



Semantic Analysis to Support Peace Analytics

Roy Gardner



PEACE ANALYTICS SERIES



THE UNIVERSITY
of EDINBURGH



Author: Roy Gardner

PeaceRep: The Peace and Conflict Resolution Evidence Platform
School of Law, Old College, The University of Edinburgh
South Bridge, Edinburgh EH8 9YL

Tel. +44 (0)131 651 4566
Fax. +44 (0)131 650 2005
E-mail: peacerep@ed.ac.uk
PeaceRep.org

✉ [@Peace_Rep_](https://twitter.com/Peace_Rep_)

f <https://www.facebook.com/PeaceRepResearch>

in <https://www.linkedin.com/company/peacerep/>

This research is supported by the Peace and Conflict Resolution Evidence Platform (PeaceRep), funded by the UK Foreign, Commonwealth & Development Office (FCDO) for the benefit of developing countries. The information and views set out in this publication are those of the authors. Nothing herein constitutes the views of FCDO. Any use of this work should acknowledge the authors and the Peace and Conflict Resolution Evidence Platform.

About the author: Roy Gardner is a consultant specialising in data analysis and natural language processing.

Thanks to Allyson Doby, and Rick Smith of Smith Design Agency for production work.

Cover images: Getty Images. All images may be subject to copyright. ©2023
Design: Smith Design Agency

Contents

	Foreword	2
	Definitions	6
1	Introduction	8
	Semantic Projects	8
	Project One: Peace Agreements and Constitutional Change Project	8
	Project Two: Implementation Narratives Project	9
	Project Three: Topic Design Project	10
2	Methodology	11
	Sentence-Level Semantic Similarity	11
	Programming Language	11
	Infrastructure	12
3	Sources of Text for Semantic Analysis	13
	Documents	13
	Topics	13
4	Preparing Text for Semantic Analysis	14
	Extraction	14
	Segmentation	14
	Encoding	15
5	Computing Semantic Similarity	16
	Similarity Matrices	16
6	The Semantic Analysis Toolkit	19
	Content Analysis	19
	Methods	20
	Use Cases	21
	Semantic Mapping	21
	Methods	22
	Use Cases	22
	Clustering	25
	Methods	26
	Use Cases	28
	Semantic Alignment	42
	Methods	42
	Use Cases	43
7	Next Steps	47
	Vocabulary Development	47
	Verification	47
	Automated Threshold Setting	47
	Contextualisation and Styling	48
	Infrastructure	49
	Conclusion	50



Peace Analytics Series

PeaceRep's Peace Analytics Series features the research methodology underlying the PeaceTech innovations of the PeaceRep programme.

The series includes: data scoping research; 'how to' discussions relating to particular challenges in the field of visualisations and geocoding; and other proof-of-concept tech-based innovations, such as the use of natural language processing. It is intended to present the methodologies and decisions behind our PeaceTech digital research, to make it transparent, and to contribute to establishing a new research digital infrastructure in the field of peace and conflict studies, by supporting others to reuse and repurpose our methodologies and findings.



Foreword

By Dr Sanja Badanjak, PeaceRep Data Director

Text is everywhere. For PeaceRep researchers in peace and conflict studies, textual data are often at the core of our work. Our own data production in the [PA-X Peace Agreement Database and Dataset](#) starts with the texts of peace agreement documents which are read, translated, digitised, and transformed into qualitative-textual and quantitative data. In addition to PA-X, we work with many other types of text, all found in great abundance and even greater variety: documents issued by governments, international organisations, NGOs, armed groups; press releases; reports; news items; academic research papers and books. Advances in automatic translation have made even more texts available. Each of these is potentially a source of systematic data for understanding conflict and its prevention and resolution. Yet it could take years for researchers to collect, read, assess, and transform into a data resource. Many of these texts are now available in online repositories, and many have been digitised into machine-readable formats, making them accessible for computer-based text analysis applications.

This report, produced by long-time PeaceRep associate, Dr Roy Gardner, describes the process and demonstrates the potential of computer-based text analysis, and more specifically, models of semantic similarity, for texts generated by peace processes. Such texts include peace agreements and drafts of peace agreements, as well as post-agreement implementation reports.

Whilst necessarily technical, the report is intended to provide a glimpse into the experimentation in which Dr Gardner and the PeaceRep team have been engaged over the past three years. This collaboration has enabled PeaceRep to produce a reliable tool for automatic data coding of texts. This form of analysis has enabled us to quickly match agreement provisions with pertinent text from implementation reports. This tool will streamline the process through which future PA-X datasets are produced, enabling researchers to focus on analysis, and providing long term sustainability to the database itself. However, the tool also has other research uses, as illustrated herein.

The promise of computer-based text analysis and various text-as-data approaches lies in the technology's ability to ingest, organise, and analyse large collections of text, all with focused researcher engagement and with exceptional speed. Rather than reading and organising each text, the researcher can set parameters for the selection of text, enabling, at the very least, automation of the initial separation of relevant text from less relevant text.

Language models designed for semantic similarity tasks, such as those described in this report, go even further. Our method employs the Universal Sentence Encoder (Cer et al., 2018), to identify text similar in meaning to a user's search requirement. This mode of searching for relevant text is more efficient than simple dictionary searches based on word combinations and synonyms.

Unless one is interested in specific terminology, a dictionary search will yield fewer relevant results than a search based on the semantic similarity between expressions and sentences. For example, a semantic search of the PA-X Peace and Transition Agreement Database for "The welfare of pregnant women and babies" not only finds sentences such as "The State shall provide maternity and child care and medical care for pregnant women", but also "Care will be provided to expectant and breastfeeding mothers and, in general, maternal-infant care".

Further problems with dictionary searches arise when the terms being searched for are burdened with additional meaning and expectations. In armed conflict, signing a "ceasefire" may carry connotations of longer-term commitment than signing an agreement that pauses hostilities. A semantic search can be written in such a way as to clarify the search and reduce possible bias. For example, searching the PA-X database for "Ceasefires and any other modes of pausing or cessation of hostilities" not only finds "The two sides commit to a ceasefire and to join the state of cessation of hostilities", but also finds "Extension of the Humanitarian Pause". The UN definition of a humanitarian pause is a *temporary cessation of hostilities purely for humanitarian purposes*.



For researchers, automatic text analysis opens a world of opportunity to systematically explore texts that would otherwise require exceptional effort and take a considerable amount of time and be subject to human error. This is not to say that this type of research is without cost; some investment is still needed. The first and key prerequisite of this research, and a major area of investment, is the digitisation of texts and their conversion into machine-readable materials. It is not enough for a document to be scanned and digitised, it must also be subjected to some form of optical character recognition or transcription for the text to be readable by computers. Over the years, PeaceRep researchers have found that comprehensive, cohesive, and machine-readable collections of text are increasingly available. One of the aims of our own work on the PA-X Peace Agreement Database and Dataset has been to create such a collection of texts that can be used by researchers looking to explore peace agreement texts. Similar collections of court judgments are now provided by other researchers (Fobbe, 2022; Frankenreiter and Livermore, 2020) and by a variety of organisations seeking to make records such as UN Security Council debates, more readily available (Schönfeld et al., 2019; Baturo et al., 2017; Ziemski et al., 2016). Provision of these data resources in the long term will require significant levels of buy-in and support from partners who hold data, be they large organisations such as the United Nations and its agencies, or smaller organisations such as NGOs. Seeking cooperation with such organisations, and planning for research in support of their mission, is a potential route for expanding the range of available textual data resources.

The second area where investment is needed in the research community is in computing skills, particularly skills in statistical computing and programming, using tools such as Python and R. Additionally, textual data and their processing are often beyond the abilities of personal computing machines, and require access to increased processing power, spacious servers, virtual machines, and similar. Such processing is typically provided by research universities or large technology companies but is often unavailable to researchers outside such organisations. These are significant barriers to entry into this form of research, and we hope that reports such as this one, will demystify the technology and encourage the uptake of methodological training in the field of computer-based text analysis among scholars of peace and conflict processes.

This report is part of the PeaceRep series on PeaceTech, aiming to provide practical introductions to a variety of computing methodologies that we have relied on during the research programme. Alongside this report on semantic similarity models, the series includes reports on methodologies of geocoding (Farquhar, 2023) and named entity recognition (Henry, 2023), among others.

References

- Baturo, Alexander, Niheer Dasandi, and Slava J. Mikhaylov. 2017. "Understanding State Preferences with Text as Data: Introducing the UN General Debate Corpus." *Research & Politics* 4(2): 2053168017712821.
- Cer, Daniel et al. 2018. "Universal Sentence Encoder for English." *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*: 169–74.
- Farquhar, A. (2023). A Primer on Geocoding for Peace and Conflict Processes [Peace Analytics Series]. PeaceRep: The Peace and Conflict Resolution Evidence Platform, University of Edinburgh. <https://peacerep.org/publication/geocoding-primer>
- Fobbe, Seán. 2022. "Introducing Twin Corpora of Decisions for the International Court of Justice (ICJ) and the Permanent Court of International Justice (PCIJ)." *Journal of Empirical Legal Studies*.
- Frankenreiter, Jens, and Michael A Livermore. 2020. "Computational Methods in Legal Analysis." *Annual Review of Law and Social Science* 16(1): 39–57.
- Henry, N. (2023). Extracting Named Actors from Text: Using Named Entity Recognition in Peace and Conflict Studies [Peace Analytics Series]. PeaceRep: The Peace and Conflict Resolution Evidence Platform, University of Edinburgh. <https://peacerep.org/publication/extracting-named-actors-from-text>
- Schönfeld, Mirco, Steffen Eckhard, Ronny Patz, and Hilde van Meegdenburg. 2019. "The UN Security Council Debates 1995-2017." *arXiv*.
- Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. "The United Nations Parallel Corpus v1.0." *Language Resources and Evaluation* (LREC'16).



Definitions

Term	Definition
Segmentation	The process of segmenting text into sentences or sentence-level segments. Sentence boundaries can be customised depending upon the text source and the degree of granularity required.
Text segment/ segment	A single sentence-level text segment generated by the segmentation process.
Provision	A text segment derived from the segmentation of a peace agreement.
Section	A text segment derived from the segmentation of a constitution or implementation report.
Topic	A natural language expression of one or more concepts curated by a human being. Topics are used to search for and classify sentences.
Encoding	The process of generating the encoding vector of a text segment or topic.
Encoding vector	A 512-length numerical representation of a text segment or topic.
Similarity score	A measure of semantic similarity computed as the inverse of the angular distance between two encoding vectors. Scores are in the range 0.0-1.0 where 1.0 means two sentences are identical in meaning and comprise the same words in the same order. As the meanings of the sentences diverge, the similarity score decreases.



Term	Definition
Similarity matrix	A matrix containing rows that correspond to topics or text segments, and columns that correspond only to text segments. A cell value contains the similarity score for a topic-segment or segment-segment pair.
Threshold	A similarity score value below which topic-segment or segment-segment pairs are discarded.
Linked pair	A topic-segment or segment-segment pair that have a similarity score at or above a given threshold.
Semantic cluster	A disjoint set of text segments that are similar in meaning.
Vocabulary	A set of topics designed to capture the conceptual scope of a corpus. Topics may be organised into categories.
Sentiment	Ranks the emotional tone of a text segment in the range -1.0 to +1.0 where negative values represent negative sentiment.

1 Introduction

Peace processes are often defined and understood through the documents they produce, starting with peace agreements, but also including constitutions, implementing legislation, court decisions, and reports on implementation. Understanding peace processes as they unfold over multiple stages and arenas requires not just grappling with complex contexts, but also grappling with multiple forms of text, which is further complicated when attempting analysis over multiple countries and over long time periods.

In this report we describe the application of sentence-level semantic similarity technology for analysis of peace process documentation. Our toolkit automates resource- and time-intensive analyses of large datasets and document corpora.

The description of the toolkit includes use cases taken from three current semantic projects described in the next section.

Semantic Projects

Three projects were used to develop this work, and illustrate its potential, and are traced through this report.

Project One: Peace Agreements and Constitutional Change Project

The purpose of this project is to understand when and how issues enter and leave peace negotiations, and which elements of peace agreements find their way into permanent constitutional arrangements. This matters because peace agreements are often agreed primarily between armed actors, while constitutions are understood as 'we the people' documents. Understanding the relationship between the two is therefore both interesting and important, as it points to the nature of the political settlement.

This project therefore assesses the influence of peace agreements on constitutions and vice versa, with a view to predicting which elements of peace agreements persist over time. Predictions of influence are divided into three partitions:

- Agreement provisions related to sections of a subsequent constitution.
- Agreement provisions related to both the sections of a previous constitution and a subsequent constitution.
- Agreement provisions related to the sections of a previous constitution only.

In addition, the project identifies substantive agreement provisions that are unrelated to constitution sections, which provides a way of understanding the type of provisions that do not make their way into constitutions.

The project uses a dataset that comprises 1447 agreements and 79 constitutions from 65 countries. These data comprise PA-X Peace and Transition Agreements, from processes that reached at least a partial framework agreement for resolution, for which there were subsequent constitutions or amendments.

Project Two: Implementation Narratives Project

Peace agreements often fail, with conflict recurring, because they are not implemented. Also, the international actors supporting implementation need ways of assessing it. This project therefore assesses progress in the implementation of a peace process by measuring the semantic relationships between peace agreement provisions and the content and sentiment of implementation report sections.

At the time of writing, the project analyses peace processes in South Sudan and pre-invasion Ukraine. These countries chosen had readily available, accurate implementation reports tracing agreement implementation. In the case of South Sudan, these include reports of the [Joint Monitoring and Evaluation Mission](#) (JMEC and later 'Reconstituted' as RJMEC) which monitored the implementation of the [Revitalised Agreement on the Resolution of the Conflict in the Republic of South Sudan \(R-ARCSS\)](#), 2018. In the case of Ukraine, the [OSCE Special Monitoring Mission](#) reports on the implementation of the Minsk Agreements. In both cases, these reports had optical character recognition and could therefore be processed and machine-read as text.

Several analysis techniques are being explored, including:

- Using topics to find semantically similar agreement provisions, which in turn are used to find semantically similar report sections.
- Using manually coded provisions to find semantically similar report sections.
- Using topics to find semantically similar report sections.
- Using provisions to find semantically similar report sections without pre-selection of provisions by topics.

Project Three: Topic Design Project

The [PA-X Peace Agreements Database](#) has extracted agreement text on over 250 topics, based on what were viewed as critical issues that agreements dealt with, supplemented by inductive addition from reading peace agreement texts. This project was motivated by the recognition that codebooks often express topics in ways that rely on coders' implicit knowledge. Such topics are not best suited to the semantic applications we describe in this report. For example, codebook entries may contain instructions for coders, or descriptions of the intent of a topic without expressing the topic in natural language. Furthermore, topic descriptions may be insufficiently disambiguated. Topic design for this project was informed by three sources of data:

- [Version 5 of the PA-X codebook](#).
- [The PA-X Gender codebook](#).
- Peace agreement provisions that have been manually coded by the PA-X team.

The objective of the project is to rewrite PA-X codebook topics to create a vocabulary of topics organised under PA-X codebook categories and codes that are suitable for semantic-similarity applications. There may be more than one topic corresponding to a codebook code.

2 Methodology

Sentence-Level Semantic Similarity

All of the applications depend on identifying sentence-level semantic similarity. Sentence-level semantic similarity measures the degree to which two or more natural language sentences or clauses convey similar meaning. We use version 4 of Google's [Universal Sentence Encoder](#) (USE) (Cer, D. et al., 2018) to generate 512-length numerical representations of sentences referred to as encoding vectors or embeddings. An encoding vector defines the coordinates of a sentence in a 512-dimensional space, within which sentences with similar meaning will be found close together and sentences with dissimilar meaning further apart.

For USE models, the preferred measurement of the distance between a pair of encoding vectors is angular distance. The inverse of this distance is a measure of the semantic similarity between the sentences that the vectors represent. This measurement is referred to as a semantic similarity score; this score ranges from 1.0 to 0.0. A similarity score of 1.0 means two sentences are identical in meaning and comprise the same words in the same order. As the meaning of the sentences diverges, the similarity score decreases.

One of the key features of the USE model is that it enables efficient and accurate computation of sentence-level encoding vectors, making it possible to perform large-scale semantic similarity tasks on a range of datasets without any pre-processing of text other than segmentation and encoding. Overall, Universal Sentence Encoders are powerful tools for natural language processing with a wide range of applications, including multilingual semantic retrieval (Yang, Y. et al., 2020), and the analysis of television news captions.¹

Programming Language

The processes and toolkit described in the report are implemented in version 3 of the Python programming language. Text processing pipelines and similarity matrix generation are implemented in .py files that generate Python data objects. These data objects are stored on disk as JSON files.

Analysis of data objects read from disk is implemented in Jupyter Notebooks, which produce human-readable outputs, visualisations, and data used by other PeaceRep software.



Infrastructure

The codebase runs on the MacOS Unix operating system. The migration of the codebase to infrastructure provided by the Edinburgh Parallel Computing Centre ([EPCC](#)) is underway.



3 Sources of Text for Semantic Analysis

Documents

Currently, there are three sources of text and associated metadata across the three topics:

1. Peace and transition agreement text contained as plain text in CSV files exported from the [PA-X database](#) (Projects One, Two and Three).
2. Peace process implementation reports in PDF files. At the time of writing, sources of reports are the [Reconstituted Joint Monitoring and Evaluation Commission](#) (RJMEC), and the now closed [OCSE Special Monitoring Mission to Ukraine](#) (Project Two).
3. Pre-segmented constitution text in XLSX files provided by the Comparative Constitutions Project (CCP). There is one XLSX document per constitution (Project One).

Document data and metadata are stored in Python dictionaries using a document ID as the key. Dictionaries are stored on disk as JSON files. Documents from the same corpus are stored in the same dictionary.

Topics

Topics are short phrases that capture one or more related concepts. Topics are encoded in their entirety, and multi-sentence topics are not segmented. Topic metadata comprise PA-X codebook codes and categories. There may be several topics corresponding to a single PA-X code. The set of topics constitutes a vocabulary.

Topic text and metadata are stored in Python dictionaries using a topic ID as the key. Dictionaries are stored on disk as JSON files. Topic data are stored in a single dictionary.



4 Preparing Text for Semantic Analysis

Preparing text for semantic analysis is a three-stage process.

1. Extraction

Text in PDF files is extracted as plain text using the Python [textract](#) package.

2. Segmentation

Text from all sources, except pre-segmented CCP constitutions, is segmented into sentence-level segments using the [spaCy](#) version 3.4 sentencizer pipeline with the [spaCy English large language model](#).²

Sentence segmentation boundaries are the default punctuation characters defined by the [spaCy sentencizer](#) with the addition of custom punctuation characters. The use of custom boundary characters depends on the source of the text being processed, and the degree of granularity required. For example, it may be necessary to include colons, semicolons, and bullet points as boundary characters.

A segment entity is created for each text segment generated by the segmentation process. Every segment entity has the following properties:

- The text of the segment itself. This is referred to as the raw text.
- A sanitised version of the text ready for encoding. Sanitisation may create empty text if a segment contains only combinations of numeric, punctuation, or whitespace characters.
- A sentiment score for non-empty sanitised text obtained from [VADER sentiment analysis](#). Sentiment scores are in the range -1.0 to 1.0 representing negative to positive sentiment values. Sentiment is used to determine whether the tone of a text segment is positive or negative.
- An identifier that combines the text segment's source document ID with a numerical value that defines the text segment's ordinal position in the source document.

The entire text of a peace agreement or implementation report can be recreated by listing the raw text in identifier order.



Text segment entities are stored in Python dictionaries using the segment identifier as the key. Dictionaries are stored on disk as JSON files. Segments from the same corpus are stored in the same dictionary.

3. Encoding

Encoding vectors are obtained from the USE model for all qualifying text segment entities. To qualify for encoding, a segment's sanitised text must exceed a minimum word count. The minimum word count is in the range four to six and is corpus dependent.

The identifiers of qualifying segment entities and their encoding vectors are stored in Python lists which are stored on disk as JSON files. Identifiers or encoding vectors from the same corpus are stored in single lists.



5 Computing Semantic Similarity

Given the encoding vectors \mathbf{v} and \mathbf{w} of two segments s_1 and s_2 , the semantic similarity score sim of the segment-segment pair is given by:

$$sim(s_1, s_2) = 1 - \arccos\left(\frac{C(\mathbf{v}, \mathbf{w})}{\pi}\right)$$

Similarly, given the encoding vectors \mathbf{v} and \mathbf{w} of topic t and segment s , the semantic similarity score S of the topic-segment pair is given by:

$$sim(t, s) = 1 - \arccos\left(\frac{C(\mathbf{v}, \mathbf{w})}{\pi}\right)$$

where C is the cosine similarity of the encoding vectors:

$$C(\mathbf{v}, \mathbf{w}) = \frac{\sum_{i=1}^N v_i w_i}{\sqrt{\sum_{i=1}^N v_i^2} \sqrt{\sum_{i=1}^N w_i^2}}$$

Similarity Matrices

A similarity matrix S is a pre-compiled matrix of semantic similarity scores. A similarity matrix contains m rows and n columns where each row represents a topic or sentence-level segment, and each column represents a sentence-level segment. A matrix cell contains the similarity score for a topic-segment or segment-segment pair s_{ij} , i.e., the similarity score between the i^{th} row topic (or segment) and j^{th} column segment.

$$\mathbf{S} = \begin{bmatrix} S_{11} & \cdots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mn} \end{bmatrix}$$

Similarity matrices are key data structures that are usually generated during text processing once encoding vectors have been obtained. The number of segments in a document set and the size of our topic vocabulary means that similarity matrices containing many millions of cells must be pre-compiled. Pre-compiled similarity matrices are converted to Python lists and are stored on disk as JSON files.

A visualisation of a similarity matrix generated from the encoding vectors of four sentences is shown in Figure 1 below. Corresponding cell values are shown in Table 1.

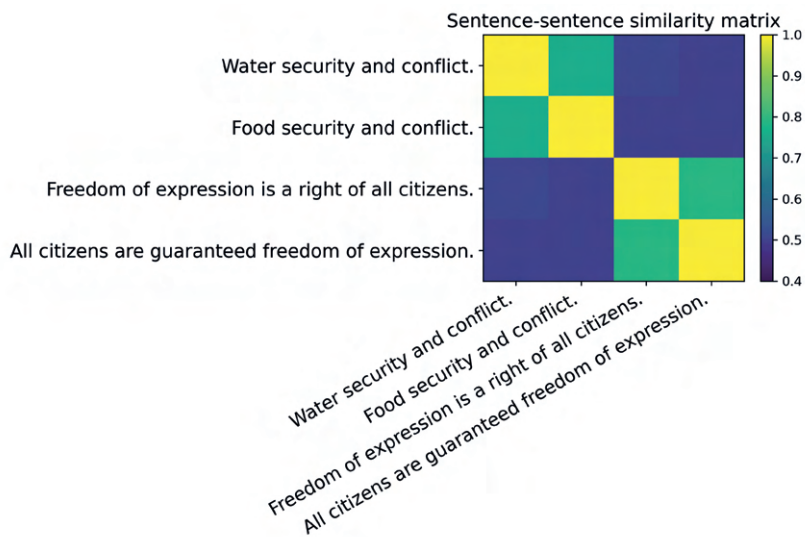


Figure 1: Heat map representing the semantic similarities between four sentences. The map is symmetric around the diagonal.



1.00	0.76	0.51	0.50
0.76	1.00	0.50	0.50
0.51	0.50	1.00	0.79
0.50	0.50	0.79	1.00

Table 1: Semantic similarity scores for the cells of the heat map in Figure 1.



6 The Semantic Analysis Toolkit

The Semantic Analysis Toolkit comprises four components that utilise either on-demand semantic similarity calculations or pre-compiled similarity matrices to implement various applications. Use cases taken from the three semantic projects listed in the Introduction are presented. Table 2 shows the relationship between the semantic projects and the components of the toolkit.

	Content Analysis	Semantic Mapping	Clustering	Semantic Alignment
Topic Design	YES	NO	YES	YES
Peace Agreements and Constitutional Change	NO	YES	YES	NO
Implementation Narratives	YES	YES	YES	YES

Table 1: Matrix of Semantic Projects and Toolkit components.

Content Analysis

Content analysis connects topics from a controlled vocabulary to parts of a document's text. The role of the content analyst is to find the textual evidence for a topic by matching the meanings of each topic with text segments. This is a time-consuming, labour-intensive activity. In addition, subjective judgments may create inconsistencies in the application of codes, and it is difficult to determine the completeness and error rates of a coded corpus.

The approach described here automates the process of matching the meaning of topics and sentences by connecting every topic-segment pair in a corpus and assigning a similarity score to each connection.

The task of a content analyst is no longer to find and create connections, but to assess the textual evidence provided by the similarity between topics and text segments.

Applications include:

- Topic search. Find text segments that are semantically similar to a topic at or above a similarity threshold.
- Machine-assisted classification. On the basis of the semantic similarity between segments and topics, automatically label text segments with one or more topics from a vocabulary.

Methods

There are two modes of topic search:

1. In on-demand topic search a topic is created and encoded. The semantic similarities between the topic and the text segments of a corpus are then calculated as semantic similarity scores.
2. In matrix-based topic search, a topic's row is extracted from a precompiled topic-segment similarity matrix. The row contains the similarity scores of all topic-segment pairs.

In both modes, text segments are sorted in descending order of similarity to the topic and are presented to the user or consumed by another process. A similarity threshold is normally applied to control the size of the result set.

Similarly, there are two modes of machine-assisted classification:

1. In on-demand classification, text is segmented and encoded. Topic search is then applied to the encoded segments for every topic in the vocabulary.
2. Matrix-based classification utilises matrix-based topic search over all the topics in the vocabulary.

In both modes, if the similarity between a segment and a topic is equal to or exceeds a threshold, then the segment is labelled with the topic. Labelled segments are presented to the user.



Use Cases

Project Three: Topic Design Project

As this project aimed in part to produce forms of automated coding for the PA-X Peace Agreement database, the principal criterion for accepting a topic is how well the topic performs in terms of finding semantically similar provisions in the PA-X peace agreement corpus. Found provisions are also compared to provisions that have been manually coded using codes that are closest in meaning to a topic.

The topic design process is highly iterative and involves reformulating a topic to improve search results. A recent advance in the assessment of topic search results is the use of clustering (see Table 6 below).

Project Two: Implementation Narratives Project

Topic search is used in the Implementation Narratives Project to find semantically similar provisions in peace agreements. These provisions are then used to find similar sections in implementation reports, and thereby provide forms of extracted implementation data of a qualitative nature for how particular provisions of agreements are being implemented.

Semantic Mapping

Semantic mapping looks for semantically similar or identical text segments in sets of documents using segment-segment matrices.

Applications include:

- Mapping the text segments of a set of documents onto the segments of a reference document (as per Projects Two and Three).
- Tracking the meaning contained in a segment across sets of documents (as per the 'new topic component of Project Three).

In both applications the document sets are usually organised as a time series.

Methods

When mapping onto a reference document, the objective is to find the text segments in the reference document that are semantically similar to segments in a set of comparison documents. This involves setting a similarity threshold and finding linked pairs at or above threshold that contain a segment from the reference document. The reference document segments in the linked pairs are the found set.

A more general case is also possible, where no reference document exists, and the objective is to build bipartite graphs that link segments in pairs of documents. This application is not discussed in the report because it has yet to be implemented in a PeaceRep project.

Mapping requires one or more pre-compiled similarity matrices. When mapping onto a reference document the segments of the reference document are in rows and the segments of a comparison documents in columns. There is usually one matrix for each comparison document.

Use Cases

Project One: Peace Agreements and Constitutional Change Project

Table 3 below lists six peace agreements in date order from the Framework/substantive stage of the Bosnia peace process. The number of sections in the constitution that are similar to the provisions of an agreement is given in the first column. Sections are not counted twice across agreements.

Our analysis is based on sequences of time-ordered peace agreement and constitutions, where a constitution is preceded by at least one agreement, and where some agreements sit between constitutions. A pre-compiled section-provision similarity matrix exists for each agreement where the constitution sections are in rows and the agreement provisions are in columns. The constitution is therefore the reference document.



Number of sections	Agreement date	Agreement ID	Agreement name
4	19920506	1472	The Public Announcement (Graz Agreement)
7	19930623	1177	Croat-Serb Constitutional Principles for Bosnia-Herzegovina
137	19930916	472	Agreement relating to Bosnia and Herzegovina (Owen-Stoltenberg Peace Plan, or 'Invincible plan')
53	19940301	608	Framework Agreement for the Federation (Washington Agreement or Contact Group Plan)
223	19940318	1198	Declaration Concerning the Constitution of the Federation of Bosnia and Herzegovina (with Proposed Constitution of the Federation of Bosnia and Herzegovina attached)
930	19951121	389	General Framework Agreement for Peace in Bosnia and Herzegovina (Dayton Peace Agreement)

Table 2: Number of sections in the 1995 constitution of Bosnia & Herzegovina related to a sample of six peace agreements that pre-date the constitution.

Project Two: Implementation Narratives Project

Figure 2 below shows a timeline marking the dates of ceasefire agreements and OSCE implementation reports from the Ukraine peace process.

The data shown are generated using a two-stage process:

1. The topic search for agreement provisions similar to the topic *Restrictions of deployment of artillery and heavy weapons including distances from civilian populations and infrastructure* is carried out.
2. Semantic mapping is then used to map provisions onto implementation report sections.

Pre-compiled provision-section similarity matrices exist for each agreement-report combination as do topic-provision and topic-section matrices.

Blue lines are date markers for agreements containing provisions found by topic search. Implementation report dates are marked as red or green lines. The colour indicates the polarity of the sentiment, and the line length reflects the mean sentiment in the sections of the report. Sentiment is overwhelmingly negative with respect to the implementation of the restrictions specified in the topic.

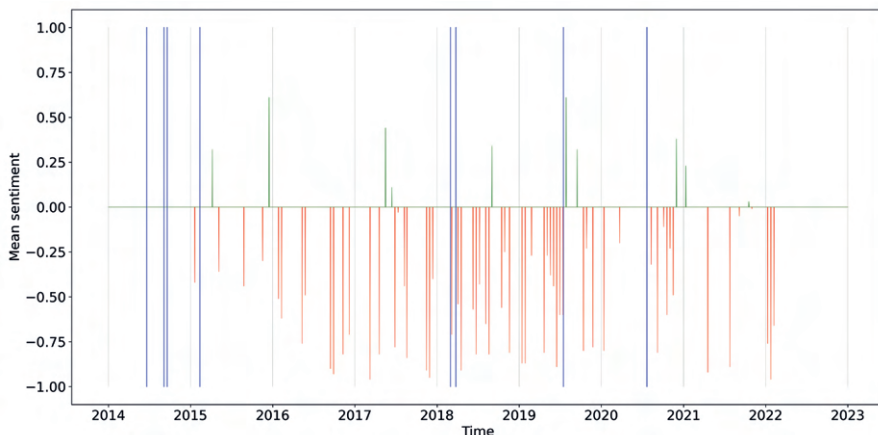


Figure 2: Timeline showing the mean sentiment of implementation report sections for the topic 'Restrictions of deployment of artillery and heavy weapons including distances from civilian populations and infrastructure'. The blue lines are date markers for ceasefire agreements. Green lines represent reports containing sections with a mean positive sentiment. Red lines represent reports containing sections with a mean negative sentiment.

Clustering

Clustering organises text segments into groups that share a common meaning.

The applications of clustering are:

- Topic discovery. The common meaning of the text segments in a cluster may provide evidence for new topics or sub-topics. This is especially true if the clustered text segments are unclassified by a vocabulary or are unrelated to text segments from a reference set (such as that of the manual coders of PA-X).
- Refining topic search results. Selecting the appropriate threshold for topic search requires experimentation. If the threshold is high, then meaningful results may be missed. If the threshold is low, then the result set will be large and could contain off-topic results. However, lower thresholds can be used if the topic search results are clustered because clustering separates on-topic and off-topic results.

Methods

We implement a connection-based clustering method that finds matrix cells containing similarity scores at or above a chosen cluster threshold. Each cell with an above-threshold score defines a linked pair of segments.

The text segments of the pairs form the set of vertices V of a graph G . Each pair of vertices (x,y) is connected by an edge $e = \{x,y\}$ to create an undirected simple graph.

Depending upon the choice of cluster threshold and the semantics of the text segments, the graph will comprise a set of components. Each component is a disjoint subgraph unconnected to other components. These components contain the clusters of semantically similar text segments.

The figure below illustrates clustering in a graph created by the linked pairs within a set of topic search results (see Table 6). Each numbered vertex represents a provision, and the edges represent semantic similarity relationships at or above a cluster threshold of 0.68.

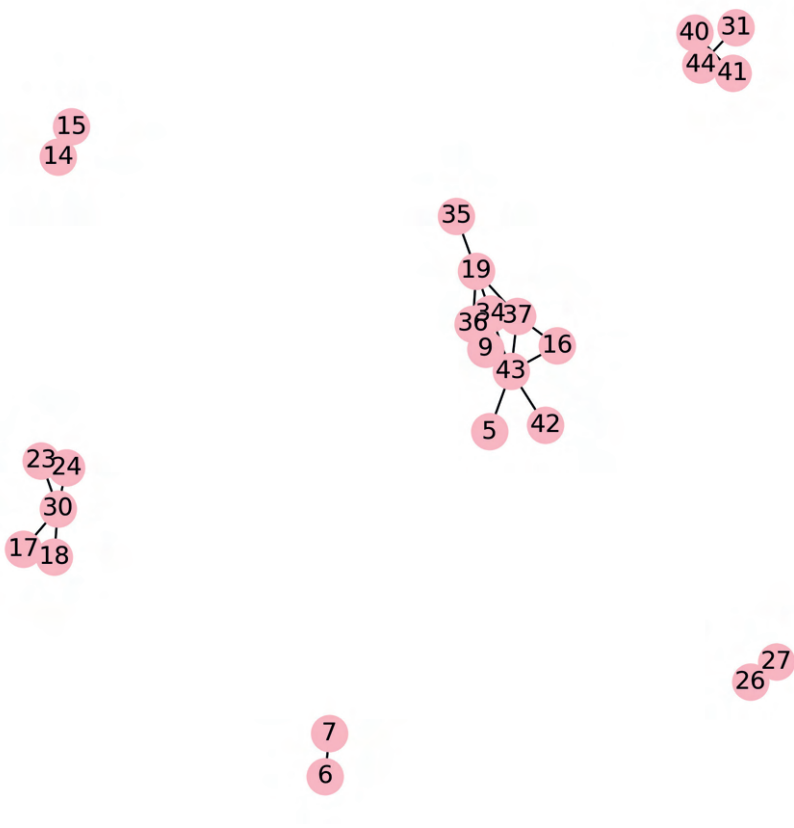


Figure 3: A graph (network) showing clusters of peace agreement provisions generated by topic search for the topic 'The use of child soldiers.' See also Table 6. The provisions are represented by numbered nodes.



Text segments that do not have an above-threshold relationship to any other segment are not contained in the graph and form the set of unconnected singleton segments referred to as the unclustered set.

Use Cases

Project One: Peace Agreements and Constitutional Change Project

One objective of this project is to find clusters of semantically similar agreement provisions that are unrelated to constitution sections.

Table 4 below shows an example of a cluster from the Darfur-Sudan peace process between 2005 and 2019. Only provisions that were not semantically similar to sections of the Sudan 2005 or 2019 constitutions were selected for clustering.

The cluster shown in the table below is one of many that illustrate the (often verbatim) similarity of the Darfur Peace Agreement of 2006 and the Doha Document for Peace in Darfur (DDPD) of 2011. Such clusters may provide evidence for substantive concepts in agreements that are not reflected in constitutions for Project One.

Agreement Name	Agreement date	Sentence number	Sentence text
Darfur Peace Agreement	20060505	249	Without prejudice to the provisions of the CPA relating to the North-South border and any international agreements in force between the Republic of the Sudan and neighbouring countries, the northern boundaries of Darfur shall return to the positions as at 1 January

Agreement Name	Agreement date	Sentence number	Sentence text
Darfur Peace Agreement	20060505	79	Without prejudice to the provisions of the CPA relating to the North-South border and any international Agreements in force between the Republic of the Sudan and neighbouring countries, the northern boundaries of Darfur shall return to the positions as of 1 January
Doha Document for Peace in Darfur (DDPD)	20110531	101	Without prejudice to the provisions of the Comprehensive Peace Agreement (CPA) relating to the North-South border and any international agreements in force between the Republic of Sudan and neighbouring countries, the northern boundaries of Darfur shall return to their positions of 1 January
Doha Document for Peace in Darfur (DDPD)	20110531	283	THE NORTHERN BORDERS OF DARFUR Without prejudice to the provisions of the Comprehensive Peace Agreement (CPA) relating to the North-South border and any international agreements in force between the Republic of Sudan and neighbouring countries, the northern boundaries of Darfur shall return to their positions as of 1 January

Table 3: A cluster of provisions in agreements from the Darfur-Sudan peace process. None of the provisions are semantically similar to sections of the 2019 constitution of Sudan.

Project Two: Implementation Narratives Project

At the time of writing, the current version of the Implementation Narratives Project tracks concepts from peace agreements to implementation reports. Topic search is used to locate provisions which are then used to find similar sections.

Clustering is used to identify topics of interest in implementation reports that are outside of the scope of the found provisions.

The table below shows two clusters inside the sections of RJMEC implementation reports on the implementation of the [2019 R-ARCSS agreement](#) from the South Sudan post-secession process. The sections in cluster 0 span three reports, whereas the sections of cluster 1 are within a single report.

Cluster ID	Provision text	Report title
0	Some Parties have released child soldiers but certainly not all and attempts by CTSAMVM to verify the forces have been impeded by some Parties.	JMEC-1st-Qtr-2019-Report
0	investigations Some Parties have released child soldiers but certainly not all and attempts by CTSAMVM to verify the forces have been impeded by some Parties.	RJMEC-2nd-Qtr-2019-Report
0	and Some Parties have released child soldiers but certainly not all and attempts by CTSAMVM to verify the forces have been impeded by some Parties.	RJMEC Quarterly Report (1 Jul - 30 Sept 2019) - 3rd Q



Cluster ID	Provision text	Report title
1	In Adviser positions, women constitute 20% in Central Equatoria, Eastern Equatoria, Northern Bahr el Ghazal, Upper Nile, Warrap and Western Bahr el Ghazal States, 40 % in Lakes State and 0% in Jonglei, Unity and Western Equatoria States.	RJMEC-1st-Qtr-2021-Report-FINAL
1	In Ministerial positions, women constitute 29.4 % in Central Equatoria State, 23.5 % in Eastern Equatoria, Unity and Western Equatoria States, 11.8 % in Jonglei, Upper Nile and Warrap States and 17.6 % in Lakes, Northern Bahr el Ghazal and Western Bahr el Ghazal States. 14 Women constitute 33.3 % of Chairpersons of Independent Commissions in Eastern Equatoria, Lakes and Northern Bahr El Ghazal States, 16.7 % in Western Bahr El Ghazal and Western Equatoria States and there are no female Chairpersons in Central Equatoria, Jonglei, Unity, Upper Nile and Warrap States.	RJMEC-1st-Qtr-2021-Report-FINAL
1	In Deputy Chairperson of Independent Commissions positions, women constitute 40 % in Central Equatoria State, 16.7 % in Eastern Equatoria, Jonglei, Northern Bahr el Ghazal, Upper Nile, Warrap and Western Bahr el Ghazal States and there are no women Deputy Chairpersons of Independent Commissions in Lakes, Unity and Western Equatoria States.	RJMEC-1st-Qtr-2021-Report-FINAL

Table 4: Two clusters from RJMEC reports on the implementation of the 2019 R-ARCSS agreement from the South Sudan post-secession process

However, the presence of these clusters reflects the restricted choice of topics used in the proof-of-concept for this work, for South Sudan. It is highly likely that these sections could be found either directly by topic search or by direct semantic mapping of agreement provisions onto report sections. A better assessment of report clusters would cluster only those sections that were not similar to any agreement provisions (cf. the Peace Agreements and Constitutional Change use case above).

Topic Design Project

Table 6 below shows clusters of peace agreement provisions found by the topic *The use of child soldiers* at a search threshold of 0.62. Altogether there are 46 provisions which contain a mixture of on-topic and off-topic results. A cluster threshold of 0.68 is applied to the results. Provisions that have a good fit to the topic are highlighted in green.

The choice of a relatively low search threshold (0.62) reduces the likelihood of missing relevant provisions but increases the presence of noise in the search results. Clustering reduces the noise of a low threshold search by segregating relevant text segments from less relevant segments and thereby makes the identification and selection of relevant sections easier.

To create clusters, the encoding vectors of the 46 found provisions are retrieved from stored data and used to construct a square similarity matrix with found provisions in both rows and columns. In other words, we are computing the semantic similarity between all pairwise permutations of the found provisions. The matrix is symmetric around the diagonal which means we only need to consider the upper triangle, excluding the diagonal.

The procedure described in the Methods section above is applied to the search result similarity matrix using a cluster threshold of 0.68. Clusters 0, 1, 2, 4, and 6 only contain provisions related to child soldiers, and half of the unclustered provisions are also related to child soldiers. Clusters 3 and 5 and the other half of the unclustered provisions are off topic.

Cluster ID	Segment text	Document name
0	recruitment and/or use of child soldiers by armed forces or militias in contravention of international conventions;	Revitalised Agreement on the Resolution of the Conflict in the Republic of South Sudan (R-ARCSS)
0	recruitment and enlistment of children;	Agreement on the Cessation of Hostilities, Protection of Civilians and Humanitarian Access, Republic of South Sudan
0	Recruitment and/or use of child soldiers by armed forces or militias in contravention of international conventions;	Agreement on the Resolution of the Conflict in the Republic of South Sudan (ARCSS)
0	Underline that both parties are committed by national and international law to prevent the recruitment and mobilization of child soldiers, and that the recruitment and use of child soldiers constitutes grave violations of the laws of war;	Re-dedication of and Implementation Modalities for the Cessation of Hostilities Agreement signed on the 23rd January 2014 between the Government of the Republic of South Sudan and the Sudan People's Liberation Army / Movement (in opposition)
0	c) Any act that would violate the rights of children, including the recruitment and use of children under 18 years of age in any direct or indirect capacity within an armed unit.	Political Agreement for Peace and Reconciliation in the Central African Republic (Khartoum Accord)

Cluster ID	Segment text	Document name
0	the recruitment and deployment of child soldiers, mercenaries or persons without Ivoirian nationality, outside the framework of the agreements regularly agreed by the Ivoirian State;	Accord de Cessez-le-Feu
0	The Parties shall refrain from recruiting children as soldiers or combatants, consistent with the African Charter on the Rights and Welfare of Children, the Convention on the Right of the Child (CRC) and the Optional Protocol to the CRC on the Involvement of Children in Armed Conflict.	Protocol between the Government of the Sudan (GoS), The Sudan Liberation movement/Army (SLM/A) and the Justice and Equality Movement (JEM) on the Enhancement of the Security Situation in Darfur in Accordance with the N'Djamena Agreement
0	The Parties recognize that the recruitment and use of children by armed forces and armed groups is a violation of children's rights.	Agreement on Disarmament, Demobilization and Reintegration, Juba, Sudan
0	COGNISANT of the fact that conscription of children into the army and their involvement in war is a serious violation of the Rights of the Child;	Intercongolèse Negotiations: The Final Act ('The Sun City Agreement')
0	recruitment and use of child soldiers;	Ceasefire Agreement (Lusaka Agreement)

Cluster ID	Segment text	Document name
1	Respect humans rights, particularly those of women and children, and to abstain from acts of sexual violence towards women, and from recruiting children as soldiers;	Déclaration de principe des parties aux négociations de Libreville sur la crise Centrafricaine
1	Protection of human rights, including the release of all detained persons, cessation of sexual violence and the conscription of child soldiers;	Accord de cessez-le-feu entre le Gouvernement de la République Centrafricaine et la Coalition Seleka
2	demobilisation and reintegration of child soldiers and vulnerable persons;	Intercongolense Negotiations: The Final Act ('The Sun City Agreement')
2	DEMOBILISATION AND REINTEGRATION OF CHILD SOLDIERS AND VULNERABLE PERSONS;	Intercongolense Negotiations: The Final Act ('The Sun City Agreement')
3	supervising the withdrawal of foreign troops;	Draft Constitution of the Transition
3	supervising the withdrawal of foreign troops;	Global and Inclusive Agreement on Transition in the Democratic Republic of Congo ('The Pretoria Agreement')

Cluster ID	Segment text	Document name
3	Servicemen in ongoing military service as part of military units;	Agreement between the Russian Federation and the Republic of Moldova on Matters Related to Jurisdiction and Mutual Legal Assistance on Issues Regarding the Russian Federation Military Formations Temporarily Situated in the Territory of the Republic of Moldova (Agreed in Moscow 21.10.1994)
3	Servicemen in ongoing military service as part of military units;	Agreement between the Russian Federation and the Republic of Moldova regarding the legal status, procedure and period for the withdrawal of the Russian Federation Military Units/Formations, temporarily situated in the territory of the Republic of Moldova
3	Deployment of troops in the Units.	Protocol of Agreement between the Government of the Republic of Rwanda and the Rwandese Patriotic Front on the Integration of the Armed Forces of the Two Parties



Cluster ID	Segment text	Document name
4	The use of children who are 18 years old and under in the armed forces.	Nepal Interim Constitution
4	The use of children who are 18 years old and under in the armed forces.	Agreement on the Monitoring of Arms and Armies
5	The armed forces belong to the people;	National Dialogue Conference Outcomes Document
5	be serving in the Army;	Protocol of Agreement between the Government of the Republic of Rwanda and the Rwandese Patriotic Front on the Integration of the Armed Forces of the Two Parties
6	j) To support efforts by relevant organisations to solve the problems of child soldiers, children who have disappeared, children who have been detained, and other children in Darfur.	Darfur Peace Agreement
6	d) Persons detained in relation to the armed conflict in Darfur and child soldiers shall be released.	Darfur Peace Agreement
Un-clustered	Refers to persons recruited and trained for the purpose of being employed as soldiers.	Lusaka Protocol

Cluster ID	Segment text	Document name
Un-clustered	soldiers for support services and administration;	The Protocol of Estoril (Bicesse Accords)
Un-clustered	soldiers, of whom 7,500 are to be operations personnel;	The Protocol of Estoril (Bicesse Accords)
Un-clustered	An indication of male, female, child soldier;	Comprehensive Ceasefire Agreement between the Government of the Republic of Burundi and the Palipehutu - FNL
Un-clustered	g. Special attention to such target groups as child soldiers, women soldiers and the disabled;	Arusha Peace and Reconciliation Agreement for Burundi
Un-clustered	Removing children from the armed conflict.	Acuerdo de 'Agenda Comun por el Cambio hacia una Nueva Colombia', Gobierno Nacional-FARC-EP
Un-clustered	lack of education for children;	Declaration finale du forum sur la PAIX dans le territoire de NYUNZU
Un-clustered	enrolment of children in the ranks of malicious forces;	enrolment of children in the ranks of malicious forces; Declaration finale du forum sur la PAIX dans le territoire de NYUNZU



Cluster ID	Segment text	Document name
Un-clustered	a presence of non-accompanied children:	Declaration finale du forum sur la PAIX dans le territoire de NYUNZU
Un-clustered	the Mixed Military Sub-Commission;	Ordonnance N° 08/008 du 02 Fev 2008 portant organisation et fonctionnement du programme national de sécurisation, pacification, stabilisation et reconstruction des provinces du Nord-Kivu et du Sud-Kivu, dénommé « Programme Amani »
Un-clustered	training and use of terrorists;	Ceasefire Agreement (Lusaka Agreement)
Un-clustered	Children and Young People's Strategy;	New Decade, New Approach
Un-clustered	Thorough knowledge of ECOMOG soldiers by the soldiers of the warring parties.	Agreement on Cessation of Hostilities and Peaceful Settlement of Conflict between the Armed Forces of Liberia, and The National Patriotic Front of Liberia, and the Independent National Patriotic Front of Liberia (Lome Ceasefire Agreement)

Cluster ID	Segment text	Document name
Un-clustered	The term "demobilized soldier" means an individual who:	General Peace Agreement for Mozambique
Un-clustered	The use of underage grazers.	Kafanchan Peace Declaration between Grazers and Farmers
Un-clustered	Approve plans for the utilization of the Army.	Protocol of Agreement between the Government of the Republic of Rwanda and the Rwandese Patriotic Front on the Integration of the Armed Forces of the Two Parties
Un-clustered	Arrests in the Military Prison:	Protocol of Agreement between the Government of the Republic of Rwanda and the Rwandese Patriotic Front on the Integration of the Armed Forces of the Two Parties
Un-clustered	ARTICLE XXX CHILD COMBATANTS The Government shall accord particular attention to the issue of child soldiers.	Peace Agreement between the Government of Sierra Leone and the Revolutionary United Front of Sierra Leone (RUF/SL) (Lome Agreement)
Un-clustered	Require the regular military and the White Army to demobilize all children under age fifteen;	Waat Lou Nuer Covenant



Cluster ID	Segment text	Document name
Un-clustered	The Parties shall abide by DDR principles related to the handling of child soldiers, persons with special needs, and women recruited in connection with the war.	Sudan peace agreement (Juba Agreement)
Un-clustered	The number of soldiers in this regiment should not be less than 500 soldiers to be supplied with all necessary means and weapons to enable them to implement the plan.	National Dialogue Conference Outcomes Document

Table 5: Clustering of topic search results for the topic 'The use of child soldiers.' Relevant provisions are marked in green. Clustering identifies noise in search results and improves the quality of search results and therefore classification.

Semantic Alignment

Semantic alignment is an estimation of the semantic similarity of two populations of topics and or text segments.

Applications include:

- Estimating the alignment of two documents.
- Estimating the alignment of topics in a vocabulary with a corpus of documents.

Methods

Measures of semantic alignment are based on the observed distributions of semantic similarity scores. For example, to measure the alignment of two documents, a similarity matrix is pre-compiled with the segments of one document in rows and the segments of the other document in columns. The observed distribution is generated by calculating the frequency of segment pairs at each similarity score.

Similarly, to measure the alignment of a topic with a population of text segments (from one or more documents), topic-segment similarity scores are retrieved from a pre-compiled similarity matrix where the topics are in rows and the segments of documents are in columns. The observed distribution is generated by calculating the frequency of topic-segment pairs at each similarity score.

We apply kernel density estimation (KDE) to observed distributions. KDE is a non-parametric method used to estimate the probability density function (PDF) of observed data (Silverman, B.W., 1986). Using integration, we determine the area under a PDF p within a similarity score interval I with limits $[a,1.0]$. The area p is the probability that a similarity score x is in the interval and is our measure of semantic alignment (A).

$$A = \Pr(x \in I) = \int_a^{1.0} p(x) dx$$

The larger the area, the higher the probability that a segment-segment pair or topic-segment pair exists with a similarity score in the interval, and therefore the greater the semantic alignment of the two populations from which the pairs are derived.

The reason for using KDE in our applications is as follows:

- We can estimate the shape of the underlying distributions of the observed similarity scores without assuming the scores follow a known distribution.
- We can control for differences in sample sizes when comparing topic distributions from two populations of text segments, e.g., comparing agreement provisions to implementation report sections.
- We can apply mathematical operations to PDFs, e.g., integration.
- We can create clearer visualisations of the distributions.

Use Cases

Peace Agreements and Constitutional Change Project

Table 7 below ranks peace agreements from the Framework/substantive stage of the Bosnia peace process by their semantic alignment with the 1995 constitution of Bosnia & Herzegovina.

A pre-compiled section-provision similarity matrix exists for each agreement where the constitution sections are in rows and the agreement provisions are in columns. The constitution is the reference document.

The top-ranking document is the Dayton Peace Agreement. The constitution was enacted soon after this agreement, but its text was in fact included as Annex 4 in the Dayton Peace Agreement. This ranking is therefore consistent with the finding from Table 3 above, showing that 930 sections of the constitution were semantically similar to provisions of the Dayton agreement. However, there is significant overlap between earlier agreements, including ones that failed to end the conflict, often being rejected shortly after being proposed or accepted.

Rank	Segment text	Date	Stage
1	General Framework Agreement for Peace in Bosnia and Herzegovina (Dayton Peace Agreement)	1995/11/21	Framework/substantive - comprehensive
2	Croat-Serb Constitutional Principles for Bosnia-Herzegovina	1993/06/23	Framework/substantive - partial
3	Declaration Concerning the Constitution of the Federation of Bosnia and Herzegovina (with Proposed Constitution of the Federation of Bosnia and Herzegovina attached)	1994/03/18	Framework/substantive - partial
4	Agreement relating to Bosnia and Herzegovina (Owen-Stoltenberg Peace Plan, or 'Invincible plan')	1993/09/16	Framework/substantive - comprehensive
5	Framework Agreement for the Federation (Washington Agreement or Contact Group Plan)	1994/03/01	Framework/substantive - comprehensive
5	The Public Announcement (Graz Agreement)	1992/05/06	Framework/substantive - partial

Table 6: The semantic alignment between six peace agreements from the Framework/substantive stage of the Bosnia peace process and the 1995 constitution of Bosnia & Herzegovina. Agreements are ranked in descending order of alignment.



Implementation Narratives Project

Figure 4 below illustrates the alignment of a pre-selected set of topics with the ceasefire agreement provisions and OSCE report sections from the Ukraine Minsk process from 2014 to 2022.

A pre-compiled topic-provision similarity matrix exists for the ceasefire agreements where topics are in rows and agreement provisions are in columns. A pre-compiled topic-section similarity matrix exists for OSCE implementation reports where topics are in rows and report sections are in columns.

The left-hand side of Figure 4 (yellow) shows the semantic alignment of topics with agreement provisions. The right-hand side (blue) shows the semantic alignment of the same topics with report sections.

Clearly topics are not equally represented in the agreements where some topics feature more than others, and some are not mentioned at all. As in the agreements, topics are not equally represented within the reports.

The figure also shows that the distribution of topics in the agreements is very different from that in the reports. Statistical analysis confirms that these two distributions are uncorrelated. This fact is interesting, because it shows that issues relating to ceasefire modalities preoccupied implementation.



Figure 4: The semantic alignment of a set of topics with ceasefire agreements and OCSE monitoring reports from the Ukraine Minsk process. The x-axis represents the probability of a topic-segment similarity score at or above a similarity threshold of 0.63.

7 Next Steps

Vocabulary Development

A set of topics optimised for semantic analysis were created in the 2022-2023 period. These topics are organised under PA-X codebook categories and codes. Further development is recommended as follows:

- Integration of clustering into the topic design process.
- Exploration of the relationship between topic design and agreement type. The language of agreements varies across different types of agreement and designing vocabularies for different types of agreement may be beneficial.
- Use other vocabularies (e.g., the UN's Language of Peace) to identify potential new topics and refine existing topics.

Verification

To build confidence in the technology there needs to be a robust system for verification.

Manually coded provisions are a potential gold standard against which to compare the results of topic search. However, reading large spreadsheets containing manually coded provisions and provisions found by topic search places a heavy burden on users, and so a new interface can be useful to compare results from manual and machine-learning coding. Furthermore, the formatting of manually coded provisions is inconsistent, meaning that an electronic 'match' to compare cannot always be used. Algorithms developed for reverse tagging³ that attempts to match manual coding to the segment of the text in the corpus, may provide a way of generating quantitative and qualitative comparison data with which to verify topic search results and improve the topic design process.

Automated Threshold Setting

As illustrated in Table 6, the performance of a topic cannot be measured simply by the number of topic search results. Whilst clustering can be used to locate and select relevant results, two thresholds are required: search and cluster. The automation of threshold settings would help further reduce the work of content analysts. Some promising early investigations suggest optimum cluster thresholds can be detected automatically.

Contextualisation and Styling

Contextualisation of text segments is already possible because the segmentation process ensures that the entire text of a document can be reconstructed from an ordered set of segments. The Implementation Narratives web application demonstrates one approach to contextualisation using HTML anchors. In this approach, clicking on a selected text segment opens the entire text of the segment's document in a panel, and scrolls the document text to the location of the selected segment which is highlighted.

What is currently lacking is any sense of the position of a text segment in a document's hierarchy of headings because segmentation removes any styling information that might indicate the position of headings. Extracting headings from numeric or alphanumeric labels is extremely difficult if there is no consistent formatting in a corpus.

If styling were available, it would greatly improve the user experience of the new PA-X tagging system as well as other applications based on the semantic toolkit described in this report. For example, locating the header of a text segment would help contextualise a segment without having to display preceding, and possibly redundant, segments. An ellipsis would indicate the presence of any preceding but omitted segments.

The development of a styling system based on text segmentation is recommended. The key features of this system would be:

- The separation of text styling from the text itself, i.e., no markup or markdown is added to the text.
- The allocation of a default styling to every segment during segmentation.
- The storage of segment styling in the PA-X tagging system database.
- Interfaces for styling segments as and when resource is available.
- Whenever peace agreement text (either the complete agreement or a set of segments) is displayed on a page, segment styles would be applied automatically. For example, headers could appear in bold larger font sizes.
- Separately configurable mapping of style onto HTML properties.



Infrastructure

Completing the move to EPCC infrastructure is essential for scaling the applications discussed here, in terms of sizes of corpora, speed of computation, and parallelisation of matrix builds. This is because the scale of data and computation is beyond that of ordinary laptops.

In addition, other encoding models can be explored. Currently, USE version 4 is our preferred model because it trades a small reduction in accuracy over version 5 for significantly improved speed and reduced hardware requirements.⁴

Models based on the BERT Transformer architecture are also slower than USE version 4 but might be feasible alternatives on EPCC infrastructure if significant improvements in semantic performance can be demonstrated.

Conclusion

The processes described above are now well developed with reference to PA-X data in terms of providing forms of automated coding, exploration of topics not coded for, tracing which agreement terms come from earlier agreements and persist into constitutional frameworks, and for extracting qualitative information on agreement of particular provisions, in reports over time. These will form part of the PeaceTech research methods of the PeaceRep team going forward.



Reference List

- Bell, Christine, Sanja Badanjak, Juline Beaujouan, Tim Epple, Robert Forster, Astrid Jamar, Sean Molloy, Kevin McNicholl, Kathryn Nash, Jan Pospisil, Robert Wilson, and Laura Wise (2021). PA-X Peace Agreements Database and Dataset, Version 5. www.peaceagreements.org
- Bell, Christine, Sanja Badanjak, Juline Beaujouan, Robert Forster, Tim Epple, Astrid Jamar, Kevin McNicholl, Sean Molloy, Kathryn Nash, Jan Pospisil, Robert Wilson, Laura Wise (2020). PA-X Codebook: Women, Girls, and Gender (PA-X Gender), Version 4. Political Settlements Research Programme, University of Edinburgh, Edinburgh. www.peaceagreements.org/wsearch
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil (2018). Universal Sentence Encoder. [ArXiv:1803.11175 \[Cs\]](https://arxiv.org/abs/1803.11175).
- Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.
- Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil (2020), Multilingual Universal Sentence Encoder for Semantic Retrieval, [Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 87–94 July 5 - July 10, 2020](#).

Endnotes

- ¹ <https://blog.gdeltproject.org/announcing-the-global-similarity-graph-television-news-sentence-embeddings-using-the-universal-sentence-encoder/>
<https://blog.gdeltproject.org/fact-checking-television-using-universal-sentence-encoder-embeddings-to-scan-news-for-fact-check-claims-at-scale/>
- ² At the time of writing https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.4.1
- ³ Reverse tagging attempts to integrate manually coded provisions into the new PA-X tagging system by matching manually coded provisions from the existing PA-X database to agreement text segments used by the new tagging system. Successful matching would ensure backward compatibility of the new system.
- ⁴ <https://blog.gdeltproject.org/experiments-using-universal-sentence-encoder-embeddings-for-news-similarity/>



✉ info@peacerep.org

f [PeaceRepResearch](#)

X [@Peace_Rep_](#)

in [PeaceRep](#)

www.peacerep.org

www.peaceagreements.org

About Us

PeaceRep: The Peace and Conflict Resolution Evidence Platform is a research consortium based at Edinburgh Law School. Our research is rethinking peace and transition processes in the light of changing conflict dynamics, changing demands of inclusion, and changes in patterns of global intervention in conflict and peace/mediation/transition management processes.

Consortium members include: Conciliation Resources, Centre for Trust, Peace and Social Relations (CTPSR) at Coventry University, Dialectiq, Edinburgh Law School, International IDEA, LSE Conflict and Civiness Research Group, LSE Middle East Centre, Queens University Belfast, University of St Andrews, University of Stirling, and the World Peace Foundation at Tufts University.

PeaceRep is funded by the Foreign, Commonwealth and Development Office (FCDO), UK.



THE UNIVERSITY
of EDINBURGH



PeaceRep
Peace and Conflict
Resolution Evidence
Platform

PeaceRep: The Peace and Conflict Resolution Evidence Platform
info@peacerep.org | <https://peacerep.org> | @Peace_Rep_

University of Edinburgh, School of Law
Old College, South Bridge EH8 9YL