The Cambodian Genocide

Text Mining of Transcripts from the ECCC
for Downstream Atrocity Prevention

William Bange, Kyle Braman, Donovan Dutcher, Nate Plummer
Binghamton University

Abstract

In this study, TF-IDF analysis and Top2Vec topic modeling programs were built to analyze the transcripts from the ECCC genocide tribunal,, gathered through the Genocide Transcript Corpus assembled by Miriam Schirmer. Through these methods, word counts were created from both the testimony of Kaing Guev Eav and witness statements. TF-IDF for Eav's testimony was enhanced via comparison with the testimony of Zdravko Tolimir present in the ICTY transcripts. This was used to highlight words which correlated to Eav's admission of guilt versus Tolimmir, who denied responsibility. Witness statements were used to show that trauma related words were common enough to drop out if TF-IDF results across the corpus. This was a further reinforcement of the truth of Eav's testimony. Top2Vec was utilized to organize documents based on similarity and generate topics containing lists of words that were deemed to fit those in the document clusters. This allowed for determination of broad themes in relation to the types of language that was present across the ECCC corpus.

1. Introduction and Background

Quantitative text analysis provides a powerful tool for downstream prevention of genocide and mass atrocity. In this study, the word counting measure "TF-IDF" (Text Frequency × Inverse Document Frequency) and the Top2Vec algorithm were used to examine line-by-line transcripts drawn from hybrid genocide tribunals in the *Extraordinary Chambers in the Courts of Cambodia* (ECCC). These methods allowed for the extraction of common words from testimonies of the witnesses and accused for comparison as well as algorithmic grouping of trial data by generated topic. Implications and possible applications of these methods for prevention and reconciliation of genocide and mass atrocity will be discussed.

From 1975 to 1979 the Khmer Rouge, a radical communist movement, held control of the nation of Cambodia. This period was characterized by violence against intellectuals and religious (Muslims, Buddhists, Christians) and ethnic (Chinese, Vietnamese) minorities. During this period, 1.5 to 3 million people died from a combination of famine, disease and mass killing. Violence is not born in a vacuum, and the Khmer regime, while bearing the majority of the responsibility, was itself a product of history.

The Cambodia of the early 1970's existed in the shadow of French colonial rule, occupation by the Japanese during World War II, and widespread bombings during the Vietnam War. As noted in *Fundamentals of Mass Atrocity Prevention*, "...the strongest macro-level indicator of the onset of genocide and mass killing is the presence of large-scale instability." (Straus, 2016). This extensive history of violence and oppression culminated in a civil war from which the Khmer Rouge emerged victorious.

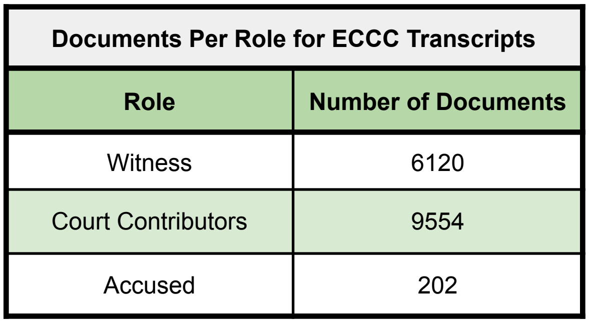
Following its ascent to power, the regime enacted policies to shape Cambodia into a utopic, atheist and agrarian socialist state. (*Khmer Rouge Ideology*, 2024) These strong ideologies to which the regime subscribed primed them for the violence that was to follow. Citizens were forcibly removed from cities and made to work in agricultural settings in which they had no experience with the goal of securing a self-sufficient Cambodia. Dissenters, intellectuals, and those who did not conform to the atheist worldview of the Khmer Rouge were imprisoned in camps and ultimately executed en masse, earning the resulting burial grounds the apt moniker "the Killing Fields." (Witness History)

Justice for the atrocities committed in Cambodia would not begin to be realized until the 1990's. At this time, the People's Republic of Kampuchea (PRK) signaled interest in holding trials clear to the United Nations, motivated by the advancing age of both witnesses and perpetrators of the violence. Historically, these trials had been conducted in international courts, but the Cambodian government was insistent that the accused be tried in their own courts. This led to a political standstill until 2006, when the trials were finally organized and able to commence (*Timeline of ECCC Events,* 2024). Three individuals stood trial: Kaing Guev Eav (also known as Comrade Duch), Nuon Chea and Khieu Samphan. Kaing faced charges of war crimes and crimes against humanity. The latter two, Chea and Samphan, stood trial together and were later charged with the additional crime of genocide against the Cham muslim religious minority and Vietnamese prisoners. Of the three, Kaing alone admitted wrongdoing. He was found guilty in 2010 and sentenced to life imprisonment upon an appeal of his conviction, and died in prison. The initial trial for Chea and Samphan resulted in sentences of life imprisonment for their crimes in 2014. Both were later tried for genocide, with Chea dying before a verdict was able to be reached. In 2018, Samphan was found guilty of genocide against the Vietnamese; a verdict that was seen as the closing chapter in legal efforts against the Khmer Rouge.

2. Data

Data for the text mining of this report is from the Genocide Transcript Corpus (GTC), a dataset constructed using web scraping to create an organized collection of the Cambodian trial transcripts. This is public data constructed by an external researcher, Miriam Schirmer.

The Dataset contains transcripts from the ECCC (*Extraordinary Chambers in the Courts of Cambodia)*, the ICTR (*International Criminal Tribunal for Rwanda)*, and the ICTY (*International Criminal Tribunal for the Former Yugoslavia*). These sets contain per-row information on the date of a case, who's being accused, if a case is trauma related or not, the role of a person speaking, what's being said, along with file information for storage and reference of court documents. The dataset contains 52,485 segments of text from 90 different transcripts of criminals and witnesses involved in the three tribunal. Of these, 15,876 documents pertain specifically to Cambodia. Within that number, 202 documents were testimony specific to Kaing, representing a significant minority of the overall data for analysis. The witness testimony was the single largest source of documents in the dataset with 6120 entries and constituted the bulk of the text analysis performed.



Documents Per Role for ECCC Transcripts

Role	Number of Documents
Witness	6120
Court Contributors	9554
Accused	202

Fig 1. This chart shows the documents per role within the ECCC transcripts

3. Results and Discussion

3.1 Word Counting - Kaing Guev Eav

To better understand what Kaing Guev Eav said during his trial, we used TF-IDF on the ICTY tribunal to ascertain the most significant words spoken in that tribuna for comparison with Kaingl. The testimony from Zdravko Tolimir, a convicted war criminal of the Bosnian Genocide was used as a reference because Tolimir pleaded not guilty to his crimes and denied the fact that Sbrencia was a civilian massacre, which would show up in his testimony. (*Bosnian Serb Zdravko Tolimir Convicted Over Srebrenica*, 2012).

From this, we could determine if there were words specific to Kaing's testimony that showed his acceptance and admittance of guilt, as those words would only be found in Kaing's testimony, not Tolimir's. After sorting the dataset to only include testimonies from when either Kaing or Tolimir spoke, we calculated the TF-IDF values for each word. Introduction of testimony from Tolimir increased the total number of documents used for the TF-IDF analysis. This served the purpose of pushing unique words in Kaing's testimony higher in their TF-IDF value by decreasing their document frequency. After this 'weighting' had been performed, the data was then filtered to only see the words with the highest TF-IDF spoken by Kaing, displaying words that are highly frequent in Kaing's testimony and not frequent in others.

Resulting from this analysis, many words were identified that Kaing used in his testimony that provide insight to the horrors that went on at S-21 and M-13, prison camps run during the Khmer Rouge regime infamous for their brutality, along with words describing his own actions at the prisons. Words like “arrested”, “confession”, “interrogation”, “detainee”, and “biography” refer to people who were imprisoned at S-21 where they were interrogated, beaten, and tortured until they confessed their opposition or “crimes” against the Khmer Rouge (*S-21, Tuol Sleng*, n.d., Mydans, 2020). Other words like “died”, “blood”, and “smashed” are all indicative of the death and destruction of the prisoners, which detail the extremity of the killings that occurred. “Smashed” has even more significance as the word was used by Kaing and the Khmer Rouge when ordering someone to be executed (Rueters, 2009).

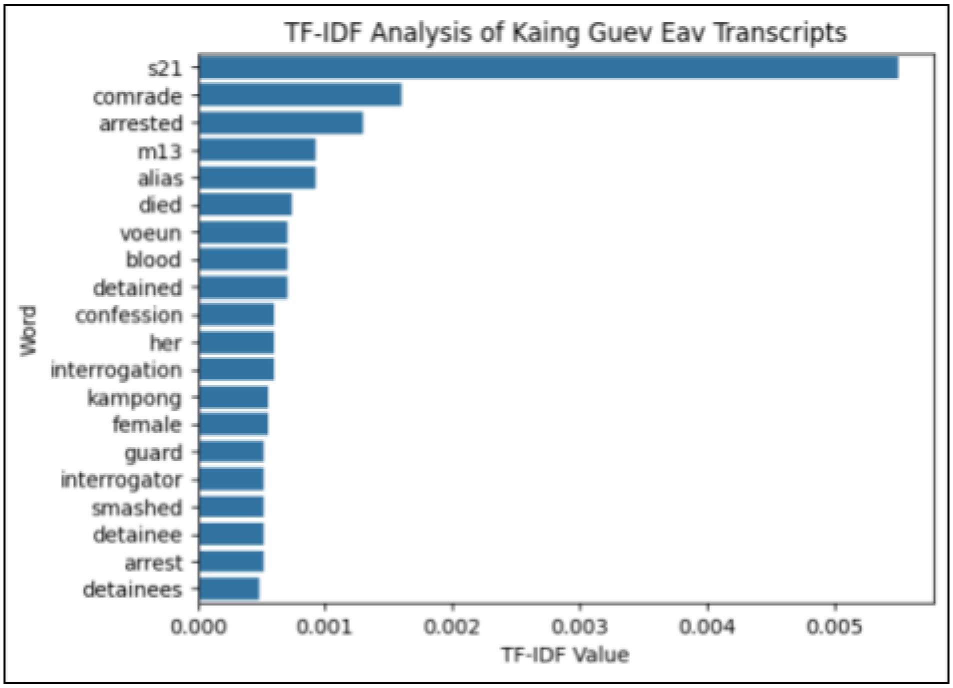


Fig 2. This plot shows the top 20 words in terms of TF-IDF score as spoken by Kaing in his testimony, after the removal of names.

Additionally, many specific names appeared with high TF-IDF, showing specific people of interest in Kaing’s testimony alone. While there were many among these, a very frequent name, “Chan”, highlighted an interesting part of Kaing’s testimony. Chan Vouen was a guard at the M-13 prison who provided testimony to the tribunals; however, in Kaing’s testimony he refuted the fact that Chan Vouen was ever a part of M-13 at all. Strangely, in many of these cases Kaing did not refute the actual crime itself that Chan testified for, often claiming a worse offense occurred instead. It is possible that Kaing wished to deny involvement of others to face conviction alone, but additional research would be necessary to fully confirm this theory. However, we can look for this behavior in other testimonies to make similar findings.

Using the ICTY as a reference helped with these words to become apparent in our analysis of the ECCC and Kaing’s testimony. While these events may have occurred in Bosnia, the strong words associated with them were not used by Tolimir, who did not admit to anything. So, the TF-IDF analysis shows us that not only did he admit to the horrors that occurred under his control, he addressed them in detail. By using meaningful words that are directly related to the action that occurred at S-21, he takes direct ownership of them. If Kaing were trying to diminish, or even lie about these events, this strong and accurate language would not be as apparent in terms of TF-IDF. Our analysis could be applied to future studies of genocide in the sense that the more frequent words associated with genocides or mass atrocities are used by an accused, the more ownership they take of their involvment in the events.

3.2 Word Counting - Witnesses

After completing a TF-IDF analysis of the accused and their testimony, it would make sense to do the same for the witnesses in the case. The same grouping metrics were applied as the previous section. The data was cut down to only include the testimonies of the witnesses from the ICTY and ECCC. We included the ICTY dataset as a reference to be able to make the same comparisons as we did in the previous section.

After sorting the results of the TF-IDF analysis in order of highest to lowest, we filtered for only the words spoken in the ECCC witness statements.

After reviewing the words, it was observed that an overwhelming number of them were names of people involved with the Khmer Rouge, aside location names. Some of these names include “Duch”, “Chan” (as mentioned before,) “Pol”, “Chea”, and “Nuon”. Duch is the short version of Comrade Duch, the nickname of Kaing Guev Eav. Chan Voeun is the same guard referenced earlier, who was also a witness in this trial. Pol is referencing Pol Pot, the leader of the Khmer Rouge. Chea Nuon is another high ranking Khmer Rouge leader being tried in this trial. Location names include “Phnom Penh”, the capital city of Cambodia where S-21 is located, and “Tuol Sleng”, an alternate name for the S-21 prison.

These are just some of the top names that occur frequently in the documents, and are specific to the documents of the ECCC, which reinforces our analysis as these people were not involved in the Bosnian Genocide. Unlike the TF-IDF analysis of Kaing, removing these names did not help to show more significant information as each testimony centered on a specific figure would cause all high TF-IDF words to just be more names and Cambodia-specific details.

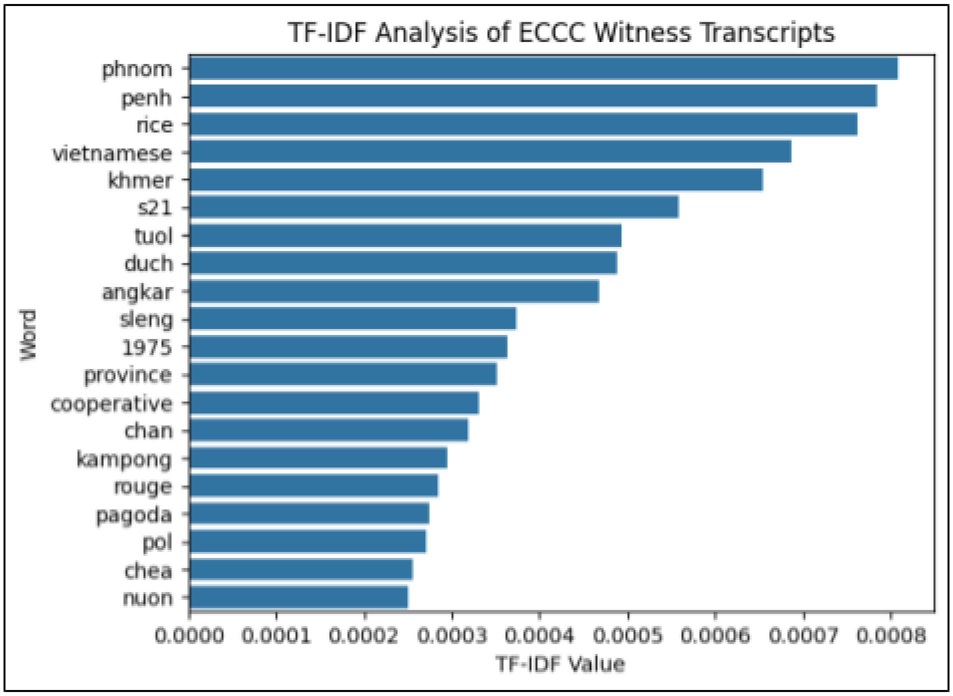


Fig 3. This plot displays the top 20 words from ECCC witness testimony in order of TF-IDF score.

What is important here is not the words we found are uninteresting, but the fact that there are violent words missing from the top of the TF-IDF analysis. This shows that while the violent words may have been frequent in the text, they were also frequent amongst the ICTY testimonies, giving them a high document frequency (and thus low inverse document frequency) in our TF-IDF calculation. Witnesses and victims to crimes tend to tell the truth in detail about what they experienced, so violent words, regardless of genocide should have a low TF-IDF score. If we compare the fact that these violent words did show up in Kaing's testimony, it further shows that of the accused, he was the only one to admit and own up to his crimes, and not avoid them. This avenue of analysis, where we look at the words that are not there, and compare them to other documents, can be useful in further studies of genocides. It can allow for a new way to sort through the vast amounts of information given by witnesses, which is going to be more detailed and lengthy in comparison to the information given by the accused. Using this new approach in conjunction with the analysis of the accused could be vital to ascertaining conclusions about the trials themselves.

3.3 Top2Vec: Topic Modeling

We performed topic modeling with Top2Vec on the witness testimonies from the ECCC tribunal. What follows are several topics with names assigned by the programmer. Topic words were generated using Top2Vec, which identified words it deemed most similar to those found within vectorized documents. These documents were spatially clustered based on perceived similarity by Top2Vec. Example documents for each topic are also displayed to give context as to what the program was considering to be items that fit into the respective topics.

A total of 63 topics were generated, with the 5 documents displayed in the table being those from topics that the Top2ec analysis found to be most similar to the keywords "beat" and "blood." This type of search is able to be further extended to identify documents which correspond to words of the user's choice. With the myriad of words that could be used, the decision was made to present those topics that indicated trauma experienced by witnesses as downstream prevention is primarily associated with a reckoning of society with the true experiences of survivors in a raw format.

After some testing with a few different sets of keywords, "blood" and "beat" were the ones used in this project. Through these tests, it was found that these 2 words were continuously being represented amongst the topics, and to get some more information, were tested as keywords themselves. Using "blood" and "beat" gave topics that were most consistent with known violent actions and served to reinforce the scope of the horrors that Cambodian citizens, Vietnamese prisoners of war, and members of the Cham muslim ethnic group faced.

The first pertains to Phnom Penh, which is the capital of Cambodia, and witnesses recounted the collapse and rebellion within the capital alongside the Vietnamese Soldiers occupying the city.

Topics and Documents	
Topic Name	**Example Document**
Phnom Penh's collapse and the Vietnamese	"...the first preparation was about the advancement of the Vietnamese troops toward Phnom Penh…"
Torture and Interrogation	"...the first person who tortured me was Sieng after he fed up with beating me up…"
Hospitalized soldiers and citizens	"...I was injured by a-105mm shelling of the Vietnamese; then I was sent to be hospitalized at Wat Totuem…"
Pits for prisoners and dead bodies	"At that time there were four pits and I did not know the number of dead bodies…"
Family Separation	"My two brothers dies in Cambodia… the second is a Sister who lives in France. A third brother Ou Rene died in Cambodia…"

Table 1. Each Topic was created using Top2Vec and manually given a name we believed to summarize the topic best. These topics were created using the keywords "blood", and "beat".

Within the documents under Torture and Interrogation, prisoners of war recounted their experiences while in captivity, even referring to people like Comrade Sieng and Comrade Tith as those taking part in the witness's torture.

The next topic gives witness insight into the hospitalization of different civilians and soldiers. Citizens and soldiers alike would be moved from hospital to hospital depending on different factors such as safety of the hospital, medication needed, and the availability of space within the hospital. The example document is from a witness who did end up moving between hospitals after sustaining injury.

The pits were holes dug into the ground by the Khmer to store prisoners and dead bodies in. Within this topic, witnesses were the surviving prisoners taken by the Khmer, and they discussed walking by these holes. Some documents recounted seeing live prisoners while others recounted seeing a countless number of dead bodies

The last topic in this table contains documents pertaining to the familial struggles from the genocide. Witnesses recounted on how they were separated from their families in one way or another due to the genocide and ongoing wars. Some families would split apart when choosing between fleeing the area or staying, and others would have to deal with the all too common deaths of family members during that time.

Through using Top2Vec, these witness recounts were categorized into coherent groups and allowed us to make more informative conclusions for our final analysis.

4. Final analysis

The ECCC has a lot of important information in the form of testimonies from both sides of the genocide. Both are useful when trying to determine the most important actions that occurred during the genocide, and the reactions of both sides, post genocide.

Using the ICTY as a base reference, we were able to find nuances within Kaing Guev Eav’s testimony. We are able to see how he specifically mentioned violence against the Cambodian people that he himself did or ordered to be done. These actions included interrogating, killing, getting false confessions, and “smashing” enemies. We know that these actions are not specific to Cambodia, due to the very nature of genocides. For them to be specific when only looking at the testimony of the accused in the case of Kaing, demonstrates his understanding, and acknowledgement of his crimes.

Going beyond ascertaining the level of acknowledgment by Kaing in this matter, we can also model different topics regarding the genocide. We used Top2Vec to model topics found within the witness testimony, producing powerful results. Within the vast amounts of death and destruction, with seemingly no direction, the model was able to produce distinct topics that elaborated on the violence in a meaningful way. Topics such as family separation, torture/interogation, and the fall of Phnom Pen were all able to be found within the testimony, allowing for a more digestible and understandable presentation of the events in Cambodia

4. Conclusion

Healing from genocide and mass atrocity is a long and painful journey. Convictions were achieved in the trials, but many of the senior members of the Khmer Rouge never faced justice. The task that is left to survivors is reconciling the events of the past with the lives they must still live today. Downstream prevention of genocide and mass atrocity necessitates a reckoning be had with history. Through quantitative analysis of genocide trials, data is generated which allows for detailed examination of the words of both survivors and perpetrators, and patterns in this data helps shed light on the thoughts of both parties. Still there exist survivors of the genocide in Cambodia whose stories have never been heard. Memorializing and honoring the past through words is a key way in which the damages caused by genocide may be worked through in a healthy way. Text analysis is an important tool in the arsenal of downstream prevention techniques, offering a data-driven method by which language may be examined to provide insight into the complexities of the human experience. It is hoped that, in the future, these types of analyses will be used more widely and proactively to anticipate and possibly avert such atrocities from being committed.Text analysis serves to enhance the human ability to examine language and, as such, even if that future arrives it will still be in human hands to effect true change in the world.

References

Bosnian Serb Zdravko Tolimir convicted over Srebrenica. (2012, December 12). BCC. Retrieved April 12, 2024, from https://www.bbc.com/news/world-europe-20700387

Former Bosnian Serb general Tolimir goes on trial over Srebrenica massacre. (2010, 02 26). France24. Retrieved April 12, 2024, from https://www.france24.com/en/20100226-former-bosnian-serb-general-tolimir-goes-trial-over-srebrenica-massacre

Khmer Rouge ideology. (2024). Holocaust Memorial Day Trust. Retrieved May 8, 2024, from https://www.hmd.org.uk/learn-about-the-holocaust-and-genocides/cambodia/khmer-rouge-ideology/

Ly, R. Persecuting the Khmer Rouge: Views From The Inside. https://www.nurembergacademy.org/fileadmin/user_upload/Cambodia.pdf

Mydans, S. (2020, September 1). *Duch, Prison Chief Who Slaughtered for the Khmer Rouge, Dies at 77 (Published 2020)*. The New York Times. Retrieved April 13, 2024, from https://www.nytimes.com/2020/09/01/world/asia/duch-kaing-guek-eav-dead.html

Rueters. (2009, 04 23). *Khmer Rouge jailer says ordered to "smash" prisoners*. Reuters. Retrieved April 12, 2024, from https://www.reuters.com/article/idUSTRE53M37F/

S-21, Tuol Sleng. United States Holocaust Memorial Museum. Retrieved April 12, 2024, from https://www.ushmm.org/genocide-prevention/countries/cambodia/s-21

Schirmer, M. (2023, June 2). *MiriamSchirmer/genocide-transcript-corpus: Dataset of text sections of genocide-related court transcripts.* GitHub. Retrieved April 12, 2024, from https://github.com/MiriamSchirmer/genocide-transcript-corpus

Straus, S. (2016). *Fundamentals of Genocide and Mass Atrocity Prevention*. United States Holocaust Memorial Museum.

Timeline of ECCC Events. Extraordinary Chambers in the Courts of Cambodia. Retrieved May 8, 2024, from https://www.eccc.gov.kh/en/keyevents

Witness History (Director). *Surviving Cambodia's 'Killing Fields'* [Film; Web]. BBC. https://wspartners.stage.bbc.com/clip/p07frcwg