PreechVis: Visual profiling using multiple-word combinations

Seongmin Mun, Gyeongcheol Choi, Guillaume Desagulier, Kyungwon Lee

PreechVis: Visual profiling using multiple-word combinations

Seongmin Mun[1], Gyeongcheol Choi[2], Guillaume Desagulier[3] and Kyungwon Lee[4]
[1]*Sciences du Langage(Computational linguistic), Life media Interdisciplinary Program, UMR 7114 MoDyCo - CNRS, University Paris Nanterre, Ajou University, 200 avenue de la République 92000, Paris, France*
[2]*Life media Interdisciplinary Program, Ajou University, 206, World cup-ro, Yeongtong-gu, Suwon, Republic of Korea*
[3]*English grammar and linguistics, UMR 7114 MoDyCo - University Paris 8, CNRS, University Paris Nanterre, 200 avenue de la République 92000, Paris, France*
[4]*Digital Media, Ajou University, 206, World cup-ro, Yeongtong-gu, Suwon, Republic of Korea*
{stat34, ckc6842, kwlee}@ajou.ac.kr, gdesagulier@univ-paris8.fr

Keywords: Data Clustering, Data Filtering, Hierarchy Data, Text and Document Data, Interaction Design, Coordinated and Multiple Views, Task and Requirement Analyses, Visualization in the Humanities

Abstract: Words in the corpus include features and information, and the visualizing of such words can improve the user's understanding of them. Those words may be consist of one-word or they may be a combination of words that together. The latter is referred to as a multiword expressions (MWEs). And if we analyze both single word and multiword with visualization, we can get more accurate results and more information than when we analyze only single word from corpus. An interactive-visualization can be useful for analyzing multiword expressions, because the following features are of interest to linguistics scholars: (1) Showing the combinations of POS pattern, (2) exploring the results according to the POS combination pattern, and (3) searching the source corpus for the verification. Therefore, we propose PreechVis, an interactive visualization tool that includes all of the requisite functions for an analysis using multiwords (http://ressources.modyco.fr/sm/PreechVisMWE/). For the present study, we used a total of 957 speeches, 164,646 sentences and 3,698,617 tokens of 43 U.S. Presidents from George Washington to Barack Obama as the corpus. PreechVis is divided into two views. In the first view, the system consists of a combination of Sunburst and RadVis. Through the Sunburst, we present the POS and its combination patterns for each gram. In RadVis, the Presidents were positioned according to their frequency value. In addition, when the President was selected, the frequency value was displayed on Sunburst to improve the user's understanding. In the second view, the user can simultaneously confirm and verify the details of the result using the Wordcloud. The two different views are synchronized each other and easy to change by the selected grams, issues, and presidents. With the experiments and case studies on the U.S. President speeches, we verified the effectiveness and usability of PreechVis.

1 INTRODUCTION

A visual analysis of textual data can support users to have a general understanding of the information of the text without actually reading. This can be very helpful for tasks with large volumes of text. Word-based research is very common for visual corpus analyses (Lu et al., 2016, Wang et al., 2016, Heimerl et al., 2016).

The words in the corpus include features and information, and visualizing a word can help user's understanding of the corpus easily. These words can be split into two types. The first type is the word with a one-word meaning, whereas the second type is the word with a combined-word meaning. The second word is called multiword and is generated by a combination of different parts-of-speech (POS) (Ramisch, 2015). Simply, the multiword is a habitual recurrent word combination of everyday language (JR, 1957). For example, when people say that *someone sets the bar high*, it is understood as a metaphor that his or her competitors will find it difficult to win against him or her. If we analyze both single word and multiword with visualization, we can get more accurate results and more information than when we used an only single word in the analysis.

An interactive visualization can be useful for analyzing multiword expressions because the following features are of interest to linguistics scholars: (1) Showing the combinations of POS

pattern, (2) exploring the results according to the POS combination pattern, and (3) searching the source corpus for the verification.
Therefore, PreechVis, an interactive visualization tool that covers all of the necessary functions for the exploration of larger amounts of the multiword corpus, has been created for this study.
This work provides the following contributions: (1) this study reveals a data-processing method that can obtain more accurate results than when we do analysis without multiword, and (2) multiwords are consist of different POS combination pattern and usually used POS combinations have existed. Presidential Address is used to verify the utility of PreechVis.

2 RELATED WORKS

For this work, it is assumed that with a textual multiword analysis, the information is easier discovered. Supporting this hypothesis, Carlos Ramisch summarized a textual-analysis technique for multiword expressions in his recent book (Ramisch, 2015). Further, many studies on the visual tools for textual analyses have been proposed (Koch et al., 2014, Sun et al., 2014).
Numerous visualizations have been created to extract and explore more information in the large corpus with a number of visualizations such as EvoRiver (Sun et al., 2014) and OpinionFlow (Wu et al., 2014. The word-based analysis is employed regardless of the multiword combinations. The focus here is the multiword-based analysis and the definition of how the multiword result can be presented in an interactive visualization.

2.1 Word-based corpus visualization

Word-based corpus visualization, which aims to understand and explore words-based text corpus, has received considerable attention in recent years (Sun et al., 2014, Cui et al., 2014a, Wu et al., 2014).
EvoRiver (Sun et al., 2014) is a time-based visualization that allows users to explore competition-related interactions and to detect dynamically evolving patterns, as well as their major causes. Cui et al. (Cui et al., 2014a) presented an interactive visual textual-analysis approach that allows users to progressively explore and analyze the complex evolutionary patterns of hierarchical topics. Wu et al. (Wu et al., 2014) introduced a visual-analysis system called OpinionFlow to empower analysts to detect opinion-propagation patterns and glean insights. Further, for OpinionFlow, a Sankey graph is combined with a tailored density map in one view to visually convey the diffusion of opinions among many users.
These related works focus on visual explorations of words without a consideration of the multiword expressions. Whereas, our present work includes multiword expressions in its word-based illustrations.

2.2 Visual Graph Comparison

A visual-graph comparison aims to analyze the similarities and differences between variables. A number of visual-graph-comparison methods have been proposed by many studies (Andrews et al., 2009, Cui et al., 2014b, Collins and Carpendale, 2007).

Andrews et al. (Andrews et al., 2009) presented a technique and a prototype tool to support the visual comparison of graphs and the interactive reconciliation of candidate graphs into a single reference graph. Cui et al. (Cui et al., 2014b) introduced a novel flow-based visualization design for the summarization of high-level evolution patterns in a dynamic graph. Collins et al. (Collins and Carpendale, 2007) described VisLink, a visualization environment in which one can display multiple two-dimensional (2D) visualizations, reposition and reorganize them in a three-dimensional (3D) form, and display the relationships between them by propagating the edges from one visualization to another.

These works assume that visual graph comparison can help the user to understand the similarities and differences between two different variables In this paper, the PreechVis allows for wider comparison, across analysis results extracted from user selections.

2.3 Verification of the visual findings

A visual analysis of the corpus can help the user to understand the corpus without actually reading corpus. However, to assess the utility of a visual-analysis tool, derived insights must be verified through comparison with the real corpus (Koch et al., 2014, Stasko et al., 2007).

Koch et al. (Koch et al., 2014) presented a method that supports visual-analytics tasks on large text documents that is particularly useful in situations where scrutiny is required and the textual source must be used to verify the findings. Stasko et al. (Stasko et al., 2007) developed a visual-analytics system called Jigsaw that visually represents documents and their entities to help analysts examine reports more efficiently and to develop potential-action theories

more quickly. Further, this system provides multiple coordinated views of the document entities with a special emphasis on a visual illustration of the connections across different documents.

For the verification of visual findings, PreechVis offers visual results with a real corpus to simultaneously verify visual-analytics insights.

3 DESIGN CONSIDERATION

3.1 User tasks

To understand the needs of the multiword analysis, regular cooperation was sought from computational-linguistic researchers to learn about their hypotheses and goals in the study of the multiword. Also, they were presented with a variety of designs, and important feedback was collected from them to refine the disambiguation results. The experience revealed that researchers typically need to accomplish the following tasks in their explorations of corpus:

- **Task1: Exploration of the word-combinations patterns for each gram.**
  Researchers are interested in this task to discover the different word-combinations patterns for each gram. N-gram is a contiguous sequence of n items from a given sequence of text or speech (Sidorov et al., 2013). In this paper, the word-combination patterns from the unigram to the trigram were analyzed.

- **Task 2: Identification of usage patterns of the part of speech and part of speech combination by each U.S. President.**
  The linguistic researchers wanted to know typical usage patterns of the part of speech and part of speech combination. To find out usage patterns of the part of speech and part of speech combination, the U.S. Presidents that exhibit in the visualization were identified.

- **Task 3: Verification of visual findings with real corpus data.**
  A visual analysis of corpus data can support the user in their attainment of an understanding of the corpus information. However, visual-analysis results are required to verify and prove the findings (Koch et al., 2014).

3.2 Design objectives

In response to the previously mentioned tasks, the following design objectives were built to guide the proposed approach:

- **Design Objective 1: We should create visual part to show word-combinations patterns.**
  Computational-linguistic researchers need to know what kind of word-combinations patterns they have and how different those of each gram are.

- **Design Objective 2: The visualization have to show usage patterns of the part of speech and combination pattern of part of speech.**
  To support the interactive comparison for the usage patterns of the part of speech and combination pattern of part of speech. among the presidents (T.2), a interactive-view needs to be included in the main view of the visualization. Through this part, the user can easily confirm the usage patterns of the part of speech and combination pattern of part of speech by different selection of the presidents. Through this interaction, the user can select each president and issues, afterwards the usage patterns will be changed according to the selected president and issue.

- **Design Objective 3: We need to make the verification part to verify between the visual findings and the real corpus.**
  The interactive visualization provides various supports for the exploration, analysis, and understanding regarding the corpus. However, it difficult to obtain all of the information in the corpus by using the visualization analysis; for this reason, a view is created for a corpus-based verification.

4 DATA PROCESSING

In this section, a data-processing structure for the extraction of the information from corpus is presented. The datasets used in this paper are taken from the Miller Center(https://millercenter.org/), one of the representative databases of the U.S. history and civil discourse. During the data processing, each piece of information in the database is extracted into several descriptive attributes including the personal information of each President, the public speeches of the President, and pictures of the President.

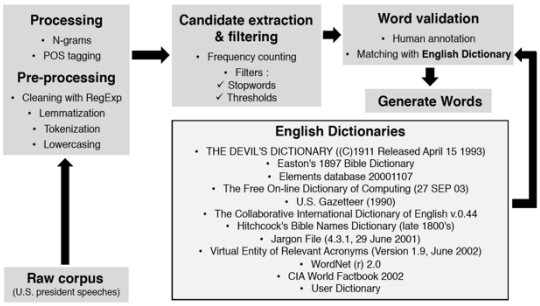


Figure 1: Structure of the Data Processing. Framework for the Word Acquisition from the Corpus Data.

Figure1 summarizes the architecture of the proposed data processing, which will be subsequently described in detail.

4.1 Processing

This section is divided into two parts. The first part is the pre-processing part, wherein Cleaning with RegExp, Lemmatization, Tokenization, and Lowercasing were conducted. Then, the N-gram analysis and the Part Of Speech (POS) tagging for the word candidate extraction were conducted.

4.2 Candidate Extraction and Filtration

Using the previously mentioned procedure, the N-gram results were obtained word candidates with the POS-tagging result. These results were counted according to the frequency value, and the data were filtered through the application of a threshold (frequency value greater than or equal to 10). In addition, word candidates were extracted without the stop-words for each gram. For instance, in the case of the bigram, word like *'house i'* and *'power we'* became stop-words and were removed from the word candidates.

4.3 Word Validation

In this section, the filtered word candidates are verified using several English dictionaries and User Dictionary.

The output of the word-candidate filtration is required for the verification. Also for the verification, an algorithmic working base was developed with several English dictionaries, as shown in figure 2. The proposed algorithm automatically compares the results with the several English dictionaries, and if the dictionary shows a result, it returns the word candidate as the meaningful result. Otherwise, It is discarded. However, Sometimes there are words have no definition in the dictionary but have a meaning. For this case, we made user dictionary collection in our database. If we store the words and meaning of words in the User dictionary, the algorithm can recognize more words.

Data: A list of word candidates $T_i : (k_i, p_i, c_i)$ sorted ascending by frequency value c_i
Result: A list of validated words $V_j : (k_j, p_j, c_j)$
start ⟵ 0, list of validated words ⟵ [ ];
foreach word candidates T_i in the input **do**
  if list of word candidates T_i isn't empty **then**
    boolean checked is true
    if dictionary haven't a word candidate **then**
      checked is false
    else
      continue
    end
    if checked is equal as true and $10 \leq c_i$ **then**
      add word candidates T_i in the list of validated words as V_j
    end
  end
end

Figure 2: Process of the Algorithmic Working Base on Several English Dictionaries.

The word candidate was extracted from single words, bi-grams, and tri-grams according to the previously described procedure; that is, 45,995 candidates were extracted from the uni-gram, 729,552 candidates were extracted from the bi-gram, and 2,089,617 candidates were extracted from the tri-gram. Among them, 8,910 in the uni-gram, 901 in the bi-gram, and 301 in the tri-gram were extracted as validated words with meaning and were analyzed.

4.4 POS combination Patterns

The computational-linguistic researchers want to find out the different word-combination patterns by each gram. For this purpose, the word results with the POS tagging were produced in the processing section. Each of the validated words comprises a different POS combination. There are 8 large and 36 detailed of POS in the uni-gram, 50 large and 200 detailed POS combinations in the bi-gram, and 65 large and 97 detailed POS combinations in the tri-gram.

4.5 Applying Term-weighting

There are many ways to calculate weights, including the Local mutual-information function, the Logarithm function, the Entropy function, and the Term Frequency-Inverse Document Frequency (TF-IDF) function. The TF-IDF function was used to identify the POS and the POS combinations by each

gram (Qu et al., 2008, Zhang et al., 2008). The TF-IDF formula is as follows:

$$\text{TF-IDF}_{t,d} = (1 + \log \text{tf}_{t,d}) \times \log(\frac{N}{\text{df}_t}) \qquad (1)$$

The main idea of the TF-IDF algorithm is as follows: In a case where the words describe the meaning of the sentence, the more times that a word appears in a sentence, the greater the contribution; furthermore, the greater the number of documents wherein the words appear, the smaller that the word result for a document contribution should be. By applying the TF-IDF algorithm, users can see more accurate analysis results.

5 VISUALIZATION DESIGN

From the previously mentioned user tasks and design objectives, an interactive visualization was designed to extract further information from the U.S. President Speech corpus. For this visualization, the display of the POS-combination patterns and the extraction of more accurate results were considered. The design of each visualization component is introduced as follows.

5.1 Emphasis of the Word-combination Pattern

A multiword is generated by a combination of the different POS patterns that computational-linguistic researchers search for. In the example of figure 4, Sunburst (Rodden, 2014) is used to demonstrate this clearly.

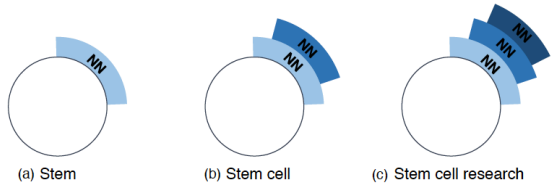


Figure 3: Emphasis of the word-combination patterns with Sunburst: (a) Presenting words in the uni-gram with the first layer, (b) showing the multiword in the bi-gram, and (c) the multiword in the tri-layer tri-gram.

Sunburst is a carrier of a spatial-information visualization in a circular layout, and it will extend outward with the increasing of the number of layers (Liu and Wang, 2015. Further, its value is high in terms of the exploration and analysis of the public information for large data amounts as a typical method for the visualization of hierarchical data. As shown in figure 3, a glyph was designed to represent the POS combination patterns with larger cost values.

5.2 Representation of the Usage Patterns

Computational-linguistic researchers look for typical usage patterns of POS and POS combinations. The combination pattern of POS is the hierarchical data, and the President's POS Usage Patterns are multivariate data.

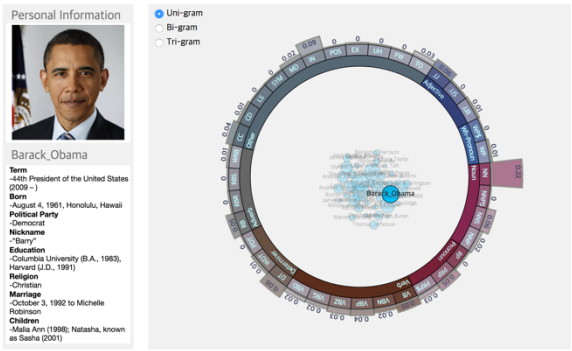


Figure 4: PreechVis Visualization Design. Displaying the information of Barack Obama by the usage frequency of the POS in the uni-gram.

It is a very experimental task to present both of them in one view, so for the design of the first view, RadVis (Sharko et al., 2008, Rubio-Snchez et al., 2016) and Sunburst were combined to determine the usage pattern of each President and their grouping. For example, in figure 4, the frequency bar that is located on the above Sunburst shows the information from Barack Obama's speech regarding the POS type that he used in his speech (Mahyar and Tory, 2014). And the nodes of the President that are inside the circle show their groups according to the usage frequency of the POS and POS combinations.

5.3 Verification for a Comparison with the Real Corpus

Previous research has shown that the visual-analysis result for corpus data is required to verify and prove findings in a comparison with real documents. In the example of figure 5, this visual tool is synchronized with real corpus, and so that it is simultaneously verified using the corpus data.

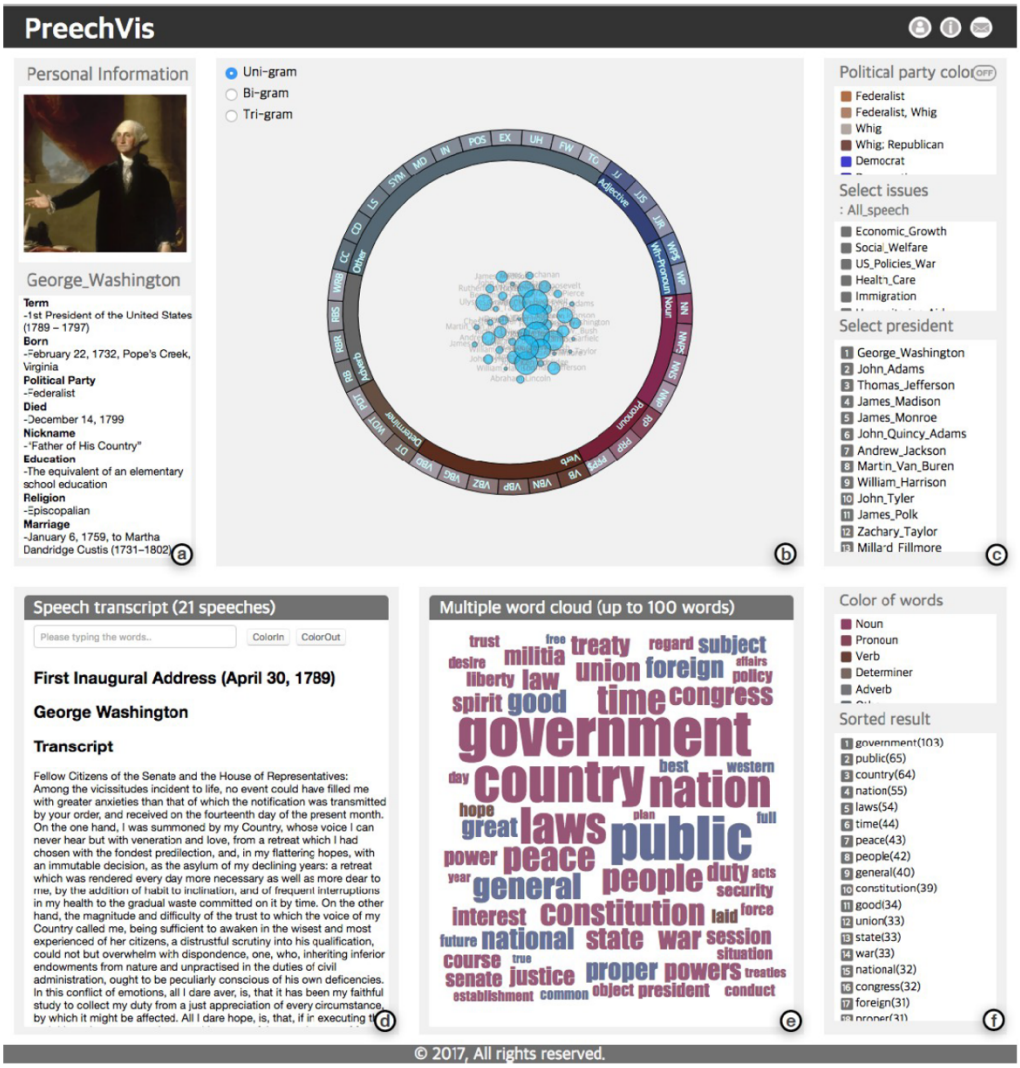


Figure 6: PreechVis Visualization Interface. The interface of the proposed visual system representing the corpus data about the speeches of 43 U.S. Presidents from George Washington to Barack Obama.

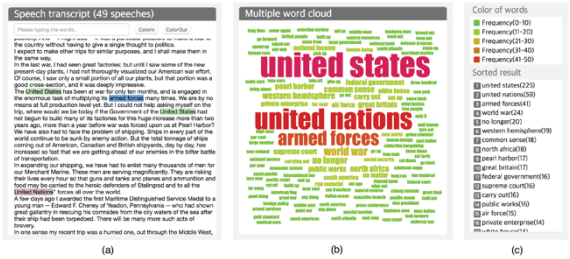


Figure 5: PreechVis Visualization Design: (a) Showing the real corpus and coloring by tying words, the (b) visualization-analysis result as a Word Cloud, and the (c) color legend and sorted results.

PreechVis uses a Word Cloud (b) to provide a visually distinguishable overview of the President's speech. This visual method is useful for learning about the number and kinds of words that are present in the corpus data. In the side view (c), the user can easily search the color legend about the used Word Cloud and the sorted word results. Additionally, PreechVis can also display the view of the real speech for the verification (a) that is on the left side of figure 5 for the verification.

5.4 PreechVis

Figure6 describes the main workspace of the proposed visual system after the loading of all of the

Presidential speeches. The three buttons in the layer headers (Figure6 (b)) provide the option to change the word combinations of the word for each gram. Additionally, the user can change the visual result by selecting the right-side option (Figure6 (c)). This makes the PreechVis approach very flexible because the different visual results of any of the Presidents' speeches can be viewed easily. The analyst can therefore determine the answer of their research question quickly.

Personal Information. Figure6 (a) presents the information about the selected president, wherein the information of George Washington, including the picture, term, birth details, and political party, are displayed. The user can easily change this information using the several options on the right side (Figure6 (c)).

Visual Result. This graph pane (Figure6 (b)) was developed through a combining of Sunburst with RadVis. In the circular Sunburst, the POS and its combination patterns are presented for each gram. And the Presidents that are located inside were positioned according to their frequency value by the POS or the POS-combination patterns of the word. In addition, when the user selects a President, the frequency value is displayed on Sunburst.

Selection. Figure6 (c) helps the user to change the visual result according to their research question. This view is divided into the following three views: Political-party color, Select issues, and Select president. The buttons in the layer headers (Figure6 (c)) provide the option to change the node color according to the President's political party. It is also possible to remove the node color according to the President's political party in the visual result. The President's speech can be categorized into the following 11 issues (Hughps, 2009), (Andrade and Young, 1996): Economic growth, social welfare, U.S. policies (war), health care, immigration, humanitarian aid, protection of the U.S. (terror), establishment of democracy, promotion of U.S. strength, U.S. priorities, and all of the speeches. Through this selection view, the user can derive a visual result of his or her issue of interest.

Speech. Figure6 (d) presents the real speech according to the user selection, and this view can display the word through a highlighting of it in the speech. A user can also highlight words or an highlighting word and turn on or off the highlight words in the speech (Figure5 (a)).

Word Cloud. Figure6 (e) presents the result words using the Word Cloud (b) function to support the user in his or her attainment of a better understanding of the analysis results of the President's speech. This can be very helpful for the user's learning of the number and kinds of words that are present in the speech. And figure6 (f) provides more information such as a color legend about the words in Word Cloud and the sorted results. Additionally, when the user hovers over each word, it shows the support view including the POS type and the word count.

6 EVALUATION AND CASE STUDIES

Case studies and usage scenario were conducted to evaluate the effectiveness of the proposed interactive visualization and its usability. In this research, we worked with computational-linguistic researchers who study the multiword and possess expert knowledge on the subject. They used PreechVis to find information regarding their research questions and compared the details from the U.S. President Speech corpus.

6.1 Case Studies

This section demonstrates the exploratory use of the system regarding several research questions.

- **Q1: What kind of POS combination pattern exist and which POS combination is commonly used?**

This question was answered using a proposal of an interactive visual system that facilitates the display of the POS combination patterns of each gram. In the example of figure 7, each of the multiwords comprise a different POS combination. Further, 8 large and 36 specific word combinations are evident in the unigram, 50 large and 200 specific word combinations are evident in the bigram, and 65 large and 97 specific word combinations are evident in the trigram.

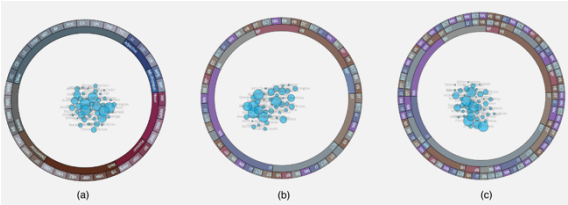
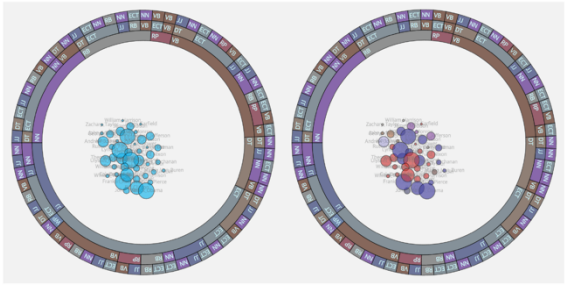


Figure 7: Word-combination Patterns by Part-of-Speech (POS). (a) POS combination in the uni-gram, (b) POS combination in the bi-gram, and (c) POS combination in the tri-gram.

- **Q2: Can we find groups that are grouped or divided by the POS and the POS combination?**

PreechVis presents the visual result based on the previously mentioned visual techniques.



(a) (b)

Figure 8: Visual Result in the tri-gram. (a) Displaying the President's group by the usage pattern and (b) changing the node color by the political party.

Figure 8 represents all of the Presidents that are inside the circle as nodes, where they are grouped in a single formation. This indicates that the patterns of the POS and the POS combination according to the Presidents are very similar. From the colors according to the political party, *'Republican'* is located in the middle of the group and *'Democrat'* is located both above and below; however, they are so close to each other, a significant difference is not evident.

- **Q3: Can we get more accurate results from the U.S. President Speech corpus?**

(a) (b)

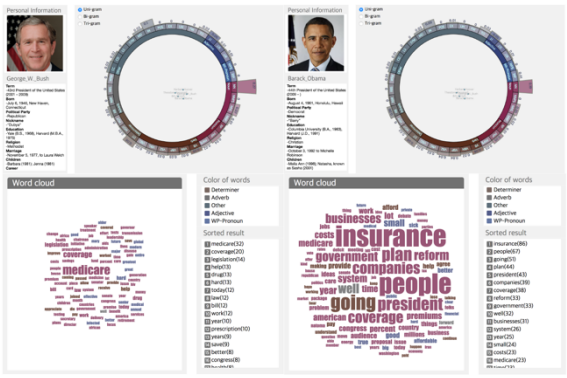
Figure 9: Analysis result of Harry Truman's speech by: (a) uni-gram and (b) bi-gram.

A serious error will occur in the analysis result if a researcher uses words with only a one-word meaning; for example, the word *'United States,'* and this word frequently appears in the speeches. However, if a multiword analysis is not used, the words *'United'* and *'States'* will account for a large proportion of the analysis results. The proposed visual tool, however, has addressed this problem, as shown in figure 9.

6.2 Usage Scenario

In the following subsection, a usage scenario that demonstrates the suitability of PreechVis for analysis tasks is presented. Additionally, a usage case that shows how the U.S. President's public speeches can be analyzed with the proposed visual system is described. The analysis of the usage case illustrates the effectiveness regarding a comparison of the Presidents' speeches.

For this part, the two U.S. Presidents George W. Bush and Barack Obama were compared in terms of two main issues. The first issue is *'Health Care'*. Figure 10 shows the visual result of the two Presidents for Health Care. In the case of Bush, he used words like *'Medicare (32)'*, *'coverage (20)'*, *'legislation (14)'*, and *'help (13)'* frequently in his Health Care speech. Alternatively, Obama commonly used words like *'insurance (86)'*, *'people (67)'*, *'going (51)'*, *'plan (44)'*, and *'president (43)'* in his speech.



(a) (b)

Figure 10: Comparison of Two Presidents for "Health Care" in the uni-gram.

Figure 11 represents the visual results of the bigram and the trigram. The users can change the visual result by using either of the three gram buttons in the top of the main view. In the bigram result, Bush and Obama frequently used *'health care (20, 71)'* in their speeches. However, Obama used more multiword (*'health insurance (31),' 'insurance company (9),' 'right thing (7),'* etc.) regarding Health Care than Bush. In the trigram, the proposed system

shows the multiword consisting of three words of each President. In the case of Obama, *'to make sure (4)'* and *'in the way (3)'* are frequently evident, as well as his passion on the word of Health Care.

Figure 11: Comparison of Two Presidents for "Health Care" in the bi-gram and the tri-gram.

To summarize, Obama is evidently more interested in health care than Bush according to a tally of the words and the size of the Word Cloud.

The second issue for the usage scenario is *'Protection of the U.S. (terror).'* The United States has previously received many terrorist attacks, including the *'September 11 attacks'* and the *'Oklahoma City bombings'*. The protection of the U.S. from terrorism is therefore a very important issue for the country's Presidents.

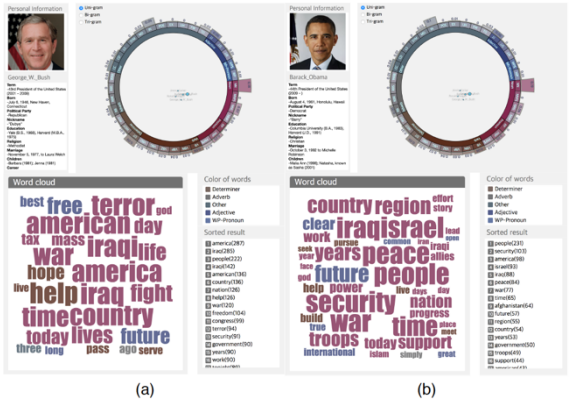


Figure 12: Comparison of Two Presidents for "Protection of the U.S. (terror)" in the uni-gram.

PreechVis was used to compare the two Presidents' opinions on the issue *'Protection of the U.S. (terror).'* Figure 12 represents the comparative analytic results between Bush and Obama on the Protection of the U.S. (terror) issue. In the results of Bush, words like *'America (287),' 'Iraq (285),' 'People (222),'* and *'Iraqi (142)'* feature many times in his speech. Alternatively, Obama commonly used words such as *'people (231),' 'security (103),' 'America (98),'* and *'Israel (93)'* in his speech. In these results, the different countries that were discussed in the issue Protection of the U.S. (terror) are evident.

Figure 13: Comparison of Two Presidents for "Protection of the U.S. (terror)" in the bi-gram and the tri-gram.

Figure 13 shows the visual results of the bigram and the trigram. In the visual results of Bush, terms like *'United States (77),' 'Al Qaeda (61),' 'Saddam Hussein (50),'* and *'Middle East (48)'* often appear in his speech regarding the Protection of the U.S. (terror). In particular, *'Al Qaeda (61)'* and *'Saddam Hussein (50)'* appear frequently in the results. Therefore, the user can easily identify the organization and the person that are involved in this case. In the visual results of Obama, terms like *'United States (76),' 'Al Qaeda (50),' 'Bin Laden (21),'* and *'Middle East (17)'* frequently appear. It is also possible to recognize the change of the person who is involved in the case depending on the President. In the trigram result, Bush and Obama used two of the same terms (*'in the middle (22,10)'* and *'Osama bin Laden (3,8)'*) in their speeches. However, *'in the middle (22,10)'* may be the result of *'in the Middle East,'* which is a four-word combination. Therefore, this matter needs to be supplemented in a future work.

In summary, more accurate information can be recognized, and the information in the corpus data can be understood quickly through the use of PreechVis.

7 DISCUSSION

Several interviews were conducted with domain experts who study computational-linguistic and natural-language processing, and these researchers agreed that the exploration of the multiword for each gram is a major strength of PreechVis. Further, the proposed visual tool can facilitate a quick exploration of the corpus information and more accurate results can be obtained, as shown in the presented case studies. The case studies three important implications and usage scenarios confirm the usability and effectiveness of the system. This study presents two important implications.

First, the study reveals a data-processing method that can obtain more accurate results than when we do analysis without multiword. This study uses a linguistic approach to obtain more accurate word-combination words, and this was explained in the data-processing without manual work. As a result of this study, the proposed visual system shows more accurate analysis results because it cared both single word and multiword together.

Second, multiwords are consist of different POS combination pattern and commonly used POS combinations have existed. Through our visual results, we found the multiwords have a certain pattern of part of speech combination. For example, *'united states'* is generated by a combination of *Verb* and *Noun*. And we found commonly used POS combinations have existed. For instance, *'Noun'* is usually used part-of-speech in uni-gram, *'Adjective+Noun'* is usually used POS combination pattern in bi-gram and *'Noun+Preposition+Noun'* is usually used POS combination pattern in tri-gram.

The computational-linguistic and natural-language-processing researchers plan to develop a system that automatically recognizes the multiword without manual works. this is the beginning of the study of the generalization to recognize the multiword.

And the proposed system can be made more generic by finding other usage cases of this tool; for example, authors instead of U.S. Presidents, and writings instead of speeches.

The present work, however, is hampered by limitation that this study covers the expressions of words consisting of three-word combinations. However, the absence of words consisting of four-word combinations is problematic; for example, *'in the Middle East,'* etc. Therefore, this matter needs to be supplemented in a future work.

Overall, though, the feedback is positive, and the experts extracted new findings and gained more information from the corpus.

8 CONCLUSION

For this work, a new interactive-visualization approach, PreechVis, was designed and demonstrated regarding the analysis of corpus data. This visual tool can help users to understand the corpus. The proposed system was developed using the multiview and a novel technique to show the analysis result of the corpus with the multiword. PreechVis supports a flexible exploration of the multiword in the corpus, and the POS and the combination patterns of the POS for each gram can be identified. The case studies and the usage scenario demonstrate how this tool can be used.

In the near future, we will do the analysis to find out more POS combination pattern in multiword and use it in the data processing part.

ACKNOWLEDGEMENTS

This work was supported by the Ajou University research fund, the BK21 Plus project in 2018 and National Research Foundation of Korea (NRF-2015S1A5B6037107)

REFERENCES

Andrade, L. and Young, G. (1996). *Presidential agenda setting: Influences on the emphasis of foreign policy.* Political Research Quarterly, 49(3):591–605.

Andrews, K., Wohlfahrt, M., et al. (2009). *Visual graph comparison.* In 2009 13th International Conference Information Visualisation.

Collins, C. and Carpendale, S. (2007). *Vislink: Revealing relationships amongst visualizations.* IEEE Transactions on Visualization and Computer Graphics, 13(6):1192–1199.

Cui, W., Liu, S., et al. (2014a). *How hierarchical topics evolve in large text corpora.* IEEE Transactions on Visualization and Computer Graphics, 20(12):2281–2290.

Cui, W., Wang, X., et al. (2014b). *Let it flow: a static method for exploring dynamic graphs.* In 2014 IEEE Pacific Visualization Symposium. IEEE.

Heimerl, F., Han, Q., et al. (2016). *Citerivers: Visual analytics of citation patterns.* IEEE Transactions on Visualization and Computer Graphics, 22(1):190–199.

Hughps, J. (2009). *Analyzing policy issues in presidential speeches and the media: An agenda-setting study.* In University of Nevada, Las Vegas, USA.

JR, F. (1957). Papers in linguistics 1934-1951. Oxford University Press.

Koch, S., John, M., et al. (2014). *Varifocal reader in depth visual analysis of large text documents.* IEEE Transactions on Visualization and Computer Graphics, 20(12):1723–1732.

Liu, C. and Wang, P. (2015). *A sunburst-based hierarchical information visualization method and its application in public opinion analysis.* In 2015 8th International Conference on BioMedical Engineering and Informatics (BMEI 2015).

Lu, Y., Steptoe, M., et al. (2016). *Exploring evolving media discourse through event cueing.* IEEE Transactions on Visualization and Computer Graphics, 22(1):220– 229.

Mahyar, N. and Tory, M. (2014). *Supporting communication and coordination in collaborative sense making.* IEEE Transactions on Visualization and Computer Graphics, 20(12):1633–1642.

Qu, S., Wang, S., et al. (2008). *Improvement of text feature selection method based on tfidf.* In 2008 International Seminar on Future Information Technology and Management Engineering.

Ramisch, C. (2015). *Multiword Expressions Acquisition.* Springer.

Rodden, K. (2014). *Applying a sunburst visualization to summarize user navigation sequences.* In IEEE Computer Graphics and Applications. IEEE.

Rubio-Snchez, M., Raya, L., et al. (2016). *A comparative study between radviz and star coordinates.* IEEE Transactions on Visualization and Computer Graphics, 22(1):619–628.

Sharko, J., Grinstein, G., et al. (2008). *Vectorized radviz and its application to multiple cluster datasets.* IEEE Transactions on Visualization and Computer Graphics, 14(6):1444–1451.

Sidorov, G., Velasquez, F., et al. (2013). *Syntactic dependency based n-grams: More evidence of usefulness in classification.* In Springer Verlag Berlin Heidelberg 2013. Springer.

Stasko, J., Gorg, C., et al. (2007). *Jigsaw: Supporting investigative analysis through interactive visualization.* In IEEE Symposium on Visual Analytics Science and Technology 2007. IEEE.

Sun, G., Wu, Y., et al. (2014). *Evoriver: Visual analysis of topic coopetition on social media.* IEEE Transactions on Visualization and Computer Graphics, 20(12):1753– 1762.

Wang, X., Liu, S., et al. (2016). *Topic panorama: A full picture of relevant topics.* IEEE Transactions on Visualization and Computer Graphics, 22(12):2508–2521.

Wu, Y., Liu, S., et al. (2014). *Opinion flow: Visual analysis of opinion diffusion on social media.* IEEE Transactions on Visualization and Computer Graphics, 20(12):1763–1772.

Zhang, W., Yoshida, T., et al. (2008). *Tfidf, lsi and multi-word in information retrieval and text categorization.* In 2008 IEEE International Conference on Systems, Man and Cybernetics. IEEE.