ISSN No:-2456-2165

Analyzing the Semantic Structure of Discourse Algorithmically

Balaanz Dagiimaa
University of the Humanities, Ulaanbaatar, Mongolia

Abstract:- Studying the main idea and tone of speech using an algorithm is of theoretical and methodological significance for determining the phenomena and their meaning structure in discourse studies. In this study, the frequency of keywords (topics) in the meaning structure of speech is mathematically ranked as an index, and the methods of creating a macrostructure (mapping) are used. We numerically and indexically ranked the frequency and strength of relevance of themes to create a structure map of the discourse's meaning, exemplifying the ideas of "first ladies" about women's rights. In doing so, we intended to determine the main ideas of speech and speech by keywords and to determine the spread and impact of those keywords using an algorithm-based method.

Keywords:- *Macrostructure, mental models, semantic structure, keyword distribution, topic modeling, mathematical-statistical methods.*

I. INTRODUCTION

Using artificial intelligence to identify the content and tone of speech is emerging as a creative way to analyze speech and text in applied linguistics. Algorithm-based neural network (text networking) is useful for defining phenomena and semantic macrostructure by expressing ideas, conveying ideas, understanding text and speech tone (nuance), determining mental models, and finding solutions to problems.

This method is based on the method of creating a mapping made of semantically related nodes by ranking the frequency of topics and the strength of correlation in the structure of the discourse value as an index or numerical value.

Mapping is the main idea of discourse which can be defined as macrostructure, global structure, superstructure, text network, etc. In other words, the key concepts expressing the idea are their oriented and weighted network of meanings. This is defined as topic modeling, mental model, etc. in artificial intelligence. This research work has selected the topic modeling method for grouping keywords that express the main ideas of text and speech from mental modeling and optimization methods.

- The purpose of this research: using open-source programs to determine the content of the discourse and speech by keywords, and to study the distribution and influencing power of those words in an algorithm-based method, we aimed to determine "the ideas put forward by the first ladies about women's rights and the meaning structure of the discourse"
- Research method: "Bag-of-Words" (BoW), co-occurrence, semantic or syntactic information (Distributional Hypothesis) in order to create a text meaning network using NLP neural network, determine the probable meaning by algorithm, and the meaning mapping. Dimensional lambda matrices (Rousseau and Vazirgiannis, 2013) and document classification (Rousseau et al., 2015; Malliaros and Skianis, 2015) methods were used.
- Research materials: Speech[1] by Hillary D. Clinton, (American politician, diplomat, lawyer, 67th US Secretary of State, US senator, former US First Lady, and presidential candidate) the speech[2] by Michelle Obama. (American politician, lawyer, former first lady, graduate of Princeton University and Harvard Law School,) selected speeches are used for speech analysis.

II. USING ALGORITHMS FOR ANALYZING TEXT SEMANTIC NETWORKS (TSN)

NLP is an interdisciplinary knowledge science that combines linguistics, computers, and artificial intelligence and is a branch of cognitive science that studies human speech as a programming language. It is a method of studying language through the language database, such as recognizing and understanding the content of any information, speech, or discourse, recognizing the tone of the topic, analyzing the text, etc. In other words, it is developing as an artificial intelligence method because it can imitate all human mental and cognitive functions. The basic principle is to determine the keywords that express the main content of the text and speech and determine its meaning range. This means that algorithms can identify the gist of human speech. Moreover, algorithms are models that determine the sequence of operations to be performed on any dimension and create a general pattern of content. In other words, it is a mathematical method that models common patterns and standards as a sequence of operations.

---

[1] Hillary Clinton's Women's Rights Are Human Rights' (https://academyatthelakes.org/),

[2]Macrostructure of the Speech by Michelle. 'O (Keynote Address at Young African Women Leaders Forum - June 22, 2011)" (https://obamawhitehouse.archives.gov/)

ISSN No:-2456-2165

Humans use many algorithms in their daily activities, but they are not always aware of them. For example, a fixed order is followed for the day's work and a specific task. The concept of the algorithm first arose in mathematics in connection with the attempt to find a general way of thinking about similar problems. The term algorithm was coined by the Persian mathematician Muhammad Ibn Mussa-Al Khorezm. In modern times, the theory of algorithms has developed as a branch of mathematical sciences, while algorithmic models are used to connect interdisciplinary problems of social sciences with artificial intelligence. Algorithmic modeling of discourse and speech structure is a mathematical method of modeling the human mind. It is being used as a way to model ideas with artificial intelligence.

In other words, it means that textual and semantic methods such as classifying texts, determining meaning by topics, ranking keywords, automatically generating discourse, analyzing literature, and writing essays can be done by machine (Saito, 2013).

- ***Semantic Neural Networks*:** Semantic neural networks are semantic macrostructures that mathematically describe the distribution, placement, and relational links of words in a given discourse. "Embedding" (word embedding) used by us is a vector measure that defines and encodes the value of the word in the context of the topic. In other words, the meaning of the word is mathematically transferred to the vector space, and a neural network of meaning is created. Since words occur in the same contexts and have the same vector space, it means that the semantic network of text and speech is determined by the order of words that occur in close space.
- ***Propagation (propagation of ideas)***: Determining the vector space of approximate values, which are sorted by algorithms, is important for the study of text networking analysis. **See (fig.1)** how to understand the main idea by determining the location of keywords that represent the main content of any text.

A. Integrating knowledge optimization and discourse meaning structure

➢ *Topic modeling:*

One of the artificial intelligence methods used in discourse studies is knowledge modeling optimization and text mining (TM). It is a machine-based information integration method that identifies key concepts by optimizing and modeling text and speech. Converting a large amount of information into numerical values is useful in many ways, such as mapping the value network or mapping text, determining the distribution of keywords, and determining sentiment. Among the many algorithm-based methods for determining the semantic structure of the text, Probabilistic-Semantic (LDA) is the main one. (Papadimitriou, Raghavan, Tamaki, and Vempala, 1999), (David Blei, Andrew Ng, Michael I. Jordan, p. Dirichlet prior distributions), (Suyakur). The mental model that defines the general content of the text is not a fixed model but can be formed according to the structure of meaning. One of the ways to optimize knowledge is to define the meaning network of the given discourse and consider nodes with meaning relationships as mental models.

A mental model is the author's ideas (idea) in the text and can have any number of models (topics). In this TM method that we have adopted, the meaning structure of the discourse can be determined by several methods and the ranked meaning of each group can be determined. I tried to analyze and evaluate those methods on the example of the speeches of the first ladies.

Discourse is a branch of knowledge inextricably linked to the semantics of identity and the psychology of identity. This logically connected knowledge is called a mental model (Garnham 1987; Johnson-Laird 1983; van Dijk & Kintsch 1983; Van Oostendorp & Zwaan 1994). A mental model is knowledge-based on underlying concepts, implications, and facts that emerge from context. A mental model is a fundamental problem of language, thinking, behavior, and cognition, and it is a set of thoughts and minds, including the perception, perspective, and worldview of each individual.

In short, they are ideas that are imagined in the human brain (Hall et al. 1994, Swan and Newell 1998, Sterman 2000). A person's mind is a psychological problem, including multifaceted views, attitudes, perceptions, and ways of expression. The information stored in the human mind is long-term memory (LTM), which is defined as "Mental models and semantic knowledge" in the field of discourse studies and cognitive studies, and is used as a basic concept for the study of ideas.

B. Study section for analyzing macrostructure

The macrostructure of the speeches taken as our example and the author's ideas are mapped by the text network (Fig.3) and the keywords are determined by the power of correlation. Also, the author's ideas are made into several mental models and each part is highlighted in different colors so that the main idea of the speech can be seen from each color. For example, "...the issue of women's rights is the main issue of human rights, and politics, government, and the whole world are involved in it..." and it is shown as a group of ideas in different colors.

ISSN No:-2456-2165

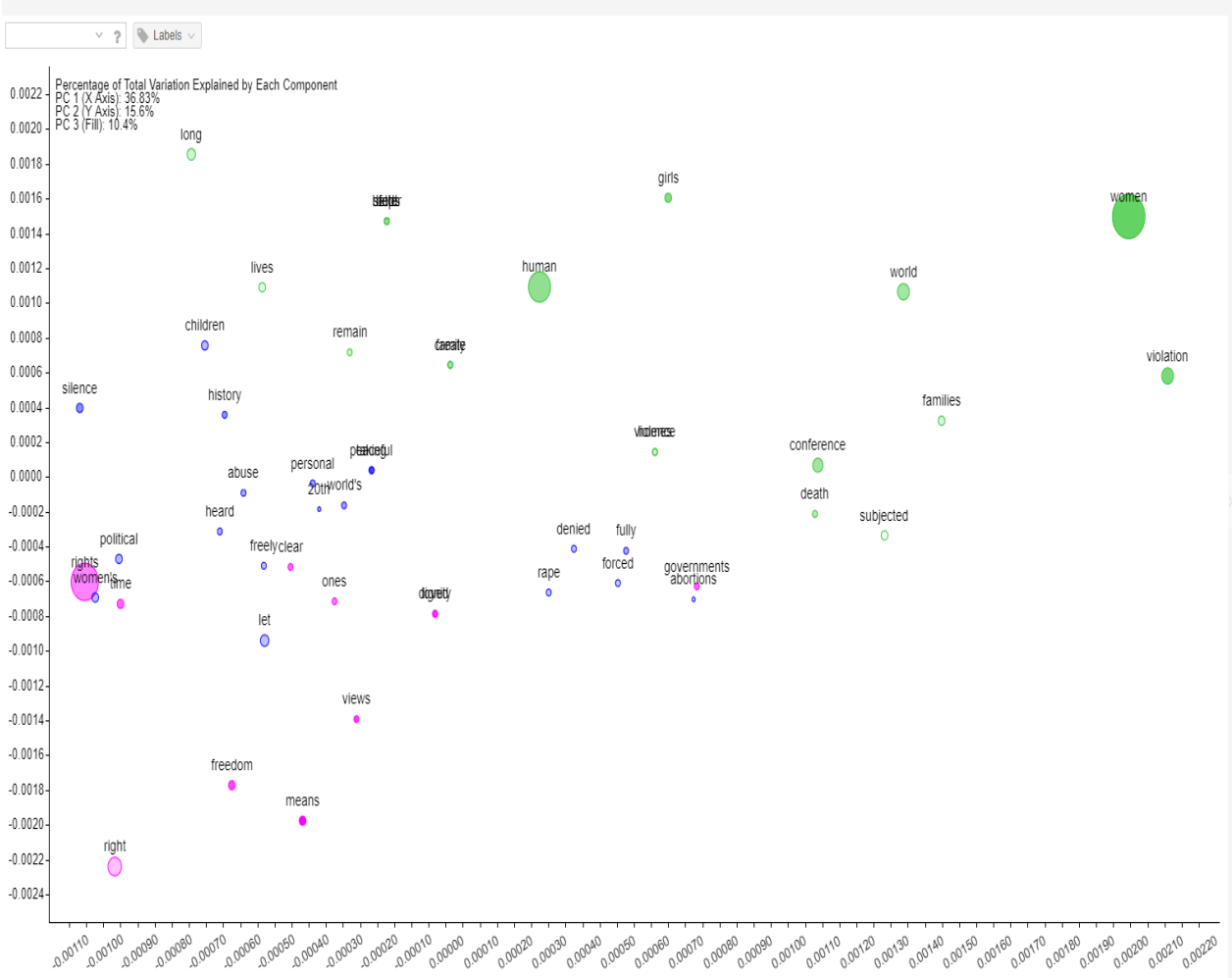


Fig. 1: The locality of the keywords in speech-1

The vector space of the keyword in the text value network is determined by the speech "Hillary Clinton's Women's Rights Are Human Rights" as shown in (Fig.1), the distribution of the keyword by the method of creating a distribution polygon using the PSA algorithm PC 1: (X-Axis) 36.83%, PC 2 (Y-Axis) is 15.6%, PC 3 (Full) is 10.4% or normal distribution.

Table 1: Index of the structure of the discourse

Nodes	degree	frequency	Co-occurrence	conductivity	locality	diversity
Total vocabulary	**826**	**261**	**3.106458**	**3176.4**	**365**	**11115**
142 nodes	**5.82**	**1.84**	**0.021876**	**22.37**	**2.57**	**78.27**

Fig. 3: Macrostructure and its mental model's speech by Hillary Clinton 'Women's Rights are human right'

The macrostructure of the speech was defined by the algorithm *tfidf(t,d,d)=tf(t,d)*idf(t,d)*. According to the structure of the speech (fig.4.), 24% of the total information is about political actions and groups; 19% is about women's rights and long-term goals; 14% is about government and institutional policies; 11% is global approach and engagement; a group of values is created as a node, which is depicted in different colors depending on the strength of the relationship and the location, multi-angle feature. They are mental models that express the main ideas of the author.

The results of the discourse structure analysis are as follows. *Semantic structure*: modeled (modulation) state (mean >0.4), (multifaceted, the higher the alpha of the influence distribution, the higher its mental strength).

The entropy of nodes: "high" level is considered high "*propagation*" of speech. *Discourse structure:* "dispersed" dispersed (0.77) is a style with very high internal power. The structure of the discourse in this network is "normally distributed", not only does it show the multiple angles of generating new ideas and the further dissemination of ideas, but also the speaker's manner of speaking moves from topic to topic. The high internal strength of the discourse structure is considered to be a style in which few main ideas are used, secondary ideas are resolved, and many points of view are presented. A number of mental models that form the main idea of the discourse are considered to be the author's own point of view, attitude, and position. Frequency: total vocabulary and topic frequency networks are not all visible above, but they are angles to show the diversity of the narrative and transfer ideas to other issues. The topicality of the main content and the percentage of information content (percentile) are shown in.

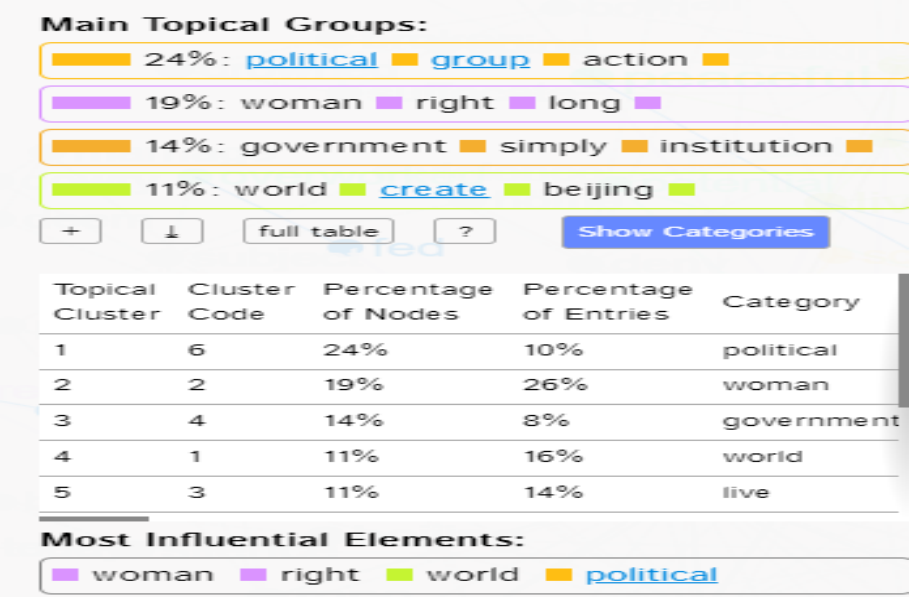


Fig. 4: Graphical representation of the main content percentile

Fig. 5: Shows the graphic of the text into ten parts (paragraphs) and the appearance of the main keyword propagation in each paragraph

High-Frequency Themes: (Fig.4. and Fig.5.) We can see how the main theme and the most influential keywords change over the speech is, X-axis: *Lemma level*: (divided into 10 sections). Y-axis: cumulative number of occurrences (change in influencing topics) (slower, but more accurate). A smoother, more paced spread means the main idea or agenda is stronger. (See alpha exponent (~ 0.5 and below), *Distribution Dynamics and Cyclical Variability*: alpha exponent: 0.53 | middle ks: 0.64, d: 0.27 <= cr: 0.58.

Table 2: Index of the structure of the discourse

nodes	degree	frequency	Co-occurrence	Conductivity	locality	diversity
Total	**1768**	**749**	**2.34004**	**1032.5**	**1195**	**2426.6**
Nodes 150	**11.79**	**4.99**	**0.0156**	**6.88**	**7.97**	**16.18**

ISSN No:-2456-2165

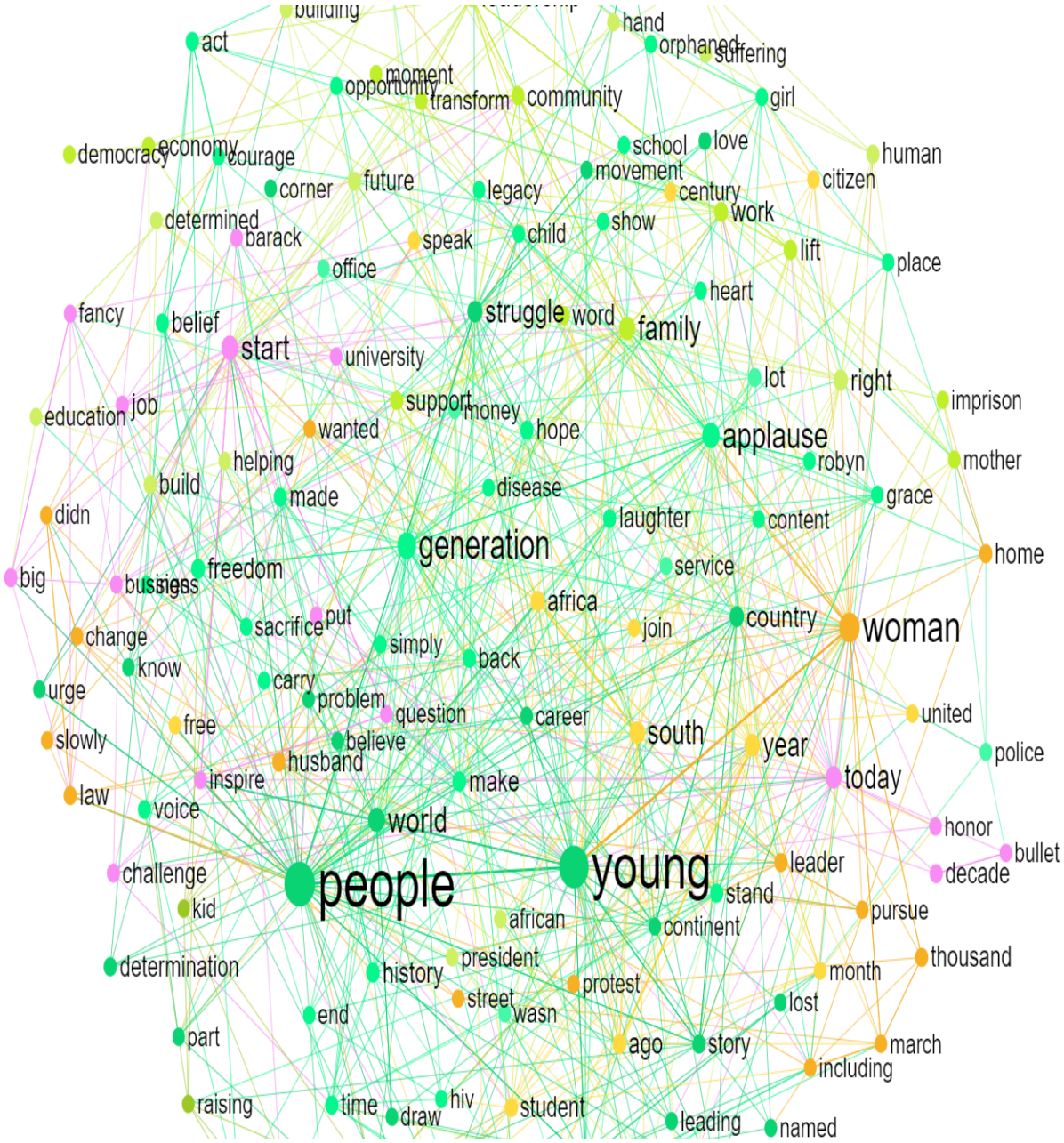
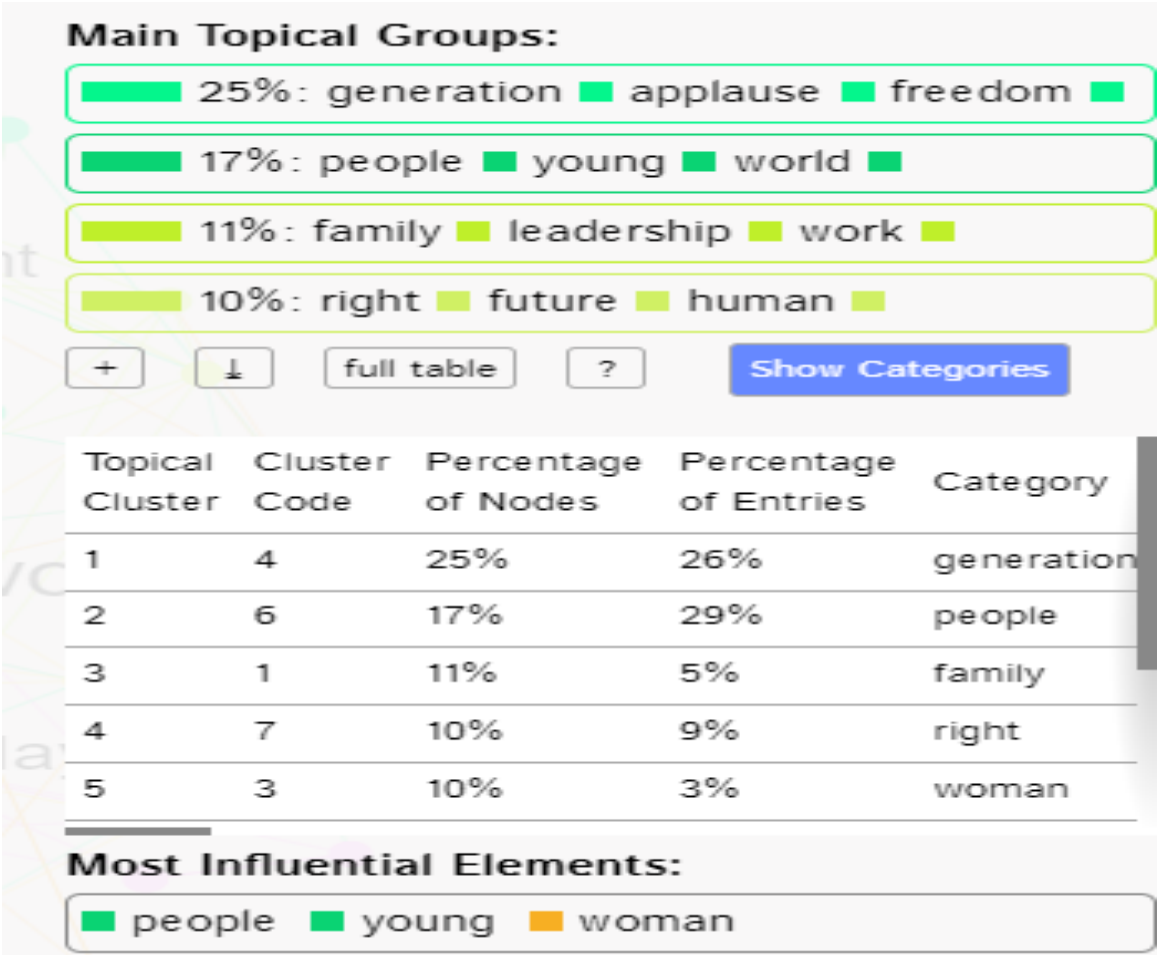


Fig. 6: Macrostructure of the Speech by Michelle. 'O

The semantic macrostructure of the speech (**Fig.6.)** is showing the relationship between its main idea and content with tabular information. According to the example (**Fig.7**), 25% of the youth and listening to their voices, about the youth and women and girls' problems, the response of girls and women; 17% about giving the youth a voice and taking the fight against violence to the world; 11% of families raised communication and workplace issues; 10% of the rights of girls addressed the future of humanity. A group of values and several nodes were created. It is explained above (picture 3) that they are mental models that express the main ideas of the author. This discourse structure is a type of "*focused*". This pattern (**Fig.6**.) is characterized by a long focus on one issue (high proportion of superclusters, high focus on one topic despite the presence of some diversity of views) alpha exponent; fractal (|0.85 < alpha < 1.15). *Text (network) structure*: (**Fig.6**) *influence distribution alpha*: modeled (modulated) state (mean >0.4).

ISSN No:-2456-2165



Topical Cluster	Cluster Code	Percentage of Nodes	Percentage of Entries	Category
1	4	25%	26%	generation
2	6	17%	29%	people
3	1	11%	5%	family
4	7	10%	9%	right
5	3	10%	3%	woman

Fig. 7: Main content thematic status and discourse structure

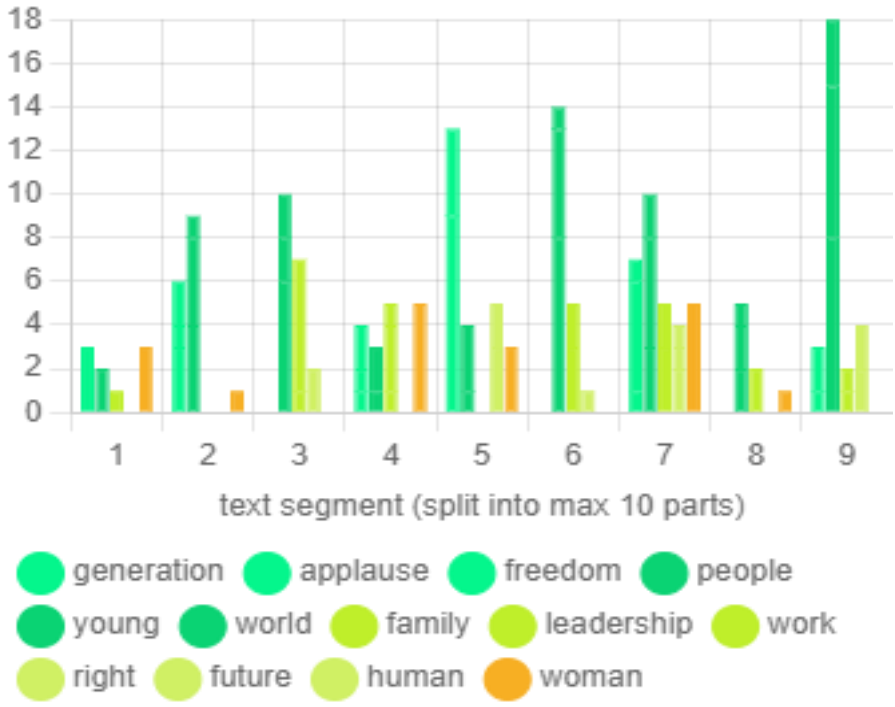


Fig. 8: The graphic is showing the most overlapping topics of the discourse. It's divided into ten main parts (paragraphs) and the occurrence of the main keyword propagation in each paragraph.

ISSN No:-2456-2165

The entropy of the nodes (**Fig.7**) occurred at a high level between the top clusters. Nodes that were tending to co-occur within the same theme (betweenness) are marked by color. *Word frequency*: the total vocabulary of a topic cannot be seen on the text grid, but it is important to show the diversity of the narrative and to transfer ideas to different issues. *An analysis for text networking*: (X) axis: (**Fig.8**) Lemma level Transition from lemma to lemma (divided into 10% sections). (Y) axis: (Fig.8) *Impact distribution*: 50%, *distribution dynamics*: cyclical fluctuations | alpha exponent: 0.56 ks: 1.88, d: 0.47 > cr: 0.34 | (based on the Kolmogorov-Smirnov test).

Macrostructure (**Fig.6**)"Speech by Michelle. O)", the normal distribution of influencing power of speech indicates the level of speaking ability, speech strategy, and knowledge.

The structure of the speech has peripheral and core value sections. There are common patterns of content prototypes that define the main idea in the content structure. The core structure is centered on "the position of young people and girls in society and the social factors that lead to violence". According to the content of speech, the issue of modern speech, speaking, and writing author's use of language has the model of "knowledge-optimization-expressiveness".

The entropy of the nodes (**Fig.7**) occurs at a high level between the top clusters. Nodes that tend to co-occur within the same theme (betweenness) are marked by color. *Word frequency*: the total vocabulary of a topic cannot be seen on the text grid, but it is important to show the diversity of the narrative and to transfer ideas to different issues. (X) axis: (**Fig.8**) Lemma level Transition from lemma to lemma (divided into 10% sections). (Y) axis: (**Fig.8**) Topic influencing power: Impact distribution: 50% distribution dynamics: cyclical fluctuations | alpha exponent: 0.56 ks: 1.88, d: 0.47 > cr: 0.34 | (based on the Kolmogorov-Smirnov test).

Macrostructure (Figure 6) "Speech by Michelle. O, the normal distribution of influencing power of speech indicates the level of speaking ability, speech method, and knowledge.

The structure of the presentations has peripheral and core value sections. There are common patterns of content prototypes that define the main idea in the content structure. The core structure is centered on "the position of young people and girls in society and the social factors that lead to violence". According to the reports, the problem of modern speaking, speaking, and writing author's use of language has the model of "knowledge-optimization-expressiveness".

III. RESEARCH RESULTS

Comparing the two speeches considered to be the best articles on women's issues by first ladies of the United States, "Speech.1" or First Lady Hillary Clinton's speech network (network): multifaceted, high alpha of influence distribution, the high inner strength of mind, entropy of nodes: Discourse structure at a high level: *Dispersed* (0.77) (very high internal power), with few primary ideas, secondary ideas, and many points of view are considered to have high influencing power. Recurring topics: The main topic and most influential keywords change over the course of the speech, leading to different ideas. On the *Y-axis*: impact force (slower, but more accurate). Distribution dynamics and cyclical volatility: alpha exponent: 0.53 | *Variability*: more based on core concepts.

But "Speech.2" or former first lady Michelle Obama's speech *network structure*: (multifaceted problems shown) influence distribution alpha: (more adaptive network) modeled state: (mean >0.4), *Discourse structure:* "focused" one problem long focused on (a high proportion of superclusters, a certain degree of diversity of opinion, but a focus on one topic) alpha exponent; multifaceted “fractal” pattern (|0.85<alpha<1.15) *Entropy of nodes*: occurs at a high level between the top clusters. Within the same theme (betweenness), the relative power of the nodes is marked by the color (Y) axis: Power of the theme: *Distribution of influence*: 50% Distribution dynamics: cyclical fluctuations | alpha exponent: 0.56.

The use of keywords in political speech is considered to be the main method of speech to express ideas. Mental models that form the structure of meaning have a "normal" distribution within the problem to be solved, forming the core of the content. It also shows that the selective use of eloquent language tools gives importance to the speaker's speech strategy and rhetorical elements and forms discursive features.

IV. CONCLUSION

The main ideas of the discourse and speech were determined by keywords, and the spread and influencing power of those words were studied, in the framework of the algorithm-based method, on the speeches about women's rights. One of the artificial intelligence methods used in discourse studies is optimization and text mining (TM) methods for modeling this knowledge. This method has the practical significance of optimizing the integration of information or idea by identifying key concepts and optimizing text and speech.

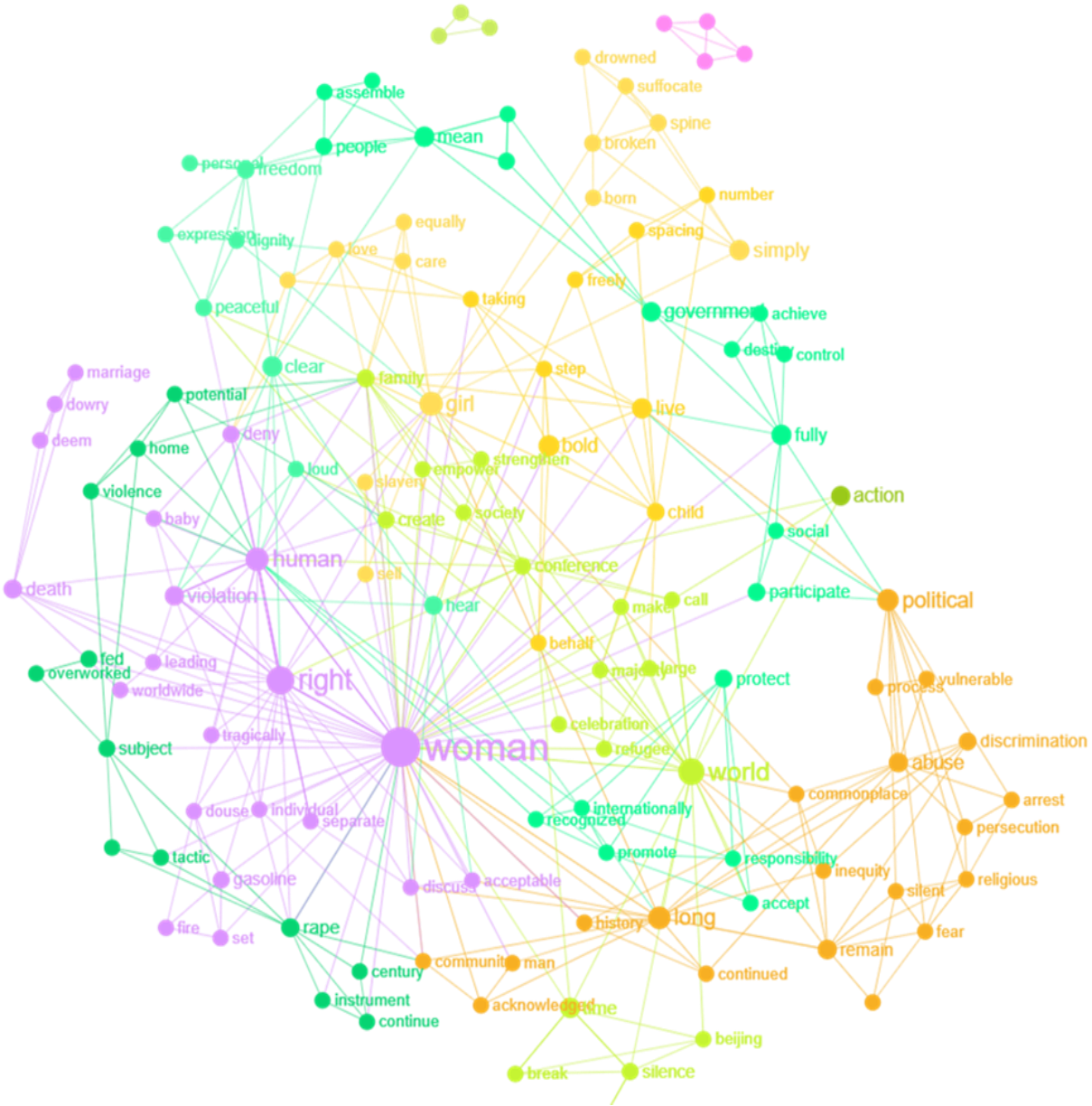
The methods of mapping the structure of discourse meaning by ranking the frequency of topics and the strength of correlation as indices using a quantitative method were used in this study. It is common in speech that the main idea is determined by keywords, and the distribution and influencing power of those words are determined. The structure of the "ideas proposed by the first ladies about women's rights" has peripheral and core meaning parts and the main keywords. In the structure of the content, the general content of defining the original character (prototype) of the core value is similar to the speeches. The core content is centered on "the position of young people and girls in society and the social factors that lead to violence." Studying the content and tone of speech using an algorithm is methodologically important to identify phenomena and semantic connections in discourse studies.

Science, 1989.

REFERENCES

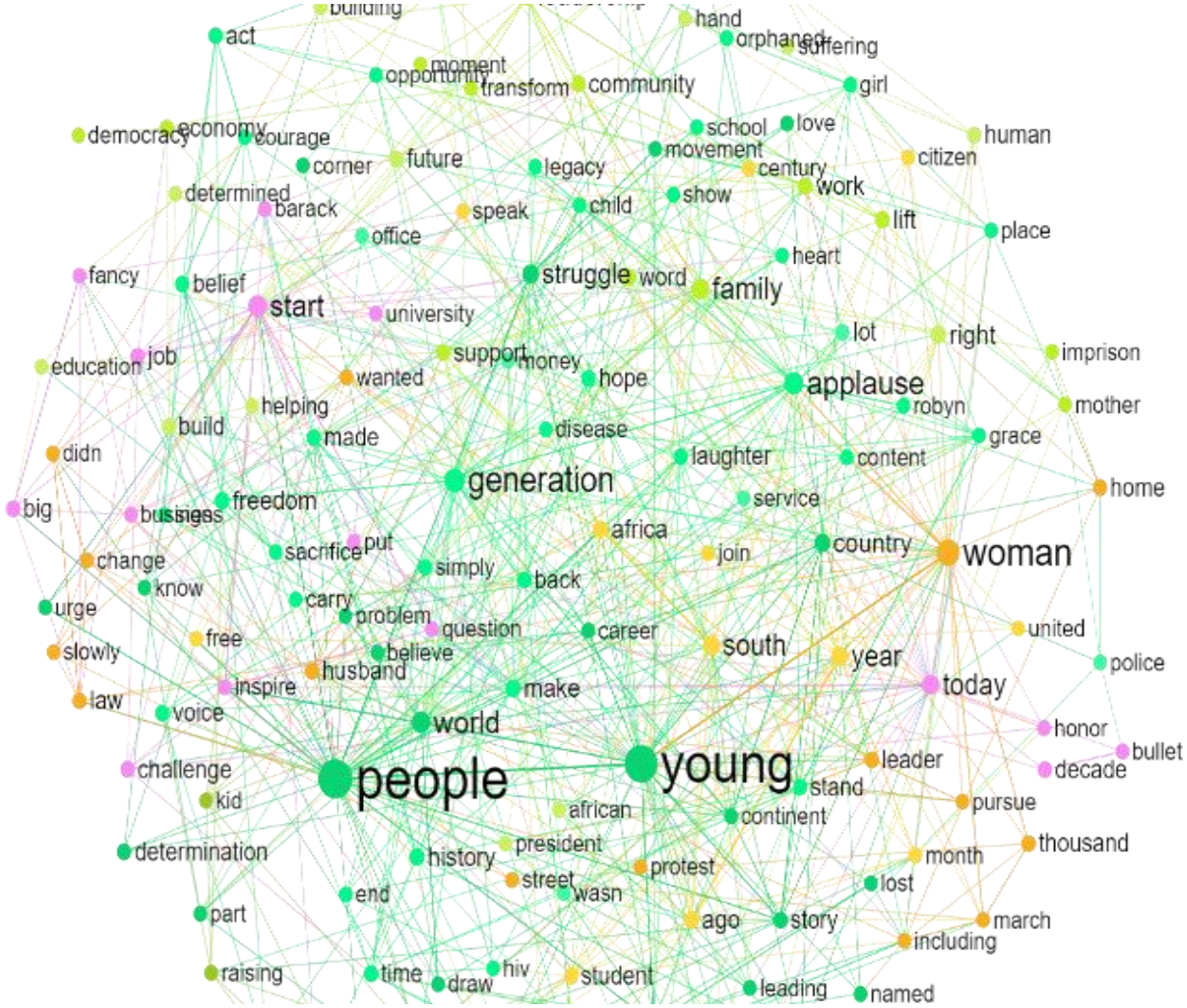
[1.] B.Dagiimaa(2019). Mental space and mental lexicon: results of an experiment. Journal of European Studies.
[2.] Bloomfield. (1998). The MIT Encyclopedia of Cognitive Science
[3.] Van Dijk in Wodak & Chilton, 2005, p. 71-72. Critical Discourse Analysis.
[4.] Bloomfield. (1998). The MIT Encyclopedia of Cognitive
[5.] Van Dijk in Wodak & Chilton, 2005, p. 71-72. Critical Discourse Analysis.
[6.] Staggers, Nancy; Norcio, A.F. (1993). "Mental models: concepts for human-computer interaction research". International Journal of Man-Machine Studies.
[7.] M. Minsky, "A Framework for Representing Knowledge," Artificial Intelligence, Memo no. 306, 1975.
[8.] Sowa, "Semantic Networks," in Encyclopedia of Cognitive Science, 1992.
[9.] K. Carley, "Extracting Team Mental Models through Textual Analysis," Journal of Organizational Behavior, 1997.
[10.] D. Jonassen and Y. Cho, "Externalizing Mental Models with Mindtools," in Understanding Models for Learning and Instruction, Boston, Springer, 2008.
[11.] S. Sonawane and P. A. Kulkarni, "Graph-based Representation and Analysis of Text Document: A Survey of Techniques.," International Journal of Computer Applications, vol. 96, no. 19,
[12.] D. Paranyushkin, "Addresses to the Federal Assembly of the Russian Federation by Russian presidents, 2008–2012: comparative analysis," Russian Journal of Communication, vol. 5, no. 3, pp. 265-274, 2013.

ISSN No:-2456-2165



Nodes	degree	frequency	Co-occurrence	conductivity	locality	diversity
Total vocabulary	**826**	**261**	**3.106458**	**3176.4**	**365**	**11115**
142 nodes	**5.82**	**1.84**	**0.021876**	**22.37**	**2.57**	**78.27**

ISSN No:-2456-2165



nodes	degree	frequency	Co-occurrence	Conductivity	locality	diversity
total	**1768**	**749**	**2.34004**	**1032.5**	**1195**	**2426.6**
Total nodes 150	**11.79**	**4.99**	**0.0156**	**6.88**	**7.97**	**16.18**