Introducing the PeaceKeeping Operations Corpus (PKOC)

Elio Amicarelli
Independent Researcher

Jessica Di Salvatore
Department of Politics and International Studies, University of Warwick[*]

Abstract

Scholars have used United Nations Secretary General's (UNSG) reports to extract information on peacekeeping operations (PKOs). As key peacekeeping political documents, UNSG reports contain much more information on the politics of peacekeeping. Furthermore, manually extracting information is costly and time-consuming. By providing a machine-readable collection of the UN Secretary General's Reports on PKOs (1994-2020), the PeaceKeeping Operations Corpus (PKOC) offers highly structured and multiformat text data that connects the peace and conflict research community to recent advancements in *text-as-data* techniques. Besides paving the way for the first quantitative content analyses on PKOs, PKOC speeds up and expands the range of information analysable from these documents and allows researchers to query them in a quicker, systematic and reproducible way. In this article, we discuss PKOC's core characteristics. As illustration of the innovative potential of PKOC, we show how *text-as-data* approaches provide more nuanced understanding on PKOs evolution toward multidimensionality, both over time and within missions. While last generation PKOs are assumed to be multidimensional, we show how they vary in multidimensionality and how their complexity also changes throughout their life-cycle.

Keywords: text-as-data; peacekeeping; United Nations

[*] Corresponding author: jessica.di-salvatore@warwick.ac.uk.

Introduction

The UN produces few statistics on peacekeeping operations (PKOs)[1], but plenty of documents that talk about them. Little efforts have been made, however, in using and analysing reports from the UN Secretary General (UNSG) on PKOs as text data. To this aim, this article introduces the PeaceKeeping Operations Corpus (PKOC), a digitised, structured collection of all UNSG reports on PKOs from 1994 to 2020. In the quantitative literature on peacekeeping, these reports are widely used to extract information on missions' activities, cooperation with the government and location of blue helmets (Clayton et al., 2017). We propose to use these documents to shed light on political and strategic aspects of PKOs.

PKOC enables and expands the spectrum of the UNSG reports' usage in peacekeeping research along three main dimensions. First, PKOC enables agile interrogation of a relatively large corpus at the researchers' fingertips. For example, to evaluate research feasibility by checking whether a piece of information is contained in the corpus, it is not necessary to download manually and then read hundreds of documents; as any other digital text, PKOC allows to quickly query the entire body of UNSG reports (1994-2020). Second, PKOC can be used to explore the dynamics of peace operations and test hypotheses by extracting relevant information at the mission-report level. For example, one can test whether different degrees of prioritization of civilians' protection enable the mission to save more lives. Notably, the corpus makes the task of variable creation faster and, more importantly, reproducible.

The third and most innovative contribution of PKOC is that it represents an entry point for conflict and peace research community into quantitative text analysis. Quantitative content analysis treat words and language as data themselves, and performs statistical analysis on large collections of texts. Other research fields in the Social Sciences have benefited form the *text-as-data* wave, particularly with the development of Natural Language Processing functionalities made available for most used statistical software (e.g. *quanteda* package for R by Benoit et al. (2018)). PKOC connects the peacekeeping research community to these exciting trends in quantitative political science (**?**Grimmer & Stewart, 2013). Within extant peacekeeping studies, analysing UNSG reports as *political texts* allows to trace the development of important narratives and discourses around peacekeeping as a policy tool. Some of these developments, such as the prioritization of protection of civilians, are assumed as common and relevant to all recent missions. Hence, PKOC paves the way to novel theoretical insights regarding the UN decision-making process (e.g. how does UNSG reporting influence decisions to extend or adjust mandates?), missions' bias (e.g. does UN reporting of civilian victimization vary by perpetrator?) and institutionalization of norms in International Organizations (e.g. how and when do gender issues enter the peacekeeping discourse?). This article focuses on quantitative uses of the corpus but it is worth pointing out that the analysis of corpora is suited for qualitative analysis as well, particularly discourse analysis (Baker, 2006).

In the next Section (2), we briefly review existing data on PKOs relying on the coding of UNSG reports.

[1] These are also referred to as peace operations, particularly after the Department of Peacekeeping Operations became Department of Peace Operation in 2019.

Then we introduce PKOC and describe the structure and main features of the corpus in Section 3. In Section 4, we explore the evolution of peacekeeping by comparing trends in reporting and multidimensionality over time and, more importantly, across missions. We propose new, more fine-grained measures of missions' features, and use UNMIK in Kosovo as case study to explore within-mission variation in multidimensionality.

Data on peacekeeping operations

The growing academic interest in peacekeeping has been accompanied by important data collection efforts in the last decade. Scholars moved from measuring presence or absence of peacekeeping missions in a given conflict-year to more disaggregated information on the timing, location, personnel type and activities they carry out. The International Peace Institute (IPI) has compiled data on missions' composition (by personnel type) and contributions.[2] In the Geo-PKO project, Cil et al. (2018) estimate the number of peacekeepers deployed subnationally using maps from UNSG reports. Data on UN peacekeeping missions' leadership has also become available (Bove et al., 2020). Two additional sets of data contribute to the disaggregation trend we currently see in the peacekeeping literature. The PKO Location Event Data (PKOLED) is an event-based dataset that codes cooperative and conflictual events involving the UN mission in the host country either as actor or target (Dorussen & Ruggeri, 2007). Finally, the PKO Governance (PKOGOV) dataset codes events involving both UN peacekeepers and the host government (Dorussen & Gizelis, 2013). As for Geo-PKO, PKOLED and PKOGOV data use UNSG reports as key source. To our knowledge, there are three more ongoing data collections using UNSG reports, namely by Smidt on peacekeeping activities during elections, by Hultman on activities related to protection of civilians (see Clayton et al., 2017: for details), and by Kjeksrud (2019) on the use of violence for civilians' protection.

This brief data review indicates that UNSG report are a commonly used source of data among peacekeeping scholars. The reliance on UNSG reports, however, has some shortcomings that are worth pointing out. The reports aim at informing the UN Security Council (UNSC) about some key developments in the host country and policy areas UN troops focus on. However, not only there is variation in the number of reports published each year for each missions, but also the richness and accuracy of reporting vary. After all, UNSG reports are political documents, and may suffer from bias in reporting specific events or developments on the ground. One of the aim of PKOC is analysing UNSG reports as political and strategic rather than operational documents with semantic and linguistic features that reveal more about the politics of peacekeeping. Reports are not expected to be comprehensive accounts of specific events in the field; the purpose of UNSG reports is to inform the UNSC about the overall situation, not to provide the exact account of missions' activities. We provide an overview of what UNSG reports are, what they contain, how they are drafted and the reporting line from the field to the Headquarters (see Appendix A.3).

[2] Available at: `http://www.providingforpeacekeeping.org/contributions/`.

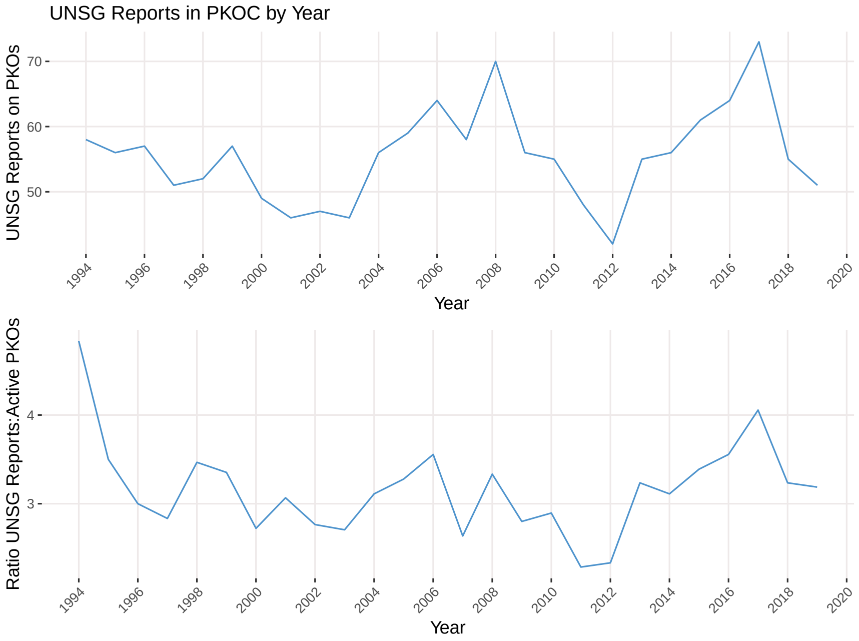

Figure 1. Number of reports by year

PeaceKeeping Operations Corpus (PKOC)

Overview

PKOC is a collection of textual data on peacekeeping retrieved from the UNSG reports. It includes 1,455 reports covering 68 missions from 1994 to 2020. Of these 68 missions, 61 have complete coverage in the corpus since they started after 1994.[3] The corpus is regularly updated via an automated pipeline that downloads new reports available on the UNSG webpage, process them, and add them to the existing corpus.[4] Figure 1 shows the number of UNSG reports included in the corpus by year, as absolute number (top plot) and as ratio relative to active missions (bottom plot). Excluding 2020 (ongoing), the number of reports varies from a minimum of 42 in 2012 to a maximum of 73 in 2017. Despite some fluctuations over time, no clear time trend emerges in the overall reporting activity; even when adjusting by the number of active missions in each year, the trend does not change significantly (bottom graph).

It could be the case, though, that while the yearly number of reports has not changed significantly, their length has increased as missions have become broader in scope. Figure 2 displays the distribution of yearly reports' length as word count. The mean number of tokens (i.e. list of unique words)[5] per year more than doubled (approximately from 3,000 in 1994 to 9,700 in 2019). These figures suggest that PKOs reporting has become either more extensive or more verbose (or both), which may also be due to the rise of multidimensional missions with more tasks, thus more to report on. Figure 3 shows the frequency of reporting for all months of deployment of PKOs authorized since 1994. Most missions receive a good coverage, and frequency of reports is

[3] See list in Appendix A.1.

[4] Since the UN does not have an API to query the documents, the procedure requires a human-in-the-loop approach where a curator checks each step of the automated pipeline.

[5] Tokens are defined as 'a sequence of characters - such as *hairy*, *his*, or :)- that we want to treat as a group' (Bird et al., 2009: 7).

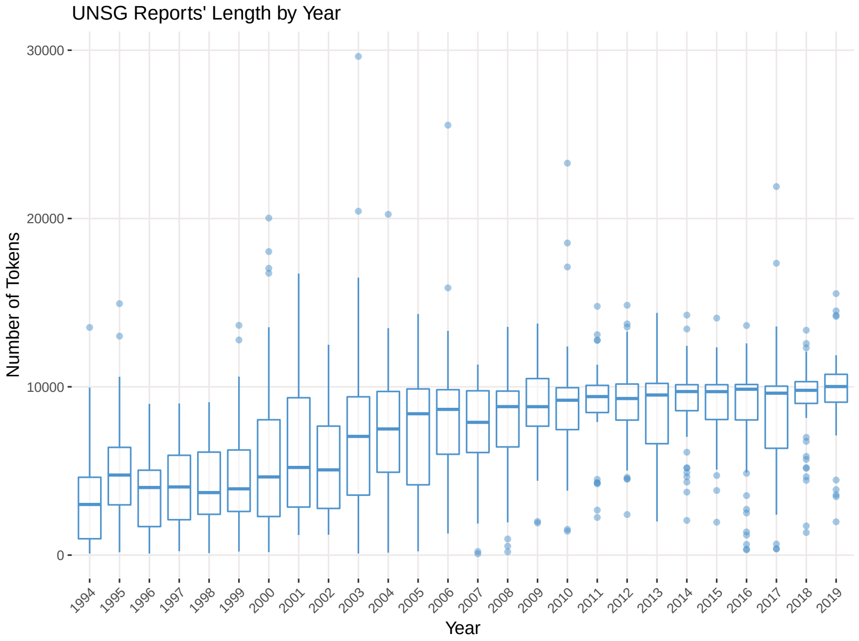

Figure 2. Distribution of UNSG reports' length by year
Boxplots shows yearly interquartile ranges and medians. Circles are values > 1.5 times the interquartile range.

generally consistent within missions. Yet, important variation exists across missions, which may be explained by deployment phase, operational needs or UNSC oversight. For example, reports for the UNAMID (Darfur) or UNMISS (South Sudan) are frequent, probably because of the scale of violence and humanitarian crisis in the region. For data projects collecting using UNSG reports as sources, the problem of reporting frequency does not seem to be a major issue within most missions, but one should be careful when comparing missions to each other.

Next, we provide information on PKOC's structure and content. To simplify the use of PKOC, we have designed it as a highly structured corpus. This means that the documents in PKOC have a rich indexing structure that supports the formulation of *ad-hoc* retrieval tasks. In other words, structured corpora are easier to query thanks to their metadata, a key feature that distinguishes a corpus from a simple archive. In addition, PKOC comes in three different formats and each of them can be seen as a platform that makes certain tasks or types of analysis easier to perform. The *structured* and *multiformat* nature of PKOC is described in the next subsections.

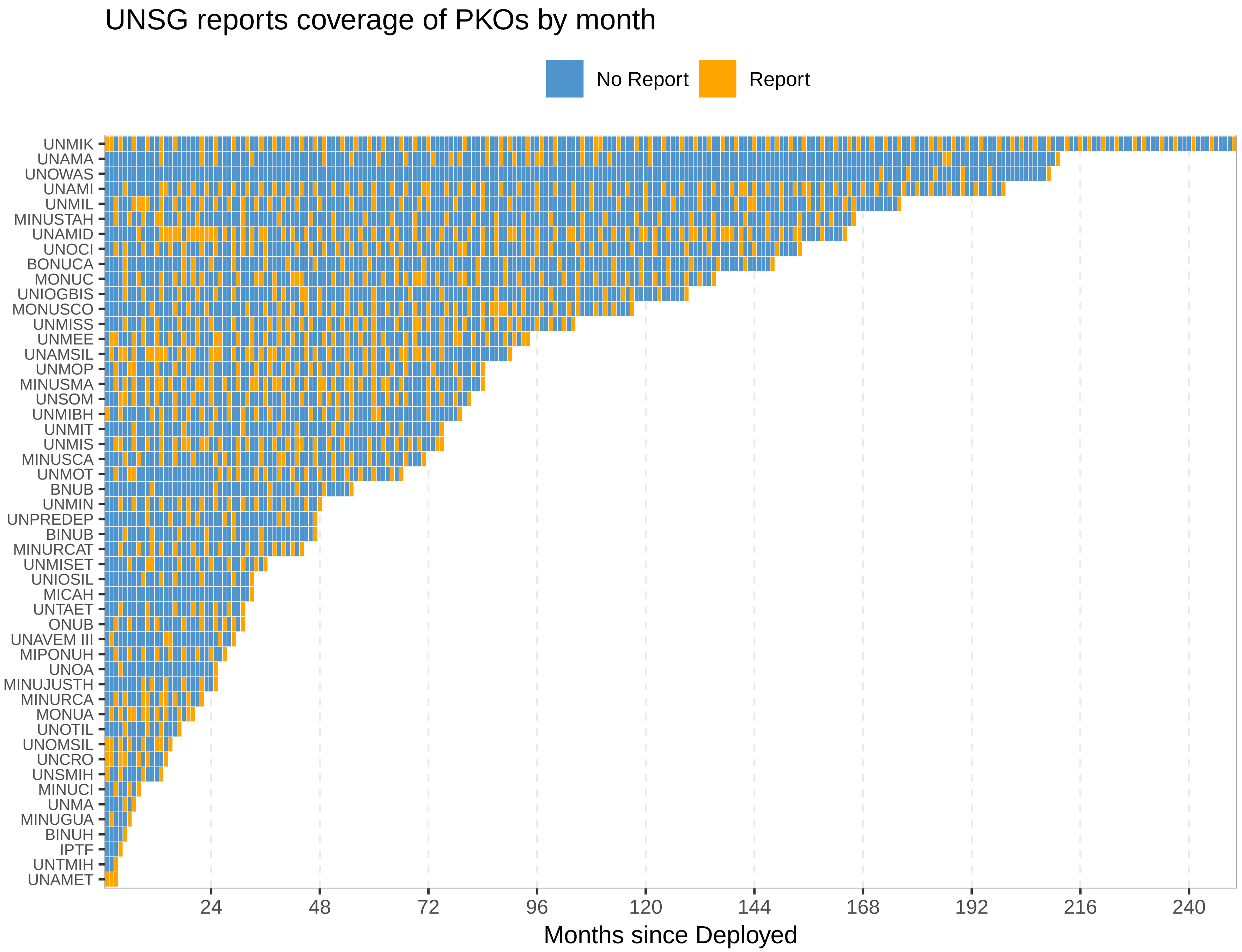

Figure 3. UNSG reporting based on PKOC.

Structure

Digital collections of documents differ on the level of meta-information associated with the documents they are made of. Without documents' meta-information, it would be difficult to retrieve specific texts from a corpus according with a set of desired criteria. PKOC's indexing structure provides meta-information at the report-level on several dimensions: mission acronym, report code, period of deployment, month, year and host country code. Indexes can be combined so that users can easily subset the data according to different needs. If one is interested in exploring how a pre-specified set of missions have evolved over time, she could leverage the mission name and month-year indexes to quickly extract only the relevant documents from the corpus and order them over time. Or, if interested in comparing missions in a specific deployment phase,it is possible to retrieve information by subsetting the corpus using the period-of-deployment index. Since each document in PKOC is linked to country codes, years, months and mission indexes, it can be easily linked to existing peacekeeping and conflict data using the same identifiers.

Formats

When considered as data, text is a ductile resource. It can be used for different types of analysis that require specific pre-processing intermediate steps. For example, researchers interested in discovering topics in documents, usually need to prepare the data such that uninformative words are stripped out from the corpus and to reduce similar words to a common root (this latter procedure is called stemming, Manning et al., 2008). Alternatively, to investigate relationships among actors or extract events from documents, one needs a corpus in a format where words are associated with word classes such as nouns, verbs and adjectives. To give the researcher a head start on different types of analysis, PKOC comes in three different formats. The first format is called plain PKOC (pPKOC). This is basically a digital version of raw UNSG reports 'as they are'. pPKOC has two main advantages: first, it allows to move across reports and to easily query them in their original format, and second it offers to experienced users the possibility to highly customise the corpus. The second format is called reduced PKOC (rPKOC). This contains a pre-processed version of the reports where stopwords[6], punctuation, numbers, symbols and annexes are removed, words are lowercased and stemmed.[7] These pre-processing steps are commonly applied before performing some types of analyses relying on document-term matrices and bag-of-words models (Manning et al., 2008). Many Natural Language Processing methods require text to be pre-processed, hence we provide two ready-to-use reduced versions of the corpus. In version one, only general stopwords are removed, while in version two domain-specific stopwords are also excluded (e.g. 'United Nations', 'Security Council').[8] Finally, the tagged version called tPKOC is similar to the plain one in the sense that text is complete (excluding annexes) but the distinctive feature here is that each word has a set of grammatical annotation. This means that each word is associated with a grammatical class and with named entities. Tagging the corpus is the first step toward automatic information extraction as it allows algorithms to interpret sentences as relationships between entities. To do so, researchers need a corpus that flags which words are actors (i.e. nouns), actions (i.e. verbs). Hence tPKOC is a format useful for all applications (and questions) focusing on the relationships between entities mentioned in the reports. As we are not in a position to assess the accuracy of the UNSG reporting in absence of an alternative data source to compare, we advice great caution in extracting information on missions' activities for the reports. As we discuss in Appendix (A.3), UNSG reports are more accurate to study peacekeeping on a strategic rather than operational level. We return to this issue later on. The tagging has been performed using the Stanford CoreNLP Toolkit (Manning et al., 2014).[9] For further clarification about the different characteristics of the p/r/t formats of PKOC, Table I compares the same sentence as it is represented in the three different formats. As a final note, to facilitate the interoperability of the data across different platforms, PKOC can provide data in different extensions, including Excel spreadsheets.

[6] Stopwords are common words in a language or set of documents (e.g. 'the').

[7] Stemming was performed using the Porter stemming algorithm.

[8] Domain-specific stopwords are words which are common in UNSG reports, i.e. they appear in more than 98% of the documents.

[9] For a complete list of word classes, see Table A.2 in Appendix.

Table I. Example of PKOC formats

Format	*Sentence*
Plain	The Government of the Sudan and the Sudan Peoples Liberation Movement/Army (SPLM/A) agreed to share responsibility, over a period of six and a half years, for creating a new model of governance by restructuring the political system on the principles of democracy and respect for human rights;
Reduced	agre share respons creat govern restructur polit system principl democraci respect human right
Tagged	The/DT Government of the Sudan/NNP and/CC the/DT Sudan Peoples' Liberation Movement/NNP SPLM/NNP agreed/VBD to/TO share/VB responsibility/NN ,/, over/IN a/DT period/NN of/IN six/CD and/CC a/DT half/NN years/NNS ,/, for/IN creating/VBG a/DT new/JJ model/NN of/IN governance/NN by/IN restructuring/VBG the/DT political/JJ system/NN on/IN the/DT principles/NNS of/IN democracy/NN and/CC respect/NN for/IN human/JJ rights/NNS ;/:

The evolution of PKOs

In this section, we use PKOC to analyse the evolution of PKOs and their multidimensionality. Recent peacekeeping missions have ambitious mandates that go beyond security-related tasks and include assistance to governments in reforming institutions, holding elections, promoting human rights and foster economic recovery. Such missions are thus multi-dimensional to the extent that they require more than a military presence focused on the security dimension. They are also referred to as integrated, with other UN agencies working alongside blue helmets to coordinate on development, humanitarian and a variety of policy domains (Fortna & Howard, 2008).

Figure 4 gives a glance into the evolution of protection of civilians (PoC) in peacekeeping since 1994. The plot shows the percentage of sentences in all UNSG reports in a given year that contains expressions related to protection of civilians. There is a common consensus around the fact that PoC has become increasingly important within UN missions, and the increasing use of the key terms in UNSG reporting supports this clearly.

The first UN mission with a PoC mandate was UNAMSIL (Sierra Leone), authorized at the end of 1999.[10] Indeed, mentions to PoC increase since early 2000s[11] (with a peak in 2002), but only pick up after 2009. Unsurprisingly, in 2009 the UNSC agreed that protection activities 'must be given priority in decisions about the use of available capacity and resources, including information and intelligence resources, in the implementation of mandates' (UNSC, 2009: para. 19). The peacekeeping literature has classified missions based on whether they had a PoC mandate using a dichotomous measurement. But with the majority of ongoing PKOs having a PoC mandate, this measurement does not capture much variation across recent missions. MINUSTAH (Haiti) and MONUC (DRC) both had PoC mandates but they they implemented this task. MINUSTAH was *encouraged* to assist the government in protecting civilians, while MONUC's mandate *required* to use all necessary means to implement PoC, with little references to assisting the government in doing so. Not to mention that the security threats to civilians in DRC and Haiti are different. Figure 5 depicts how prominent is protection of civilians

[10] There are rare mentions of PoC before 1999, for example in UNAMIR (1994), UNOMIL (1995) and UNIFIL (1996). Being so rare, they are not visible in the line trend.

[11] Interestingly, and in line with trends shown, Hultman (2013) finds that the implementation of PoC as a norm by the UNSC manifested clearly only after 1999.

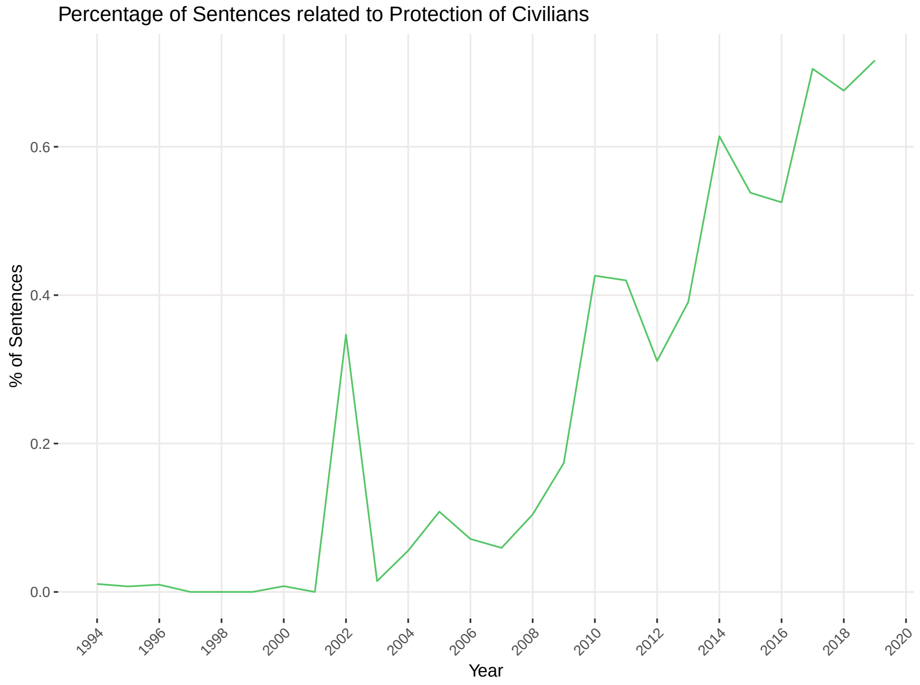

Figure 4. Percentage of sentences containing 'Protection of Civilians' or 'Protection to Civilians' or 'Protecting Civilians'.

in UNSG reporting by mission, to further show how tokens' occurrency highlights cross-missions' variations in PoC focus. In fact, if we compare MINUSTAH and MONUC on this measure of PoC centrality, MINUSTAH is less focused on PoC than MONUC, even though all have a mandate to protect civilians. It is not surprising to see how important is PoC for the UN mission in South Sudan (UNMISS), in light of the current humanitarian crisis the country is facing.

Notably, UNMISS has been struggling in fulfilling its PoC mandate, to the point that the former UNSG Ban Ki Moon warned about the possibility of ethnic cleansing escalating to genocide.[12] If we assume that UNSG reports will not mention issues that missions are failing at, then we would have likely seen less PoC references in UNMISS reports. While tasks where blue helmets do not meet expectations set in mandates are likely underreported, the overall strategic focus of UNMISS is still evident. One reason for this, is that the UNSC may request the UNSG to report on specific aspects of the mandate. We would expect that the more challenges missions face on specific tasks, the more likely the UNSC will make these requests to maintain more oversight. Hence, while the UNSG will report on all policy areas strategically relevant to the mission, what will be exactly reported is a different matter, and pertains to the distinction we draw on the Appendix regarding UNSG as strategy-oriented vs. operation-oriented documents. In the next section, we also briefly touch upon another issue concerning the relationship between the UNSC and the UNSG on peacekeeping, more specifically the relationship between UNSC mandates and UNSG reports.

This illustration points out that dummy variables for PoC mandate are not always well-suited to capture variation across recent missions. We also suggest to use PKOC to construct alternative, continuous measures of missions' PoC focus. Most recent missions are also often referred to as multidimensional. The progressive expansion of the domains in which UN missions operate is often discussed and taken as given, but it has never been examined from a quantitative standpoint. In the next section, we investigate this topic by capturing missions' multidimensionality via the degree of diversifications of the documents' reporting structure.

[12] Available at `https://rb.gy/kqdiar`.

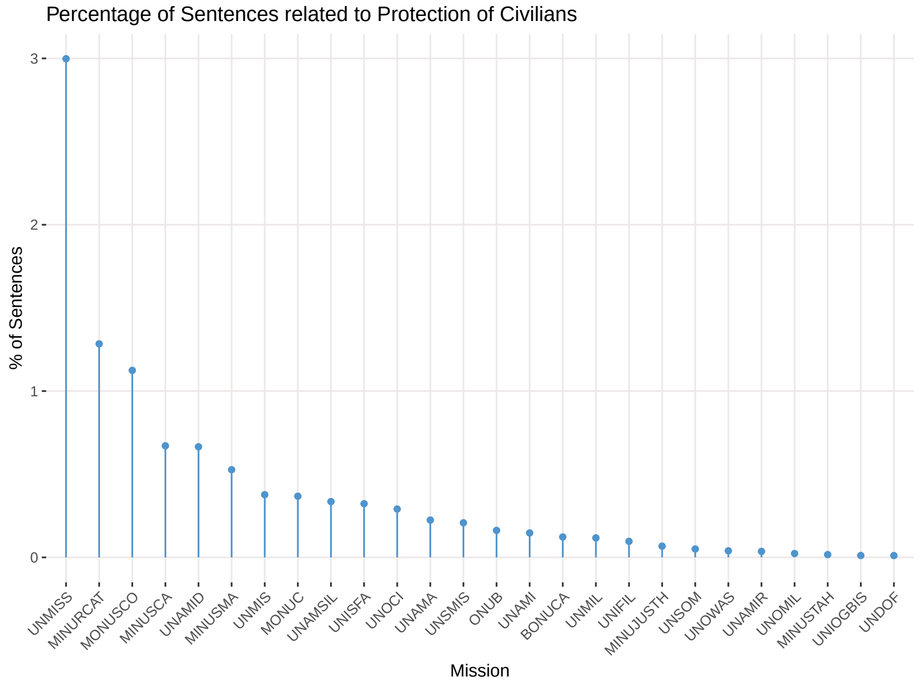

Figure 5. Percentage of sentences containing 'Protection of Civilians' or 'Protection to Civilians' or 'Protecting Civilians' by mission (only % above 0 included).

Multidimensional peacekeeping? A quantitative textual analysis approach

In this section we use PKOC for three goals that can only be achieved with text-as-data approaches. First, we propose and calculate a measure of peacekeeping multidimensionality derived from the structure of UNSG reporting on PKOs; second, we investigate how peacekeeping multidimensionality has changed over time; finally, we zoom in one specific mission to see how multidimensionality changes within the lifespan of a UN peace missions.

UNSG reports are structured in sections that cover a specific topic summarized in a commonly used heading. In Appendix (A.3) we describe the recurrent structure of reports that we leverage to explore strategic developments of PKOs. Once headings are extracted from the corpus, we process them to reduce each report to a concatenation of words. The procedure is detailed in Appendix (A.4). Briefly, once all titles are extracted, we clean the set of titles by removing entries that do not refer to missions' activities (e.g. 'Introduction') or do not refer to a clear domain of activity (e.g. 'Status of Deployment'). The list of informative titles is then stemmed (see section on Formats). Ultimately, each section title is reduced to the minimum parts that are the most informative about the domain of blue helmets' activities. As an example, this four-step procedure would look as follows:

(a) introduction; status of deployment; economic regeneration; reconstruction and development; demobilization and reintegration

(b) ~~introduction~~; ~~status of deployment~~; economic regeneration; reconstruction and development; demobilization and reintegration

(c) economic ~~regeneration~~; reconstruction ~~and~~ ~~development~~[13]; demobilization ~~and~~ reintegration

(d) econom reconstruct demobilizat reintegrat

Groupings of words represents policy domains the UNSG is reporting on. Once stemmed, grammar and ordering of these groupings are no longer informative; occurrences and multiplicity of words are maintained,

[13] See A.4 for the explanation of why development is deleted in this case.

and used to construct an index of mission's multidimensionality. Since ordering is not relevant, we can use a Bag of Words (BoW) model as representation of the documents in the corpus, which allows us to calculate Shannon's index of information entropy to measure documents' complexity. For each document i, we calculate an information entropy index H to quantify the information contained in the document. This quantification considers the probability that a domain appears in the text, so that domains that are less likely to be present are much more informative when they are detected in a document. In the literature on political communication and media framing, Shannon's index is used to understand and measure media's attention diversity, 'that is, the degree to which attention on an agenda is distributed across items, from complete concentration (a single item receiving all attention) to complete diversity (all items receiving an equal level of attention)' (Boydstun et al., 2014: 174). We select Shannon's entropy index H as it is found to be one of the most appropriate measurement of issue attention (Boydstun et al., 2014).[14]

Shannon's index never takes negative values, and equals 0 when the probability of seeing a domain in a document is 1 (i.e. there is no uncertainty). Lower H among documents can also be interpreted as documents having a 'more focused message' (Munger et al., 2019: 826). We use the entropy index H to (i) compare missions to each other, (ii) describe temporal trends in the evolution of peacekeeping, and (iii) explore the evolution of activities within the life-cycle of one PKO.

Figure 6 shows Shannon's information entropy for all missions that started after 1990. Missions that are commonly considered complex and multidimensional such as UNMIK, MINUSCA, MONUC, UNOCI, UNAMSIL, MINUSMA and UNMIL do report high levels for entropy; in other words, they are more likely to report on different domains. At the bottom of the list, among others, we find UNAMET. Given that the mission in East Timor was deployed to organise and verify the referendum on autonomy from Indonesia, it is clear why the mission score so low on the H index. Similarly, UNPREDEP or UNAVEM II had monitoring and verification as main tasks, few domains if compared to recent, more complex operations.

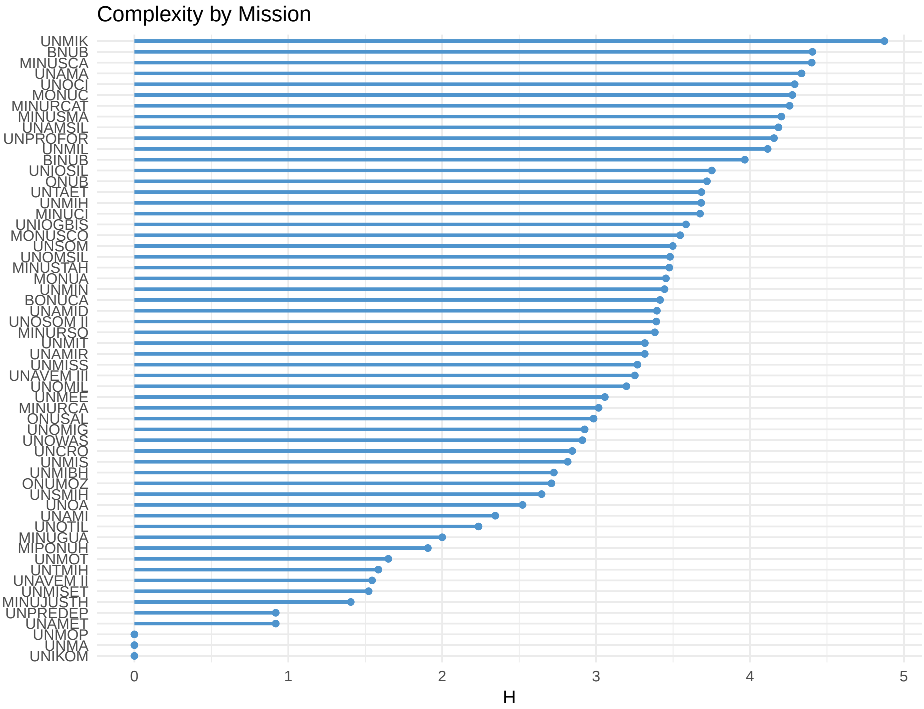

Figure 6. Information entropy by mission.

[14] The formula for Shannon's entropy is as follows:

$$H(X_i) = -\sum_{i=1}^{n} P(d_i) log_2 P(d_i) \tag{1}$$

where $P(d_i)$ is the probability that a domain d (e.g. 'econom') appears in reports for mission X; alternatively, $P(d_i)$ can be thought of as the proportion of total attention the domain d receives in the UNSG reporting for a given mission.

To further understand how missions are classified based on the reporting, Figure 7 compares domains of activity for a relatively simple mission, UNAVEM II (Angola), and a more multidimensional mission, MINUCI (Ivory Coast). It emerges that UNAVEM II focused significantly on the humanitarian crisis in Angola, followed by political developments and the military situation. This is in line with a traditional peace operation. Interestingly, the correspondence with its mandate is not obvious. The mandate encourages support to humanitarian agencies, but the UNSC never explicitly requests it. The dire humanitarian crisis, however, explains why UNAVEM II reports focus predominantly on this issue. The inverse dynamic is also possible. UNAVEM II had a mandate to support the electoral process in Angola, which however something that does not emerge from the reports. Indeed, in 1994, the UNSG reported that: 'Once the outstanding issues on national reconciliation are resolved, the discussions will concentrate on the conclusion of the electoral process and on the future mandate of the United Nations and the role of the three observer States' (S/1994/374), suggesting that other issues (i.e. humanitarian crisis) had to be addressed before carrying out the electoral mandate.

MINUCI's reporting, on the other hand, had human rights and economic development as key concerns. These domains suggest MINUCI was less military-focused than UNAVEM. Furthermore, the number of domains is larger, and includes elections, police and judiciary, most of which were in fact part of MINUCI's mandate. But as UNAVEM II, MINUCI reports also substantially focus on a non-mandated task, namely the economic situation (without a mandate to support economic recovery). This example illustrates two key dynamics. First, UNSG reports do not strictly follow UNSC mandated tasks. They may report on issues that are not part of the mandate or may not report on tasks blue helmets are expected to carry out.

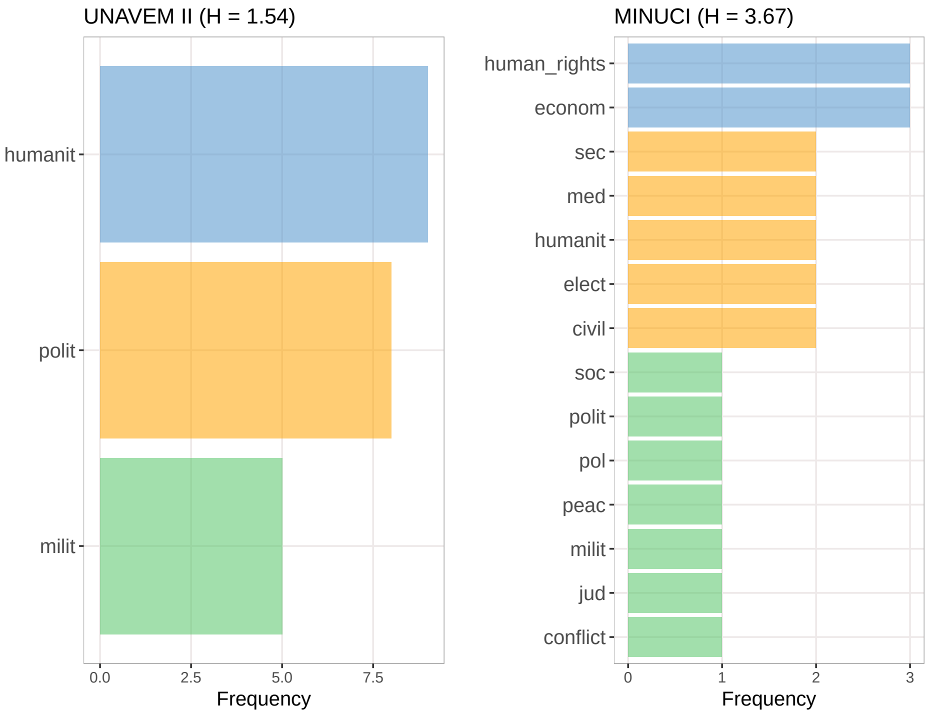

Figure 7. Comparing domains of activity in UNAVEM II and MINUCI.

Second, Figures 6 and 7 show that missions vary in terms of attention they pay to different domains of activities, and it is possible to classify them accordingly.

Besides these static comparisons, we select UNMIK to explore how the domain of activities changes in the life-cycle of PKOs. Figure 8 plots all activity domains of UNMIK on the vertical axis, and the month of reporting on the horizontal axis. Notice each column is a reporting-month, not a deployment-month. Orange tiles indicate that a certain domain is present in reports released in a given month. Fully blue columns indicate reports with no informative titles (see Appendix A.4).

In the first 5 years of deployment, UNMIK activities involved economic, humanitarian and reconstruction

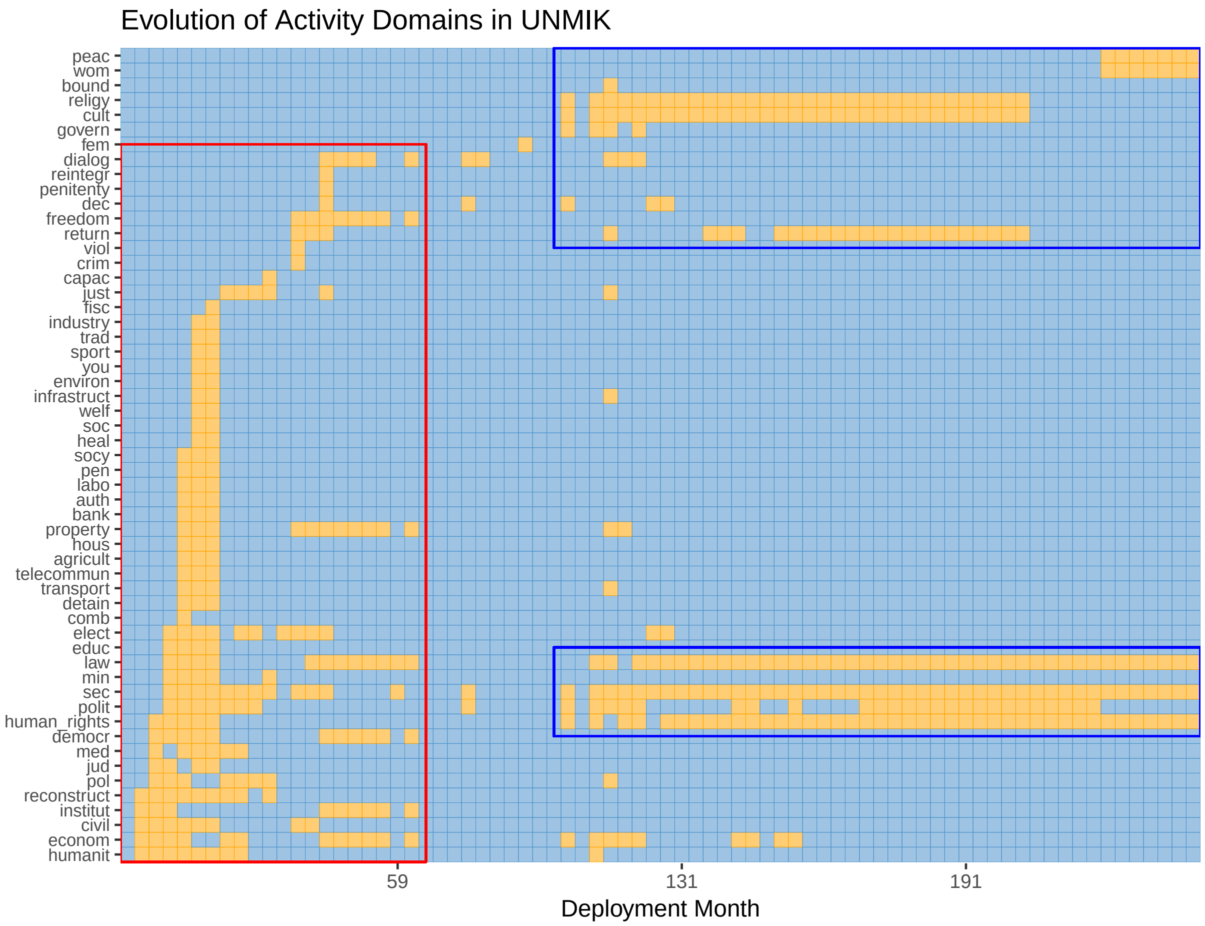

Figure 8. Evolution of UNMIK's multidimensionality.

tasks. Overall, UNMIK performed activities in a wide range of domains in the first phase of deployment (red rectangle). This ambitious approach changed later on, when the mission apparently specialized in fewer domains (blue rectangle), some of which were completely new to the mission, such as religion, culture and women. Other domains that were already present before to some extent, also became steadily prominent, such as rule of law, returnees and human rights.

The evolution of UNMIK suggests that the mission went through some important adjustments. In the first five years, the domains of activities kept changing and the mission seemed to stabilize only in the second half of its (ongoing) life-cycle. As highlighted by Di Salvatore & Ruggeri (2017: 20), '[s]ystematic identification of stages in peacekeeping operations would provide interesting insights into which main phases the most successful missions go through'. PKOC allows to examine whether similar patterns, or 'phases', exist in other PKOs, and to quantify their dynamics in a way that capture variation across missions, which currently remain bundled up in dichotomies that rely on mandates rather than actual activities.

Conclusion

In this article, we introduced PKOC as the first step toward the use of text-as-data on UNSG reports on peacekeeping. The combination of PKOC with other ongoing projects on corpora of text from other UN institutions, such as the General Assembly's speeches (Baturo et al., 2017) and UNSC debates (Schoenfeld et al., 2019), represents a fertile ground for future research that aims at understanding UN dynamics among and within UN key institutions. Indeed, given the strategic and political orientation of UNSG reporting, we believe that this is where the main contribution of PKOC lies. Analysing UNSG report as objective reporting of PKOs should not overlook that the intent of these documents is not to accurately document what blue helmets do, and the risk of reporting bias is not negligible. Hence, PKOC can support research at different stages, from preliminary exploration to measurement of key variables and more advanced content analysis. We discussed how its indexing structure increases the flexibility of interrogation and data extraction, making possible to query the corpus by mission, deployment period, month-year, host country and report code. Furthermore, we discussed how the characteristics of the three corpus' formats (plain, reduced and tagged) enhance the flexibility of researchers on the analytical side by widening the range of tools and perspectives by which peacekeeping can be studied.

References

Baker, Paul (2006) *Using corpora in discourse analysis.* London: Continuum.

Baturo, Alexander; Niheer Dasandi & Slava J. Mikhaylov (2017) Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics* 4(2).

Benoit, Kenneth; Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller & Akitaka Matsuo (2018) quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30): 774. `https://quanteda.io`.

Bird, Steven; Ewan Klein & Edward Loper (2009) *Natural language processing with Python: analyzing text with the natural language toolkit.* Sebastopol: O'Reilly Media, Inc.

Bove, Vincenzo; Chiara Ruffa & Andrea Ruggeri (2020) *Composing Peace: Mission Composition in UN Peacekeeping* . Oxford: Oxford University Press.

Boydstun, Amber E.; Shaun Bevan, & Herschel F. Thomas III (2014) The importance of attention diversity and how to measure it. *Policy Studies Journal* 42(2):173–196.

Cil, Deniz; Hanne Fjelde, Lisa Hultman & Desirée Nilsson (2018) Mapping Blue Helmets: Introducing the Geocoded Peacekeeping Operations (Geo-PKO) dataset. *Journal of Peace Research* OnlineFirst: `https://doi.org/10.1177/0022343319871978`.

Clayton, Govinda; Jacob Kathman,Kyle Beardsley, Theodora-Ismene Gizelis, Louise Olsson, Vincenzo Bove, Andrea Ruggeri, Reemco Zwetsloot, Jair van der Lijn, Timo Smit & others (2017) The known knowns and known unknowns of peacekeeping data. *International Peacekeeping* 24(1):1–62.

Di Salvatore, Jessica & Andrea Ruggeri (2017) Effectiveness of peacekeeping operations. In *Oxford Research Encyclopedia of Politics.* Oxford: Oxford University Press.

Dorussen, Han & Theodora-Ismene Gizelis (2013) Into the Lion's Den: Local responses to UN peacekeeping. *Journal of Peace Research* 50(6):691–706.

Dorussen, Han & Andrea Ruggeri (2007) Introducing PKOLED: A peacekeeping operations location and event dataset. *Conference on Disaggregating the Study of Civil War and Transnational Violence, University of Essex.*

Fortna, Virginia Page & Lise Morjé Howard (2008) Pitfalls and prospects in the peacekeeping literature. *Annual Review of Political Science* 11:283–301.

Grimmer, Justin & Brandon M. Stewart (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297.

Hultman, Lisa (2013) UN peace operations and protection of civilians: Cheap talk or norm implementation?. *Journal of Peace Research* 50(1): 59–73.

Kjeksrud, Stian (2019) Using force to protect civilians. *Replication Data Available at: `https://doi.org/10.18710/FZAVCN`.*

Manning, Christpher; Prabhakar Raghavan & Hinrich Schutze (2008) *Introduction to Information Retrieval* . Cambridge: Cambridge University Press.

Manning, Christopher; Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard & David McCloskye (2014) The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 55–60.

Munger, Kevin; Richard Bonneau, Jonathan Nagler, & Joshua A. Tucker (2019) Elites tweet to get feet off the streets: Measuring regime social media strategies during protest. *Political Science Research and Methods* 7(4): 815–834.

Schoenfeld, Mirco; Steffen Eckhard, Ronny Patz & Hilde van Meegdenburg (2019) The UN Security Council Debates. *Data Available at:* *`https://doi.org/10.7910/DVN/KGVSYH`*.

UNSC (2009) Resolution 1894. *S/RES/1894*

UNSG (2005) In larger freedom: Towards development, security and human rights for all. *A/59/2005*

UNSG (2018) Report of the Secretary-General to the Security Council. *S/2018/747*