked multiple avenues for verifying or deepening the already familiar arguments. Delivering reliable answers is a crucial task for the study of

China as a near-peer competitor to the US, and which also has a tremendous impact on the world.

3. The advance of the text-as-data approach

Text has always been a source of information and resource for analysis. But there is an essential difference between an interpretivist approach to text, which focuses on what the text means and treats the text as content to be evaluated, and a text-as-data approach (Benoit, 2020). Instead, the text-as-data approach "views text as a source of data that can be transformed into a quantitative representation and analyzed using statistical and computational methods to make systematic inferences about political, social, or psychological processes" (Grimmer et al., 2022). There is a conceptual shift, as classical content analysis requires reading and understanding the text to extract meaning, while the text-as-data approach treats texts as datasets for analysis. Analysis often seeks to quantify words, phrases or syntactic structures by counting them, exposing topic distribution or sentiment scores, and often this is computational or computer-assisted analysis.

This change has not occurred overnight. In the field of social sciences, the first systematic attempt to approach text as machine-readable data was made by Stone (1962) and then by Stone et al. (1966), who explored the possibility of supporting human text coding with computer automation. It combined dictionaries – word lists – with a program that could apply them across large corpora and count the use of those words. Later, in the 1980s, Roderick Hart applied these dictionary-based methods to rhetorical analysis of presidential speeches. But it was not until 2003, when Laver et al. (2003) explicitly approached texts "not as discourses to be understood and interpreted but rather as data in the form of words." Aimed at overcoming the limitation of traditional computer-assisted techniques, which still required human coding to generate dictionaries, Laver et al. (2003) suggested a technique, which "extracts data from [a textual document] in the form of word frequencies and uses this information to estimate the policy positions of texts about which nothing is known."

These breakthroughs paved the path for the text-as-data approach, the methods of which have rapidly advanced to this day. King and Lowe (2003), equally intrigued to overcome the costs of human labor in manually collecting data on daily political events, introduced a pioneering method, where a computer program was "able to extract information from Reuters news reports on a level equal to trained Harvard undergraduates." Such a tool, the efficiency and accuracy of which surprised even the authors at the time, has undergone significant advancements over the last two decades. For example, the *Global Database of Events, Language, and Tone* (*GDELT Project*) monitors news media (print, broadcast, web) globally in over 100 languages. It then converts this unstructured news content into structured data, identifying events, actors, themes, emotions, geographic locations, and other relevant information to update its database every 15 minutes. There is also an ongoing effort to integrate social media into the database, which is already free and accessible to all, offering downloadable datasets reaching as far back as 1979.

Since the advance of machine learning into social sciences, there has been a proliferation of methods and tools to analyze textual data. Topic modeling, sentiment analysis, and semantic network analysis have been the most common techniques, and their tools continue to evolve. Some tools require advanced computational methods and programming skills, while others are more accessible to social scientists.

In China studies, text analysis is by no means a new method. China scholars have long sought to understand the essence of Chinese politics and the intent of its foreign policy through a careful examination of state media, official documents, and leadership speeches. This methodological tendency reflects both structural and ideological features of the Chinese political system. Limited transparency and restricted access to decision-making processes have long made official texts – such as Party documents, leaders' speeches, and MFA statements – primary sources of empirical evidence. At the same time, the CCP's ideological tradition assigns a performative role to language: political speech not only describes policy but enacts and legitimizes it. As a result, shifts in rhetorical formulations often signal changes in policy orientation or elite consensus. This has encouraged a persistent interpretive focus on discourse. This approach has also defined the Japanese tradition of China studies.

Nevertheless, as online data sources have proliferated, Chinese students have also turned to utilizing computational text analysis. Besides some early computer-assisted text analysis, computational methods of the text-as-data approach reached Chinese studies in the 2010s. Some of them are reviewed further below to illustrate the advantages and potential of text-as-data approaches.

4. Text-as-data analysis in China studies

4.1 Comprehensive analysis of large volumes of textual data from different sources

One of the most significant advantages of the text-as-data approach is that it enables a comprehensive analysis of vast volumes of textual data. Computational or computer-assisted analysis enables the researcher to process corpora – the sets of texts under study – that exceeds anything that can be processed manually. Moreover, different data – leadership speeches, MFA statements, press conferences, newspaper articles, social media posts, etc.– can be collected from a wide variety of sources that are increasingly available online. Online data scraping techniques, although not possible for all websites due to their data protection regulations, have further increased the costs of data collection. Finally, and very importantly, regularly advancing research tools and techniques already enable the researchers to analyze text not only in English but also in other languages, including Chinese.

As expected, China scholars use the text-as-data approach to study documents produced by official institutions. Leader speeches, just as before, are still a valuable resource for analysis, and, notably, they are available online. Lim et al. (2025) analyze 9,016 speeches, reports and statements in Chinese by Xi Jinping in the period from 2012 to 2022 from the online database of Xi Jinping's important speech series (习近平系列重要讲话数据库). Another data source is the regular press conference of the Chinese MFA. The advantage of it as a data source is that it is held regularly and the transcripts are available online in both original Chinese and English translation. In the existing literature, Mochtak and Turcsanyi (2021) present the corpus of nearly 23,000 Q&A dyads from the MFA press conferences in the period of 2002 to 2020, which they then use to visualize the evolution of the US-related foreign policy discourse. Based on a similar time period of two decades, Dai and Luqiu (2024) demonstrate that the combatant tone of China's "wolf warrior diplomacy" increased after 2012.

Media has always been a source for understanding Chinese politics and the communication surrounding it, and the same is true of the new approaches and techniques. Mattingly et al. (2024) use a corpus of 19,791 CGTN video segments posted on the broadcaster's YouTube channel to explore the media narratives about the Chinese regime. Fisher et al. (2022) survey how Chinese media – *People's Daily* and *Global Times* – along with MFA narrate relevant topics. Based on the corpus of over one million media articles and 762 MFA documents from the *FOCUSdata Project*, they observe that media reporting is more negative than that of MFA. Weston and Rauchfleisch (2021) analyze 53,000 articles from *People's Daily* to survey relative attention to different regions in the media in the period between 2016 and 2020.

Needless to say, social media has provided scholars with endless opportunities in this field. King et al. (2017), who studied the full list of 43,757 leaked so-called 50c posts on *Weibo*, show the logic of regime-supported communication on Chinese social media. The goal of coordinated posts by *Weibo* users, who are almost always state officials, is not to counter critical arguments but to distract the public and change the subject during critical periods – their posts often involve cheerleading for China, the revolutionary history of the CCP, or other symbols of the regime. Guo and Qin (2024) survey China's diplomatic engagement with foreign audiences on Twitter through sentiment analysis of 14,000 *Twitter* posts. The same sentiment analysis is used by Tao et al. (2024) for over 2,8 million *Sina Weibo* posts for a three-year period to explore populist discourse on Chinese social media.

Among the less common data sources is the Chinese court data updated daily online. Liebman et al. (2020) reveal the untapped potential of the digitalization of Chinese court decisions as they use over one million documents from Henan Province.

4.2 New findings about familiar problems

Studies that use a text-as-data approach suggest that analysis of large, comprehensive datasets may produce research results that contradict the findings of the already existing studies. For example, Liebman et al. (2020) challenge the conventional wisdom that administrative lawsuits are an extension of contentious politics, where Chinese citizens challenge the state in court. Instead, they find that administrative lawsuits are the venue of seeking the state's help in resolving an underlying civil dispute between two private parties. Similarly, Mattingly et al. (2024) demonstrate that "Chinese messages promoting its system to a global audience are strikingly successful," in stark contrast to the common argument that the CCP has promoted a narrative of national rejuvenation and it had little international appeal (Weiss 2019). More studies report their findings as contrary to the existing arguments.

4.3 The tool for discovery

In what became the first textbook for the text-as-data approach, Grimmer et al. (2022) make the case for a more inductive approach in social sciences, as new observations happen when analyzing data rather than using data to test the existing theories. Indeed, different methods in text-as-data approach offer tools for discovering new insights in the text that researchers might otherwise miss. First, this approach points to new ways to organize texts, prompting the researcher to read texts differently. Topic modeling, a common technique in computational analysis, is a type of statistical model to determine abstract "topics" that occur in the collection of textual data. It is employed to discover a hidden semantic structure within a body of text allowing researchers to explore their corpus without a preexisting hypothesis. For example, Mattingly et al. (2024) apply topic modeling to their CGTN media posts to extract prevalent topics, and then group them into clusters, primarily including China's political model, China's economic model, international news, Chinese domestic news, Chinese culture, and pandemic news. Topic models have been used to trace the rise and fall of themes in Chinese political communication. Zhang et al. (2023) demonstrate that reporting of hard news on *People's Daily's Weibo* account during the initial stage of the pandemic in 2019 softened toward 2021. Li et al. (2025) chart the evolution of language policy in China, showing how attention to specific themes rises or falls over time. Lim et al. (2025) capture the change of topic prevalence between Xi Jinping's two terms, as they find that, in addition to the three major agendas identified in previous research, poverty alleviation and domestic innovation gained significant rhetorical attention during Xi's second term. All these analyses were not guided by a hypothesis of the expected findings but instead explored the topics that emerged from the data.

Text-as-data tools enable one to measure not only the prevalence of particular concepts but also the strength of the relationship between them. The case study below shows how semantic network analysis is used to grasp the strength of the argument of China *vis-à-vis* the United States in the context of the war in Ukraine.

5. China's official communication on the war in Ukraine

5.1 The choice of the case study

To illustrate the potential of the text-as-data approach for the study of China, specifically its foreign policy, this paper offers a pilot study of China's official communication in the context of Russia's invasion of Ukraine since February 2022. It surveys the communication of the Chinese MFA spokesperson at the daily press conferences in the first few weeks after the invasion and a similar period of time approximately one year later. A larger corpus would be necessary for an exhaustive analysis; however, the purpose here is to demonstrate how the text-as-data approach can be useful in identifying patterns and their subtle yet significant changes in China's official narrative.

This case is especially interesting to look into, because Russia's invasion of Ukraine put Chinese foreign policy at odds. Careful analysis of China's official communication on the war can reveal how China navigates such contradictions. After Russia invaded Ukraine on February 24, 2022, the attention of Western observers quickly turned to how China would respond. In particular, questions were asked about how Beijing would reconcile its stated fundamental foreign policy principle of "respect for state sovereignty and territorial integrity" with its "no limits partnership" with Russia, proclaimed in a joint statement of the two sides just a few weeks before the invasion (Presidential Executive Office, 2022). Meanwhile, political elites in Western countries grew increasingly impatient with China's so-called "pro-Russian neutrality," the term widely adopted in Western media to describe China's reluctance to condemn Russia's violation of Ukraine's territorial integrity. Early into the war, US President Joe Biden warned China against supplying material assistance to Russia, and European leaders have repeatedly called on China to put pressure on the aggressor over the war in Ukraine.

Against the background of Russia's invasion of Ukraine, China needs to maneuver between ostensibly incompatible foreign policy objectives or principles. Beijing has declared its neutrality with respect to the war in Ukraine, yet this has failed to shield it from international criticism. China finds itself in a position where it must, on the one hand, be seen to stand for its long-time adherence to respect for sovereignty and territorial integrity – which Russia has clearly violated in Ukraine – while, on the other, it cannot openly condemn Russia due to their long-term strategic partnerships. As Trush (2022) points out, while China must demonstrate support for its "quasi-ally" Russia, it also needs to avoid becoming fully entangled in Moscow's decision to resolve its conflict with Ukraine and NATO through exclusively military and coercive means. China's principal challenge, then, is to "harmonize" its tacit support for Russia as its "strategic partner" with its other geoeconomic priorities – a diplomatic task apparently impossible to accomplish. In this regard, sanctions imposed by the West for supporting Russia could bear rather painful implications for China. Under such circumstances, China had to be very careful in how it communicated its position on the war in Ukraine as well as its partnership with Russia, the aggressor in the conflict.

5.2 The dataset and research methodology

Among the many existing tools in text-as-data approaches, this analysis employs Semantic Network Analysis (SNA), which is used to explore and visualize how concepts and words are related within a body of text. SNA presents text as a network, where nodes represent words or concepts in the text, and the edges connecting the nodes represent their co-occurrences in a sentence, paragraph or so, depending on the research design. In other words, the SNA identifies the dominant narrative in the corpus and exposes the strength of relations – co-occurrence – of specific nodes. In this analysis, we utilize an online analytical tool for SNA, *neTxt,* developed by Segev (2021).

The data used here are taken from the English-language transcripts of the daily press conferences held by the spokesperson of the Chinese MFA. We chose to focus specifically on these documents for several reasons. First, the daily press conferences are held in a Q&A format, where journalists, including those from Western media outlets, are able to ask the spokesperson questions; the transcripts recording these exchanges thus differ from other official statements inasmuch as the agenda is also introduced by the journalists, not only by a Chinese state institution. And thus it may contain information that would be omitted in other documents. Second, since the press conferences recur daily, the transcripts provide a substantial body of data over a period of time, making text analysis possible. Third, MFA spokespersons can be regarded as directly articulating the official position of the Chinese government at any given moment; this is in contrast to official or semi-official newspapers, such as the CCP mouthpiece *People's Daily* or *Global Times*, where the reporting is more indirect and will likely involve sensationalized language on the aspects of the editor's choice.

Two datasets were created for our analysis: (1) for the period from February 22 to March 30, 2022 (consisting of 187 Q&A dyads; and 37,384 words and phrases in total); (2) for the period from February 8 to March 30, 2023 (consisting of 77 Q&A dyads; and 15,080 words and phrases in total). The smaller size of the second dataset can be attributed to a reduction in interest in the war by February-March 2023 from the side of the journalists. The full transcripts available from the website of the MFA of China contain information about topics other than the war in Ukraine; hence these datasets include only the questions and answers directly related to the war. Further, while all data recorded in the transcripts were used to determine the list of words to be analyzed, only the answers given by the Chinese side were subjected to Semantic Network Analysis, as we are interested in exploring China's official communication.

5.3 Centrality of the US in China's narrative surrounding the war in Ukraine

China's response to the contradiction in its stance on the war in Ukraine is argued to have been two-fold. On the one hand, in the early months of the war, China sought to demonstrate its neutrality and independence by avoiding any direct statements of support or approval for Russian actions in Ukraine, while also indirectly providing diplomatic cover via expressions of "its understanding of Moscow's interpretation of the origins of this conflict" (Trush, 2022). On the other hand, it was already clear at an early stage in the war that the way China had chosen to manage the conflict between its support for sovereignty and territorial integrity on the one hand, and its partnership with Russia on the other, was to frame the war as a result of US domination of the international system. This narrative can be observed in a plain look at official discourse (McCarthy, 2022) and on Chinese social media (Repnikova and Zhou, 2022). Several observers agree that in the early months of the war, official Chinese statements focused on the responsibility of the West and NATO as the main line of its narrative (Repnikova, 2022).

SNA analysis of China's official statements supports the latter argument, demonstrating that the "neutrality" aspect was overshadowed by blame shifting on the US. China's communications regarding the war in Ukraine have been consistent in terms of the language they employ. The US was frequently said to be "adding fuel to the fire" or "fanning the flames," while references to the "US" with respect to the war in Ukraine occur 350 times in the sampled statements from 2022, and 83 times in 2023. While the Chinese narrative framing of the US role in the war in Ukraine, taken at face value, indicates support for Russia's position on the causes of the conflict, if we take a step back to consider the bigger picture, other, more nuanced, aspects of the Chinese viewpoint come into view.

To start with, the results of the SNA show that the central axis in China's official discourse regarding Ukraine is, in fact, US-China bilateral relations. This is evident in 2022 (*Figure 1*) and to a lesser extent in 2023 (*Figure 2*). In 2022, the link between "the US" and "we", i.e., China, is the strongest – the blue edge in the semantic network visualization is the thickest, indicating that these two words appear next to each other more frequently than any other dyad of words. For China, the war in Ukraine provides a context or background where it can present the US as a culprit or instigator of conflicts, and China as playing a constructive role in the world. As can be seen in the semantic networks, "the US" is linked with negative words, such as Cold War, invasion, hegemon, confrontation, and flames (part of the phrase "fanning the flames"). By repeatedly using the phrases suggestive of arson – e.g., "fanning the flames" or "adding fuel" – China is able to seed the idea of the US as a malicious actor, stirring trouble for its own ends. In turn, the Chinese narrative attributes this tendency in US foreign policy to a lingering "Cold War mentality," as well as double standards when it comes to conflicts in general; Afghanistan, Iraq, Syria, and the Federal Republic of Yugoslavia (referring to the Yugoslav wars in the 1990s) are among the examples presented by the spokesperson. These country names appear on the semantic networks for both 2022 (*Figure* 1) and 2023 (*Figure* 2). Thus, the narrative of the US as an instigator of conflicts around the world clearly emerges from Chinese messaging on Ukraine in 2022, and to a lesser but still significant degree in 2023.

5.4 The basis of China's stance: moral condemnation of the US

China framed itself as constructive and the US as destructive, driven by Cold War mentality and bloc politics, apparently seeking to present moral condemnation of the US. *Figure 3* below presents the original semantic network for the 2022 corpus before network sparsification. It shows that the strongest connection in the network (i.e., the thickest edge) is "the US-should," which often in turn link to the term "Cold War" (as in "should discard the Cold War mentality"). Using moral and normative language, mostly the modal verb "should," and the image of the destructive US as shown above, foreign ministry spokespersons frame the situation in terms of morality.

Such a narrative reflects Rathburn's (2023) argument that morality matters in international relations, particularly when states feel other actors have wronged them. States may have a moral self-perception as "right" acting in accordance with certain norms but they hold a moral perception of being mistreated by others. And this perception binds a group together against that "immoral" actor in the international arena.

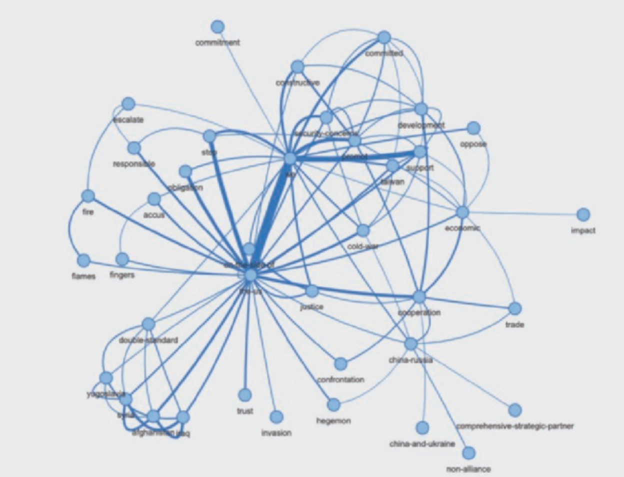


Fig. 1 Semantic network for the 2022 dataset after network sparsification.
This semantic network is the original network for the 2022 dataset (created based on the original list of 47 words in *Appendix 1*; showing 100 most prevalent links) after sparsification, i.e., the removal of less common words. The thickness of the edges (blue lines) represents the strength of the links between the edges – two words – it connects.

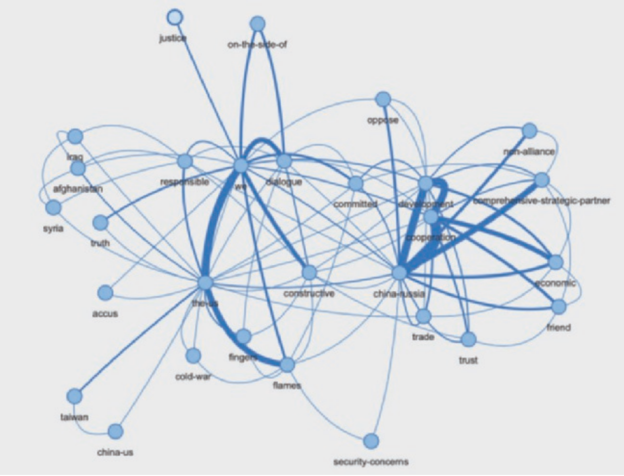


Fig. 2 Semantic network for the 2023 dataset after network sparsification.
This semantic network is the original network for the 2023 dataset (created based on the original list of 47 words in *Appendix 1*; showing 100 most prevalent links) after network sparsification, i.e., the removal of less common words. The thickness of the edges (blue lines) represents the strength of the links between the edges – two words – it connects.

It is this kind of "binding morality," which brings states together and upholds their loyalty to that group (Rathburn, 2023), not Beijing's close relations with Moscow, that China presents as the basis for its stance on the war in Ukraine. And this is further reinforced by communication on China-Russia relations. In the 2023 corpus, Sino-Russian relations emerged as a distinct theme in addition to the China *vs*. the US narrative (*Figure 2* above). At that time, "China and Russia" or "China-Russia" occurred at least 72 times. The framing of bilateral relations between the two countries is consistently positive,

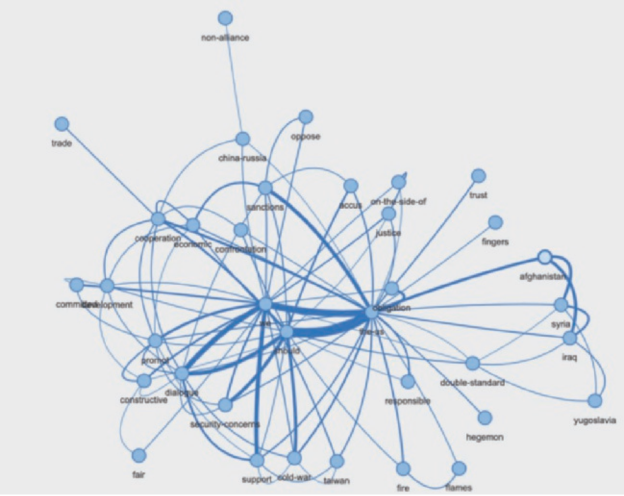


Fig. 3 Original semantic network for the 2022 corpus before network sparsification.
This semantic network is created based on the original list of 47 words (*Appendix 1*) and shows 100 most prevalent links.

focusing on elements like cooperation, development, economic ties (including trade), friendship and trust, and security. This is contrasted with the negative image of the US, linked with "fanning the flames" and different conflicts across the world. On the other hand, China-Russia cooperation in the framework of a comprehensive strategic partnership sits alongside references to "non-alignment." This trend underscores the pragmatic nature of Chinese foreign policy but also emphasizes that alliance politics is not what keeps China from condemning Moscow's actions against Ukraine. That is, while political leaders in Western countries repeatedly called on China to condemn Russia or to exert more pressure on it over the war in Ukraine, China is bound together with Russia based on moral principles, and alongside that, it advances its own pragmatic interests.

5.5 China's changing tone on the terms of settlement: "dialogue" replaces "justice"

SNA analysis of corpora across time enables us to observe some nuanced changes in China's stance and even draw inferences about the terms of settlement of the war that China may support. In 2022, the phrase "on the side of" was linked with the word "justice," but in the following year "justice" in the context of the war in Ukraine disappeared. Instead, a new correlation between "on the side of" and "dialogue" emerged. A look at the datasets reveals that Chinese foreign ministry spokesperson's phrase "stand on the side of peace and justice," mentioned five times in 2022 (where it defines China's position three times and twice calls for the US to do so) indeed was replaced by a different one "stand on the side of peace and dialogue," which, in 2023, was also linked with the "right side of history." That phrase was present in the 2022 corpus but was not particularly related to any other words.

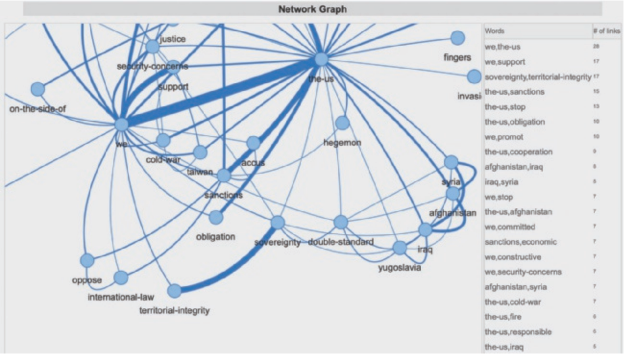
The fading of "justice" as a solution for the war narrative is significant. In *China's Position on the Political Settlement of the Ukraine Crisis*, published on the MFA website on February 24, 2023, i.e., at the same time as the 2023 corpus of this analysis, it is stated that "All parties should jointly uphold the basic norms governing international relations and defend international fairness and *justice*. Equal and uniform application of international law should be promoted, while double standards must be rejected" (*Italics* added by the author). Nevertheless, the context here implies justice of the international system, not for Ukraine as such. In the 2022 corpus, such a broad international context was also present, but it was linked to China's position on the Ukraine war too: "China plays a constructive role in the Ukraine issue. We speak for *justice* and work for peace with a long-term vision." In 2022, China repeatedly emphasized that "the Ukraine issue has a complex historical context. On this issue, China has all along upheld an objective and *just* position," but, in 2023, the emphasis shifted to "dialogue." On the one hand, this suggests that China's stance on the war in Ukraine has become more cautious. On the other hand, such a subtle change in the narrative also represents China's new more active position on the issue. In *China's Position on the Political Settlement of the Ukraine Crisis* published exactly one year after the start of the invasion, China outlines specific points that, in Beijing's view, define "political settlement" conditions.

6. Will computational analysis replace human researchers?

As computational text analysis advances and more research tools are developed almost on a daily basis, it becomes clear that no human analyst can ever outperform a computer. There are some cases, when large data sets are gathered and coded manually. For instance, Yang and Yang (2025) scraped from the website and coded manually 4,556 Q&A dyads in Chinese from MFA press conferences (four years in total) to demonstrate that China's "wolf warrior diplomacy" approach significantly increases due to aggressive questioning from foreign journalists. While this is an impressive dataset for manual analysis, yet it is nevertheless extremely demanding both in terms of cost and effort. But automated or computer-assisted text analysis can process the size and volumes of text corpora far beyond human ability in terms of minutes not days. In such large text corpora, a machine can identify patterns spreading across all the text, group it along machine-identified topics, and offer the researcher a new way to look at the way they look at those text documents. Moreover, some recent analyses on China suggest that computational text analysis using a text-as-data approach can help overcome existing bias (e.g., Kang et al., 2025). Thus, there naturally emerges a question whether computational text data will replace human researchers.

The answer in the literature to date is straightforward – such methods "augment humans, not replace them" (Grimmer and Stewart, 2013). Marcellino (2023) highlights that there is a fundamental tension between a human and machine approach to text. While machines are good at finding patterns throughout large corpora, humans have an advantage in close reading of individual texts to make sense of them (Marcellino, 2023). And this is exactly how machines augment human capacity but are not able to replace a human researcher. Automated analysis introduces an additional step of comprehensive large text data analysis allowing human researchers to later make sense of it.

A concrete case to illustrate this argument can be drawn from the above case study on China's communication in the context of Russia's war in Ukraine. SNA of the two corpora from 2022 and 2023 revealed that, across the two corpora, China's emphasis on territorial integrity diminished. The stated principle of "respect for sovereignty and territorial integrity" has been one of the fundamental tenets of Chinese foreign policy since the establishment of the People's Republic of China (PRC) in 1949. The so-called *Five Principles of Peaceful Coexistence*, which in Beijing's own words "are not only the basis of the Chinese independent foreign policy of peace, but also constitute important principles in regulating state-to-state relations," explicitly outline "mutual respect for sovereignty and territorial integrity." Thus, it comes as no surprise that Beijing repeatedly emphasized its continued commitment to these principles. The official statement issued by the foreign minister Wang Yi, entitled *China's Five-Point Position on the Current Ukraine Issue*, two days after the start of the Russian invasion, stated that: "1. China maintains that the *sovereignty and territorial integrity* of all countries should be respected and protected and the purposes and principles of the UN Charter abided by in real earnest. This position of China is consistent and clear-cut, and *applies equally to the Ukraine issue*" (*Italics* added for emphasis). During the first few weeks of the war, the spokesperson of the foreign ministry regularly referred to the "sovereignty and territorial integrity" principle, usually in reference to "all states" rather than Ukraine specifically. In the corpus for 2022, this phrase recurs seventeen times. As the *Figure 4* below shows, the "sovereignty-territorial integrity" link was one of the strongest in the dataset.



Words	# of links
we,the-us	28
we,support	17
sovereignty,territorial-integrity	17
the-us,sanctions	15
the-us,stop	13
the-us,obligation	10
we,promot	10
the-us,cooperation	9
afghanistan,iraq	8
iraq,syria	8
we,stop	7
the-us,afghanistan	7
we,committed	7
sanctions,economic	7
we,constructive	7
we,security-concerns	7
afghanistan,syria	7
the-us,cold-war	7
the-us,fire	6
the-us,responsible	6
the-us,iraq	5

Fig. 4 A fragment of the semantic network for the 2022 corpus before network sparsification.
(*source:* Created by the author).

While communication in 2022 is in line with China's officially stated principle, it is important to highlight a semantic change in the 2023 corpus. There, "sovereignty" is mentioned by the foreign ministry spokesperson six times, three of them as "sovereignty of all states." But in this phrase, the term "territory" is missing. The same is true for the Chinese-language transcripts of the same press conferences on the MFA website. That is, in 2023, the term "sovereignty and territorial integrity" (*zhuquan he lingtu wanzheng* 主权和领土完整) appears to diminish to "sovereignty" only (*zhuquan* 主权). In 2023, the term "territorial integrity" has all but disappeared from the official language of the Chinese foreign ministry spokesperson at the regular press conferences.

Based solely on the SNA results, China was adjusting its narrative to accommodate the potential settlement of the war, where Ukraine would have to cede territory to Russia. As the Russian invasion of Ukraine began in February 2022, observers of China in the West wondered how Beijing would reconcile its strategic partnership with Russia – which had just violated Ukraine's sovereignty and territorial integrity – with its long-standing foreign policy principles. On the one hand, Chinese officials sought to do so by stressing the wrongdoing of NATO with its eastward expansion after the Cold War had ended. On the other hand, from this SNA result it may have been concluded that Beijing had quietly altered the wording of its principles, which now omitted the phrase "territorial integrity," This would be a logical explanation considering that, toward the end of the first year of war in Ukraine, Russia controlled 18 percent of Ukraine's territory (Allison and Davidson, 2023). In September

2022, Russia annexed four eastern provinces of Ukraine, and Russian President Vladimir Putin declared Russia would use its nuclear arsenal to defend those territories. In that context, the wording of *China's Position on the Political Settlement of the Ukraine Crisis*, China's "peace plan" presented one year after the start of the war, also stands out. In the document, the usual wording of China's principles – "respect for sovereignty and territorial integrity" – is changed to "respecting the sovereignty of all countries." Furthermore, "territorial integrity" is only mentioned in the third sentence of the peace proposal after the words "sovereignty" and "independence," thus giving the appearance of its being of secondary importance to the latter two.[1]

Nevertheless, one of the key principles when using automated text analysis is validation of its results. When applied to a specific problem, some analytical models may provide misleading or incorrect results. "Therefore, it is incumbent upon the researcher to validate their use of automated text analysis" and blind use of any method without validation should be avoided (Grimmer and Stewart, 2013). Validation can be performed in different ways. Literature (Grimmer and Stewart, 2013) suggests complex models of validation, that sometimes combine experimental, substantive, and statistical evidence. Here in our case study, which was based on a relatively small corpus, interpretative analysis through manual content analysis, commonly used in China studies, is sufficient, yet necessary, to verify and cross-check the results of the SNA.

Manual content analysis of the datasets does not extend firm support for the output of the SNA. Two out of six times, when "sovereignty" was mentioned, the spokesperson referred to the so-called peace plan by China, which "offers 12 propositions, including respecting the sovereignty of all countries." In other cases in this corpus, the word appeared in a different context, which also underscores the importance of a larger corpus for making reliable observations. And that is not sufficient to draw a reliable conclusion. Furthermore, the Chinese readout published by the Chinese MFA after President Xi Jinping's call to Ukrainian President Volodymyr Zelensky explicitly states that "Mutual respect for sovereignty and territorial integrity is the political foundation of China-Ukraine relations," thus reaffirming that stance of China at the highest level.

That is to say that automated text analysis offers a significant advantage and augmentation to human analysis as it offers efficiency, yet automated methods "are no substitute for careful thought and close reading and require extensive and problem-specific validation" (Grimmer and Stewart, 2013). Here, a careful combination of computational and manual approaches can maintain the strengths of traditional content analysis at the same time maximizing the capacity of computational analysis (Lewis et al., 2013).

7. Conclusion

This paper attempted to shed light on the potential of the text-as-data approach in China studies. As access to the field has narrowed over the past decade and more digital text material has been produced in China, the advance of computational text analysis tools has opened a promising avenue for the research of Chinese domestic politics as well as foreign policy. Nearly any issue can be explored using the text-as-data approach, and the tasks lying ahead for China scholars seem to be testing and validating already existing conclusions about China using big textual data.

The first part of this paper presented the text-as-data approach in reference to existing research. It underscored the advantage of this approach in conducting a comprehensive analysis of extensive text data and its potential in discovering new patterns.

The pilot case study using SNA of China's communication in the context of the war in Ukraine, presented in the latter part of the paper, sought to illustrate how narratives are detected in the text and how their prevalence can be measured. It showed that the centrality of the US and China in the narrative is also accompanied by China's framing of the situation in terms of "binding morality," where it refuses to renounce Russia's war of aggression, not due to its support for Russia but because of China's perceived injustice from the side of the US. While this narrative can possibly be grasped through an interpretive study of text data, the SNA shows its centrality and also reveals the normative dimension of China's official communication. SNA of the data from 2022 and 2023 also detected a nuanced but significant change in China's official communication that could provide insight into China's expected settlement of the war, as it shifted away from "justice" toward "dialogue" as a solution to the war.

This analysis also made a case for the argument that computational analysis cannot replace human researchers. Needless to say, it offers an extraordinary improvement beyond human capacity, but the knowledge that China scholars have accumulated over the years of their careful study of China is essential for drawing valid conclusions about the real world.

8. Acknowledgments

This work has been supported by JSPS KAKENHI Grant Number 20K13441.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the author's organization, JSPS or MEXT.

9. References

Allisson, G. and Davidson, K. (2023) "Russia-Ukraine Report Card", February 23, 2023, *Belfer Center for Science and International Affairs, Harvard Kennedy School*, https://www.belfercenter.org/publication/russia-ukraine-report-card (2026年1月3日アクセス)

Benoit, K. (2020) "Text as data: An overview", In Curini, L., Franzese, R. (Eds.) *SAGE Handbook of Research Methods in Political Science and International Relations*, London: SAGE, p.461-497.

Dai, Y. and Luqiu, W. R. (2024) *Wolf Warrior Diplomacy and China's Ministry of Foreign Affairs: From Policy to Podium*, Lanham, MD: Lexington Books.

Economist (2023) "It is Getting even Harder for Western Scholars to do Research in China", 5 April, https://www.economist.com/china/2023/04/05/it-is-getting-even-harder-for-western-scholars-to-do-research-in-china (2026年1月3日アクセス)

Fisher, S., Klein, G. R., Codjo, J. (2022) "Focusdata: Foreign Policy through Language and Sentiment, *Foreign Policy Analysis,* 18 (2).

Grimmer, J., Roberts, M. E., Stewart, B. M. (2022) *Text as Data: A New Framework for Machine Learning and the Social Sciences*, Ofxorf University Press.

Grimmer, J., and Stewart, B. M. (2013) "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts", *Political Analysis*, 21, p.267-297.

Guo, L. and Qin, Q. (2024) "Wolf Warrior Spreads Superior: The Narrative and Effectiveness of Chinese Public Diplomacy Behaviours on Twitter", *Journal of Chinese Political Science*.

Kang, D. J., Wong, S. H., Chan, Z. T. (2025) "What Does China Want?", *International Security* 50 (1), p.46-81.

King, G. and Lowe, W. (2003) "An Automated Information Extraction Tool for International Conflict Data", *International Organization* 57, p.617-642.

King, G., Pan, J., Roberts, M. E. (2017) "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction", *American Political Science Review,* 111 (3), p.484-501.

Laver, M., Benoit, K., Garry, J. (2003) "Extracting Policy Positions from Political Texts Using Words as Data", *The American Political Science Review* 97 (2), p.311-331.

Lewis, S. C., Zamith, R., Hermida, A. (2013) "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods", *Journal of Broadcasting & Electronic Media,* 57 (1), p.34-52.

Li, T., Ke, X., Shi, H. (2025) "Topic Modeling and Evolutionary Trends of China's Language Policy: A LDA-ARIMA Approach", *PLoS One*, 20 (5).

Liebman, B. L., Roberts, M. E., Stern, R. E., Wang, A. Z. (2020) "Mass Digitization of Chinese Court Decisions: How to Use Text as Data in the Field of Chinese Law", *Journal of Law and Courts*. p.177-201.

Lim, J., Ito, A., Zhang, H. (2025) "Uncovering Xi Jinping's Policy Agenda: Text-As-Data Approach", *The Developing Economies,* 63 (1), p.9-46.

Marcellino, W. (2023) "Public Policy Research Applications of DocuScope's Linguistic Taxonomy Mining Style and Stance for Sociocultural Insight", in Brown, D. W, Wetzel D. Z. (eds.) *Corpora and Rhetorically Informed Text Analysis: The Diverse Applications of DocuScope (Studies in Corpus Linguistics, 109),* John Benjamins Pub Co.

Mattingly, D., Changwook, J., Moreshead, C., Tanaka, S., Yamagaishi, H. (2024) "Chinese State Media Persuades a Global Audience that the "China model", is Superior: Evidence from a 19-country Experiment", *American Journal of Political Science,* 69 (3), p.1029-1046.

McCarthy, S. (2022) "As War Breaks out in Europe, China Blames the US", *CNN*, 25 February, https://edition.cnn.com/2022/02/25/china/china-reaction-ukraine-russia-intl-hnk-mic/index.html (2026年1月3日アクセス)

Messingschlager, S. (2025) "Decoding Xi's China: The Return of Pekingology", *The Interpreter*, 1 July, https://www.lowyinstitute.org/the-interpreter/decoding-xi-s-china-return-pekingology (2026年1月3日アクセス)

Ministry of Foreign Affairs of the PRC (2023) "China's Position on the Political Settlement of the Ukraine Crisis", 20 March, https://www.mfa.gov.cn/eng/zy/gb/202405/t20240531_11367485.html (2026年1月3日アクセス)

Ministry of Foreign Affairs of the PRC (2022) "Wang Yi Expounds China's Five-Point Position on the Current Ukraine Issue", 26 February, https://www.fmprc.gov.cn/eng/zxxx_662805/202202/t20220226_10645855.html (2026年1月3日アクセス)

Mochtak, M. and Turcsanyi, R. (2021) "Studying Chinese Foreign Policy Narratives: Introducing the Ministry of Foreign Affairs Press Conferences Corpus", *Journal of Chinese Political Science* 26 (4), p.743-761.

Parton, C. (2022) "China Watching in a 'New Era': A Guide", *Council on Geostrategy,* https://www.geostrategy.org.uk/research/china-watching-in-the-new-era-a-guide/ (2026年1月3日アクセス)

Presidential Executive Office (2022) "Joint Statement of the Russian Federation and the People's Republic of China on the International Relations Entering a New Era and the Global Sustainable Development", February 4, 2022, http://en.kremlin.ru/supplement/5770 (2026年1月3日アクセス)

Rahal, C., Verhagen, M., Kirk, D. (2024) "The Rise of Machine Learning in the Academic Social Sciences", *AI & Society,* 39, p.799-801.

Rathburn, B. C. (2023) *Right and Wronged in International Relations: Evolutionary Ethics, Moral Revolutions, and the Nature of Power Politics*, Cambridge University Press.

Repnikova, M. (2022) "China's Propaganda on the War in Ukraine", *China Leadership Monitor,* 72.

Repnikova, M. and Zhou, W. (2022) "What China's Social Media Is Saying About Ukraine", *The Atlantic*, 11 March, https://www.theatlantic.com/ideas/archive/2022/03/china-xi-ukraine-war-america/627028/ (2026年1月3日アクセス)

Segev, E. (2021) *Semantic Network Analysis for Social Sciences*, Routledge.

Stone, P. J. (1962) "The General Inquirer: A Computer System for Content Analysis and Retrieval", *Behavioral Science,* 7 (4), p.484-498.

Stone, P. J., Dunphy, D. C., Smith, M. S., Ogilvie, D. M. (1966) *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press.

Tao, Y., Zhan, Z., Zhou, H., Kang, J., Sun, S. (2024) "Measuring Chinese Online Populist Discourse: An Automated Semantic Text Analysis Method", *Chinese Journal of Communication,* 18 (2), p.121-141.

Trush, S. M. (2022) "Crisis between Russia and Ukraine: The China Factor", *Herald of the Russian Academy of Sciences,* 92 (7), p. S595-S600.

Weiss, J. C. (2019) "A World Safe for Autocracy: China's Rise and the Future of Global Politics", *Foreign Affairs,* 98(4), p.92-102.

Weston, M. J. and Rauchfleisch, A. (2021) "Close to Beijing: Geographic Biases in People's Daily", *Media and Communication,* 9 (3), p.59-73.

Yang, L. and Yang, H. (2025) Aggressive Journalistic Questioning and China's "Wolf Warrior Diplomacy", The China Quarterly, 263, p.758-775. doi:10.1017/S0305741025000396

Zhang, C., Zhang, D., Shao, H. L. (2023) "The Softening of Chinese Digital Propaganda: Evidence from the People's Daily Weibo

Account during the Pandemic", *Front. Psychol,* 14.

Note

[1] The first point of *China's Position on the Political Settlement of the Ukraine Crisis* reads as follows. "1. Respecting the sovereignty of all countries. Universally recognized international law, including the purposes and principles of the United Nations Charter, must be strictly observed. The sovereignty, independence and territorial integrity of all countries must be effectively upheld. All countries, big or small, strong or weak, rich or poor, are equal members of the international community. All parties should jointly uphold the basic norms governing international relations and defend international fairness and justice. Equal and uniform application of international law should be promoted, while double standards must be rejected."

〔受付日　2025. 10. 10〕

Appendix. List of words selected for the SNA of the 2022 and 2023 corpora.

Some words have been combined into phrases, while others are shortened to their stems to capture different words sharing the same root (e.g., “accus-”: accuse, accusing, accusation).

	Search-word in the text	**Original words and phrases in the text**
1.	accus	accusations; accusing; accuse
2.	afghanistan	Afghanistan
3.	china-and-ukraine	China and Ukraine
4.	china-russia	China-Russia; China and Russia
5.	china-us	China-the US; China and the US
6.	cold-war	Cold War
7.	commitment	
8.	committed	
9.	comprehensive-strategic-partner	comprehensive strategicpartnership
10.	confrontation	
11.	constructive	
12.	cooperation	
13.	democracy	
14.	development	
15.	dialogue	
16.	double-standard	double standard
17.	economic	
18.	escalate	
19.	fair	
20.	fingers	
21.	fire	
22.	flames	
23.	friend	
24.	hegemon	Hegemon; hegemony
25.	impact	
26.	invasion	
27.	iraq	Iraq
28.	justice	
29.	non-alliance	note: also covers *no-alliance* used once in the same context
30.	obligation	
31.	on-the-side-of	on the side of
32.	oppose	
33.	promot	promote, promotion, promoting
34.	putin	
35.	responsible	
36.	sanctions	
37.	security-concerns	security concerns
38.	should	
39.	support	
40.	syria	Syria
41.	taiwan	Taiwan
42.	the-us	the US
43.	trade	
44.	trust	
45.	truth	
46.	we	
47.	yugoslavia	Yugoslavia