

ANALYSIS OF RESEARCH ON THESAURUS IN COMPUTATIONAL LINGUISTICS

<https://doi.org/10.70728/phil.v05.i03.034>

Saparova Mokhira Fayzullayevna
Mamun University

Abstract: This article provides a comprehensive analysis of research on thesaurus development and application within the field of computational linguistics. Thesauri, as structured lexical-semantic resources, play a crucial role in natural language processing (NLP), information retrieval, and knowledge representation systems. The study examines the evolution of thesaurus-based approaches, including traditional lexicographic models and modern computational frameworks such as WordNet and ontology-based systems. Special attention is given to corpus-based methods, semantic similarity modeling, and automated thesaurus construction techniques, which have significantly improved the accuracy and scalability of linguistic databases. Recent studies highlight the importance of identifying inconsistencies in lexical relations and enhancing semantic networks through data-driven approaches. Furthermore, the article explores multilingual and domain-specific thesauri, emphasizing their relevance in modern AI-driven applications. The findings demonstrate that thesauri remain fundamental tools for organizing linguistic knowledge and facilitating intelligent text processing, while ongoing research continues to refine their structure, interoperability, and practical implementation in computational systems.

Keywords: computational linguistics, thesaurus, WordNet, natural language processing (NLP), semantic relations, corpus-based analysis, ontology, lexical databases, information retrieval, semantic networks

The term *computational linguistics* refers to a cognitive science and a field of artificial intelligence that studies and models the processing of natural language information in electronic form, focusing on language production and comprehension. The development and analysis of computational linguistics, particularly through the practical advancement of electronic computing machines, began in the 1940s. This new direction is an interdisciplinary field that integrates theoretical and methodological approaches from computer science, psychology, linguistics, lexicography, philosophy, and mathematical logic. It encompasses text processing, interaction with databases, and both spoken and written language processing.

As a relevant and rapidly developing field, computational linguistics combines the descriptive-classificatory methods of traditional grammar with the mathematical approaches of theoretical linguistics, requiring comprehensive morphological, lexical, syntactic, semantic, and pragmatic analysis of natural language. Additionally, computational lexicography deals with the principles of encoding and constructing lexical entries and represents an applied boundary of natural sciences. The study of thesaurus dictionaries plays an important role in computational linguistics. Such dictionaries have been extensively studied by various scholars.

In G.Carling's work, the history of corpus technology, collections of written sources from different peoples, as well as phraseology and dialectology are discussed. In the works of K. Nakayama, T. Hara, and Sh. Nishio, information is provided about WordNet-type thesauri, noting that such resources include around 80,000 nouns and 90,000 verbs, and addressing issues related to synonymy and NLP mapping. [1]

A.Kilgariff and C.Yallop, in their research, distinguish four types of thesauri, including Roget's Thesaurus, WordNet, EuroWordNet, and automatically corpus-based thesauri, highlighting their differences. They also explain the concept of taxonomy, hierarchical units, and their roles in thesaurus dictionaries with examples.[1]

H.G.Oliveira and P.Gomes, in their works, clarify the structure of words in thesauri and the concept of ontology. S.A.Zhabotynska discusses the use of cognitive linguistic methodology in building conceptual models of thesauri. The works of S.Matlatipov, M.Aripov Mirsaid, and N.Abdurakhmonova emphasize the importance of linguistic databases and the creation of Uzbek language corpora as key tasks in computational linguistics.[2]

Computational linguistics integrates the descriptive-classificatory approach of traditional grammar with mathematical methods of theoretical linguistics. It also enables the implementation of formal grammars as computer programs and their effective testing on real data. Functional analysis of natural language at morphological, lexical, syntactic, and pragmatic levels contributes to modeling communication mechanisms.

In computational linguistics, automatic language analysis is carried out using "parsers," which are computer programs that break down complex structures into simpler components. Although natural language grammar analysis is methodologically significant, computational linguistics ensures the automatic and objective validation of descriptive adequacy using real data.

It is important to note that computational linguistics is closely related to machine translation. Machine translation serves as a bridge between languages and represents an interactive and adaptive system for automatic text generation. However, translation errors often occur, and to address this issue, human involvement is

required for post-editing. In determining semantic meanings of words, thesaurus dictionaries play a crucial role.

The thesaurus was developed in 1984 at Princeton University under the leadership of psycholinguist George Miller. [3] In 1995, WordNet became publicly available on the Internet, leading to increased research in automatic text processing applications. WordNet is one of the most widely used lexical resources in computational linguistics. Version 3.0 contains approximately 155,327 words and phrases organized into 175,979 synsets, with over 207,016 token-value pairs.[3]

Since the mid-20th century, thesauri have been used as tools for creating information retrieval systems across various domains. These dictionaries serve not only as lists of words but also as sources of structured knowledge and are referred to as information retrieval thesaurus. [5]

Thesaurus dictionaries have become essential tools for information retrieval systems, and computational linguistics has played a direct role in their development. An information retrieval thesaurus (IRT) is a vocabulary of terms within a specific domain, clearly defining relationships between them and used to describe document content and search queries. A major challenge is that users must know the exact form of a term in the thesaurus for successful retrieval.

Today, automated document processing systems use information retrieval thesauri in applications such as machine translation, information retrieval, and question-answering systems. However, creating electronic linguistic resources presents challenges, especially because only parts of the vocabulary can be structured within such relational frameworks.

Relations such as “type-of” and “part-whole” are considered the least problematic, and therefore nominal vocabulary (nouns and noun phrases) dominates in such systems.

The development of electronic thesauri representing complete knowledge systems began in the late 20th century, with two main approaches: [6]

1. Creating electronic versions of existing thesauri
2. Developing new types of electronic resources with different structures and search mechanisms

Examples of the first type include electronic versions of Roget’s Thesaurus and Baranov’s Ideographic Dictionary of Russian. These resources replicate the structure and content of printed dictionaries but enhance search speed through hyperlinks, though not always improving comprehension.

One of the main problems of traditional thesauri-quickly finding the correct word-is effectively addressed by electronic thesauri. However, ease of search is not a linguistic mechanism itself but a result of system design.

It is believed that both linguistic mechanisms for knowledge representation and software tools for word retrieval are necessary. Language itself does not search but provides a limited set of options. Identifying appropriate words requires understanding intralinguistic relationships, which involve not only mathematical formulas but also semantic relations between words.

The "MorcPMel" system is an example that addresses many limitations of printed thesauri by presenting lexical systems as hierarchical structures suitable for automated text processing.

Experts highlight that one major advantage of such systems is the classification of words into semantic structures based on parts of speech. In WordNet, this results in four independent systems (nouns, verbs, adjectives, adverbs) with minimal interconnection.

However, WordNet has limitations. Fine-grained semantic distinctions in polysemous words complicate automatic processing. Another issue is the difficulty of identifying domain-specific knowledge due to distant semantic relations between words.

Currently, thesauri similar to WordNet are being developed for European languages. Notable projects include RussNet and RuThes in Russian, as well as WordNet-type systems based on Princeton's model. Uzbek thesaurus projects have also been developed based on these models.

Thesaurus dictionaries are valuable tools for indexing and retrieving lexical items, especially in selecting appropriate search terms. Modern online thesauri provide advanced search and query reformulation capabilities.

According to A. Gilchrist, thesauri differ significantly from learner dictionaries. He identifies three types of relationships in thesauri:[3]

1. Equivalent relations (synonymy and quasi-synonymy)
2. Hierarchical relations (hypernyms, hyponyms, meronyms, etc.)
3. Non-hierarchical relations (conceptually related terms)

Finally, general and specialized methodologies exist for creating thesauri for information description and retrieval, such as those developed by the Getty Research Institute and the Arts and Humanities Data Service. A thesaurus is ultimately a standardized set of terms agreed upon for representing general concepts.

The development of electronic thesauri has followed two primary trajectories: the digitization of existing print thesauri and the creation of novel digital resources with enhanced structural and functional capabilities. While electronic versions of traditional thesauri improve search efficiency through hyperlinking, they do not necessarily enhance cognitive accessibility.

A key challenge in thesaurus construction is the representation of semantic relationships, particularly in the case of polysemous words and domain-specific

terminology. While hierarchical relations such as hypernymy and meronymy are relatively stable, more complex semantic associations present ongoing difficulties.

Contemporary research efforts are focused on developing multilingual and language-specific thesauri, including projects such as RussNet, RuThes, and Uzbek WordNet-type systems. These initiatives aim to expand lexical coverage and improve the representation of semantic networks across languages.

The practical significance of thesauri lies in their ability to support indexing, information retrieval, and query formulation. As noted by A. Gilchrist, thesauri differ fundamentally from learner dictionaries in both structure and function. He identifies three principal types of semantic relations: equivalence (synonymy), hierarchical relations, and associative (non-hierarchical) relations.

In conclusion, thesaurus-based lexical resources represent a fundamental component of computational linguistics, providing structured representations of lexical knowledge that are essential for a wide range of natural language processing applications. Their continued development is critical for advancing both theoretical research and practical implementations in the field.

References

1. Killgariff and Yalooop. What is in Thesaurus? - 2000 P. 7.
2. Foskett D. Thesaurus. Readings in Information Retrieval. Morgan Kaufmann Publishers, Inc., 1997. - P. 111- 134.
3. Aitchison J, D. Bawden A. Gilchrist Thesaurus construction and use. A practical manual, 2000. - P. 34.
4. Oliveria H.G, Gomes P. Automatically enriching a thesaurus with information from dictionaries - Cognitive & Media Systems Group CISUC, Universidade de Coimbra October 11, 2011. - P. 45.
5. Matlatipov S, Aripov M, Abdurakhmonova N. Modeling WordNet type thesaurus for Uzbek language semantic dictionary - International Journal of Systems Engineering. Vol.2, No.1,2018, P. 26-28. doi: 10.11648/j.ijse.20180201.16
6. Hausser R. Foundation of computational linguistics, 1994. - P. 20-22.