International Journal of Social Research Methodology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tsrm20

Natural Language Processing in Mixed-methods Text Analysis: A Workflow Approach

Louisa Parks & Wim Peters

To cite this article: Louisa Parks & Wim Peters (2022): Natural Language Processing in Mixed-methods Text Analysis: A Workflow Approach, International Journal of Social Research Methodology, DOI: 10.1080/13645579.2021.2018905

To link to this article: https://doi.org/10.1080/13645579.2021.2018905

INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
https://doi.org/10.1080/13645579.2021.2018905

Natural Language Processing in Mixed-methods Text Analysis: A Workflow Approach

Louisa Parks[a] and Wim Peters[b]

[a]School of International Studies and Department of Sociology and Social Research, University of Trento, Trento, Italy; [b]Department of History, Johannes Gutenberg University, Mainz, Germany

KEYWORDS Digital humanities; mixed methods; natural language processing; research workflow

1. Introduction

The ever-increasing application of Digital Humanities techniques to social scientific research questions calls for continuous reflection on how they can contribute to scholarly research in combination with other more common text analysis methods. This article explores the various dimensions along which scholarly text analysis can be performed, using combinations of Natural Language Processing and qualitative content analysis methods.

Language technology, also called Natural Language Processing (NLP), provides valuable text and knowledge base-derived information for a variety of analytical aims including description, classification, data reduction, topic and sentiment analysis, and more (e.g. Azzopardi et al., 2016). NLP uses techniques covering statistical and linguistic analysis to produce conceptual text interpretation, for example, in the form of named entities, terms and their relations. These provide useful information for analysis and interpretation at various stages of research. As a computational component of a digital humanities research workflow, NLP forms a bridge between the linguistic surface structure and the underlying conceptual content of textual resources via the computer-based, automatic acquisition of content. In this way NLP can help to unlock large amounts of data for focused manual analysis by experts. When used appropriately, it provides useful tools for conceptual exploration, analysis and knowledge acquisition. To integrate NLP into scholarly research successfully, NLP techniques should be applied in a way that both enhances and focuses text exploration and supports the scholarly research workflow (Peters & Wyner, 2016; Woolf & Silver, 2018).

This article describes a practical methodological position based on a mixed-methods approach, arguing that a careful mix of automated language analysis using NLP and manual qualitative text analysis can help respond to the depth/breadth dilemma often encountered in case study research, whereby more detailed studies cover fewer cases, and vice versa (e.g. Gerring, 2007). Beyond this, it demonstrates that using NLP and qualitative text analysis within a research workflow or cycle can allow researchers to move beyond the examination of the same data through the lenses of different methods for purposes such as corroboration, and towards a deeper exploration that reacts to the findings arising from different methods to hone and generate new avenues of research.

The approach is illustrated with an example from socio-legal research on environmental governance. Specifically, we discuss a triangulation of methods of qualitative text analysis and NLP tools conceived of within a cycle or workflow, as opposed to more traditional linear or stepwise research design. Using mixed methods in a workflow approach allowed the exploration of findings from qualitative analyses of limited text samples to contribute to the resolution of the breadth-depth

CONTACT Louisa Parks Louisa.parks@unitn.it School of International Studies, University of Trento, Via T. Gar, 14, Trento 38122, Italy

trade off by underpinning claims to generalizability (in the illustrative case in the statistical-probabilistic sense, though the approach may also aid with other types: see, Smith, 2018). NLP-based analyses subsequently raised fresh questions to be explored in depth with qualitative techniques. On this basis, the article argues that the attributes of NLP bring possibilities for more focused and fine-grained quantitative text analysis, and that this can be combined with tailored qualitative text analysis methods in a flexible workflow to allow researchers to produce more robust results and identify innovative research avenues.

Before situating NLP in the broader literature on mixed-methods text analysis and describing the illustrative example, some words on our approach to mixed-methods within a workflow are in order. The basis for our claim about the advantages of flexible workflow approaches to mixing methods is rooted in scholarly debates on mixed methods, including questions of how to mix methods with due attention to research paradigm compatibility. It also responds to calls to move beyond methods triangulation for corroborating or challenging findings and increasing their validity, or producing different views of the same data (Hammersley, 2008), towards 'a multi-method logic' that can blend 'the sensibilities of both computational and interpretive approaches to social analysis' (Espinoza-Kulick, 2020, p. 51). Scholars have also called for more attention to be paid to how we engage with computers and analytical software, by unpacking the ontological and epistemological assumptions of research projects and/or computer software (e.g. Jacobs & Tschötschel, 2019) and by leaving space for critical and self-conscious engagement with computers (Hitchcock, 2013), research questions, data and findings, and indeed between social and computer scientists. To take these calls seriously, we place our mixed methods approach to text analysis using NLP and qualitative approaches within a research workflow. As we explore when situating the approach in the wider literature on mixed methods and text analysis, existing scholarship generally uses different language technology and qualitative text analysis approaches in a stepwise approach. This produces valuable and informative results, and has allowed social scientists to come closer to solving breadth and depth trade-offs by providing strategies to test findings from small sample text analyses using automated analysis techniques that can be applied to much larger text corpuses. This generates evidence for, or may challenge, the different types of generalizability that may be sought from findings in smaller scale and detailed qualitative analyses (Smith, 2018).

Placing mixed methods approaches within a workflow brings a further set of advantages. A workflow approach allows researchers to combine methods in a dialogue that answers the above-mentioned calls for more reflection on how we mix methods. A workflow first brings us semantically closer to this kind of flexible, collaborative and dynamic approach. It gives space to allow us to redefine research questions, explore avenues opened by findings suggested by the results of one analytical method using another, and to move back and forth between methods. It builds the re-specification of research questions on the basis of findings, often a task left to conclusions, into the research design itself. There are certainly practical limitations to this: often projects are limited in time and resources, and the research workflow has no clear end point. Nevertheless, to the extent that it forces us to think in terms of a dialectical approach to mixing methods (though not research paradigms), it is a useful heuristic that lies at the basis of the approach discussed and illustrated in the remainder of this article.

In the following section we situate the approach in the literature on mixed methods and text analysis, before describing our application of NLP in a mixed-methods workflow. We then illustrate the approach with an example from an interdisciplinary research project on environmental law and governance. The article concludes by summarizing the suggested approach and reflecting on learnings, limitations and the scope for applications in other social scientific endeavours. Its central contention is that the example of NLP in a mixed-methods text analysis workflow can allow researchers not only to address the breadth-depth trade off, but also to fully exploit the qualities of different tools in a reflexive dialogue between methods, their design, and research questions.

2. Situating mixed-methods text analysis

Mixed methods research was established against the backdrop of heated debates between proponents of constructivist and positivist research paradigms from the 1970s onwards (Denzin, 2012). Although mixed methods approaches have long been used, scholarly reasoning about this 'third paradigm' grew from the 1990s (Greene, 2007). Mixed methods research concerns the mixing of more than methods, though initial debates focused on this and viewed mixing as a triangulation strategy, understood in the sense of using different methods to increase validity (Hammersley, 2008). Yet this obscured fundamental questions that remain under debate. A central theme concerns the mixing of research paradigms, and whether this is possible. Scholars maintain different positions on this according to how fundamental they consider the differences between positivist and constructivist research. The view from major scholars in the area is that these differences remain important, but do not rule out the possibility of mixed methods research. Research paradigms flow from certain ontological bases that shape epistemology and methodology, and mixing these can create incommensurability that must be addressed (Denzin, 2012; Greene, 2007, 2008).

In the approach to mixed methods text analysis suggested here, this issue is tackled in the careful 'tailoring' of methods to research aims. We subscribe to the view that mixing paradigms can be problematic, but that there is no automatic link between methods and research paradigms. A close reading, qualitative text analysis may be used in a positivist paradigm, indeed much doctrinal legal analysis could be classified as such, but also in a constructivist paradigm, as is the case here. The same applies to a distant reading, quantitative text analysis methods. This underlines the importance of methods being carefully designed for specific research tasks (Boréus & Bergström, 2017). It is this attention to how a method is tailored that contributes to the flexibility and dynamicity of the workflow suggested here, which may be used in different research paradigms.

Another central question in scholarship on mixed methods research concerns social justice, and the use of mixed methods specifically in service of social justice by combining different perspectives, including those of marginalized communities (Denzin, 2012). Text analysis has a long tradition of emancipatory aims when used within a constructivist paradigm, particularly in the Foucauldian tradition (Moses & Knutsen, 2012) of revealing the power exercised by societal discourses. In the illustration used here, this aim is central: the discourses that guide decision-making in environmental governance have been argued to exclude the worldviews of indigenous peoples in particular. Using different methods within a constructivist paradigm is intended to reveal whether and how these discourses shape decisions.

Concerning text analysis strategies, in the Digital Humanities in general, and social sciences in particular, these revolve around a number of methodological choices within various analytical dimensions which in turn determine the nature, granularity and quality of knowledge acquisition within the methodological workflow.

Many traditional scholarly approaches in the social sciences involve close reading activities relying on manual analysis. This is the foundation of, for example, formal legal analytical approaches (black letter law or doctrinal analysis) as well as much qualitative content analysis in sociology and political science – whether frame analysis, claims analysis, discourse analysis or other types (Boréus & Bergström, 2017). Depending on their research agenda, scholars tend to take either a predominantly deductive approach, investigating the textual attestation of pre-formed hypotheses, or a predominantly inductive approach, working empirically towards the formulation of one or more hypotheses on the basis of evidence arising from the textual material. The results from these close, manual analyses are rich and detailed, and of vital importance for new areas of study where little is yet known. Content analysis methods are also closely linked to research questions at the ontological and epistemological levels, and crucial for testing and developing theory. Their advantages for these goals are clear and undisputed in terms of the granularity of their results and operationalizability.

Nevertheless, it is also widely acknowledged that the limited quantity of textual material it is possible to interpret using qualitative, manual techniques can create a significant bottleneck for an exhaustive scholarly understanding of the content of the textual source material and the domain it stems from. Close, qualitative textual analysis provides rich and detailed information, but is time consuming and can be 'tedious', which creates significant challenges in terms of researcher errors (Crowston et al., 2012) and unintentional cognitive bias (Boréus & Bergström, 2017; Crowston et al., 2012). The time required for manual analysis means it is difficult or impossible to apply to large amounts of data (Chakrabarti & Frye, 2017). Sampling is necessary, which can limit the generalisability of results, at least in the sense of the term generally applied in positivist paradigms, often demanded of qualitative methods despite the diverse reasoning behind them (Smith, 2018). In addition to this, the plasticity of content analysis methods can affect the perceived validity and replicability of the research. Though these concepts are not applicable to qualitative manual analysis in the same ways as quantitative analyses, if researchers fail to fully explain their methods and decisions taken in their analyses, results can be challenged. This also makes results difficult to replicate or verify through applications to other data (e.g. Chakrabarti & Frye, 2017; Jacobs & Tschötschel, 2019).

Distant reading techniques (Moretti, 2005) involving quantitative analysis resolve many of these problems. These often involve fully automated text analysis – essentially the counting of words, or combinations thereof, instead of manual qualitative analysis. They solve problems linked to the quantity of data to be analysed, since they allow for large text corpora to be handled: whole 'universes' of data can be dealt with, making sampling unnecessary, while issues of cognitive bias and researcher error are reduced. In addition, methods can more easily be made clear to readers, making replicability simpler, and statistical-probabilistic generalization possible (Smith, 2018). Yet this comes at the cost of many of the advantages associated with qualitative manual analysis listed above. While such techniques allow for large amounts of data to be analysed, they do so at higher levels of abstraction, producing results that are often less meaningful for social scientists and the research questions they pose. Mixing methods is thus an attractive solution that allows the resolution of problems associated with each technique. Mixing methods can overcome research errors, involve both representative samples and close attention to contexts, and allow researchers to test results from one analysis using another.

In the reality of doing research, most methods are positioned somewhere between inductive vs deductive approaches, close vs distant reading, and manual vs fully automated text analysis. They are customized to scholarly requirements, and operationalized to intersect and complement each other in order to flexibly address research questions depending on the research domain, the text corpus size, and the research questions of interest. Methods that combine these techniques in mixed approaches allow researchers to switch from more distant and close reading, automated and manual, and deductive and inductive methods. Existing examples of mixed methods approaches in text analysis vary along these lines, for example, building on knowledge gained from close reading analyses using digital story grammars for discourse analysis (Andrade and Andersen 2020); employing multiple methods to view texts from different angles (Di Giammaria & Faggiano, 2017); verifying the results of manual coding using natural language processing (Crowston et al., 2012); or taking stepwise approaches following topic modelling with discourse analysis (Jacobs & Tschötschel, 2019) or network analysis (Espinoza-Kulick, 2020).

Linking these observations back to the literature on mixed methods research, it has been underlined that more attention needs to be paid to exactly how to mix methods by pinpointing the drawbacks of methods that need to be offset by mixing, and to questions around mixing paradigms and interpretive strategies (Greene, 2008; Johnson et al., 2007). These themes are not often explicitly addressed in existing mixed methods studies combining text analysis methods, which we try to remedy here.

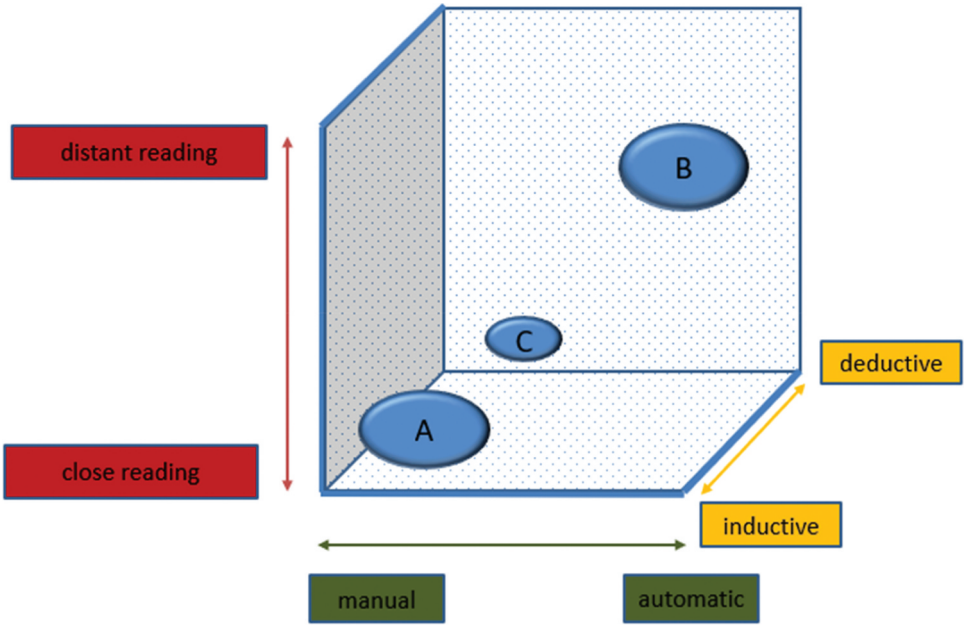

Figure 1. A three-dimensional space situating mixed methods approaches to text analysis.

To situate the mixed-methods approach to text analysis illustrated here, which combines qualitative close reading methods (both inductive and deductive) in a dialogue with quantitative distant reading NLP methods (again both inductive and deductive) Figure 1 describes a three-dimensional space along these axes. Different mixed-methods approaches can be positioned within this space.

In Figure 1, approach A represents a qualitative, manual close reading method within an inductive setting that aims at incremental theory formation on the basis of manual analysis. This would apply, for example, to a discourse analysis approach in an understudied area where the goal is to build knowledge and hypotheses. Approach B represents a fully automatic distant reading analysis, again within an inductive approach. An example here would be when new large sources of textual data become available (such as social media data in recent years) and researchers wish to take a broad approach to see what the data might reveal. Approach C can be characterized as a predominantly close reading activity, with some computational support, in which hypotheses are deductively attested. This is the case for research where software for qualitative text analysis is used to support the identification of patterns and relationships within a coded body of texts.

Many mixed-methods approaches to text analysis in the social sciences, including those cited above, combine a qualitative and manual method (Approach A/C) and a quantitative and distant method (Approach B) in a stepwise manner, beginning with either a deductive or an inductive approach before passing to the other. In the recent examples cited, Andrade and Andersen first use close reading analysis, then a digital story grammar method to build on emerging findings (2020). Similarly, Crowston et al suggest that Natural Language Processing methods be used to verify results emerging from qualitative, close reading analysis (2012). Other stepwise combinations begin with quantitatively-oriented and distant methods: topic modelling results are deepened with discourse analysis by Jacobs and Tschötschel (2019), and network analysis by Espinoza-Kulick (2020). Di Giammaria and Faggiano take a pluralistic approach, approaching texts using different methods to uncover different findings (2017). Bucchi et al. (2019) take an approach closer to that suggested here, using qualitative text analysis to elucidate findings from a quantitative analysis.

The approach we propose here seeks to move beyond stepwise approaches towards a dialogical approach within a workflow. It relies on qualitative, manual close reading methods supported by tailored computational support to enable focused close reading, quantitatively based interpretation. This approach addresses the question of research paradigm commensurability by tailoring methods to research agendas situated in specific ontologies. The combination of close and distant reading techniques is developed to respond to the specific strengths and drawbacks of each, thus paying attention to offsetting and allowing the corroboration of findings. Beyond this, we also suggest that following this type of methodological workflow instigates a dialogue between findings based on close and distant reading techniques. This can allow social science researchers to find new ways to move beyond stepwise mixed methods approaches designed to test or deepen a particular set of findings. It seeks to respond differently to the longstanding question of depth vs breadth, and open up to questions posed in the literature on mixed methods research about how to take a dialectical approach to the research process (Greene, 2008). The approach is based on a living, cyclical dialogue between research questions, tailored methods, and interpretation, moving research agendas forward by deepening findings. It fully exploits the sensibilities and strengths of the methods employed, and in that sense goes beyond merely offsetting their weaknesses.

Given the relative scarcity of literature on NLP applications in the social sciences (Crowston et al., 2012), an overview is useful before describing our approach. Crowston et al. (2012) provide a succinct introduction to NLP and illustrate how it may be applied in qualitative research in the social sciences. They define it as 'a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis' (Liddy 2003 cited in Crowston et al., 2012, p. 525). NLP is not, then, a single method *per se* but a set of tools and underlying linguistic and statistical assumptions about language. Referring back to Figure 1, NLP tools also vary in terms of where they fall on the continua between close and distant reading, and manual and automatic analysis. They can equally be combined to develop methods within inductive and deductive approaches. They include tools that allow researchers to deal with large text corpora by applying various techniques such as linguistic analysis, named entity recognition, term extraction and relation extraction (Peters & Wyner, 2016). The extracted information is then associated with source texts through text metadata in the form of annotations. Given the variety of NLP tools available, it is important to underline that their use in social science research depends on careful tailoring or customization of the analysis to research questions and expert knowledge, though of course this is a generally applicable point when using software (see, e.g. Woolf & Silver, 2018). The importance is perhaps greater for NLP as a mixed tool comprising different techniques. Each of these needs to be chosen in light of the research interest and on the basis of its positioning along the axes in Figure 1. This customization process also mirrors the dialogue that then takes place amongst research findings and research questions to some extent, underlining once more the importance of moving between research questions and findings both in the preparation of methods and the interpretation of results. Indeed, while NLP offers more automated tools, the interpretation of findings remains in close reading techniques and the expertise of social scientists.

3. NLP in a mixed methods text analysis workflow

In general, fully automated text analysis using NLP techniques creates material for distant reading and does not generate knowledge that is rich, focused and detailed enough to answer many of the research questions posed by social scientists. If used by researchers as the sole source of textually derived knowledge, the results obtained from automatic analysis risk misinforming, overloading and/or confusing the scholar, thus obfuscating their research targets (Chakrabarti & Frye, 2017; Peters et al., 2019). There is equally a danger that the researcher relies on automatic tools to 'do' analysis, rather than tailoring the way they use such tools to generate data they then interpret: in other words, tools may be used in ways that are not appropriate to the research agenda in hand, and may be expected to carry out the analysis of the data they produce to too much of an extent (Woolf

& Silver, 2018). In addition, we argue that beyond the necessity to shape automatic tools to fit with specific research questions, a fine-tuned trade-off between manual and automated analysis is required to address the knowledge acquisition bottleneck (Peters et al., 2019; Peters & Wyner, 2016).

A flexible and targeted collaboration is thus needed, between research aims and tools, but also among scholars with expertise in automatic text analysis and the social scientists using them. Conceiving of mixed methods research as a stepwise and linear process, where one method produces results that are then verified or further probed with a second can be unhelpful in this view, as it leaves any adjusting or respecifying of research questions to the end of the research process. In other words, reflections on promising research avenues emerge in the conclusions of a project, where social scientists traditionally pose new questions on the basis of their data and interpretations. A mixed methods research design that pays attention both to selecting and honing automated and manual tools for text analysis throughout the research process, thus building in reflection and dialogue between researchers, research questions and findings, resembles a research workflow rather than a stepwise approach. This mixed methods workflow allows for the full exploitation of the potential of mixed methods beyond a stepwise, linear application while retaining the benefits related to resolving aspects of the breadth-depth trade off and ensuring paradigm commensurability.

In more detail, because the approach is a workflow that focuses on back-and-forth dialogue between researchers, tools, findings and questions throughout, there is no single 'starting point'. Rather, on the basis of the qualities of the field the research addresses, different starting points may be chosen. Research questions on themes where little knowledge is available might likely begin by seeking to generate in-depth knowledge using qualitative close reading methods in an inductive approach, while better known fields where research questions aim to test pre-defined hypotheses might likely begin from broad sources of existing knowledge and probe these using quantitative distant reading methods in a deductive approach (see, e.g. Della Porta & Keating, 2008). While our description here begins with NLP tools, this is not necessarily the starting point.

The first step in the research process is the specification and tailoring of tools and methods to the questions and aims in hand. This is no different for NLP tools. Giving precedence to specific research questions requires the application of NLP microtasks at more fine-grained junctures within the workflow when compared to computational applications. Engaging in a collaborative approach to the application of NLP, social and computer scientists must first work together to clarify how NLP should be used for a specific research problem in a particular paradigm, given the interdisciplinarity inherent to mixing methods from computer and social sciences. If the research problem is broad, selected NLP techniques will be used to provide partial data in an inductive approach which are then evaluated to allow the specification of more targeted questions. A second scenario sees social and computer scientists collaborate to use NLP techniques to verify research questions or hypotheses, whether they emerge from existing literature, or (as in the illustrative case presented here) from qualitative, close reading text analysis. The data provided via NLP techniques may reveal information about types of terminology, their frequency, high-level information about patterns, omissions, and the juxtaposition of terms within texts, or other research question-driven information. These data can be further pursued using NLP techniques before interpreting the findings with a view to honing or respecifying research questions and avenues.

Turning to qualitative, close reading methods, once again the first step is tailoring these to the research paradigm, aims and questions of interest. Social scientists build the text analysis approach most appropriate to their aims, given the term covers a range of methods with varying ontological and epistemological roots (Boréus & Bergström, 2017). Amongst the qualitative close reading variants are discourse analysis, frame analysis, claims analysis, narrative analysis, and more. The careful choice of an approach, and how it is constructed for the purposes of the research question, is perhaps taken for granted by qualitative researchers but, as noted, is sometimes overlooked where computer software comes into play. Once again, the choice of qualitative close-reading method may

take an inductive turn in response to the need to generate detailed knowledge, or it may be deductive, following avenues suggested by a previous round of NLP-generated data or literature-guided hypotheses. The variety of qualitative, close-reading text analysis methods makes further specification counteractive. However, following Boréus & Bergström (2017) they may be analyst, producer, addressee or discourse-oriented in their interpretations, and will produce detailed data that underpin nuanced findings about the patterns and features within a limited number of texts. Qualitative data analysis software tools may also be used to apply these methods.

Regardless of the starting point in the research workflow, the next stage includes a fresh dialogue involving the interpretation of the data, research aims and questions, and the subsequent tailoring of tools to follow promising (and feasible) avenues. This is in line with a pluralist approach to social science research (Schmitter, 2008), as well as the call for dialectical mixed methods approaches (Greene, 2008), both being characterized by dialogue between data and interpretation. Undoubtedly, there are potential drawbacks here: the workflow approach requires more effort and teamwork, and may well be more time-consuming. Researchers may be wary of proposing research that appears open-ended, not least because funders may not be convinced by such designs (Parks & Morgera, 2019). In this vein, scholars within digital humanities rightfully object to regarding automated analysis as a one stop solution that drives the scholarly research process (see, e.g. Hitchcock, 2013). Incorporating automated techniques into humanities research is still a contentious issue. A completely automatic analysis does not meet requirements of quality and rigour, because it presupposes scholars are comfortable setting wider tolerances for error and assessing large amounts of data potentially irrelevant to research questions. Across disciplines, it is recognised that technical fixes will achieve little unless they are embedded in, and customized to, a broader understanding of the rationale and assumptions behind qualitative research (e.g. Zelik et al. 2007). The workflow approach takes this issue seriously. In addition, if applied with clear research aims and a clear timetable there is no reason that the advantages of mixed methods in terms of flexible dialogues and addressing the breadth/depth trade off cannot be delimited. The payoffs, compared to these concerns, can be significant. As our illustration in the following section demonstrates, this mixed methods approach allowed us to uncover new information about how an international treaty is understood and applied. Figure 2 provides a condensed representation of the mixed methods text analysis workflow, underlining that the research process is wholly driven by the research questions, which can be flexibly (re)formulated within a methodological feedback loop.

4. An illustrative case: socio-legal research on indigenous peoples and the Convention on Biological Diversity

We illustrate the use of NLP in a mixed-methods text analysis workflow using research on the discursive construction of the importance and role of indigenous peoples and local communities (IPLCs)[1] in the CBD. The research context was a 5-year project on benefit-sharing in international environmental law that included interest in the portrayal of this group in the decisions of the CBD's Conferences of the Parties (CBD COPs) and its implications.[2] Our illustration focuses on one moment of dialogue between research questions, methods and results, beginning with a qualitative close reading discourse analysis in a deductive approach and moving into a quantitative distant reading application of NLP tools applied inductively. We also describe further dialogue, albeit more cursorily.

The research question emerged from a qualitative comparative case study of local community discussions around issues of benefit-sharing as well as existing literature.[3] This raised questions about how to understand whether, and where, these issues were reflected in the decisions of the CBD COPs, addressed with a qualitative, close reading discourse analysis within a deductive approach. The results of that analysis raised a further question: would the development of the discourse about IPLCs be confirmed in the entirety of the texts of CBD COPs? To move towards this aim, a test was developed using NLP tools to reconstruct the analysis

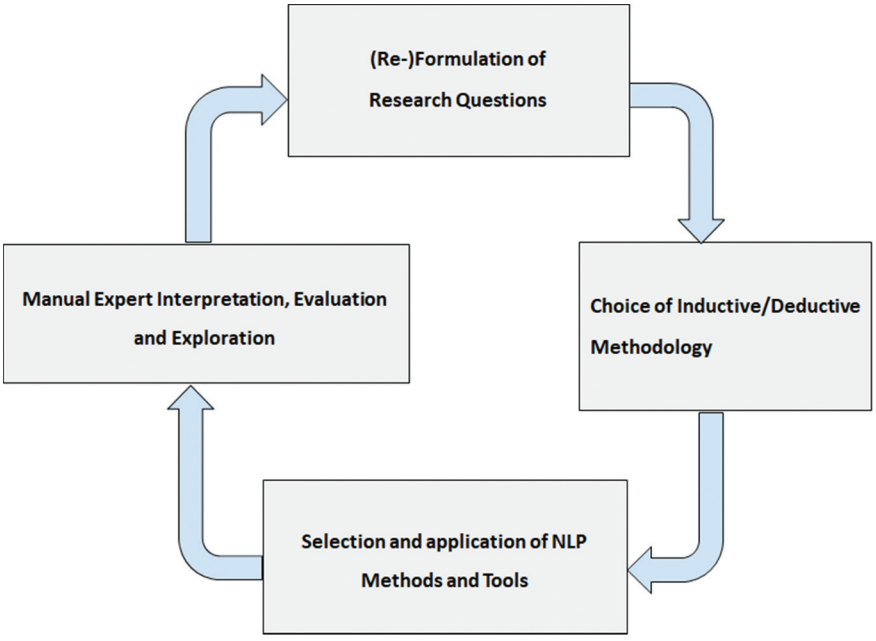

Figure 2. NLP in a mixed-methods text analysis workflow.

using 3 key COPs. In further steps, the dialogue between questions, methods and findings also continued around questions of participation (Zelik et al. 2007). Here we focus on the former moment of dialogue within the research workflow, which addresses the breadth vs depth conundrum.

In more detail, the initial research question raised by the comparative case studies revealed some common issues, as well as confirming an enduring question arising from the literature about whether or not the discursive construction of arenas of global environmental governance leaves spaces for issues considered important by IPLCs. Given that much of the political and sociological literature on this question is based on analyses of treaty texts, while legal scholars argue that the CBD evolves through its subsequent application via COPs, it was decided that an analysis of the CBD COPs would be a useful first step. A qualitative close-reading approach was chosen to generate in-depth and fine-grained findings about discourses around IPLCs over time. Given the, sometimes opposing, claims about these discourses in the literature, a deductive approach was developed following these. A purposive sample drawn from the CBD COPs was constructed given the size of the text corpus, which comprises 14 sets of decisions that run to hundreds of pages. The sample was built using a list of keywords selected by legal experts to identify all passages referring explicitly or implicitly to IPLCs. These texts were then manually coded by two researchers, following inter-coder reliability tests, according to their expressions of different 'discourse categories'.

This analysis produced fine-grained, nuanced findings about the evolution of the discourse on IPLCS. It focused on discourse categories about the recognition of traditional knowledge, participation, local initiatives for protecting biodiversity, the valorisation of diverse worldviews, and (more or less) opposed categories on the emphasis of science in a modern or western view, exclusion from decision-making and implementation, the imposition of decisions, and an emphasis on market or capitalist reasoning. It showed that these discourses had not remained fixed over time. Decisions recognising the value and importance of the traditional knowledge held by IPLCs for protecting biodiversity increased with clear peaks at the seventh, tenth and fourteenth COPs. Similarly, talk

about IPLCs' participation increased over time, with peaks in the seventh and fourteenth COPs. The discourse categories that were more negative for IPLCs were much more infrequent, and rare compared to these.

Having begun from this deductive approach and a qualitative, close reading method, we then interrogated our findings. A number of questions arose, including whether the evolution of the discourse around IPLCs would hold in an analysis comprising all of the texts of the CBD COPs. To investigate, a quantitative distant reading method using NLP tools was developed. This required a collaborative dialogue between computer and social scientists to tailor NLP tools to the research in hand. The first phase was an inductive domain exploration, where social and computer scientists worked to sketch the semantic scope of the text corpus. First, salient concepts were identified using the TermRaider tool, part of the General Architecture for Text Engineering (Cunningham et al., 2002).[4] This tool combines linguistic analysis, such as the addition of part of speech information to words and grammars, defining the possible combination of parts of speech in phrases. In this way, it distinguishes term candidates, in this case terms such as 'biodiversity', 'social benefit', and 'delivery of environmental benefit'. A termhood score was then computed and assigned to each candidate, then discussed in decreasing order by the computer and social scientists to assess relevance for further analysis. When necessary, the team looked for context information using AntConc (Anthony 2016). The manually approved terms were linked to the texts as textual metadata in the form of annotations to allow their further use in text analysis. This first part of the research intended to shape the NLP tools to the specific research question by forming a relevant terminological vocabulary to inform a subsequent deductive exploration drawn from the results of qualitative, close reading text analysis (i.e. did the story of the discursive evolution around IPLCs apply when considering the entirety of the texts of the CBD COPs?).

The results of this first application of NLP tools were broad, providing information at a high level of abstraction. We therefore decided to proceed with an initial test case to probe the accuracy of the data produced, again with tailored tools. We zoomed in on three texts accounting for the peaks revealed in the first, qualitative close reading discourse analysis, and which legal scholars advised us were similarly identified in their discipline as important sources.[5] As detailed in Figure 1, we thus moved along the quantitative/qualitative axis, but retained the use of distant reading NLP tools in order to see whether our results would reproduce accounts similar to those in the legal literature. If this were the case, it would confirm the accuracy of the tailored NLP tools and allow further steps in the research workflow. If not, we would proceed with further tailoring.

NLP tools were thus used in this second phase to compare and contrast the selected texts. The terminology contained within each document provided a semantic signature made up of vocabulary terms as well as actors, which highlighted differences in perspectives on those actors between the documents. Given the research question, the interpretation of this data concentrated on IPLCs. Through text and its annotation with terminology and actor information, co-occurrences of terms within each paragraph were identified. The data were interpreted proceeding on the assumption that these co-occurrences denoted thematic relatedness and would inform us about the shape and evolution of discourse in the three documents. The interpretations were then shared with legal experts in the wider research team to see how far they tallied with their understandings.

The overall story that emerged from the analysis indicated that the language used about IPLCs appears to evolve from a more instrumental view towards one of a more or less independent group of rights-holders. This emerged from the differences in term overlap between the different texts. For example, the term 'involvement', which appears 5 times in the Akwé: Kon Voluntary Guidelines (Akwé: Kon) (Convention on Biological Diversity, 2004) but 21 times in the Mo'otz Kuxtal Voluntary Guidelines (Mo'otz Kuxtal) (Convention on Biological Diversity, 2016), indicates an increase in talk about what 'involvement' should entail for IPLCs over time. Unique terms linked to participation were also informative. In the first text, Akwé: Kon (Convention on Biological Diversity, 2004), common terms included 'consultation', 'public consultation', and 'stakeholder'. In the Tkarihawié:ri Code of Ethical Conduct (Tkarihawié:ri) (Convention on Biological Diversity, 2010) these terms disappear,

replaced with language that suggests reflection on why participation is needed such as 'ethical conduct', 'respect', and 'sacred sites and species'. Later, in Mo'otz Kuxtal (Convention on Biological Diversity, 2016), the emphasis shifts to the theme of making participation 'effective'. On this basis, we suggested that there was evidence of an evolution from a rather bureaucratic view of IPLC involvement as 'consultation', to an emphasis on ethics, to an effort to impart meaning to IPLC participation as distinct from mere consultation.

This story tallies with key findings in the legal literature, and in this sense confirmed the accuracy of our tailored NLP tools, paving the way for further steps in the research workflow. For reasons of space these steps are not described in detail here. Briefly, they included dialogue on findings and the tailored development of a second qualitative, close reading frame analysis in an inductive approach which generated in-depth findings about the meaning of 'participation' in CBD COP decisions, and preliminary research about the uniqueness of the discourse surrounding IPLCs in CBD COPs when compared with those relating to other actors. What these further steps in the research workflow illustrate are the continued advantages to be drawn from the dialogue between research questions, methods, and findings. The findings in this workflow not only answered breadth/depth trade off issues by increasing the validity of findings based on qualitative close reading analysis of sampled text. They also responded and evolved in dialogue, allowing for fresh research questions and deeper findings to emerge. Arguably, this means that methods were thoughtfully mixed both to offset blind spots and fully exploit their strengths, producing more than the sum of their parts, as they would have done in a stepwise application.

5. Conclusion

This article detailed an application of Natural Language Processing in a mixed-methods text analysis workflow. It began by arguing we need to conceive of mixed methods approaches as workflows and as cyclical rather than linear. This obliges us to think carefully about how we construct methods, and allows findings and research questions to enter into dialogue. We then discussed the application in light of the existing literature. We noted that mixed-methods text analysis approaches can be understood along three axes: distant to close reading, manual to automatic analysis, and inductive to deductive approaches. Much work using mixed-methods approaches to text analysis use linear, stepwise approaches in one direction or another along these axes, producing important findings and resolving the bottlenecks represented by qualitative, close reading techniques on one hand, and the challenges posed by abstract and high-level findings emerging from quantitative, distant reading techniques on the other. To gain even more from a mixed-methods approach, researchers can pay attention to important questions raised in the literature on mixed methods research by increasing the dynamicity of their approaches, and engaging in a dialogue between findings, research questions and methods in order to carefully mix methods with attention to research paradigms and offsetting to exploit the strengths of the methods they use to the full. We illustrated with such an approach using NLP and discourse and frame analysis to explore the roles of indigenous peoples and local communities portrayed in decisions of the Convention on Biological Diversity. This demonstrated how the application of NLP in a mixed-methods text analysis workflow allowed findings to frame emerging research questions and feed into attention to the construction of methods for scholarly interests. Although this type of research may involve some drawbacks, payoffs can be substantial and applicable in the many areas of the social sciences where dynamic responses to the implications of texts are needed. Not least, in the current pandemic such a dynamic approach could be crucial to understanding emerging political, legal and social aspects of the crisis.

Notes

1. This is the language used in the Convention at present. The history of the grouping and definition of 'indigenous peoples and local communities' is complex and problematic for various reasons. We use it here, but acknowledge the politics and sensitivities around it.

2. The research cited to illustrate the approach was carried out within the project 'Benefit-sharing for an equitable transition to the green economy (BeneLex)' supported by the European Research Council (grant 335592).
3. On the research illustrated see: Parks (2018, 2020); Parks et al. (2019); Parks and Schröder (2018).
4. TermRaider, information available at https://gate.ac.uk/sale/talks/gate-course-jun15/module-6-applications/termraider.pdf; on GATE see http://www.gate.ac.uk.
5. Namely the Akwé: Kon Voluntary Guidelines (Convention on Biological Diversity, 2004), the Tkarihawié:ri Code of Ethical Conduct (Convention on Biological Diversity, 2010), and the Mo'otz-Kuxtal Voluntary Guidelines (Convention on Biological Diversity, 2016).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

European Research Council [grant 335592], project 'Benefit-sharing for an equitable transition to the green economy' (BeneLex)

Notes on contributors

Louisa Parks is Associate Professor in Political Sociology at the University of Trento's School of International Studies and Department of Sociology and Social Research. Her research has focused on the impacts of social movements on European Union legislation, framing in the environmental movement, global civil society, local community activism and the inclusion of local communities and indigenous peoples in global environmental governance with a focus on the Convention on Biological Diversity.

Wim Peters has been working as a senior researcher and research fellow in the areas of Natural Language Processing (NLP) and Digital Humanities at different universities, amongst which the University of Sheffield (UK), Aberdeen (UK) and Mainz (GER). He specializes in information extraction and other types of text analysis using NLP techniques customized to scholarly requirements from a range of humanities subdisciplines.

ORCID

Louisa Parks http://orcid.org/0000-0002-3921-5397

References

Andrade, S. B., & Andersen, D. (2020). Digital story grammar: A quantitative methodology for narrative analysis. *International Journal of Social Research Methodology*, *23*(4), 405–421. https://doi.org/10.1080/13645579.2020.1723205

Anthony, L. (2016). AntConc Version 3.5.7. (2018) *Computer software*. Retrieved June 2018, from 2016 http://www.laurenceanthony.net/software

Azzopardi, S., Gatt, A., & Pace, G. J. (2016) *Integrating Natural Language and Formal Analysis for Legal Documents* [Paper presentation]. Paper presented at the Conference on Language Technologies & Digital Humanities, Ljubljana: Slovenian Language Technologies Society.

Boréus, K., & Bergström, T. (2017). *Analyzing text and discourse. Eight approaches for the social sciences*. Sage.

Bucchi, M., Loner, E., & Fattorini, E. (2019). Give science and peace a chance: Speeches by Nobel laureates in the sciences, 1901–2018. *PLoS ONE*, *14*(10), e0223505. https://doi.org/10.1371/journal.pone.0223505

Chakrabarti, P., & Frye, M. (2017). A mixed-methods framework for analyzing text data: Integrating computational techniques with qualitative methods in demography. *Demographic Research*, *37*, 1351–1382. https://doi.org/10.4054/DemRes.2017.37.42

Convention on Biological Diversity. (2004) *Akwé: Kon voluntary guidelines*, Retrived July 2018, from https://www.cbd.int/traditional/guidelines.shtml

Convention on Biological Diversity. (2010) *Tkarihawié:riCodeof ethical conduct*. Retrieved July 2018, from https://www.cbd.int/traditional/code.shtml

Convention on Biological Diversity. (2016) *Mo'otz-Kuxtal voluntary guidelines*. Retrieved July 2018, from https://www.cbd.int/decisions/cop/13/18

Crowston, K., Allen, E. E., & Heckman, R. (2012). Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, *15*(6), 523–543. https://doi.org/10.1080/13645579.2011.625764

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002, July 7–12) GATE: An architecture for development of robust HLT applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 168–175). ACL '02. Stroudsburg, PA, USA, Association for Computational Linguistics.

Della Porta, D., & Keating, M. (2008). How many approaches in the social sciences? An epistemological introduction. In D. Della Porta & M. Keating (Eds.), *Approaches and methodologies in the social sciences, a pluralist perspective* (pp. 19–39). Cambridge University Press.

Denzin, N. K. (2012). Triangulation 2.0. *Journal of Mixed Methods Research*, *6*(2), 80–88. https://doi.org/10.1177/1558689812437186

Di Giammaria, L., & Faggiano, M. P. (2017). Big text corpora & mixed methods – The Roman five star movement blog. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *133*(1), 46–64. https://doi.org/10.1177/0759106316681088

Espinoza-Kulick, A. (2020). A multimethod approach to framing disputes: Same-sex marriage on trial in Obergefell V. Hodges. *Mobilization: An International Quarterly*, *25*(1), 45–70. https://doi.org/10.17813/1086-671X-25-1-45

Gerring, J. (2007). *Case study research: Principles and practices*. Cambridge University Press.

Greene, J. C. (2007). *Mixed methods in social inquiry* (Vol. 9). John Wiley & Sons.

Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research*, *2*(1), 7–22. https://doi.org/10.1177/1558689807309969

Hammersley, M. (2008). Troubles with triangulation. In M. M. Bergman (Ed.), *Advances in mixed methods research* (pp. 22–36). Sage.

Hitchcock, T. (2013). Confronting the digital. *Cultural and Social History*, *10*(1), 9–23. https://doi.org/10.2752/147800413X13515292098070

Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: A quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), 469–485. https://doi.org/10.1080/13645579.2019.1576317

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, *1*(2), 112–133. https://doi.org/10.1177/1558689806298224

Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. Verso.

Moses, J. W., & Knutsen, T. L. (2012). *Ways of knowing. Competing methodologies in social and political research*. Palgrave Macmillan.

Parks, L., & Morgera, E. (2019). Research note: Reflections on methods from an interdisciplinary research project in global environmental law. *Transnational Environmental Law*, *8*(3), 1–14. https://doi.org/10.1017/S204710251900027X

Parks, L., Peters, W., & Lennan, M. (2019) Guidelines and codes on the participation of indigenous peoples and local communities of the convention on biological diversity: A comparative analysis using natural language processing (BENELEX Working Paper 23). https://ssrn.com/abstract=3384691 .

Parks, L., & Schröder, M. (2018). What we talk about when we talk about 'local' participation in international biodiversity law. The changing scope of Indigenous peoples and local communities' participation under the Convention on Biological Diversity. *Partecipazione e Conflitto*, *11*(3), 743–785. https://doi.org10.1285/i20356609v11i3p743

Parks, L. (2018). Spaces for local voices? A discourse analysis of the decisions of the Convention on Biological Diversity. *Journal of Human Rights and the Environment*, *9*(2), 141–170. https://doi.org/10.4337/jhre.2018.02.02

Parks, L. (2020). *Benefit-sharing from the bottom up: Local experiences of a global concept*. Routledge.

Peters, W., Parks, L., & Lennan, M. (2019). Integrating language technology into scholarly research workflows. In L. Pitcher & M. Pidd (Eds.), *Proceedings of the Digital Humanities Congress 2018. Studies in the Digital Humanities*. The Digital Humanities Institute. https://www.dhi.ac.uk/openbook/chapter/dhc2018-peters

Peters, W., & Wyner, A. (2016). Legal text interpretation: Identifying Hohfeldian relations from text. In *Proceedings of LREC 2016* (pp. 379–384).

Schmitter, P. (2008). The design of social and political research. In D. Della Porta & M. Keating (Eds.), *Approaches and methodologies in the social sciences, a pluralist perspective* (pp. 263–295). Cambridge University Press.

Smith, B. (2018). Generalizability in qualitative research: Misunderstandings, opportunities and recommendations for the sport and exercise sciences. *Qualitative Research in Sport, Exercise and Health*, *10*(1), 137–149. https://doi.org/10.1080/2159676X.2017.1393221

Woolf, N., & Silver, C. (2018). *Qualitative analysis using ATLAS.ti*. Routledge. https://doi.org/10.4324/9781315181684

Zelik, D., Patterson, E. S., & Woods, D. D.*et al* (2007). Understanding rigor in information analysis. In K. Mosier, and U. Fischer (Ed.), *Proceedings of the Eighth International Naturalistic Decision Making Conference*. Pacific Grove, CA .