CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES
(ISSN –2767-3758)
VOLUME 05 ISSUE 12 Pages: 27-32
SJIF IMPACT FACTOR (2022: 6. 041) (2023: 7. 491) (2024: 8.235)
OCLC – 1242423883
Publisher: Master Journals

Journal Website: https://masterjournals.com/index.php/crjps

Research Article

WORDNET – A LEXICAL DATABASE FOR LINGUISTIC ONTOLOGIES

Submission Date: December 02, 2024, **Accepted Date:** December 07, 2024,
Published Date: December 12, 2024
Crossref doi: https://doi.org/10.37547/philological-crjps-05-12-06

Pulatova Gulhayo
Independent researcher at Namangan State University, Uzbekistan

ABSTRACT

This article analyzes the concept of a thesaurus and its capabilities, comparing the differences between the notions of WordNet and a thesaurus. It highlights the structure, composition, and challenges faced in creating linguistic resources like the Turkish WordNet, Arabic WordNet, and the Uzbek UzWordNet, which is based on WordNet. Preliminary linguistic models serve as a foundation for the Uzbek WordNet. Additionally, ideas regarding linguistic research aimed at creating the Uzbek UzNet network are presented.

KEYWORDS

Thesaurus, WordNet structure, WordNet, KeNet, AWN (Arabic WordNet), merging, splitting.

INTRODUCTION

One of the main issues in global linguistics in recent years has been the rapid development of universal language resources. These include various resources such as software and linguistic databases (lexicons, lexical data repositories, grammars, and corpora marked in different ways), which are used in both research and industrial applications. At present, comprehensive ontologies are available for knowledge-based NLP tasks. Princeton WordNet, BabelNet, FramNet, and the European WordNet are among the most renowned ontologies.

Although manually constructing WordNet for any language is a time-consuming and labor-intensive process, it is notable for its accuracy. Existing dictionaries in a language are limited to augmenting databases. For other languages, some authors have emphasized the importance of using XML-type specially extended languages, which, in turn, facilitate

CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES
(ISSN –2767-3758)
VOLUME 05 ISSUE 12 Pages: 27-32
SJIF IMPACT FACTOR (2022: 6. 041) (2023: 7. 491) (2024: 8.235)
OCLC – 1242423883
Publisher: Master Journals

working with structured data [1]. Princeton's model has been utilized in creating WordNets for Turkish (Turkish WordNet), Arabic (Arabic WordNet), and many other languages. KeNet, another WordNet created for the Turkish language, is based on the Turkish lexicon.

METHODOLOGY

In 1985, a group of linguists and psychologists at Princeton University conducted research for the WordNet lexical database project, also referred to as "lexical ontology." The project was inspired by psycholinguist George Miller's experiments with artificial intelligence aimed at understanding human semantic memory. According to Miller, WordNet contains approximately 95,600 word forms (51,500 words and 44,100 multi-word expressions) with 70,100 meanings represented through synonyms. Among these, approximately 57,000 are nouns, of which 47,000 have recorded meanings.

The variability in the number of entries is due to the dynamic nature of the online system, where these figures are constantly evolving. Unlike traditional dictionaries, WordNet is exclusively composed of nouns, verbs, adjectives, and adverbs [2]. WordNet is a psycholinguistic lexical database created at Princeton University, widely used in language engineering research [3].

WordNet serves as a lexical database for the English language, where nouns, verbs, adjectives, and adverbs are grouped into synonym sets (Synsets) representing a specific concept. Synsets are interconnected based on conceptual-semantic and lexical relationships.

WordNet is one of the most user-friendly tools for natural language processing and computational linguistics. While it resembles a thesaurus in some aspects, significant differences exist. Notably, a thesaurus does not include semantically related words. In contrast, WordNet groups words based on semantic relationships such as synonyms, hyponyms, and meronyms. The central relationship in WordNet is synonymy. Synonyms are organized into Synsets, which include brief definitions and usage examples. Therefore, WordNet can be seen as a combination or extension of a dictionary and thesaurus.

Analyzing the WordNet system using the noun category as an example reveals that the database contains over 80,000 nouns. This includes opportunities to use lexemes as speech units or word combinations. WordNet is among the most innovative achievements in computational lexicography [3].

DISCUSSION

Robert M. Loseelar's research on thesaurus development discusses methods for incorporating thesaurus features into indexing and retrieval systems. He emphasizes that understanding how people use terms simplifies this problem and facilitates decision-making, which should rely on hypotheses about future research by the author or others [4].

Traditionally, dictionaries store information about words and their definitions. With the development of natural language processing research, the need for machine-readable dictionaries has emerged [Miller, 1995:39-41]. To meet this need, lexical networks have been created, adapting lexical data for modern computing. WordNet is one of the earliest, most comprehensive lexical databases for the English language, serving as a prototype thesaurus in some sources and as open electronic ontology in others.

WordNet is part of the lexicographical class of ontologies and is extensively used in information retrieval. Thousands of experiments in this domain have been conducted using its framework [5]. WordNet, in a broader sense, is a comprehensive dictionary built around definitions of specific word meanings.

CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES
(ISSN –2767-3758)
VOLUME 05 ISSUE 12 Pages: 27-32
SJIF IMPACT FACTOR (2022: 6. 041) (2023: 7. 491) (2024: 8.235)
OCLC – 1242423883
Publisher: Master Journals

WordNet's foundational concept is the Synset—a group of synonymous words sharing the same meaning. Given that words can have multiple meanings, a lemma may belong to several Synsets representing its various senses. Additionally, WordNet defines relationships among Synsets, such as hyperonymy, hyponymy, and meronymy. Synsets form the core of WordNet and have inspired the creation of similar electronic resources for dozens of other languages worldwide.

For instance, Turkish WordNet (KeNet), Finnish WordNet (FinnWordNet), Polish WordNet, Norwegian WordNet, Danish WordNet, and French WordNet (WOLF) have been developed. Below, we will examine the development methods and structure of Turkish WordNet (KeNet), Arabic WordNet (AWN), and Uzbek UzWordNet, which were modeled on English WordNet resources.

Currently, two word networks exist for Turkish: BalkaNet TR-wordnet and KeNet. Unlike BalkaNet, which was based on the Princeton WordNet model, KeNet was constructed from scratch using a bottom-up approach and is the most comprehensive word network for Turkish, containing 76,757 Synsets. The BalkaNet Turkish WordNet (TR-wordnet) was the first WordNet created for Turkish and encompasses approximately 14,626 Synsets and 19,834 internal semantic relationships [8].

In the creation of KeNet, managing the semantic relationships within Synsets posed significant challenges. To address these, two additional principles were implemented: merging and splitting Synsets. During the merging process, multiple Synsets were combined into groups, resulting in the addition of new definitions for 10,612 Synsets.

As with other WordNets, the first step in developing KeNet was creating Synsets, defined as groups of interchangeable words sharing the same meaning and part of speech (POS). A sample structure of a Synset in KeNet is as follows:

Table 1. Sample Synset.

Synset ID	Synset Members	Definition	Example Sentence
TUR10-0000030-n	su ab âb "water"	Hidrojenle oksijenden oluşan, oda sıcaklığında sıvı durumunda bulunan, renksiz, kokusuz, tatsız madde	
TUR10-0000220-a	abajurlu "with lampshade"	Abajuru olan	Üstünde lacivert abajurlu, parlak bir madenden lamba.
TUR10-0000350-v	abanmak "to lean over"	Egilerek bir şeyin, bir kimsenin uzerine kapanmak	Efendi, sen de ne üstume abanıyorsun?
TUR10-0000520-adv	abartısız mubalagasız "without exaggeration"	Abartmadan, abartısız olarak, mubala gasız bir biçimde	

Using Examples from the Table to Illustrate KeNet's Frequent Speech Parts

KeNet includes four major parts of speech frequently encountered: noun, adjective, verb, and adverb. For each example in the table, the first column represents the Sinset ID. The identifiers are separated by a "-" symbol, indicating the Sinset POS (n–noun, v–verb, a–

CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES
(ISSN –2767-3758)
VOLUME 05 ISSUE 12 Pages: 27-32
SJIF IMPACT FACTOR (2022: 6. 041) (2023: 7. 491) (2024: 8.235)
OCLC – 1242423883
Publisher: Master Journals

adjective, adv–adverb). The second column lists synonyms within the respective Sinset.

According to KeNet's creators, some necessary components were missing in certain Sinset groups within WordNet. While KeNet addressed some of these deficiencies, others remain unaddressed. Existing Sinset group members in KeNet were merged with WordNet synonym sets. Sinsets absent in KeNet were added to existing Sinsets as separate components. This process enriched 122 Sinsets with new members [5].

Splitting Process in KeNet

During the splitting process, Sinsets with different meanings were separated, and new Sinsets were created for each group. Semantically unrelated or incorrect Sinsets were also separated, and new definitions were assigned to them. These Sinsets are provided in Google Sheets tables for clarity.

Table 2. Number of Sinsets in KeNet

Part of Speech	# of Synsets
Nouns	44,074
Verbs	17,791
Adjectives	12,416
Adverbs	2,550
Interjections	3342
Pronouns	68
Conjunctions	60
Postpositions	29
Total	77,330

Specifically, KeNet has served as a foundational lexicon for projects like the Turkish PropBank TRopBank (Kara et al., 2020), Turkish SentiNet HisNet (Ozcelik et al., 2020), Turkish FrameNet (Marsan et al., 2020), domain-specific terms for estate (Parlar et al., 2019), and tourism (Arican et al., 2020). According to the researchers behind the KeNet project, the lack of necessary resources in the Turkish language and the relatively well-structured KeNet program make it a vital resource for further studies in this field [7].

Arabic WordNet (AWN): Limited Research and Expanding Applications

Although Arabic is the official language for hundreds of millions across 20 countries in the Middle East and North Africa, it has seen minimal research in the fields of computerization and lexical resources. Elkateb (2005) emphasized that the richness of the Arabic language necessitates the development of a lexical resource such as an Arabic WordNet. Established in 2005, the Arabic WordNet (AWN) is based on the widely recognized Princeton WordNet (PWN) in design and content. AWN is directly mapped to PWN 2.0 and EuroWordNet (EWN), enabling translation between Arabic, English, and several other languages at the lexical level.

CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES
(ISSN –2767-3758)
VOLUME 05 ISSUE 12 Pages: 27-32
SJIF IMPACT FACTOR (2022: 6. 041) (2023: 7. 491) (2024: 8.235)
OCLC – 1242423883
Publisher: Master Journals

Key criteria for selecting Sinsets for AWN include:

Connectivity: AWN must be densely interconnected through chains like hyperonymy and hyponymy. Most AWN Sinsets should correspond to English WN analogs, and the general topology of the two networks should be similar.

Relevance: Frequently encountered and salient concepts are prioritized. Criteria include the frequency of lexical elements in Arabic and English, as well as the frequency of Arabic roots in relevant corpora.

Generality: Sinsets from the highest levels of the WordNet hierarchy are given preference.

The construction of AWN occurs in two stages:

English to Arabic: For each Sinset in English, all corresponding Arabic variants (if available) are selected.

Arabic to English: For each given Arabic word, its various senses are identified, and corresponding English Sinsets are chosen.

AWN development also includes preparatory and expansion tasks. AWN is not only aligned with PWN (Fellbaum 1998) but also accommodates multi-language support via interlingual indexes (ILI) or ontologies, offering a multilingual user interface rather than a merely bilingual one. Nevertheless, to streamline the development and validation processes, existing resources are leveraged as much as possible [9].

UzWordNet: Expanding Uzbek into the Digital World

One of the projects aimed at integrating Uzbek into the digital sphere and enhancing its IT applications is UzWordNet. Created by Uzbek linguists and developers, it serves as a corpus for the largest lexical network available for the language. UzWordNet is the initial WordNet-like resource for Uzbek, corresponding to Princeton WordNet and encompassing 28,140 Sinsets, 64,389 senses, and 20,683 words [10]

The construction of UzWordNet is divided into three stages

Selection and preprocessing of lexical resources.

Automated creation of a PWN-like structure for Northern Uzbek (UzWordNet).

Human verification of the automated structure [10].

Challenges in Creating UzWordNet

Authors encountered multiple challenges during UzWordNet's development. Initially, the quality of scanned dictionaries was poor, leading to errors when converting electronic copies into text for analysis. This issue was addressed by expanding each page visually and processing it using free OCR (optical character recognition) services. The dictionary was then converted into a machine-readable format with a tabular structure comprising three columns.

Like all word networks derived from PWN, UzWordNet groups and classifies nouns, verbs, adjectives, and adverbs into Sinsets based on lexical relationships. The semantic tree structures of noun and verb Sinsets rely on hyperonym-hyponym relations

Research by Alessandro Agostini, N. Abdurahmonova, T. Usmanov, et al., titled "A Lexical-Semantic Database for the Uzbek Language", presented initial results for UzWordNet's development through expansion to Northern Uzbek. The study's evaluations achieved an accuracy of 71.64% to 75.98% for over 280 sample entries per part of speech (nouns, verbs, adjectives, adverbs) [10].

CONCLUSION

The creation of thesaurus dictionaries laid the groundwork for the development of WordNet databases in global linguistics. WordNet opened a new era in building extensive linguistic resources. Its creation sparked interest in developing WordNets for various national languages, contributing to significant advancements in fields like machine translation, computational lexicography, and corpus linguistics. It also partly addressed translation challenges and language diversity issues.

CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES
(ISSN –2767-3758)
VOLUME 05 ISSUE 12 Pages: 27-32
SJIF IMPACT FACTOR (2022: 6. 041) (2023: 7. 491) (2024: 8.235)
OCLC – 1242423883
Publisher: Master Journals

REFERENCES

1. Matlatipov S., Abdurahmonova N. Modeling Wordnet type thesaurus for uzbek language semantik dictionary International Journal of Systems Engineering 2018; 2(1): 26-28) http://www.sciencepublishinggroup.com/j/ijse
2. Fellbaum Ch. A Semantic Network of English Verbs. – In: Fellbaum, C (ed) WordNet – An Electronic Lexical Database. – The MIT Press. 1998. – pp. 69-104.
3. Abdurahmonova N. Muhammad Haydarov On the tasks of creating a wordnet in uzbek language "Ўзбекистонда хорижий тиллар" илмий-методик электрон журнал journal.fledu.uz № 4/2019
4. Robert M. Losee Decisions in thesaurus construction and use Informating Processsing $ Management Volume 43, issue 4. July 2007. Pages 958-968
5. Abjalova M. WordNet – lingvistik ontologiyalar uchun tayanch baza. O'zbekiston milliy universiteti xabarlari. Toshkent – 2023. №1.
6. Bakay O., Ergelen O., Sarmıs E. Turkish WordNet KeNet. January, 2021 Conference: Global Wordnet Conference.
7. Abjalova M. O'zbek tili ontologiyasi: yaratish texnologiyasi va konsepsiyasi. Monografiya. Toshkent – 2022. – B.54.
8. Ehsani R. Kenet: a comprehensive turkish wordnet and its applications in text clustering, isik university – 2018.
9. William B., Elkateb S. Introducing the Arabic WordNet Project fellbaum-alkhalifa-206.pdf (upc.edu)
10. Alessandro A., Abdurahmonova N. va boshqalar "A Lexical-Semantic Database for the Uzbek Language". 11th International Global Wordnet Conference. January – 2021.