Understanding Electoral Violence through Complex Textual Data: OSCE Monitoring Missions in Different Contexts

Michal Mochtak [†]

Institute of Political Science, University of Luxembourg

[†] Michal Mochtak, Ph.D. is a post-doctoral research fellow at the Institute of Political Science, University of Luxembourg (michal.mochtak@uni.lu), Maison des Sciences Humaines, 11 Porte des Sciences, L-4366 Esch-sur-Alzette, Luxembourg. ORCID: https://orcid.org/0000-0001-5598-5642.

Abstract

The paper analyses more than 20 years of evidence on electoral violence as reported by OSCE monitoring mission reports. It identifies prevailing trends of electoral violence in the OSCE participating states in order to better understand how the phenomenon is understood and framed by leading international monitoring organizations in the region. The analysis utilizes a unique approach based on automated content analysis employing counting algorithms and latent semantic indexing. The results of the analysis show how electoral violence differs throughout the region while highlighting the qualitative variations in regional patterns of the reported incidents of election-related violence.

Keywords

electoral violence; OSCE; elections; LSI

Funding

The paper was supported by an ERC Starting Grant No. 714589 for the project “Electoral Legacies of War: Political Competition in Postwar Southeast Europe" (ELWar).

Understanding Electoral Violence through Complex Textual Data: OSCE Monitoring Missions in Different Contexts

Introduction

Academics and practitioners have studied elections in democratizing and transforming societies, repressive autocracies, and stable democracies for decades. Number of them have shown that electoral competition as a central component of any democratization attempt is often affected by internal tensions fueled by a strong winner/loser logic election is inherently based on.[1] However, as election and democracy have become the most important concepts defining the 21st century political discourse in general, electoral violence as an extreme form of electoral competition is getting more attention as well. As many political actors and entrepreneurs approach electoral violence as a unique strategy for acquiring the political power or maintaining the political arena's status quo, its relevancy rises with the existing tension and/or polarization present in a society.[2] In this context, electoral violence can be defined as acts of physical or psychological violence that disturb the electoral competition and its outcome. These instances are distinguished from other criminal activities by their direct relationship to the events, personalities and issues comprising an electoral contest.[3] It is understood as a unique subcategory of political violence based on timing, a relationship to specific issues, instrumentality (violence intended to achieve defined goals) and consequences, all in connection to electoral arenas.[4]

Although existing literature postulates that electoral violence does occur in all sort of regimes and geographical regions, its empirical basis remains rather limited.[5] The primary focus lays on conflict-ridden societies in Africa and Southeast Asia with many ethno-cultural

cleavages affecting political as well as societal arenas.[6] Methodologically speaking, quantitative studies applying regression models dominate the debate with most cited works focusing on drivers and agents of electoral violence and its dynamics. The paper attempts to approach the study of electoral violence from a different perspective. We focus on the region of OSCE participating states that is generally overlooked by the mainstream research initiatives while applying novel methods based on advances of computational text analysis. The results are then combined with a social network perspective allowing us to visualize complex data in a transparent and comprehensive way while uncover patterns of electoral violence as reported by one of the most respected monitoring bodies – the Organization for Security and Cooperation in Europe (OSCE) .

The goal of the paper is to analyse electoral violence as reported by the OSCE in regions the organization monitors and to detect the prevailing patterns defining regional variations there. Based on the results, we argue that electoral violence significantly differs throughout most of the regions under study and as such reflects the socio-economic, cultural, and political patterns present there. The paper shows how the concept stretching might be reflected in different regions and highlights the empirical richness of the observed patterns in the OSCE countries. It provides an additional analytical layer for better understating of electoral violence and its regional differences, a perspective often neglected by the mainstream comparative research. Utilizing a multi-regional approach also provides us with a valuable insight on ongoing sophistication of coercive practices that have been observed in the past decade all around the world. This is especially relevant with the recent rise of modern authoritarian regimes that actively learn from each other and export their successful practices abroad.[7]

The paper is formally covered by a research question Q: *How does electoral violence differ in different regions monitored by the OSCE (?)*. We apply a novel approach to study of electoral violence based on computational text analysis and process a corpus of OSCE documents covering 362 monitoring/assessment reports and counting over 9200 pages of text in the period of 1996-2018. The analysis utilizes a vector space model based on latent semantic indexing (LSI) which constructs a semantic space defining the context of keywords we are interested in (in our case the keywords referring to electoral violence). In contrast to more commonly known keywords in context approach (KWIC), the LSI model produce simplified vector space model summarizing the *contexts of keywords* over the whole corpus. Context here does not refer to a relative frequency with which words co-occur in the same document, but the extent to which they have the same effect in the construction of total passage meanings.[8] Using cosine similarities and tf-idf weighting, we are able to identify contextually similar words that are close (similar) in the vector space of a corpus while describing the context of the keywords on average. In comparison to simple counting of n-grams, the LSI vector space model can preserve the information on semantic space and further process it. Although available alternatives like word2vec and doc2vec models might be superior in terms of quality of the results, the LSI model can generally outperform them on a smaller corpus like this one. The results allow us to understand what concepts look like in different geographical regions and map the observed variations.

The results show that electoral violence, as reported by OSCE monitoring/assessment missions, is not a static but rather dynamic phenomenon typologically following the characteristics of a specific sub-region in which it occurs. Patterns observed in Eastern Europe (post-Soviet region; Western Balkans) transparently show how incidents of electoral violence may differ through the studied sub-regions reflecting the nature of conflicts present in their

political arenas. It stands in stark contrast to the patterns reported in Western democracies where cyber-attacks represent one of the main pools of violence-related incidents. Although empirically-oriented, the paper also presents a novel approach in the study of electoral violence and discusses it as a complementary tool to more mainstream methodologies which dominate the ongoing debate.

Studying electoral violence from text

Most of the studies published on electoral violence in the recent years have focused on identification of context- and election- related drivers of electoral violence potentially responsible for the occurrence of the phenomenon. Traditionally, it has been seen as a way of manifesting political instability during transformation processes or as a strategy for advancing the goals and policies of authoritarian regimes.[9] In recent years, scholars have approached electoral violence not as a sole by-product of troubled political liberalization, but also as an integral part of electoral campaigning.[10] Traditional theories emphasise the importance of the structure of the electoral system and the effectiveness of related institutions[11], ethnic tensions[12], and a country's socio-economic situation as factors which may ignite violent confrontation during elections.[13] Pevehouse, Straus, and Taylor[14] found that most election-related violence reported in Sub-Saharan Africa is committed by incumbents seeking re-election. According to them, the pre-existing social conflict and the quality of founding elections shape pre-election violence, while the stability of democratic institutions and weaker economic growth shape violence after polling stations close.[15] Arriola and Johnson[16] identify clientelistic corruption as a potential driver of electoral unrest while Hafner-Burton, Hyde, and Jablonski[17] see the loss of incumbent elites as a risk factor which may escalate violent clashes

and deepen existing tensions. Recent study by Sandra Ley showed that the strategic use of violence by organized crime groups during electoral campaigns in Mexico demobilizes voters at large. She suggests that the impact of a criminal context on turnout transcends personal victimization experiences. As a result, regions where criminal organizations attempted to influence elections and politics by targeting government officials and party candidates exhibited significantly lower levels of electoral participation.[18]

Although different forms of textual data certainly are important for most of the findings presented in these studies, almost none of them use it as a primary data. Many of them are based on large-N datasets among which the most popular are National Elections Across Democracy and Autocracy (NELDA)[19], Varieties of Democracy (V-Dem)[20] and recently Countries at Risk of Electoral Violence (CREV).[21] Although comprehensive and very complex in nature, the level of aggregation, coding, and generalization necessarily affect the amount of data these datasets are losing. A partial exception in the context of a *text as data* approach is the CREV dataset by Birch and Muchlinski utilizing the automated coder platform implemented as a part of the Integrated Crisis Early Warning System (ICEWS), an event data project developed by Lockheed Martin Advanced Technology Laboratories.[22] Although Birch and Mulchinski certainly push the bar higher in terms of standards of natural language processing and provide a new direction for processing data on electoral violence, their effort is limited by the level of aggregation and overall goals that match the expectations set by large-N datasets, still so popular among political scientists.

To the author's best knowledge, there is currently no study implementing natural language processing tools for analyzing textual data on electoral violence or related issues. The truth is that beauty of this approach is penetrating the mainstream political science only slowly, mostly present in studies on parliamentary and campaign speeches, media outlets, and

social networks.[23] However, the available tools are not reserved solely to these domains and can be applied to all sort of social science problems. Articles, transcripts, blogs, reports, posts, and tweets are unique and invaluable source of information that can be processed and analyzed. The *text as data* approach provides a whole new set of analytical possibilities for issues and sensitive research problems that are often hard to study and are in constant danger of being biased. Moreover, the costs of these studies can be significantly reduced through automated algorithms that allow systematic scraping, sorting, coding, and analysis. In political science, most of the analyzed corpora are not that big requiring massive computational power of super computers, allowing scholars to process their data on regular machines.

Electoral violence in this context represents a unique opportunity where number of these tools can be applied. This is not the place for review of the available tools but rather highlighting their primary benefits. Through them we can trace the development of topics we are interested in (counting and weighting of n-grams), identify these topics (topic-modeling), scale the evidence on a single (Wordfish) or multiple scales (Wordscores), understand the contexts in which concepts occur (keywords in context, LSI models, word2vec models), detect actors, organizations, or places (named entity recognition), understand the semantic structure of a written and/or spoken text (semantic network analysis), or ultimately build a complex machine-learning algorithms based on neural networks in order to get closer to a state when computer programs can read and understand the natural language for us. This is just a fraction of opportunities we have at hand and can be used for study of electoral violence, conflicts, and politics in general. New technologies have made available vast quantities of political texts, documenting an increasing share of political communication, interaction, and culture. These data sources are unique and might provide new answers to already addressed questions or come up with new ones we have not thought about yet. One way or another, the information

encoded in text is a rich complement to more structured kinds of data traditionally used in political science research.

Text as data: OSCE corpus

For the purpose of this paper, we have compiled an original corpus consisting of OSCE monitoring/assessment reports in English covering 362 elections organized between 1996 and 2018.[24] The corpus consists of a unique set of documents harvested from the official OSCE website using R package *rvest*.[25] Although the structure, focus, and content are mostly consistent throughout the documents, the corpus is a combination of final reports, preliminary reports, assessment reports, and various types of expert reports. The corpus in its raw form (no cleaning or preprocessing) consists of 4.5mil words, with over 45.000 unique tokens and roughly 21.000 sentences. Simple document term matrix with term frequency weighting at this stage has 96% sparsity.

This meta-corpus is further cleaned and preprocessed in order to build a vector space model that can be analyzed. Each document is cleaned off its title page, footers and headers and lemmatized using R package *udpipe*, a natural language processing toolkit providing language-agnostic tokenization, parts of speech tagging, lemmatization and dependency parsing of raw text.[26] Lemmatized documents are further cleaned with R package *tm,* which removes stop-words, numbers, punctuation, and whitespaces. All uppercase characters are converted to lowercase.[27]

The corpus represents a comprehensive database of electoral monitoring reports unique in its complexity and size. After cleaning and preprocessing, the *stage two* version of the corpus consists of 362 documents with 539.734 words and 39.005 unique tokens (number of

sentences is not reported anymore as we removed the punctuation in the pre-preprocessing phase). Simple document term matrix with term frequency weighting has again 96% sparsity (Non-/sparse entries: 539.734/13.580.076).

Methods and models

The pre-processed corpus still needs to be analyzed in order to answer the aforementioned research question. The following section will briefly discuss applied tools, which can be generally divided into three groups. The first one is based on corpus level counting algorithms, which monitor occurrence of keywords as unigrams, bigrams, trigrams, and 4-grams. Besides that, it allows us to identify empirical relevant instances of the phenomenon under review – the electoral violence. Based on the definition presented in the introduction, a general list of all the words in our corpus and their frequencies are reviewed so those terms falling under the definition of electoral violence can be selected to a working dictionary for further processing. This step is repeated on the level of unigrams, bigrams, trigrams, and 4-grams. For an n-gram tokenization (n = [1, 4]) and actual counting, we use R package *tm*.

The second group of analytical tools (and the most important for this study) is based on principles of latent semantic indexing (also known as latent semantic analysis, LSA). It is based on the assumption that texts have a higher (=latent semantic) order which, however, is obscured by word usage (e.g. through the use of synonyms or polysemy). By using conceptual indices that are derived statistically via a truncated singular value decomposition (a two-mode factor analysis) over a given document-term matrix, this variability problem can be overcome.[28] To construct a semantic space for language, LSI first casts a large representative text corpus into a rectangular matrix of words by coherent passages, each cell containing the

number of times that a given word appears in a given text. The matrix is then decomposed in such a way that every text is represented as a vector whose value is the sum of vectors standing for its component words. Similarities between words and words, texts and words, and texts and texts are then computed as dot products, cosines or other vector-algebraic metrics.[29] The approach simultaneously models the relationships among documents based on their constituent words, and the relationships between words based on their occurrence in documents.[30]

LSI is a fully automated statistical approach for extracting relations among words by means of their contexts of use in documents, passages, or sentences. As such, it belongs to a group of unsupervised learning techniques. It starts with a large collection of texts, builds a term-document matrix, and tries to uncover existing similarity structures that are useful for information retrieval and related text-analysis problems.[31] The LSI algorithm consists of four main steps: 1) in the first step, a large collection of text is represented as a term-document matrix. Rows are individual words and columns are documents represented in corpus. Individual cell entries contain the frequency with which a term occurs in a document. The order of words in a document represented in the matrix does not matter as the approach is based on a bag of words representation; 2) as raw frequencies can be often misleading, the next step transform them into their weighted form representing their relevancy to a document, in a corpus or both. For the purpose of this study, we use term frequency – inverse document frequency (tf-idf) score which reflects how important a word is to a document in a collection or corpus; 3) in the next step a reduced-rank singular value decomposition (SVD) is performed on the matrix, in which the *k* largest singular values are retained, and the remainder set to 0. The resulting reduced-dimension SVD representation is the best k-dimensional approximation to the original matrix, in the least-squares sense. Each document and term is now represented as

a k-dimensional vector in a space derived by the SVD; 4) in the last step similarities are computed among entities in the reduced-dimensional space, rather than in the original term-document matrix. Because both documents and terms are represented as vectors in the same space, document-document, term-term, and term-document similarities are all easy to compute. The cosine or angular distance between vectors can be then used as the measure of their similarity for many information retrieval applications, which has been shown to be effective in practice.[32] The whole analysis is performed with R packages *lsa* and *tm.*[33]

The last applied tool in this paper is based on visualization capabilities of the social network analysis in order to unveil complex relations among concepts we follow. Keywords (concepts) selected from the list of n-grams and their closest co-occurring words represent nodes. The non-zero cosine distances of the keywords (concepts) and the co-occurring words represent edges. Cosine distance is a measure of similarity between two non-zero vectors of an inner product space represented as the cosine of their angle. As some of the keywords share co-occurring words in their vector space, the structure creates a network of ties, which can be visualized and analysed. The visualization is done with Gephi v0.92.[34]

Each of the presented tools contributes to the overall goal of the paper and the answering of the aforementioned research question. Following sections apply all of them in a sequential order on the level of whole corpus followed by different regions.

OSCE monitoring missions and electoral violence

The OSCE Office for Democratic Institutions and Human Rights (ODIHR) carries out election observation in OSCE participating States to assess whether the elections comply with OSCE

commitments, other international obligations and standards for democratic elections, and with national legislation.[35] From its start in 1996, the OSCE has, to these days (as of May 1, 2018), issued 362 monitoring/assessment reports for countries all around the world. The pre-processed corpus is analysed in order to understand how electoral violence is seen and understood by OSCE monitoring missions in the OSCE participatory states. Although not without problems and shortcomings (e.g. criticism about political bias), the mission/assessment reports are one of the most valid sources of evidence on quality of election in the OSCE countries following more or less the same standards for the past 20 years. As such, the results can be seen as generally objective or at least having a higher level of objectivity than news, blogs, posts, tweets or any other available written sources that are usually used by experts when assessing quality of elections in general.

In order to identify concepts associated with electoral violence under assessment, we construct a list of all unigrams that occur in our corpus at least twenty times and can be seen as empirically relevant enough. The list has 5.425 items. They are inspected manually for words potentially referring to electoral violence as defined in this paper. We identify 25 such words, which will be further tracked and analyzed (see Table A in Annex). Only the word "intimidat" is not clearly lemmatized and needs to be handled manually. It is a result of an imperfect lemmatization that, according to the authors, has only 97% accuracy.[36] We need to look for the original word before the lemmatization was applied in order to identify its original form. Most of the word's pre-processed form is "intimidating" which is already present in the list and can be easily recoded. The final list has 24 keywords referring to electoral violence that can be tracked throughout the corpus. The same approach is applied for identification of keywords on the level of bi-grams (20), tri-grams (8), and 4-grams (10). The summary is

presented in Tables B – D in Annex. With these dictionaries, we can proceed to corpus level analysis and then move on to parsing the corpus into smaller unites based on regional clusters.

Electoral violence in the OSCE participating states (1996-2018)

Taking a better look at the corpus, it is easy to summarize it through the most used words, which are present in almost all documents. Table 1 presents ten most used unigrams, bigrams, trigrams, and 4-grams. If we would not know anything about the corpus, we can be quite certain it is about elections, candidates, voting, campaigns, OSCE, and human rights. If we compare it with the ten most frequent ngrams related to electoral violence (Table 2), we can see that concepts are (relatively) not that common in the corpus yet still relevant enough to be studied. The most frequent word is *intimidation* followed by *violence*, *attack*, and *detention*. The frequency of these words and their presence in many documents however does not mean that electoral violence actually took place. References like "violence-free election", "without fear" or "no intimidation" might be quite common in the reports and generally refer to what is seen as a good conduct of election. In order to assess the actual meaning of the violence-related words we need to process them further. In order to do so, we build a vector space model based on latent semantic indexing (LSI model) and visualize it.

[Table 1. Summary of ten most used unigrams, bigrams, trigrams, and 4-grams]

[Table 2. Summary of ten most used violence-related unigrams, bigrams, trigrams, and 4-grams]

The LSI model is built on top of a cleaned and preprocessed corpus with term document matrix weighted on tf-idf scores, 95% sparsity (only words occurring at least in 5% of documents are included), and k dimensions based on statistically best-fitting number of singular values for the dimensionality reduction. [k] is selected based on the examination of the singular-value decomposition of a rectangular matrix plotted in Figure 1. The knee of the curve defines the cut-off point for the number of dimensions to retain (k = 50).

[Figure 1. Visualization of singular-value decomposition]

A document-term matrix M is constructed with a function textmatrix() from a given text base of n documents containing m terms. This matrix M of the size m x n is then decomposed via a singular value decomposition into a term vector matrix T (constituting left singular vectors), a document vector matrix D (constituting right singular vectors) being both orthonormal, and a diagonal matrix S (constituting singular values).

$$M = T\,S\,D^T \qquad \{1\}$$

These matrices are then reduced to the given number of dimensions $k=dims$ to result into truncated matrices T_k, S_k and D_k — the latent semantic space. As such, the matrices (T_k, S_k, D_k) are multiplied to give a new matrix M_k (of the same format as M, i.e., rows are the same terms, columns are the same documents), which is the least-squares best fit approximation of M with k singular values.[37]

$$M_k = \sum_{i=1}^{k} t_i . s_i . d_i^T \quad \{2\}$$

The document-term matrix is then converted into a text matrix so comparison of terms and documents with common correlation measures can be performed. With the final text matrix we can inspect words we are interested in and look for their co-occurring words with the closest cosine distances. This creates lists of words with identified ties to keywords we want to understand better. Table 3 summarizes the first five unigram-keywords with their five closest words in the constructed vector space. Although we might already sense the context of the incidents, the level of complexity of the observed relations (links) is still relatively high.

We can say that *intimidation* is contextually mentioned together with words like *violence*, *irregularity*, or *evidence* indicating that most of the intimidation is discussed as an irregularity, often together with violence supported by evidence. *Observer* on the other hand might be referred to either as a target or a witness. This tells us a great deal on how *intimidation* as a form of electoral violence is reported over the period of twenty years. It can help us to summarize the context of the keyword without reading over 9000 pages of text and trying to generalize the prevailing patterns. Although different contexts in which *intimidation* may occur in the corpus are possible, they are not dominant or even significant (e.g. reference to intimidation-free election). Similarly, we can assess the average context of *violence*, *attack*, *detention*, *fear* or any of the 24 keywords we have selected as violence-related. It does not mean that all the keywords must be present in the overview, as not all of them pass the sparsity threshold filtering the infrequent words in the corpus.

[Table 3. Cosine distances of the co-occurring words of the five most frequent unigrams]

The words that pass the threshold do not exist in the vector space separately and as Table 3 suggests, some of them might be interconnected, meaning they are contextually close to each other (e.g. *intimidation – violence*). To better understand these structures, we employ a network perspective, which is capable of visualizing concept relations on the level of a corpus. We take all the unigram-keywords and plot them together with the 20 closest co-occurring words. The visualization is done with Gephi v0.92 using Yifan Hu layout (with a default setting). In order to show the main concept and clusters of words we adjust the size of the nodes based on their degree and color them based on their modularity class (resolution = 1.0, modularity: 0.756)[38]. The thickness of the edges refers to their cosine distance where thicker edge means closeness in the vector space while a thinner one the opposite. The results are visualized in Figure 2.

[Figure 2. Visualization of network of co-occurring unigram-keywords]

This provides us with an additional layer of insight describing evidence on electoral violence in OSCE participating states over the period of more than 20 years. The graph can be discussed from three perspectives: 1) the macro- structure of the whole property space; 2) the meso- structure based on the identified clusters and structural similarities of the concepts; and 3) the micro- structure of the individual concepts and their immediate neighborhood.

The first perspective is showing that not all concepts are equally well connected and some of them appear to be more cohesive than others. We can see this in the case of words like *intimidate*, *intimidation,* and *intimidating* on one hand and *threat* and *threaten* on the other. It is not a surprise as they contextually refer to the same instances of behavior. Although all these concepts are appearing together in a similar context, they are not the same (otherwise, their position would be identical sharing most or all of the co-occurring words). Although the connections are based on tf-idf weighting, we cannot interpret the graph centrality as a ratio of importance. We can however discuss the context under which these words occur and are connected through. The most important finding here is that words referring to serious forms of physical violence like *harm* and *kill* are contextually very weakly connected to any of the less violent yet more common concepts like *intimidation* or *threat* or the general concept of *violence*. As we are using 20 closest co-occurring words for describing context of the followed keywords, we are virtually covering segments of two to three sentences in natural language. As these concepts do not share many words, it suggests that they are not very close and are potentially discussed in different contexts (criminal activities, historic events, un-related incidents, etc.). On the other hand, keywords like *threat*, *threaten, intimidation* and *intimidating* are strongly related and close to general concepts of *violence* and *violent* showing what might be the prevailing pattern of electoral violence in the region if discussed so. In other words, when OSCE is talking about violence, most often it is connected to acts of intimidation directly or through its shared property vector space. This is a robust summarization of the existing patterns observed in the corpus already providing us with a great deal of insight based on cumulative knowledge harvested from over 9000 pages of text.

The second perspective (the meso-level), operating on structural similarities, shows clusters of concepts, which are close to each other and distinguish them as such by color. These

communities can be analyzed based on their closeness and ties they share in order to better understand how the concepts are interconnected. Previously discussed keywords like *threat, threaten, intimidation* and *intimidating* are close to each other because they share co-occurring words in their neighborhood like *bad, supporter, pressure, observer*, or *instance* or they are connected directly (e.g. *violence* - *intimidation*). These words indicate the context under which the keywords are discussed telling us what kind of violence was prevailing in the region, who was involved and how strong the relations are. Another example is the connection between words *attack* on one hand and *destruction, coercion,* and *threat* on the other. We can assume that attack as a form of destructive action actual caused destruction and damage to the attacked side. What is interesting here, the *attack* and *destruction* do not have to infer a brute force but also cyber-attacks and crimes on the internet (this will be more apparent when analyzing electoral violence in different regions; see below).

The last perspective (micro- level) focusses on single concepts and their unique position in the network structure. Similar to previously discussed concepts and their cosine distances (Table 3), any keyword can be analyzed in order to better understand its (average) meaning in the corpus. Let us discuss the keyword *harm* as a representative of brute force and one of the most severe concepts followed. As we can see, it is not connected to the bulk of other keywords in the network as reported through its first 20 closest words. Although the closest word (*iccr* standing for International Covenant on Civil and Political Rights) does not say much, rest of the neighbors are quite descriptive. Words like *odd, unhrc, paragraph, proportionate, effective, refrain, unduly, obligation,* or *standard* most probably refer to general international standards often cited as a framework for conducting free and fair elections. In this context, *harm* is most often not discussed in terms of incidents of electoral violence rather formal provisions that either need to be implemented (as a form of recommendation) or

alternatively, are discussed as being violated. With this information we can go back to a list of n-grams and look for more complex constructions telling us a bit more about the actual incidents. It is important to stress that this approach is not good for uncovering single incidents rather some prevailing patterns. Moreover, higher n-grams can show only the most repeating patterns which might be limiting in terms of interpretation but are good for generalization. To better demonstrate this, we can go back to the list of 4-grams and select those, which contain the keyword *harm*. There are three of them: *design – restore – reputation – harm*; *strictly – proportionate – actual – harm*; and *harm – cause – law – prioritize*. All three 4-grams refer to the same thing describing institutional setting that causes or caused harm to electoral process. It is a different context than *harm* as an instance of physical injury, especially that which is deliberately inflicted. With a bigger corpus, the result might be more complex and insightful but even with this one we can already see some patterns otherwise hidden.

Similar analysis can be performed on the level of higher ngrams, although we can expect less complex results as the number of analyzed items would be smaller. The number of items in the lists of higher ngrams is higher in comparison with unigrams but the term-document matrix is much sparser as many combinations of words are uniquely present only in small number of documents. With bigger corpus, this does not have to be an issue and can provide additional insight on contexts we are interested in.

Electoral violence in different regions

Although identifying general patterns on the level of whole corpus is interesting and can provide us with valuable insights about patterns of electoral violence in the OSCE countries, the existing differences in geographical sub-regions might reveal variations we are interested

the most. We specifically focus on reporting on electoral violence in three regions of OSCE participating states in order to show how electoral violence may differ based on regional predispositions and varying historical paths. We slice the original corpus into three sub-corpora covering Western World, post-Soviet countries, and the Western Balkans and compare them with each other. The decision is based on the number of available reports in each of the regions, so latent semantic modeling and its comparison makes sense, and the fact that we want to highlight the difference between *East* and *West* as well as to show how Eastern Europe can differ in its constituting regions. The regions are delimited quite straightforwardly: 1) Western World covers all the western democracies that were not part of the Eastern Bloc before 1989 (86 elections); 2) post-Soviet region covers all the former countries of the Soviet Union minus Baltics (87 elections); and 3) the Western Balkans covers all the countries of the former Yugoslavia minus Slovenia, plus Albania (81 elections). Analytical steps are similar to previous section. Table 4 summarizes most frequent unigrams, bigrams, trigrams, and 4-grams in order to see what does dominate in each of the regions. We do not standardize them, as we are only interested in the most frequent words and not their frequencies per se. As we can see, the reports are quite coherent in terms of most frequent words, differing only slightly in their order and relative frequency. The only true outlier is the reference to *cec* (central electoral commission) which is very prominent in the reports assessing election in post-Communist countries. Reading through the reports can reveal it is a result of persistent criticism of independence and performance of the election management bodies (EMBs) in the region as well as relatively high number of complaints the EMBs have to deal with.

[Table 4. Summary of ten most used unigrams, bigrams, trigrams, and 4-grams per region]

More interesting are the results of latent semantic indexing and their visualization in network layouts. We again build LSI model for each of the studied regions and follow each of the 24 keywords and their 20 closest co-occurring words. The number of dimension is again set at k = 50 so we can compare the results of the modeling on the level of regions with the modelling on the level of the OSCE countries presented in the previous section. The networks are summarized in Figure 3 – 5. Full-scale visualizations are available in Annex.

As we can see, the constructed vector space models quite differ in their respective structures, showing how electoral violence is reported by the OSCE in different contexts. This is crucial for the understanding of the electoral violence as a phenomenon that is strongly affected by contextual settings under which the election is organized. Each of the contexts are different and concepts associated with electoral violence are not placed in the same positions and do not have identical relations. The results show that either OSCE intentionally talk about similar incidents of electoral violence differently, or the incidents simply differ naturally. Either way, the concepts are framed differently which is obvious through Figures 3 – 5.

Figure 3 visualizes the semantic space of reports covering elections in post-Soviet countries. As we can see the concepts related to electoral violence in the region are much more densely connected than in the other two regions under study or the overall space defined on the level of whole corpus. Although reference to brute force is present (*injury*, *death*), it is not discussed together with mainstream notions of violence which is more associated with coercive strategies and non-physical violence (*threat*, *intimidation*). Interestingly, word *attack* moves to the center of the cluster explicitly referring to violence showing that OSCE often refers to attacks when talking about violence. The whole graph shows that electoral violence in the region is often described through intimidation practices that involves not only threats but also detention and harassment. This is a picture of electoral violence we would expect in

the region where most of the ruling elites use their position to influence the outcome of the competition and actively prosecute independent journalists, political activists, and the opposition. Electoral violence is often just one of the available strategies that is used wisely. The instances referring to brute force like injury, harm, kill, or bomb are not that important for the overall picture. They occur in the reports but on average are not discussed as something integrally connected to more common forms of electoral violence. In other words they occur and are relevant enough to be reported in the graph, but their isolated position tells us that this is not something systematically affecting electoral competition on the level of the region. Good connectedness of the central nodes on the other hand indicates variability of electoral violence and its multi-dimensional nature.

[Figure 3. Post-soviet region]

The situation is a little bit different in the Western Balkans (Figure 4) where the concepts are less connected yet still tell an interesting story. As we can see, the central group is similar but not identical to the one discussed in the post-Soviet countries. Especially close connection of *injury* to *intimidate* and *intimidating* is showing that patterns of electoral violence are different and brute force is contextually more common (or more commonly reported as such). *Injury* is well connected to *attack* which might refer to interpersonal and intergroup confrontation, so common in some of the countries in the region (e. g. Albania, Macedonia, Bosnia and Herzegovina). Most serious acts of electoral violence with human causalities (*death*, *murder*), although reported, seem to be discussed in their own contexts not connected to the mainstream notion of electoral violence as presented by OSCE. The

interesting finding is that the central part of the graph seems to have two branches connected through two central nodes (*threat – violence*) showing on one hand the pattern we already saw in the previous graphs, where violence is relatively strongly connected to *intimidation* and *threat/threaten*, and on the other, the dynamics where injury as a form of inflicting harm is vitally present. As we can see electoral violence in the Western Balkans is different from the one discussed in the post-Soviet countries. Although authoritarian practices are certainly present in the region (e.g. see the position and the context of the keyword *detention*), the contentious interaction exist also on a grass-root level capturing the intercommunity tensions in the region.

[Figure 4. Western Balkans]

The picture discussed in the previous paragraphs fall apart entirely in the Western democracies (Figure 5), where political and electoral violence is not that common yet still exist (e.g. post-electoral clashes between pro- and anti- Trump activists in 2017). In the graph, we can spot two main communities of concepts co-occurring close to each other. The first one combining triad *attack – destruction – coercion* while the second one being a tetrad of *violence – fear – intimidation – death*. The first one refers mostly to cyber-attacks and communication problems associated with elections. Although present in different regions as well, here the framing dominates the context. The second community is however more interesting. It is not a surprise to see the general concept of *violence* connected to *fear* and *intimidation* as the most dominant pattern in the OSCE countries. However, the association of *death* to *intimidation* and *fear* is not very intuitive in the region. Again, it does not say anything about the intensity or the frequency of the incidents, rather describing the context. If we take a better look at the

keyword *death*, we can see some hints potentially explaining the observed connections. Words like *exemption, argument, deceased, midterm, white, ninth,* and *street* might refer to elections in the United States with potential street violence and racially motivated clashes mentioned in the reports. Especially the last presidential election in the country showed how divided the society is and what mutual hostility can cause in the time of election.[39] Similar tensions do exist in other Western democracies as well (e.g. Spain or the UK/Northern Ireland) and as such could cause tension if the social or political turmoil prevail.[40] The picture is however nowhere near the picture presented in post-Soviet countries or Western Balkans. We can unsurprisingly conclude that electoral violence is quite rare in Western democracies and is represented by different patterns typical for the level of advancement the region is specific of (e.g. the attacks in cyber space).

[Figure 5. Western Democracies]

Based on the presented overview, we can argue that different regions do experience different forms of electoral violence. The important finding here is that brutal force and sever coercion strategies we know from elections in Africa are contextually not that common in the OSCE countries as a whole or its sub-regions. However, as we have showed, the relevancy of the argument differs throughout the regions as the incidents of electoral violence are not the same. Advancements we can observe refer mostly to sophistication of strategies regimes use against their opponents and critics (already mentioned shift from brute force to non-physical coercive strategies) as well as penetration of cyber space where electoral violence overlaps with the conceptual space of electoral integrity and security of strategic infrastructure. Text as

data is a powerful tool how these patterns can be detected and understood. It does not mean that all the co-occurring word have to have apparent meaning in the vector space and can be interpreted literally. The advantage of this approach is to see the macro structure of concepts we want to understand and study them in comparative perspective.

Conclusion

The paper presents a novel approach to study of electoral violence combining natural language processing tools and social network analysis. We have analyzed a corpus of OSCE monitoring reports and tracked down concepts associated with electoral violence in OSCE states. Based on the presented analysis we can claim that electoral violence is not a static phenomenon and varies across countries as well as regions. The paper analyzed the pre-processed corpus in two different settings – first, on the level of the whole corpus, and second, on the level of three regions (post-Soviet countries, Western Balkans, Western democracies). Each of the constructed models provided slightly different results showing variation of how electoral violence is discussed/seen by international authority the OSCE represents.

In order to answer the research question, the results can be summarized in three points. First, based on the observed patterns at the level of the whole corpus, most of the incidents reported in the OSCE reports are discussed more in the context of intimidation and threats than physical force. Second, this pattern shows a certain level of sophistication that differs from typical instances of electoral violence reported in Africa and Southeast Asia. Political entrepreneurs as well as activists in the region use coercive strategies tactically rather strategically and implement them in their strategic portfolio as one of the available tools of electoral campaigning. Third, the regional differences in post-Soviet countries, Western

Balkans, and Western democracies show that electoral violence is potentially affected by cultural, historical, and political contexts of each of the regions and the countries defining them. We are talking here mainly about top-down strategies of controlling the public space in post-Soviet modern autocracies (Russian, Ukraine, Belarus), political activism and social polarization in the Western Balkans (Macedonia, Albania, Bosnia and Herzegovina), and although significantly less violent yet still present, social and cultural polarization in the Western World (the United States, Spain).

This assessment aims to open a new chapter in the study of electoral violence employing tools that proved to be powerful in helping us to understand the world around us in the recent years. The contribution of this paper in this direction is three-folded: 1) the paper provides an alternative perspective on electoral violence and its understanding through analysis of primary textual sources where text is treated as data; 2) based on presented analysis we can claim that electoral violence varies through regions and as such needs to be understood – through non-static, constantly evolving and advancing coercive strategies that are often employed as a way of campaigning; 3) electoral violence is moving to virtual space with attacks being performed on critical infrastructure, political opponents, and general public. Elections are organized in 21st century and so is the electoral violence. Text as data approach is a methodological advancement that allows us to study phenomena otherwise hidden, complicated, or ethically challenging. Although it should not be taken as a universal approach, it definitely represents a valuable complement to the mainstream research designs based on statistical modelling and substantial case-oriented knowledge.

Notes

1. Pippa Norris, Richard W Frank, and Ferran Martínez i Coma, *Contentious Elections: From Ballots to Barricades* (New York: Routledge, 2015).
2. Thad Dunning, "Fighting and Voting: Violent Conflict and Electoral Politics," *Journal of Conflict Resolution* 55, no. 3 (2011): 327–39.
3. Kristine Höglund, "Electoral Violence in Conflict-Ridden Societies: Concepts, Causes, and Consequences," *Terrorism and Political Violence* 21, no. 3 (2009): 412–27; David Rapoport and Leonard Weinberg, *The Democratic Experience and Political Violence*, ed. David Rapoport and Leonard Weinberg (London: F. Cass, 2001).
4. Norris, Frank, and Martínez i Coma, *Contentious Elections: From Ballots to Barricades*.
5. Norris, Frank, and Martínez i Coma; Sarah Birch and David Muchlinski, "The Dataset of Countries at Risk of Electoral Violence," *Terrorism and Political Violence*, September 26, 2017, 1–20.
6. Hanne Fjelde and Kristine Höglund, "Electoral Institutions and Electoral Violence in Sub-Saharan Africa," *British Journal of Political Science* FirstView, no. Supplement-1 (2015): 1–24; Dorina Bekoe, *Voting in Fear: Electoral Violence in Sub-Saharan Africa*, ed. Dorina Bekoe (Washington D.C.: United States Institute of Peace, 2012); Michael Wahman, "Democratization and Electoral Turnovers in Sub-Saharan Africa and Beyond," *Democratization* 21, no. 2 (2014): 220–43.
7. Stephen G. F. Hall and Thomas Ambrosio, "Authoritarian Learning: A Conceptual Overview," *East European Politics* 33, no. 2 (April 3, 2017): 143–61; Valerie Bunce and Sharon L Wolchik, *Defeating Authoritarian Leaders in Postcommunist Countries* (Cambridge: Cambridge University Press, 2011); Larry Jay Diamond, Marc F Plattner, and Christopher Walker, *Authoritarianism Goes Global: The Challenge to Democracy* (Baltimore: John Hopkins Universtiy Press, 2016).
8. Thomas K Landauer, Darrell Laham, and Marcia Derr, "From Paragraph to Graph: Latent Semantic Analysis for Information Visualization.," *Proceedings of the National Academy of Sciences of the United States of America* 101 Suppl, no. suppl 1 (April 6, 2004): 5214–19.
9. Samuel P Huntington, "Democracy's Third Wave," *Journal of Democracy* 2, no. 2 (1991): 12–34; Steven Levitsky and Lucan Way, "The Rise of Competitive Authoritarianism," *Journal of Democracy* 13, no. 2 (2002): 51–65; Andreas Schedler, "The Nested Game of Democratization by Elections," *International Political Science Review* 23, no. 1 (January 2002): 103–22.
10. Sarah Birch and David Muchlinski, "Electoral Violence Prevention: What Works?," *Democratization* 25, no. 3 (April 3, 2018): 385–403; Norris, Frank, and Martínez i Coma, *Contentious Elections: From Ballots to Barricades*.
11. Arend Lijphart, *Democracy in Plural Societies : A Comparative Exploration* (New Haven: Yale University Press, 1977); Juan J Linz, "Transitions to Democracy," *The Washington Quarterly* 13, no. 3 (September 1990): 143–64; Timothy D Sisk, "Conclusions and Recommendations," in *Elections and Conflict Management in Africa*, ed. Timothy D Sisk and Andrew Reynolds (Washington: United States Institute of Peace Press, 1998), 145–71.
12. Donald L Horowitz, *Ethnic Groups in Conflict* (Berkeley: University of California Press, 1985); Steven Wilkinson, *Votes and Violence Electoral Competition and Ethnic Riots in India* (Cambridge; New York: Cambridge University Press, 2004).

13. John B Londregan and Keith T Poole, "Poverty, the Coup Trap, and the Seizure of Executive Power," *World Politics* 42, no. 02 (June 2011): 151–83; Paul Collier, *Wars, Guns, and Votes : Democracy in Dangerous Places* (New York: Harper, 2009).
14. "Perils of Pluralism: Electoral Violence and Competitive Authoritarianism in Sub-Saharan Africa," *Working Paper, Department of Political Science, University of Wisconsin - Madison*, 2012.
15. cf. Patrick M Kuhn, "Do Contentious Elections Trigger Violence?," in *Contentious Elections: From Ballots to Barricades*, ed. Pippa Norris, Richard W Frank, and Ferran Martínez i Coma (New York: Routledge, 2015).
16. "Election Violence in Democratizing States," *APSA 2011 Annual Meeting Paper*, 2012, pscourses.ucsd.edu/poli120n/ArriolaJohnson2012.pdf.
17. "When Do Governments Resort to Election Violence?," *British Journal of Political Science* 44, no. 01 (February 2013): 149–79.
18. Sandra Ley, "To Vote or Not to Vote," *Journal of Conflict Resolution*, May 22, 2017, 002200271770860.
19. Susan Hyde and Nikolay Marinov, "NELDA - National Elections Across Democracy and Autocracy," 2015, http://www.nelda.co/.
20. Michael Coppedge et al., "Measuring High Level Democratic Principles Using the V-Dem Data," *International Political Science Review* 37, no. 5 (2015): 580–93.
21. Birch and Muchlinski, "The Dataset of Countries at Risk of Electoral Violence."
22. Birch and Muchlinski.
23. Sven-Oliver Proksch and Jonathan B. Slapin, *The Politics of Parliamentary Debate: Parties, Rebels and Representation* (Cambridge: Cambridge University Press, 2015); Julio Cesar Amador Diaz Lopez et al., "Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data," *Statistics, Politics and Policy* 8, no. 1 (January 26, 2017): 85–104; David Nicolas Hopmann et al., "Effects of Election News Coverage: How Visibility and Tone Influence Party Choice," *Political Communication* 27, no. 4 (October 29, 2010): 389–405.
24. 206 parliamentary elections, 17 general elections, 103 presidential elections, 10 referendums, 46 early elections (parliamentary and presidential combined), 3 repeated elections (parliamentary and presidential combined), 5 federal elections, 18 local elections, 12 municipal elections, 2 provincial elections.
25. Hadley Wickham, "Package 'Rvest,'" 2016, https://github.com/hadley/rvest.
26. Milan Straka, Jan Hajič, and Jana Straková, "UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing," Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (Portorož, 2015).
27. Ingo Feinerer, Kurt Hornik, and David Meyer, "Text Mining Infrastructure in R," *Journal of Statistical Software* 25, no. 5 (March 31, 2008): 1–54.
28. Fridolin Wild, "Package Lsa," 2015, https://cran.r-project.org/package=lsa.
29. Thomas K. Landauer and Susan Dumais, "Latent Semantic Analysis," *Scholarpedia* 3, no. 11 (November 13, 2008): 4356.
30. Susan T. Dumais, "Latent Semantic Analysis," *Annual Review of Information Science and Technology* 38, no. 1 (September 22, 2005): 191.
31. Dumais, 191.
32. the four steps procedure was adopted from Dumais, 192–93.
33. Wild, "Package Lsa"; Feinerer, Hornik, and Meyer, "Text Mining Infrastructure in R."
34. Sébastien Heymann, "Gephi," in *Encyclopedia of Social Network Analysis and Mining*, ed. Reda Alhajj and jon Rokne (New York: Springer-Verlag, 2014), 612–25.

35. OSCE/ODIHR, “Election Observation Handbook: Sixth Edition” (OSCE/ODIHR, 2010).
36. Straka, Hajič, and Straková, “UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing.”
37. Wild, “Package Lsa”; adopted from Scott Deerwester et al., “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science* 41, no. 6 (September 1, 1990): 391–407.
38. Modularity measures how well a network decomposes into modular communities. A high modularity score indicates sophisticated internal structure. This structure, often called a community structure, describes how the network is compartmentalized into sub-networks.
39. Michael Barkun, “President Trump and the ‘Fringe,’” *Terrorism and Political Violence* 29, no. 3 (May 4, 2017): 437–43.
40. Joan Barceló, “Batons and Ballots: The Effectiveness of State Violence in Fighting against Catalan Separatism,” *Research & Politics* 5, no. 2 (April 25, 2018): 205316801878174; Emilie M. Hafner-Burton, Susan D. Hyde, and Ryan S. Jablonski, “Surviving Elections: Election Violence, Incumbent Victory and Post-Election Repercussions,” *British Journal of Political Science* 48, no. 02 (April 19, 2018): 459–88.

References

Amador Diaz Lopez, Julio Cesar, Sofia Collignon-Delmar, Kenneth Benoit, and Akitaka Matsuo. “Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data.” *Statistics, Politics and Policy* 8, no. 1 (January 26, 2017): 85–104.

Arriola, Leonardo R, and Chelsea Johnson. “Election Violence in Democratizing States.” *APSA 2011 Annual Meeting Paper*, 2012. pscourses.ucsd.edu/poli120n/ArriolaJohnson2012.pdf.

Barceló, Joan. “Batons and Ballots: The Effectiveness of State Violence in Fighting against Catalan Separatism.” *Research & Politics* 5, no. 2 (April 25, 2018): 205316801878174.

Barkun, Michael. “President Trump and the ‘Fringe.’” *Terrorism and Political Violence* 29, no. 3 (May 4, 2017): 437–43.

Bekoe, Dorina. *Voting in Fear: Electoral Violence in Sub-Saharan Africa*. Edited by Dorina Bekoe. Washington D.C.: United States Institute of Peace, 2012.

Birch, Sarah, and David Muchlinski. “Electoral Violence Prevention: What Works?” *Democratization* 25, no. 3 (April 3, 2018): 385–403.

———. “The Dataset of Countries at Risk of Electoral Violence.” *Terrorism and Political Violence*, September 26, 2017, 1–20.

Bunce, Valerie, and Sharon L Wolchik. *Defeating Authoritarian Leaders in Postcommunist Countries*. Cambridge: Cambridge University Press, 2011.

Collier, Paul. *Wars, Guns, and Votes : Democracy in Dangerous Places*. New York: Harper, 2009.

Coppedge, Michael, Staffan Lindberg, Svend Erik Skaaning, and Jan Teorell. “Measuring High Level Democratic Principles Using the V-Dem Data.” *International Political Science Review* 37, no. 5 (2015): 580–93.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. “Indexing by Latent Semantic Analysis.” *Journal of the American Society for Information Science* 41, no. 6 (September 1, 1990): 391–407.

Diamond, Larry Jay, Marc F Plattner, and Christopher Walker. *Authoritarianism Goes Global: The Challenge to Democracy*. Baltimore: John Hopkins Universtiy Press, 2016.

Dumais, Susan T. “Latent Semantic Analysis.” *Annual Review of Information Science and Technology* 38, no. 1 (September 22, 2005): 188–230.

Dunning, Thad. “Fighting and Voting: Violent Conflict and Electoral Politics.” *Journal of Conflict Resolution* 55, no. 3 (2011): 327–39.

Feinerer, Ingo, Kurt Hornik, and David Meyer. “Text Mining Infrastructure in R.” *Journal of Statistical Software* 25, no. 5 (March 31, 2008): 1–54.

Fjelde, Hanne, and Kristine Höglund. “Electoral Institutions and Electoral Violence in Sub-Saharan Africa.” *British Journal of Political Science* FirstView, no. Supplement-1 (2015): 1–24.

Hafner-Burton, Emilie M., Susan D. Hyde, and Ryan S. Jablonski. “Surviving Elections: Election Violence, Incumbent Victory and Post-Election Repercussions.” *British Journal of Political Science* 48, no. 02 (April 19, 2018): 459–88.

Hafner-Burton, Emilie M, Susan D Hyde, and Ryan S Jablonski. “When Do Governments Resort to Election Violence?” *British Journal of Political Science* 44, no. 01 (February 2013): 149–79.

Hall, Stephen G. F., and Thomas Ambrosio. “Authoritarian Learning: A Conceptual Overview.” *East European Politics* 33, no. 2 (April 3, 2017): 143–61.

Heymann, Sébastien. “Gephi.” In *Encyclopedia of Social Network Analysis and Mining*,

edited by Reda Alhajj and jon Rokne, 612–25. New York: Springer-Verlag, 2014.
Höglund, Kristine. “Electoral Violence in Conflict-Ridden Societies: Concepts, Causes, and Consequences.” *Terrorism and Political Violence* 21, no. 3 (2009): 412–27.
Hopmann, David Nicolas, Rens Vliegenthart, Claes De Vreese, and Erik Albæk. “Effects of Election News Coverage: How Visibility and Tone Influence Party Choice.” *Political Communication* 27, no. 4 (October 29, 2010): 389–405.
Horowitz, Donald L. *Ethnic Groups in Conflict*. Berkeley: University of California Press, 1985.
Huntington, Samuel P. “Democracy’s Third Wave.” *Journal of Democracy* 2, no. 2 (1991): 12–34.
Hyde, Susan, and Nikolay Marinov. “NELDA - National Elections Across Democracy and Autocracy,” 2015. http://www.nelda.co/.
Kuhn, Patrick M. “Do Contentious Elections Trigger Violence?” In *Contentious Elections: From Ballots to Barricades*, edited by Pippa Norris, Richard W Frank, and Ferran Martínez i Coma. New York: Routledge, 2015.
Landauer, Thomas K., and Susan Dumais. “Latent Semantic Analysis.” *Scholarpedia* 3, no. 11 (November 13, 2008): 4356.
Landauer, Thomas K, Darrell Laham, and Marcia Derr. “From Paragraph to Graph: Latent Semantic Analysis for Information Visualization.” *Proceedings of the National Academy of Sciences of the United States of America* 101 Suppl, no. suppl 1 (April 6, 2004): 5214–19.
Levitsky, Steven, and Lucan Way. “The Rise of Competitive Authoritarianism.” *Journal of Democracy* 13, no. 2 (2002): 51–65.
Ley, Sandra. “To Vote or Not to Vote.” *Journal of Conflict Resolution*, May 22, 2017, 002200271770860.
Lijphart, Arend. *Democracy in Plural Societies : A Comparative Exploration*. New Haven: Yale University Press, 1977.
Linz, Juan J. “Transitions to Democracy.” *The Washington Quarterly* 13, no. 3 (September 1990): 143–64.
Londregan, John B, and Keith T Poole. “Poverty, the Coup Trap, and the Seizure of Executive Power.” *World Politics* 42, no. 02 (June 2011): 151–83.
Norris, Pippa, Richard W Frank, and Ferran Martínez i Coma. *Contentious Elections: From Ballots to Barricades*. New York: Routledge, 2015.
OSCE/ODIHR. “Election Observation Handbook: Sixth Edition.” OSCE/ODIHR, 2010.
Pevehouse, Jon C, Scott Straus, and Charles Taylor. “Perils of Pluralism: Electoral Violence and Competitive Authoritarianism in Sub-Saharan Africa.” *Working Paper, Department of Political Science, University of Wisconsin - Madison*, 2012.
Proksch, Sven-Oliver, and Jonathan B. Slapin. *The Politics of Parliamentary Debate: Parties, Rebels and Representation*. Cambridge: Cambridge University Press, 2015.
Rapoport, David, and Leonard Weinberg. *The Democratic Experience and Political Violence*. Edited by David Rapoport and Leonard Weinberg. London: F. Cass, 2001.
Schedler, Andreas. “The Nested Game of Democratization by Elections.” *International Political Science Review* 23, no. 1 (January 2002): 103–22.
Sisk, Timothy D. “Conclusions and Recommendations.” In *Elections and Conflict Management in Africa*, edited by Timothy D Sisk and Andrew Reynolds, 145–71. Washington: United States Institute of Peace Press, 1998.
Straka, Milan, Jan Hajič, and Jana Straková. “UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing.” Proceedings of the Tenth International Conference on Language Resources

and Evaluation (LREC 2016). Portorož, 2015.
Wahman, Michael. “Democratization and Electoral Turnovers in Sub-Saharan Africa and Beyond.” *Democratization* 21, no. 2 (2014): 220–43.
Wickham, Hadley. “Package ‘Rvest,’” 2016. https://github.com/hadley/rvest.
Wild, Fridolin. “Package Lsa,” 2015. https://cran.r-project.org/package=lsa.
Wilkinson, Steven. *Votes and Violence Electoral Competition and Ethnic Riots in India*. Cambridge; New York: Cambridge University Press, 2004.

Figure 1. Visualization of singular-value decomposition.

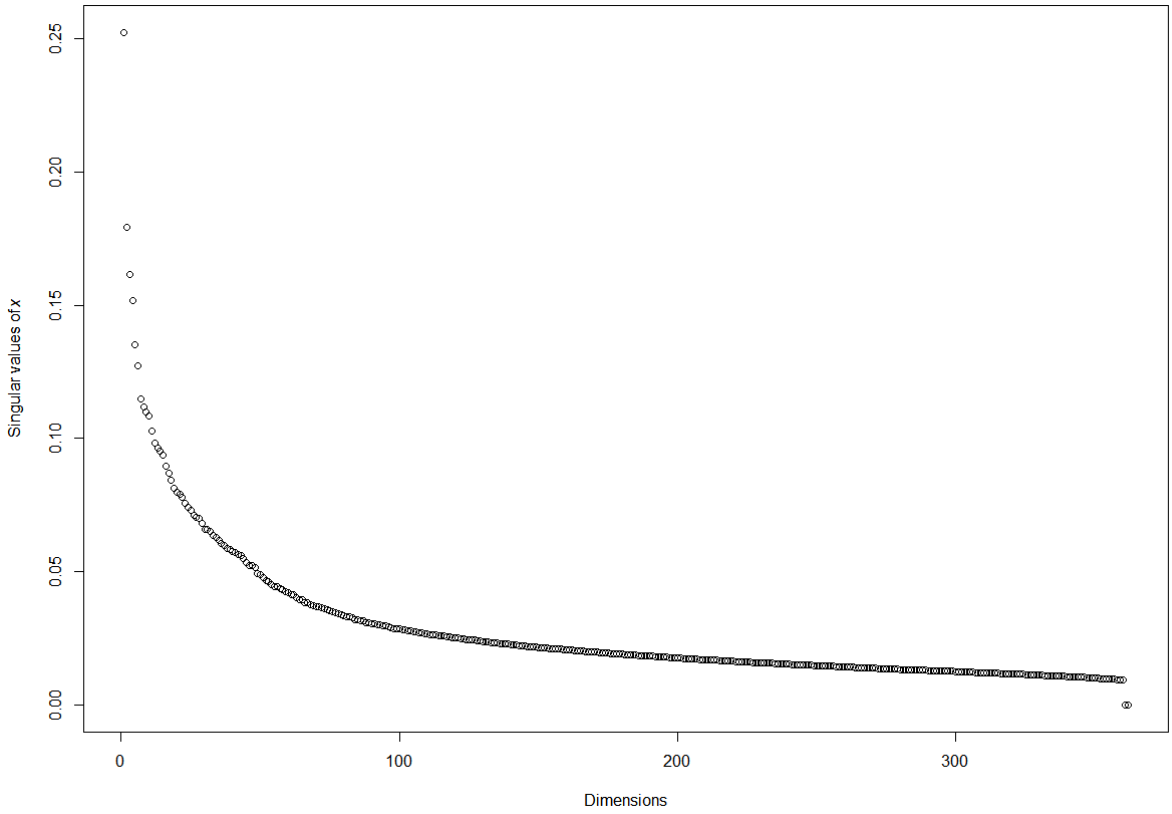

Figure 2. Visualization of network of co-occurring unigram-keywords

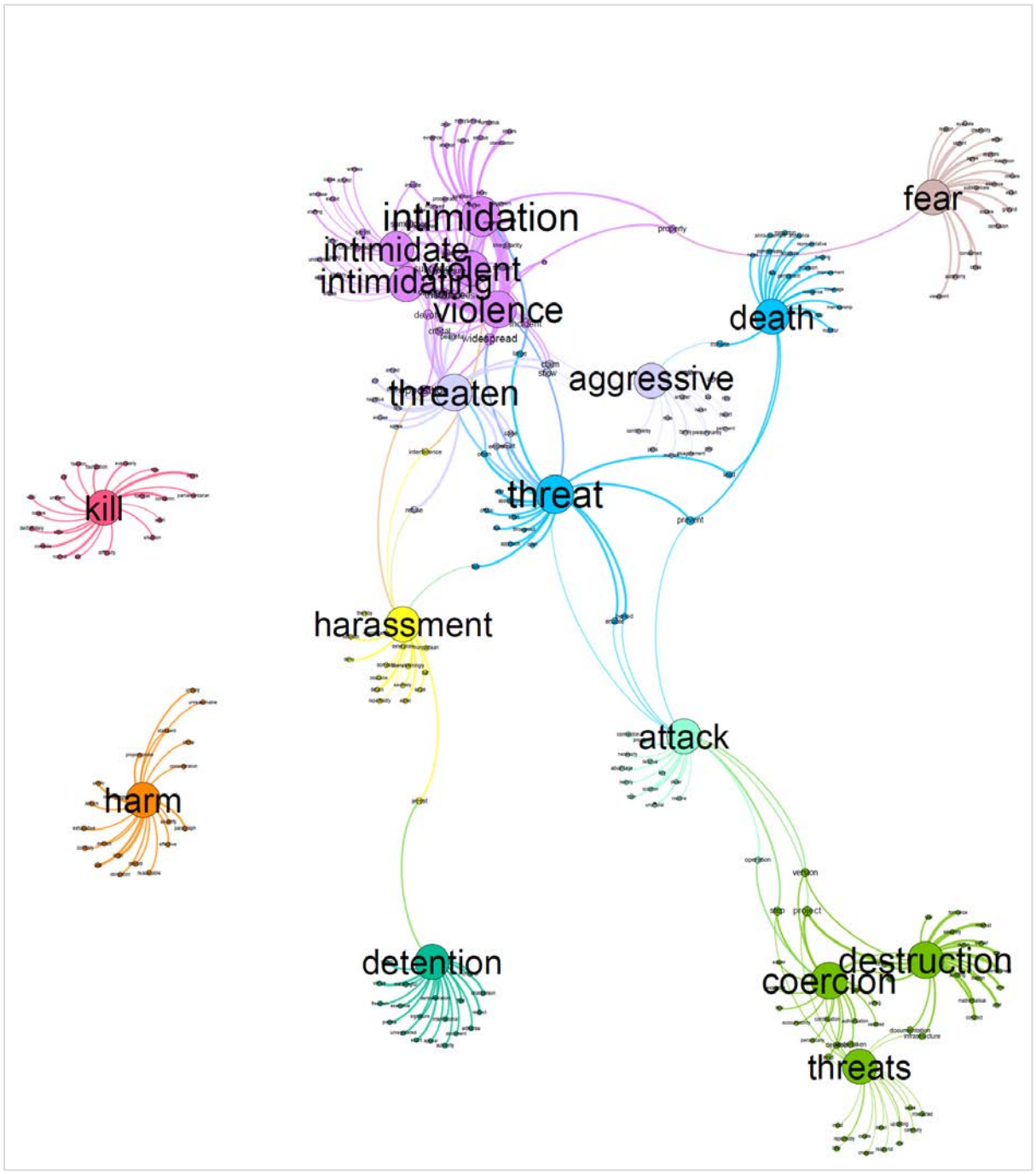

Figure 3. Post-soviet region

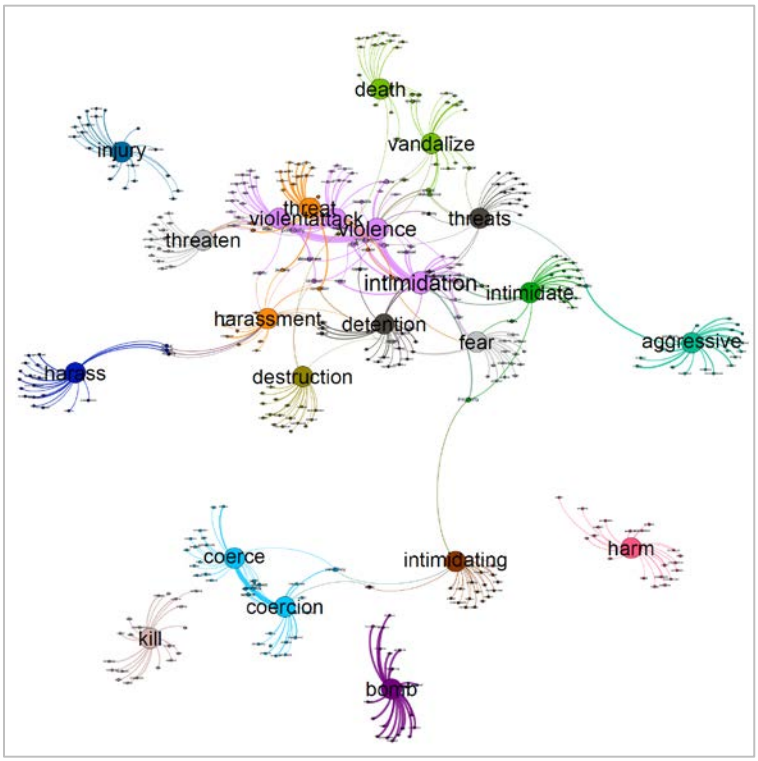

Figure 4. Western Balkans

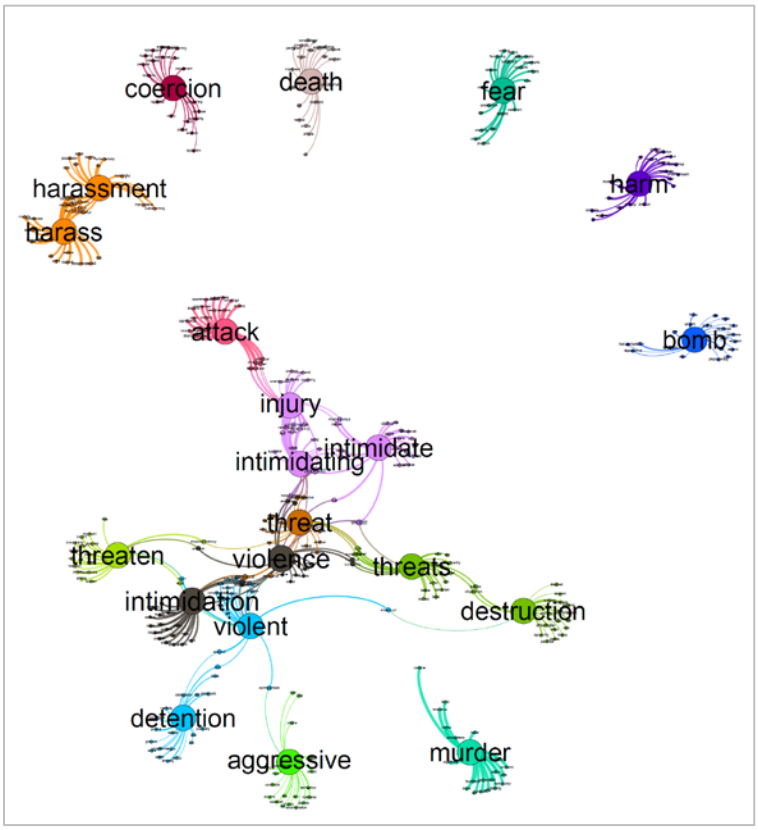

Figure 5. Western Democracies

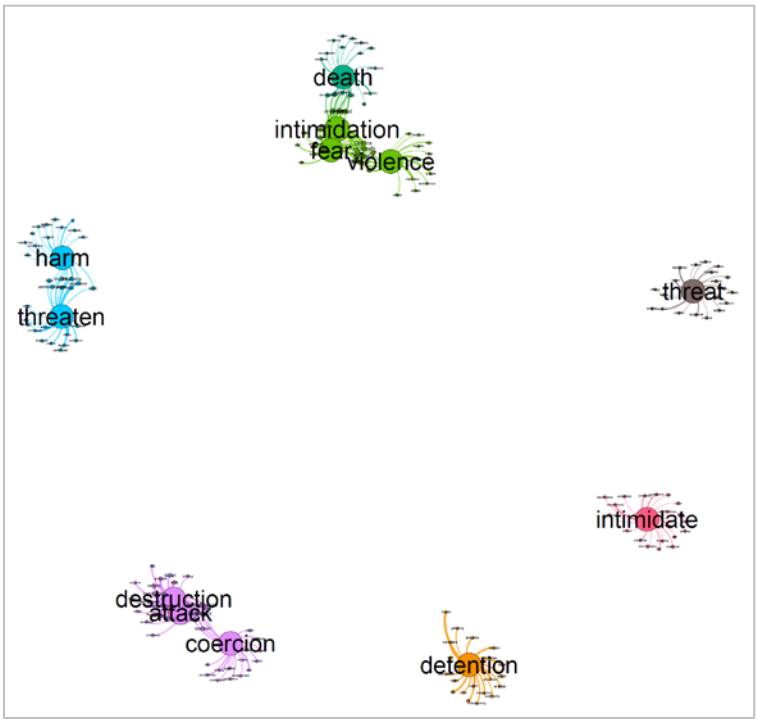

Table 1. Summary of ten most used unigrams, bigrams, trigrams, and 4-grams

Unigram	*wordfreq*	*docfreq*	*Bigram*	*wordfreq*	*docfreq*
election	58879	362	osce_odihr	13633	351
party	28331	361	polling_station	9698	352
voter	25601	362	per_cent	9066	272
candidate	22093	357	political_party	8100	356
osce	21187	362	election_day	7115	354
campaign	17227	359	voter_list	4191	317
odihr	15768	360	election_commission	3540	297
political	15262	361	human_rights	3145	350
law	14772	362	odihr_eom	3058	127
vote	14575	362	election_administration	2801	346

Trigram	*wordfreq*	*docfreq*	*4-gram*	*wordfreq*	*docfreq*
osce_odihr_eom	2882	120	paragraph_osce_copenhagen_document	997	209
osce_odihr_eam	1468	67	per_cent_polling_station	767	114
osce_copenhagen_document	1148	248	office_democratic_institutions_human	701	339
paragraph_osce_copenhagen	1003	210	democratic_institutions_human_rights	701	339
election_observation_mission	977	224	human_rights_osce_odihr	527	254
osce_odihr_leom	962	33	institutions_human_rights_osce	523	253
polling_station_visit	886	206	osce_commitment_international_standard	498	187
osce_odihr_nam	877	59	osce_odihr_nam_interlocutor	460	36
osce_commitment_international	871	222	osce_office_democratic_institutions	411	278
per_cent_polling	783	116	line_osce_commitment_international	398	193

Table 2. Summary of ten most used violence-related unigrams, bigrams, trigrams, and 4-grams

Unigram	*wordfreq*	*docfreq*	*Bigram*	*wordfreq*	*docfreq*
intimidation	761	205	pressure_intimidation	94	46
violence	362	138	violent_incident	94	53
attack	359	140	pretrial_detention	86	46
detention	291	131	intimidation_voter	84	56
fear	226	105	fear_retribution	80	39
threat	225	120	free_fear	59	38
violent	171	84	violence_intimidation	59	48
threaten	147	93	voter_intimidation	50	30
intimidate	101	55	detention_centre	50	38
harassment	97	55	incident_violence	43	32

Trigram	*wordfreq*	*docfreq*	*4-gram*	*wordfreq*	*docfreq*
free_fear_retribution	57	36	vote_free_fear_retribution	55	35
vote_free_fear	55	35	administrative_action_violence_intimidation	29	29
administrative_action_violence	29	29	neither_administrative_action_violence	28	28
action_violence_intimidation	29	29	cast_vote_free_fear	27	16
violence_intimidation_bar	25	25	caste_vote_free_fear	27	27
pretrial_detention_centre	23	17	action_violence_intimidation_bar	25	25
fear_retribution_require	23	14	violence_intimidation_bar_party	23	23
intimidation_bar_party	23	23	free_fear_retribution_require	22	13
			intimidation_bar_party_candidate	22	22
			fear_retribution_require_paragraph	20	14

Note: Counting algorithm identified only eight trigrams occurring at least 20 times in the corpus.

Table 3. Cosine distances of the co-occurring words of the five most frequent unigrams

intimidation	1.000	violence	1.000	attack	1.000	detention	1.000	fear	1.000
violence	0.890	intimidation	0.890	location	0.788	venue	0.815	confusion	0.833
irregularity	0.878	violent	0.828	step	0.766	restrict	0.799	ground	0.812
evidence	0.870	pressure	0.809	unofficial	0.744	executive	0.794	substantiate	0.811
observer	0.867	bad	0.797	prominent	0.718	event	0.791	concerned	0.810
instances	0.859	instances	0.793	deter	0.717	chairperson	0.787	autonomy	0.809

Table 4. Summary of ten most used unigrams, bigrams, trigrams, and 4-grams per region

	Post-soviet countries	*Western Balkans*	*Western World*
Unigrams	election (16689) candidate (6969) cec (6921) voter (6747) party (6294)	election (13272) party (7099) voter (5891) candidate (4242) osce (4155)	election (11052) party (6188) voter (5442) osce (4645) candidate (3891)
Bigrams	osce_odihr (3473) polling_station (3062) per_cent (3043) election_day (2232) political_party (1728)	osce_odihr (2470) polling_station (2330) per_cent (2324) political_party (1755) election_day (1485)	osce_odihr (3304) political_party (1953) polling_station (1332) election_day (1285) per_cent (1042)
Trigrams	osce_odihr_eom (1338) election_observation_mission (395) per_cent_polling (384) cent_polling_station (382) central_election_commission (346)	osce_odihr_eom (737) election_observation_mission (307) osce_odihr_leom (266) per_cent_polling (190) osce_copenhagen_document (190)	osce_odihr_eam (791) osce_odihr_nam (578) odihr_nam_interlocutor (332) osce_copenhagen_document (246) paragraph_osce_copenhagen (219)
4-grams	per_cent_polling_station (382) paragraph_osce_copenhagen_document (307) office_democratic_institutions_human (175) democratic_institutions_human_rights (175) osce_odihr_eom_observer (154)	per_cent_polling_station (182) paragraph_osce_copenhagen_document (160) office_democratic_institutions_human (140) democratic_institutions_human_rights (140) osce_commitment_international_standard (115)	osce_odihr_nam_interlocutor (301) paragraph_osce_copenhagen_document (219) osce_odihr_eam_interlocutor (211) office_democratic_institutions_human (169) democratic_institutions_human_rights (169)