GRADUATE PROGRAMS INSTITUTE

MERONYMS AND SYNONYMS EXTRACTION FROM ENGLISH DICTIONARY

Oula ALSHEIKH

MASTER'S THESIS

Istanbul, March 2025

MERONYMS AND SYNONYMS EXTRACTION FROM ENGLISH DICTIONARY

Oula ALSHEIKH

MASTER'S THESIS
COMPUTER ENGINEERING DEPARTMENT
COMPUTER ENGINEERING MASTER'S PROGRAM WITH THESIS

MASTER'S THESIS ADVISOR
Prof. Dr. Ercan SOLAK

MASTER'S THESIS JURY MEMBERS
Asst. Prof. Dr. Özlem Feyza ERKAN
Prof. Dr. Taner ESKIL

ACKNOWLEDGMENT

In the Name of Allah, the Most Gracious, the Most Merciful.

It is with profound gratitude and a deep sense of fulfillment that I present this work. This achievement would not have been possible without the unwavering support and encouragement of several extraordinary individuals, to whom I owe my heartfelt thanks.

First and foremost, I extend my sincerest appreciation to my advisor Prof. Dr. Ercan SOLAK. His expert guidance, thoughtful mentorship, and steadfast commitment have been instrumental in shaping this work. I am deeply grateful for his patience, insightful feedback, and the wealth of knowledge he has shared with me throughout this journey.

I would also like to express my heartfelt gratitude to my family. To my beloved parents, your unconditional love, sacrifices, and faith in me have been my greatest source of strength and inspiration. This accomplishment is a reflection of your endless support and encouragement. To my two wonderful sisters, your kindness, understanding, and constant reassurance have been a tremendous source of comfort and motivation, for which I am eternally grateful.

Finally, I extend my deepest thanks to my friends, whose unwavering support and encouragement have been a beacon of positivity during this journey. Their companionship and kind words have lifted my spirits during challenging times and have made this experience all the more memorable.

CONTENTS

ABSTRACT

Meronyms And Synonyms Extraction from English Dictionary

Relations extraction is the process of identifying connections between entities. It is used in wide range of applications including information retrieval, text summarization, machine translation, question answering, thesaurus building, and ontology building. This thesis presents a method for detecting meronyms and synonyms relationships between nouns within sentences from a given English dictionary.

This study presents an analyzer designed to detect pairs of meronyms and sets of synonyms (synsets) between noun words. It depends on OPTED dictionary as a data source. OPTED dictionary consists of words with their tags and definitions. The analyzer tokenizes the dictionary into sentences and then extracts words' POS tags using NLTK library to put the base data structure. The data structure is used to find meronyms relations by using predefined keywords and rules. The data structure is used also to find synsets and their intersected, transitive, and similarity generations.

Keywords: Dictionary Processing, Meronyms, Synonyms, POS Tags, NLTK
Date: March 2025

ABSTRACT

İngilizce Sözlükten Meronymler Ve Eş Anlamlılar Çıkarımı

İlişki çıkarma, adlandırılmış varlıklar arasındaki bağlantıları belirleme sürecidir ve bilgi alma, metin özetleme, makine çevirisi, soru cevaplama, eş anlamlılar sözlüğü oluşturma ve ontoloji oluşturma dahil olmak üzere çok çeşitli uygulamalarda kullanılır. Bu makale, belirli bir İngilizce sözlükten cümleler içindeki isimler arasındaki meronim ve eş anlamlı ilişkilerini tespit etmek için bir yöntem sunmaktadır.

Bu çalışma, OPTED Sözlüğü'nde, etiketleri ve tanımlarıyla birlikte sözcüklerden oluşan isim sözcükleri arasındaki meronim çiftlerini ve eş anlamlı eş anlamlı kümelerini tespit etmek için tasarlanmış bir analizör sunar ve daha sonra NLTK kütüphanesini kullanarak sözcüklerin eş anlamlı etiketlerini çıkararak, önceden tanımlanmış ilişki anahtar sözcüklerini ve belirli önceden tanımlanmış kuralları kullanarak meronim ilişkilerini bulmak için kullanılan temel veri yapısını oluşturur ve eş anlamlı eş anlamlı kümelerini ve bunların kesişimsel, geçişli ve benzerlik nesillerini bulur.

anahtar kelimeler: Dictionary Processing, Meronyms, Synonyms, POS Tags, NLTK
Tarih: Mart 2025

ABBREVIATIONS

NLP : Natural Language Processing
OPTED : Online Plain Text English Dictionary
NLTK : Natural Language Toolkit
POS : Part of Speech

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Semantic relations are fundamental aspects of Natural Language Processing (NLP) and play a crucial role in various NLP tasks. They involve finding the main entities that contribute to relations. They can be used in multiple domains, such as query optimization, semantic sentiment, and ontology building.

English is the most used language. In our study, we focused on finding two kinds of semantic relations: meronyms and synonyms, using the OPTED English dictionary as a resource.

OPTED is a public-domain English dictionary based on the 1913 edition of Webster's Unabridged Dictionary. It is structured for efficient computer processing with simple HTML formatting. It offers improved accuracy and includes inflections and plurals. It is available for personal and research use and is not copyrighted. Our study concentrated on nouns only with the tag (n) in their definitions. We chose this dictionary after looking at many dictionaries on the internet. The OPTED dictionary has a very well-defined structure, is very wide, and has multiple definitions for each word. Words with multiple definitions can enrich our data set and enable us to discuss more scenarios. The dictionary defines each word in a detailed context compared to other existing dictionaries.

Our rule-based approach is clear and precise. It is particularly effective for static and well-defined datasets like OPTED. Unlike machine learning models, which often require large annotated datasets and may struggle with overfitting. Our method uses structured dataset and ensures high accuracy.

We employed NLTK [5] as our chosen tool to get the main entities in our relations through NLTK tokenizer and tagger. We put a wide collection of manual rules, that can be used to catch most meronym relations cases. Our approach is rule-based, which has a high accuracy according to our static structure. Our static structure includes

multiple domains and has a moderate size.

Semantic relations like meronyms and synonyms are important and used to enhance information retrieval, build knowledge graphs, and optimize query results in search engines [11]. Our study focused on finding these relations.

We found a collection of rules that determine keyword usage, whether a word really expresses a relation, and whether it is just a normal word. We also extracted a rich collection of rules that can catch the most meronyms in dictionary definitions.

In this study, we offer a strong algorithm for generating an ontology of synonyms for nouns. The algorithm goes through multiple collections' phases, such as union, intersection, and similarity match.

To check our results, we compared them with the synonyms of WordNet, a well-known database of words and their meanings and synonyms. This confirmed that our method works well and matches established standards. We also used the thesaurus dictionary to evaluate our results. Thesaurus is a very rich dictionary that provides definitions of words and synonyms.

In our study we were able to prove that following a rule-based approach can have high accuracy, when it is used with rich and well-structured dictionary, and with a wide set of manual defined rules.

2 LITERATURE SURVEY

Semantic relations are important in multiple domains because they are used to understand the meaning of the context and clarify ambiguity [1]. Semantic relations depend on identifying the relations that exist in the dataset, and defining objects that contribute to these relations. They help to get the main objects that distinguish one meaning from another. That will increase the speed of information retrieval and achieve higher accuracy [2].

Semantic relations' objects can have different natures like text, multi-modal data, multi-attribute data, and images. In [3] they explained the semantic relations process to a multi-data model that exists in social media.

Semantic relations applications are wide, and the most noticeable field is query optimization. In [11] they used the entropy-weighted similarity method and score normalization to solve the problem of query limitation by replacing query tokens with synonyms or homonyms.

Our work concentrated on extracting the semantic meronyms and synonyms relations from text files. We worked with the OPTED dictionary that has a very rich collection of English nouns and their multiple definitions. Because our dataset is well structured, we followed the rule-based approach, which can have better results than the machine learning approach. Extracting relations can be done by putting predefined manual stable rules, which identify existing relations clearly. In [14] they applied the rule-based approach on social network dataset. They extracted the main relations, which helped to clarify the emotions in the text and determined if they are positive, negative, or neutral. In [15] they used a linguistic rule-based approach to enhance sentiment analysis by extracting the features from people reviews, which removed context complexity and enhanced classification accuracy.

There are a variety of relations that could be extracted from the data, like

meronyms (part-to-whole), hyponyms (subtype), hypernymy (super type), synonyms (similar meanings), and antonymy (opposite meanings). In [4] a detailed description of each relation with examples was mentioned. Extracting the relations starts by understanding the structure of the sentence. That needs tokenizing the sentence depending on specific punctuations and patterns, that identify the nature of the word. Then the POS tagger is used to get each word category. Multiple methods are used, and NLTK is a popular method, many works use NLTK. [5] used patterns and POS auto-learning approach to extract terms from scientific texts. They started with basic patterns, refined them by matching POS sequences in sentences, and created new patterns to find more terms. This process works well for extracting terms from paper abstracts.

[6] used also NLTK to process tweets. They used the hashtags and URLS found in tweets in the tokenization process after excluding the hashtag sign. [7] created an automatic grammatical tagger for a Spanish–Mixtec parallel corpus with 12,300 sentences. They applied GPT-4, CRF, and BERT models after tokenizing and annotating the corpus according to EAGLES standards.

Multiple languages can be parsed by NLTK. [8] used NLTK tokenizer and tagger to realize Maithili, Magahi and Bhojpuri languages.

Ontologies are structured representations of knowledge that specify concepts and the relationships between them within a particular domain. WordNet is a very common ontology, and we used it in our work in synonyms evaluation process. [9] used WordNet ontology to build Latvian WordNet ontology. [10] explained how the Bulgarian WordNet is created and focused on its structure and the additional details compared to the Princeton WordNet. It discussed the inclusion of words and phrases, and improvements in organizing synsets with labels for grammar, meaning, and style. It also described the various connections between synsets and words in the Bulgarian WordNet.

We concentrated in our work on getting the meronyms and synonyms relations, because they are very strong relations to clarify the sentence meaning and getting multiple same meanings, which can broaden the semantic relation so much. Many researchers concentrated on extracting these relations. In [12] they used lexico syntactic patterns to extract the synonyms from biomedical texts, which enhanced the retrieval of biomedical information. In [13] the authors used meronyms relations and represented

them as formulas with fuzzy logic to extract the main objects, such as colors and body parts. Meronyms are important relations and also synonyms are the same. Synonyms are the set of words that have similar meanings for specific word [16]. Synonyms can be a rich set because each language is also affected by other languages. In [17] they explained how Russian language was affected by Greek and Turkic languages. [18] used the SynMICdb database which has 659,194 synonyms, to explore cancer mutation frequency and tumor distribution.

We used Wordnet ontology and thesaurus dictionary in our evaluation, which provided a rich collection of relations. Wordnet ontology in used in multiple projects. [19] used WordNet ontology in their research and evaluated its hypernyms and synonyms. In [20] researchers used the WordNet ontology and its relations to discover the domain of a topic by using TF-IDF method.

3 LANGUISTIC FUNDAMENTALS

3.1 DICTIONARY FORMAT

There are many dictionaries available on the internet, but I selected the most used one in other people's work and a rich one, which is OPTED. It is not copyrighted and is available for use.

It consists of 28 separate sections, each dedicated to one letter, which I unified into one section.

The dictionary is easy to read and comprehend. It is well-structured and uses plain punctuation. The word being defined appears first in each entry, followed by its part of speech (for example, n. for a noun or v. for a verb) in parenthesis. Definitions are rich and frequently include information about the word's origins, such as Old English, Latin, or Greek. Commas are used to keep things clear by separating lists or parts of a definition, whereas parentheses are used to give further information. Semicolons are used to separate closely related concepts or examples. Quotation marks are used to illustrate sentences, and periods are used to finish each definition. The dictionary also covers specialized terms from areas like music, architecture, and science, giving helpful examples and links to related words. The punctuation and layout make this dictionary detailed and straightforward to follow, helping readers learn about words, their meanings, and their history.

Each word may have different entries. Each entry could express a normal meaning or a relation, like a synonym, meronym, or hyponym. Each relation has its special keywords that refers to its existence.

The dictionary is wide and follows an organized structure. It handles the case of multiple definitions and multiple parts of speech for each word.

In our study, we concentrated only on nouns and avoided other parts of speech that may have different processing methods.

Malarian (a.) Alt. of Malarious

Malarious (a.) Of or pertaining, to or infected by, malaria.

Malashaganay (n.) The fresh-water drumfish (Haploidonotus grunniens).

Malassimilation (n.) Imperfect digestion of the several leading constituents of the food.

Malassimilation (n.) An imperfect elaboration by the tissues of the materials brought to them by the blood.

Malate (n.) A salt of malic acid.

Malax (v. t.) Alt. of Malaxate

Malaxate (v. t.) To soften by kneading or stirring with some thinner substance.

Malaxation (n.) The act of softening by mixing with a thinner substance; the formation of ingredients into a mass for pills or plasters.

Malaxator (n.) One who, or that which, malaxates; esp., a machine for grinding, kneading, or stirring into a pasty or doughy mass.

Malay (n.) One of a race of a brown or copper complexion in the Malay Peninsula and the western islands of the Indian Archipelago.

Malay (a.) Alt. of Malayan

Malayan (a.) Of or pertaining to the Malays or their country.

Malayan (n.) The Malay language.

Malayalam (n.) The name given to one the cultivated Dravidian languages, closely related to the Tamil.

Malbrouck (n.) A West African arboreal monkey (Cercopithecus cynosurus).

Figure 3.1 Dictionary Format Snapshot

(Figure 3.1) shows a section of the dictionary. There are same words with different tags and definitions. OPTED dictionary has a lot of definitions for each word. It is a good data source for extracting semantic relations.

3.2 Meronyms (Part of Relations)

Meronyms, or "part-whole" relationships, are connections in a language where a word represents a part of something larger. Essentially, meronyms identify components or elements of a whole. Meronyms help break down complex ideas into smaller, more understandable pieces.

For instance, "wheel" is a meronym of "car," as a wheel is part of a car. Similarly, "head" is a meronym of "body," since a head is part of a body. Meronyms can describe physical parts, like the "engine" in a "car," but also abstract elements, such as a "chapter" in a "book."

Meronyms are not limited to just physical components. In some cases, they also represent functional or conceptual parts. For example, in a "team" a "player" can be seen as a meronym of the "team," since each player is a part of the team. In language, individual words or sounds can act as meronyms of a larger phrase or sentence.

Meronyms are usually categorized into two types: componential and member meronyms. Componential meronyms refer to physical parts that make up a whole, like the "engine" in a "car" or the "keyboard" in a "computer." Member meronyms, on the other hand, refer to individuals within a group, such as a "player" in a "team" or a "student" in a "class".

Meronym relationships can be recognized by specific words like "part of," " has," "includes," "is part of," or "is made of." These words clearly indicate how parts relate to a whole.

For example, in the sentence "A car has wheels and an engine," the word "has" shows the connection between the car and its parts. Similarly, in "A wheel is part of a car," the phrase "is part of" explains the relationship.

The structure, or syntax, used to express these relationships is simple. Usually, the whole is mentioned first, followed by its parts, like "A house includes walls and a roof." Alternatively, the part may come first, such as "A roof is part of a house."

3.3 Synonyms And Synsets

Synonyms are words that have similar or identical meanings. They add variety to language and help avoid repeating the exact words. For example, house, home, and residence all refer to where someone lives, but each word can have a slightly different tone or usage. The choice of synonym often depends on the situation or the feeling you want to express.

Synsets, or synonym sets, group nouns that share the same meaning. These are particularly useful in fields like linguistics and computer language tools. For example, the synset {car, automobile} refers to “a vehicle with four wheels.”

Synonyms are similar words but don’t always mean the same thing. Some, like attorneys and lawyers, are almost the same. Others, like happy and joyful, have the same feelings. Synonyms help in many ways. In translation, they match words with the correct meanings in another language. In text analysis, they help systems understand content. Search engines use them to link related words, making searches more effective. The word "tree" can have different meanings. It can mean a big plant with a trunk and branches or a diagram with branching parts. Synsets help sort these meanings so language tools know which one to use. In conclusion, synonyms make language more flexible and expressive, while synsets organize these meanings to improve understanding. Together, they enhance communication for both humans and technology.

4 METHODS AND TOOLS

To extract the meronymy and synonyms, we used the OPTED dictionary, which has a predefined rich structure. We followed a rule-based approach rather than a machine-learning approach. We applied multiple steps, starting by preprocessing our dictionary and then parsing and tokenizing to get the data structure that we handled. We used the data structure to find sentences that could have a relation, then matched them with exact patterns that we defined according to the dictionary's content. We used the same dictionary to find synonyms and synsets of nouns.

Because our dataset consists of nouns with their definitions, there were static rules that we could define to find meronyms and synonyms relations. This more suited the rule-based approach and produced a high-accuracy result compared to the machine learning approach, which needs more training and a bigger dataset.

4.1 Meronyms Extraction

There was no rich ready dictionary to use, so we used a dictionary called OPTED, which consists of a separate dictionary for each English alphabet letter. We gathered all the letters dictionaries, which have the same format, in one dictionary. Then, we removed unnecessary characters and marks such as "!" and extra unneeded spaces.

After preparing our dictionary we applied the NLTK tokenizer, which depends on specific punctuation to detect the start and end of each token. NLTK knows common abbreviations, so the percentage of parsing mistakes is rare.

NLTK POS tagger was used to discover the right tags for each word, which is the most sensitive step in extracting relations from the dictionary. Wrong POS tags can mislead an existing right relation. We applied NLTK POS tagger for each sentence by itself, and not for whole dictionary, which also gave us higher accuracy.

After applying NLTK tokenizer and getting POS tags, we could start to detect sentences that may have a relation, and that is according to specific keywords that are so common to detect meronym relations.

We processed one keyword to detect meronym relations which is "part of."

Because it is the basic key to detect meronym relation. We did put some rules that we recognized according to the structure of meronym relations, to say if the word "part of" is really a keyword, or it is a normal word, that is no indicator for meronym relation.

After having the sentences that could have a meronym relation, we were able to start finding the relations by matching the sentence with the right relation pattern. We had previously put these relation patterns through looking into dictionary by our eyes. We had defined 11 patterns which have specific tags for each word like noun is part of noun (we will mention these patterns in chapter 5).

4.2 Synonyms Extraction

We used the same OPTED dictionary because it was rich of synonyms. We started by a simple method for extracting synonyms, by realizing the punctuation (semicolon), which comes in the definition and refers to possible synonym relation. We processed the nouns and their noun synonyms only.

Synsets are groups of each noun synonyms together and can be used to get more synonyms relations. We used these synsets to find nouns which have intersected synonyms, and as a result they become synonyms for each other. And we used the words' synsets to find nouns which have transitive relations. After getting the transitive and intersected synonyms, we got the nouns that are similar to each other. Similarity was defined by getting the synsets of each word in the definition and adding it to other words in the same definition. Then we applied the same process for all nouns. Then we looked for nouns that have 70% similarity or higher in final definition synsets.

5 SOLUTION AND IMPLEMENTATION

We processed the OPTED dictionary, and removed not needed characters. We tokenized the dictionary text file by using NLTK tokenizer. We used NLTK POS tagger to get POS tags for nouns' definitions sentences. The final structure consists of each word and its multiple definitions, that can be used in getting desired relations.

We used the most famous keyword for realizing meronym relation "part of," to get sentences that may have a meronym relation. We classified these sentences according to some rules, that we wrote to decide if they might have a synonym relation or not. After that we processed the sentences that have the meronym relations and matched them with the most suitable rule to get the final relations.

For getting synonyms relations, we also tokenized the dictionary text and got POS tags by using NLTK library. We concentrated on nouns only. We looked for the nouns that have the semicolons in their definitions, and got the synonyms from there. After getting the synsets, we applied the intersection and transitive relation between them and then found similar synsets. Similar synsets should have exact synonyms shared between them or synonyms of these shared synonyms according to specific percentage that we defined.

5.1 Dictionary Preprocessing

The first step for finding relations in dictionary is to use a rich dictionary, which has a formal definition way for defining the words and has the tag of each word that is defined.

OPTED dictionary is a very good English dictionary. It is used in multiple ontologies and APIs. It has a lot of definitions and multiple definitions for each word. It consists of 28 separate sections, one for each alphabetic letter. We gathered these 28 dictionaries in one dictionary and had 180160 definitions for multiple tags. We were only interested about nouns in defining the main objects of our relations, so we

extracted the nouns only.

We removed extra characters like extra spaces and not needed punctuations to not confuse the tokenizer.

Then we extracted the words, which have the noun type and have (.n) next to the word that is defined, and we built our base structure, which consists of each noun and its multiple definitions.

5.2 Getting POS Tags

The POS tag is the category of the word. Getting the right tags is the most important phase in getting the right relations.

We used the NLTK to get the tags by passing each definition for each word to the NLTK tokenizer. The tokenizer depends on specific words to know the start and end of the sentence. It splits between words, even if there are no direct spaces, which could be special characters, like dashes, for example.

NLTK tokenizer has many tools and we used the tool word_tokenizer. This tool divides the text into words and uses the rule-based approach. It depends on finding patterns after dividing the text according to punctuations or spaces. It processes special cases like abbreviations and mixed text. (Figure 5.1) shows the tags for words in the sentence "the angry bear chased the frightened little squirrel."

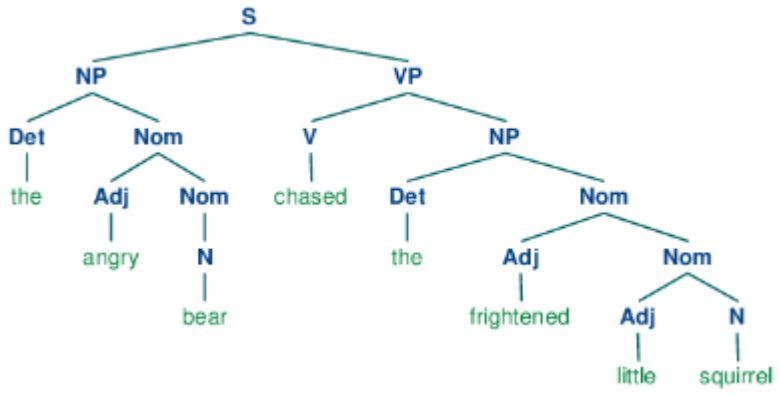

Figure 5.1 Words' Tags Example

After tokenizing we were able to apply the NLTK POS tagger, that gives a tag for each token in the sentence. We applied that for each sentence by itself to get the last

tokens, that will be used in getting our desired relation.

For example, for the sentence "Ablution The act of washing or cleansing specifically the washing of the body or some part of it as a religious rite," NLTK will give the following output for POS tags:

Ablution/NNP The/DT act/NN of/IN washing/VBG or/CC cleansing/VBG specifically/RB the/DT washing/NN of/IN the/DT body/NN or/CC some/DT part/NN of/IN it/PRP as/IN a/DT religious/JJ rite/NN.

Where NNP Proper Noun, DT Determiner, NN Noun, IN Preposition, VBG Gerund, CC Conjunction, RB Adverb, PRP Pronoun, JJ Adjective.

There is a difference between parsing the whole dictionary and parsing each sentence by itself. Parsing the whole dictionary at one step ignores some punctuations, which could result in wrong tags.

For example, we will see the difference between whole file parse and each definition parse for three definitions in the dictionary.

Milliampere (n.) The thousandth part of one ampere.

Milliweber (n.) The thousandth part of one weber.

Nape (n.) The back part of the neck.

1- Result of Parsing the whole file by one parse.

Milliampere/NNP The/DT thousandth/JJ part/NN of/IN one/CD ampere/JJ.

Milliweber/NNP The/DT thousandth/JJ part/NN of/IN one/CD weber/NN.

Nape/NNP The/DT back/JJ part/NN of/IN the/DT neck/NN.

2- Result of Parsing each line by itself.

Milliampere/NNP The/DT thousandth/JJ part/NN of/IN one/CD ampere/NN.

Milliweber/NNP The/DT thousandth/JJ part/NN of/IN one/CD weber/NN.

Nape/NNP The/DT back/JJ part/NN of/IN the/DT neck/NN.

The difference is found in word ampere, which had the wrong tag JJ in the first parse, but the right tag NN in the second parse.

5.3 Building Meronym Structure

We tokenized the text and got the POS tags for each word. After this phase we were able to use the dictionary for getting multiple relations, starting with the meronym relations. This relation is labeled by multiple keywords and we depended on the basic keyword which is "part of." This word is a strong indicator for meronym relation.

We selected each word with at least one definition containing the word "part of." We depended on dictionary data structure and considered each word as a key in this dictionary, so no word could be repeated. Every word definition containing the word "part of" can be a value from multiple values for this key.

(Table 5.1) shows a representation of our structure by having three words with their definitions as an example.

Table 5.1 Meronyms Data Structure Shape

Key	**Values**
Hair	is part of head and it covers skin above the skull is part of body and it grows in different places
Wheel	the lower part of car
Door	Basic part of a closed place which is a way to be in that place

5.3.1 Evaluating Relation Existence

After getting the base structure, which consist of each keyword which is a noun in the English dictionary. We were able to get the meronym relations or any other relations.

We have only the sentences the have the keyword “part of.” This keyword can refer to meronym relation or it can be a normal word that doesn’t have the meaning of part of whole and so doesn’t indicate a relation.

We had to follow some rules to determine if the keyword has the part-whole meaning. We had these rules after looking into dictionary sentences and having some statistics. We did put a test collection consisting of 100 sentences that were matched with rules and approved that the rules are 90% correct. We will mention these seven rules that we constructed after a manual look at the dictionary.

First rule: if the length of the sentence is less than 7 words so, it is definitely a keyword. Because the keyword "part of" will not contribute into other idea. If we have this length, the keyword "part of" will have the part whole meaning only.

example: "Barrel (n.) The hollow basal part of a feather."

Second rule: if the length of the sentence is more than 7 words, and the keyword comes in the first section of the sentence, it is a keyword. When the length is more than 7 words, we should look at the location of the keyword, and if it comes in the first half of the sentence, it is describing the noun by whole part relation, while if it is in the second half, it is a big probability that it is not.

Example: "Bar (n.) An ordinary, like a fess but narrower, occupying only one fifth part of the field."

Here, the word "part of" is not a meronym keyword relation because it comes in the second section.

Third rule: when the keyword is proceeded by the proposition, it has no whole part meaning.

Example: "Act (n.) A performance of part of a play; one of the principal divisions of a play or dramatic work in which a certain definite part of the action is completed."

In this sentence, the keyword "part of" comes in the first section of the sentence, but it is preceded by a proposition, so it doesn't indicate a part-whole relation.

Fourth rule: when the keyword is proceeded by a relative clause, like "which" and "that," it doesn't indicate a whole part relation.

Example: "Backlash (n.) The distance through which one part of connected machinery, as a wheel, piston, or screw, can be moved without moving the connected parts, resulting from looseness in fitting or from wear; also, the jarring or reflex motion caused in badly fitting machinery by irregularities in velocity or a reverse of motion."

The keyword "part of" here is preceded by "which," that describes another word in the sentence and not the word "Backlash" and it does not indicate a whole-part relation.

Fifth rule: if the sentence has the word "Act of" then the keyword "part of," it doesn't indicate a part-whole relation.

Example: "Bath (n.) The act of exposing the body, or part of the body, for purposes of cleanliness, comfort, health, etc., to water, vapor, hot air, or the like; as, a cold or a hot bath; a medicated bath; a steam bath; a hip bath."

Sixth rule: if the keyword is proceeded by one of the following words: "about,"

"above," "across," "after," "against," "along," "among," "around," "at," "before," "behind," "below," "beneath," "beside," "between," "beyond," “by," "during," "except," "for," "from," "in," "inside," "into," "near," "of," "on," "over," "past," "since," "through," "to," "under," "until," "up," "with," "without," it doesn’t indicate a part-whole relation (meronym).

Example 1: "Yoke (n.) A tie securing two timbers together, not used for part of a regular truss, but serving a temporary purpose, as to provide against unusual strain."

Example 2: "Goldfinch (n.) A beautiful bright-colored European finch (Carduelis elegans). The name refers to the large patch of yellow on the wings. The front of the head and throat are bright red; the nape, with part of the wings and tail, black; -- called also golds pink, Goldie, fool's coat, draw bird, draw-water, thistle finch, and sweet William."

Seventh Rule: If there is an indefinite and general object after the keyword, there is no relation. The words that are indefinite are "anything," "thing," "whole," and "object."

Example: "Back, The outward or upper part of a thing as opposed to the inner or lower part as the back of the hand the back of the foot the back of a hand rail. "

5.3.2 Right Rule Matching

After getting the definitions that have the keywords which express a meronym relation and not a normal meaning, we moved to get the right shape of the meronym relation and to get the final pairs, which have the shape of A is part of B.

We applied the rule-based approach which depends on a core. The core consists of multiple rules. We had these rules after evaluating multiple meronym relations by eye to conduct the rules. We ended with 11 rules for meronym relations that we describe here.

5.3.2.1 Noun Keyword Noun Rule

If the word that is defined is a noun and the word that comes after the keyword is a noun. If so, the word that is defined is part of the word that comes after the keyword, and they have a meronym relation. (Figure 5.2) shows the representation of the rule, where the dark blue circle represents the tag, and the light blue circle represents the noun in the sentence.

Example: "Wheel (n.) part of Car. "

Meronym Relation: Wheel part of car.

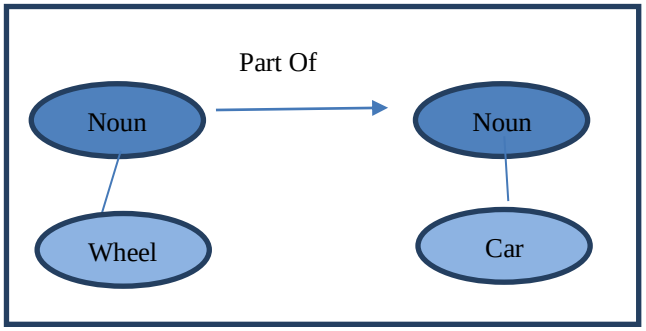

Figure 5.2 Noun Keyword Noun Rule Representation

5.3.2.2 Noun Keyword Determiner Noun Rule

If the word that is defined is a noun and the word that comes after the keyword is the determiner word, then followed by a noun.

If so, the defined noun is part of the word that comes after the determiner word. (Figure 5.3) shows the representation of the noun keyword determiner noun rule. The dark blue circle represents the tag, and the light blue circle represents the noun in the sentence.

Example: "Back (n.) The upper part of a lode, or the roof of a horizontal underground passage. "

Meronym relation: Back part of lode.

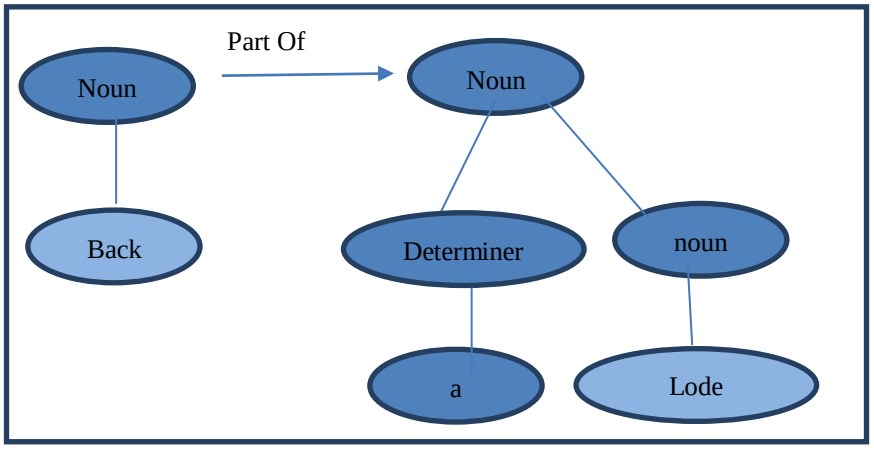

Figure 5.3 Noun Keyword Determiner Noun Rule Representation

5.3.2.3 Noun Keyword Determiner One Adjective Noun Rule

If the word that is defined is noun and the word that comes after the keyword is determiner then followed by adjective and then followed by noun.
If so, the defined noun is part of the noun that comes after the adjective but described by this adjective because the advective limits the item.
Example: "Beakhead (n.) A small platform at the fore part of the upper deck of a vessel, which contains the water closets of the crew. "
Meronym Relation: Buckhead part of upper deck.

5.3.2.4 Noun Keyword Determiner Multiple Adjectives Noun Rule

If the word that is defined is a noun and the word that comes after the keyword is a determiner, then followed by multiple adjectives and then followed by a noun.
If so, the defined noun is part of the noun that comes after the adjective but described by these adjectives because the adjectives limit the item.
Example: "Base (n.) The lower part of a complete architectural design, as of a monument; also, the lower part of any elaborate piece of furniture or decoration. "
Meronym Relation: Base part of complete architectural design.

5.3.2.5 Noun Keyword Verb-Gerund Noun Rule

If the word that is defined is a noun and the word that comes after the keyword is verb-gerund, then followed by a noun.
If so, defined noun is part of the noun associated with the verb-gerund that describes it.
Example: "Face (n.) That part of the acting surface of a cog in a cog wheel, which projects beyond the pitch lines."
Meronym Relation: Face part of acting surface.

5.3.2.6 Noun Keyword Comparative Proposition Noun Rule

If the word that is defined is a noun and the word that comes after the keyword is a comparative adjective (adjective +er) followed by proposition then by noun.
If so, the defined word is part of the noun that is preceded by the comparative adjective.
Example: "Tome (n.) As many writings as are bound in a volume, forming part of a larger work; a book; -- usually applied to a ponderous volume."
Meronym Relation: Tome part of larger work.

5.3.2.7 Noun Keyword Adjective Noun Rule

If the defined word is a noun and the word that comes after the keyword is an adjective followed by a noun.
If so, the defined word is part of the noun that is preceded by the adjective that comes directly after the keyword.
Example: "Nonage (n.) The ninth part of movable goods, formerly payable to the clergy on the death of persons in their parishes."
Meronym Relation: Nonage part of movable goods.

5.3.2.8 Noun Keyword adjective Verb Rule

If the defined word is a noun and the word that comes after the keyword is an adjective followed by a verb as a tag but acts as a noun in categorization.
If so, the word that comes before the keyword is part of the verb-noun that is preceded by the adjective.
Example: "Tenth (n.) The tenth part of annual produce, income, increase, or the like; a tithe."
Meronym: Tenth Part of annual produce.

5.3.2.9 Noun Keyword Comparison Noun Proposition Determiner Noun Rule

If the defined word is a noun and the word that comes after the keyword is a determiner word + comparison + noun1 + proposition + determiner + noun2.
If so, the defined word is part of the comparison + noun1 + proposition + determiner + noun2 phrase.
Example: "Malar (n.) The cheek bone, which forms a part of the lower edge of the orbit."
Meronym: Malar Part of lower edge of the orbit.

5.3.2.10 Noun Keyword "One" Noun Rule

If the defined word is noun and the word that comes after the keyword is the word "one" followed by noun.
If so, the defined word is part of the noun that comes after the word "one".
Example: "Micron (n.) A measure of length; the thousandth part of one millimeter. "
Meronym: Micron is part of millimeter.

5.3.2.11 Noun Keyword Noun's Noun Rule

If the defined word is noun and the word that comes after the keyword is possessive phrase containing of noun + apostrophe (') + noun.
If so, the defined word is part of the possessive phrase.
Example: "Taffrail (n.) The upper part of a ship's stern, which is flat like a table on the top, and sometimes ornamented with carved work; the rail around a ship's stern. "
Meronym: Taffrail is part of ship's stern.

We had an order for these rules in our program. If a sentence has a rule that could be part of a longest rule, it will be discussed first in our program and will be matched first. After selecting the correct rule, we were able to get the meronym relations in our program, which are added to the output text file. The text file is the output of our program. It contains two lines for each relation, the first one is to mention each word with its definition, and the second line is to mention the meronym relation objects according to our program.

Here are some lines of our meronym file, that the program produces after evaluating relation existence and matching with the correct rule for each definition that has the keyword “part of.”

Adenography, That part of anatomy which describes the glands

Adenography->Adenography Part of anatomy.

Scampavia,A long low war galley used by the Neapolitans and Sicilians in the early part of the nineteenth century

Scampavia->no relation:keyword in second half.

Footing,The finer refuse part of whale blubber not wholly deprived of oil

Footing->Footing Part of whale blubber.

```
Therapeutics,That part of medical science which treats of the discovery and application of remedies for diseases
Therapeutics->Therapeutics Part of medical science 6

Thesis,The accented part of the measure expressed by the downward beat  the opposite of arsis
Thesis->Thesis Part of measure 4.2

Thesis,The part of the foot upon which such a depression falls
Thesis->Thesis Part of foot 4.2

Third,The sixtieth part of a second of time
Third->Third Part of second of 2

Third,The third part of the estate of a deceased husband which by some local laws the widow is entitled to enjoy du
Third->Third Part of estate 4.2
```

Figure 5.4 Program Meronyms Snapshot

(Figure 5.4) shows our program output for the meronym relations. The number which is mentioned near each relation is a reference to the rule that the relation matches in our program.

5.4 Synonyms Extraction

Extracting synsets can’t be done through one phase because synonyms have a collection structure. Multiple operations can be applied on this collection like intersection, transitive relations, and similarity relations. Now we will describe each step that we applied until we got the last synsets.

5.4.1 Base Synonyms

Synonyms are words that have the same meaning. The OPTED dictionary is rich in synonyms, which are expressed by putting a semicolon ";" to indicate that the phrase is a synonym of the base word.

"Taenioid (a.) Ribbonlike; shaped like a ribbon. " This an example of one word synonym which is "Ribbonlike."

"Levite (n.) A priest; -- so called in contempt or ridicule. " This is an example of a synonym that contains a determiner + word, which is "A priest.”

Because we focused in our search on nouns, so, we took the synonyms which are nouns only. We looked for the definitions that have the semicolon and then split them by semicolon to get the multiple synonyms if there is more than one synonym.

“Earth (n.) A part of this globe; a region; a country; land.”

In this sentence, the program will find multiple synonyms for the word "earth" which are " region" and "globe. "

After getting the nouns with their multiple synonyms, we went through the next stages to get more synonyms (See Table 5.2).

Table 5.2 Base Synonyms

Noun	Synset
Noun1	A,B,C
Noun2	B,D,E
Noun3	F,C

Table 5.3 Synonyms Synsets Phase1 Representation

Noun	Synsets
A	B, C,
B	
C	F
D	B, H, E
E	
F	E
H	

We will use Table 5.3 as an example in our next phases.

5.4.2 Synonyms Intersection

After getting the base synsets, we were able to increase our synonyms collection for each word by following inference.

If A is a synonym of B, A is a synonym of C => B is a synonym of C.

Because of intersection, which means if two words have shared synonyms between their synsets, they would be synonyms to each other.

We took each word from our synsets and then looked in every other word to find if there is any intersection between the synsets of both words.

If we have n words, we can find n-1 other extra synonyms, if the intersection in synsets happened in each word with the other.

From (Table 5.3) we can see that A and D are synonyms because they have the shared synonym B, so we will add the noun D to the synonyms of A. D and F are synonyms because they have E shared between them.

After we applied the intersection, we got the following synsets represented in (Table 5.4)

Table 5.4 Synonyms Intersection

Noun	Synsets
A	B, C, D
B	
C	F
D	B, H, E, F
E	
F	E
H	

5.4.3 Transitive Synonyms

After getting the base synonyms and applying the intersection between them, we moved to the third step, which is applying the transitive relation between synsets.

If A is a synonym of B and B is a synonym of C => A is a synonym of C.

So, for each noun we have and that has synonyms, we will go into its synonyms and then for each one we will add its synonyms to the base noun if it wasn't there.

For a noun we can have more n-1 synonyms, if each other noun has one synonym at least.

From (Table 5.5) we can see that C has the synonym F. C is one of the synonyms of A, so by applying transitive relation, F will be another synonym of A.

F is synonym of C, C is synonym of A => F is synonym of A

B is synonym of D, D is synonym of A=> B is synonym of A

H is synonym of D, D is synonym of A=> H is synonym of A

E is synonym of D, D is synonym of A=> E is synonym of A

E is synonym of F, F is synonym of C=> E is synonym of C

After applying first level of transitive relation, we got the following Table (See Table 5.5).

Table 5.5 Transitive Synonyms

Noun	Synsets
A	B, C, D, F, H, E
B	
C	F, E
D	B, H, E
E	
F	E
H	

We can get more and more synonyms by applying transitive relations for second time, but we stooped at first transitive level to have the strongest synonyms.

5.4.4 Similarity Synonyms

After getting the base synonyms and then applying the intersection between the synonyms and then applying the transitive relations we moved to the fourth step. Fourth step includes applying the transitive relations between synsets to get a larger set.

There are some words that are not direct synonyms of each other and which are not the result of intersection or transitive relations, but have a similar meaning which might be expressed by different words or by synonyms of the words in the meaning.

Example:

Let say we have the two nouns word1, words2 defined by the following definitions

Word1: noun1 noun2 verb1 noun3 noun4.

Word2: noun5 verb2 noun6 noun7 noun8.

Let’s suppose that noun1 is a direct synonym of noun5, so it is in the direct synset of noun1. Let's suppose noun2 and noun6 have intersected synonyms, so as a result, they are synonyms of each other, and noun3 and noun7 are the same word, so we have three shared words between the definitions of the nouns word1 and word2.

These shared words might be the same exact word, or one word occurs in the wide synset of the second word, which will include the direct synonyms or intersected synonyms or transitive synonyms. Because we have three words, so the percentage of shared words is 3/4 = 75, which means that the words are synonyms of each other, because they have similar words in their definitions. We specified the percentage of

similarity in our work as higher than 70%. So, words might be similar to each other when they have synonyms that are shared between them according to a specific percentage which we set as higher than 70%. These shared synonyms can be direct synonyms or can be synonyms that are a result of intersection or transitive relations.

6 Evaluation

We evaluated our work accuracy according to accuracy factors and did that for both meronyms and synonyms.

6.1 Meronyms Evaluation

As we previously mentioned in getting our meronyms relations, we used two steps. The first one was deciding if the keyword is really a relation keyword or it is a normal keyword, and the second step was to match the relation with the correct rule. We had to evaluate if the first step was correct and to measure our factors for keyword correctness if it really expresses a meronym relation. Then, we tested if the relation we got is really successful and if the correct rule is matched.

To do our evaluation, we selected 100 random definitions from the dictionary. We labeled these definitions by our hand by 0 or 1. 0 means that the definition doesn't have a meronym relation and the keyword has a different meaning than the meronym meaning. 1 means that the definition has a meronym relation and the keyword expresses a meronym relation. Here is a part of our random 100 samples (see Table 6.1).

Table 6.1 Manual Keyword Evaluation

Definition	**Manual Keyword Evaluation**
Tenth, The tenth part of annual produce income increase or the like a tithe	1
Basin, A circular or oval valley or depression of the surface of the ground the lowest part of which is generally occupied by a lake or traversed by a river	0
Yellow, A bright golden color reflecting more light than any other except white the color of that part of the spectrum which is between the orange and green	0

Hearth,The floor of a furnace on which the material to be heated lies or the lowest part of a melting furnace into which the melted material settles	1
Table,The part of a machine tool on which the work rests and is fastened	0

In the next (Table 6.2) we will look at our program evaluation. There is a section in our program that is responsible for comparing manual results with automatic results.
This section wasn't a human process, which is a strong point for our evaluation. It made our evaluation flexible, and whenever we changed any rule, it was easily reflected in our evaluation results (See Table 6.2).

Table 6.2 Program Keyword Evaluation

Definition	**Program Evaluation**
Tenth, The tenth part of annual produce income increase or the like a tithe	1
Basin, A circular or oval valley or depression of the surface of the ground the lowest part of which is generally occupied by a lake or traversed by a river	0
Yellow, A bright golden color reflecting more light than any other except white the color of that part of the spectrum which is between the orange and green	0
Hearth,The floor of a furnace on which the material to be heated lies or the lowest part of a melting furnace into which the melted material settles	0
Table,The part of a machine tool on which the work rests and is fastened	1

6-1-1 Relation Existence Evaluation

In order to evaluate the relation existence and if the program found that the keyword is really a meronym keyword if it should be, or it is not a meronym keyword if it expresses a normal meaning, we defined the following metrics:

True Positives (TP): manual evaluation says that the keyword expresses a relation, and our program also says the same.

We will use the examples from (Table 6.1) and (Table 6.2).

Example "Tenth" definition, which is at the first row of the both tables.

False Positives (FP): manual evaluation says that the keyword doesn't express a relation, but our program says the keyword expresses a relation.

Example "Table" definition that is at the fifth row of both tables.

False Negatives (FN): manual evaluation says that the keyword expresses a relation, but our program says the keyword doesn't express a relation.

Example "Hearth" definition, which is at the fourth row of both tables.

True Negatives (TN): manual evaluation says that the keyword doesn't express a relation, and our program says the keyword doesn't express a relation too.

Example "Basin" definition, which is at the second row of both tables.

After defining our four metrics we can calculate the F factors as following:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

$$\text{FScore} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

We got the following results for the 100 random samples, precision=0.884, recall=0.91, accuracy=0.86, and fscore=0.89.

6-1-2 Right Rule Match Evaluation

In order to evaluate that if the program was able to extract the correct shape of the relation, and define the sub and whole objects in meronym relation correctly, we defined the following metrics:

True Positives (TP): manual evaluation mentions the relation A is part of B and our program also extracts the same relation A is part of B.

We will use the examples from (Table 6.3) and (Table 6.4).

Example "Tenth" definition, which is at the first row of both tables

False Positives (FP): manual evaluation mentions no relation and the program extracts a relation.

Example “Table” definition that is at the fifth row of both tables.

False Negatives (FN): we have this value in one of two scenarios:

1- manual evaluation mentions the relation A is part of B and our program extracts a different relation.

Example “Bare” definition which is at the third row of both tables.

2- manual evaluation mentions the relation A is part of B and our program doesn’t extract any relation.

Example “Microampere” definition that is at the fourth row of both tables.

True Negatives (TN): manual evaluation mentions no relation and our program doesn’t extract a relation.

Example “Basin” definition, which is at the second row of both tables.

After getting our four metrics we can calculate the F factors as following:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

$$FScore = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

We got the following results: precision=0.882, recall=0.89, accuracy=0.85, and fscore=0.88.

In (Table 6.3) we see the rules that we received from meronyms manual evaluation.

Table 6.3 Manual Meronym Rules

Definition	**Manual Rule**
Tenth, The tenth part of annual produce, income, increase or the like a tithe	Tenth Part of annual produce
Basin, A circular or oval valley or depression of the surface of the ground the lowest part of which is generally occupied by a lake or traversed by a river	0
Bare,That part of a roofing slate shingle tile or metal plate which is exposed to the weather	Bare part of roofing slate shingle tile
Microampere,One of the smaller measures of electrical currents the millionth part of one ampere	Microampere part of ampere
Table,The part of a machine tool on which the work rests and is fastened	0

In (Table 6.4) we see the rules that we received from the program.

Table 6.4 Program Meronym Rules

Definition	**Program Rule**
Tenth, The tenth part of annual produce ,income ,increase or the like a tithe	Tenth Part of annual produce
Basin, A circular or oval valley or depression of the surface of the ground the lowest part of which is generally occupied by a lake or traversed by a river	No relation keyword in second half

Bare,That part of a roofing slate shingle tile or metal plate which is exposed to the weather	Bare Part of roofing
Microampere,One of the smaller measures of electrical currents the millionth part of one ampere	No relation:keyword in second half
Table, The part of a machine tool on which the work rests and is fastened	Table Part of machine

In the following figure (Figure 6.1) we see a section of our program evaluation.

```
0-Adjunct,Something joined or added to another thing but not essentially a part of it
0-no relation:keyword in second half

0-Backboard,A board serving as the back part of anything as of a wagon
0-undefined object

0-Act,A performance of part of a play one of the principal divisions of a play or dramat
0-no relation:proposition or conjuction

1-Bare,That part of a roofing slate shingle tile or metal plate which is exposed to the
1-Bare Part of roofing 4.2

1-Barrel,The hollow basal part of a feather(Barrel part of feather)
1-Barrel Part of feather 4.2

1-Beak,The upper or projecting part of the shell near the hinge of a bivalve(Beak part o
1-Beak Part of shell 4.2

0-Cam,A projecting part of a wheel or other moving piece so shaped as to give alternate
1-Cam Part of wheel 4.2

1-Face,That part of a body having several sides which may be seen from one point or whic
faces(face part of body)
1-Face Part of body 4.2
```

Figure 6.1 Program Evaluation

Where the first line is manual evaluation and the second line is program output.

6.2 Synonyms Evaluation

Synonyms are words that have similar meanings. Synsets can be a wide collection. Finding exact measures is a hard operation because two words may have different synsets and have no shared synonyms between them, but they are synonyms in meanings.

In our evaluation, we depended on a popular website for finding synonyms, which is www.thesaurus.com and on the ontology Wordnet which has many relations between words. For getting synonyms from Wordnet, there is a ready API, which we used. For www.thesaurus.com, there is no API, so we had to make a program that sends a request to the website for each word and then analyzes the response to get the synonyms.

In (Table 6.5) we see the result for this program for 5 random words which were taken from thesaurus.com website by using our program.

Table 6.5 Thesaurus Program Result

Word	**Synset**
Sectator	adherent, apostle, cohort, devotee, enthusiast, fanatic, learner, pupil, zealot
Tempera	acrylic, color, cosmetic, dye, enamel, latex, makeup, oil, pigment, stain, varnish, wax
Troop	armed forces, army, contingent, forces, military, soldiers
Fomentation	cataplasm, compress, dressing, plaster
Tendon	hamstring, ligament, sinew, tissue

In (Table 6.6) we see the result that we had by using our program, which generates the synsets through four steps, which are dictionary lookup, then synsets intersection, and then transitive relation for synsets, and then finding similarities between synsets.

Table 6.6 Synonyms Program Result

Word	Synset
Sectator	follower, disciple, scholar
Tempera	distemper
Troop	company, number, multitude, army, battery, bevy, fellowship, fraternity, legion, livraison, meiny, military, numero, partnership, race, school, thrave, turm, war
Fomentation	excitation, instigation, encouragement
Tendon	sinew, fibre, thew

In order to evaluate our synsets results we used a collection of 100 random words. We got the synsets for these 100 random words from our program synsets. We also got the synsets for same 100 random words from thesaurus website by using a program that we developed.

After getting the synsets from our program and from thesaurus we calculated two metrics. The first metric is the total numbers of intersected synonyms for the whole 100 words between our program and thesaurus. The second metric is the total count of synonyms that we got from thesaurus for these 100 words.

We measured the accuracy of our program synonyms by the following recall formula:

$$\text{Recall} = \frac{\text{Total Number Of Intersected Synonyms}}{\text{Total Number Of Thesaurus Synonyms}}$$

We got 10.4% for evaluating 100 random samples.

For example, the number of intersected synonyms in the five words which are mentioned in (Table 6.5) and (Table 6.6) is 3 synonyms (army, military, sinew). The number of thesaurus synonyms for the five words in (Table 6.5) is 35.

Recall is 3/35 * 100 =8.5%.

(Figure 6.2) shows a section of our evaluation program that we developed for 100 random words. The first line represents the synonyms for the word from our program. The second line represents synonyms from thesaurus website. The third line represents the intersection between the first and second lines.

```
Veil
Program Synset List: cover, curtain, cover, disguise, mask, pretense, fence, guise, jest, lodging, masque, masquerade,
Thesaurus Synset List: cloak, curtain, mask, shroud
Intersected Synonyms: curtain, mask

Peg
Program Synset List: reason, step, footing, logos, pace, sake, score, walk
Thesaurus Synset List: fix
Intersected Synonyms:

Fidget
Program Synset List: uneasiness, restlessness, dysphoria, impatience, unease, unquietude, unrest
Thesaurus Synset List: fret, jiggle, squirm, twiddle, twitch
Intersected Synonyms:

Indecision
Program Synset List: indetermination, irresolution, vacillation, hesitation, faltering, misdoubt, pause
Thesaurus Synset List: uncertainty
Intersected Synonyms:

Parturition
Program Synset List: delivery, childbirth, birth, geniture, nascency, nativity, parage, parentage, seed, tradition
Thesaurus Synset List: birth, delivery
Intersected Synonyms: birth, delivery
```

Figure 6.2 Thesaurus And Program Synonyms Evaluation

We followed the same evaluation method for our program synonyms and Wordnet synonyms. We got synonyms from Wordnet ontology by using their API. (Table 6.7) shows the synonyms for the five words by using Wordnet ontology C# library, which is called Syn.WordNet.Extension, which has multiple functions, and we used the function GetSynSets.

Table 6.7 WordNet Synonyms

Word	**Synonyms**
Sectator	follower, disciple, scholar
Tempera	poster-paint, poster-color. poster-colour
Troop	Flock, scout-group, scout-group, parade, promenade, parade
Fomentation	instigation
Tendon	sinew

We measured the accuracy of our program synonyms comparing to Wordnet ontology synonyms by using the same metric that we used to compare our program synonyms with thesaurus synonyms which is the recall formula.

$$\text{Recall} = \frac{\text{Total Number Of Intersected Synonyms}}{\text{Total Number Of Thesaurus Synonyms}}$$

We got the following value for this metric 10.2% for evaluating same 100 random samples that we used in thesaurus evaluation.

(Figure 6.3) shows a section of our evaluation program for the 100 random words. The first line represents the synonyms for the word from our program. The second line represents synonyms from Wordnet ontology. The third line represents the intersection between the first and second lines.

```
Veil
Our Program Synset List: cover, curtain, cover, disguise, mask, pretense, fence, guise, jest, lodging, masque, masquerade,
WordNet Synset List: humeral_veil, veil, caul, embryonic_membrane, head_covering, obscure, blot_out, obliterate, hide
Intersected Synonyms:

Peg
Our Program Synset List: reason, step, footing, logos, pace, sake, score, walk
WordNet Synset List: peg, pin, thole, tholepin, rowlock, oarlock, wooden_leg, leg, pegleg, stick, nog, peg_down, nail_down,
Intersected Synonyms:

Fidget
Our Program Synset List: uneasiness, restlessness, dysphoria, impatience, unease, unquietude, unrest
WordNet Synset List: fidget, fidgetiness, restlessness
Intersected Synonyms: restlessness

Indecision
Our Program Synset List: indetermination, irresolution, vacillation, hesitation, faltering, misdoubt, pause
WordNet Synset List: indecisiveness, indecision, irresolution
Intersected Synonyms: irresolution

Parturition
Our Program Synset List: delivery, childbirth, birth, geniture, nascency, nativity, parage, parentage, seed, tradition
WordNet Synset List: parturition, birth, giving_birth, birthing
Intersected Synonyms: birth
```

Figure 6.3 Wordnet And Program Synonyms Evaluation

Our program synsets result were rich, but we depended on the following argument to defend our result. If we look at the most famous two resources for getting synsets, which are the thesaurus and Wordnet, we find that they have different synonyms for the same words, but they are close in meanings. Because the synsets could be different but have similar meanings, there will be no exact way to evaluate. We can see from (Table 6.5) and (Table 6.7) that for five words, there was only one word that has shared synsets, which is “Tendon” and that was in most famous synonym’s

resources, WordNet and thesaurus. That supports our result and clarify that there is no accurate way to measure synonyms. In our program, we got very big and rich synsets for each word.

We calculated the recall value for the same 100 random samples between thesaurus and Wordnet, and found the following value 12.6% which is so close to our previous recall values.

That means that our program synsets were good results but there is no chance to get high recall values for two different resources. Because they will not have the exact synsets. They could have different synsets and that is the principle of synonyms (different words with same meanings).

(Figure 6.4) shows a section of our evaluation program for 100 random words. The first line represents the synonyms for the word from Wordnet. The second line represents synonyms from thesaurus. The third line represents the intersection between the first and second lines.

```
Veil
Wordnet Synse List: humeral_veil, caul, embryonic_membrane, head_covering, obscure, blot_out, obliterate, hide
Thesaurus Synset List: cloak, curtain, mask, shroud
Intersected Synonyms:

Peg
Wordnet Synse List: pin, thole, tholepin, rowlock, oarlock, wooden_leg, leg, pegleg, stick, nog, peg_down, nail_down, n
Thesaurus Synset List: fix
Intersected Synonyms:

Fidget
Wordnet Synse List: fidgetiness, restlessness
Thesaurus Synset List: fret, jiggle, squirm, twiddle, twitch
Intersected Synonyms:

Indecision
Wordnet Synse List: indecisiveness, irresolution
Thesaurus Synset List: uncertainty
Intersected Synonyms:

Parturition
Wordnet Synse List: birth, giving_birth, birthing
Thesaurus Synset List: birth, delivery
Intersected Synonyms: birth
```

Figure 6.4 Wordnet And Thesaurus Synonyms Evaluation

7 CONCLUSION

This thesis has demonstrated the importance of semantic relations in Natural Language and presented an implementation for extracting two kinds of semantic relations: meronyms and synonyms. Extracting the relations from text focuses on the most important words. Our thesis showed that the rule-based approach is highly effective for identifying meronyms and synonyms in a well-structured dataset like OPTED. By utilizing OPTED's organized structure and NLTK tools, we achieved high accuracy.

We created a comprehensive set of rules for identifying meronyms and designed a robust algorithm for finding synonyms. These efforts simplify the process of extracting key relationships. They offer a dependable method for building ontologies and enhancing tasks like information retrieval and search optimization.

Comparing our results with reliable resources like WordNet and thesaurus dictionary validated the accuracy and dependability of our approach.

Future work will focus on implementing more keywords, more semantic relations, such as hyponyms, and it will implement other parts of speech, such as nouns, adjectives, and more.

By expanding the scope of this research, we can unlock new insights into semantic relations and contribute to the field of Natural Language Processing.

BIOGRAPHY

OULA ALSHEIKH is a master's student in the Computer Engineering program at Beykoz University, where she took courses in various computer engineering fields, such as Image Processing, Database, Information Retrieval, Software Reliability, and Theory of Computation.

Oula earned her bachelor's degree in software engineering from Governmental Damascus University in Syria. During her study, she was passionate about how semantic relations are extracted from data with its different types, especially texts and social media sources like tweets. She always wanted to know how a query on google can fetch multiple sources even if it doesn't have the searched keywords. She was able to learn the basics in this domain by completing this thesis and she plans to continue with other semantic relations and other kinds of data sources.

Date: March 2025

REFERENCES

[1] Singh, S., & Siddiqui, T. J. (2015). Role of semantic relations in Hindi word sense disambiguation. *Procedia Computer Science, 46,* 240–248. https://doi.org/10.1016/j.procs.2015.02.017

[2] Sergeev, A. (2020). An application of semantic relation extraction models.

[3] Kou, F., Du, J., He, Y., & Ye, L. (2016). Social network search based on semantic analysis and learning. *CAAI Transactions on Intelligence Technology*, *1(4),* 322–329. https://doi.org/10.1016/j.trit.2016.12.001

[4] Hegedűs, I., & Martsa, S. (2009). Cognitive approaches to the study of english. *9th HUSSE conference*.

[5] Shao, W., Hua, B., & Song, L. (2021). A pattern and POS auto-learning method for terminology extraction from scientific text. *Data and Information Management, 5(3),* 329–335. https://doi.org/10.2478/dim-2021-0005

[6] Ramachandran, D., & Parvathi, R. (2019). Analysis of Twitter-specific preprocessing technique for tweets. *Procedia Computer Science*, *165,* 245–251. https://doi.org/10.1016/j.procs.2020.01.083

[7] Santiago-Benito, H., Córdova-Esparza, D. M., Castro-Sánchez, N. A., Terven, J., Romero-González, J. A., & García-Ramirez, T. (2025). Automatic grammatical tagger for a Spanish–Mixtec parallel corpus. *SoftWareX, 29*. https://doi.org/10.1016/j.softx.2024.101985.

[8] Mundotiya, R. K., Mishra, S., & Singh, A. K. (2022). Hierarchical self-attention based sequential labeling model for Bhojpuri, Maithili, and Magahi languages. *Journal of King Saud University - Computer and Information Sciences, 34*(10, Part A) , 8739–8749. https://doi.org/10.1016/j.jksuci.2021.09.022

[9] Paikens, P., Grasmanis, M., Klints, A., Lokmane, I., Pretkalniņa, L., Rituma, L., Stāde, M., & Strankale, L. (2022). Latvian WordNet. *Proceedings of the 12th Global Wordnet Conference*ion (LREC 2022), 187–196. Marseille, France. https://aclanthology.org/2023.gwc-1.23/

[10] Koeva, S. (2021). The Bulgarian WordNet: Structure and specific features. *Papers of BAS: Humanities and Social Sciences, 8(1).*

[11] Li, X., Lipp, J., Shakir, A., Huang, R., & Li, J. (2024). BMX: Entropy-weighted similarity and semantic-enhanced lexical search. *Arxiv*.
https://arxiv.org/abs/2408.06643.
[12] Roberts, A. (2005). Learning meronyms from biomedical text. *Proceedings of the ACL Student Research Workshop* (pp. 49–54).
https://aclanthology.org/P05-2009.pdf
[13] Okal, B. O., & Miruka, F. A. (2016). Application of set theory formulas in the presentation of lexical meronymy. *Saudi Journal of Humanities and Social Sciences*, *1*(2), 62–72. 10.21276/sjhss.2016.1.2.5
[14] Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining, 11(81)*.
https://doi.org/10.1007/s13278-021-00776-9.
[15] Saraswathi, N., Rooba, T. S., & Chakaravarthi, S. (2023). Improving the accuracy of sentiment analysis using a linguistic rule-based feature selection method in tourism reviews. *Measurement: Sensors, 29*, 100888. https://doi.org/10.1016/j.measen.2023.100888.
[16] Gadimova, L. S. (2023). Synonyms as the semantic component of conversion. *International Journal of Social Science and Human Research, 6(3)*, 1365–1368.
https://doi.org/10.47191/ijsshr/v6-i3-03
[17] Ziyayeva, M. F. (2023). Synonyms in Russian philology (Literature review). *International Journal of Advance Scientific Research, 3(11)*, 19–26.
https://doi.org/10.37547/ijasr-03-11-05.
[18] Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M., Groß, M., Backofen, R., & Diederichs, S. (2019). A pan-cancer analysis of synonymous mutations. *Nature Communications, 10*, Article 2569. https://doi.org/10.1038/s41467-019-10489-2
[19] Nadig, R., Ramanand, J., & Bhattacharyya, P. (2008). *Automatic evaluation of Wordnet synonyms and hypernyms*. In *Proceedings of ICON-2008: 6th International Conference on Natural Language Processing Macmillan Publishers, India*.
http://ltrc.iiit.ac.in/proceedings/ICON-2008
[20] Wargärde, N**.** (2020). Using WordNet synonyms and hypernyms in automatic topic detection. *Linnaeus University, Faculty of Technology, Department of Computer Science and Media Technology (CM). Linnaeus University Press*. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1461669

[21] https://www.thesaurus.com/

[22] https://www.mso.anu.edu.au/%7Eralph/OPTED/

[23] https://wordnet.princeton.edu/