NLP of UNGA 78th Session Statements

Analysing statements from African countries during the 78th United Nations General Assembly Debate through text mining, employing Natural Language Processing techniques and predefined algorithms.

Table of Contents

List of Tables

List of Figures

1 Manipulating Text

The process begins by loading the speeches' clean data. The data underwent cleaning, stemming, lemmatization, and categorization using the Universal Part-of-Speech (UPOS) system with the assistance of the udpipe R package.

```
# Read your CSV file
db<- read.csv("UNGA_78_clean_corpus.csv")

# Tokenize and annotate the text
annotated_texts <- udpipe_annotate(ud_model, x = db$Cleaned_Text)


# db being existing data frame
db$doc_id <- paste0("doc", 1:nrow(db))
```

1.1 Frequency Tokens

To provide a concise and expeditious analysis of the speech, a tabular representation of word and sentence frequencies and the syntax (code) is presented herein.

```
# Convert 'Tokens' to a data frame with one row per word
word_data <- speeches %>%
  unnest_tokens(word, Cleaned_Text)

# Calculate word count per country
word_count_per_country <- word_data %>%
  count( Country_code, word, sort = TRUE) %>%
  group_by( Country_code) %>%
  summarise(Sentences = n(), Words = sum(n))

# Display the result
print(word_count_per_country)
```

Table 1.1: descriptive statistics

Country	Sentences	Words
SIERRA LEONE	971	1651
ANGOLA	855	1465
KENYA	902	1444
EGYPT	844	1419
GAMBIA	760	1336
MAURITIUS	784	1225
MOZAMBIQUE	738	1178
ESWATINI'	695	1150
UGANDA	607	1086
LESOTHO	674	1050
GHANA	642	1033
SOUTH AFRICA	545	1029
NAMIBIA	582	939

SEYCHELLES'	565	936
TANZANIA	608	931
ETHIOPIA	614	912
NIGERIA	624	902
BOTSWANA	593	888
LIBERIA	514	869
ZIMBABWE'	520	794
CAPE VERDE	463	777
MALAWI	458	775
GUINEA-BISSAU	336	512
SOUTH SUDAN	366	510
ERITEA	368	459
RWANDA	352	457

Based on the statistics shown in Table 1, it is apparent that Sierra Leone exhibited the greatest frequency of sentences and words in their speech, followed by Angola, Kenya, Egypt, Gambia, Mauritius, and Mozambique. The countries of Rwanda, Eritrea, South Sudan, and Guinea-Bissau presented the shortest sentences and words in their speech, as stated in given table.

1.2 Universal part of speech (POS) tag

For a comprehensive list of the parts of speech (POS) tags and their corresponding definitions, please refer to this resource: https://universaldependencies.org/u/pos/index.html. The provided code examines the distribution of each category independently.

```
pos_counts <- merged_df %>%
  group_by(Country_Name) %>%
  summarise(
    SpeechLength = n(),
    NounCount = sum(upos == "NOUN"),
    VerbCount = sum(upos == "VERB"),
    AdjectiveCount = sum(upos == "ADJ"),
    AdverbCount = sum(upos == "ADV")
  )

print(n=26, pos_counts)

library(ggplot2)

# Reshape data for better plotting
pos_counts_long <- tidyr::gather(pos_counts, key = "PartOfSpeech", value = "Count",
                                 -Country_Name, -SpeechLength)

# Plot
ggplot(pos_counts_long, aes(x =Count , y = PartOfSpeech, fill = PartOfSpeech)) +
  geom_bar(stat = "identity") +
  facet_wrap(~Country_Name, scales = "free") +
  labs(title = "Part of Speech Distribution by Country",
       x = "Part of Speech",
       y = "Count") +
  theme_minimal()
```

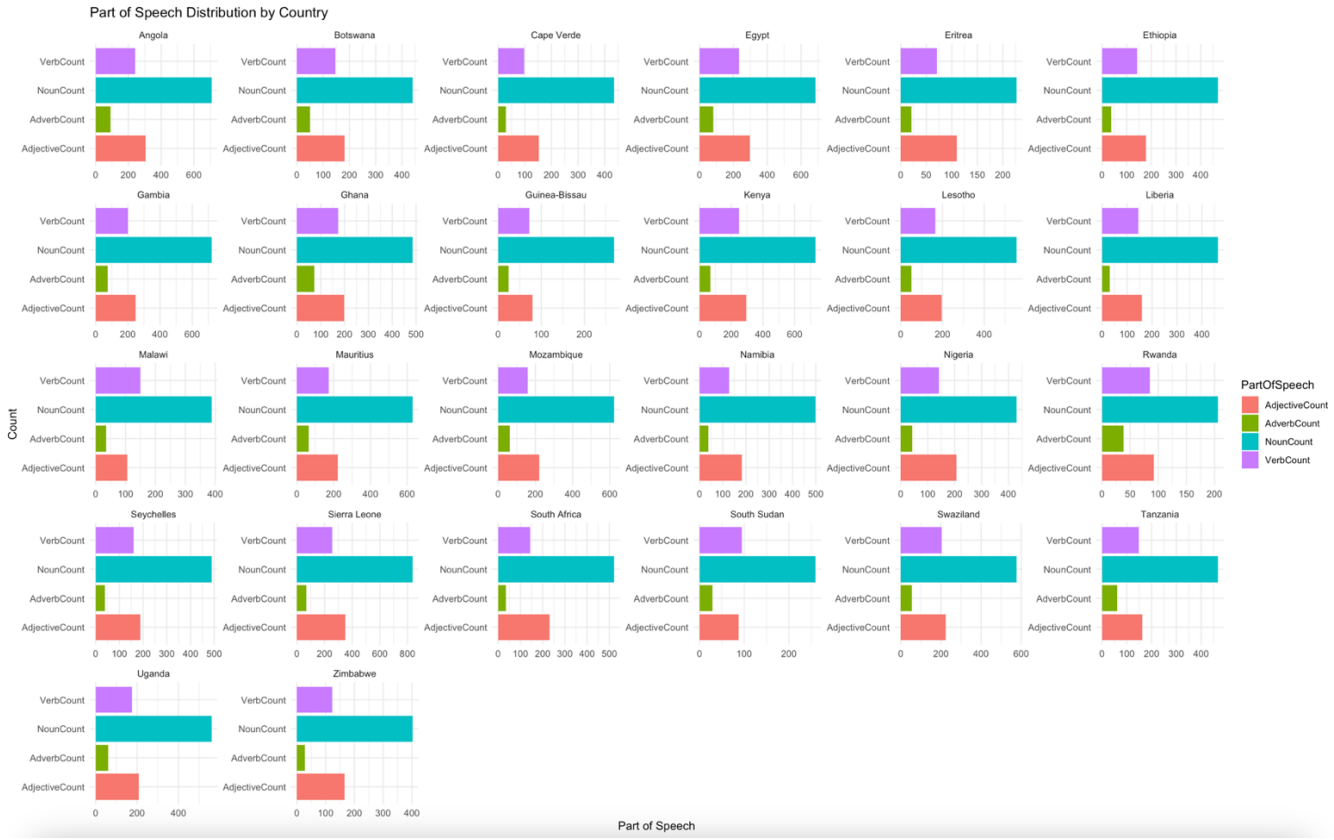

Figure 1.1: Part of Speech distribution by country

From figure 1.1, it is evident that Sierra Leone possesses the longest speech, surpassing all other countries in this regard. Subsequently, Figure 1.1 also illustrates the frequency distribution of each UPOS type. Most of the speeches primarily comprise of nouns, adverbs, verbs, and adjectives.

1.2.1 Nouns

```
# Filter for NOUNs only
noun_counts <- merged_df %>%
  filter(upos == "NOUN") %>%
  group_by(Country_Name) %>%
  count(term = lemma, sort = TRUE) %>%
  top_n(6, wt = n)  # Adjust 10 to the desired number of top nouns

# Plot the most used nouns by country
ggplot(noun_counts, aes(x = n, y = fct_reorder(term, n), fill = term)) +
  geom_col() +
  geom_text(aes(label = n), hjust = -0.2, size = 3, color = "black") +  # Add count labels on y-axis
  facet_wrap(~ Country_Name, scales = "free_y") +
  labs(title = "Top Nouns by Country",
       x = "Count",
       y = "Noun") +
  theme_minimal()+
  guides(fill = FALSE)+ # Remove legend
  theme(axis.text.y = element_text(size = 10, margin = margin(0, 0, 0, 0)))  # Adjust y-axis label size and add space between labels
```

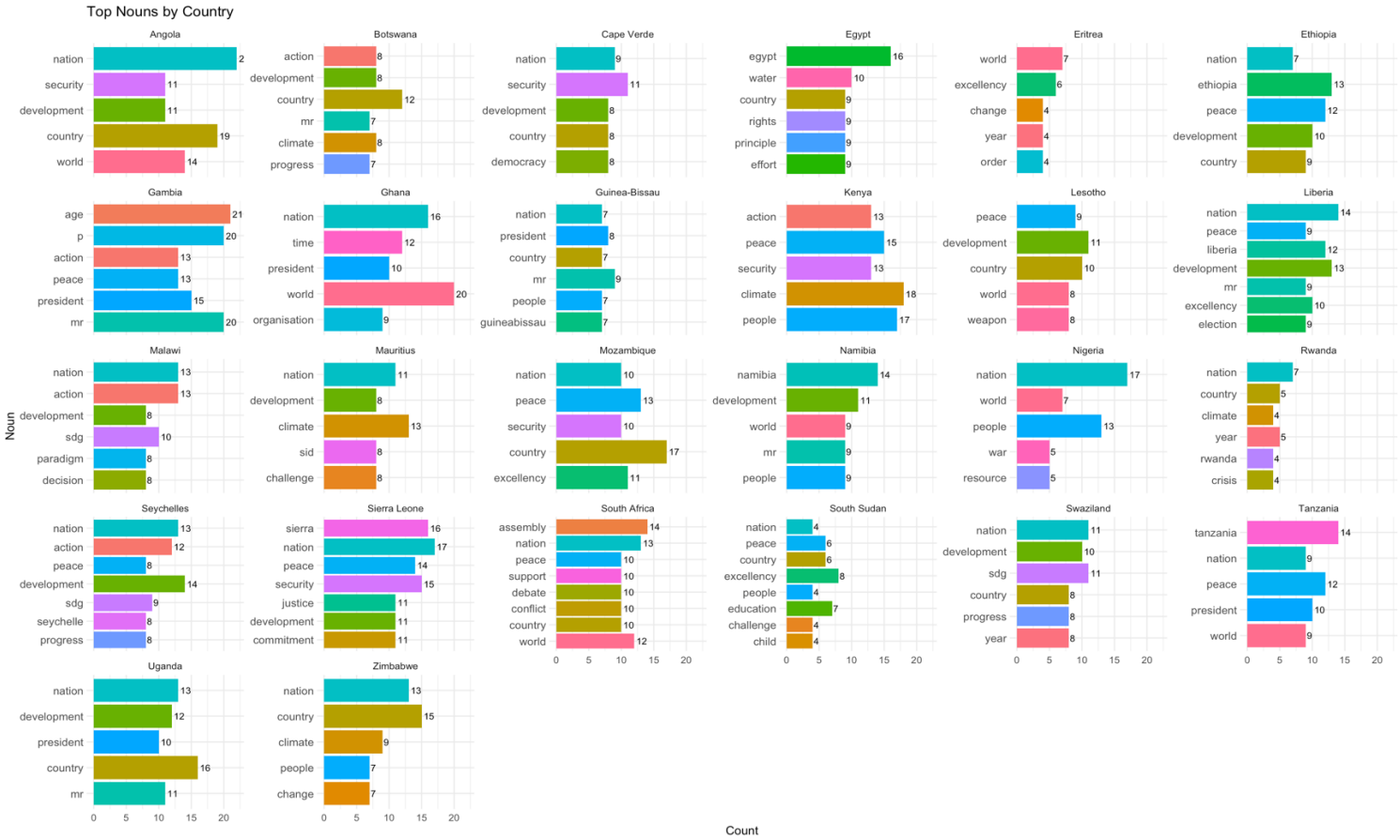

Figure 1.2: Top nouns by country

1.2.2 Verbs

```
# Filter for VERBs only
verb_counts <- merged_df %>%
  filter(upos == "VERB") %>%
  group_by(Country_Name) %>%
  count(term = lemma, sort = TRUE) %>%
  top_n(6, wt = n)  # Adjust 10 to the desired number of top verbs

# Plot the most used verbs by country with separate plots for each country
ggplot(verb_counts, aes(x = n, y = fct_reorder(term, n), fill = term)) +
  geom_col() +
  geom_text(aes(label = n), hjust = -0.2, size = 3, color = "black") +  # Add count labels on y-axis
  facet_wrap(~ Country_Name, scales = "free_y") +
  labs(title = "Top Verbs by Country",
       x = "Count",
       y = "Verb") +
  theme_minimal() +
  guides(fill = FALSE)  # Remove legend
```

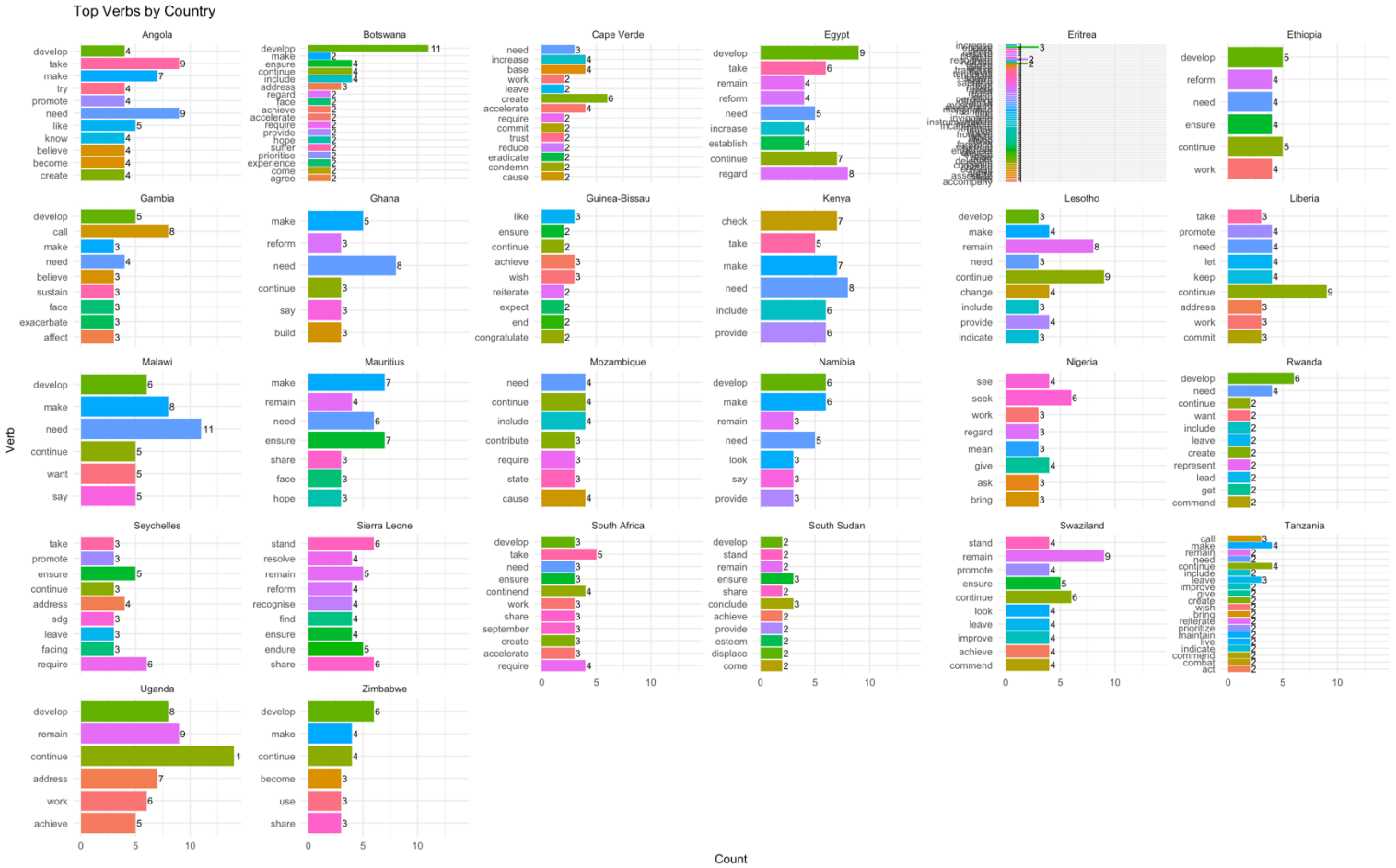

Figure 1.3: Top Verbs by Country

1.2.3 Adjectives

```r
# Filter for ADJECTIVEs only
adjective_counts <- merged_df %>%
  filter(upos == "ADJ") %>%
  group_by(Country_Name) %>%
  count(term = lemma, sort = TRUE) %>%
  top_n(4, wt = n)  # Adjust 10 to the desired number of top adjectives

# Plot the most used adjectives by country with separate plots for each country
ggplot(adjective_counts, aes(x = n, y = fct_reorder(term, n), fill = term)) +
  geom_col() +
  geom_text(aes(label = n), hjust = -0.2, size = 3, color = "black") +  # Add count labels on y-axis
  facet_wrap(~ Country_Name, scales = "free_y") +
  labs(title = "Top Adjectives by Country",
       x = "Count",
       y = "Adjective") +
  theme_minimal() +
  guides(fill = FALSE)  # Remove legend
```

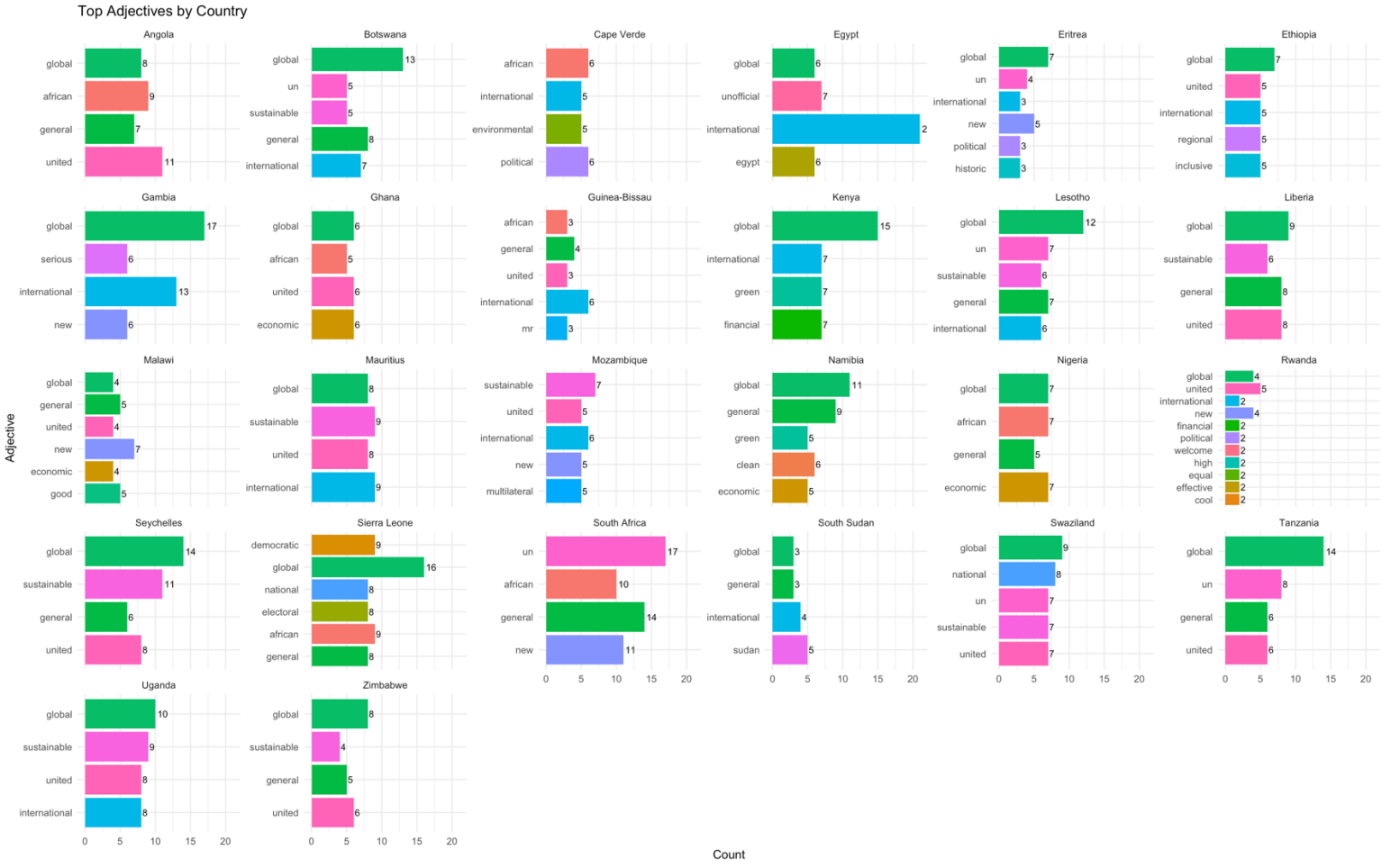

Figure 1.4: Top Adjectives by country

1.2.4 Adverbs

```
# Filter for ADVERBs only
adverb_counts <- merged_df %>%
  filter(upos == "ADV") %>%
  group_by(Country_Name) %>%
  count(term = lemma, sort = TRUE) %>%
  top_n(4, wt = n)  # Adjust 10 to the desired number of top adverbs

# Plot the most used adverbs by country with separate plots for each country
ggplot(adverb_counts, aes(x = n, y = fct_reorder(term, n), fill = term)) +
  geom_col() +
  geom_text(aes(label = n), hjust = -0.2, size = 3, color = "black") +  # Add count labels on y-axis
  facet_wrap(~ Country_Name, scales = "free_y") +
  labs(title = "Top Adverbs by Country",
       x = "Count",
       y = "Adverb") +
  theme_minimal() +
  guides(fill = FALSE)  # Remove legend
```

Figure 1.5: Top Adverbs by Country

1.3 Summary

It is evident that Sierra Leone possesses the longest speech, surpassing all other countries in this regard. Subsequently, this sections also illustrates the frequency distribution of each UPOS type. Most of the speeches primarily comprise of nouns, adjectives, verbs, and adverbs. Based on the depicted picture, it can be observed that most national speeches exhibit a higher frequency of nouns and adjectives, followed by verbs, and finally adverbs. For a comprehensive list of the parts of speech (POS) tags and their corresponding definitions, please refer to this resource: https://universaldependencies.org/u/pos/index.html .

2 Visualizing Words

The initial segment pertains to the frequency of words within each statement. This procedure calculates the words that exhibit the highest degree of exclusivity in a specific speech. This metric quantifies the degree of specificity exhibited by individual speeches in terms of their respective vocabularies. The second portion will incorporate a word cloud that encompasses all the aggregated statements, serving as an additional visual representation of the frequency of words used. The frequency of a word in the text is shown by its size.

2.1 Frequent words

2.1.1 Top most Frequent words per country

```
#frenquent used words per country

# Create a data frame with word frequencies
word_freq <- merged_df %>%
  group_by(Country_Name, lemma) %>%
  summarise(freq = n()) %>%
  arrange(desc(freq)) %>%
  group_by(Country_Name) %>%
  top_n(10, wt = freq)

# Plot the bar chart with facets
ggplot(word_freq, aes(x = freq, y = fct_reorder(lemma, freq), fill = lemma)) +
  geom_col() +
  facet_wrap(~ Country_Name, scales = "free_y") +
  labs(title = "Ten Most Frequent Words in Country Statements",
       x = "Word",
       y = "Frequency") +
  theme_minimal() +
  guides(fill = FALSE)  # Remove legend
```

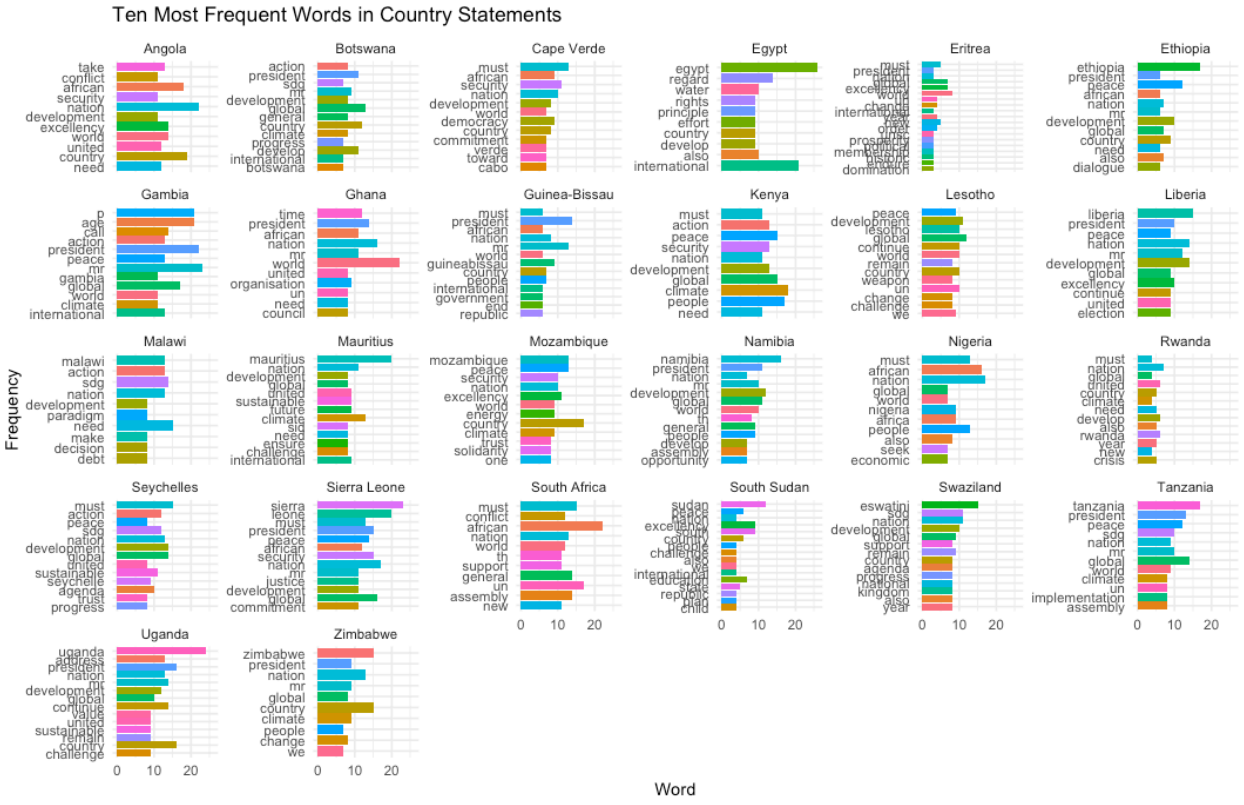

Figure 2.1: Ten Most Frequent Words in Country Statement

2.1.2 Top most Frequent words as a whole

```
# Select the top 20 most frequent words
top_words <- head(sort(word_frequencies, decreasing = TRUE), 20)

# Create a data frame for plotting
plot_data <- data.frame(word = names(top_words), freq = top_words)

# Plot a bar graph
ggplot(plot_data, aes(x = fct_reorder(word, freq), y = freq, fill = word)) +
  geom_col() +
  coord_flip() +
  facet_wrap(~"", scales = "free_y", ncol = 1) +
  labs(title = "Top 20 Most Frequent Words in Country Statements",
       x = "Word",
       y = "Frequency") +
  theme_minimal() +
  guides(fill = FALSE)  # Remove legend
```

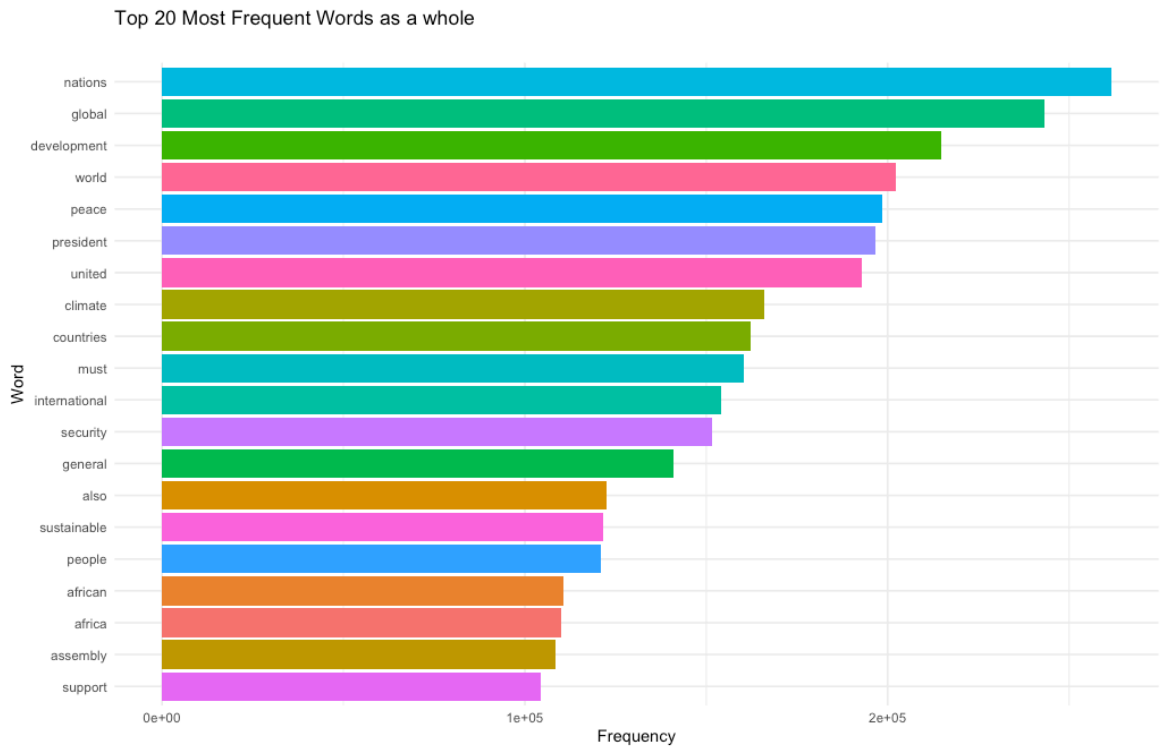

Figure 2.2:Top 20 Most Frequent words as a whole

Upon analyzing the top 20 most frequent words, it becomes evident that the words "nations," "global," "development," "word," "peace," "president," "united," and "climate" hold the highest

frequency. However, the presence of these terms in isolation provides minimal assistance or guidance since they lack contextual information.

2.2 word could

```
# Installing and loading the wordcloud package
if (!require(wordcloud)) {
  install.packages("wordcloud")
  library(wordcloud)
}
library(wordcloud)

# Assuming you already have 'merged_df' data frame
merged_df %>%
  count(lemma, sort = TRUE) %>%
  with(wordcloud(words = lemma, freq = n, max.words = 200, random.order = FALSE,
                 rot.per = 0.35, colors = brewer.pal(8, "Dark2")))
```

Figure 2.3: word could

2.3 Summary

To summarize, the examination of word frequency, frequency statistics in the initial part yielded significant findings regarding the predominant usage of words in individual speeches. Nevertheless, it became apparent that specific terms acquired value alone when used in conjunction with others. To tackle this issue, the subsequent phase of the study placed emphasis on the identification and extraction of significant keyword combinations, recognizing the significance of contextual factors in facilitating a thorough comprehension of the claims. This methodology enriches the comprehensiveness of our analysis by encompassing subtle nuances that may not be readily discernible through the sole utilization of isolated word frequency metrics.

3 Key word Identification

Frequency statistics of words are revealing, but one may find words which only make sense in combination with other words. Hence the goal of finding and extracting keywords which are a combination of words. The udpipe R package provides three method to identify keywords in text :

- RAKE (Rapid Automatic Keyword Extraction)
- Collocation ordering using Pointwise Mutual Information
- Parts of Speech phrase sequence detection

Both RAKE and PoS techniques are used to generate rankings of common keywords across all combined speeches. Using different algorithms, for the same purpose, are a useful ways of testing if different models perform in an expected, comparable way:

3.1 Using Rake method

```
# Using the RAKE method

stats <- keywords_rake(mrged_df, term = "lemma", group = "Country_Code",
                       relevant = mrged_df$upos %in% c("NOUN", "ADJ"))
stats$key <- factor(stats$keyword, levels = rev(stats$keyword))
barchart(key ~ rake, data = head(subset(stats, freq > 3), 20), col = "cadetblue",
         main = "Keywords identified by RAKE",
         xlab = "Rake")
```

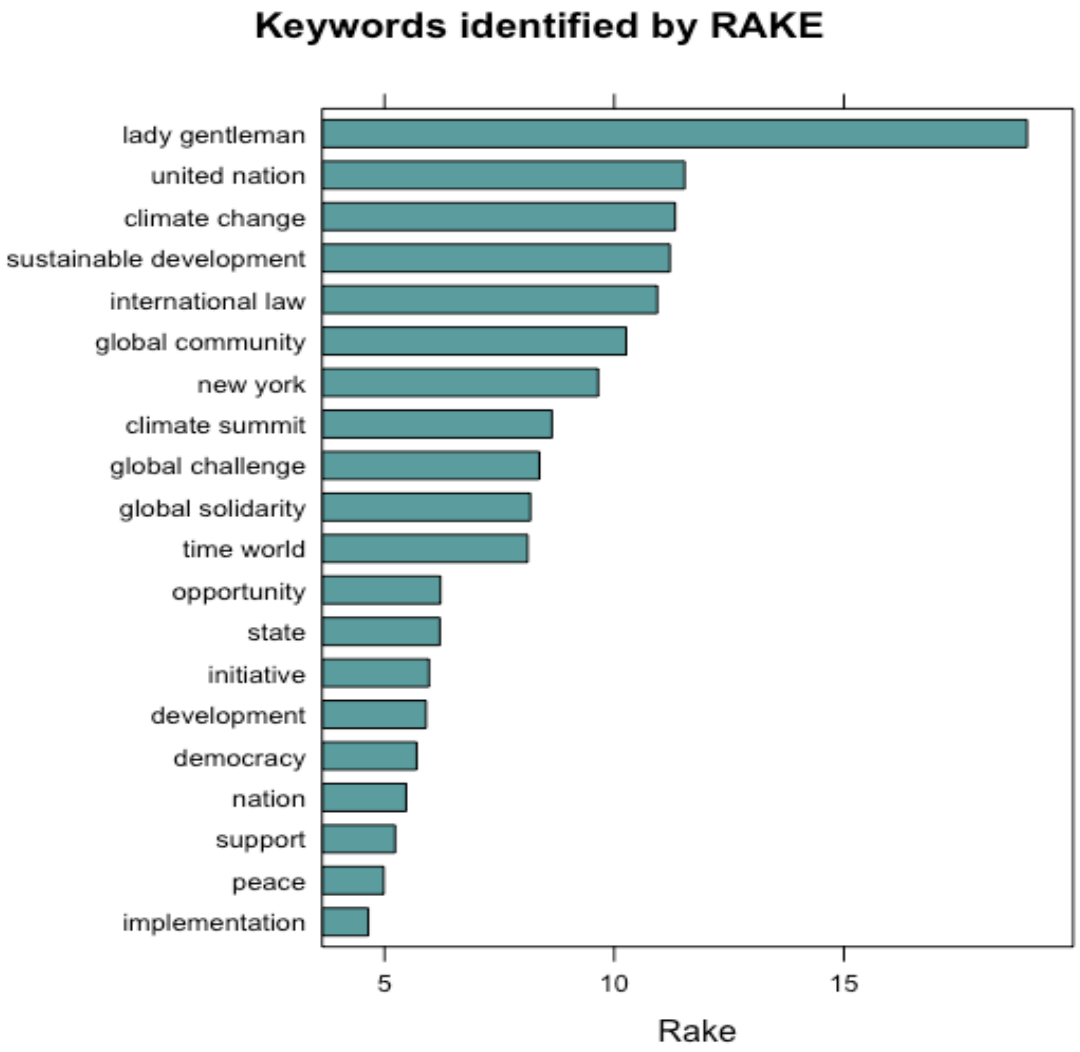

Figure 3.1:Key word identified by RAKE

The RAKE algorithm was employed to identify the prevailing keywords, which encompassed terms such as "lady" and "gentlemen," "United Nations," "climate change," "sustainable development," "international law," "global community," and "climate summit."

3.2 Pointwise mutual information

```
# Using Pointwise Mutual Information Collocations

mrged_df$word <- tolower(mrged_df$token)
stats <- keywords_collocation(x = mrged_df, term = "word", group = "doc_id")
stats$key <- factor(stats$keyword, levels = rev(stats$keyword))
barchart(key ~ pmi, data = head(subset(stats, freq > 3), 20), col = "cadetblue",
         main = "Keywords identified by PMI Collocation",
         xlab = "PMI (Pointwise Mutual Information)")
```

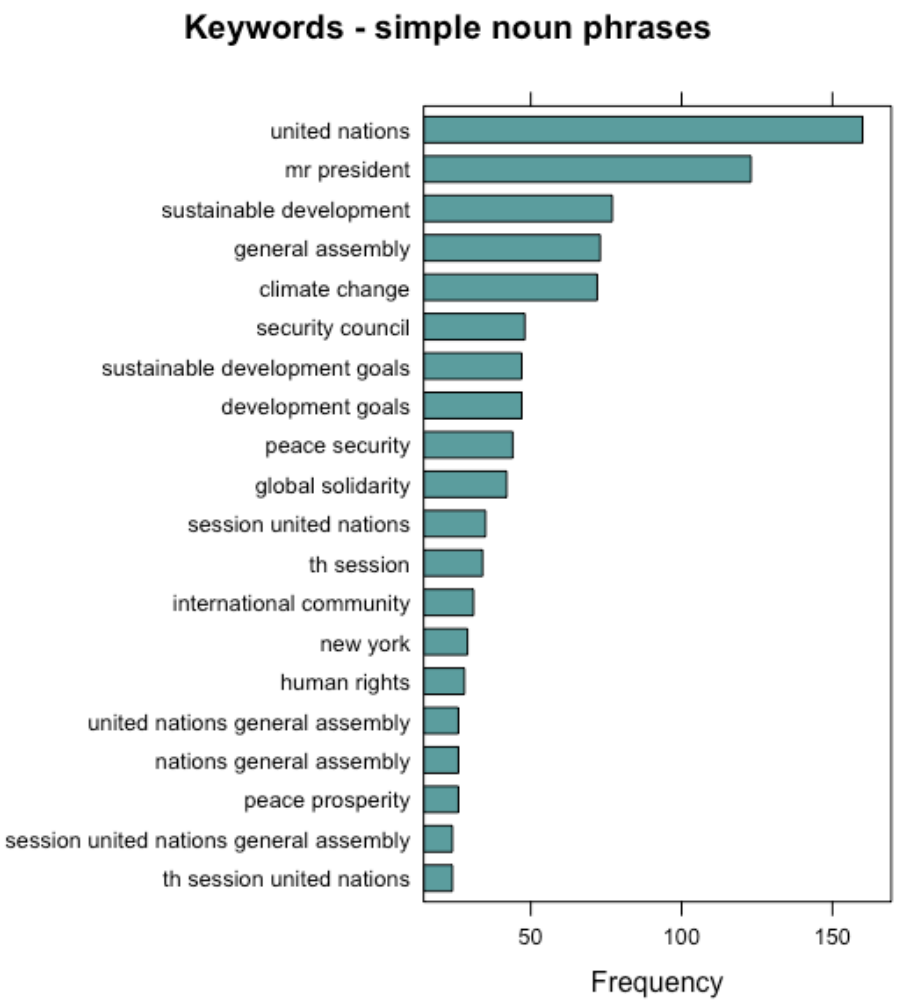

Figure 3.2: PMI (Pointwise Mutual Information)

Utilizing the PMI (Pointwise Mutual Information) framework, the prevailing keywords identified encompass the United Nations, the President, sustainable development, the General Assembly, climate change, the Security Council, sustainable development goals, development goals, and peace security.

3.3 Using POS method

```
#Using POS

mrged_df$phrase_tag <- as_phrasemachine(mrged_df$upos, type = "upos")
stats <- keywords_phrases(x = mrged_df$phrase_tag, term = tolower(mrged_df$token),
                          pattern = "(A|N)*N(P+D*(A|N)*N)*",
                          is_regex = TRUE, detailed = FALSE)
stats <- subset(stats, ngram > 1 & freq > 3)
stats$key <- factor(stats$keyword, levels = rev(stats$keyword))
barchart(key ~ freq, data = head(stats, 20), col = "cadetblue",
         main = "Keywords - simple noun phrases", xlab = "Frequency")
```

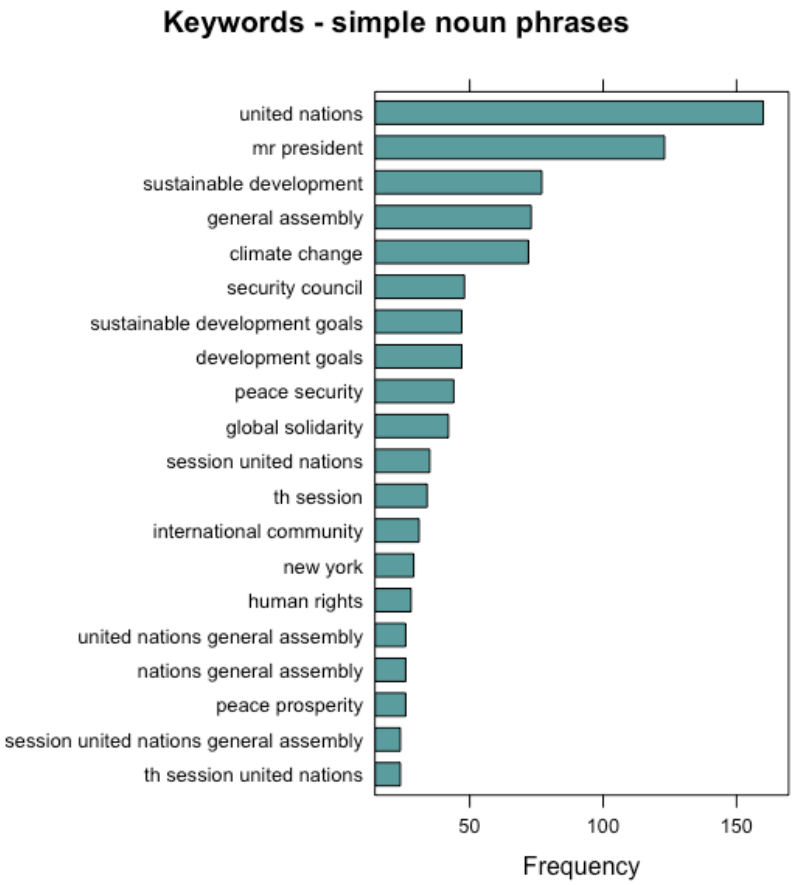

Figure 3.3: Part-of-Speech (PoS)

The results obtained with the Part-of-Speech (PoS) method exhibit similarities with those obtained through the Pointwise Mutual Information (PMI) method. The most frequently occurring terms identified are "United Nations," "Mr. President," "sustainable development," "General Assembly," "climate change," "Security Council," "sustainable development goals," "development goals," and "peace security."

4 Analysing n-gram

An n-gram refers to a consecutive sequence of n words extracted from a given text. For instance, a bigram is a combination of two words, where the value of n is equal to 2. This analysis provides an initial examination of the occurrence frequencies of the most commonly observed bigram (n=2) and trigram (n=3).

4.1 Nouns / adjectives used in same sentence

```
#Nouns / adjectives used in same sentence

cooc <- cooccurrence(x = subset(mrged_df, upos %in% c("NOUN", "ADJ")),
                     term = "lemma",
                     group = c("doc_id", "paragraph_id", "sentence_id"))
head(cooc)

library(igraph)
library(ggraph)
library(ggplot2)
wordnetwork <- head(cooc, 30)
wordnetwork <- graph_from_data_frame(wordnetwork)
ggraph(wordnetwork, layout = "fr") +
  geom_edge_link(aes(width = cooc, edge_alpha = cooc), edge_colour = "pink") +
  geom_node_text(aes(label = name), col = "darkgreen", size = 4) +
  theme_graph(base_family = "Arial Narrow") +
  theme(legend.position = "none") +
  labs(title = "Cooccurrences within sentence", subtitle = "Nouns & Adjective")
```

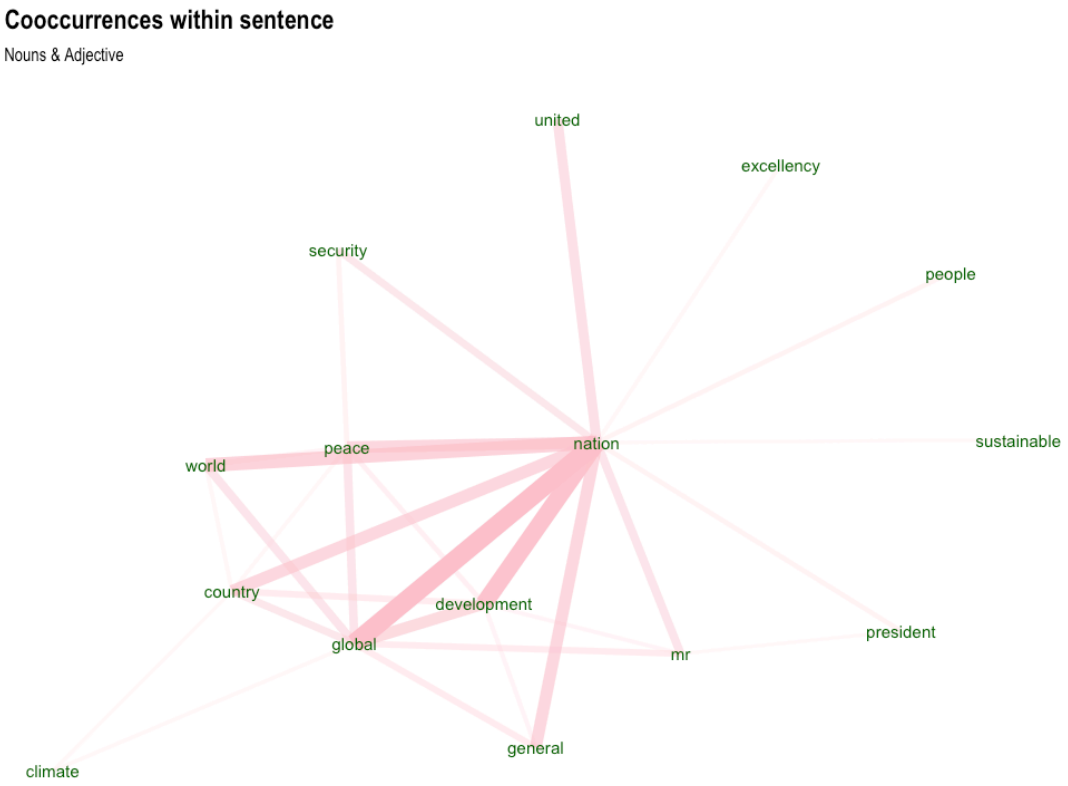

Figure 4.1: Nouns / adjectives used in same sentence

4.2 Nouns / adjectives which follow one another

```
#Nouns / adjectives which follow one another
cooc <- cooccurrence(mrged_df$lemma, relevant = mrged_df$upos %in% c("NOUN", "ADJ"), skipgram = 1)
head(cooc)

library(igraph)
library(ggraph)
library(ggplot2)
wordnetwork <- head(cooc, 30)
wordnetwork <- graph_from_data_frame(wordnetwork)
ggraph(wordnetwork, layout = "fr") +
  geom_edge_link(aes(width = cooc, edge_alpha = cooc)) +
  geom_node_text(aes(label = name), col = "darkgreen", size = 4) +
  theme_graph(base_family = "Arial Narrow") +
  labs(title = "Words following one another", subtitle = "Nouns & Adjective")
```

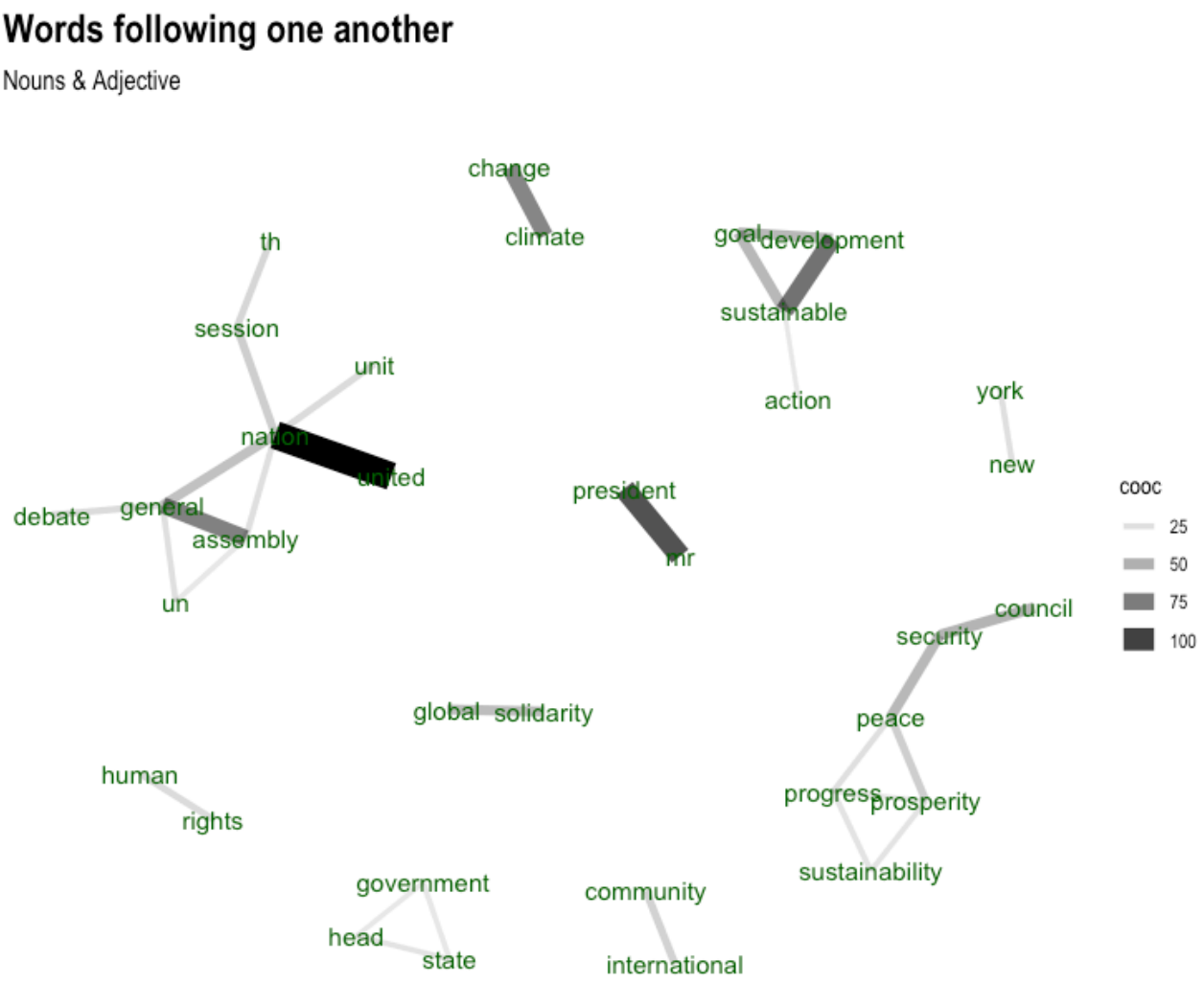

Figure 4.2: Nouns / adjectives which follow one another

4.3 Summary

In contrast to examining individual characteristic words, the analysis now shifts focus to characteristic n-grams per document. This approach visualizes combinations of words that are most representative of each country's statement. The resulting network graph illustrates common co-occurrences, reaffirming findings from previous analyses (RAKE, PMI, PoS). Notably, terms like 'President,' 'Sustainable development,' 'Progress,' 'Prosperity,' 'Peace,' 'The government,' 'Global solidarity,' and variations of 'United Nation' consistently emerge. This aligns with the earlier frequency analysis, highlighting the persistence of key thematic elements across different analytical perspectives.

5 Sentiment Analysis

From this analysis we can tell the sentiment of each speech as delivered by each country.

```
db_bing<- mutate(sentiment=positive - negative,
                 .data=db_bing)

ggplot(db_bing, aes(x = n, y = fct_reorder(sentiment, n), fill = sentiment)) +
  geom_col() +
  geom_text(aes(label = n), hjust = -0.2, size = 3, color = "black") +  # Add count labels on y-axis
  facet_wrap(~ Country_Name, scales = "free_y") +
  labs(title = "Country speech sentiment",
       x = "Count",
       y = "Verb") +
  theme_minimal() +
  guides(fill = FALSE)  # Remove legend
```

Figure 5.1: Speech Sentiment

Upon observation of the aforementioned figure, it becomes evident that the speeches, on average, had a favorable emotion. Examining Eritrea's speech is of significant importance, as it stands as the sole speech expressing a negative viewpoint.