Comparative Study on Currently Available WordNets

Sreedhi Deleep Kumar
PG Scholar
Department of Computer Science and Engineering
Vidya Academy of Science and Technology
Thrissur, India.

Reshma E U
PG Scholar
Department of Computer Science and Engineering
Vidya Academy of Science and Technology
Thrissur, India.

Sunitha C
Associate Professor
Department of Computer Science and Engineering
Vidya Academy of Science and Technology
Thrissur, India.

Amal Ganesh
Assistant Professor
Department of Computer Science and Engineering
Vidya Academy of Science and Technology
Thrissur, India.

Abstract

WordNet is an information base which is arranged hierarchically in any language. Usually, WordNet is implemented using indexed file system. Good WordNets available in many languages. However, Malayalam is not having an efficient WordNet. WordNet differs from the dictionaries in their organization. WordNet does not give pronunciation, derivation morphology, etymology, usage notes, or pictorial illustrations. WordNet depicts the semantic relation between word senses more transparently and elegantly. In this work, a general comparison of currently browsable WordNets are done. There are many WordNets available globally. However, the users are able to browse few among them. Hence this work will enable to analyze the statistics and the usage difference of each of the WordNets. It also includes their differences in the storages, structures, user accessibility etc.

Keywords: WordNet, indexed file system, Synsets, Multi-lingual

INTRODUCTION

In the area of Natural Language Processing, WordNet plays an important role. Wordnet is a semantic dictionary that was designed as a network following the idea that representing words and concepts as an interrelated system is consistent with evidence for the way speakers organize their own mental lexicons[1]. Nowadays, WordNets are available in many Languages. But in Malayalam there is not a good wordnet available yet.

India is a country with diverse culture, language and varied heritage. Due to this, it is very rich in languages and their dialects. Being a multilingual society, a dictionary in multiple languages becomes its need and one of the major resources to support a language. There are dictionaries for many Indian languages, but very few are available in multiple languages. WordNet is one of the most prominent lexical resources in the field of Natural Language Processing. There are numerous languages in India which belong to different language families. These language families are Indo-Aryan, Dravidian, SinoTibetan, Tibeto-Burman and Austro-Asiatic. The major ones are the Indo-Aryan, spoken by the northern to western part of India and Dravidian, spoken by southern part of India. The Eighth Schedule of the Indian Constitution lists 22 languages, which have been referred to as scheduled languages and given recognition, status and official encouragement.

A Dictionary can be called as are source dealing with the individual words of a language along with its orthography, pronunciation, usage, synonyms, derivation, history, etymology, etc. arranged in an order for convenience of referencing the words. Various criterions used for classifying this resource are - density of entries, number of languages involved, nature of entries, degree of concentration on strictly lexical data, axis of time, arrangement of entries, purpose, prospective user, etc. Some of the common types of dictionaries are

- **Encyclopedia**: Single or multi-volume publication that contains accumulated and authoritative knowledge on a subject arranged alphabetically. E.g. Britannica encyclopedia.
- **Thesaurus:** Thesaurus is a dictionary that lists words in groups of synonyms and related concepts
- **Etymological Dictionary**: An etymological dictionary discusses the etymology/origin of the words listed. It is the product of research in historical linguistics.
- **Dialect Dictionary**: These dictionaries deal with the words of a particular geographical region or social group which are non standard.
- **Specialized Dictionary**: These dictionaries covers relatively restricted set of phenomena.
- **Bilingual or Multilingual Dictionary**: These are linguistic dictionaries in two or more languages.
- **Reverse Dictionary**: These dictionaries are based on the concept/idea/definition to words.

- **Learners Dictionary**: These dictionaries are meant for foreign students/tourists to learn the usage of the word in language.
- **Phonetic Dictionary**: These dictionaries help in searching the words by the way they sound.
- **Visual Dictionary**: These dictionaries use pictures to illustrate the meaning of words.

WordNet is a semantic dictionary that was designed as a network following the idea that representing words and concepts as an interrelated system. WordNet groups words into sets of synonyms and provides short definitions and usage examples, also records a number of relations among these synonym sets or their members.

WordNet can be seen as a combination of dictionary and thesaurus. WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms,but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

Properties of WordNet

A WordNet can provide the following information:

• **Synonymy:**

This one is easy and links words that have similar meanings, e.g. happy and glad.

• **Antonymy:**

The opposite of synonymy, e.g. happy and sad

• **Hypernymy**:

Hypnernymy refers to a hierarchical relationship between words. For example, furniture is a hypernym of chair since every chair is a piece of furniture (but not vice-versa).

• **Hyponymy**:

Hyponymy is the opposite of hypernymy. Dog is a hyponym of canine since every dog is a canine.

• **Meronymy**:

Meronymy refers to a part/whole relationship. For example, paper is a meronym of book, since paper is a part of a book.

• **Troponymy**:

Troponymy is the semantic relationship of doing something in the manner of something else. For example, walk is a troponym of move and limp is a troponym of walk.

RELATED WORKS

This section enumerate the available WordNets and their properties

I. Princeton WordNet

This is the first WordNet to be developed. WordNet was created in the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George Armitage Miller starting in 1985 and has been directed in recent years by Christiane Fellbaum. WordNet is a lexical database for the English language. It groups English words in to sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. As of November 2012 WordNet's latest Online-version is 3.1. The database contains 155,287 words organized in 117,659 synsets for a total of 206,941 wordsense pairs; in compressed form, it is about 12 megabytes in size. WordNet includes the lexical categories nouns, verbs, adjectives and adverbs but ignores prepositions, determiners and other function words. Words from the same lexical category that are roughly synonymous are grouped into synsets. Synsets include simplex words as well as collocations like ”eat out” and ”car pool.” The different senses of a polysemous word form are assigned to different synsets. The meaning of a synset is further clarified with a short defining gloss and one or more usage examples. An example adjective synset is: good, right, ripe (most suitable or right for a particular purpose; ”a good time to plant tomatoes”; ”the right time to act”; ”the time is ripe for great sociological changes”) All synsets are connected to other synsets by means of semantic relations. These relations, which are not all shared by all lexical categories, include: Nouns.

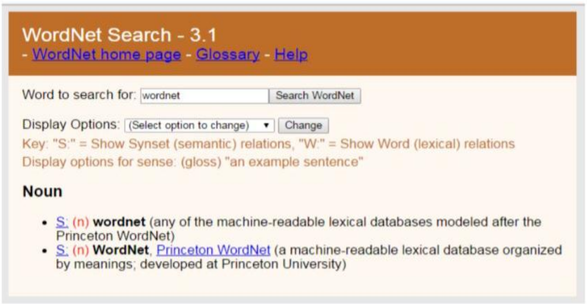
These semantic relations hold among all members of the linked synsets. Individual synset members (words) can also be connected with lexical relations. For example, (one sense of) the noun ”director” is linked to (one sense of) the verb”direct” from which it is derived via a ”morphosemantic” link.

The morphology functions of the software distributed with the database try to deduce the lemma or stem form of a word from the user's input. Irregular forms are stored in a list, and looking up ”ate” will return ”eat,” for example.

The initial goal of the WordNet project was to build a lexical database that would be consistent with theories of human semantic memory developed in the late 1960s. Psychological experiments indicated that speakers organized their knowledge of concepts in an economic, hierarchical fashion. Retrieval time required to access conceptual knowledge seemed to be directly related to the number of hierarchies the speaker needed to ”traverse” to access the knowledge.

Thus, speakers could more quickly verify that canaries can sing because a canary is a songbird, but required slightly more time to verify that canaries can fly and even more time to verify canaries have skin . While such experiments and the underlying theories have been subject to criticism, some of WordNet's organization is consistent with experimental evidence.

The figure of Princeton WordNet is as shown below.



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: wordnet [Search WordNet]

Display Options: (Select option to change) ▾ [Change]
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **wordnet** (any of the machine-readable lexical databases modeled after the Princeton WordNet)
- S: (n) **WordNet**, Princeton WordNet (a machine-readable lexical database organized by meanings; developed at Princeton University)

Figure : Princeton WordNet

II. EuroWordNet

EuroWordNet is a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The wordnets are structured in the same way as the American wordnet for English (Princeton WordNet, Miller et al 1990) in terms of synsets (sets of synonymous words) with basic semantic relations between them. Each wordnet represents a unique language-internal system of lexicalizations. In addition, the wordnets are linked to an Inter-Lingual-Index, based on the Princeton wordnet. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. The index also gives access to a shared top-ontology of 63 semantic distinctions. This top-ontology provides a common semantic framework for all the languages, while language specific properties are maintained in the individual wordnets. The database can be used, among others, for monolingual and cross-lingual information retrieval, which was demonstrated by the users in the project. The EuroWordNet project was completed in the summer of 1999. The design of the database, the defined relations, the top-ontology and the Inter-Lingual-Index are now frozen. Nevertheless, many other institutes and research groups are developing similar wordnets in other languages (European and non-European) using the EuroWordNet specification. If compatible, these wordnets can be added to the above database and, via the index, connected to any other wordnet. The EuroWordNet format is defined by the EuroWordNet Database Editor Polaris.

Multilingual WordNet Database :

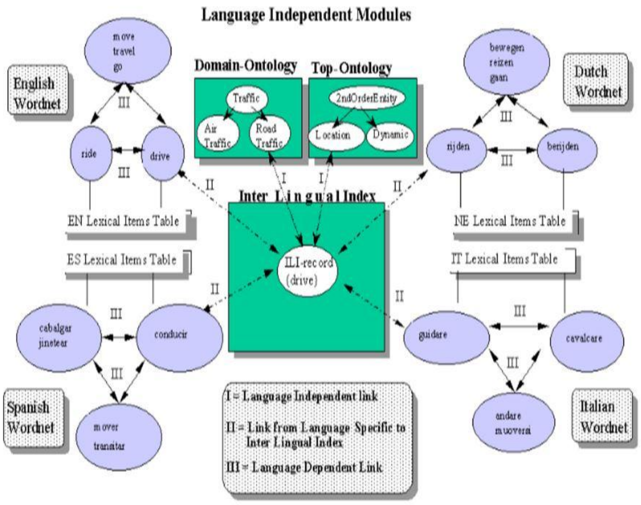
Unfortunately, such resources are not available for other languages than English, let alone are source in which multiple wordnets are combined and interlinked. This severely holds back developments in language engineering and the information society in Europe.

The aim was to develop such a multilingual database with wordnets for several European languages which can be used to improve recall of queries via semantically linked variants in any of these languages. These European wordnets have as much as possible been built from available existing resources and databases with semantic information developed in various national and European projects (Acquilex, Sift).

This is not only more cost-effective but also made it possible to combine information from independently created resources. This made the database more consistent and reliable, while keeping the richness and diversity of the vocabularies of the different languages.

The wordnets have been stored in a central lexical database system and the word meanings have been linked to meanings in the Princeton WordNet1.5, which functions as the so-called Inter-Lingual-Index. Furthermore, they merged the major concepts and words in the individual wordnets to form a common language-independent ontology (an ontology is the set of semantic relations between concepts). This guarantees compatibility and maximizes the control over the data across the different wordnets while language-dependent differences can be maintained in the individual wordnets.

The overall architecture of EuroWordNet is as shown in the diagram.



I. IndoWordNet

IndoWordNet is an integrated multilingual WordNet for Indian languages. These WordNet resources are used by researchers to experiment and resolve the issues in multilinguality through computation. However, there are few cases where WordNet is used by the non-researchers or general public. IndoWordNet is a linked lexical knowledge base of wordnets of 18 scheduled languages of India, viz., Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Meitei (Manipuri), Marathi, Nepali, Odia, Punjabi, Sanskrit, Tamil, Telugu and Urdu. IndoWordNet Dictionary or IWN Dictionary is an online interface to render multilingual IndoWordNet information in the dictionary format. It allows user to view the results in multiple formats as per the need. Also, user can view the result in multiple languages simultaneously.

The look and feel of the IWN Dictionary is kept same as a traditional dictionary keeping in mind the user adaptability. So far, it renders WordNet information of 19 Indian languages. Dictionary words are included in the wordnet according to the frequency of their use. Transliteration, Short phrase, Coined word are typically needed in expanding from a culture or region specific concept. However, these options should be used with discretion, respecting the native speakers sensitivities. The Indo WordNet uses linked structure for storing the data.

The different views in IWN Dictionary are:

• **Sense Based view**: All the meanings of an input word are displayed with respect to the senses available in the IWN database. Here, each sense is shown in a different card, where user can click or unclick to get the corresponding senses in other languages.

• **Thesaurus Based view**: In thesaurus based view, synonymous words in each language are rendered. Here, user can click on any of the words in the list to go to see other senses of that word.

• **Word Usage Based view** : In word usage based view, usage of an input word with respect to the languages is rendered. Here, the examples of a synset from IWN database are rendered.

• **Language Based view**: In language based view, meaning of a word is rendered with respect to the language. Here, for each sense of a word, the meaning in all the languages is rendered in a horizontal tabbed format or a card format. IndoWordNet is publicly browsable.

In developing the IndoWordNet the following considerations have been kept. Wordnets central concern is to express a concept unambiguously. To express concepts with a set of word (s) we can follow these options:

– Dictionary words

– Transliteration

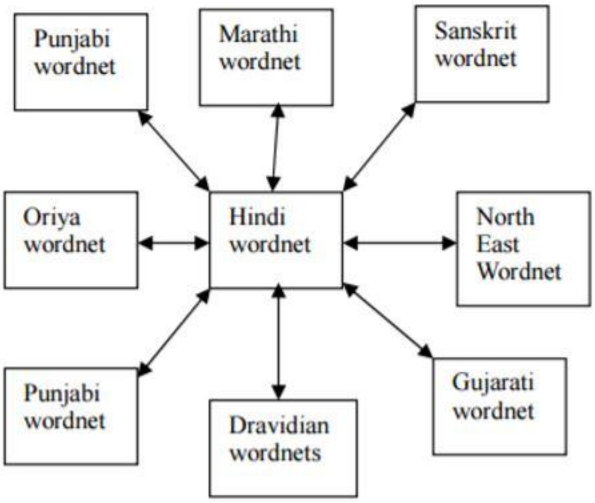
– Short phrase

– Coined word

• Dictionary words are included in the wordnet according to the frequency of their use. Transliteration, Short phrase, Coined word are typically needed in expanding from a culture or region specific concept.

The Indo WordNet uses linked structure for storing the data. The basic structure of linked Wordnet is as depicted in the figure.

Linked IndoWordnet structure is as below:



The IndoWordNet is as shown below.

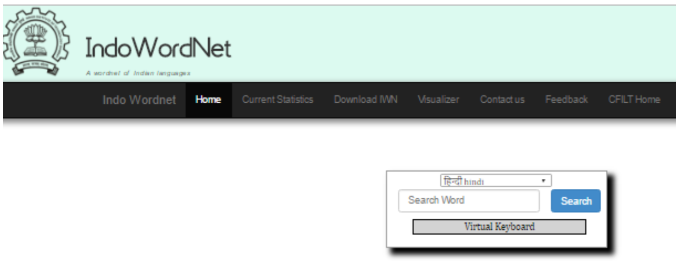
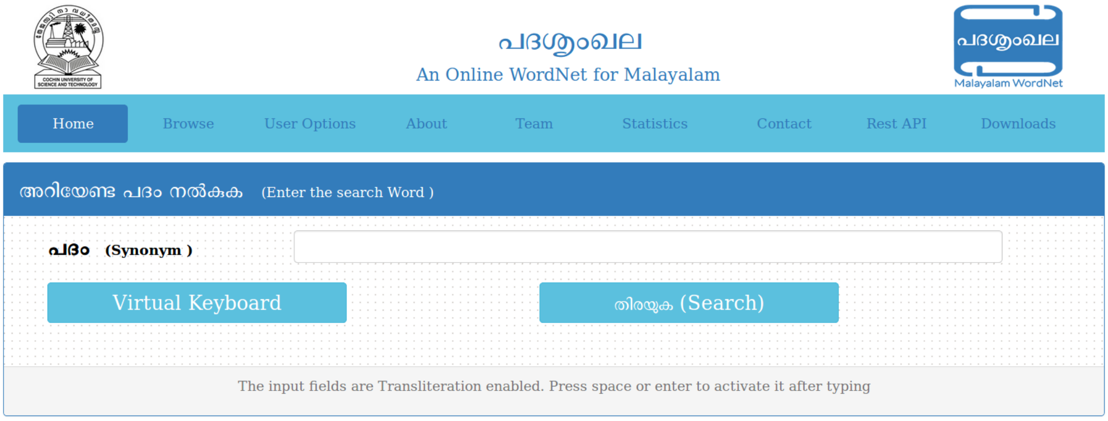


Figure: IndoWordNet

II. *Padasringala (Malayalam WordNet)*

Malayalam WordNet is a component of Dravidian WordNet which in turn is the component of IndoWordNet. Malayalam WordNet is an online lexical database. Malayalam WordNet aims to capture the net work of lexical or semantic relations between lexical items or words in Malayalam. Malayalam WordNet is an crowd sourced project. IndoWordNet is publicly browsable, but it is not available to edit. Malayalam WordNet allows users to add data to the WordNet in an controlled crowd sourcing manner. Either a set of experts or users itself could review the entries added by other members which helps in maintaining consistent data throughout. It also has a JSON and XML interfaces which helps the programmers to interact with the WordNet. It would be highly useful for the researchers, language experts as well as application developers.

The figure of "padasringala" as shown below:



The input fields are Transliteration enabled. Press space or enter to activate it after typing

Figure: Padasringala

COMPARISON TABLE

The comparison table is given below

Table 1. Comparison of Methods

Sl.No	Review of the existing WordNets		
	WordNets	**Language**	
1	Princeton WordNet	English	The database contains 155,287 words organized in 117,659 synsets for a total of 206,941 wordsense pairs; in compressed form, it is about 12 megabytes in size.
2.	EuroWordNet	European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian)	The wordnets are linked to an Inter-Lingual-Index, based on the Princeton wordnet. The languages are interconnected.
3.	IndoWordNet	Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Meitei (Manipuri), Marathi, Nepali, Odia, Punjabi, Sanskrit, Tamil, Telugu, Urdu	Dictionary words are included in the wordnet according to the frequency of their use.
4.	Malayalam WordNet	Malayalam	It has a JSON and XML interfaces which helps the programmers to interact with the WordNet.

CONCLUSION

A brief summary of different currently available WordNets are presented in this paper. Also a comparison of the properties of the different WordNets are discussed. It can be inferred that the methods will lean on the complexity of the language that we choose. It is difficult to discover a single method for all the different languages. The creation of the WordNets will depend on the type of the language also. Different methods used for implementing the WordNets are also mentioned.

REFERENCES

[1] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller (Revised August 1993) The Burning Velocity of Methane-Air Mixtures, Introduction to WordNet: An On-line Lexical Database.

[2] Piek VossenEuroWordNet: a multilingual database for information retrieval ,University of Amsterdam Delos Workshop on Cross-language Information Retrieval, Chicago, 1997

[3] Dr. Pushpak Bhattacharyya,IIT Bombay IndoWordNet Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC10)

[4] S. Rajendran(Tamil University) G.Shivapratap, V.Dhanlakshmi, KP. Soman(Amrita Vishwa Vidyapeeth) Building a WordNet for Dravidian Languages Fifth International Conference of the Global WordNet Association (GWC-2010)

[5] Mujeeb Rehman o, P. C. Reghu Raj Malayalam Wordnet: A Relational DatabaseApproachInternationalJournalofLatestTrendsin Engineeringand Technology (IJLTET)

[6] Venkatesh Prabhu,Shilpa Desai,Hanumant Redkar, Neha Prabhugaonkar,Apurva Nagvenkar,Ramdas Karmali An Efcient Database Design for IndoWordNet Development Using Hybrid Approach Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language

[7] Processing (SANLP), pages 229236, COLING 2012, Mumbai, December 2012

[8] Pushpak Bhattacharyya, Christiane Fellbaum, Piek Vossen 2010. Principles, Construction and Application of MultilingualWordNets, Proceedings of the 5th Global WordNet Conference(MumbaiIndia), 2010

[9] GeorgeA.Miller1995. WordNet: A Lexical Database for EnglishCommunication Technologies (ICT), 2013 IEEE Conference on, pp. 1138–1143, IEEE, 2013.