CHAPTER 4

WORD SENSE DISAMBIGUATION

Human beings can analyze instances of linguistic ambiguity(ies) and overcome them with considerable ease. And human beings can teach a machine to do so. To achieve this objective, we need to describe in detail the process of how humans overcome linguistic ambiguity. This study aims at setting out the techniques to do so. This being the first attempt to enter into this complex subject, we concentrate on ambiguity only at the word level. With this intention we first look at WHAT A WORD IS

1 WHAT IS A WORD?

The concept of a word has been considered to be of central importance in all schools of grammar. The definition of word develops a consistent and coherent approach to central questions about morphology and its relation to syntax. The word ‘word’ relates to three concepts, namely, listemes (linguistic objects permanently stored by the speaker, (concept parallel to lexeme)), morphological objects (objects whose shape can be characterized in morphological terms of affixation and compounding) and syntactic atoms (objects that are not analyzable units with respect to syntax). There are several ways of defining words which are not equivalent as these are based on different views and so theories. We therefore need to examine the several definitions, and to understand the differences among them.

Etymologically, English word is directly taken from Old English word, and has cognates in all branches of Germanic family of languages (Old High German wort, Old Norse orð, Gothic waurd). It is derived from Proto-Germanic *wurđa, continuing a virtual PIE *werdhom. Cognates outside Germanic include Baltic (Old Prussian wīrds "word", and with different ablaut Lithuanian varðas "name", Latvian vārds "word, name") and Latin verbum. The PIE stem *werdh- is also found in Greek. The PIE root is *ūer-, ūrē- "say, speak" also found in Greek. The original meaning of word is utterance, speech, verbal expression. Until Early Modern English, it could more specifically refer to a name or title.

Wittgenstein approached words philosophically, as if they are tools in a toolbox. Words are instruments of language which may have varying uses, according to the purposes for which language may be used. Wittgenstein describes language as a game in which words may be used in a multiplicity of ways, for example, to describe things, to ask questions, to report events, to speculate about events, to make requests, to give commands, to form hypotheses, to solve problems, and to perform other acts of communication.
Attempting a linguistic definition, Leonard Bloomfield\(^4\) said a word is the smallest meaningful unit of speech. This correlates phonemes (units of sound) to lexemes (units of meaning). The five ways to determine where the word boundaries of natural language should be placed include potential pause, phonetic boundaries, indivisibility, semantic unity and pragmatic directives.

The questions often asked are the following:
How many words are there in a language?
Are dog and dogs the same word or different words?

A word, a unit of language, carries meaning (lexeme) and consists of one or more morphemes which are linked more or less tightly together, and has a phonetic value. Typically a word will consist of a root or stem and zero or more affixes (morphological object). Words can be combined to create phrases, clauses, and sentences (syntactic atom). A word consisting of two or more stems joined together forms a compound. A word combined with another word or part of a word forms a portmanteau. Concisely, a sound or a combination of sounds, or its representation in writing or printing that symbolizes and communicates a meaning and may consist of a single morpheme or of a combination of morphemes is called as word.

The term word is defined in different ways considering different levels of analysis. It is usual to distinguish four definitions of word, namely, orthographic word (section 1.1), phonological word (section 1.2), lexical word (section 1.3) and grammatical word (section 1.4). These are related to four levels of analysis respectively. In each such definition word is considered as the unit for analysis of a natural language.

1.1 ORTHOGRAPHIC WORDS

An orthographic word is a written sequence which has a white space at each end but no white space in the middle. Very obviously, orthographic words exist only in written texts, and they have no existence in speech. Word separators (typically space marks) are common in modern orthography of languages using alphabetic scripts. If we consider the sequence ice cream, there are two orthographic words, not one. We can write it as a single word by using hyphen as ice-cream.

The idea of separating words was slow to take hold among the ancients, but eventually some speech communities began to write their languages (Oscan\(^5\), an ancient language of Italy) with dots or strokes to separate the words as spaces do now-a-days. Even today, there are languages that use different criteria for orthography. For example, Vietnamese orthography, although using the Latin alphabet, delimits monosyllabic morphemes, not words. Logographic scripts as for Chinese use single signs (characters) to express a word. German orthography does not separate compound members by space marks as in English.\(^5\)

Usually, the rules of orthography simply do not specify which compounds should be written with a white space and which not, and individual preferences vary considerably. Of course, this freedom to choose where to put white spaces is far
from absolute. In most cases, the rules of orthography dictate where the white
spaces should go, and failure to conform these rules manifest illiteracy.

However, orthographic words are generally of very little linguistic interest in
real world. They are important in learning to write a language. Also, they are
important in natural language processing (NLP).

The concept of orthographic words fails in case of number of languages like
ancient languages, Phoenician, Greek and Latin. For example, if Greek text present
in chart 1 is divided into words in the way that later became conventional in Greek,
it would look like that present in chart 2 below.

| TOMΝΗΜΑΤΟΔΕΠΙ |
| ΟΙΗΣΑΤΟΣΙΔΑΡΙΟΣΠΑΡΜΕΝΟ |
| ΝΤΟΣΥΙΟΣΕΑΥΤΩΙΚΑΙΓΥΝΑ |
| ΙΚΙΚΑΙΥΙΩΠΠΒΙΛΑΗΙ |

Chart 1: Greek Text

| ΤΟ ΜΝΗΜΑ ΤΟΔ ΕΠΟΙΗΣΑΤΟ ΣΙΔΑΡΙΟΣ ΠΑΡΜΕΝΟΝΤ |
| ΟΣ |
| ΥΙΟΣ ΕΑΥΤΩΙ ΚΑΙ Γ ΥΝΑ ΙΚΙ ΚΑΙ ΥΙΩΙ ΠΠΒΙΛΑΗΙ |

Chart 2: Words in above Greek Text

Shorthand Writing is beyond the scope of any discussion here about
orthography.

1.2 PHONOLOGICAL WORDS

A phonological word is a piece of speech which behaves as a unit of
pronunciation according to criteria which vary from language to language. For
example, in English, the most useful criterion is that a phonological word contains
only one main stress.

Consider the following sentence, as one would pronounce it in relaxed
colloquial speech, ‘The rest of the books’I’ll have to go here.’ There are five main
stresses here, falling on the words rest, books, have, go and here. This sentence
therefore contains five phonological words. One can break up the utterance into
phonological words as [the rest] [of the books’I’ll] [have to] [go] [here] or [the rest]
[of the books] [’I’ll have] [to go] [here].
It is clear that not all of these units correspond to units that we might want to recognize as a word for other purposes, for example, of the books’II is certainly not a unit of grammatical structure, but people nevertheless pronounce it as a single phonological word.

In Sanskrit and other languages originated from and influenced by Sanskrit, like, Marathi and Hindi, pause serves as the criteria for delineating phonological word.

Phonological words are important in the study of pronunciation, but they are irrelevant to the study of grammar, syntax.

Phonologically each natural language has a repository of onomatopoeic words which imitate human, animal or other sounds and have due place in the lexicon of the languages.

1.3 LEXICAL ITEMS (LEXICAL WORDS)

A lexical item (or lexeme) is an abstract unit of the lexicon (vocabulary) of a language that has a more or less readily identifiable meaning or function. A lexical word refers to a single concept; it may consist of multiple orthographical, phonological or grammatical words.

A lexical item is a word in the sense in which a dictionary contains words. It is also the sense of word in sentences like ‘I learned twenty new words of French today.’ The abstract lexical item must be represented in speech or writing by one of the possibly several forms it can assume for grammatical purposes. For example, if we want to mention canine animals, we must use either the singular form dog or the plural form dogs. These two grammatical forms both represent the same single abstract unit, the same lexical item. We can conveniently represent that lexical item as DOG. Then dog and dogs are the two possible forms of the lexical item DOG. A dictionary provides entries for lexical items. So, for example, we do not expect to find separate entries in the dictionary for dog and dogs. We expect to find only one entry for the lexical item DOG which is usually represented by the form ‘dog’ or ‘Dog’ in printed dictionaries and by ‘DOG’ in electronic dictionaries.

However, in grammatically more elaborate languages, such as French, Russian, Latin Arabic, Hindi and Marathi a lexical item may have several dozen different forms (usually verbs may have couple of hundred forms).

1.3.1 Citation Forms

We are obliged to choose one or another of the various forms of a lexical item. The form we choose is called the citation form of the lexical item. The citation form (or dictionary form) of a lexical item is the particular grammatical form of it which we use in naming it, talking about it, and entering it in a dictionary. The base form, which carries no grammatical marking, is the natural choice for the citation form. For example, singular form for noun (Dog) is used as its citation form or
infinitive form for verb (Take) or positive form for adjective (BIG but not BIGGER) is usually preferred.

There is a practice in general to say that any inflectional or derivational form of a word is originated from its base form although some linguists like Rajendra Singh do not agree with it, may be for convenience in computation.

A problem may arise with lexical items which are defective. A defective lexical item is one that lacks some of the forms normally exhibited by a lexical item of its class, for example, the nouns without plural form (e.g., the word furniture). If one of the missing forms, in such cases, happens to be the citation form; then one is forced to make a different choice for the same.

1.4 GRAMMATICAL WORD-FORMS

A grammatical word-form (or GWF, or grammatical form) is one of the several forms that may be assumed by a lexical item for grammatical purposes. For example, wake, wakes, woke, woken and waking are grammatical forms of the lexical item WAKE. Sometimes a lexical item like BEWARE exhibits only one grammatical form.

Nevertheless, it may still be useful to distinguish the lexical item BEWARE from its sole grammatical form beware, and likewise for the others.

All the grammatical forms denote grammatical function of the word. Such grammatical forms other than base form of a lexical item are the inflected forms of that lexical item. These are the ones which carry grammatical markings. There is one complication in this regard. The forms like police carry no overt marking though they bear an abstract inflection for plurality or otherwise. Some linguists prefer to identify them as inflected forms.

The number of grammatical forms exhibited by a lexical item may be larger than the number of overtly distinct forms. In other words, for example, the -ing form of an English verb has at least three different grammatical functions as being participle, progressive (continuous) and gerund. Additionally -ing form can be verbal noun, a derivational form. Further, one should not confuse gerund form with verbal nouns with –ing suffixation. Gerund can take an object and verbal noun acts as a noun i.e. it can have plural form.

Here is the point where it is essential to distinguish the process of inflection from the quite different process called ‘derivation’.

1.4.1 Inflection and Derivation

Inflection is the variation in form of a lexical item for grammatical purposes. Derivation is the construction of a new lexical item from another lexical item, usually by the addition or the deletion of an affix (a prefix or a suffix). For example, CAT (noun) and CATTY (adjective).
There can be zero-derivation, or conversion, by which a lexical item is simply shifted from one word-class to another, without adding or deleting any material, e.g., pump as a noun and as a verb. In all these cases, we must regard the zero-derived word as belonging to a different lexical item from the source (PUMP1 & PUMP2), because it belongs to a different word-class.

The direction of derivation is not always obvious (KISS as a noun and as verb). In such cases one has to make an arbitrary decision.

1.5 TYPES OF WORDS

There are various issues regarding recognizing a string of letters as a word. These issues can be well covered by discussing various types of word as follows.

1.5.1 Multi-part and Discontinuous Words

In some of natural languages, certain items exhibit awkward problems from the point of view of word-hood. These are the items which appear to be words by some criteria, but which exist in two or three pieces, sometimes separated by other text material e.g., English Phrasal Verbs like turn on in turn the light on. There is another serious issue of identifying such words. In prepositional verb phrase, call on somebody, there are two possible analyses, namely, call on – somebody or call – on somebody. In such cases one must recognize call on ('visit') considering its meaning. Here, the item on is, syntactically, a part of a prepositional phrase, but, lexically, a part of a multi-part verb. Similarly, the preposition ‘into’ is expressed in Mandarin Chinese by the discontinuous item dào...li (dào guòn li).

1.5.2. Content Words and Grammatical Words

Lexical items are commonly divided into content words and grammatical words.

A content word (or full word, in the Chinese terminology) is a lexical item which has semantic content. It has a readily identifiable meaning. A grammatical word (or empty word) has little or no identifiable meaning, but has one or more grammatical functions. Here meaning refers to lexical item’s relation to a concept.

A content word can be defined / glossed and it can often be translated into another language with some ease. Thus, we can say that Marathi ‘ghara’ is equivalent to Hindi ‘ghara’ and English ‘house’. It is worth noting that, because of lexical differences among languages, not all content words can be translated quite straightforwardly.

A grammatical word cannot be defined; only an account of its grammatical functions can be given. The grammatical words may not have definable equivalents in another language. For example, in equivalent phrases, ‘chahaacha bhaNDa’
(Marathi, M) / ‘chaayakaa bartana’ (Hindi, H) / tea pot (English, E), there is no equivalent of ‘cha’ (M) / ‘kaa’ (H) in English in such translations.

Some words appear to be on the border-line in this classification. They have clear grammatical functions with some identifiable semantic content, such as pronouns. The best policy is to classify such items as grammatical words, since their grammatical functions are usually much more important than any traces of lexical meaning they may possess.

1.5.3 Clitics

A clitic is an item which represents a lexical item and a grammatical word form. (Word form is the phonological sound or orthographic appearance of a word that can be used to describe or identify something.) Since a clitic (e.g., a(n), the and ’ll) cannot form a phonological word by itself, it must be phonologically bound to a more substantial item, its host, with which it forms a phonological word, possibly together with other clitics. A clitic which precedes its host, like the English articles is called as a proclitic. A clitic which follows its host, like English ‘ll, is an enclitic.

In Marathi, the ‘ya’ ending of a verb is considered as a clitic. It does not have a status of phonological word and is not yet accepted as a verb suffix. It is a contracted form of ‘aahe’ / ‘hai’ / ‘is or are’. It is bound to a preceding host; it is enclitic. For example, ‘aalee aahe’ is usually written and more frequently spoken as ‘aaleeaa’ / ‘aayaa hai’ / ‘have come’. In English, clitic-article is bound to a following host, as in a book and the clitic-auxiliary is bound to a preceding host, as in John’ll do it.

From a historical point of view, a clitic is perhaps to be seen as an item which has lost its status as an autonomous word, but which has not yet been reduced to the status of an affix, a prefix or a suffix.

1.5.4 Short Forms

Natural languages exhibit a number of short forms of several kinds and these short forms merit some discussion from the point of view of word-hood.

Among non-linguists, there is a habit of labeling all or most of these short forms as “abbreviations”, but this practice is misconceived. It is possible to distinguish true abbreviations from the other cases.

1.5.4.1 Abbreviations

An abbreviation is a short way of writing a word or a phrase, using only letters of the alphabet and possibly full stops. An abbreviation is strictly a written form. It has no pronunciation of its own, and it can only be pronounced by pronouncing the full form which it abbreviates or in some cases, by spelling it out letter by letter, for example, Dr. for Doctor (only as titles accompanying names)
and etc. for and other things. The exceptional forms Mrs. and Ms. are treated as abbreviations, even though no longer forms exist.

An abbreviation usually qualifies as an orthographic word, but it is not a word in any other sense. With a few marginal and idiosyncratic exceptions, an abbreviation is not a part of the spoken language at all.

1.5.4.2 Logograms

A logogram is a written character which is not a letter of the alphabet but which conventionally represents a word, or rarely a sequence of words e.g., ‘5’ for five or ‘%’ for per cent. A logogram is a representation of a lexical item or of a grammatical word-form. It fits the definition of an orthographic word given above, but some linguists prefer to restrict that definition to cases consisting of letters of the alphabet, in which case a logogram would not be an orthographic word.

1.5.4.3 Contractions

A contraction is a conventional brief ways of pronouncing a sequence of two (or rarely three) words that often occur together. A contraction always has a distinct written form e.g., I’m for I am. Some contractions have unexpected forms, such as won’t for will not. A contraction is always a single orthographic word and a single phonological word. But it represents two (or three) lexical items, and two (or three) grammatical word-forms.

1.5.4.4 Acronyms and Initialisms

There is a practice in many languages to coin new words by extracting the initial letters of the most important words in a phrase; and resulting sequence of letters together to form the new word which almost always has the same meaning as the original phrase e.g., BBC for British Broadcasting Corporation in English. In Marathi / Hindi, ‘manapaa’ stands for ‘mahaa nagara paalikaa’ which means municipal corporation; it is used in the phrases like, ‘manapaachii basa’ / manapaa kii basa’ / bus run by municipal corporation.

We may distinguish two kinds of outcome here. Sometimes the resulting form can only be pronounced by spelling it out letter by letter, as with BBC or ‘manapaa’. A formation of this kind is initialism in British tradition. In other cases the resulting form can be pronounced like an ordinary word, as with NATO. A formation of this kind is an acronym in British tradition. However, in the American tradition, the label “acronyms” is applied to all such formations without distinction.

There are usually two ways to write such forms. The first is to write all letters in capital (NATO) and the second is to write first letter capital and others small (Nato). But not all acronyms are written with capitals. Some acronyms have become perfectly ordinary lexical items, and they behave accordingly. For example, the phrase, light amplification by the stimulated emission of radiation has given rise to the acronym, laser which is an ordinary lexical items, entirely unremarkable
apart from its origin. The canonical pattern of forming acronyms, like, laser ignores the small grammatical words in constructing them. But several acronyms are deliberately constructed in a non-canonical manner, in order to obtain a result which can be easily pronounced. A good example is radar from radio detection and ranging. It is notable here that in case of languages like Marathi and Hindi, as there is no system of capitalizing alphabets, such issues are irrelevant.

An acronym does not differ from any other lexical item, except perhaps in its unusual written form. Initialism has both an unusual written form and an unusual pronunciation, but otherwise it is an ordinary lexical item grammatically.

Initialisms are sometimes confused with abbreviations, but they are not abbreviations. First, initialism always has its own pronunciation, distinct from the pronunciation of the longer form which it represents. Abbreviations do not usually have their own pronunciations. Second, initialism, being a lexical item, can appear in a structural position in a sentence, in which a lexical item is appropriate, e.g., plural forms like BBCs. No abbreviation can behave like this.

1.5.4.5 Clipped Forms

A clipped form is an item which is obtained by extracting a piece from a longer word or phrase, e.g., gym from gymnasium. The process of extraction is clipping. In all but very rare cases, a clipped form has the same meaning as the longer form from which it is obtained. Even discontinuous pieces can be clipped if convenient, e.g., biopic for biographical picture.

A clipped form is not an abbreviation. It is a genuine lexical item, just like any other lexical item, and it is unusual in no way apart from its origin. A clipped form accepts the grammatical inflections which are typical of its word class, e.g., plural form of noun as gyms. A clipped form can enter into compounds like any other lexical item, e.g., gym-shoes. In some cases, the longer form would be abnormal or impossible in the same position, e.g., the longer word, gymnasium shoes, is not used. Clipped forms are entered in dictionaries like other lexical items, and they are legal in Scrabble, a popular game of forming words.

Sometimes a clipped form displaces its original longer form as piano has completely supplanted its source, pianoforte so that many people do not even know the longer forms.

One should not confuse between abbreviation (as in Prof. Chomsky) and clip (as in an appointment with my prof.)

The eminent query often made is do all languages have words. Our intuitions suggest that a word is a unit which is much smaller than a sentence, and that a sentence typically consists of a sequence of words. But this account does not hold straightforwardly for all languages. In some languages, it can be difficult to draw a distinction between sentences and words. Such languages are known as polysynthetic languages. Marathi is a synthetic language, (its origin, Sanskrit is
polysynthetic), as there exist (orthographic, phonological and grammatical) word-like adverb phrases in Marathi, e.g., ‘jhaaDaawaruun → jhaaD-aa-war-uun’ which means from the top of the tree / from the tree / (from) above the tree. This happens as even postpositions (e.g., ‘wara’) are attached to the stem like case markers do. In Hindi it is different; it is not polysynthetic / synthetic. Even case markers are written disjoint from the stem in Hindi though it is originated from and influenced by Sanskrit to a great extent. Still Marathi is not as polysynthetic as language, Yup’ik, an Eskimo language of Alaska.

The written form of the Yup’ik language sentence, ‘kaipiallrulliniuk’ (The two of them were apparently really hungry) somehow represents the ordinary pronunciation of this sentence. But, at a somewhat more abstract level, we can analyze this sentence into a sequence of structural units, or morphemes as ‘kaig- ‘be hungry’ --- piar- ‘really’ + llru- PAST --- llini- ‘apparently’ + -u- INDICATIVE --- -k ‘they two’. (The string following ‘---’ is the stem (otherwise) and that following +, is grammatical morpheme.) Thus the sentence consists of a verb stem kaig- ‘be hungry’ followed by a string of suffixes. In effect, the whole sentence is merely a grammatical form of this verb. One can notice that in the sentence of this kind, the forms of some of these morphemes are altered when they are joined in Sequence. In a sentence of Yup’ik language, there is nothing that we can call a word, except for the sentence itself. Such sentences (comparable with orthographic word in ordinary sense) can be formidably as long as ayaqaququaryuuumitqapiallruryugnarquq-qaa (I guess she probably didn’t really want to go for those short little trips, did she?)

Some sentences of Yup’ik language consist of shorter units which can be reasonably called as words, say, Maurluqa ayunek pitllallruuq waten amllervkenaki qillertaqluki enemuun agartaqluki (‘My grandmother used to pick Labrador tea leaves, just a few like this, and tie them together and hang them inside the house). The words here may be glossed as follows. [Maurluqa - my grandmother, ayunek - Labrador tea leaves, pitllallruuq - used to pick, waten - like this, amllervkenaki - just a few’, qillertaqluki - and tie them together, enemuun - inside the house, agartaqluki - and hang them]. So, Yup’ik does have words after all, though you can see that even the shortest Yup’ik words appear to be more complex than most words of non-polysynthetic languages.

In short, words can be orthographical words, phonological words, lexical words and grammatical words. All these may not match in size and shape. We mean measure of size in terms of number of characters / letters within a word. The shape of word can be continuous or discontinuous / broken as seen in section 1.5.1.

While undertaking any natural language processing, machine can read only orthographical words, strings of letters separated by spaces and without a space in between. The natural language processing task of machine translation necessarily transfers meaning of a text from the source language to the target language through words. The dynamic nature of a word directs us to the issue of establishing correspondence between a word, a unit of linguistic analysis and the unit of meaning.
This is the point where we should know about the concept of *concept*.

## 2 WHAT IS A CONCEPT?

A concept is the basic element of thought serving as a unit of material storage of information. (Information is stored in neurons of human beings or electrons in machines.)

A concept is an idea that brings diverse elements into a basic relationship. It is an abstract, universal idea, notion or entity that serves to designate a category or class of entities, events or relations. Concepts are abstract in that they omit the differences of the things in their extension, treating them as if they were identical. They are universal in that they apply equally to each thing in their extension. Concepts are also the basic elements of propositions, much the same way a word is the basic semantic element of a sentence.

A concept is a stable mental representation of objects, classes, properties, and relations. When we encounter a new object or event for the first time, we draw upon our mental store of concepts in order to identify it. One of the most important parts of the human learning process is concept-formation, where, after a number of distinct experiences of the ‘same’ object or event, we acquire, by a process of induction, a concept for it. Thus, all the concepts in memory are interrelated; they form a web, a net.

Concepts are bearers of meaning, as opposed to agents of meaning. Concepts do not create meaning; they just carry meaning. A single concept can be expressed by any number of languages. The concept of *cat* can be expressed as ‘billi’ in Hindi, as ‘maanjara’ in Marathi, and ‘maarjaara’ in Sanskrit in as many ways as number of natural languages is there. The fact that concepts are in some sense independent of language makes the translation possible. There can be synonyms (having identical meaning) across the various natural languages, because they express one and the same concept.

The distinction between *concept* and *object* is due to the German philosopher Gottlob Frege. He differentiates concept and object by drawing an analogy with mathematical expressions. In a mathematical expression, the sign of the argument and the expression of the function are dissimilar (e.g., ‘x’ is different from the function on it, say, x + 1). He held that the referent of a predicate is a concept. Predicates have a semantic role in determining the truth value of sentences in which they occur.

According to Gottlob Frege, any sentence that expresses a singular proposition consists of an expression (a proper name or a general term plus the definite article) that signifies an object together with a predicate (the copula "is", plus a general term accompanied by the indefinite article or an adjective) that signifies (*bedeuten*) a Concept.
Thus, "Socrates is a philosopher" consists of "Socrates", which signifies the Object *Socrates*, and "is a philosopher", which signifies the concept of *being a philosopher*. Any expression that purports to signify a concept (e.g., the concept *horse*) grammatically signifies an object. The object is an instantiation of the concept. A concept is distinguished categorically from objects. The set of attributes in each concept can be interpreted as a set of singly necessary and jointly sufficient conditions for defining the set of objects in the concept, represented by the concept.

3 THE THEORETICAL BACKGROUND OF WORD SENSE DISAMBIGUATION (WSD)

From the above discussion, we can conclude that whatever may be the approach towards the definition of a *word*, any concept can be expressed either by a word or a group of words in every natural language (so, translation is possible). We can say further that a concept provides a room for to describe an object, its instantiation, in terms of the defining features of such objects that can be instantiation of the concept under consideration. When a word or a group of words refers to an object, it means it refers to an instantiation of a concept. The reference of a word or a group of words to an object and the way in which it is referred (sense) together constitutes the meaning of that word or that group of word. When we refer to the meaning of the lexical word (item), may be formed of one or more orthographic words, we mean the content of that lexical word. Content of the lexical word refers to a concept, a unit of meaning or a unit of information involved in thought process.

An object, an instantiation of a concept is often referred by several words in a language with different senses (e.g., Mahatma Gandhi and M. K. Gandhi). Even a word or a group of words; that constitutes a lexical item (word) can refer to different concepts, units of meaning. This linguistic phenomenon is known as linguistic ambiguity (section). The possibility of having multiple meanings of a lexical item in monolingual situation is referred as polysemy.

Machine translation, a linguistic transaction in multilingual environment, involves the handling of transfer of meaning of any linguistic unit from the source language into the target language. The transfer of meaning stands for a process of finding a synonymous expression in the target language for an expression in the source language. In other words, the transfer of meaning is expressing the meaning of a linguistic unit; say a lexical word, in the source language in terms of its synonym in the target language. These synonyms are popularly known as *translation equivalents*. The instances of translation equivalences cover Multi-part and Discontinuous / Broken Words, Content Words & Grammatical Words, Clitics and Short Forms (Abbreviations, Logograms, Contractions, Acronyms and Initialisms, Clipped Forms), along with regular words. All these should be stored as lexical items irrespective of their size and shape. We mean measure of size of a lexical item in terms of number of characters / letters and / or number of orthographic or phonological words within that lexical item. The shape of such lexical item can be continuous or discontinuous / broken (section 1.5.1 & 1.5.4.5).
Thus, in the situation of machine translation, a (lexical) word in the source language will be identified as ambiguous if there is a possibility of finding more than one (lexical) word in the target language as its translation (concluding comments of chapter 3).

Like all other natural language processing tasks, machine translation also faces the bottle-neck problem of resolving linguistic ambiguity. Before proceeding to actual techniques of disambiguation we will take a brief view of what is word sense disambiguation in terms of its definition, the salient efforts taken so far, along with the possible ways of evaluation of the word sense disambiguation task and some resources developed for the same.

3.1 DEFINITION OF WORD SENSE DISAMBIGUATION

Word sense disambiguation plays an important role in almost all areas of natural language processing, like, machine translation, information retrieval, sense analysis, and speech recognition, message understanding, man-machine communication, hypertext navigation, content and thematic analysis, grammatical analysis, speech processing, text processing, and question-answering system. Research on word sense disambiguation has great theoretical and practical significance.

Word sense disambiguation assumes word senses. In Computational Linguistics, Word Sense Disambiguation, popularly called as WSD, is the identification of the intended sense of a word-token considering its usage in a bigger linguistic expression, i.e., a sentence. Within the lexicography and linguistics literature, word senses are known to be very slippery entities.

Word sense disambiguation is the problem of determining in which sense a word having a number of distinct senses is used in a given sentence. For example, the two distinct senses of the word bass are a type of fish (sense 1) and the tones of low frequency (sense 2). Consider the sentences ‘The bass part of the song is very moving’ (sentence 1) and ‘I went fishing for some sea bass’ (sentence 2). To a human it is obvious that the first sentence is using the word bass in sense 2 above, and in the second sentence it is being used in sense 1. But although this seems obvious to a human, developing algorithms to replicate this human ability is a difficult task.

Thus, the prominent problem with word sense disambiguation is making decision about the appropriate sense among others. In cases like the word bass above, at least some senses are obviously different. In other cases, however, the different senses can be closely related (one meaning being a metaphorical or metonymic extension of another), and there division of words into senses becomes much more difficult.

Close correlation has been found between lexical meaning and its distribution. According to a study in the field of cognitive science, people often
disambiguates word sense, accounting only a few other words in a given context, frequently & usually, one additional word.

The relation between the given word and the other words in its context can be effectively used for word sense disambiguation. The context usually provides the necessary information. The first step of the task of developing word sense disambiguation model is to fix the sort of useful contextual information and the size of the context required for the same.

A multi-level framework of syntagmatic analysis can be designed to describe the syntactic and semantic constraints of the given word. During the survey of 5793 words, it was found that different senses have different and complementary distributions at the syntax and / or collocation levels. This serves as a foundation for establishing word sense disambiguation model by using grammatical information and a thesaurus provided by the linguists.

In short, the proper context is the basic tool for working on word sense disambiguation; irrespective of whether it is done by humans or by machines.

The automatic disambiguation of word senses has been an interest and concern since the earliest days of computer treatment of language in 1950’s. Sense disambiguation is considered to be an “intermediate task” as it is a very necessary task for getting good results out of any task related to natural language processing. The crucial question here is; ‘is word sense tagging much more than part-of-speech tagging’.

It is instructive to compare the word sense disambiguation problem with the problem of part-of-speech tagging. The algorithms used for one do not tend to work well for the other, mainly because the part of speech of a word is primarily determined by the immediately adjacent 1-3 words, whereas the sense of a word may be determined by words a fair way further away. The success rate for part of speech tagging algorithms is at present much higher than that for word sense disambiguation, state-of-the art being around 95% accuracy or better, as compared to less than 75% accuracy in word sense disambiguation with supervised learning. Figures are typical for English, and may be very different from those for other languages.

The problem of word sense disambiguation has been described as AI-complete, that is, a problem which can be solved only by first resolving all the difficult problems in artificial intelligence (AI), such as the representation of common sense and encyclopedic knowledge. This facilitates to catch the exact word sense in context.

Word sense disambiguation assumes word sense and hence necessitates the study of the meaning of different linguistic entities like words, phrases, sentences, texts and compounds. The de-compositional perspective towards meaning holds that the meaning of words can be analyzed by defining meaning atoms or primitives, which establish a language of thought. Traditionally, semantics has
included the study of connotative sense and denotative reference, truth conditions, argument structure, thematic roles, discourse analysis, and the linkage of all of these to syntax.

The study of the meaning of words encompasses the study of the relations between the words such as homonymy, synonymy, antonymy, polysemy, paronymy, hypernymy, hyponymy, meronymy, metonymy, holonymy. It also includes the study of the exocentric, and endocentric relations between different linguistic expressions. Further, it is the study of thematic roles, argument structure, and it’s linkage to syntax. It deals with sense and reference, truth conditions and discourse analysis.

The efforts for developing models for word sense disambiguation had been and are being made considering such theoretical aspects of meaning of linguistic entities, especially, words.

The survey of various efforts made so far to develop word sense disambiguation methods will guide to know the present state of art. A few of such efforts are described below.

Additionally, the comparison between world-wide views and the approach adapted in this research will be supported by the information. (In India, the work on word sense disambiguation is in infant state.) Word sense disambiguation methods are broadly divided into early methods, AI-based methods such as symbolic or frame-based methods, knowledge based methods and corpus based methods. The discussion will conclude in possibility of evaluation of word sense disambiguation task.

3.2 SURVEYS OF WORD SENSE DISAMBIGUATION METHODS

In general terms, word sense disambiguation involves the association of a given word in a text or discourse with a definition or meaning (sense) which is distinguishable from other meanings potentially attributable to that word. The task therefore necessarily involves two steps. The first is the determination of all the different senses for every word relevant (at least) to the text or discourse under consideration. The second, but not less important, is to assign each occurrence of a word to the appropriate sense among others. Much recent work on word sense disambiguation relies on pre-defined senses for the first step, including, a list of senses, a group of features, categories, or associated words (e.g., synonyms, as in a thesaurus), an entry in a transfer dictionary which includes translations in another language, etc. The second step of the assignment of a correct sense to words necessitates the context, external knowledge sources, information from an external knowledge source (knowledge-driven WSD), or information about the contexts of previously disambiguated instances of the word derived from corpora (data-driven or corpus-based WSD).

However, since the earliest days of word sense disambiguation work there has been general agreement that the problems of morpho-syntactic disambiguation
and sense disambiguation can be disentangled\(^\text{15}\). That is, for homographs with different parts of speech (e.g., *play* as a verb and as a noun), morpho-syntactic disambiguation accomplishes sense disambiguation, and therefore (especially since the development of reliable part-of-speech taggers), word sense disambiguation work has since focused largely on distinguishing senses among homographs belonging to the same syntactic category.

### 3.2.1 Early Word Sense Disambiguation Work in Machine Translation

The first attempts at automated sense disambiguation were made in the context of machine translation. Weaver\(^\text{16}\) discusses the need for word sense disambiguation in machine translation and outlines the basis of an approach to it (word sense disambiguation) which underlies all subsequent work on the topic. Kaplan\(^\text{17}\) observed that sense resolution given two words on either side of the word was not significantly better or worse than when given the entire sentence.

The complexity of the context, and in particular, the role of syntactic relations, was also recognized. Weaver’s *Memorandum\(^\text{16}\)* discusses the role of the domain in sense-disambiguation, making a point that was reiterated several decades later.

The need for knowledge-representation for word sense disambiguation was also acknowledged from the outset. The need of undertaking statistical semantic studies as a necessary primary step was then felt. Estimations of the degree of polysemy in texts and dictionaries were made.

Based on approaches toward word sense disambiguation, Pimsleur\(^\text{19}\) introduced the notion of levels of depth for a translation. The first level uses the most frequent equivalent (e.g. Marathi & Hindi ‘bhaarii’ and English heavy), producing a text where 80% of the words are correctly translated and the second level distinguishes additional meanings (e.g., Marathi & Hindi ‘bhaarii’ and English costly), producing a 90% correct translation. (The examples are purposely given which refer triplet of Marathi, Hindi and English)

The striking fact about this early work on word sense disambiguation is the degree to which the fundamental problems and approaches to the problem were foreseen and developed at that time.

### 3.2.2 AI-based Methods

AI methods began to flourish in the early 1960’s and began to attack the problem of language understanding. As a result, WSD in AI work was typically accomplished in the context of larger systems intended for full language understanding. AI methods are modeled after some theory of human language understanding. These methods include symbolic methods, frame based methods, graph based methods, preference semantics based methods etc. as narrated below.
3.2.2.1 Symbolic Methods

The semantic networks were developed in the late 1950’s and were immediately applied to the problem of representing word meanings. Masterman\textsuperscript{20} working in the area of machine translation, used a semantic network to derive the representation of sentences in an interlingua comprised of fundamental language concepts; sense distinctions are implicitly made by choosing representations that reflect groups of closely related nodes in the network.

3.2.2.1 Frame-based Methods

Subsequent AI-based approaches exploited the use of frames which contained information about words and their roles and relations to other words in individual sentences. Masterman\textsuperscript{20} introduces “polaroid words,” a mechanism which progressively eliminates inappropriate senses based on syntactic evidence provided by the parser, together with semantic relations found in the frame network.

3.2.2.3 Graph-based Methods

Knowledge representation by using graphs was tried by researchers, such as Pierce’s “existential graphs”\textsuperscript{22} and the graphs of the psychologist Selz\textsuperscript{23} which represent patterns of concepts and inheritance of properties.

3.2.2.4 Preference Semantics-based Methods

Preference semantics specifies selectional restrictions for combinations of lexical items in a sentence that can be relaxed when a word with the preferred restrictions does not appear, thus enabling, especially, the handling of metaphor (as in my car drinks gasoline, where the restrictions on drink prefer an animate subject but allow an inanimate one).

3.2.2.5 Word Expert Parser

A rather different approach to language understanding which contains a substantial sense discrimination component is the Word Expert Parser\textsuperscript{24}. The approach derives from the somewhat unconventional theory that human knowledge about language is organized primarily as knowledge about words rather than rules. Their system models what they feel is the human language understanding process, a coordination of information exchange among word experts about syntax and semantics as each determines its involvement in the environment under question. Each expert contains discrimination net for all senses of the word, which is traversed on the basis of information supplied by the context and other word experts, ultimately arriving at a unique sense which is then added to a semantic representation of the sentence. The well-known drawback of the system is that the word experts need to be extremely large and complex to accomplish the goal, which is admittedly greater than sense disambiguation.
3.2.2.6 Information-based Methods

Dahlgren’s\textsuperscript{25} language understanding system includes a sense disambiguation component which uses a variety of types of information such as fixed phrases, syntactic information (primarily, selectional restrictions) and commonsense reasoning. Reasoning often involves traversing ontology to find common ancestors for words in context.

Dahlgren’s\textsuperscript{25} work anticipates Resnik’s\textsuperscript{26,27,28} results by determining that ontological similarity, involving a common ancestor in the ontology, is a powerful disambiguator. Dahlgren\textsuperscript{25} also notices that verb selectional restrictions are an important source of disambiguation information especially for nouns.

3.2.2.7 Connectionist Methods

Work in psycholinguistics in the 1960’s and 70’s established that semantic priming, a process in which the introduction of a certain concept influence and facilitate the processing of subsequently introduced concepts that are semantically related, plays a role in disambiguation by humans\textsuperscript{29}. This leads to several connectionists’ approaches.

Waltz & Pollack\textsuperscript{30} and Bookman\textsuperscript{31} hand-encode sets of semantic “microfeatures,” corresponding to fundamental semantic distinctions (animate/inanimate, edible/inedible, threatening/safe, etc.), characteristic duration of events (second, minute, hour, day, etc.), locations (city, country, continent, etc.), and other similar distinctions, in their networks.

In addition to these local models (i.e., models in which one node corresponds to a single concept), distributed models have also been proposed\textsuperscript{32}. However, whereas local models can be constructed \textit{a priori}, distributed models require a learning phase using disambiguated examples, which limits their practicality.

Many of the AI-based disambiguation results involve highly ambiguous words and fine sense distinctions (e.g., \textit{ask}, \textit{idea}, \textit{hand}, \textit{move}, \textit{use}, \textit{work}, etc.) and unlikely test sentences (\textit{The astronomer married the star}), which make the results even less easy to evaluate in the light of the now-known difficulties of discriminating even gross sense distinctions.

3.2.3 Knowledge-based Methods

The AI-based work of the 1970’s and 80’s was theoretically interesting but not at all practical for language understanding in any but extremely limited domains. A significant roadblock to generalizing word sense disambiguation work was the difficulty and cost of hand-crafting the enormous amounts of knowledge required for word sense disambiguation, the so-called “knowledge acquisition bottleneck”\textsuperscript{33}. Several knowledge bases are used as given below.
3.2.3.1 Machine-readable Dictionaries

Machine-readable dictionaries (MRDs) became a popular source of knowledge for language processing tasks following theses written by Amsler’s. The attempts to automatically extract lexical and semantic knowledge bases from MRDs are made by Michiels, Calzolari, Chodorow, Markowitz, Byrd, Nakamura and Nagao, Klavans, Wilks, etc.

All of these methods (and many of those cited elsewhere in this thesis) rely on the notion that the most plausible sense to assign to multiple co-occurring words is the one that maximizes the relatedness among the chosen senses. This metric is also used with the help of a vector method that relates each word and its context.

3.2.3.1.1 The Box Codes and the Subject Codes

The box codes and the subject codes are provided for each sense in machine readable dictionaries. Box codes include primitives such as ABSTRACT, ANIMATE, HUMAN, etc. and encode type restrictions on nouns and adjectives and on the arguments of verbs. Subject codes use another set of primitives to classify senses of words by subject (ECONOMICS, ENGINEERING, etc.).

The use of LDOCE (Longman Dictionary of Contemporary English) box codes, however, is problematic as the codes are not systematic. In later work, Braden-Harder showed that simply matching box codes or subject codes is not sufficient for disambiguation. For example, in I tipped the driver, the codes for several senses of the words in the sentence satisfy the necessary constraints (e.g. tip-money + human object or tip-tilt + movable solid object).

The inconsistencies in dictionaries become a major source of the limitations for word sense disambiguation. So, corpora become a primary source of information for the same (word sense disambiguation).

3.2.3.2 Thesauri

Thesauri provide information about relationships among words, most notably synonymy.

Roget’s International Thesaurus is used in variety of applications including machine translation, information retrieval, and content analysis. It also supplies an explicit concept hierarchy consisting of up to eight increasingly refined levels. Patrick uses Roget’s International Thesaurus for “word-strong neighborhoods,” comprising word groups in low-level semicolon groups, which are the most closely related semantically in the thesaurus, and words connected to the group via chains.
3.2.3.3 Computational Lexicons

WordNet\textsuperscript{48}, CyC\textsuperscript{49}, ACQUILEX\textsuperscript{50}, COMLEX\textsuperscript{51} are some examples of computational lexicons.

There exist two fundamental approaches to the construction of semantic lexicons. The first is the enumerative approach, wherein senses are explicitly provided. The second is the generative approach, in which semantic information associated with given words is underspecified, and generation rules are used to derive precise sense information\textsuperscript{52}. WordNet, the enumerative lexicon, is at present the best known and the most utilized resource for word sense disambiguation in English.

The attempts are made to create a knowledge base from WordNet's hierarchy and apply a semantic similarity function developed by Resnik\textsuperscript{53, 54, 55}. The hypothesis is that for a given set of terms occurring near each other in a text, choosing the senses that minimize the distance among them selects the correct senses. Resnik computes the shared ‘information content’ of words, which is a measure of the specificity of the concept that subsumes the words in the WordNet IS-A hierarchy. The more specific the concept that subsumes two or more words, the more semantically related they are assumed to be.

3.2.3.4 Generative Lexicons

In generative lexicons\textsuperscript{56}, related senses (i.e., systematic polysemy, as opposed to homonymy) are not enumerated but rather are generated from rules which capture regularities in sense creation, as for metonymy, meronymy, etc. A large syntactic and semantic lexicon which provides detailed information about selectional restrictions, etc. for words in a sentence is accessed and then a richly-connected ontology is searched to determine which senses of the target word best satisfies the constraints.

3.2.4 Corpus-based Methods

Since the end of the Nineteenth Century, the manual analysis of corpora has enabled the study of words & graphemes and the extraction of lists of words & collocations for the study of language acquisition or language teaching. Corpora have been used in linguistics since the first half of the Twentieth Century, some of this work concerned word senses. Several important corpora were developed during this period, including the Brown Corpus, the Trésor de la Langue Française, and the Lancaster-Oslo-Bergen (LOB) Corpus etc. Corpus study is done by using a parallel corpus or two monolingual corpora & a bilingual dictionary.

In the area of natural language processing, the ALPAC report (1966) recommended intensification of corpus-based research for the creation of broad-coverage grammars and lexicons. Weiss\textsuperscript{57} demonstrated that disambiguation rules can be learned from a manually sense-tagged corpus. Black\textsuperscript{58} developed a model
based on decision trees using a corpus of 22 million tokens, after manually sense-tagging approximately 2000 concordance lines for five test words.

Corpus-based methods are discussed in terms of tools applied (e.g., automatic sense tagging) as well as their application for defining context of the target word for word sense disambiguation.

3.2.4.1 Automatic Sense-tagging

Manual sense-tagging of a corpus is extremely costly. The overall strategy for automatic sense-tagging is more or less that of most subsequent work on bootstrapping. The problem of data sparseness, which is common for much corpus-based work, is especially severe for work in word sense disambiguation. A class-based bootstrapping method for semantic tagging in specific domains has been proposed as an answer in part, to the problem of data sparseness as it eliminates the need for pre-tagged data. Firstly, the words in the text are automatically clustered and each target word is represented by a vector, a sense is then assigned manually to each cluster, rather than to each occurrence.

Corpus study can use corpus itself alone or take help from other sources. Resnik\textsuperscript{59} uses the taxonomy of WordNet. Yarowsky\textsuperscript{60} uses the categories of Roget's Thesaurus. Slator\textsuperscript{61} and Liddy and Paik\textsuperscript{62} use the subject codes in the LDOCE. Luk\textsuperscript{63} uses conceptual sets built from the LDOCE definitions.

3.2.4.2 Corpora for Context

The 'context' means the nearby words that are present in the sentence containing the ambiguous word.

There are no ‘stand-alone chunks’. Words exist in context, as a concept denoted by a word is contextual (and context is conceptual). Context bears the omni-connecting power for both the electronic and traditional texts. Technically, we can speak of non-linearity of translation, but contexts and concepts ensure mental linearity.

In identifying the meaning of a polysemous word, the context plays a great role. The contexts are highly reliable for disambiguation in 8 out of 10 cases. The context is used in two ways, the bag of words\textsuperscript{4} approach and relational information approach.

The bag of words approach considers the context as words in some window surrounding the target word, regarded as a group without consideration for their relationships to the target in terms of distance, grammatical relations, etc. the bag of words approach has been shown to work better for nouns than for verbs.

In the relational information approach, context is considered in terms of some relation to the target, including distance from the target, syntactic relations, selectional preferences, orthographic properties, phrasal collocation, semantic
categories, etc. Information from micro-context, topical context, and domain contributes to sense selection, but the relative role and importance of information from the different contexts and their inter-relations are not well understood.

Often questions are asked about the sort of useful contextual information and the size of the context that is required for word sense disambiguation.

3.2.4.2.1 Micro-context

Most disambiguation work uses the local context of a word occurrence as a primary information source for word sense disambiguation. Local or “micro” context is generally considered to be some small window of words surrounding a word occurrence in a text or discourse, from a few words of context to the entire sentence in which the target word appears.

3.2.4.2.2 Distance

Use of context for disambiguation has been the basis of word sense disambiguation work in machine translation, content analysis, AI-based disambiguation, dictionary-based word sense disambiguation, as well as the more recent statistical, neural network, and symbolic machine learning, etc.

The distance refers to the number of words that serve as the context for disambiguation of a target word. It is observed that local ambiguities need only a window of 3 or 4 words while semantic or topic-based ambiguities require a larger window of 20-50 words. For different ambiguous words, different distance relations are more efficient.

3.2.4.2.3 Collocation

The term “collocation” was popularized by J. R. Firth. The collocation is not simple co-occurrence but is “habitual” or “usual”. Halliday defines collocation as the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at n removes (a distance of n lexical items) from an item x, the items a, b, c... it is assumed that the word has one sense per collocation.

3.2.4.2.4 Syntactic Relations

Ear used syntax exclusively for disambiguation in machine translation. Yarowsky determined various behaviors based on syntactic category, for example, that verbs derive more disambiguating information from their objects than from their subjects, adjectives derive almost all disambiguating information from the nouns they modify, and nouns are best disambiguated by directly adjacent adjectives or nouns.
3.2.4.2.5 Topical context

Topical context includes substantive words which co-occur with a given sense of a word, usually within a window of several sentences.

3.2.4.2.6 Domain

The use of domain for word sense disambiguation is first evident in the micro-glossaries developed in early machine translation work.

3.2.4.2.7 Sense Division

Word sense disambiguation work is currently relying on the sense distinctions provided by established lexical resources, such as machine-readable dictionaries or *WordNet*. The bank model of sense division attempts to extend the clear delineation between bank-money and bank-riverside to all sense distinctions.

3.3 EVALUATION: WORD SENSE DISAMBIGUATION TASK

The evaluation can be carried out both automatically and manually. Automatic evaluation is done by means of four objective measures, namely, error, coverage, support and novelty. The evaluation is worked out manually by means of a subjective analysis which depends on the level of interest of the evaluator.

However, human beings do not agree on the task at hand. Given a list of senses and sentences, human beings do not always agree on which word belongs in which sense. A computer cannot be expected to give better performance on such a task than a human being. Indeed, since the human being serves as the standard, the computer being better than the human being is incoherent. So, the human performance serves as an upper bound. Human performance, however, is much better on coarse-grained than fine-grained distinctions. This makes the research on coarse-grained distinctions more successful.

It is generally agreed that word sense disambiguation is an application-dependent task. The strategy to aim at the development of word sense disambiguation systems without considering their application is not appropriate, since some aspects, such as the sense repository and the disambiguation process itself, vary according to the application if we consider machine translation as application and focus on the sense repository. For example, objectives of word sense disambiguation in queries are two folded. First is to find appropriate sense (meaning) of each content word in a query and second is to add terms like synonyms, hyponyms etc of the determined sense to the query so as to improve retrieval effectiveness.

3.4 RESOURCES FOR WORD SENSE DISAMBIGUATION

It is clearly and plainly evident that the task of word sense disambiguation calls for determining exactly the one and the only one sense of a given (lexical)
word in a given context, usually, a sentence. To achieve a better result out of such task, one cannot rely only on a bilingual dictionary. The real world knowledge also plays an important part in the scene. A human possesses the real world knowledge through N number of ways. When it comes to automatic word sense disambiguation, researchers have tried several options to build the real world knowledge. The real world knowledge, we are referring as resources for the task of word sense disambiguation. Various types of corpora are popularly used to serve the purpose. The relational information approaches look for different nets such as, word-net, concept net, frame net, how net etc. We will discuss few of them.

3.4.1 Corpus / Corpora

A corpus is simply a body of text. A corpus can be described as a large collection (a greater number of texts than one would be able to easily collect and read in printed form) of authentic texts (example of live language) that have been gathered in electronic form (that can be processed by a computer) according to a specific set of criteria (depending on the purpose of the study). The plural of corpus is corpora. Corpus can be monolingual or multilingual.

Any corpus is constructed on the basis of period of text and language. The criteria for selection of text data are subjects, authors (considering representation of academic merit, social class, economic strata, caste, gender, educational qualifications, geographical location, influence on the language and membership in beneficiary class), the form of presentation (e.g., narrative or informative text, locution (dramatic text), reports, news, textbooks, instructive text, thesis, articles, research papers) and the demographic sampling of readers based on reading habits.

The sources for collection of data to construct a corpus include libraries of universities and colleges, major public libraries, publishers, private collections, online texts. Generation of corpus involves various aspects. There are three modes to collect written texts; from the web, by scanning the printed material, by typing the text manually. Each above referred mode has its own requirements.

Corpus has been applied in a wide range of disciplines and has been used to investigate a broad range of linguistic issues through study of authentic examples of language use (ex.: lexicography, language learning, language teaching, sociolinguistic studies, historical linguistics, grammatics, computational linguistics, natural language processing, stylistics, terminology, translation and technical writing etc.).

Corpus can offer a number of benefits over other types of resources such as printed dictionaries, printed texts, subject field experts and intuition though corpus is not so perfect that it contains all the answers. There is no physical constraint imposed on electronic media such as corpus as in case of printed media. Corpus has the potential to be more extensive than other resources. In addition, its electronic form makes it easier to update than printed resources, and it is easier to consult, as it is less labor-intensive and time-consuming task. It is possible to conduct
exhaustive searches without exhausting the researcher. Moreover, with the help of
corpus analysis tools, one can sort these contexts so that meaningful patterns are
revealed. The discovery that certain usages of words do not occur in a corpus may
help to establish that even though words may appear in dictionaries, they cannot
be used in certain contexts, and that even if a sentence is grammatical, it may not
be idiomatic. Frequency information is another type of data that is much more
easily obtainable when using an electronic corpus along with the corpus analysis
tool. Knowledge about frequency allows one to analyze the lexical patterns
associated with words in a more objective and consistent way, but such
observations are difficult to make when working with other resources. A corpus
once compiled, is constantly available to the researchers unrestrictedly. In addition,
researchers turn to corpus is for reassurance. A corpus can be seen as a test bed
that one can use to verify or reject one’s hypotheses by accessing corpus based
observable and verifiable data. A corpus can act as an objective frame of reference
than intuition. An additional advantage of the corpus-based approach is that it is
more efficient to consult a single corpus-based resource than multiple types of
conventional resources.

A corpus is a single yet broad-ranging resource that can meet the majority of
needs of researchers as one can retrieve information about both lexical and non-
lexical (e.g. style, punctuation, grammar, register) elements of language. In other
words, a corpus is a ‘one-stop shop’.

The data is compiled in a manner, which supports goal-oriented
rearrangements. The database supports selection and retrieval of information. For
more specific studies, goal oriented analysis of data is carried out.

Once a corpus has been compiled, one can use corpus analysis tools to help
with investigations. Most corpus-analysis tools come with two main features,
namely, a feature for generating word lists and a feature for generating
concordances. Corpus based studies include calculating the total number of words
in the corpus, ‘tokens’ & each different word in the corpus, ‘type’, sort the words in
the list in different ways to access information more easily (e.g., in alphabetical
order, in order of frequency), count all the occurrences of a particular word in its
immediate contexts that is the keyword in context (KWIC) and list collocations of
words etc. Thus the corpus-analysis tools vary according to their applications.

The analysis of the context is the essential part of word sense disambiguation
task. Such analysis becomes authentic and objective when corpus serves as the
base for the same, however it may be the nature of ‘bags of words” or ‘node of
relational information’.

3.2.4 Ontology

Ontology is “(the set of) explicit formal specifications of the terms in the
domain and relations among them”\(^6\). It defines concepts, terms and vocabularies
in a domain, and also the relationship among these concepts. Concepts are
organized in a taxonomic structure, with subclasses inheriting properties and specializing from super-classes.

Ontology, together with a set of instances of the classes or concepts defined, constitute a knowledge base about the domain being described\(^7\). Current semantic web technologies also have the added capability of inferring new facts from old facts already captured in the ontology.

Using taxonomies in lexical resources is not a new idea. Roget's Thesaurus groups words with similar meanings in hierarchies (with few number of levels) of classes and sections, while Word-Net is well-known for its “is-a” relations (amongst other types of relations) between “synsets”, or groups of synonymous words. The construction of an ontology-based multilingual lexicon involves four tasks; building the taxonomic structure of the ontology, preparing lexical entries and the information they contain, categorizing the lexical entries under the appropriate semantic classes in the ontology and specifying suitable lexical relations among the lexical entries. The lexical entries are classified under their respective classes. The relations can be specified to link various concepts and instances (in this case, lexical entries). Firstly, the taxonomic structure implies that super-class & sub-class relations already exist in our lexicon.

In fine, ontology implies hierarchical relations between its nodes. These nodes can be concepts, (lexical) words or grammatical frames. Approaching ontologies in different ways, different nets like Word-Net, Frame-Net, Concept-Net etc. are being developed.

### 3.4.3 Word-Net

Word-Net\(^7\) versions for several western and eastern European and Indian languages are currently under development. Word-Net combines the features of many of the other resources commonly exploited in disambiguation work.

It includes definitions for individual senses of words within it, as in a dictionary. It defines “synsets” of synonymous words representing a single lexical concept, and organizes them into a conceptual hierarchy, like a thesaurus. It includes other links among words according to several semantic relations, including hyponymy/ hypernymy, antonymy, meronymy/holonymy, gradations, troponymy and entailment and the cross parts of speech linkages.

When a group of synonymous words put together, a unique meaning of a polysemous word emerges. Each entry in the wordnet carries a synset, a "gloss" and a simple example sentence as shown below.

The glosses in the wordnet explicate the synset senses, but cannot really be encyclopedic, scientific or legal definitions. In explicating the senses, they are assisted by the members of the synset and also the accompanying example sentences. The gloss is used for linking and creating the synsets.
The synsets are constructed abiding by the principles of **Minimality** (the minimal set of words to make the concept unique), **Coverage** (the maximal set of words, ordered by frequency in the corpus, to include all possible words standing for the sense) and **Replaceability** (the example sentence should be such that the most frequent words in the synset can replace one another in the sentence without altering the sense).

The componential approach to the wordnet creation proposes that words are bundles of semantic features which are binary and parallel to those in phonetics. A space of semantic features are designed and the words sharing ALL and ONLY a set of common semantic features be inserted into the same synset.

The wordnet can be constructed based on the ontology (Tamil wordnet at AU-BAK, Chennai which is motivated by Nida's concept classification). It is created by extracting words from the dictionary, grouping words into domains and subdomains and arranging the groups hierarchically.

Automatically create document specific dictionaries, obtain disambiguators and semantic attributes for given words are some of the applications of the wordnet. The glosses and the example sentences contribute to multilingual corpora. Currently, exploiting Word-Net provides the foundation for the attempts for sense disambiguation in the field of natural language processing.

### 3.4.4 Frame-Net

'The Berkeley Frame-Net\(^2\) project is a computational lexicography project based on the principles of *Frame Semantics*\(^3\). Frame semantics is considered as the best way to provide semantic structure to the lexicon. Frame-Net is a lexicon-building effort that describes the frames or conceptual structures with an aim to study words by examining sentences. It is made available using a very large corpus of contemporary English that contains these words and then recording the ways in which information from the associated frames are expressed in these sentences. Corpora used to develop Frame-Net are The British National Corpus (for Frame-Net version 1) and *LDC North American Newswire* corpora in American English along with *BNC* (for Frame-Net version 2).

Frame-Net aims to provide a body of semantically and syntactically annotated sentences from which reliable information can be reported on the valences or combinatorial possibilities of each item analyzed. It embraces *Semantic Classification* and *Syntagmatic Information*. Fine-grained semantic classification as in Frame-Net captures syntactic facts better than syntactic alternation information. Syntagmatic information of the sort provided in Frame-Net serves the purpose of sense discrimination. (This is explained in detail in following discussion.)

Frame-Net is the quest to describe semantics and usage in the Lexicon, focusing on words and their contexts. The output of the Frame-Net project is a database for core lexicon of English which connects Frames, Frame Elements,
words, and semantic annotations. It is useful for second-language learners and ultimately crucial for a variety of natural language applications.

### 3.4.4.1 Constructing Frames

While constructing frames, distinctions based on delivery context such as register, dialect, attitudinal, deixis (aspects of a communication whose interpretation depends on knowledge of the context in which the communication occurs) is ignored. The attention is paid to the number and kinds of entailed entities, both realized and implicit and to the presuppositions of words and other kinds of relationships to other concepts.

Before proceeding to procedures of constructing Frame-net, it is worthwhile to have a quick review of Frame Semantics.73

#### 3.4.4.1.1 Frame Semantics

Frame Semantics presupposes Frames. Frames are based on recurring experiences. So the commercial transaction frame is based on recurring experiences of commercial transaction. A Frame, a schematic representation of a situation, involves various participants, props, and other conceptual roles. Each of this is a Frame Element. A frame, a conceptual structure or prototypical situation, is defined by the inter-relations of Frame Elements (FE). Frame element (e.g., verb, noun etc.) has frame-specific semantic roles. Frame Element has grammatical function. The elements of a frame are decided inductively. The linguistic realizations of frame element highlight different participants and props of a frame. A semantic frame is activated by Lexical Units (LU) (i.e., words in a particular sense). A semantic frame provides the background and motivation for the existence and use of a lexical item in a language. Lexical unit consists pairing between a lemma (i.e., head word) and a frame as hot in 'It’s hot outside today.’ / ‘The curry is really hot.’ / ‘She’s one hot lady’.

A frame is uniquely defined by its Core frame elements, namely, Speaker, Addressee and Message, for example, consider the word (verb) told in ‘Charlie told Jo that Mike was late’. It is marked as ‘Speaker: Charlie’, ‘Addressee: Jo’ and ‘Message: that Mike was late’. The Non-Core frame elements of a frame describe aspects of events more generally such as Time and Place.

Frame is an intuitive construct that links between semantics and syntax. Every sense of every word (i.e., every lexical unit) has its own frame. Lexical units that have similar frame structure are clustered into groups. This allows generalizations and reduces effort.

Semantic Frame is a schematic representation of a situation, object or event providing the background structure against which words are understood. The semantic arguments correspond to the frame elements of the frame associated with that word. A word activates a frame of semantic knowledge relating to the specific concept it refers to. Words not only highlight individual concepts, but also specify a
certain perspective in which the frame is viewed. For example "sell" views the situation from the perspective of the seller and "buy" from the perspective of the buyer. Frame marks Speaker, Manner, Topic, Addressee etc.

A semantic frame is defined as a coherent structure of related concepts that are related in such a way that without knowledge of all of them, one does not have complete knowledge of any one of them. For example, the word "sell" can be understood knowing situation of commercial transfer, a seller, a buyer, goods, money, relation between money and goods, relations between seller and goods and money, relation between buyer and goods and the money.

Semantic roles have been important in natural language processing for many years. Most significant are the relation between lemmas and frames and the relations between frames. Frame-to-Frame relations include (1) composition, by which a complex frame is shown to be decomposable as a temporal structure is often a structured procedural sequence of simpler frames - and (2) inheritance, by which a single frame can be seen as an elaboration of one or more other frames, with bindings between the elements of co-inherited frames. Polysemy is seen as a one-to-many relation between lemmas and the frames that express their meanings.

Frame semantics is a theory that relates linguistic semantics to encyclopedic knowledge and further, it analyzes sentences as constituted by the combination of a verb plus a set of grammatical cases and semantic roles like Agent, Location, and Instrument etc. It assumes that in order to understand the meanings of the words in a language one must first have knowledge of the semantic frames.

The basic idea behind frame-semantics is that one cannot understand the meaning of a single word without access to all the essential knowledge that relates to that word. For example, one would not be able to understand the word "sell" without knowing anything about the situation of commercial transfer, which also involves, among other things, a seller, a buyer, goods, money, the relation between the money and the goods, the relations between the seller and the goods and the money, the relation between the buyer and the goods and the money and so on. It links syntax to the semantic structure. Its main semantic principle is to extend the notion of symbolic units to the grammar of languages. The primary unit being shifted from syntactic unit to grammatical constructions.

There are several typographic conventions for constructing a frame. Frame-Net provides frames for nouns, verbs, adjectives and prepositions. The multi-word expressions are covered in it. Single sentences are annotated in Frame-Net rather than running text. It does not offer information about frequency of occurrence.

### 3.4.4.2 Application of Frame-Net

Frame-Net is unique because of its unique base of frame semantics, its adaptation of semantic annotations and its covering syntagmatic information. Frame-Net is used by researchers in the field of natural language processing for word sense disambiguation, machine translation, information extraction, enriching
semantic representation, rule-based refinement of semantic representation, logical representation and reasoning, LFG-based parsing and question answering. It is also used by lexicographers, language teachers and advanced language learners.

3.4.5 Concept-Net

3.4.5.1 What is Concept-Net?

Concept-Net is a freely available commonsense knowledgebase and natural-language-processing toolkit which supports many practical textual-reasoning tasks over real-world documents right out-of-the-box i.e., without additional statistical training including topic-jisting (e.g. a news article containing the concepts, “gun,” “convenience store,” “demand money” and “make getaway” might suggest the topics “robbery” and “crime”). It is developed in collaboration with over 14,000 authors.

Commonsense knowledge in ConceptNet encompasses the spatial, physical, social, temporal, and psychological aspects of everyday life. Whereas similar large-scale semantic knowledge-bases like Cyc and Word-Net are carefully handcrafted, Concept-Net is generated automatically from the 700,000 sentences of the Open Mind Common Sense Project which is World Wide Web based. The ConceptNet knowledgebase, a semantic network is presently available in two versions: concise (200,000 assertions) and full (1.6 million assertions).

Concept-Net consists of nodes (concepts) and arcs (relation between concepts). Somehow it looks as shown in Figure 1 below.

In excerpt from ConceptNet's semantic network of commonsense knowledge, Compound (as opposed to simple) concepts are represented in semi-structured English by composing a verb (e.g. "drink") with a noun phrase ("coffee") or a prepositional phrase ("in morning").

Figure 1: Concept-Net
Concept-Net is a unique resource in that it captures a wide range of commonsense concepts and relations, such as those found in the Cyc knowledgebase, yet this knowledge is structured not as a complex and intricate logical framework, but rather as a simple, easy-to-use semantic network, like WordNet. While Concept-Net still supports many of the same applications as WordNet, such as query expansion and determining semantic similarity; its focus on concepts-rather-than-words, its more diverse relational ontology, and its emphasis on informal conceptual-connectedness over formal linguistic-rigor allow it to go beyond WordNet to make practical, context-oriented, commonsense inferences over real-world texts.

### 3.4.5.2 Applications of Concept-Net

Concept-Net is not as easily accessible as WordNet for common humans. Concept-Net is meant to be simply useful to AI Researchers and computer enthusiasts who want to experiment with adding commonsense to make their smart robots and programs smarter.

Due to a large number of contributors from the Concept-Net community, Concept-Net is now implemented in several programming languages as it has a variety of Concept-Net-related tools and browsers. Research based on Concept-Net includes analogy-making (e.g. “scissors,” “razor,” “nail clipper,” and “sword” are perhaps like a “knife” because they are all “sharp,” and can be used to “cut something”), text summarization, contextual expansion, causal projection, document classification and other context-oriented inferences.

It is thus possible to construct an ontology-based multilingual lexicon, from various existing language resources, as an alternative to hierarchical lexicons. There are many more Nets, like How-Net, are being build to represent real world knowledge on the basis of principles of ontology.

### CONCLUDING COMMENTS

In fine, word sense disambiguation refers to the task of determining the correct meaning or sense of an ambiguous word in context. This requires first establishing a list of all different meanings (senses) for all the words under consideration. Disambiguation is then performed by evaluating the context of an occurrence of an ambiguous word and the sense entries in the said list, in order to assign the correct sense to the word occurrence under consideration. Accordingly, the selection of equivalent words in a target language, often termed word selection, is done to translate ambiguous words.

The state of the art in the world, as we have seen so far, is much more advanced than in India. There is a huge scarcity of resources in India. Even the basic corpus of each Indian language is not yet developed to an extent to which one needs to perform natural language processing. Few languages like, Marathi, Hindi, Tamil, are fortunate as even Word-Nets (including at least few thousand synsets) are being built in these languages.
In short, Indian languages are poor in resources for natural language processing. And we have therefore to rely on grammatical rules to begin with. This researcher worked on rule-based spell checker and part-of-speech tagger for Marathi. While working on rule-based part-of-speech tagger, she felt the need for developing rule base for disambiguation of lexical words.

In the next chapter, the algorithm for, or the key words for, or the techniques for the disambiguation of lexical words are discussed considering words belonging to all parts of speech. One can develop a system for word sense disambiguation based on it.