Production Systems and Information Engineering
Volume 12 (3), pp. 36–46.
https://doi.org/10.32968/psaie.2024.3.4

# SEMICONDUCTIVE INTERPRETATION AND MEASUREMENT OF DEFINITION SENTENCES

LÁSZLÓ BEDNARIK
University of Miskolc, Hungary Institute of Information Technology
laszlo.bednarik@uni-miskolc.hu

PÉTER MILEFF
University of Miskolc, Hungary Institute of Information Technology
peter.mileff@uni-miskolc.hu

**Abstract**: Computer processing of text documents for years stirs people's imaginations, as it can be used to greatly increase the efficiency of the given process. As the central theme and goal of this article, it aims to present the grammatical foundations for research in this area. The basic grammatical elements in the documents are explained, then the examination of the meaning and similarity of the words is presented using examples. Investigation were carried out in two areas: firstly, on the basis of a manually selected word from a definition sentence, during which we analyzed the substitutability of the words given by the thesaurus, the second investigation was how far the similar word is compared to a given word on a pre-prepared scale.

***Keywords:*** *word, sentence, meaning, measurement*

## 1. Introduction

Meaning in language is a dynamic and evolving property of linguistic elements. The words and structures we use in communication do not remain static; they continually adapt and change over time and across different contexts and cultures. For a computer system to effectively process and understand natural language, it must be capable of managing this fluid property of meaning, which shifts with time and varies across different spatial and cultural dimensions.

Currently, semantics stands at the forefront of research within the broader field of language technology. Semantics, which is fundamentally the science of meaning, delves into how we comprehend, interpret, and utilize the meanings of words, sentences, and larger texts. This field of study explores the relationships between signifiers—like words, phrases, signs, and symbols—and what they stand for or represent in particular contexts.

In recent years, the importance of semantic analysis has surged dramatically, particularly within the domain of natural language processing (NLP). NLP encompasses the interaction between computers and human language, aiming to equip machines with the ability to understand, interpret, and generate human language in a way that is both meaningful and useful. The role of semantic analysis in NLP is crucial as it allows for deeper understanding and more nuanced processing of language beyond mere syntactic or lexical analysis.

Several key research areas illustrate the breadth and depth of current semantic analysis endeavors:

1. **Word Sense Disambiguation (WSD):** This involves identifying which sense of a word is used in a given context, a challenge essential for accurate language understanding and translation.
2. **Semantic Role Labeling (SRL):** This technique determines the relationships between a sentence's predicates and its arguments, essentially understanding who did what to whom, when, and where.
3. **Distributional Semantics:** This area examines how the meaning of words can be derived from their context in large text corpora, leading to models that represent words as high-dimensional vectors.
4. **Ontology Learning and Knowledge Representation:** This focuses on creating structured representations of knowledge from unstructured text, allowing for the development of comprehensive frameworks that support reasoning and inference.
5. **Sentiment Analysis:** Here, the objective is to determine the sentiment expressed in a piece of text, whether it's positive, negative, or neutral, which has applications in everything from market research to social media monitoring.
6. **Machine Translation and Cross-Language Semantics:** Advancements in this area are crucial for breaking down language barriers and improving communication between speakers of different languages.
7. **Semantic Search and Information Retrieval:** This involves developing search engines and retrieval systems that understand the intent behind queries and can provide more accurate and relevant results.

These areas, among others, illustrate the profound impact that semantics has on the development and enhancement of technologies that strive to bridge the gap between human and machine communication. As we continue to advance in our understanding and application of semantics in language technology, the capabilities of machines to process and generate natural language will become increasingly sophisticated and human-like.

The research of semantic analysis has made significant advancements in the development of technologies such as chatbots, virtual assistants, automatic text summarizers, as well as search engines and recommendation systems. These technologies rely heavily on the precise and nuanced understanding of language, which semantic analysis aims to provide. Advanced models like BERT *(Bidirectional Encoder Representations from Transformers)* and GPT *(Generative Pre-trained Transformer)* have revolutionized the field, significantly enhancing the ability of computers to understand and handle language at a deep semantic level. These models leverage large-scale datasets and sophisticated algorithms to capture the complexities of human language, enabling more accurate and context-aware interactions.

The impact of these advancements is far-reaching. In chatbots and virtual assistants, for instance, improved semantic understanding allows for more natural and effective communication with users, making these tools more useful and user-friendly. In the realm of automatic text summarization, semantic analysis ensures that the summaries generated are coherent and accurately reflect the key points of the original texts. Search engines benefit from enhanced semantic capabilities by delivering more relevant and contextually appropriate search results, while recommendation systems become better at suggesting items that truly match user preferences and interests.

In this paper, we address the meaning and measurement of words and sentences, providing the appropriate grammatical foundations for the computational processing of textual documents. We explore the methodologies and techniques used to capture and quantify semantic information, and discuss their applications in various language technology systems. By understanding how meaning is constructed and interpreted, we can develop more sophisticated algorithms that improve the performance and accuracy of NLP applications.

## 2. The meaning of the word

A word is a unit of language that has its own independent form and meaning. It can also be said that a word is a unit of the linguistic sign system, i.e., a linguistic sign. Words, as linguistic signs, can be examined from many perspectives, such as their form, parts of speech, and their place in the vocabulary. The study of their meanings is the domain of semantics [1]. Semantics is the branch of linguistics that describes the meaning of word elements, words, structural units, and sentences [2].

The meaning of a word can be interpreted on multiple levels:

- Lexical Meaning: The meaning found in the dictionary, which reflects the word's basic, generally accepted sense.
- Contextual Meaning: The meaning of words often depends on the context in which they are used.
- Semantic Networks: The meaning of words is related to the meanings of other words. For example, the meaning of the word “dog” is connected to concepts like “pet”, “animal”, “barking”, and others.
- Polysemy: A word can have multiple meanings, which are determined by the context.
- Syntactic Meaning: The position of words in a sentence can also influence their meaning. For instance, the meaning of the word “lead” differs in the sentences “John leads” and “John is led”.

Understanding the meaning of a word is crucial in linguistic communication, and several fields of study address this issue, including lexicology, semantics, and pragmatics. In natural language processing, accurately determining the meaning of a word is essential for machine translation, text comprehension, and information retrieval.

Linguistics examines the meanings of words and sentences separately. A word always consists of multiple components in relation. These are: phonetic form (signifier), concept (signified), and referent (the object existing in the world). One of the best-known formulations of this is Ogden and Richards' "semiotic triangle", [3], [3a], illustrated in *Figure 1*.

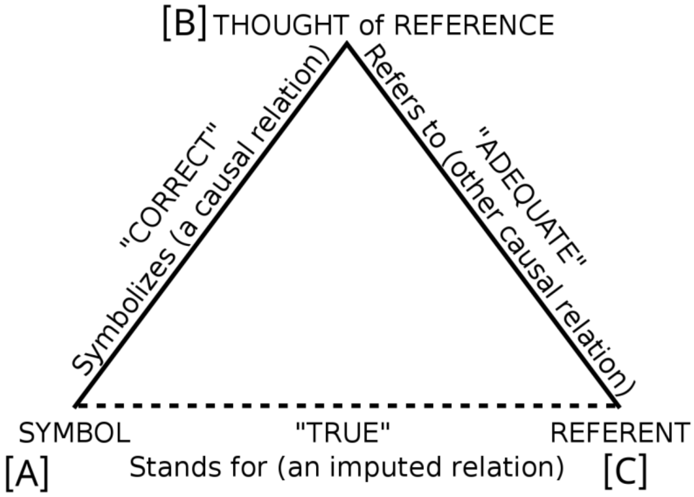
[B] THOUGHT of REFERENCE
"CORRECT"
Symbolizes (a causal relation)
Refers to (other causal relation)
"ADEQUATE"
SYMBOL
"TRUE"
REFERENT
[A]
Stands for (an imputed relation)
[C]

**Figure 1.** The semiotic triangle

The schema shows that in a given communication situation, the mutual relationship between A and B extends and becomes a three-way connection. The relationship between the sign and the referent passes through the category of "concept", but in our actual communications, the signifier always reaches the actual referent, or the object in the world, not just the concept.

The relationship between A and B is "denotation", the conceptual labeling, where A is the linguistic expression (for example, a word), while B is the specific sense of the linguistic expression: its "denotatum".

The relationship between A and C is "reference", where A is the linguistic expression (possibly a word), and C refers to the actual referents of the linguistic expression, its "referents".

Linguistics defines the meaning of a word in various ways. For instance, it breaks down the meaning of some words into elements or components, lists the attributes of the referent combined in the concept, provides the rules of word usage, or designates the class of things that the word represents and singles out the individual belonging to that class. "The meaning of a word is something complex, constituted by the peculiarities of the various meanings of the word." In other words, "meaning is not simply a reference to a single concept, but generally a complex relationship" [2].

The various types of meaning, which typically operate simultaneously, are as follows:

- denotative meaning: the primary relationship between the signifier and the signified. All other types of meaning are related to this. For example, “hill”: a surface elevation that is sloping on at least one side and is not higher than 200 meters [4], [4a].
- lexicological meaning: the relationship between the phonetic form and the meaning of the word. In this respect, “hill” is a word with two meanings associated with its phonetic form [4], [4a], meaning it is a polysemous word.
- pragmatic or stylistic meaning: indicates the possible uses of the word, originating from the style of words. For instance, the word "hill" is emotionally neutral in everyday language and is also used as a geographical term.
- connotative meaning: adds typical characteristics, values (the word’s associative connotations) to the denotative meaning but is not part of it. For example, the word “hill” has a connotative meaning in Sándor Reményik’s poem “Csigadomb”.
- collocative meaning: determines which other words’ meanings the word is most commonly associated with, i.e., in which word combinations it occurs and in which it does not. For example, the word “hill” can occur in structures like “little hill” or “hill rises”, but not with the word “clear”.
- grammatical meaning: the part-of-speech meaning of the word. The word ”hill” is a noun, signifying a conceptual notion. This is associated with syntactic meaning. For example, the noun “hill” can fulfill any role in a sentence.

From the types of meanings, we highlight the lexicological meaning of the word, which we intend to examine further. Our knowledge regarding lexicological meaning was summarized by Sándor Károly in a matrix [5].

**Table 1**. The knowledge related to lexicological meaning

| **Meaning form** | **Single meaning** | **Multiple related meanings** | **Multiple non-related meanings** |
|---|---|---|---|
| One form (mononymy) | Single-meaning words | Polysemous words | Homographs |
| Multiple similar forms (polynymy) | Form variations | Shared meanings | Separate meanings |
| Multiple different forms (heteronymy) | Synonymous words | Field relationships | Context-independent words |

In the following, we will review phenomena arising from the relationship between phonetic form and meaning, namely synonymy, ambiguity, and homonymy. Their domains of occurrence are:

1. The level of words, word elements
   - Homonymy (identical form): castle, wait [4], [4a]
   - Polysemy (multiple meanings): virus 1,2., knight 1,2,3,4,5. [4], [4a]

- Synonymy (related meanings): humor (noun): wit, cheerfulness, joke (archaic), jest, jesting, fun, fooling around, nonsense, clowning, diversion (archaic) [6]

2. <u>The level of sentences</u>
   - There are no sentences with multiple meanings in the same way as there are words with established multiple meanings. Occasionally, there are sentences with ambiguous or unclear meanings, but this characteristic arises from imprecision in the given speech situation. There are pairs of sentences with identical form, sometimes resulting in misunderstandings, for example: “I saw you sitting on the terrace.”

3. <u>The level of text</u>
   - The intention of the message, the communication situation, and the relationship between the speaker and the listener influence the meaning of the message, for example: “But I love you!”

What is the purpose of semantic analysis? Knowledge of sentence structure alone does not provide enough information for analysis. From the perspective of the message sender, a clear message can be filled with ambiguities that the message recipient must resolve [9].

## 3. The meaning of the sentence

A word in use enters into a dynamic relationship with reality as part of sentences. This relationship is meaning. The meaning of sentences is of a different nature than that of words. The association of words is determined by their fields of meaning and their semantic compatibility. Therefore, the meaning of a sentence is not identical to the sum of the meanings of the words composing it, but rather arises from the constructed association of the meanings of the words. Semantic analysis of sentences always relies on syntactic analysis. In the sentence “Julie learned to bake a cake”, the active, directed meaning of the verb “learned” determines that this action can only be attributed to Julie, not to the cake. The relationship between “learned” and “bake” is determined by the importance of the sense of “completion” in the example, which is why “learned” becomes the predicate and “to bake” becomes the object. The signifier of sentences does not indicate a single concept but rather the relationship between concepts. In the sentence “The eagle is a bird”, the signifiers of two concepts (“eagle” and “bird”) are read. These two concepts are linked in a subject-predicate structure, thereby describing a basic situation (fact). This is the most important part of the sentence’s meaning. However, other factors also contribute to the meaning of the sentence, such as the order of words within the sentence.

## 4. The measurement of meaning

Psychologists are also concerned with how to determine the semantic properties of individual words. Numerous concepts have been developed to represent the differences between words and measure the psychological distance between words.

C. E. Osgood, G. Suci, and P. Tannenbaum conducted pioneering work in this field in their book "The Measurement of Meaning" [7], where they examined the affective meaning, the emotional reactions evoked by words [3a].

Words were subjected to a study called semantic differential. Semantic differential is a special type of attitude measurement, a well-known and frequently used method for researching emotional meaning. Its development is attributed to C. Osgood. Osgood likened the method to a question-and-answer game, in which each question (e.g., "Is this good or bad?", "Fast or slow?", "Small or big?") serves the purpose of placing the concept in semantic space. Based on these questions, the participants in the study had to place the words on a scale. This was a kind of qualification of the words. If they felt, for example, that the car was "good", they marked towards the "good" end of the scale; if they felt it was "bad", they marked towards the other end [3a].

<u>The practical applications of the Osgood method:</u>

- Data collection: Based on the data collected from participants, multidimensional scores are assigned to each concept or object. These scores reflect the meanings and attitudes perceived by the participants.
- Data processing and analysis: The scores are analyzed using statistical methods to determine similarities and differences between concepts. The results are often presented in graphs or semantic maps, which visually demonstrate the semantic differences.
- Areas of application: The semantic differential method is widely used in marketing (e.g., consumer perceptions of products), psychology (e.g., assessment of personality traits), sociology (e.g., examination of social attitudes), and other fields of science.

Using the method presented above, we conducted two analyses to measure the semantic interpretation and meaning of words.

In one study, we manually selected synonyms for words from definitional sentences and examined where additional words, assignable by the synonym dictionary, were individually placed on a given scale in relation to the highlighted word.

The results of the experiment were presented through examples. Consider the following definition:

"Data refers to a sequence of signals stored in the computer, from which information can be obtained during processing." [10]

Close and distant synonyms for the word "data" include: news, detail, file, information, addition, fact, notice, information, evidence [6]. The question was given on a seven-point scale. If we feel that the word "data" can be most appropriately replaced with the word “news”, we mark it on the scale as “appropriate”; if we feel the opposite, we mark it as “inappropriate” on the scale.

Let’s place the concepts in the semantic space *(Figure 2)*, according to how replaceable we feel they are with the highlighted word from the definition!

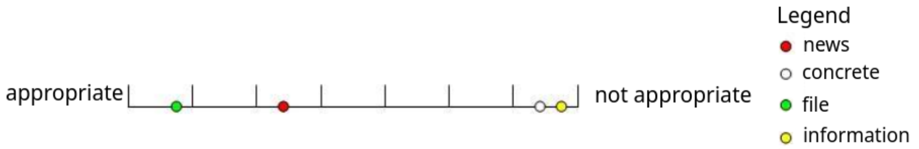

**Figure 2.** Concepts in semantic space

The result of the study: The concepts “file” and “news” are most appropriate for the word “data”. The closer the examined word is to the “appropriate” end of the scale, the better it fits into the definitional sentence. This could also mean that the word “data” can be replaced with the word “news” – the sentence would still be meaningful. The question is how replaceable words such as “news”, “detail”, “file”, and “information” are when generating complementary question types automatically by the computer for the given definition. The semantic differential method has its limitations. It only provides information about the emotions associated with the word, not its meaning.

In the other analysis, we examined the distance between a given word and similar words. This was demonstrated through a procedure where we judged the similarity between words.

Again, the results of the experiment were confirmed through examples. Consider the following definition:

“An entity is an object type, something clearly distinguished from the rest of the external world.” [10] Here, we manually selected the word “clearly”, which is the subject of the study.

Let’s place the synonyms of the word “clearly” in semantic space! evident (red), indisputable (yellow), clear (black), trivial (green), self-evident (red), clear, unmistakable, unambiguous, correct, real, explicit, understandable, indisputable, undeniable, irrefutable, beyond doubt, obvious, self-evident, visible, tangible, adequate, exact, evident, definite, precise, open, clear [6].

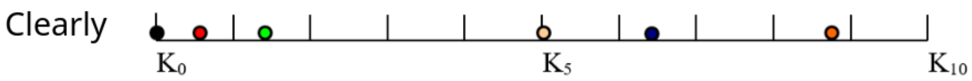

**Figure 3.** Synonyms of words in the semantic space

Let’s denote the word “clearly” with $K_0$. Let’s measure distance on a scale of ten units! Which word is closest to $K_0$?

The result of the measurement is contained in *Table 2*.

**Table 2.** Measurement results

| evident | indisputable | clear | trivial | obvious |
|---|---|---|---|---|
| K = 0.5 | K = 5 | K = 6.3 | K = 1.3 | K = 8.9 |

The results of the measurement confirm that the words “evident” and “trivial” are closest to the word “clearly”, while “clear” and “obvious” are farthest away. If we place the farthest synonyms in the definitional sentence, it becomes meaningless.

## 5. Current Technologies

“In recent years, new technologies have emerged in the field of natural language processing (NLP). One such advancement is the self-attention mechanism, which enabled one of the latest breakthroughs in NLP: the development of transformer models. Transformer models are currently considered advanced solutions for many NLP tasks, including [11]:

- Machine translation
- Question answering
- Text summarization
- Natural language inference”

**Definition of Self-Attention:** Self-attention is a technique used to create vector representations of words. This final representation is crucial for helping machines understand how different words in a sentence relate to one another.

Self-attention is closely linked to other concepts in machine learning and artificial intelligence:

- **Transformer:** Self-attention is a key component of the Transformer model, a powerful architecture that has achieved state-of-the-art results in various NLP and computer vision tasks.
- **Self-Attention Mechanism:** Self-attention is a specific type of attention mechanism that allows the model to selectively focus on relevant information.
- **BERT (Bidirectional Encoder Representations from Transformers):** BERT is a pre-trained transformer model that uses self-attention to capture contextual information in natural language. [12]

The self-attention mechanism is a technique primarily used in the field of NLP, especially in transformer models like BERT. The core idea of the mechanism is to place each word in a given sequence in context with other words, without prioritizing their positional order. This allows the model to efficiently focus on the parts of the input that are most important for the task at hand.

The self-attention mechanism uses three main components, each performing linear transformations [14]:

- **Query (Q):** A vector associated with a specific word that indicates which words it is most closely related to.

- **Key (K):** Another vector associated with every word that shows what type of information the word carries.
- **Value (V):** The third vector, which stores the actual information of the word.

These vectors are generally derived from transformations applied to the input words through a matrix.

### *5.1. Definition of Dot-Product Attention*

For each word, we calculate the scaled dot-product between the query vectors and the key vectors of the other words. The attention value is determined using Equation 5.1:

$$Attention(Q, K)= Q*K/\sqrt{d}^{k} \tag{5.1}$$

where $Q$ is the query vector, $K$ is a key vector, and $d^k$ is the dimensionality of the vectors [13].

The Dot-Product Attention method examines how similar a given word (the query) is to other words (the keys) using word vectors. This similarity is measured by the dot product between the vectors, which accounts for both the angle and magnitude between them.

Using the previous concepts (words), let’s look at an example where we determine which of the given words is closest to the word “trivial” (the query): “evident”, “indisputable”, “clear”, “trivial”, “obvious”.

**Example:** Given word vectors, which in this case are 3-dimensional:

- trivial (Q): (0.8, 0.1, 0.3)
- evident (K1): (0.7, 0.2, 0.4)
- indisputable (K2): (0.1, 0.9, 0.1)
- clear (K3): (0.6, 0.3, 0.5)
- trivial (K4): (0.8, 0.1, 0.3) (same as the query)
- obvious (K5): (0.7, 0.1, 0.5)

The attention values between the words are determined using Equation 5.1:

- Attention(Q, K1)= 0,7/1,732 = 0,404
- Attention(Q, K2)= 0,2/1,732 = 0,115
- Attention(Q, K3)= 0,66/1,732 = 0,381
- Attention(Q, K4)= 0,7/1,732 = 0,404
- Attention(Q, K5)= 0,72/1,732 = 0,416

Using Dot-Product Attention, the word “obvious” is the closest to “trivial”, indicating that “trivial” and “obvious” are very similar in meaning. This is followed by “evident” and “clear”, while “indisputable” is the least related to “trivial”. This method allows us to precisely quantify the semantic relationships between words and easily compare which words are closer to each other in a given conte.

## 6. Conclusion

To process textual documents computationally, it is essential to have a basic understanding of the words and sentences in the document. Additionally, it is important to understand the semantics of these linguistic signs, which examines the occurrence frequencies of information as carrier signs.

Using the Osgood method, we conducted two analyses to measure the semantic interpretation and meaning of words. One method resulted in substituting the highlighted word in the sentence with the closest synonym word (based on emotion), resulting in a meaningful sentence. The results of the other method were validated with numbers – also on a given scale – determined by the distance value of the words. These findings (knowledge) can be utilized in the computational processing of digital documents, such as clustering, classification, and providing possible answers to a specific question.

## References

[1] Adamikné J. A., Hangay Z. (1995). *Nyelvi elemzések kézikönyve*. Mozaik Oktatási Stúdió. Szeged.

[2] Tolcsvai N. G. (2000). *Nyelvi fogalmak kisszótára A-Zs*. Korona Kiadó, Budapest.

[3] Ogden, C. K., Richards, I. A. (1923). The meaning of meaning London: Kegan Paul, 99. p.

[3a] Forgács T (2009). *Jelentéstan (Szemantika)*. SzTE BTK, pp. 5–32.

[4] Pusztai F., Csábi Sz. (2003). *Magyar értelmező kéziszótár*. Akadémiai Kiadó Rt., Budapest.

[4a] Pusztai F. (2014). *Magyar értelmező kéziszótár + NET*. Akadémiai Kiadó Zrt., Budapest.

[5] Károly S. (1970). *Általános és magyar jelentéstan*. Akadémiai Kiadó, Budapest, pp. 78–79.

[6] Kiss G. (2004). *Magyar szókincstár. Rokon értelmű szavak, szólások és ellentétek szótára*. Tinta Könyvkiadó, Budapest.

[7] Osgood, C. E., Suci, G., Tannenbaum, P. (1957). *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.

[8] MTA Mesterséges Intelligencia Kutatócsoport (2010). *Szemantikai elemzés*.

[9] Prószéky Gábor (2016). *A nyelvtechnológia alapjai 6. Számítógépes szemantika*. Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar.

[10] Kovács L.: *Adatbázis rendszerek I.* http://moodle.iit.uni-miskolc.hu