Working Paper
SWP Working Papers are online publications within the purview of the respective Research Division. Unlike SWP Research Papers and SWP Comments they are not reviewed by the Institute.

SWP

RESEARCH DIVISION ASIA | WP NR. 01, JUNE 2019

Introducing UNSCdeb8 (beta)

A Database for Corpus-Driven Research on the United Nations Security Council[1]

Paul J. Kohlenberg, Nadine Godehardt, Stephen Aris, Fred Sündermann, Aglaya Snetkov, and Juliet Fall

[1] UNSCdeb8 (pronounced *UNSC debate*) was developed within the project "*Which region? The politics of the UN Security Council P5 in inter-national security crises*," jointly run by the Center for Security Studies at ETH Zürich, the Department of Geography and Environment at the University of Geneva and the German Institute for International and Security Affairs (SWP). The project is co-funded by the Swiss National Science Foundation (SNF, Project Number: 162925) and the Deutsche Forschungsgemeinschaft (DFG, Project Number: 28295967). Fred Sündermann wrote the algorithms for compiling the database and this research project would not have been possible without him. We also thank David Schulze, Ilja Sperling as well as the participants of the RUN-2018 workshop in Geneva for their valuable support and feedback. This working paper is archived in Zenodo (doi: 10.5281/zenodo.3234498).

Stiftung Wissenschaft und Politik
German Institute for International and Security Affairs

Contents

1. Introducing UNSCdeb8

UNSCdeb8 (pronounced *UNSC debate*) is a database of all verbatim statements given by the representatives of permanent and non-permanent members during public meetings of the United Nations Security Council (UNSC) between 2010 and 2017. It is comprised of all the records of these statements that have been made publically available by the UN, in either their English original or the UN's translation of a statement into English. All the textual data in UNSCdeb8 is aligned with metadata about the particular statement and meeting in which it was articulated, such as speaker-nationality, speaker-role, date/time, meeting number and agenda title of the meeting. It is primarily designed to facilitate corpus-driven social science research, but can also be used for other methodological approaches.

UNSCdeb8 is accessible via a web-based query tool (https://unscdeb8.swp-berlin.org & unscdeb8.ethz.ch) that enables researchers to quickly and intuitively assemble sub-corpora of UNSC-debates, based on the metadata-filters they are interested in. Users have the possibility to analyse and compare UN Security Council debates and then jump back to the web-tool to assemble new sub-corpora based on their previous insights. It is thereby possible, for example, to use UNSCdeb8 in tandem with corpus-linguistic analysis software such as Antconc or Wordsmith,[2] in order to calculate the most statistically significant keywords that differentiate the members of the UNSC. Similarly, one can examine how the discourse of a particular UNSC member evolves over time or changes with regard to specific thematic issues. Put differently, it is possible to comparatively analyse UNSC debates according to varied factors and perspectives, from the evolution of UNSC member-states' discourse in general to tracing more subtle discursive changes within a particular member state's outlook on a particular crisis within the eight-year period covered by the database.

UNSCdeb8 currently consists of 6.8 million-word tokens. In the future, yearly updates are planned to continuously expand the corpus by including the most recent 12 months of debates.

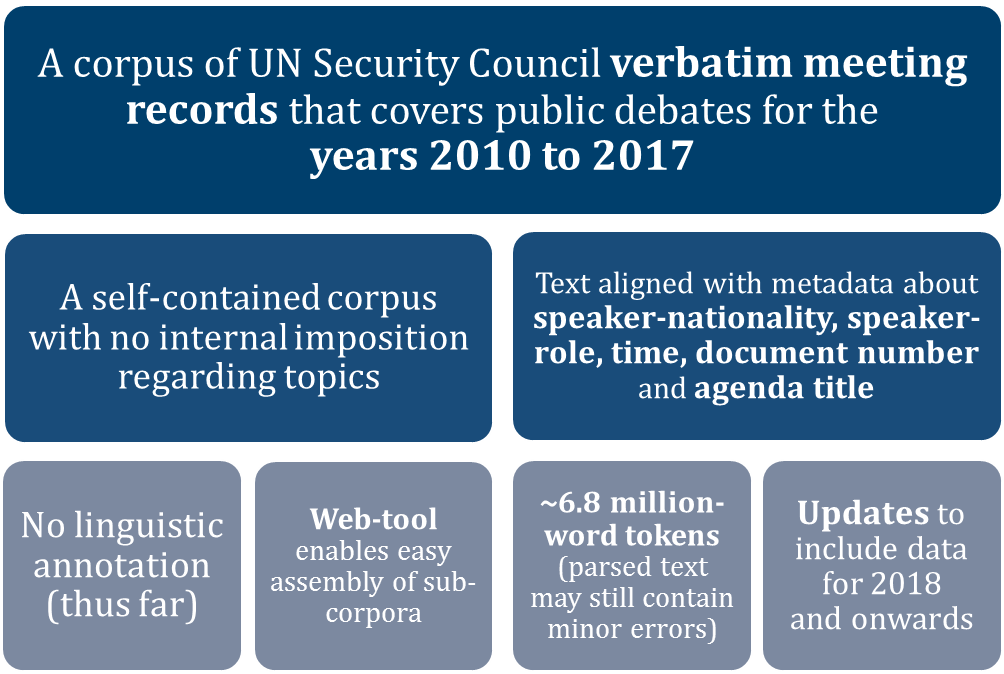


Figure 1: UNSCdeb8 (beta) at-a-glance

[2] Anthony, L. (2014) AntConc (Version 3.4. 3) [Computer Software]. Tokyo, Japan: Waseda University.

Following best-practice guidelines for corpus building, the database is assembled from a single and continuous source of documents, i.e. verbatim statements that are publicly available from the UNSC across the eight-year timeframe (i.e. a "specialized corpus"[3]). It is, thus, a self-contained corpus of verbatim UNSC statements, with no external pre-selections, interventions in or mark-up of this textual data.

It is important to note that the database contains statements by diplomats and official representatives of permanent and non-permanent member states of the UNSC. **Statements by invited guests, special rapporteurs, witnesses, or UN leadership personnel are not included in the database**. UNSCdeb8 can, therefore, not be used to reconstruct individual debates in their entirety.[4]

The corpus primarily lends itself to comparisons of statements by particular UNSC members-states over time, especially its permanent members. Generally speaking, the oftentimes highly formulaic language employed by UNSC diplomats is conducive to the lexicometric tracing of small differences over longer periods of time, which could hardly be spotted by any close (qualitative) reading of these statements.

First, we outline functionality of UNSCdeb8 and how it can contribute to the study of important questions about international politics? Second, we briefly introduce our understanding of discourse, the relevance of the database to our research, and our method of analysing discourse, with a particular focus on corpus-linguistic research. Third, we illustrate how scholars can make use of UNSCdeb8, by introducing examples from our research on Chinese discourse in the UNSC. This introduction is geared towards an audience of social science researchers that depend on software tools with graphical user interfaces (GUI). It is nonetheless important to emphasize that the UNSCdeb8 database can be usefully employed for a great variety of other scholarly approaches and methods – with or without a GUI.

2. The value and functionality of UNSCdeb8 and how it was developed

The database is a product of the collaborative research project "Which region? The politics of the UN Security Council P5 in international security crises", jointly run by Center for Security Studies at ETH Zürich, Department of Geography and Environment at the University of Geneva and SWP Berlin. The project is co-funded by the Swiss National Science Foundation (SNF) and the Deutsche Forschungsgemeinschaft (DFG)[5]. In this project, we are interested in analysing the use of spatial terms to justify, explain and dispute responses to international security crises within UNSC debates. We seek to investigate this topic using corpus linguistic analysis. However, we discovered that there were a number of challenges to conducing such analysis of UNSC debates, which the development of UNSCdeb8 serves to overcome.

Primarily, we have constructed this new database of UNSC debates, because the transcripts for UNSC debates are currently only available in variously formatted PDF-files that can be downloaded from the official UN website(s). The UN's existing search tools do not allow a user to differentiate or filter textual data within a given PDF (i.e. a particular UNSC meeting). Therefore, it is not possible to gather textual data of UNSC debates using key parameters, such as speaker nationality or keywords. As a result, research into UNSC debates has until now been limited by the search and data collation restrictions of the UNSC official website. In other words, while the PDF transcripts of UNSC debates are already in

[3] Hunston, S. (2002) Corpora in applied linguistics. Ernst Klett Sprachen.

[4] Full (PDF) transcripts of individual debates are readily available on undocs.org.

[5] SNF Project Number: 162925; DFG Project Number: 28295967

the public domain, they are largely unsuited for computational processing. Therefore, lexicometric research on UNSC debates is constrained to analysis of the totality of words spoken in a given debate. For example, it is possible to compile a corpus of statements by a specific nation across a number of UNSC meeting. However, this would have to be done manually, by extracting the required text from each PDF for a specific meeting, cleaning it and collating it for further analysis. To overcome this obstacle, the UNSCdeb8 database offers researchers a tool with which to search, sort, and gather text from UNSC debates through a simple interface, opening up new opportunities for analysis of the UNSC.

Relevance of UNSCdeb8 to research on international politics

The ability to more easily and flexibly analysis UNSC debates using UNSCdeb8 is an important resource for scholars of IR and other disciplines, because the Security Council is arguably the most high-profile body in international politics. Key debates receive broad coverage. Contentious statements by the council's ambassadors reverberate across all levels of politics and shape the perceptions of actors across international organizations, regions and nations. The Security Council mostly focuses on security crises, which tend to be moments in which the political and territorial status quo is challenged and (re)negotiated. Irrespective of whether it is focused on an interstate war, civil war, transnational conflict or some more issue-specific topic, debates in the UNSC regularly produce key initial statements by high ranking diplomats about political hotspots in a state of flux. As opposed to the diverse and often ad hoc responses of officials to the media or in research interviews, statements made to the Security Council are generally not made lightly. They are often repetitively phrased and changes in the use of language are much more likely to reflect conscious or unconscious policy (or strategy) changes, rather than mere happenstance.

Security Council debates, thus, represent a useful resource for academic inquiry into shifts, similarities and divergences in how important international actors seek to represent international affairs, especially when analysed over a longer time-period, or across multiple topical areas. Indeed, some scholars have conducted small scale corpus linguistics analyses of the UNSC, but they have focused primarily on the language of resolutions and not debates.[6] In the contrast to research on the UNSC, researchers have already parsed large amounts of data from the UN General Assembly. Most of these works have concentrated on certain types of debate, and thus have only parsed data from those instances. One such example is the UN General Debate Corpus (UNGDC), which contains all statements made in the annual General Debate between 1970 and 2014.[7] While other corpora of UN documentation have been compiled, these have not been developed to support social science research. Most notably, the United Nations Parallel Corpus[8], which is designed to serve the needs of linguists interested in advancing the cause of machine translations, due to the assumption that UN resolutions and debates are translated with great care and professionalism. Hence, this corpus is not conducive for asking social science questions about actors, conflicts or intentions. Overall, there have been only a few attempts to develop textual corpora about the UNSC that would facilitate corpus linguistic analysis. We posit that an important technical obstacle to such work is the uneven nature

[6] E.g.: DiCarlo, G. (2011) "Indeterminacy and Vagueness in UN Resolutions Relating to the Second Gulf War." International Journal of Humanities and Social Sciences 1.21:, p. 46-57.

[7] Baturo, A., N. Dasandi and S. Mikhaylov (2017). "Understanding State Preferences with Text As Data: Introducing the UN General Debate Corpus" Research and Politics 4(2):2053168017712821.

[8] Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B., (2016) "The United Nations Parallel Corpus", Language Resources and Evaluation (LREC'16), Portorož, Slovenia.

of the PDF-formatting in the published records of UNSC meeting that present numerous obstacles to algorithmic parsing. UNSCdeb8 has been built to overcome this obstacle.

Against this background, UNSCdeb8 can be said to expand the toolbox available for researchers of the UN Security Council, who are interested in asking questions that relate to the fields of political science, IR and cognisant disciplines. Social scientists, particularly those interested in discourse theory, tend to work with inductive or "corpus-driven" methods, as compared to the more deductive and positivist "corpus-based" approaches of linguistics.[9] While the former approach is often based on the examination of a single corpus that represents a discourse in its totality, the latter tends to be better served by the development of multiple sub-corpora to compare and contrast the varied discourse of speakers (actors), topics and time-periods within the meta-discourse. In other words, we suggest that if social-scientific studies limit themselves to building only one single (static) corpus of text in order to represent a particular discourse in its totality, they may be following too closely in the footsteps of established linguistic approaches.[10] For social science inquiry, it is usually more interesting to be able to differentiate between and thus compare smaller and multiple segments of a discourse, rather than studying the totality of a given discourse.

To this end, some corpora resources allow a user to flexibly filter the main corpus to create smaller sub-corpora, which can then be studied in comparison to one another. Similar to the composition of UNSCdeb8, diachronic corpora such as the Corpus of Historical American English[11] and online applications – e.g. Google's Ngram Viewer – are well-suited to research concerned with comparative analysis and the study of the evolution of language over time. As opposed to synchronic, diachronic linguistic analysis does not examine language-use and meaning-making from the perspective of a singular moment in time, but rather traces its shifting development across time. Put differently, it is a genealogical tracing of meaning-structures within a discourse.

UNSCdeb8 web-based query tool

UNSCdeb8 combines such diachronic functionality with additional filters for actors (speaker-nation) and agenda. The ability to quickly and flexibly define the parameters to create a plethora of subcorpora from UNSC discourse is, thus, a defining aspect of UNSCdeb8. Coming back to the example of diachronic analysis, it thus becomes possible to inductively study not only when a certain vocabulary became more or less frequent, but also to identify which actors were involved in the discourse and in which contexts of UNSC debates the terminology appeared.

[9] D. Biber (2009) "Corpus-Based and Corpus-Driven Analyses of Language Variation and Use" in The Oxford Handbook of Linguistic Analysis, 17 December 2009, https://doi.org/10.1093/oxfordhb/9780199544004.013.0008.

[10] P. Baker, C. Gabrielatos, and T. McEnery (2012) "Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word "Muslim" in the British Press 1998–2009" Applied Linguistics, 12

[11] The Corpus of Historical American English (COHA): 400 million words, 1810-2009. Brigham Young University, 2002.

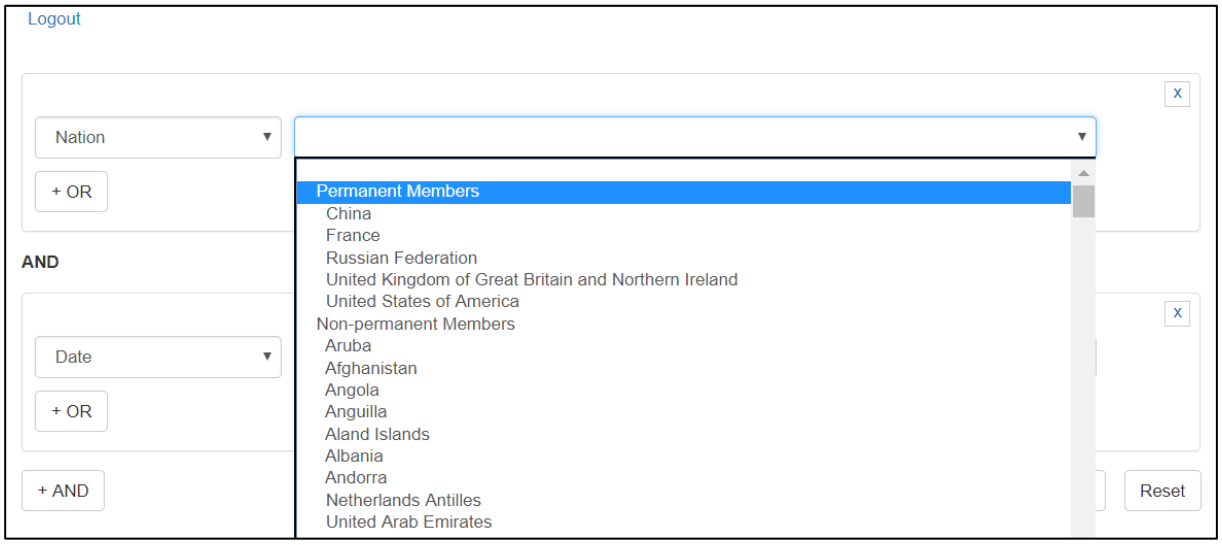


Figure 2: UNSCdeb8 Web-tool

The UNSCdeb8 web-based query tool tool (https://unscdeb8.swp-berlin.org & unscdeb8.ethz.ch) can structure the data selected for export (i.e. its sub-corpora) in various ways. For software-based corpus linguistic analysis concerned with the comparative analysis of subcorpora, we encourage users to export individual plain text (.txt) files, i.e. files separated by debate (date and document number are automatically included in the filename). This export-mode allows one to make the most of the functionality (such as concordance-plotting) available in corpus-linguistic analysis software tools. Corpora consisting of many individual .txt files also enable the user to jump back and forth between software tools and full-text source files, without losing their orientation in an ocean of text. This can be particularly important when corpus-driven research is intend to feed into qualitative social science analysis; i.e. as a useful stepping stone in a multi-method route, narrowing down relevant source material. In this sense, UNSCdeb8 can serve as helpful tool within a wide range of theoretical and methodological approaches.

How UNSCdeb8 was built

The UNSC usually publishes verbatim records of its public meetings a few days after they take place. These transcripts can be downloaded from the UN website. These records only include statements made during the official meetings. In other words, any preparatory work or meetings prior to the formal meeting are considered informal and are not recorded in the transcript. Furthermore, no verbatim transcriptions are available for closed meetings that are not open to the public. While closed meetings are required to provide a communiqué to the UNSC meetings archives, these documents usually contain very little content other than stating that a meeting took place. Documents relating to closed meetings are thus excluded from UNSCdeb8. The official UNSC meeting records are the only type of document included in UNSCdeb8. Following the classificatory system of UN documentation, UNSCdeb8 is a database of public UNSC meeting records that can be identified by the prefix "S/PV". According to an administrative instruction by the UN Secretariat,

"S/PV" records are “left in the public domain”[12] and can, therefore, be utilized for non-commercial, scholarly purposes. These documents are generally available in PDF (and sometimes MS-word) format. For these documents to be conveniently analysable with corpus-linguistic analysis tools, UNSCdeb8 converts the textual data in these PDF into outputs in .txt-format. These .txt-files can then be processed by the most common Corpus Linguistic Analysis software tools.

This is valuable because while various tools are available to convert PDF file formats to .txt these differ in quality.[13] Generally speaking, automated file conversion is error prone whenever the original formatting of the source files includes tabulation or columns. The official PDFs of UNSC meeting records are formatted using two columns. Further complicating any attempt to automatically convert these files, the formatting and layout of the PDF meeting records are often only superficially similar to one another; the meeting records employ different underlying tables and styles during different time-periods. Taking this into account, the first step we took in building our database was to convert most of these PDFs into a HTML format. This conversion to HTML is advantageous for the consistent extraction of text and its coding with metadata labels, because HTML elements (such as <b> for bold-print) can be used as start- and stop-markers for building an automated text parsing code. Hence, HTML notation could be used to signal that a new speaker is starting a statement. In some difficult cases, especially for older documents, clean conversion to html was not possible and documents could only be converted to .txt (thus lacking HTML elements). In such cases, the HTML file was only used as a resource to extract speaker-names and a separate code was used to extract speeches from .txt files. As algorithms were used to gather textual data, total accuracy cannot be guaranteed. Due to the varied formatting of the original UN-transcripts, some names with special characters might be misrepresented. Mr. Kubiš, for instance, may in some instances be displayed as “Mr. Kubi” or “Mr. Kubis”. It should also be taken into account that many statements are translations (from Russian, Chinese, etc.), while others are not.

The process of automatically assigning metadata labels to the extracted text was more difficult for some labels than other. The meta-data on the *agenda* of a meeting was relatively difficult to extract from the PDFs. It was possible to identify most of the relevant agenda fields in the database using a code that parsed the text following the statement “the agenda was adopted.” The data thus extracted was then triangulated, and in some case supplemented, by a (web-scraping) search on the agenda items as listed on the UNSC website. This technique of triangulating the data gathered by PDF-parsing with additional web-scraping was also used to ensure the accuracy of *meeting dates*, because the formatting of the date in the official UNSC PDFs is not uniform.

[12] ST/AI/189/Add.9/Rev.2, 17 September 1987 http://www.un.org/Docs/journal/asp/ws.asp?m=ST/AI/189/ Add.9/Rev.2 . Reaffirmed in ST/IC/2015/1, Index to administrative issuances, UN Secretariat (2015-01).

[13] Rayson, P. (2015) “Computational Toolsand Methods for Corpuscompilation and Analysis.” in D. Biber and R. Reppen (eds.) The Cambridge Handbook of English Corpus Linguistics, Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press, p. 32–49.

```
Iterate through HTML body:
        IF number of ‚bold-tags‘ > 0:
                save (line number, bold-text)
        // sometimes pages have a bold header/footer which can be identified by the page
number (footer) or SPV code (header). Those need to be excluded
        IF number in bold-text delete (line number, bold-text)
        SORT (line number, bold-text) BY line number ASCENDING
        // previously all speakers need to be identified and stored in the SET SPEAKERS
        LOOP THROUGH EACH (line number, bold-text) PAIR:
                IF bold-text IN SET SPEAKER:
                        FIND next PAIR with bold-text IN SET SPEAKER
                ATTACH TEXT FROM line-number to next PAIR line-number 1 AS SPEECH TO
SPEAKER bold-text
```

Figure 3: Extracting Speeches (Pseudo Code)

```python
import requests
import pandas as pd

from bs4 import BeautifulSoup

url = "http://www.un.org/en/sc/meetings/records/{}.shtml"

def parseTable(url):
    fields = ['SPV', 'Date', 'Press_Release', 'Agenda', 'Vote']
    response = requests.get(url)
    soup = BeautifulSoup(response.text, "html.parser")
    trs = soup.find_all('tr')
    entries = []
    for elements in trs:
        tds = elements.find_all('td')
        if len(tds) == 5:
            data = {k:v.text for k,v in zip(fields, tds)}
            entries.append(data)
    entries = pd.DataFrame(entries)
    return entries

for year in range(2000, 2018):
    print(url.format(year))
    data = parseTable(url.format(year))
    data.to_csv("agenda_{}.csv".format(year), index=False)

```

Figure 4: Extracting Agenda Lines

(Figure 4: assumes that speaker-names are known and already extracted.)

A HTML parser (beautiful soup (https://www.crummy.com/software/BeautifulSoup/) was used to identify the respective HTML tags. The beginning of a speech was identified by three criteria: 1) It is always a new paragraph; 2) The name of the speaker is always in bold text; 3) The bold text ends with a colon (:).

In some rare cases, speeches had to be manually selected in order to correct parsing errors. In the end, all data was merged into a PostgreSQL database and is currently hosted by ETH Zürich Scientific IT Services.

3. UNSCdeb8 in our research project: concept, method & compilation

In the "Which region?" project, we consider that "language is the medium through which meaning is generated"[14], and is thus constitutive of social (constructed) reality. More specifically, this means that we seek to analyse how and when international actors constitute spatial and scalar objects, boundaries and relations during international security crisis through their discourse. We are particularly focused on how notions of regions/regional are argued about in the UNSC. Amongst many questions, we ask how the P5 represent and argue about where the boundaries of a region begin and end, who has the legitimacy to speak for a region, and how the relationship between regional and international actors and authority should be ordered? Our methodological approach to investigating these questions centers on combining the use of corpus linguistic analysis with inductive qualitative discourse analysis, drawing on similar work from Political Geography[15].

What do we mean by "discourse"? Discursive articulation and the performativity of language

We conceptualise discourse as "a structure in which meaning is constantly negotiated and constructed",[16] and which is produced through and by actors' articulation of their perceptions and readings of themselves, their context and others around them. From this standpoint, discourse is not separate from practice. It is rather constitutive of practice, as every social phenomena and object obtains their meaning(s) through discourse.[17] Further, we look at discourse as a process of producing meaning-systems that make the complexity of the world more comprehensible. Discourse is neither purely ideational, nor reducible to language; in discourse, materiality matters in a different way. For example, new railroads routes are currently being built across Europe and Asia, in association with China's Belt & Road Initiative: How this is interpreted depends on the specific structure of the discourse through which this interpretation is framed. Thereby, these railroad tracks are viewed, on the one hand, as a purely logistical project undertaken by multinationals and, on the other hand, as part of a massive geo-political Chinese strategy to gain more control over logistical networks and standards in Eurasia. Hence, objects may exist as material brute facts, but they cannot "constitute themselves as objects outside any discursive condition of emergence"[18]. A discourse theoretical perspective is thus more interested in how objects are represented and used. In our case, this leads to a focus on how and when international

[14] Hansen, L. (2012) "Discourse analysis, post-structuralism, and foreign policy", in S. Smith, A. Hadfield and T. Dunne (eds.), Foreign Policy: Theories, Actors, Cases, Oxford: Oxford University Press.
[15] Mattissek, A. (2010) "Analyzing City Images: Potential of the "French School of Discourse Analysis"", Erdkunde 64, 4; Glasze, G. (2007) "The discursive constitution of a world-spanning region and the role of empty signifiers", Geopolitics 12, 4.
[16] Laclau, E. (1988) "Metaphor and social antagonism", in C. Nelson & L. Grossberg (dds) Marxism and the interpretation and culture, Basingstoke: Macmillan Education.
[17] Müller, M. (2008) "Reconsidering the concept of discourse for the field of critical geopolitics: Towards discourse as language and practice", Political Geography 27, 3.
[18] Laclau, E. and Mouffe, C. (2001) Hegemony and Socialist Strategy. Towards a Radical Democractic Politics, London: Verso, 2nd edition, p. 108.

actors articulate and contest spatial terms in UNSC debates, and how these notions are embedded in these actors' foreign policy.

In this context, we consider all utterances in the UNSC as playing a part in the constitution – and dissolution – of meaning systems. More specifically, discourse is a continuous practice of articulation through which differing elements (ideas, events, things, etc.) are put in relation to one other (differentiated from each other). In other words, discourse functions to organize language-elements in a particular way, reducing their surplus of meaning and temporarily stopping their fluctuations in meaning. But, discourse never reaches full closure; it is never complete because "discourse is always constituted in relation to what it excludes, that is, in relation to the field of discursivity".[19] All of the excluded possibilities of meaning are still present in the "field of discursivity", which is why discourse is always in motion and subject to modification. The act of discursive articulation, thus, highlights the impossibility of an ultimate fixity of meaning, while, at the same time, it aims to stabilize a temporary meaning-system (a specific discourse).

In order to investigate international actors' discursive constitution of region/regional and other spatial terms, we focus on the discursive articulations of foreign policy elites and diplomatic discourses that are available in the public domain, rather than the individual, private or hidden discourses of state elites. This is because we read discourses as social and collective processes, whereby "discourses are not confined to an 'inner' realm of mental phenomena, but are those publicly available and incomplete frameworks of meaning which enable social life to be conducted".[20] And, hence, "foreign policies have to be connected, through discourse, to justification for why these policies are necessary, plausible, and possible".[21] A prominent site of such justifications of foreign policies, narratives and representations is the UNSC. Hence, the UNSC is an important venue for stabilizing, challenging and modifying discursive systems of meaning in international security.

Our method of analysing "discourse": the advantages of corpus-driven (lexicometric) research

Developed as a theoretical and methodological approach within Linguistics, corpus linguistics is concerned with the analysis of large bodies of textual data, to the end of identifying macro-linguistic/semantic structures, changes and relationships within the language used in this text. In other words, it aims to account for how discourse constitutes meaning. Now predominantly based on computational analysis, it analyses the associations, disassociations and positioning of lexical items that compose the discourse's semantic meaning structures.

A major advantage of this method is that it serves to direct the researcher towards significant linguistic relationships within, otherwise unmanageably, large bodies of discourse, which the researcher can then seek "to interpret and explain" by a closer reading.[22] It also provides a tool for countering several of the established criticisms of discourse analysis. For example, the analysis of large corpora enables claims about the

[19] Jorgensen, M. and Philipps, L. J. (2002). Discourse Analysis as Theory and Method, London: Sage Publications; Laclau, E. and Mouffe, C. (2001) Hegemony and Socialist Strategy. Towards a Radical Democratic Politics, London: Verso, 2nd edition, p. 111.

[20] Howarth, D. (2000) Discourse, Buckingham: Open University Press, p.104.

[21] Hansen, L. (2012) "Discourse analysis, post-structuralism, and foreign policy", in S. Smith, A. Hadfield and T. Dunne (eds.), Foreign Policy: Theories, Actors, Cases, Oxford: Oxford University Press, p.101.

[22] Baker, P. (2006) Using corpora in discourse analysis, London/New York: Continuum, p. 11.

"representativeness" of the findings of discourse analysis to be made with greater authority. A researcher can illustrate how the particular argumentations, "empty signifiers"[23] or "rhetorical commonplaces"[24] they identify as significant are evident across the discourse under study. Furthermore, its quantitative statistical grounding provides an element of verifiability that is not normally associated with discourse analysis[25]. This helps a researcher avoid some typical pitfalls of text selection bias, whereby "the researcher only takes the texts into consideration, which fit his prejudices"[26].

Our project aims to add to the budding adoption of corpus linguistic approaches in IR.[27] It seeks to examine the application of this methodology in other fields of social science, such as political geography, and apply the lessons learned to its usage for addressing IR research questions[28]. In our project, we combine corpus-driven research with qualitative/interpretative discourse analysis. These approaches can be seen as "two techniques that mutually enrich and complement each other"[29]. To this end, we follow a two-stage analytical approach. We firstly conduct lexicometric textual analysis, including frequency analyses to identify and examine specific terms (such as "region", "community", or others), concordance analysis to highlight how specific terms are embedded in and relate to the textual context, and "keyness" analyses to assess the significance of specific terms within a textual corpus (i.e. a comparison between the primary "study corpus" and a second "reference-corpus"). Secondly, the results of the lexicometric analysis are further interpretatively analyzed, using a whole range of qualitative and "close-reading" methodological approaches, such as argumentation analysis, Foucauldian discourse analysis[30], or an application of Laclau and Mouffe's discourse theory.[31]

[23] Laclau, E. and C.Mouffe (2001), Hegemony and Socialist Strategy: Towards a Radical Democratic Politics, Verso.

[24] Jackson, P.T. (2006) Civilizing the Enemy: German Reconstruction and the Invention of the West, University of Michigan Press.

[25] Mattissek, A. (2010) "Analyzing City Images: Potential of the "French School of Discourse Analysis"", Erdkunde 64, 4.

[26] Glasze, G. (2007) "The discursive constitution of a world-spanning region and the role of empty signifiers", Geopolitics 12, 4.

[27] If this method is used in IR, these are mostly post-structural scholars with a interest in discourse theory and discourse analysis. See for instance Nabers D. (2015) A Poststructuralist Discourse Theory of Global Politics, Basingstoke: Palgrave.

[28] Dzudzek I. et al. (2009) "Verfahren der lexikometrichen Analyse von Textkorpora" in G. Glasze and A. Mattissek, (eds.) (2009), Handbuch Diskurs und Raum. Theorien und Methoden für die Humangeographie sowie die sozial- und kulturwissenschafliche Raumforschung; Baker, P. (2006) Using corpora in discourse analysis, London/New York: Continuum; Mattissek, A. (2010) "Analyzing City Images: Potential of the "French School of Discourse Analysis"", Erdkunde 64, 4; Glasze, G. (2007) "The discursive constitution of a world-spanning region and the role of empty signifiers", Geopolitics 12, 4.

[29] Uriel Abulof (2015) Normative concepts analysis: unpacking the language of legitimation, International Journal of Social Research Methodology, 18:1, p. 73.

[30] Dudzek, I. (2012) "Widersprüche kultureller Vielfalt. Eine Genealogie der Dezentirerung kultur-räumlicher Repräsentationen in der UNESCO", in I. Dudzek, P. Reuber, A. Strüver (eds.) Die Politik räumlicher Repräsentationen. Beispiele aus der empirischen Forschung, LIT Verlag: Berlin, p. 109-151.

[31] Glasze, G. (2007) "The discursive constitution of a world-spanning region and the role of empty signifiers", Geopolitics 12, 4.

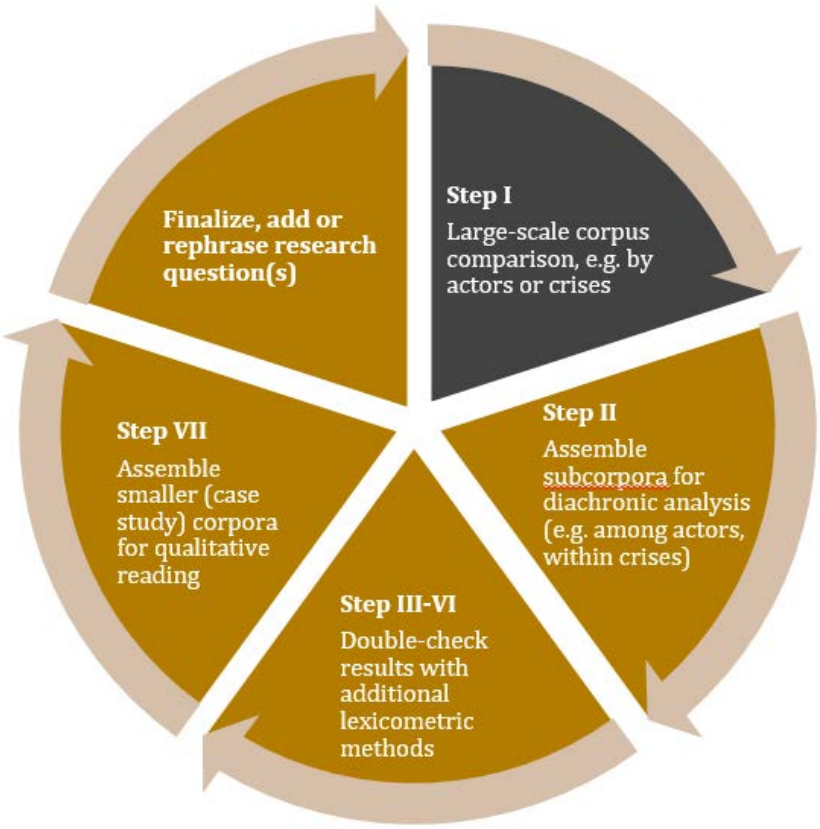


Figure 5: Suggested Workflow Using UNSCdeb8

In announcing the development of UNSCdeb8 as a new database resource for analysis of UNSC debates, we are also advocating a multi-stage research strategy to make full use of its functionality. As figure 1 visualises, we point to the way that UNSCdeb8 can be used in association with corpus linguistic analysis tools as one part of this research strategy. In this way, UNSCdeb8-informed corpus linguistic analysis may be used as a first step to guide subsequent close-reading discourse analysis, as an intermediate step that digs deeper into a line of inquiry or opens up new avenues for investigation, and/or as a final step in verifying the representative of a close-reading analysis. This multi-function utility is illustrated in the next, by way of a case study analyzing China's UNSC discourse.

4. Analysing discourse with UNSCdeb8

The statements given by the representatives of the five permanent members of the UN Security Council are among the subcorpora of UNSCdeb8 that are most likely to be of interest to social science researchers. As representatives of the "P5" are present at almost all debates, the aggregate statements by China, France, Russia, UK and USA are the largest textual subcorpora (defined by speaker nation) that can be studied in themselves and, in comparison to non-permanent UNSC members. These subcorpora of all P5 or individual P5 member state statements, thus, offer opportunities for research that breaks the discourse down into smaller units for chronological or topical comparison. For example, an analysis of how the language and arguments used by the P5 has changed over time, or an examination of whether and how a particular P5 member state uses distinct terminology

and claims in relation to different security crises. To illustrate how these national subcorpora can be used to address important questions about the UNSC, this paper sets out an analysis of how China's discourse in the UNSC, one, differs from other P5 members and, two, has evolved over recent years as Beijing has sought to play a new role in international affairs. It also represents an exemplary workflow making use of UNSCdeb8 as part of a multi-step, -approach and –tool research strategy.

Exemplary Workflow: How to extract meaning out of China's discourse in the UNSC

Step I: Comparison between Chinese discourse and all UNSC discourse (analysing keyness values)

In line with our methodological approach, it is advisable to start developing research questions with large corpora and then slowly work towards more detailed inquiries. As China's oft-cited "rise" to global power status has been ongoing over the past decade, an analysis of possible changes in China's use of language at the UNSC represents a fascinating case study to illustrate how we envision that UNSCdeb8 can be utilized to investigate interesting research questions. As a first step, the webtool was used to collate all China's statements from UNSCdeb8, without setting additional filters for time or topic.

Figure 6: Webtool

As a second step, a reference corpus was downloaded via the webtool. In this case, the complete UNSCdeb8 corpus is used as a reference corpus with which to compare the Chinese subcorpora, but it would also be plausible to download the aggregate statements of the other four permanent UNSC members for the same purpose. With the help of corpus-linguistic software, it is then possible to calculate keyness values (i.e. "log-likelihood" values[32]), in order to identify words that are unusually frequent or rare in statements given by Chinese diplomats, as compared to those used by all other speakers in the UNSC. In other words, keyness refers to words that are mentioned more (or less) often in comparison to a reference corpus. A high keyness score is seen as indicating a word or phrase is of particular semantic significance within the subcorpora under study. Table 1 outlines the keyness results of a comparison of Chinese UNSC statement vis-à-vis the totality of all other statements by the UNSC's other four permanent members (here limited to the top 15 most significant differences).

[32] Rayson, P. and Garside, R. (2000) "Comparing corpora using frequency profiling" in Proceedings of the Workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp. 1 - 6.

Table 1: China's Keywords (Compared to Language-Use of Other Permanent UNSC Member States, 2010-2017)

	KEYNESS	KEYWORD
1	13.770.851	china
2	4.165.615	chinese
3	2.503.065	and
4	2.443.617	should
5	1.868.049	countries
6	1.728.412	development
7	1.495.219	community
8	1.483.576	peace
9	1.153.999	parties
10	1.125.181	international

Most of these results are not surprising. China naturally refers to itself more frequently not least because the parsed UNSC statements often include the diplomat's name, nationality and the language spoken. The high keyness of the conjunction "and" is likely a result of translations, partly because Chinese conjunctions are often implicit, and translators might choose the most neutral English variant. "Countries" is also a word that is overrepresented because "country" and "state" are normally both encompassed in one single Chinese term (国家), but also because China is often less inclined to directly mention (and thereby potentially antagonize) the states involved in international disputes. (This is also an explanation of the centrality of "parties" (#9)). China's focus on the concept of development has also been widely studied.[33] Taking the most significant keywords among Chinese UNSC discourse as a starting point,[34] we noted that Chinese statements used the word "community" (#7) more frequently than the average statement at the UNSC. Therefore, we chose to further investigate why and how the word "community" is so frequently employed by Chinese representatives, as an exemplar of a workflow that uses UNSCdeb8 to investigate UNSC discourse.

[33] Fewsmith, J., (2004) "Promoting the scientific development concept." China Leadership Monitor 11.30, p.1-10.

[34] Their statistical significance is far above the value of 15.13 which the linguistic literature regularly employs to signal a p-value above the 99.99th percentile. However, as the exact statistical values have been criticized as "arbitrary", the keyness-numbers are here only listed to provide a general orientation relative to the values of other terms. See: Gabrielatos, Costas and Marchi, Anna (2012) Keyness: Appropriate metrics and practical issues. CADS International Conference 2012. Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing?, 13-14 September, University of Bologna, Italy.

To ascertain whether a keyword (in this case "community") is a stable or episodic phenomenon, we recommend comparing corpora that are divided into smaller chronological units. Hence, we used the webtool to download four corpora of Chinese statements at the UNSC according to different time-periods, each covering a time span of 24 months.

Figure 7: UNSCdeb8 Webtool – Statements Filtered for "USA" and "2010-2011"

Corresponding to the above outlined approach for identifying semantic keyness, corpora of all other statements in the UNSC for the same four time-periods were also downloaded. In this way, China's statements could be compared against a reference corpora for each of the four 24-month periods. The keyness values for words distinct to China's lexicon of speech were then analysed using Antconc. We found that the term "community" is a consistent keyword in Chinese UNSC discourse throughout the 8-year period, but – relative to the overall UNSC discourse – became increasingly distinct in the most recent years. Its centrality fully manifests itself post-2013.

Table 2: Top Keywords Distinct to China's UNSC Statements (Relative to Statements by All UNSC Member States)

2010-2011	2012-2013	2014-2015	2016-2017
china (1)	china (1)	china (1)	china (1)
chinese (2)	chinese (2)	chinese (2)	chinese (2)
...	...	...	...
community (20)	community (29)	community (12)	community (4)

Step II: Identification of "Community" as a significant keyword in China's UNSC discourse, by way of diachronic analysis (analysing keyness values)

The increasing keyness of "community" among China's statements could be a consequence of other countries referring to the idea of "community" more infrequently in recent years, rather than an increase in Chinese usage. To answer this question, we firstly compared China's own statements diachronically to see whether China's usage of the term is increasing. And, then, secondly, we compared each of the other P5 members' usage of the term to the reference corpus, in order to assess whether these member states had increased or decreased their frequency of usage. It was not necessary to use the webtool for the first inquiry, because the four subcorpora of Chinese UNSC discourse (for the years 2010-2011, 2012-2013, 2014-2015, 2016-2017) had already been downloaded and could thus be

used for diachronic comparison. By merging these four files into two larger ones, we compared China's statements from 2010 to 2013 (as the reference corpus) with those between 2014 and 2017 (as study corpus), in order to quickly check whether the term "community" appears as a keyword in latter period as compared to the former (see table3).

Table 3: China 2014-2017: Positive Keywords (vs. 2010-2013, personal names excluded)

	KEYNESS	KEYWORD
1	172.916	terrorist[35]
2	162.396	peninsula
3	154.556	ukraine
4	110.504	china[36]
5	104.455	and
6	90.231	win[37]
7	76.091	should[38]
8	73.611	terrorism
9	69.686	**community**
10	67.932	korean
(> 15.13 SIGNALS P-VALUE ABOVE THE 99.99TH PERCENTILE)[39]		

This diachronic evaluation of China's statements shows that "community" is, indeed, a term that appears significantly more frequently in recent years. Overall, "community" is among the ten most statistically significant keywords, after notions such as "peninsula" or "Ukraine" that are tied to the development of particular crises that increased in prominence in the latter time-period and would therefore be expected to top any list of diachronic comparison. Despite the strong evidence of China's increasing usage of the term, it is still worthwhile to double-check the extent to which a decline in usage of "community" by *other* UNSC actors may have contributed to its high keyness score in China's contributions to Security Council debates.

[35] "Terrorist" and "terrorism" have high diachronic keyness values for all five permanent UNSC members.
[36] The high keyness of "China" can be partially attributed to a conscious or unconscious decision by Chinese diplomats to decrease the use of personal pronouns such as "we" and speak as "China" instead. The latter terms are thus marked by significant negative keyness shifts for this diachronic comparison.
[37] High keyness almost exclusively due to increasingly frequent references to the idea of "win-win cooperation".
[38] High keyness partially mirrored by a high negative keyness of the verb "shall". Generally, however, Chinese statements have shifted to stronger verbs like "should" and "must", while softer verbs like "wish" decrease in keyness.
[39] Exact statistical values have been criticized as "arbitrary", the keyness-numbers are here only listed to provide a general orientation relative to the values of other terms. See: Gabrielatos, C.and Marchi, A. (2012) "Keyness: Appropriate metrics and practical issues'" CADS International Conference 2012. Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing?, 13-14 September, University of Bologna, Italy.

Step III: Verification of diachronic usage of "community" in statements by other UNSC P5 (keyness analysis)

To compare other P5 members' usage of "community" to that of China's, we first focused on US statements at the UNSC. To trace any changes in their usage of "community" at the UNSC, the webtool was employed to assemble two chronologically corresponding corpora of statements by US representatives: 2014-2017 (as study corpus) and 2010-2013 (reference corpus). This analysis revealed a significant *decrease* in US references to "community". Using the same workflow, analysis of the other three permanent UNSC members found similar results.

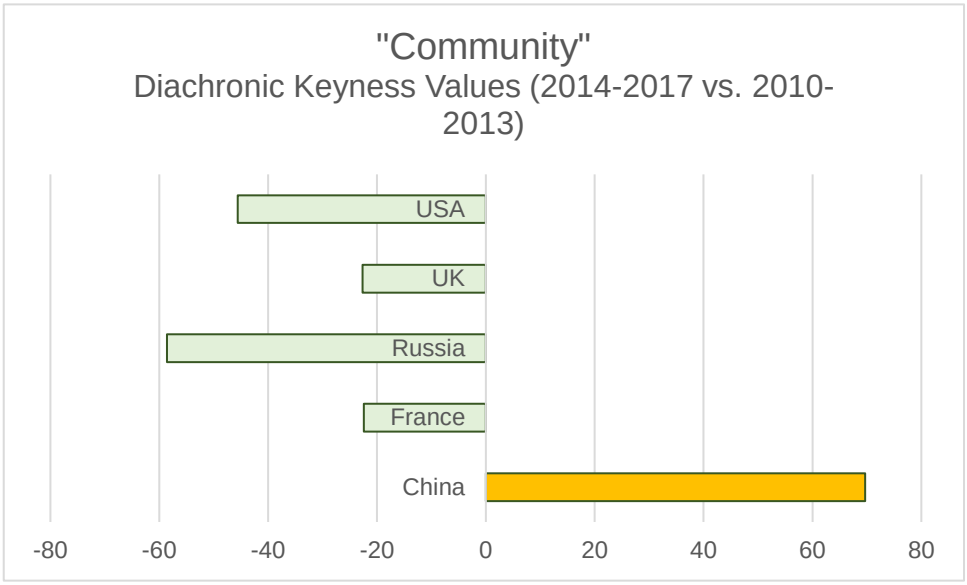


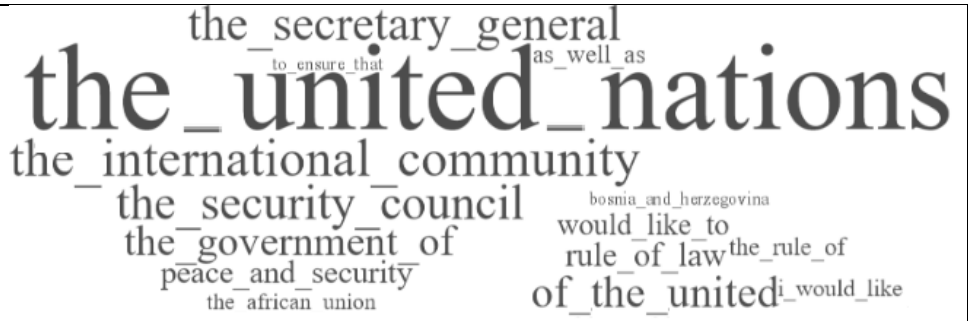
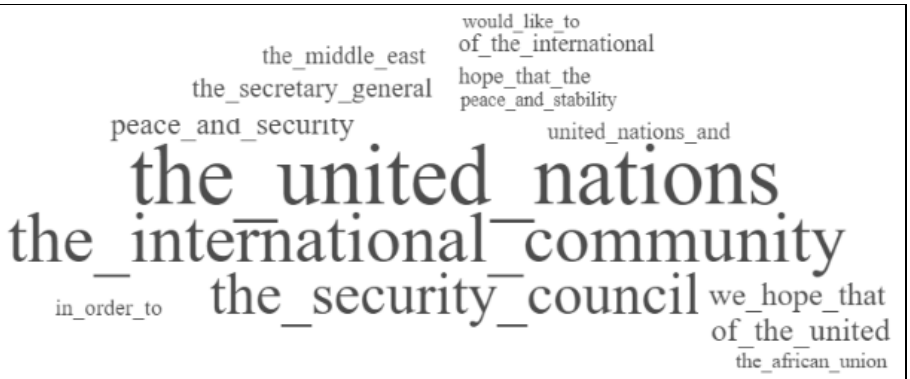
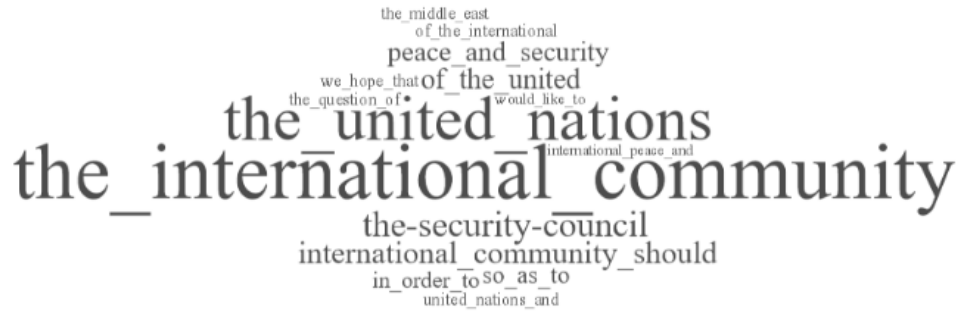
Figure 8: The use of "community" by China and other permanent UNSC members

Therefore, China was found to be the only permanent member of the Security Council that has increasingly emphasized the idea of "community" over the past eight years. Importantly, even during the period 2010-2013 in which China's usage was less frequent than the latter period, "community" was much more common in Chinese diplomats' statements to the UNSC vis-à-vis the statements by the representative of the other four members of the P5. Furthermore, while China's P5 counterparts consciously or unconsciously chose to use the term less frequently than previously between 2014 and 2017, China has moved in the opposite direction.

Step IV: Verifying the significance of "community" in China's UNSC discourse vis-a-vis the other P5 (analysing collocations through N-Grams)

Continuing to work with the numerous subcorpora that have already been downloaded though the UNSCdeb8 webtool, we can further check the accuracy of the above findings. In corpus linguistics, one way to quantitatively assess the most significant meaning structures within a piece of text is by employing N-grams that list the most frequent combinations of words (i.e. "collocations") within a corpus of text. After defining the target "size" of the N-gram (here: three words), we compared the standard phrasal formulations (3-word phrases) of the P5's discourses at the UNSC, as well as contrasting such formulations in each actor's statements over time.

Table 4: Change over time - Most Frequent Three-Word Combinations "N-Grams" of USA & China[40]

USA 2010 - 2013	
USA 2014 - 2017	
China 2010 - 2013	
China 2014 - 2017	

(See Annex 1 for corresponding word clouds of all P5.)

[40] Some terms in the N-Grams were excluded. To ensure comparability between P5 with single-word states like China or France, N-Grams such as "United States of" or "The United Kingdom" were excluded. The same applies for N-grams containing the names of Ambassadors (e.g. „Mr Delattre France").

Numerous observations emerge in the N-Gram analysis. Generally, there is significant overlap in UNSC-parlance. References to the UN are frequent in N-grams for all P5. However, in the later period, China's focus on the "international community" surpassed references to the institutional (i.e. UN) context in which all debates take place. Concerning the way its statements have developed over time, Beijing is evidently very different from the other P5 (see annex 1).

Step V: Diachronic concordance analysis of China's and the other P5's UNSC statements
Along similar lines, China's UNSC discourse also saw a steep decline in references to the UN Secretary General, a change not mirrored in any of the other P5's diachronic comparisons. Notably, half of China's references to the "international community" are followed by one specific verb: "should". Most lexicometric investigations reach their limits when it comes to furthering unpack the linguistic shifts they identify. In this case, how to better understand what China means when it refers to "community" in the context of the UNSC.[41] For instance, an analysis of the words most frequently co-located with the term "community" does not reveal any significant diachronic changes (see table). It is, thus, necessary to move towards a more qualitative engagement with the linguistic and political context, in order to answer further questions about why, how, and when the above observed shifts occur.

Table 5: Top-Collocations with "Community" in China's UNSC Statements

	2010-2013	2014-2017
1	international	international
2	should	should
3	support	support
4	efforts	efforts
5	continue	china
6	we	work
7	assistance	continue
8	china	countries
9	provide	we
10	must	assistance
11	work	must
12	respect	as
13	peace	**region**
14	united	provide
15	countries	security
16	**call**	development
17	nations	ready
18	attention	african
19	help	peace
20	common	common
(COLLOCATION RANGE: -7 TO +7 WORDS)		

[41] This is also why we decided to combine corpus-driven research with qualitative/interpretative discourse analysis in our project.

Step VI: Extract and analyse examples of collocations, further specifying the research focus

Among the only terms that drops out of the top 100 collocations in the latter period as compared with the former is "call". In other words, while there are still several recent UNSC debates during which China states that "we call on the international community,"[42] Beijing's diplomats do so less often than before. "Region", in contrast, has become much more commonly collocated with "community". By using a concordance tool (generally part of the functions of linguistic software), it is possible to take a cursory in-context look at the combined-usage of region and community in recent years. This indicates how, in numerous instances, China increasingly juxtaposes or equates the two notions (seeFigure 9).

Figure 9: Sample Collocations of Region and International Community

Undoubtedly, the frequent references to "international community" are designed to signal that China is itself a legitimate part of the said community (whatever its precise definition). However, more recently China has sought to more directly contribute to the definition of this community, not least by emphasizing the role of regions within the international community and, possibly, by focusing less on the institutions of the UN in its own statements.

[42] E.g. S.PV.7789, S.PV.7840 , S.PV.8115.

Step VII: Case Study on the discursive relation of "region" + "community" within China's UNSC statements on Afghanistan

As well as its utility for constructing subcorpora and surveying similarity/divergence in discourse, UNSCdeb8 is also a versatile launchpad for delving into structured qualitative/interpretative analysis. As a case study to illustrate this, we used the webtool to build a small corpus of China's UNSC statements on Afghanistan.

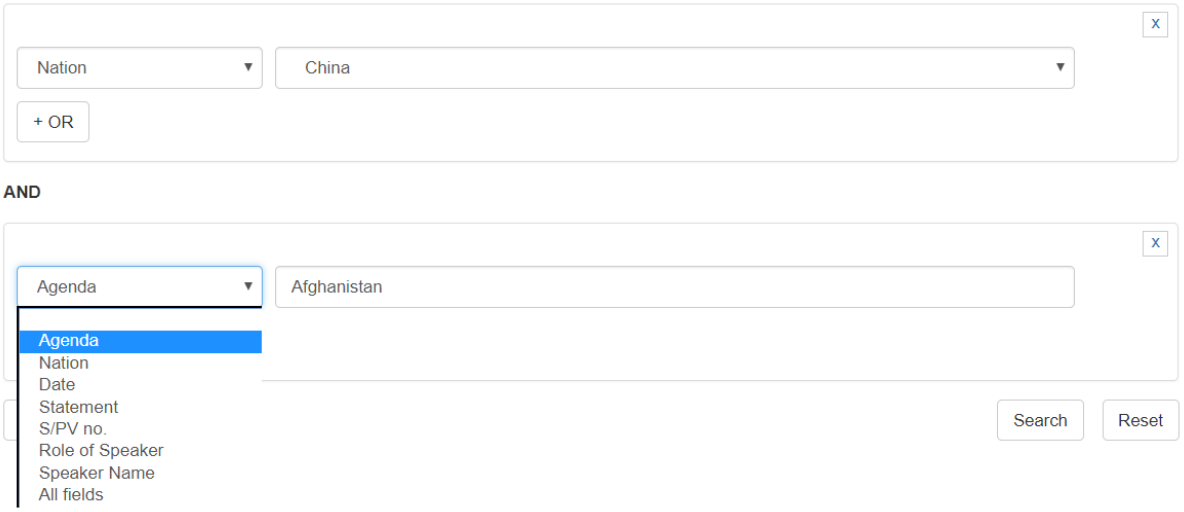
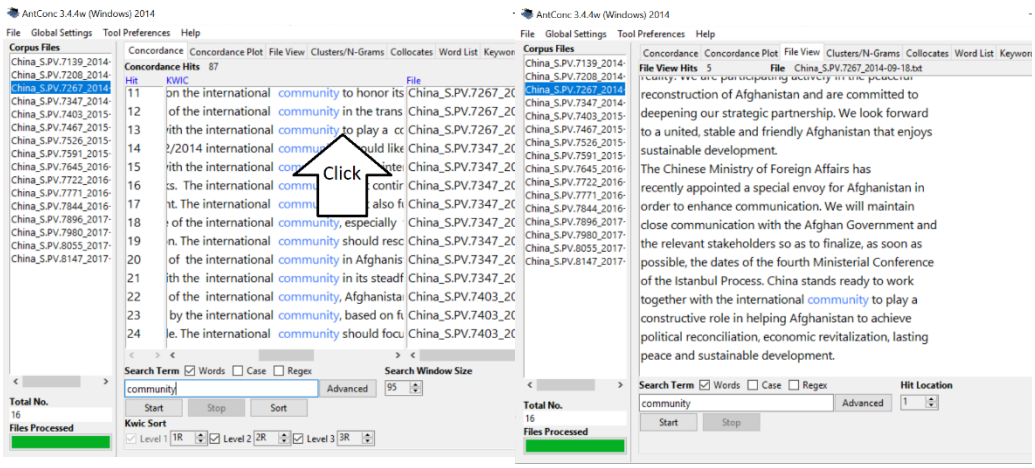


Figure 10: Agenda Selection in Webtool

Afghanistan was chosen because it is a "crisis" with which the UNSC has been continuously engaged across the timeframe covered by UNSCdeb8. Debates on Afghanistan have been scheduled to be held at least four times a year. It is also geographically close to China, so if "regional" thinking increasingly contributes to China's interpretation of the "international community", we would expect to find it reflected in this case study. Using Antconc to conduct a concordance analysis of the subcorpora of China's statements on Afghanistan allowed us to see all the instances in which the term "community" was used in their immediate semantic context.

Table 6: Jumping to File View with Antconc



Qualitative/Interpretative approaches can then be deployed to pin down meanings in more depth. Importantly, it is possible to jump from the concordance-window to the original file for close-reading, by directly clicking on any section of interest. After defining specialized subcorpora, textual content can then be quickly eyeballed and possibly even manually coded, and the analyst can jump to full text mode when necessary.[43] Accordingly, as shown in table 7, we employed a close reading approach and looked at the wider context of instances of the term "community" in China's statements during UNSC debates on Afghanistan. A tabular coding scheme was then used to group frequent juxtapositions and connotations.

Table 7: In China's UNSC Statements "Community" is Juxtaposed with…

	UN / UNSC	Political Principles	Region/Regional	China's Initiatives	China's Leadership	New Type of Community
China_S.PV.6287	1					
China_S.PV.6255						
China_S.PV.6351	1	1				
China_S.PV.6394	1	1				
China_S.PV.6464	1	1				
China_S.PV.6574	1					
China_S.PV.6625		1				
China_S.PV.6690	1					
China_S.PV.6735	1	1				
China_S.PV.6896	1	1	1			
China_S.PV.6983	0,5	1				
China_S.PV.7035	1		0,5	0,5		
China_S.PV.7139	0,5		1			
China_S.PV.7208	1	1				
China_S.PV.7267	1		2	0,5		
China_S.PV.7347	1		1		1	
China_S.PV.7403	1	1		0,5		
China_S.PV.7467	1	1	1	0,5		
China_S.PV.7526	0,5	1	0,5	1		
China_S.PV.7591		1	1	1	1	
China_S.PV.7645	0,5		1	0,5		
China_S.PV.7722		0,5	0,5	0,5		
China_S.PV.7771		1	1	1	0,5	
China_S.PV.7844		0,5	0,5	1	1	
China_S.PV.7896	0,5	1	1	0,5	1	
China_S.PV.7980	0,5	1	1	1	1	1
China_S.PV.8055	0,5	1	1	1	1	1
China_S.PV.8147	0,5	1	1	1	1	1

This analysis revealed that over the course of the eight years of China's statements on Afghanistan, Chinese discourse has moved away from explicitly associating the Security Council or the United Nations with the idea of the international community. At the beginning of the 2000s, these two nouns were still directly associated:

[43] Naturally, the text can also be transferred to other qualitative analysis tools like MAXQDA

> *"In its next phase, the work of the Security Council and the international community must focus on implementing the communiqué adopted at the international Kabul Conference on Afghanistan."* (China, 2010)[44]

> *"The promotion of peace and reconstruction in Afghanistan requires the joint efforts of that country and the international community. China supports the United Nations lead role in coordinating international efforts to address that matter."* (China, 2012)[45]

Incrementally, however, China's statement have drawn less associations between the role of the UN and UNSC bodies and the concept of the international community. Although China has continued to refer to the importance of UNAMA (UN Assistance Mission Afghanistan, here coded as "0,5"), it is less directly connected to claims about how the security challenges in Afghanistan are being resolved by the international community. Instead, in recent years, we find more statements that juxtapose the notion of regional stakeholders with the concept of the international community.

> *"China welcomes the ongoing assistance being provided by countries in the region, and we hope that the international community will support and coordinate with the Afghan Government and people's efforts with a view to creating an external environment conducive to the country's national reconciliation process."* (China, 2015)[46]

> *"Afghanistan cannot realize development and prosperity without the support and assistance of the international community, especially the countries in the region."* (China, 2014)[47]

From 2014 onwards, China increasingly presents the cornerstones of the Belt and Road Initiative as the appropriate framework of action for the international community.

> *"The international community should encourage and support Afghanistan in fully leveraging its advantages in resources and geographical location and push for progress in building infrastructure and developing trade and investment, among other areas. "* (China, 2016)[48]

Similarly, it assertively presents its own initiatives as key elements of the international community that deserves wider support.

> *"The international community should step up its cooperation in combating terrorism and support the efforts of the Shanghai Cooperation Organization and other regional actors in playing an active role."* (China, 2016)[49]

[44] SPV.6394 (2010)
[45] S.PV 6896
[46] S.PV 7467
[47] S.PV 7347
[48] S.PV 7771
[49] S.PV 8147

Most interestingly, in 2017, China introduced an alternative concept of community ("a community of a common future" / "community with a shared destiny"), which is overlapping with, and to some degree begins to bypass, the notion of the "international community" (see last row of Table 7).

> *"As resolution 2344 (2017) and the relevant General Assembly resolutions stipulate, parties should be committed to creating a community with a shared destiny in a spirit of cooperation in which everyone wins, and to promoting the economic and social development of Afghanistan and regional cooperation through the Belt and Road Initiative."* (China, 2017)

Overall, different patterns emerge in the data: all P5 regularly refer to the notion of the international community (not least to hammer home their individual take on what should be appropriate community behavior). However, from a diachronic perspective, there is significant change. For China, references to "the international community" have become *the* central reoccurring formulation of their UNSC statements, at the same time as their P5 counterparts refer to it less frequently than before (see figure 11 above). While there are probably many reasons for this change, the coincidental increase in China's usage and decrease in the other P5's usage speaks to other studies that have suggested that China is currently consciously taking advantage of a lack of ideational leadership from other global actors.[50] As the case of Afghanistan shows, China is increasingly using the stage of the UNSC to not only portray its own initiatives as projects *by* and *for* the international community, but to also introduce new and overlapping notions of global or regional "community." Scholars interested in investigating whether these alternative notions are echoed by other UNSC actors – as well as those with any other questions relating to UNSC debates since 2010 – can turn to UNSCdeb8 as an easy-to-use resource to quickly access, (re)assemble and analyze UNSC debates according to their preferred parameters. With this exemplary workflow only one of the many possible sequences of analytical steps, drawing on the functionality of UNSCdeb8, that could be pursued to investigate UNSC debates.

Points to keep in mind

There are a couple of aspects to keep in mind while using UNSCdeb8.

First, some technical issues should be noted:

All UNSC meetings are classified by the UN with one unique S/PV number. However, meetings are sometimes resumed on another day. Whenever that happens, content from different days is merged into combined database-entries, even for meetings in which different speakers represent the same state on these different days. In these cases, the correct order of paragraphs cannot be guaranteed.

Similarly, official representatives speaking on behalf of their country can change during a meeting (due to resumption or other reasons). In this case, the name field in the database contains more than one (i.e. two) name for the participant of a nation.

Speakers in the database can have one of two roles: "President" or "member." Due to the rotating UNSC presidency, when statements by "Presidents" are included, similar vocabulary concerning UNSC debate moderation ("I now give the floor to…") can be expected to increase for all subcorpora. To get clearer results in comparative analysis, it may thus make sense to filter out statements made by "Presidents".

[50] Gore, Lance LP. (2018) "Seizing the "Trump Opportunity" and Engaging the World: Chinese Foreign Policies in 2017." East Asian Policy 10.1, p. 56-67.

Statements by guests, such as experts invited to brief the UNSC, are not included in UNSCdeb8. Neither are statements by diplomats with nationalities *not* represented (at the time of the meeting) as permanent or non-permanent UNSC member-states. Therefore, the database does not mirror the full content of the official (PDF) debate transcripts. Some (possibly extensive) meetings – in which official "member" diplomats did not speak – may thus not be represented in the database.

Second, we reiterate that UNSCdeb8 has been designed primarily to facilitate corpus-linguistic research on the UNSC. Although users are invited to use UNSCdeb8 in any way that suits their purposes, the structure of the available meta-data and search filter options have been built to support both the quick creation of subcorpora and ease in importing this textual data into corpus-linguistic analysis softwares. At this (beta) stage, we do not guarantee that all debates have been parsed with complete accuracy. Errors can be reported by sending an email to UNSCdeb8@swp-berlin.org.

UNSCdeb8's use-value: the key takeaway

In a nutshell, UNSCdeb8 can be used by researchers to quickly assemble specific subcorpora according to their terms of investigation, in order to conduct in-depth discourse or content analysis of UNSC debates. Generally, and regardless of the method employed, UNSCdeb8 can serve as a complimentary information source for researchers unhappy with the existing UN databases, which only allows users to download UNSC meeting records individually, i.e. one by one.

References

Anthony., L. (2014) *AntConc (Version 3.4. 3)*[Computer Software]. Tokyo, Japan: Waseda Uni versity.

Abulof, U., (2015) "Normative concepts analysis: unpacking the language of legitimation", Inter national Journal of Social Research Methodology, 18:1, 73-89, DOI: 10.1080/13645579.2013.861656

Baker, P. (2006) *Using corpora in discourse analysis*, London/New York: Continuum.

Baturo, A., N. Dasandi and S. Mikhaylov (2017) "Understanding State Preferences with Text As Data: Introducing the UN General Debate Corpus." *Research and Politics 4*(2):2053168017712821.

DiCarlo, G. (2011)"Indeterminacy and Vagueness in UN Resolutions Relating to the Second Gulf War." In ternational *Journal of Humanities and Social Sciences* 1.21, p. 46-57.

D. Biber (2009) "Corpus-Based and Corpus-Driven Analyses of Language Variation and Use'", *The Oxford Handbook of Linguistic Analysis,* https://doi.org/10.1093/ox fordhb/9780199544004.013.0008.

Dzudzek I. *et al*. (2009) "Verfahren der lexikometrichen Analyse von Textkorpora" in G. Glasze and A. Mattissek, (eds.) (2009), *Handbuch Diskurs und Raum. Theorien und Methoden für die Humangeographie sowie die sozial- und kulturwissenschafliche Raumforschung.*

Fewsmith, J., (2004) "Promoting the scientific development concept." China Leadership Monitor 11.30, p.1-10.

Gabrielatos, C. and Marchi, A. (2012) Keyness: Appropriate metrics and practical issues. CADS International Conference 2012. Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing?, 13-14 September, University of Bologna, Italy.

Glasze, G. (2007) "The discursive constitution of a world-spanning region and the role of empty signifiers", *Geopolitics* 12, 4

Gore, L. LP. (2018) "Seizing the "Trump Opportunity" and Engaging the World: Chinese Foreign Policies in 2017." East Asian Policy 10.1, p. 56-67.

Hansen, L. (2012) "Discourse analysis, post-structuralism, and foreign policy", in S. Smith, A. Had- field and T. Dunne (eds.), *Foreign Policy: Theories, Actors, Cases*, Oxford: Oxford Univer sity Press.

Howarth, D. (2000) *Discourse*, Buckingham: Open University Press.

Hunston, S. (2002). *Corpora in applied linguistics*. Ernst Klett Sprachen.

Laver, M., K. Benoit, and J. Garry (2003) "Extracting policy positions from political texts using words as data." *American Political Science Review* 97.2, p. 311-331.

Jorgensen, M. and Philipps, L. J. (2002). *Discourse Analysis as Theory and Method*, London: Sage Publications.

Laclau, E. (1988) "Metaphor and social antagonism", in C. Nelson & L. Grossberg (dds) *Marxism and the interpretation and culture,* Basingstoke: Macmillan Education.

Laclau, E. and Mouffe, C. (2001) *Hegemony and Socialist Strategy. Towards a Radical Democratic Politics*, London: Verso, 2nd edition

Mattissek, A. (2010) "Analyzing City Images: Potential of the "French School of Discourse Analy sis", *Erdkunde* 64, 4.

Nabers D. (2015) *A Poststructuralist Discourse Theory of Global Politics*, Basingstoke: Palgrave.

P. Baker, C. Gabrielatos, and T. McEnery (2012) "Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word "Muslim" in the British Press 1998–2009", Applied Linguistics, ams048, https://doi.org/10.1093/applin/ams048.

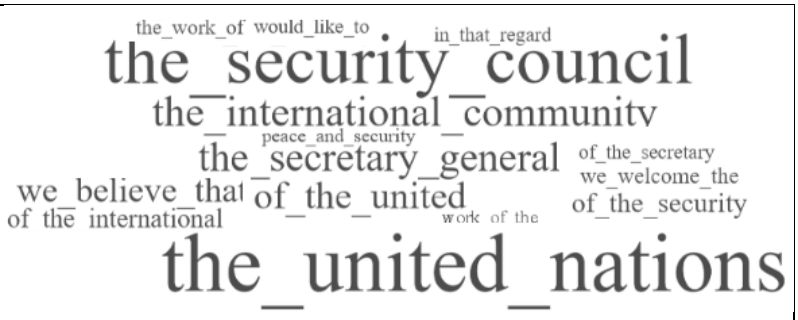
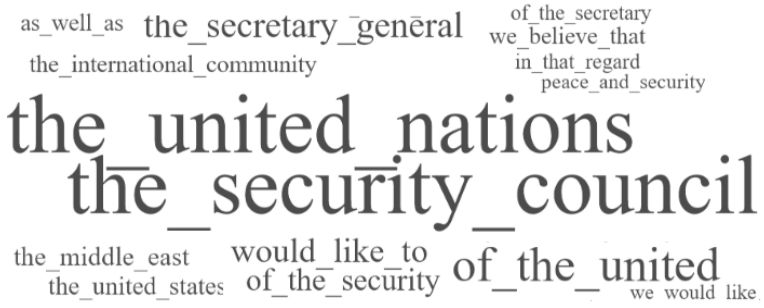
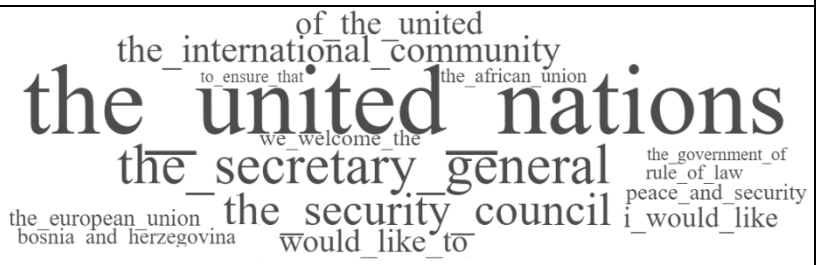
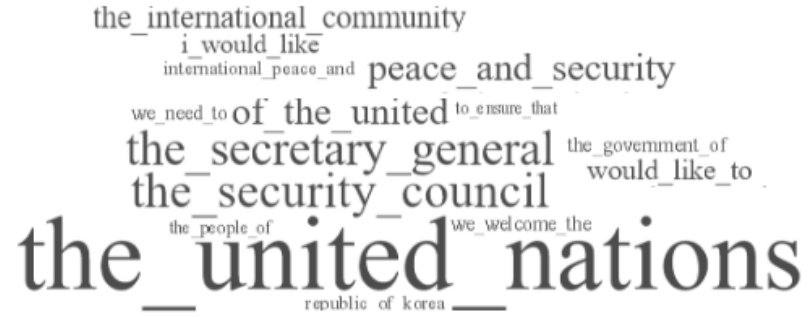
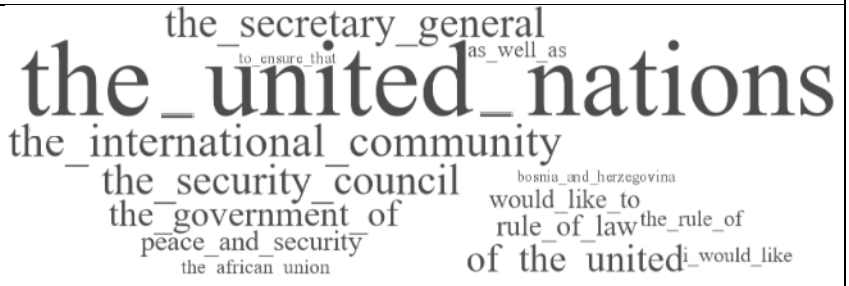
Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp. 1 - 6.

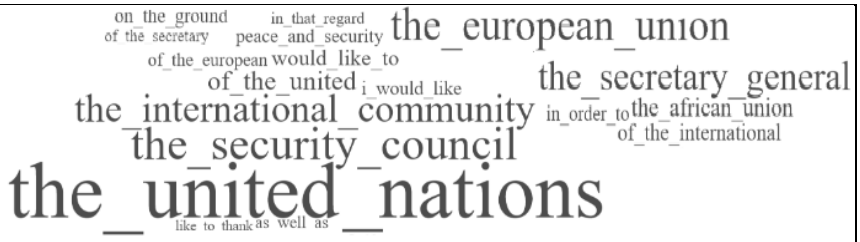
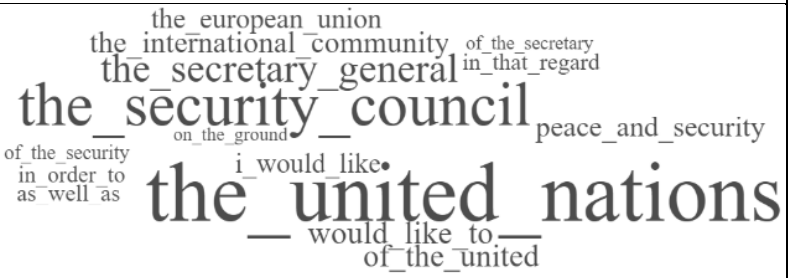
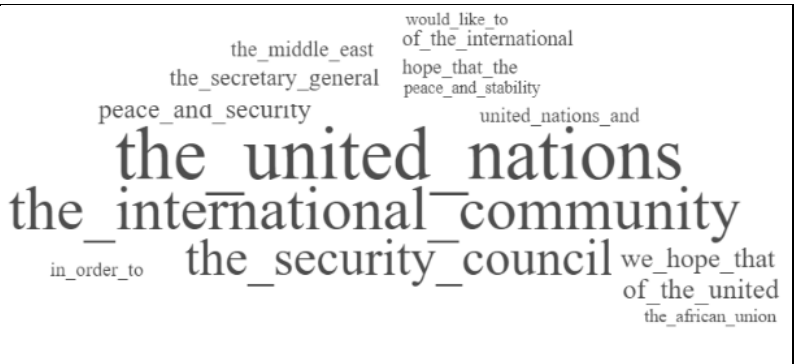
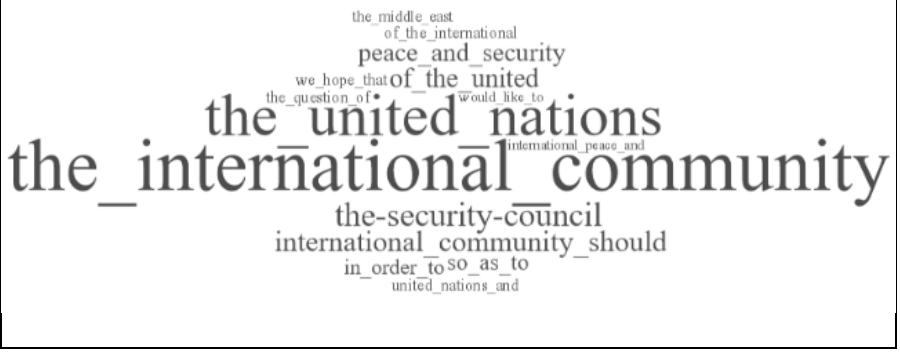
Rayson, P. (2015) "Computational Toolsand Methods for Corpuscompilation and Analysis" in D. Biber and R. Reppen (eds.) *The Cambridge Handbook of English Corpus Linguistics*, Cam bridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139764377.003.

COHA - The Corpus of Historical American English (2002): 400 million words, 1810-2009. Brigham Young University.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B., (2016) "The United Nations Parallel Cor pus", *Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.

Annex

Russia 2010 - 2013	
Russia 2014 - 2017	
UK 2010 - 2013	
UK 2014 - 2017	
USA 2010 - 2013	

USA 2014 - 2017	
France 2010 - 2013	
France 2014 - 2017	
China 2010 - 2013	
China 2014 - 2017	

Top 3-Word N-Grams 2014-2017

China		France		USA	
Freq	N-Gram	Freq	N-Gram	Freq	N-Gram
1061	the international community	1001	the united nations	1218	the united nations
857	the united nations	737	the security council	734	the security council
476	the security council	488	the secretary general	409	the secretary general
421	international community should	384	would like to	379	of the united
383	of the united	379	of the united	374	peace and security
374	peace and security	364	the international community	345	the international community
333	so as to	348	peace and security	252	would like to
307	in order to	338	the european union	216	i would like
234	of the international	330	i would like	203	as well as
231	united nations and	282	as well as	196	international peace and
231	we hope that	282	in that regard	190	members of the
230	the question of	275	in order to	187	on the ground
219	the middle east	241	of the security	187	republic of korea
205	would like to	229	of the secretary	187	the al assad
198	international peace and	226	on the ground	187	the government of

Russia		UK	
Freq	N-Gram	Freq	N-Gram
985	the security council	845	the united nations
980	the united nations	411	the security council
571	of the united	400	the secretary general
432	the secretary general	320	of the united
405	would like to	319	peace and security
370	of the security	258	the international community
318	the middle east	234	would like to
304	the united states	214	i would like
288	the international community	166	the government of
283	we believe that	163	we welcome the
278	as well as	163	we need to
255	in that regard	160	international peace and
250	of the secretary	156	republic of korea
247	peace and security	148	the people of
243	we would like	141	to ensure that

Stiftung Wissenschaft und Politik (SWP) German Institute for International and Security Affairs
Ludwigkirchplatz 3–4, 10719 Berlin TEL +49 30 88007-0 FAX +49 30 88007-100 WEB www.swp-berlin.org