Hypergraph-Based Metrics for Semantic Ambiguity and Structural Overlap in NYT Connections

Neysa Alya Mukhbita - 13525080
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail: neysalyae@gmail.com , 13525080@std.stei.itb.ac.id

***Abstract*—NYT Connections is a word puzzle where players must group 16 words into four groups. This study proposes a hypergraph-based model for representing the puzzle structure and introduces two metrics: the Ambiguity Score (AS) and Structural Overlap Score (SOS). Words are represented as vertices, while solution categories and candidate categories which are generated using a WordNet-based coherence function are represented as hyperedges.**

The proposed metrics were experimented on a dataset of 40 different NYT Connections puzzles and the result shows that both metrics successfully represent semantic structure with hypergraph, but not puzzle difficulty, since both metrics decrease from Yellow to Purple. This also indicate that higher-difficulty categories often rely on wordplay, abbreviations, and cultural references. Therefore, AS and SOS are better interpreted as measures of semantic structure rather than difficulty.

Kata Kunci—hypergraph; NYT Connections; puzzle difficulty; WordNet; ambiguity score; structural overlap score

I. Introduction

In recent years, digital word puzzle games have become very popular. One well-known example is *NYT Connections*, a daily game created by *The New York Times* and released on June 12, 2023. In this game, players must group 16 words into four groups, where each group contains four words that are related in some way. Since its release, *Connections* has become the second most popular game from The New York Times after *Wordle*, and it has been played by millions of people around the world. This shows that the game is not only entertaining but also interesting to study from a mathematical and problem-solving point of view.

Although the rules of *Connections* are simple, the difficulty of each puzzle can be very different. Some puzzles are easy because the relationships between words are clear. However, others are harder and require more thinking, guessing, and elimination. So far, puzzle difficulty is usually judged by the game designer or based on players personal opinions. There is no clear mathematical method to analyze the structural properties that may contribute to puzzle difficulty.

Fig I. Connections Logo (Source: https://www.nytimes.com/games/connections)

One thing that may affect difficulty is ambiguity in the word relationships. A word can sometimes belong to more than one possible group, which can confuse players. In addition, similarities between different groups can also make the puzzle harder to solve. This shows that difficulty does not only depend on the number of words or groups, but also on how the relationships between words are structured.

To study this structure, this research uses a hypergraph approach. Unlike a normal graph that connects two points, a hypergraph can connect more than two points in one connection. This fits the structure of *NYT Connections*, because each group contains four related words. Therefore, hypergraphs are more suitable for modeling the relationships between words.

Based on this approach, this study aims to develop hypergraph-based measurements to analyze semantic ambiguity and structural similarity in *NYT Connections* puzzles. The study further investigates how these metrics relate to the official difficulty assigned by The New York Times.

II. Literature Review

A. Hypergraph Theory

A hypergraph is an extension of a graph that allows one connection (called a hyperedge) to link more than two vertices at the same time. In contrast, regular graphs only show relationships between two points. Because of this, hypergraphs can represent group relationships in a single structure.

Hypergraphs are used in many areas such as signal processing, social networks, bioinformatics, databases, and

machine learning. Their ability to represent relationships involving multiple elements makes them suitable for modeling the structure of the *NYT Connections* puzzle, where each group of four words can be represented as one hyperedge.

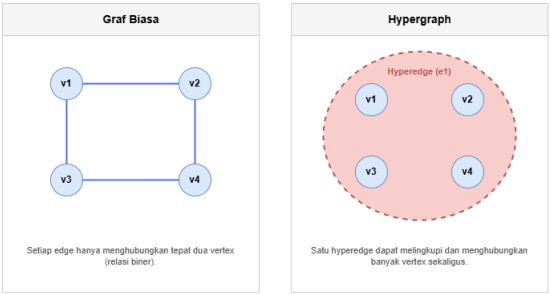

Fig II. Graph vs Hypergraph (Source: Author's Personal Documentation, made by draw.io)

B. *Puzzle Difficulty Analysis*

Many studies have been conducted on puzzle difficulty, especially for Sudoku. Pelánek found that difficulty is not only determined by how many steps are needed to solve a puzzle, but also by how complex each step is and how the constraints interact. This led to the use of Constraint Satisfaction Problem (CSP) models, which are useful for representing the rules and limits in puzzle solving.

Later, Thangamani and Regulagedda used CSP-based metrics to predict Sudoku difficulty based on constraint density. Liu et al. also studied puzzle difficulty by using features such as the number of symbols, number of constraints, and complexity of the puzzle description. In addition, research in computational complexity shows that many puzzles are NP-complete, meaning their difficulty can be analyzed mathematically.

Overall, these studies suggest that puzzle difficulty may be related to structural and mathematical features, not only to player experience.

TABLE I. SUMMARY OF PUZZLE DIFFICULTY MEASUREMENT APPROACHES IN PREVIOUS RESEARCH

Research	***Domain***	***Method***	***Subhead***
Pelánek (2014)	Sudoku	Computational model of human solving behavior	Uses data from 1700+ human-solving traces
Thangamani & Regulagedda (2026)	Sudoku	CSP symbolic modeling, constraint density	Deterministic constraint analysis
Liu et al. (2025)	PuzzleClone (general logic puzzles)	Composite scoring (symbols, constraints, description length)	Internal normalization

C. *Research Related to Word Games*

Research on *Wordle* mainly focuses on finding optimal solving strategies. Healy showed that an information theory approach can solve most *Wordle* puzzles in about 3–4 guesses on average. In addition, *Wordle* has also been modeled as a Constraint Satisfaction Problem (CSP), which allows the solving process to be analyzed using variables, domains, and constraints.

For *NYT Connections*, most existing research focuses on Large Language Models (LLMs). Todd et al. tested several LLMs and found that their performance in solving *Connections* is still worse than human players. The study also showed that the game requires semantic reasoning and general knowledge.

Other studies have used LLMs to generate *Connections* puzzles automatically and estimate their difficulty using cosine similarity from semantic embeddings. This shows that most current difficulty analysis still relies on machine learning methods. So far, there is very little research that studies the structure of the puzzle itself.

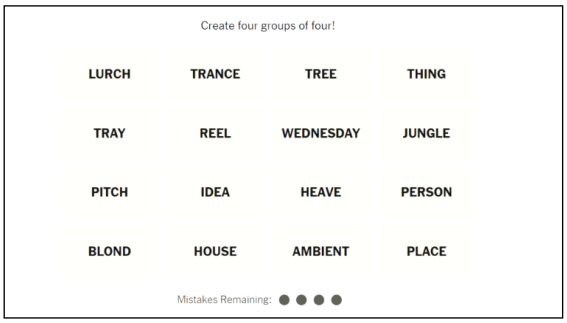

Fig III. *NYT Connections* Example, Puzzle #532 (Source: https://gamerant.com/new-york-times-connections-hints-answers-532-november-24-2024/)

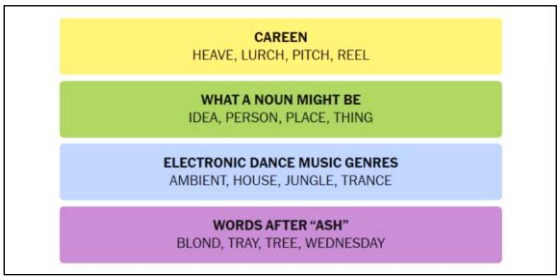

Fig IV. *NYT Connections* Answers Example, Puzzle #532. Yellow is the easiest category, and Purple is the hardest category. (Source: https://gamerant.com/new-york-times-connections-hints-answers-532-november-24-2024/)

D. *Research Gap*

Previous research shows that hypergraphs are widely used to model relationships involving multiple objects. However, there is no study that formally uses hypergraphs to model *NYT Connections*. Most existing work focuses on testing LLM performance or generating puzzles automatically, rather than analyzing the mathematical structure of the game.

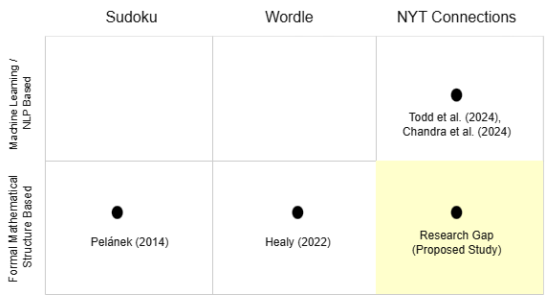

Fig V. Previous research based on puzzle domain and type of difficulty metric approach (Source: Author's Personal Documentation, made by draw.io)

In addition, current methods for measuring difficulty in *Connections* mostly use machine learning and semantic embeddings. Unlike Sudoku and *Wordle*, which already have formal mathematical models such as CSP and information theory, *Connections* still lacks a mathematical framework for studying and analyzing the relationships among words and categories.

III. Theoretical Basis

A. Graph

A graph is a structure used to represent objects and the relationships between them. Mathematically, a graph is written as $G = (V, E)$, where V is a set of vertices (nodes) and E is a set of edges that connect two vertices.

A vertex represents an object, while an edge represents a relationship between two objects. In a simple graph, each edge can only connect two vertices. Therefore, graphs are suitable for modeling pairwise relationships.

For example, if $V = \{A, B, C\}$ and $E = \{(A,B), (B,C)\}$, then A is connected to B, and B is connected to C. However, A is not directly connected to C.

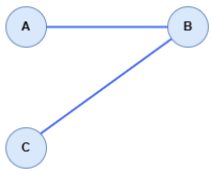

Fig VI. Simple graph with 3 vertex and 2 edge (Source: Author's Personal Documentation, made by draw.io)

In this study, graphs are introduced as a basic concept before hypergraphs. Although graphs can model relationships between two objects, they cannot represent relationships involving more than two objects at the same time. Because of this limitation, a more suitable model is needed for *NYT Connections*, where each category consists of four related words.

B. Hypergraph

A hypergraph is an extension of a graph where one edge (called a hyperedge) can connect more than two vertices at the same time. A hypergraph is written as $H = (V, E)$, where V is a set of vertices and E is a set of hyperedges, each hyperedge can contain any number of vertices.

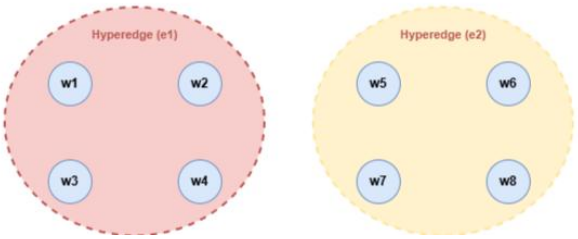

Fig VII. 4-uniform hypergraph with 8 vertex and 2 hyperedge (Source: Author's Personal Documentation, made by draw.io)

Unlike graphs, hypergraphs can represent relationships involving multiple objects at once. This makes them useful for modeling group-based relationships. A special type of hypergraph is a *k-uniform hypergraph*, where every hyperedge has the same number of vertices. In this study, a 4-uniform hypergraph is used, meaning each hyperedge always contains 4 vertices.

For example, if $V = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8\}$ and the hyperedges are $e_1 = \{w_1, w_2, w_3, w_4\}$ and $e_2 = \{w_5, w_6, w_7, w_8\}$, then this is a 4-uniform hypergraph.

In this study, hypergraphs are used to model *NYT Connections*. Each word is represented as a vertex, and each category is represented as a hyperedge containing four words. Therefore, a puzzle with 16 words and 4 categories can be represented as a 4-uniform hypergraph with 16 vertices and 4 hyperedges. This representation is more suitable than a standard graph because it directly matches the structure of the game.

C. Set Theory

Set theory is a basic part of mathematics used to explain concepts such as sets, subsets, size of sets, and partition. A set A is called a subset of B (written $A \subseteq B$), if all elements of A are also elements in B. The cardinality of a set, written $|A|$, shows how many elements are in the set.

One important concept is a partition. A partition is a way to divide a set into several smaller groups that do not overlap, where each group is not empty, and all groups together form the original set. Formally, a partition satisfies:

(i) $B_i \neq \emptyset$ for all i

(ii) $B_i \cap B_j = \emptyset$ for $i \neq j$

(iii) $B_1 \cup B_2 \cup ... \cup B_k = S$

For example, a set $S=\{1,2,3,4,5,6\}$ can be divided into $B_1=\{1,2,3\}$ and $B_2=\{4,5,6\}$. These subsets do not overlap and together form the original set.

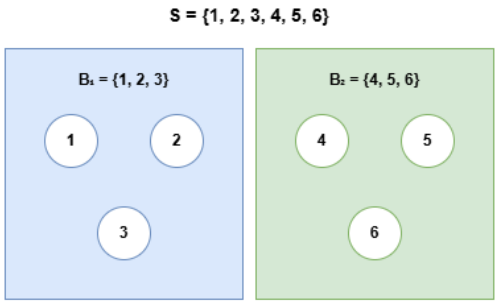

Fig VIII. Illustration of a partition of a set S into 2 disjoint blocks (Source: Author's Personal Documentation, made by draw.io)

In this study, partitions are used to represent the solution of *NYT Connections*. The 16 words are treated as one set, which is divided into four groups, each containing four words. Each group is then represented as a hyperedge in the hypergraph.

D. Constraint Satisfaction Problem (CSP)

A Constraint Satisfaction Problem (CSP) is a problem made up of variables, possible values (domain), and constraints that must be satisfied *(X, D, C),* where *X* is a set of variables, *D* is the set of possible values, and *C* as a set of constraints.

The goal of a CSP is to assign values to all variables so that all constraints are satisfied. A simple example is map coloring, where adjacent regions cannot have the same color.

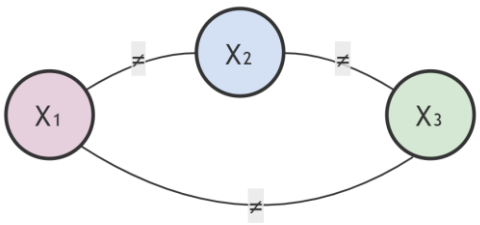

Fig IX. Example of a simple CSP (Source: Author's Personal Documentation, made by draw.io)

In this study, *NYT Connections* is modeled as a CSP. Each word is a variable, and the possible categories are its domain. The constraints are that each word must belong to exactly one category, and each category must contain exactly four words.

With this approach, solving *NYT Connections* can be seen as both a hypergraph partitioning problem and a constraint satisfaction problem. This provides a strong mathematical foundation for analyzing the game.

IV. Mathematical Modeling

A. Vertex Set

Given a *NYT Connections* puzzle, we define the vertex set *V* as:

$$V = \{w_1, w_2, \ldots, w_{16}\} \quad (1)$$

with each element $w_i \in V$ represents one word that appears in the puzzle, for i = 1, 2, …, 16. Since every Connections puzzle always consists of exactly sixteen different words, then the size of the vertex set is fixed |V| = 16.

At this stage, V only contains a collection of words without any relationships between them. Each word is treated as a single vertex, which will later be connected through hyperedges according to the correct category.

B. Solution Hyperedges and Perfect Partition

Each correct category in the Connections puzzle is defined as a solution hyperedge e_i which is a subset of *V,*

$$e_i \subseteq V, |e_i| = 4 \quad (2)$$

for i = 1, 2, 3, 4. The full set solution hyperedges is

$$E_{sol} = \{e_1, e_2, e_3, e_4\} \quad (3)$$

with $|E_{sol}| = 4$.

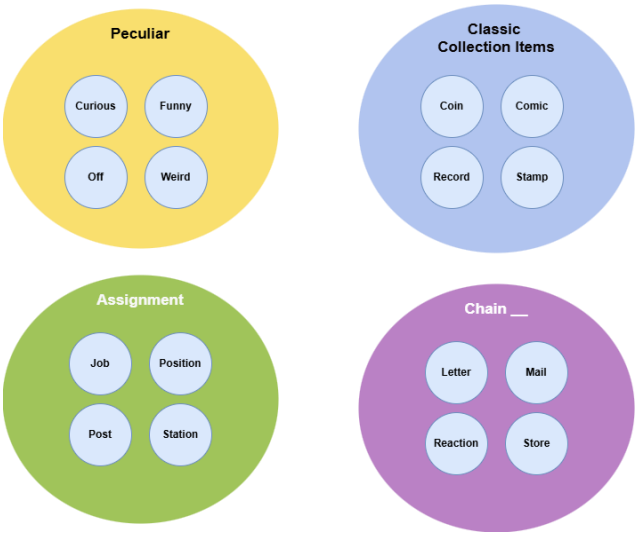

Fig X. *NYT Connections* #449 represented with hypergraph (Source: Author's Personal Documentation, made by draw.io)

Proposition 1. The set E_{sol} forms a perfect partition of V.

Proof. Based on the rules of the Connections game, each word $w_i \in V$ belongs to exactly one category, and each category contains exactly four words. Therefore E_{sol} satisfies the three partition conditions:

(i) $e_i \neq \emptyset$ for = 1, 2, 3, 4, because $|e_i| = 4 > 0$;

(ii) $e_i \cap e_j = \emptyset$ for $i \neq j$, because no word appears in more than one category

(iii) $e_1 \cup e_2 \cup e_3 \cup e_4 = V$, because all words must be assigned to one of the four categories.

As a direct consequence:

$$|V| = \Sigma_{i=1}^{4} |e_i| = 4 \times 4 = 16 \quad (4)$$

C. Candidate Hyperedges

Besides the solution hyperedges (the correct categories), we also define candidate hyperedges. A candidate hyperedge is a subset $S \in V$ such that $|S| = 4$ (meaning it contains four words), the words in S share a strong semantic relationship or theme, but $S \notin E_{sol}$ (meaning it's not one of the correct categories). Formally, the set of all candidate hyperedges is defined as

$$E_{cand} = \{S \subseteq V : |S| = 4, \kappa(S) \geq \tau, S \notin E_{sol}\} \quad (5)$$

where $\kappa(S)$ represents the semantic coherence of the set S, and τ is a threshold used to decide whether a group of words is strong enough to be considered a candidate category.

The existence of candidate hyperedges shows that a Connections puzzle can contain multiple word groups that all look reasonable. The more candidate hyperedges a word belongs to, the more likely it is to confuse the player and be placed in the wrong category. This idea will later be used to define the Ambiguity Score.

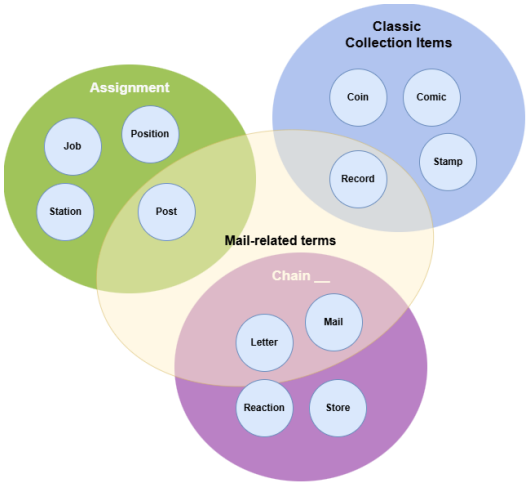

Fig XI. *NYT Connections* #449 represented with solution hyperedge (green, blue, and purple colored) and 1 candidate hyperedge (orange) (Source: Author's Personal Documentation, made by draw.io)

In this study, the semantic coherence κ is based on semantic distance using the WordNet lexical database

D. Final Hypergraph

The full structure of the puzzle is modeled as a hypergraph formed by combining both solution and candidate hyperedges. Formally:

$$H = (V, E) \quad (6)$$

where

$$E = E_{sol} \cup E_{cand} \quad (7)$$

All hyperedges in H have the same size (4 vertices), so the model is a 4-uniform hypergraph. The difference between solution and candidate hyperedges lies only in their role: one represents the correct answer, while the other represents plausible but incorrect groupings.

Proposition 2. It holds that $E_{sol} \subseteq E$, $E_{cand} \subseteq E$, and $E_{sol} \cap E_{cand} = \emptyset$

Proof. From the definition of candidate hyperedges, every $S \subseteq E_{cand}$ must satisfy $S \notin E_{sol}$. Therefore, no hyperedge can belong to both sets at the same time. Since E is defined as the union of both sets, we also have $|E| = |E_{cand}| + |E_{sol}|$

The degree of a vertex w_i in hypergraph H, denoted as $deg_H(w_i)$ is defined as

$$deg_H(w_i) = |\{e \in E : w_i \in e\}| \quad (8)$$

Every vertex must belong to exactly one solution hyperedge, so $deg_H(w_i) \geq 1$. This degree value will later be used as the basis for measuring ambiguity in the puzzle

E. Justification for Using Hypergraphs

Hypergraphs are chosen because they can represent an entire category as a single unit. Unlike standard graphs, which only show pairwise relationships between vertices, a hyperedge can directly connect four words within the same category. This matches the structure of *NYT Connections*, where every category always contains four related words.

If a normal graph were used instead, then each category of four words would need to be broken into six pairwise edges. This would lose the important information that all those relationships belong to the same group. In contrast, a hypergraph preserves this group structure through a single hyperedge.

In addition, hypergraphs allow candidate categories to be represented naturally as additional hyperedges. This makes it easier to identify words that appear in multiple semantic groups and increases interpretability when measuring ambiguity.

V. METRIC DESIGN

This chapter defines two structural metrics based on the hypergraph $H = (V, E)$ used to measure the difficulty of a *NYT Connections* puzzle in a mathematical way.

A. Ambiguity Score (AS)

The Ambiguity Score measures how many alternative word groupings can be formed besides the correct solution. In a simple puzzle, most words only belong to one group, which is the correct category. As a result, there are few or no alternative groupings. In other puzzles, some words may also appear in additional candidate hyperedges, creating more possible groupings.

Given a hypergraph $H = (V, E)$, the Ambiguity Score is defined using the degree of each vertex v, denoted as $deg_H(v)$, as follows

$$AS(P) = \frac{1}{|V|} \sum_{v \in V} (deg_H(v) - 1)$$

where:

- AS(P): Ambiguity Score of the puzzle
- |V|: number of vertices in the hypergraph
- $deg_H(v)$: number of hyperedges containing vertex v

The degree is reduced by one because every word must belong to exactly one correct group. Therefore, $deg_H(v) - 1$ only counts how many extra (incorrect or alternative) groups a word appears in.

The AS score represents the average number of alternative groupings per word in the puzzle.

- If AS = 0, every word appears only in its correct group, meaning no alternative groupings are found
- If AS is higher, more words appear in additional candidate groups, meaning more alternative groupings exist

In general, a higher AS value indicates that the puzzle contains more semantically plausible groupings in the WordNet-based hypergraph representation.

B. Structural Overlap Score (SOS)

The Structural Overlap Score measures how similar the correct solution groups are to the most similar incorrect groups. Formally, for a hypergraph $H = (V, E)$ with solution hyperedges E_{sol} and candidate hyperedges E_{cand} and the SOS is defined as:

$$SOS(P) = \frac{1}{|E_{sol}|} \sum_{e \in E_{sol}} \max_{e' \in E_{cand}} |e \cap e'|$$

where

- SOS(P): Structural Overlap Score on puzzle P
- E_{sol}: set of solution hyperedges
- E_{cand}: set of candidate hyperedges
- e ∩ e': number of words shared by the solution hyperedge and the candidate hyperedge

For each solution hyperedge, the candidate hyperedge with the largest overlap is found and then averaged to obtain the SOS score.

The SOS value ranges from 0 to 3 because each group contains 4 words, and candidate groups are not identical to the solution groups

- A high SOS value means candidate groups share many words with the correct solution groups
- A low SOS value means that candidate groups share few words with the correct solution groups

In general, a higher SOS value indicates that the candidate hyperedges are structurally more similar to the solution hyperedges in the WordNet-based hypergraph representation

C. Example Calculation

To illustrate the proposed metrics, consider the *NYT Connections* puzzle shown in Figure X. The puzzle consists of the following four solution categories:

- CURIOUS, FUNNY, OFF, WEIRD
- JOB, POSITION, POST, STATION
- COIN, COMIC, RECORD, STAMP
- LETTER, MAIL, REACTION, STORE

TABLE II. EXAMPLE OF AS AND SOS CALTULATION FOR PUZZLE #449

Component	***Value***
Number of words	16
Solution hyperedges	4
Threshold	0.15
Candidate hyperedges	854
Total hyperedges	858
Ambiguity Score (AS)	213.50
Structural Overlap Score (SOS)	2.75

This Ambiguity Score result indicates that many words participate in alternative semantically plausible groupings. And the Structural Overlap Score result is close to the theoretical maximum value of 3, showing that several candidate hyperedges strongly overlap with the correct solution groups.

TABLE III. CATEGORY LEVEL METRICS FOR PUZZLE #449

Component	*AS Category*	*SOS Category*
Yellow	288.00	3.00
Green	238.25	3.00
Blue	186.25	3.00
Purple	141.50	2.00

Category-level scores can also be computed, in this example, the Yellow category has the highest AS value (its words participate in the largest number of alternative candidate hyperedges). In contrast, the Purple category has the lowest AS value, meaning that fewer semantically plausible alternative groupings are found.

For SOS, the Yellow, Green, and Blue categories each achieve the maximum overlap value of 3, while the Purple category reaches an overlap of 2.

Then, the category-level metrics will be used in the experiment section to compare structural properties across the four official NYT difficulty levels.

VI. EXPERIMENT

A. Dataset

The data used in this study were collected from the public archive of *NYT Connections* puzzles. Each puzzle consists of 16 words divided into 4 solution categories, where each category has an official difficulty level (Yellow, Green, Blue, and Purple) defined by the New York Times. A total of 40 puzzle were selected from different publication periods to reduce potential bias.

Since each puzzle always contains four categories with different difficulty levels, the unit of analysis in this study is the category, not the puzzle itself. Therefore, the 40 puzzles produce a total of 160 categories (40 Yellow, 40 Green, 40 Blue, and 40 Purple categories). This categorization is used consistently throughout all analysis.

B. Experimental Procedure

The experiment was conducted in five main stages:

1) Hypergraph construction: each puzzle was transformed into a hypergraph consisting of 16 words as vertices and 4 solution categories as solution hyperedges

2) Forming candidate hyperedge: All possible combinations of 4 words from the 16 puzzle words were evaluated using a WordNet-based coherence function ($\kappa(S) \geq \tau$) and that were not part of the correct solution were added as candidate hyperedges.

3) Metric computation: for each puzzle, a combined hypergraph consisting of solution hyperedges and candidate hyperedges was constructed. Two proposed metrics, Ambiguity Score (AS) and Structural Overlap Score (SOS) were then calculated.

4) Difficulty-level aggregation: since the metrics were initially computed at the puzzle level, they were aggregated to the category level. The category AS was calculated from the average connectivity degree of the four words within a category, while the category SOS was calculated from the maximum overlap between a solution category and its most similar candidate hyperedge. The resulting values were then averaged for each difficulty level.

5) Result comparison: The average AS and SOS values were compared across difficulty levels using descriptive statistics

C. Coherence Function Implementation $\kappa(S)$

Subset coherence was computed using WordNet through the NLTK library. For each pair of words within a four-word subset, the path similarity score was calculated. The final coherence value was defined as the average of the six pairwise similarity scores.

- Higher values indicate stronger semantic similarity
- The score range is: (0, 1]

D. Expected Results and Evaluation Criteria

The goal of this experiment is to evaluate whether AS and SOS can explain the structure of *NYT Connections* categories and whether these properties are related to the official difficulty levels assigned by The New York Times. There are two possible results:

1) The metric values increase from Yellow to Purple: this would indicate that more difficult categories contain more alternative semantic groupings and stronger overlaps with candidate hyperedges.

2) The metric values decrease from Yellow to Purple: this would indicate that AS and SOS measure semantic relationships represented in WordNet, rather than the official difficulty ordering used by NYT.

In either case, the experiment provides insight into the relationship between semantic structure and puzzle difficulty.

VII. RESULT AND DISCUSSION

The experiment was implemented in Python using NLTK WordNet, and executed with a threshold of $\tau = 0.15$

A. Global Analysis of Puzzle Structure

Before analyzing category difficulty, all puzzles were evaluated to obtain a general overview of the hypergraph structures produced.

TABLE IV. DESCRIPTIVE STATISTICS OF ALL PUZZLES

Metrics	*Mean*	*Std*	*Median*	*Min*	*Max*
AS	55.9562	53.0587	43.25	0.0	240.5
SOS	2.4000	0.5483	2.50	0.0	3.0
E_{cand}	223.8250	212.2349	173.00	0.0	962.0

Based on Table IV, the Ambiguity Score shows substantial variation, ranging from 0 to 240.50. This variation indicates that some puzzles contain very few alternative groupings, while others contain hundreds or even thousands of candidate hyperedges that are still considered semantically plausible.

The SOS values have a mean of 2.4 and a median of 2.5. Since the theoretical maximum value of SOS is 3, these results indicate that most puzzles contain candidate hyperedges that closely resemble the actual solution groups.

The number of candidate hyperedges also demonstrates that the space of possible word groupings in *NYT Connections* is much larger than the four correct solution categories.

B. Analysis by NYT Difficulty Level

To evaluate whether the proposed metrics can represent the official *NYT Connections* difficulty levels, the values were aggregated according to the four difficulty categories.

TABLE V. AMBIGUITY SCORE BY DIFFICULTY LEVEL

Difficulty	*Mean AS*	*Std*	*Median*
Yellow	65.5438	58.0265	56.375
Green	62.1812	61.9611	34.375
Blue	51.7688	55.0042	28.875
Purple	44.3312	53.5832	28.375

The average AS values show a decreasing trend from Yellow to Purple.

TABLE VI. STRUCTURAL OVERLAP SCORE BY DIFFICULTY LEVEL

Difficulty	*Mean SOS*	*Std*	*Median*
Yellow	2.575	0.8130	3.0
Green	2.475	0.7157	3.0
Blue	2.425	0.7808	3.0
Purple	2.125	0.6480	2.0

A similar pattern can be seen for SOS, which also decreases gradually from Yellow to Purple.

C. Interpretation of Results

These findings suggest that the proposed hypergraph-based metrics primarily capture semantic relationships represented in WordNet rather than the official difficulty levels experienced by human players.

Yellow categories generally contain words with clear and direct semantic relationships. As a result, WordNet can identify these relationships effectively, generating many candidate hyperedges and consequently producing higher AS and SOS values.

In contrast, Purple categories often involve wordplay, letter patterns, idioms, abbreviations, or cultural references. Consequently, fewer candidate hyperedges are generated, leading to lower AS and SOS values.

Therefore, the decreasing trend from Yellow to Purple suggests that the metrics measure semantic relationships represented in WordNet, whereas the official NYT difficulty levels are influenced by additional linguistic and cognitive factors beyond semantic similarity alone.

VIII. CONCLUSION

This study demonstrates that hypergraphs provide a suitable representation for modeling category structures in *NYT Connections* puzzles. The proposed AS and SOS metrics successfully measure the number of alternative word groupings and the similarity between solution groups and candidate groups.

The result show that AS and SOS successfully represent semantic relationship in the WordNet-based hypergraph. Categories with stronger semantic connections tend to produce more candidate hyperedges and obtain higher metric values.

However, the results do not follow the official NYT difficulty order. Instead of increasing from Yellow to Purple, both AS and SOS decrease across the difficulty levels. This indicates that the proposed metrics measure semantic similarity rather than the official difficulty assigned by NYT.

A possible explanation for this is that Yellow categories usually contain clear semantic relationships that can be recognized by WordNet, while Purple categories often rely on wordplay, abbreviations, letter patterns, idioms, or cultural references. The findings suggest that *NYT Connections* difficulty is influenced not only by semantic relationship between words, but also by cognitive factors.

Future work may use richer semantic representations, such as Word2Vec, GloVe, or BERT, to capture relationships that cannot be represented by WordNet. Validation using human participants is also necessary to determine how well the proposed metrics correlate with actual human performance and perceived difficulty.

VIDEO LINK AT YOUTUBE

https://youtu.be/GTS4tjpY-gg?si=j5vVbRLtAA7q7c-h

SOURCE CODE & DATASET REPOSITORY

https://github.com/longlivephos/hypergraph-based-nyt-connections.git

.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to Allah SWT for His blessings which made this completion of this paper possible, and Prof. Dr. Ir. Rinaldi, M.T., whose lectures and enthusiasm for discrete math have been a constant source of inspiration throughout this semester.

Special appreciation for Billy Joel, whose music provided companionship during many hours of writing and revising this paper, and to all the *NYT Connections* content creators out there, particularly Brian K. (@briank_fromtiktok) and Savannah (@dailyxsav) whose inspired the author's interest in analyzing this topic.

Finally, the author would like to thank the creators of *NYT Connections for* designing such a creative, challenging, and brilliant game.

REFERENCES

[1] X. Ouvrard. 2020. "Hypergraphs: An Introduction and Review." arXiv preprint arXiv:2002.05014. https://arxiv.org/abs/2002.05014. Accessed: Jun. 16, 2026.

[2] R. Pelánek. 2014. "Difficulty Rating of Sudoku Puzzles: An Overview and Evaluation." https://www.fi.muni.cz/~xpelanek/publications/sudoku-arxiv.pdf.

[3] Y. Liu et al. 2025. "PuzzleClone: An SMT-Powered Framework for Synthesizing Verifiable Data." https://arxiv.org/abs/2508.15180. Accessed: Jun. 17, 2026.

[4] A. D. Healy. 2022. "On Optimal Strategies for Wordle." http://www.alexhealy.net/papers/wordle.pdf. Accessed: Jun. 17, 2026.

[5] R. T. Todd et al. 2024. "Connecting the Dots: Evaluating Abstract Reasoning Capabilities of LLMs Using the New York Times Connections Word Game." arXiv preprint arXiv:2406.11012. https://arxiv.org/abs/2406.11012. Accessed: Jun. 18, 2026.

[6] A. Chandra et al. 2024. "Making New Connections: LLMs as Puzzle Generators for The New York Times' Connections Word Game." arXiv preprint arXiv:2407.11240. https://arxiv.org/abs/2407.11240. Accessed: Jun. 18, 2026.

[7] R. Munir. 2025. "Graf (Bagian 1)." https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2025-2026/ Accessed: Jun. 18, 2026.

[8] R. Munir. 2025. "Himpunan (Bagian 1)." https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2025-2026/ Accessed: Jun. 18, 2026.

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 19 Juni 2026

Neysa Alya Mukhbita (13525080)