CHAPTER III
WORDNET AND MMD

Introduction

This chapter aims at creating a multilingual, sense-disambiguated dictionary from information harvested from WordNet. WorldNet's features, database contents, knowledge structure justifications, and the characteristic of WordNet as Ontology and so on are discussed in this chapter. Brief information regarding the project, which is working on WordNet, based on different languages like Princeton, IndoWordNet. In addition to this discussion, it will discuss about advantages of multilingual dictionary and WordNet, limitations of WordNet and multilingual dictionary. Raw data extraction for corpus from WordNet for next chapter will also be highlighted in this chapter.

3.1 WordNet

George A. Miller first started WordNet in the mid-1980s, who passed away on 22nd July 2012 at the age of 92. The origin of the tool is to build a lexical-conceptual model and database. It can define as lexical database of English language; WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human memory in lexical form. WordNet is still maintained by the Cognitive Science Laboratory. In WordNet nouns, verbs, adjectives and adverbs are organized into set of synonyms (synset), which represents one underlying lexical concept. A synset contains a brief definition or gloss and in most cases, one or more short sentence explains the uses of synset member. Synset represent the word forms with several distinct meaning, which are unique. WordNet structure makes it a useful tool for computational linguistics and Natural Language Processing. It mainly resembles a thesaurus, in the sense, it groups words together based on meaning. However WordNet interlinks not only word forms but also specific sense of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated.

The most frequently encoded relation among synset is the super- subordinate relation, i.e. hyponymy. All noun hierarchies ultimately go up the root node. And Hyponymy relation is transitive. This expressed manner depends on the semantic field;

volume is just one dimension along which verbs can be elaborated while adjectives are organized in terms of antonymy. The Princeton WordNet search page is given below:

Figure 9: Princeton WordNet page for searching a lexicon

Source: http://wordnet.princeton.edu

3.1.1 Concept of Lexical Matrix

Table 5: Concept of Lexical Matrix in WordNet

Word Meanings	Word Forms				
	F_1	F_2	F_3	. . .	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
⋮				⋱	
M_m					$E_{m,n}$

3.1.2 Database Contents

As of November 2012, WorldNet's latest Online-version is 3.1 (announced on June 2011), but the latest released version is 3.0 (released on December 2006)(Wordnet.princeton.edu, 2012). The 3.0 databases contain 155,287 words organized in 117,659 synsets for a total of 206,941 word-sense pairs; in compressed form, it is about 12 megabytes in size (Wordnet.princeton.edu, 2014).

WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. It does not include prepositions, determiners etc. Every synset contains a group of synonymous words or collocations (a *collocation* is a sequence of words that go together to form a specific meaning, such as "car pool"); different senses of a word are in different synsets. The meaning of the synsets is further clarified with short defining *glosses* (Definitions and/or example sentences). A typical example synset with gloss 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")

Most synonym sets are connected to other synsets via a number of semantic relations. These relations vary based on the type of word and include:

1. Nouns

   Hypernyms: *Y* is a hypernym of *X* if every *X* is a (kind of) *Y* (*canine* is a hypernym of *dog*)

   Hyponyms: *Y* is a hyponym of *X* if every *Y* is a (kind of) *X* (*dog* is a hyponym of *canine*)

   coordinate terms: *Y* is a coordinate term of *X* if *X* and *Y* share a hypernym (*wolf* is a coordinate term of *dog*, and *dog* is a coordinate term of *wolf*)

   Holonym: *Y* is a holonym of *X* if *X* is a part of *Y* (*building* is a holonym of *window*)

   Meronym: *Y* is a meronym of *X* if *Y* is a part of *X* (*window* is a meronym of *building*)

2. Verbs

   hypernym: the verb Y is a hypernym of the verb *X* if the activity *X* is a (kind of) *Y* (*to perceive* is an hypernym of *to listen*)

   Troponym: the verb *Y*is a troponym of the verb *X*if the activity *Y* is doing *X* in some manner (*to lisp* is a troponym of *to talk*)

   Entailment: the verb *Y* is entailed by *X* if by doing *X* you must be doing *Y* (*to sleep* is entailed by *to snore*)

   coordinate terms: those verbs sharing a common hypernym (*to lisp* and *to yell*)

3. Adjectives

   related nouns

   similar to

   participle of verb

4. Adverbs

   root adjectives

While semantic relations apply to all members of a synset because they share a meaning but are all mutually synonyms, words can also be connected to other words through lexical relations, including antonyms (opposites of each other) which are derivationally related, as well.

WordNet also provides the polysemy *count* of a word: the number of synsets that contains the word. If a word participates in several synsets (i.e. has several senses) then typically some senses are much more common than others. WordNet quantifies this by the frequency score: in which several sample texts have all words semantically tagged with the corresponding synset, and then a count provided indicating how often a word appears in a specific sense.

The morphology functions of the software distributed with the database try to deduce the lemma or root form of a word from the user's input; only the root form is stored in the database unless it has irregular inflected forms.

3.1.3 Knowledge Structure

Both nouns and verbs are organized into hierarchies, defined by hypernym or *IS A* relationships. For instance, the first sense of the word *dog* would have the following hypernym hierarchy; the words at the same level are synonyms of each other: some sense of *dog* is synonymous with some other senses of *domestic dog* and *Canis lupus familiar is*, and so on. Each set of synonyms (*synset*), has a unique index and shares its properties, such as a gloss (or dictionary) definition.

dog, domestic dog, Canis familiaris
=> canine, canid
=> carnivore
=> placental, placental mammal, eutherian, eutherian mammal
=> mammal
=> vertebrate, craniate
=> chordate
=> animal, animate being, beast, brute, creature, fauna

At the top level, these hierarchies are organized into base types, 25 primitive groups for nouns, and 15 for verbs. These groups form *lexicographic files* at a maintenance level. These primitive groups are connected to an abstract root node that has, for some time, been assumed by various applications that use WordNet. In the case of adjectives, the organization is different. Two opposite 'head' senses work as binary poles, while 'satellite' synonyms connect to each of the heads via synonymy relations. Thus, the hierarchies, and the concept involved with lexicographic files, do not apply here in same way they do for nouns and verbs.

The network of nouns is far deeper than that of the other parts of speech. Verbs have a far *bushier* structure, and adjectives are organized into many distinct clusters. Adverbs are defined in terms of the adjectives they are derived from, and thus inherit their structure from that of the adjectives.

3.1.4 Psychological Justification

The goal of WordNet was to develop a system that would be consistent with the knowledge acquired over the years about how human beings process language. Anomic aphasia, for example, creates a condition that seems to selectively encumber

individuals' ability to name objects; this makes the decision to partition the parts of speech into distinct hierarchies more of a principled decision than an arbitrary one.

In the case of hyponymy, psychological experiments revealed that individuals could access properties of nouns more quickly depending on when a characteristic becomes a defining property. That is, individuals can quickly verify that *canaries can sing* because a canary is a songbird (only one level of hyponymy), but require slightly more time to verify that *canaries can fly* (two levels of hyponymy) and even more time to verify *canaries have skin* (multiple levels of hyponymy). This suggests that we too store semantic information in a way that is much like WordNet, because we only retain the most specific information needed to differentiate one particular concept from similar concepts. (Colins A., Quillian M. R., 1972).

3.1.5 WordNet as Ontology

The hypernym/hyponym relationships among the noun synsets can be interpreted as specialized relationship between conceptual categories. In other words, WordNet can be interpreted and used as a lexical ontology in the computer science sense. However, such an ontology should normally be corrected before being used since it contains hundreds of basic semantic inconsistencies such as (i) the existence of common specializations for exclusive categories and (ii) redundancies in the specialization hierarchy. Furthermore, transforming WordNet into a lexical ontology usable for knowledge representation should normally also involve (i) distinguishing the specialization relations into *subtype Of* and *instance Of* relations, and (ii) associating intuitive unique identifiers to each category. Although such corrections and transformations have been performed and documented as part of the integration of WordNet 1.7 into the cooperatively updatable knowledge base of WebKB-2, most projects claiming to re-use WordNet for knowledge-based applications (typically, knowledge-oriented information retrieval) simply re-use it directly. WordNet has also been converted to a formal specification, by means of a hybrid bottom-up top-down methodology to automatically extract association relations from WordNet, and interpret these associations in terms of a set of conceptual relations, formally defined in the DOLCE foundational ontology (Gangemi, et al. 2003)

3.1.6 WordNet Features

The most obvious difference between WordNet and a standard dictionary is that WordNet divides the lexicon into five categories: nouns, verbs, adjectives, adverbs, and function words. Actually, WordNet contains only nouns, verbs, adjectives, and adverbs. The relatively small set of English function words is omitted on the assumption (supported by observations of the speech of aphasic patients: Garrett, 1982) that they are probably stored separately as part of the syntactic component of language. The realization that syntactic categories differ in subjective organization emerged first from studies of word associations. Fillenbaum and Jones (1965), for example, asked English speaking subjects to give the first word they thought of in response to highly familiar words drawn from different syntactic categories. The modal response category was the same as the category of the probe word: noun probes elicited nouns responses 79% of the time, adjectives elicited adjectives 65% of the time, and verbs elicited verbs 43% of the time.

The price of imposing this syntactic categorization on WordNet is a certain amount of redundancy that conventional dictionaries avoid words like back; for example, turn up in more than one category. But the advantage is that fundamental differences in the semantic organization of these syntactic categories can be clearly seen and systematically exploited.

The most ambitious feature of WordNet is its attempt to organize lexical information in terms of word meanings, rather than word forms. In that respect, WordNet resembles a thesaurus more than a dictionary, and, in fact, Laurence Urdang's revision of Rodale's *The Synonym Finder* (1978) and Robert L. Chapman's revision of *Roget's International Thesaurus* (1977) have been helpful tools in putting WordNet together. But neither of these excellent works is well suited for the printed form. The problem with an alphabetical thesaurus is redundant entries: if word *Wx* and word *Wy* are synonyms, the pair should be entered twice, once alphabetized under *Wx* and again alphabetized under *Wy*. The problem with a topical thesaurus is that two look-ups are required, first on an alphabetical list and again in the thesaurus proper, thus doubling a user's search time. These are, of course, precisely the kinds of mechanical chores that a computer can perform rapidly and efficiently. However, WordNet is not merely an

online thesaurus, however. In order to appreciate what more has been attempted in WordNet, it is necessary to understand its basic design (Miller and Fellbaum, 1991). It is characteristic of semantic relations that are reciprocated: if there is a semantic relation R between meaning {*x, x′, . . .*} and meaning {*y, y′, . . .*}, then there is also a relation R′ between {*y, y′, . . .*} and {*x, x′, . . .*}. For the purpose of the present discussion, the names of the semantic relations will serve a dual role: if the relation between the meanings {*x, x′, . . .*} and {*y, y′, . . .*} is called R, then R will also be used to designate the relation between individual word forms belonging to those synsets. It might be logically tidier to introduce separate terms for the relation between meanings and for the relation between forms, but even greater confusion might result from the introduction of so many new technical terms.

All three kinds of semantic relations- hyponymy, meronymy and antonymy are included; the result is highly interconnected network of nouns. A graphical representation is shown below

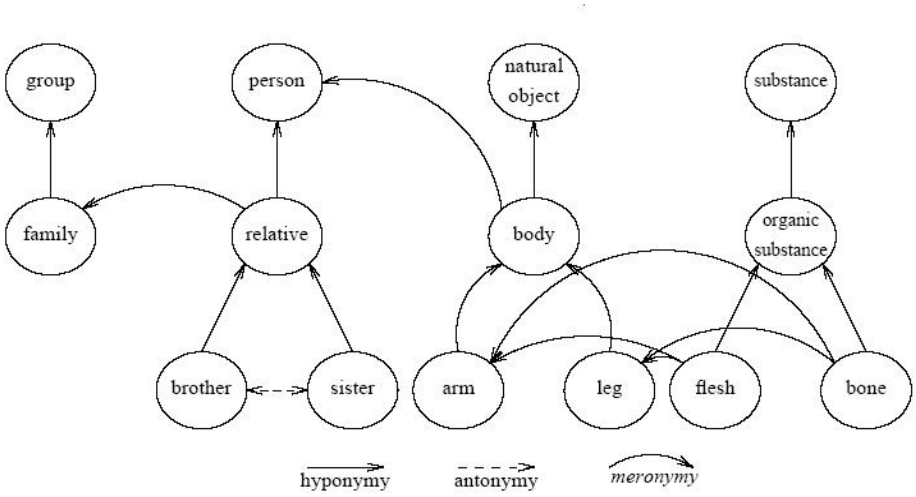

Figure 10: WordNet Semantic Relationship in English

For Manipuri language, the given figure shows the relations as the WordNet depicts. The relations between the lexicons are given by different symbols as given below.

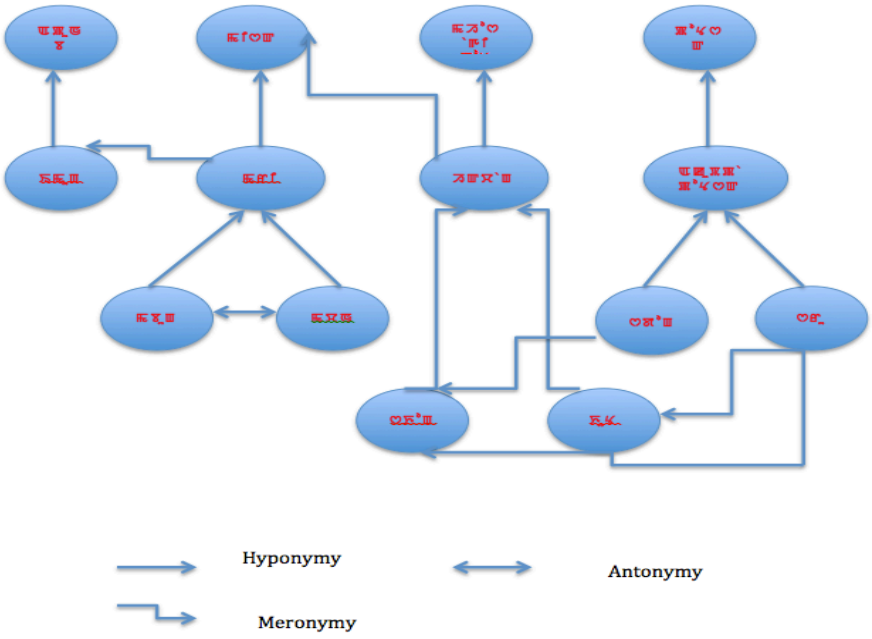

Figure 11: WordNet Semantic Relationship in Manipuri.

3.1.7 Four Kinds of Entailment

In WordNet we find four kinds of Entailment, which are Noun, Verb, Adjectives and Adverbs only from the general Part of Speech. The table below shows the kinds of entailment and its sub categories with their standard symbols in WordNet.

Table 6: Four kinds of Entailment in WordNet

Noun	Verb	Adjective	Adverb
Antonym !	Antonym !	Antonym !	Antonym !
Hyponym ~	Troponym ~	Similar &	Derived from \
Hypernym @	Hypernym @	Relational Adj \	
Meronym #	Entailment *	Also see .	
Holonym %	Cause >	Attribute =	
Attribute =			

3.1.7.1 Synonymy

From what has already been said, it should be obvious that the most important relation for WordNet is similarity of meaning, since the ability to judge that relation between word forms is a prerequisite for the representation of meanings in a lexical

matrix. According to one definition (usually attributed to Leibniz), two expressions are synonymous if the substitution of one for the other never changes the truth-value of a sentence in which the substitution is made. By that definition, true synonyms are rare, if they exist at all. A weakened version of this definition would make synonymy relative to a context: two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth-value. For example, the substitution of plank for board will seldom alter truth-values in carpentry contexts, although there are other contexts of board where that substitution would be totally inappropriate.

The definition of synonymy in terms of truth-values seems to make synonymy a discrete matter: two words either are synonyms or they are not. But as some philosophers have argued, and most psychologists accept without considering the alternative, synonymy is best thought of as one end of a continuum along which similarity of meaning can be graded. It is probably the case that semantically similar words can be interchanged in more contexts than can semantically dissimilar words. But the important point here is that theories of lexical semantics do not depend on truth- functional conceptions of meaning; semantic similarity is sufficient. It is convenient to assume that the relation is symmetric: if x is similar to y, then y is equally similar to x. The gradability of semantic similarity is ubiquitous, but it is most important for understanding the organization of adjectival and adverbial meanings.

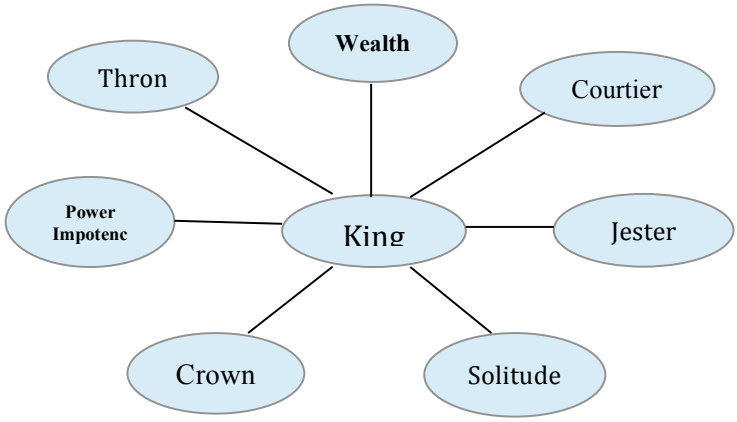

Figure 12: Synonymy Diagram

3.1.7.2 Antonymy

Another familiar relation is antonymy, which turns out to be surprisingly difficult to define. The antonym of a word x is sometimes not-x, but not always. For example, rich and poor are antonyms, but to say that someone is not rich does not imply that they must be poor; many people consider themselves neither rich nor poor. Antonymy, which seems to be a simple symmetric relation, is actually quite complex, yet speakers of English have little difficulty recognizing antonyms when they see them. Antonymy is a lexical relation between words forms, not a semantic relation between word meanings. For example, the meanings {rise, ascend} and {fall, descend} may be conceptual opposites, but they are not antonyms; [rise/fall] are antonyms and so are [ascend/descend], but most people hesitate and look thoughtful when asked if rise and descend, or ascend and fall, are antonyms. Such facts make apparent the need to distinguish between semantic relations between word forms and semantic relations between word meanings. Antonymy provides a central organizing principle for the adjectives and adverbs in WordNet, and the complications that arise from the fact that antonymy is a semantic relation between words are better discussed in that context.

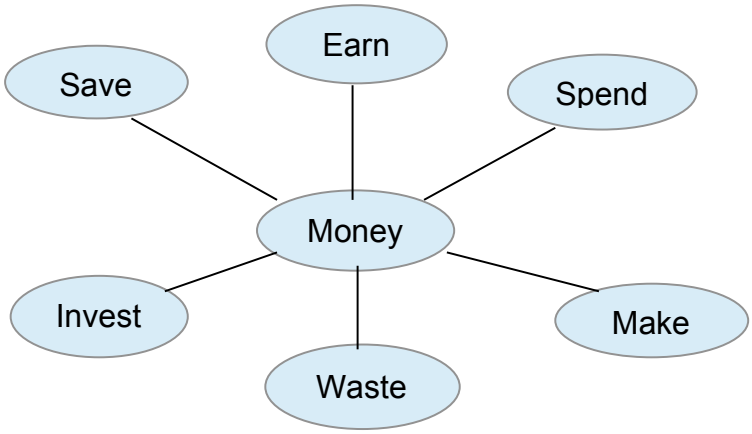

Figure 13: Antonymy Diagram

Table 7: Antonymy table.

Size	**əcəu-əpik, əthabə-əpabə**	Big-small, thick-thin
Quality	**əphə-phəttə nuŋsibə- nuŋsidəbə**	Good-bad, love-hatred
Gender	**nupa-nupi məma- məpa**	Son-daughter, father-mother
State	**əhəubə- əroibə**	Beginning-end
Personality	**ram-rabən debid-goliath**	Rama-Ravana, David-Goliath
Direction	**noŋpok-noŋcup məniŋ-məmaŋ**	East-west, front-behind
Action	**pithok-pisin ləibə-yonbə**	Give-take, buy-sell
Amount	**yambə-yamdəbə ərum-əyaŋ**	Little-much, light-heavy
Place	**ərappə-ənəkpə**	Far-near
Time	**əyuk-nuŋdaŋ əyuk-nuŋthil**	Day-night, morning-evening

3.1.7.3 Hyponymy

Unlike synonymy and antonymy, which are lexical relations between word forms, hyponymy/hypernymy is a semantic relation between word meanings: e.g., {maple} is a hyponym of {tree}, and {tree} is a hyponym of {plant}. Much attention has been devoted to hyponymy/hypernymy (variously called subordination/superordination, subset/superset, or the ISA relation). A concept represented by the synset {*x, x'*, . . .} is said to be a hyponym of the concept represented by the synset {*y, y'*, . . .} if native speakers of English accept sentences constructed from such frames as an x is a (kind of) y. The relation can be represented by including in {*x, x'*, . . .} a pointer to its superordinate, and including in {*y, y'*, . . .} pointers to its hyponyms. Hyponymy is transitive and asymmetrical (Lyons, 1977), and, since there is normally a single superordinate, it generates a hierarchical semantic structure, in which a hyponym is said to be below its superordinate. Such hierarchical representations are widely used in the construction of information retrieval systems, where they are called inheritance systems (Touretzky, 1986): a hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate and from any other hyponyms of that superordinate.

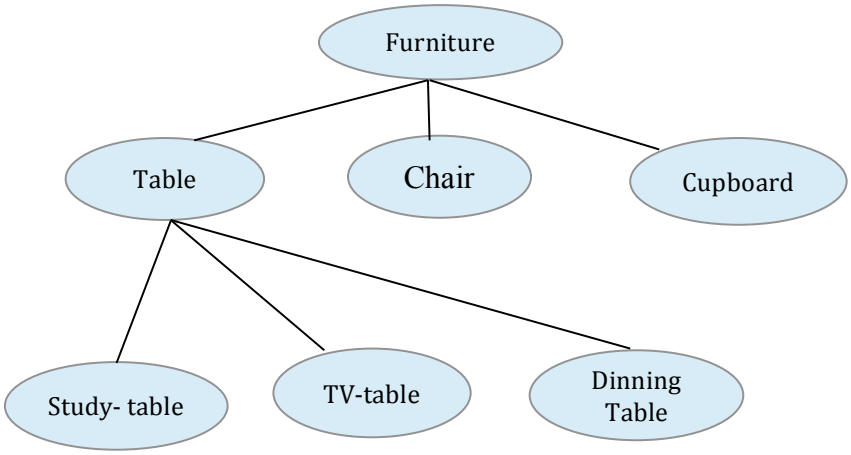

Figure 14: Hyponymy Diagram

3.1.7.4 Meronymy

Synonymy, antonymy, and hyponymy are familiar relations. They apply widely throughout the lexicon and people do not need special training in linguistics in order to appreciate them. Another relation sharing these advantages—a semantic relation—is the part-whole (or HASA) relation, known to lexical semanticists as meronymy/holonymy. A concept represented by the synset {*x, x′, . . .*} is a meronym of a concept represented by the synset {*y, y′, . . .*} if native speakers of English accept sentences constructed from such frames as Ayhasanx (as a part) or Anxis a part of y. The meronymic relation is transitive (with qualifications) and asymmetrical (Cruse, 1986), and can be used to construct a part hierarchy (with some reservations, since a meronym can have many holonyms). It will be assumed that the concept of a part of a whole can be a part of a concept of the whole, although it is recognized that the implications of this assumption deserve more discussion than they will receive here. These and other similar relations serve to organize the mental lexicon. They can be represented in WordNet by parenthetical groupings or by pointers (labeled arcs) from one synset to another. These relations represent associations that form a complex network; knowing where a word is situated in that network is an important part of knowing the word's meaning. It is not profitable to discuss these relations in the abstract, however, because they play different roles in organizing the lexical knowledge associated with different syntactic categories.

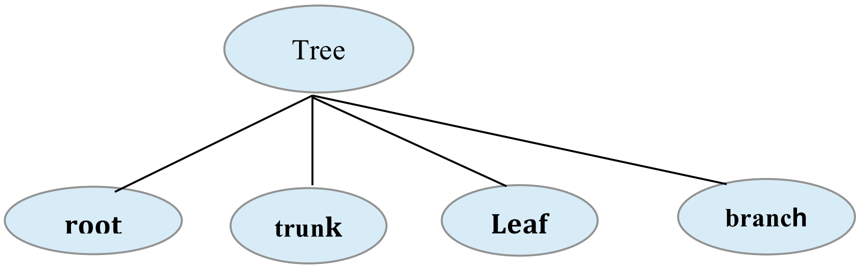

Figure 15: Meronymy Diagram

Table 8: Meronymy Relationship

Compound –object	**kok-həkcaŋ**	Head-body
Member-collection	**pambi-uməŋ**	Tree-jungle
Feature-Activity	**waŋaŋ-thəurəm**	Lecture-ceremony
Place-Area	**dilli-bharət**	Delhi-india
Phase-State	**inkhətləkpə-punsi**	Youth-age
Portion-Mass	**mətum-ləibak**	Lump-clay
Resource-Process	**muktabi-ibə**	Pen-writting
Position-Area	**doktər laiyeŋ-pəthap**	Doctor-medical treatment

3.1.8 Gradation

Gradation is another semantic relation between adjectives that is considered by WordNet is gradation. For some attributes gradation can be expresses by ordered strings of adjectives going from a weak meaning to a strong one. An example of lexicalized gradation for the lightness attribute in Manipuri would be:

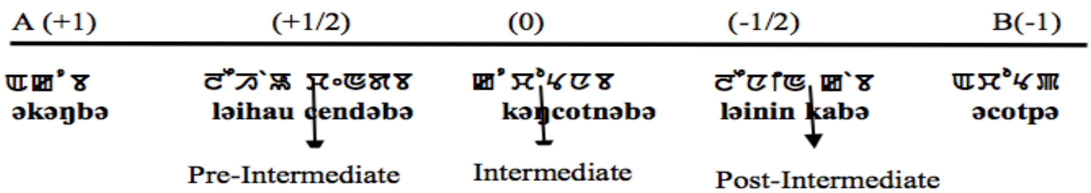

Figure 16: Gradation levels

Table 9: Gradation Relationship

Size	**əcəu-əpik, əthabə-əpabə**	Big-small, thick-thin
Quality	**əphə-phəttə nuŋsibə- nuŋsidəbə**	Good-bad, love-hatred
Gender	**nupa-nupi məma- məpa**	Son-daughter, father-mother
State	**əhəubə- əroibə**	Beginning-end
Personality	**ram-rabən debid-goliath**	Rama-Ravana, David-Goliath
Direction	**noŋpok-noŋcup məniŋ-məmaŋ**	East-west, front-behind
Action	**pithok-pisin ləibə-yonbə**	Give-take, buy-sell
Amount	**yambə-yamdəbə ərum-əyaŋ**	Little-much, light-heavy
Place	**ərappə-ənəkpə**	Far-near
Time	**əyuk-nuŋdaŋ əyuk-nuŋthil**	Day-night, morning-evening

3.1.9 Morphological Relations

An important class of lexical relations is the morphological relations between word forms. Initially, interest was limited to semantic relations; no plans were made to include morphological relations in WordNet. As work progressed, however, it became increasingly obvious that if WordNet were to be of any practical use to anyone, it would have to deal with inflectional morphology. For example, if someone put the computer's cursor on the word trees and clicked a request for information, WordNet should not reply that the word was not in the database. A program was needed to strip off the plural suffix and then to look up tree, which certainly is in the database. This need led to the development of a program for dealing with inflectional morphology.

Verbs are the major problem, of course, since there are four forms and many irregular verbs. But the software has been written and is presently available as part of the interface between the lexical database and the user. In the course of this development it became obvious that programs dealing with derivational morphology would greatly enhance the value of WordNet, but for that more ambitious projects have not yet been undertaken.

Table 10: WordNet Tables in the Database.

tbl_synset_data	tbl_noun_data	tbl_verb_data	tbl_adj_data	tbl_adv_data
Synset_id	Synset_id	Synset_id	Synset_id	Synset_id
Head_word	Hypernyms	Hypernyms	Antonyms	Antonyms
synset	hyponym	troponyms	Similar	Derived form
gloss	meronyms	entailments	Attributes	Gradation
Onto_node_id	antonyms	gradation	Gradation	
Category	Gradation	Causative		
		Compound		
		Conjunct		

3.1.10 Synset Syntax

Strings in the source files that conform to the following syntactic rules are treated as synsets. Note that this is a brief description of the general synset syntax and is not a formal description of the source file format. A formal specification is found in the manual page input (5) of the ‘‘WordNet Reference Manual’’.

1. Each synset begins with a left curly bracket ({).

2. Each synset is terminated with a right curly bracket (}).

3. Each synset contains a list of one or more word forms, each followed by a comma.

4. To code semantic relations, the list of word forms is followed by a list of relational pointers using the following syntax: a word form (optionally preceded by "filename:" to indicate a word form in a different lexicographer file) followed by a comma, followed by a relational pointer symbol.

5. For verb synsets, "frames:" is followed by a comma-separated list of applicable verb frames. The verb frames follow all relational pointers.

6. To code lexical relations, a word form is followed by a list of elements from step 4 and/or step 5 inside square brackets ([...]).

7. To code adjective clusters, each part of a cluster (a head synset, optionally followed by satellite synsets) is separated from other parts of a cluster by a line containing only hyphens. Each entire cluster is enclosed in square brackets.

3.2 IndoWordNet

IndoWordNet is a linked lexical knowledge base of WordNet of 18 scheduled languages of India, viz., Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu and Urdu. Such project indeed took off in 2000 with Hindi WordNet being created by the Natural Language Processing group at the Center for Indian Language Technology (CFILT) in the Computer Science and Engineering Department at IIT Bombay. It was made publicly available in 2006 under GNU license. The Hindi WordNet was created with support from the TDIL project of Ministry of Communication and Information Technology, India and also partially from Ministry of Human Resources Development, India. The figures:(17(a) and (b)) shows below the IndoWordNet home and one search page for Indian languages.

Figure 17(a): IndoWordNet Home page, Source (http://www.cfilt.iitb.ac.in/indowordnet/)

Figure 17(b): Search page of IndoWordNet Source (http://www.cfilt.iitb.ac.in/indowordnet/)

3.3 Advantages and Limitations of WordNet

3.3.1 Advantages of WordNet

1. Development of interlinguistic indices for multilingual conceptual equivalence, without translation. Subsidiarity, this endeavour has also been geared to perfecting integrated access to information, driven by the rapid development of multiple database access systems.

2. Use as an ideal tool to optimise the retrieval capacity of existing systems: natural language interfaces for search engines; automatic generation of tools for semantic disambiguation of concepts (corpora, dictionaries, directories, thesauri) and the creation of knowledge summaries from expanded queries.

3. Support for the design of grammatical categorisations designed to classify information by aspects and traits, but in particular to design and classify semantic ontologies that organise web data semantically, to be sure.

4. Basis for the development of audio-visual and multi-media information retrieval systems.

5. In the last 3 years ontologies construction has been one of the most dynamic areas and its applications to the semantic web.

6. Parameterisable information system i.e. extracting system.
7. Language teaching and translation application.
8. Document classifications.
9. Conceptual identifications.
10. Machine Translation.

3.3.2 Limitation of WordNet

- Although WordNet is an electronic resource, it was, after all, designed for manual consultation and not for automatic processing of natural language texts; as a result, no particular emphasis was placed on enabling the system to automatically differentiate between the various concepts involved.
- Another problem is its multidisciplinarity, which prompts flawed operation in many NLP systems, due to which processing is usually conducted with sublanguages or special records.
- Classification was performed manually, which means that the reasons and depth of classification may not be consistent.
- While the synset simplification affords obvious advantages, in the longer term it leads to shortcomings. These are particularly acute in semantic proximity calculations and may create insuperable situations whenever the context of the discourse in which the relation appears is not contained in the synset information.
- The overabundance of nuance in the concepts calls, in nearly any NLP application, for prior calculation of the frequency of the concept in a given domain. Such calculation is one of the sources of system error, especially where WordNet glosses extracted, as noted above, from the Brown Corpus are used, due to the uneven coverage afforded to the different domains.
- Unlike other dictionaries, Wordnet does not include information about etymology, pronunciation and the forms of irregular verbs and contain only limited information about usage.

- The actual lexicographical and semantic information is maintained in lexicographer files, which are then processed by a tool called grind to produce the distributed database. Both grind and the lexicographer files are freely available in a separate distribution, but modifying and maintaining the database requires expertise.

- Wordenet does not cover special domain vocabulary since it is primarily designed to act as an underlying database for different application, those applications cannot be used in specific domain that are not covered by wordnet.

- The content of Wordnet has not simply been corrected when semantic problem have been encountered. Instead, Wordnet has been used as an inspiration source but heavily re-interpreted and updated whenever suitable. This was the case when, for example, the top level ontology of Wordnet was re-structured according to the onto clean based approach or when Wordnet was used as a primary source for constructing the lower classes the census ontology.

- Wordnet is the most commonly used computational lexicon of English for word sense disambiguation, a task aimed to assign the most appropriate sense to words in context. However, it has been argued that Wordnet encodes sense distinction that is too fine grained even for humans. This issue prevents WSD system from achieving high performance. The granularity issue has been tackled by proposing clustering methods that automatically group together similar sense of the same word.

3.4 WordNet Finding

WordNet presently contains 95,000 different words forms, as 51,500 as simple words, 41,000 as collocation and 70,100 as word meaning or set of synonyms. The table below shows the number of data in WordNet.

Table 11: Number of words in WordNet.

Sl. No.	POS	TOTAL NUMBER OF WORDS
1	NOUN	57, 000(48,800 word meaning)
2	ADJECTIVES	19, 500(10,000 word meaning)
3	VERB	21,000 (13,000 are unique string and 8,400 synset)
4	ADVERB	834 as frequent words

3.4.1 Results of WordNet Data

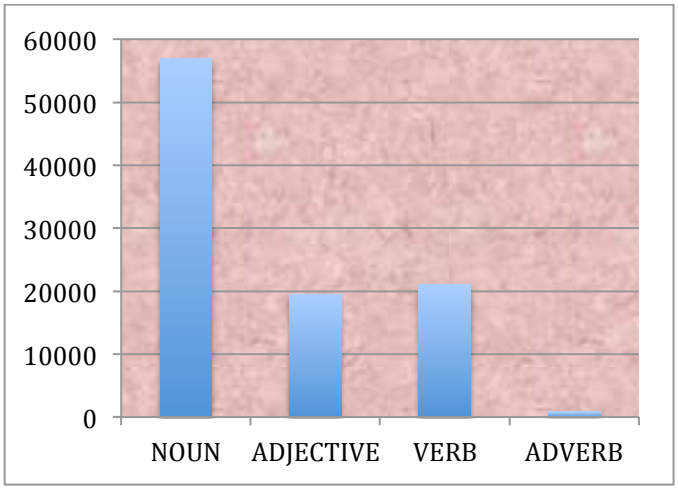

Figure 18: Graphical representation of WordNet Data

Table 12: Number of Words in MMD and WordNet

POS	English	Hindi	Manipuri	WordNet
Noun	57000	145000	57000	145103
Pronoun	39	24	11	Nill
Verb	21000	25884	25884	25884
Adverb	834	834	834	5721
Adjective	19500	31302	31302	31302
Preposition	189	189	189	Nill
Interjection	8	17	7	Nill
Conjunction	199	17	6	Nill

3.4.2 Results of MMD Data

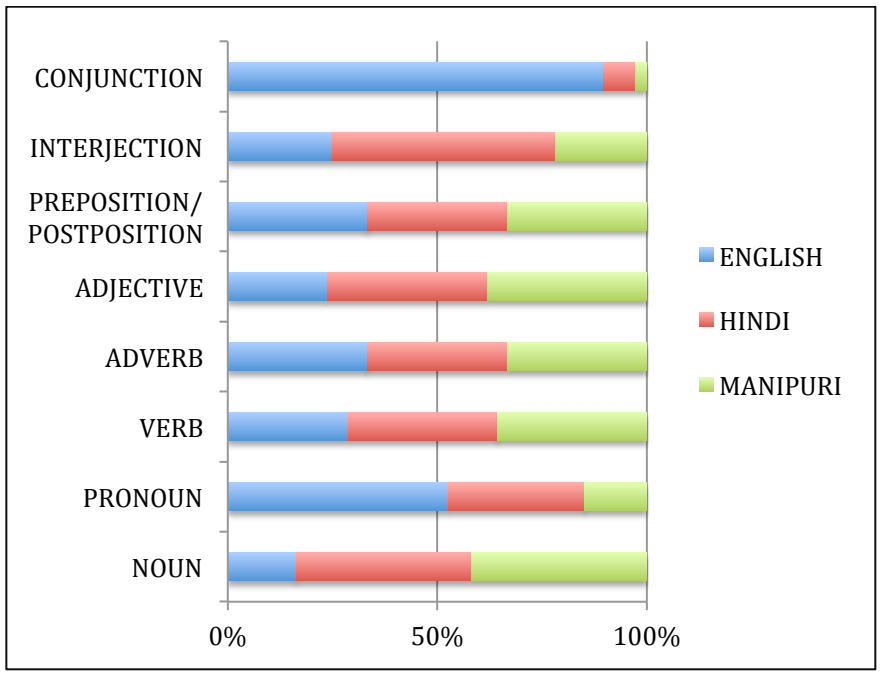

Figure 19: Graphical representation of English, Hindi and Manipuri Data

3.5 Multilingual Dictionary: MMD

Multilingual Manipuri Dictionaries (MMD) is closely related to bilingual dictionaries. In Multilingual Manipuri dictionaries, we look up a word or phrase in English or Hindi language and are presented with the translation in Manipuri language. Multilingual Manipuri Dictionaries can be arranged alphabetically or words can be grouped by topic. i.e. domain based or by other features which are available in WordNet, mainly in Indo WordNet (mainly focus on Indian Languages). It is common for a Multilingual dictionary to be illustrated. In addition to the translation, a multilingual dictionary usually indicates the parts of speech, gender, verb type, declension model and other grammatical clues to help a non-native speaker use the word. Other features sometimes present in bilingual dictionaries are lists of phrases, usage and style guides, verb tables, maps and grammar references. In contrast to the multilingual dictionary, a monolingual dictionary defines words and phrases instead of translating them. In Multilingual Manipuri Dictionary you will get the best features of

WordNet along with addition to the general dictionary features in English, Hindi and Manipuri Language.

3.5.1 MMD Process

According to Atkins & Rundell (2008) the process for building a dictionary is threefold. The three processes for building a dictionary is shown below, the first stage is Analysis stage, second stage is Transfer stage and last stage is Synthesis stage.

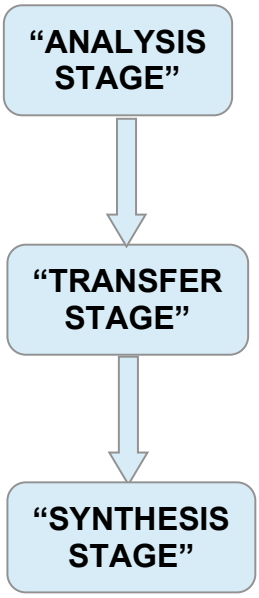

Figure 20: Stages for building Multilingual Dictionary

In building Multilingual Dictionary it will go through under three processes, which have been shown in the above figure, i.e. Analysis Stage, Transfer Stage and Synthesis stage. The exploitation of existing monolingual, bilingual dictionaries, WordNet, mono corpora of such a headword list is the first or Analysis stage.

The API development process of MMD will be in analysis process, designing, development, implementing, and then evaluation and then it will start from analysis again and the figure given below shows the process.

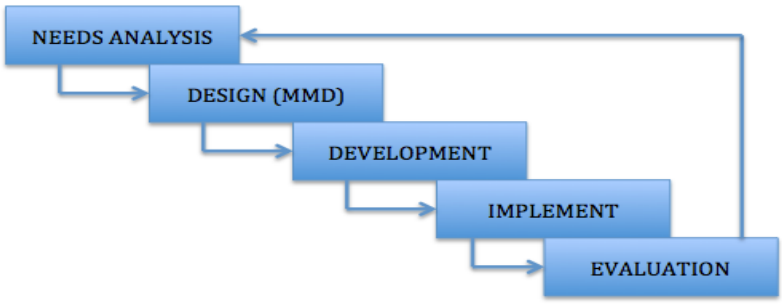

Figure 21: Step by step process for MMD development.

3.5.2 Lexicon of MMD

In linguistics, a lexicon is the vocabulary of a person, group or language. It contains all the minimal grammatical elements of a language. In a sense, it represents a speaker's knowledge of the vocabulary. This is comparable to a dictionary, however, it does not necessarily include the definitions of the entries. Furthermore, it can carry only part of words such as suffixes. The given figure shows how leximes are stored.

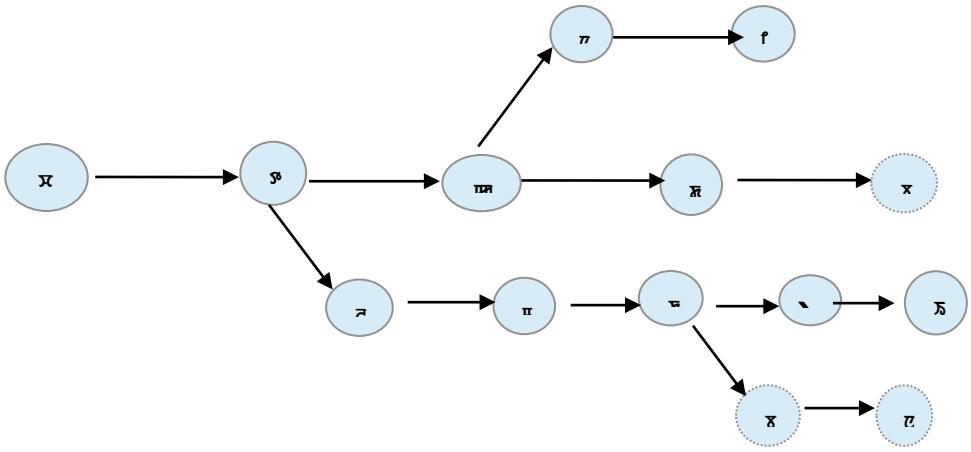

Figure 22: Lexicon of Manipuri words

3.5.3 Process Diagram of MMD with WordNet

WordNet data are extracts from the database and used for corpus, for Corpus we extract the gloss, concept and synset for developing corpus and thus the text in

Manipuri are collected and start for Corpus processing. The given figure no. 22 shows how to extract raw data from WordNet and figure no. 23 shows how the process of MMD database can be developed.

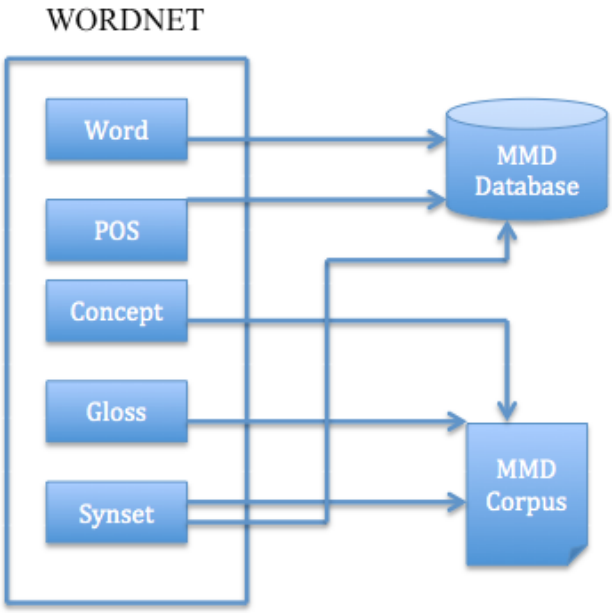

Figure 23(a): Schematic diagram how WordNet data are used for MMD

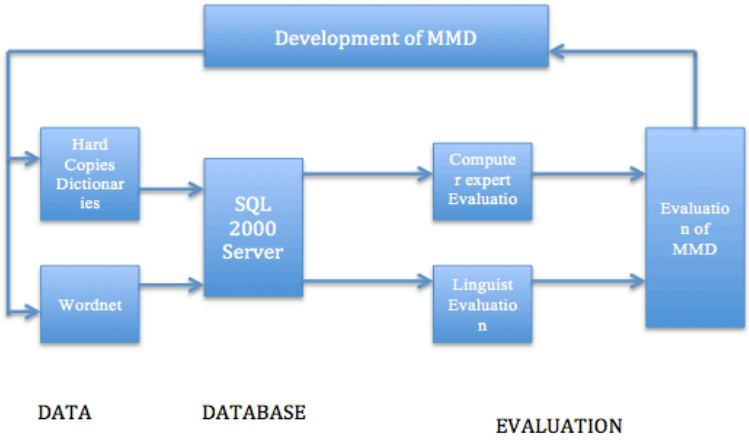

Figure 23(b): Process diagram of WordNet to MMD

3.5.4 Advantages of MMD

"**All literate people own a dictionary**" this is how Nesi, an English professor and author of various work on lexicography, starts the first chapter of her book "the use

and abuse of Learners Dictionaries". According to Philipp Haubmann, chief editor of PONS, a new dictionary is bought every nine years. Unfortunately, it is hard to determine exactly how many dictionaries are acquired, as publisher tends to keep their circulation strength and sales figures top secret (Herbst & klotz, 2003). Now-a-days Internet allows us to obtain information at all times and all over the world, but it does not have only positive aspects. Anonymity, its fast pace, its lack of reliability and structure, its insecurity regarding safety and information might leave a bitter aftertaste. One of the major advantages of online multilingual dictionaries is their updatedness. Information or the entries can be changed, edited.

1. The linguistic server of Multilingual Manipuri Dictionary is dictionary independent and language dependent. (www.morphologic.hu)

2. The MMD dictionary server has intelligent access to various sorts of dictionaries bilingual corpora, monolingual and multilingual corpora.

3. Simultaneously an unlimited number of dictionaries can be held open, thus by a single interrogation step, all the dictionaries (with translations, explanations, synonyms, etc.) can be surveyed.

4. The translators own glossaries built with the help of the system may also be disseminated among other users, if needed.

5. It has an open architecture and a well-defined API.

6. It has been implemented and is available with a gradually increasing number of dictionaries for numerous language pairs.

7. It has reputable Internet equipment included in the Internet Multilingual dictionaries, systems for grammar examining, as well as other processing equipment.

8. Words as well as their meanings are dynamic but the truth is, language by itself, no matter just what the language is, is ever switching, and this is what the uniqueness of language printed dictionaries cannot accommodate, as a result the need for one on the internet.

9. Online English dictionaries are auto-updated; in actual fact here is the primary benefit of on the web dictionaries above printed kinds.

10. An on-line Multilingual dictionary is actually a great tool of data, especially if we know how you can distinguish a fantastic on line dictionary from the bad a single.

11. This may be an essential resource for understanding an international language; here I used English source language so one can translate from this source language easily.

12. One among the very best traits of these Multilingual Manipuri dictionaries would be the translation feature, e.g. an online dictionary can translate an English word into Hindi, Manipuri and vice versa.

13. Dependability on the resource.

14. A web based Multilingual dictionary is updated quickly. It means that every new phrase or terminology might be extra during the building of database of, on the net dictionary.

15. Development of inter-linguistic indices for multilingual conceptual equivalence, with Machine Translation.

16. New ideal tool to optimize the retrieval capacity of existing systems: natural language interfaces for search engines; automatic generation of tools for semantic disambiguation of concepts (corpora, dictionaries, directories, thesauri) and the creation of knowledge summaries from expanded queries.

17. A design, which support grammatical categorizations designed to classify information by aspects and traits, but in particular to design and classify semantic ontologies that organize web data.

18. Language teaching and translation application.

19. Domain classifications of each word.

3.5.5 Disadvantages of MMD

As the multilingual dictionary is based on web enabled some of the possible limitations of MMD are as follow:

1. Promptly reduce the restrictions of printed reference resources.
2. Time consuming for updating from Administrator side.
3. Uncertainty connected with the supply.

4. Server down may be one major problem.

5. Verification and Validation depends on the Administrator.

3.6 Estimate Data for MMD with WordNet

Number of head words found in Manipuri Languages is given in the below table:

Table 13: Estimated Data for MMD with WordNet

Sl. No.	POS	English	Hindi	WordNet	No. of Words.
1	**Noun**	57000	145000	145103	57000
2	**Pronouns**	39	24	Null	13
3	**Verb**	21000	25884	25884	21, 000 +
4	**Adjectives**	834	834	5721	19500
5	**Adverbs**	19500	31302	31302	843
6	**Prepositions/ Postposition**	189	189	Null	189 (English)
7	**Interjections**	8	17	Null	8 (English)
8	**Conjunction**	199	17	Null	17(Hindi)

The total numbers of words entry for MMD from the source of WordNet is 98570(Ninety Eight thousands five hundreds seventy).

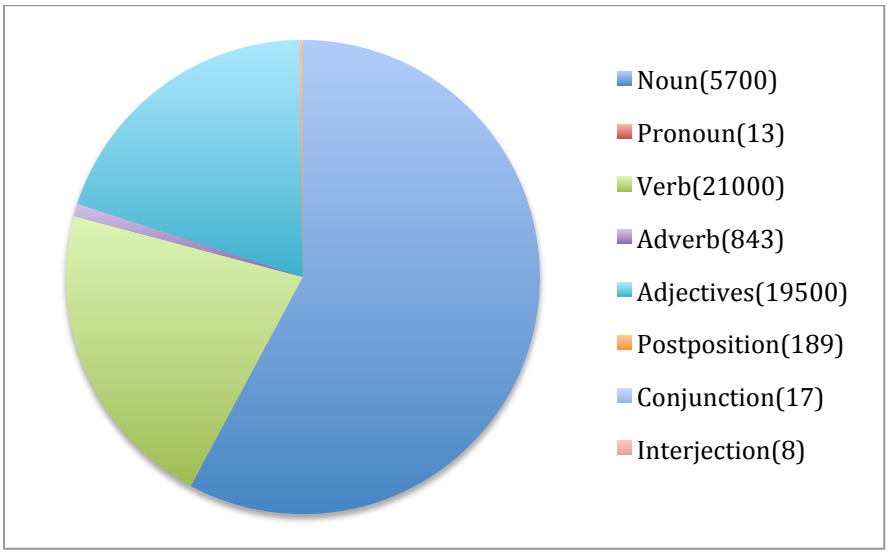

Figure 24:Graphical representation of WordNet Data

3.7 Conclusion

This chapter discussed about what is WordNet and Multilingual Manipuri Dictionary. In this chapter, we compared MMD to WordNet by means of quantitative methods in features and data analysis methods and the features of WordNet, advantages and disadvantages of WordNet. Discussion was about the advantages MMD over WordNet and the additional features of MMD. The data extracted from WordNet for MMD and the corpus text for MMD from WordNet data which is based on development of MMD corpus, is discussed in the next chapter. WordNet has been extensively used in knowledge rich natural language processing system and there is no best computational dictionary for all purposes but MMD will try to overcome the problems.