HindiTech
A Blind Double Peer Reviewed Bilingual Web-Research Journal

Corpus and its Role in NLP

Thennarasu Sakkan

thennarasus@gmail.com

Associate Professor, Department of Linguistics, Central University of Kerala, India

1. INTRODUCTION

1. 1 WHAT IS CORPUS?

'Corpus' means 'body' in Latin, and literally refers to the biological structures that constitute humans and other animals (Wikipedia). Metaphorically, it refers to collections of things that are felt to share noteworthy characteristics - the body of Hindi literature, the body of Tamil literature, the body of English literature, Indian law and so on. In the context of linguistics, such a body or corpus is a collection of recorded spoken or written text. Corpus is a collection of spoken language stored on a computer and used for language research and writing dictionaries (Macmillan Dictionary 2002). It is a collection of written or spoken texts (Oxford Dictionary 2005). Corpus means a large collection of written or spoken language that is used for studying the language (Longman Dictionary 2003). The collection of a single writer's work or of writing about a particular subject, or a large amount of written and sometimes spoken material collected to show the state of a language (Cambridge Dictionary 1995). A large or complete collection of writings or a collection of utterances, as spoken or written sentences, taken as a representative sample of a given language or dialect and used for linguistic analysis (Random House Webster's College Dictionary, 2001). Corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language (Eagles 1904). A corpus

is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair, 2005). To summarise, Corpus is a collection of texts. Spoken or Written which has been designed and compiled based on a set of pre-defined criteria.

A corpus is a collection of spoken or written texts to be used for linguistic analysis. Any collection of texts assembled in order to investigate one or more linguistic phenomena can be termed a corpus, even if it may only contain a handful of classroom transcripts, interviews or plays. Corpora are now essential tools in research and everyday practice for translators, lexicographers, second language learners, linguists, computational linguists etc. Specialists in these areas share a general goal in using corpora in their work: corpora provide the possibility to find and analyze linguistic patterns characteristic of various kinds of language users, observe language change, and reveal important similarities and divergences across different languages. For the translators who are professional, corpora present an invaluable linguistic and cultural awareness tools. For language learners, they serve as a means to gain insights into specifics of competent language use as well as to analyze typical errors of fellow learners. For lexicographers, corpora are key for observing the development of the vocabularies of languages, making informed decisions as to lexicographic relevance of the lexical material, and for general verification of all varieties of lexicographic data. For computational linguists, corpora are used for enhancing the coverage of all NLP tools (Rao &Thennarasu 2007).

While simple corpus analysis tools such as concordancers have been long in use in these specialist areas, in the past decade there have been important developments in Natural Language Processing (NLP) technologies: it has become much easier to construct corpora and powerful NLP methods have become available that can be used to analyze corpora not only on the surface level, but also on the syntactic, and even semantic, pragmatic, and stylistic levels.

2. WHY DO WE NEED CORPUS?

The corpus has paved the way to many new areas of language research which were unknown to us even a few decades ago. Language corpora, and the results obtained from them have put intuitive language study under strong challenge. In most cases, intuitive observations are proved to be wrong or inadequate while compared with the findings from corpora. Thus, corpora have proved their usefulness in empirical language analysis, theory making, as well as in theory modification which were missing in intuitive language study. However, this trend of corpus-based language research is yet to set its firm footing in India though there have been some sporadic attempts for developing corpora in Indian languages. We should realize that in a multilingual country like India we need to develop language corpora of various types not only to be at par with language related technology developed in other countries, but also to provide advanced resources and systems to our people for their education and research (N. S. Dash, 2001). We need corpus to ensure coverage and testbeds (Manning and Schütze 1999) for our descriptive, typological, theoretical and computational linguistic aim is to describing primary data, explicating precise linguistic information in the lexicon and grammar, richer linguistic typological analyses, improved linguistic analyses by rigorous testing etc.

Intuition alone is not enough

-Is “starting” always replaceable by “beginning”?

- Is “think of” vs “think about”

-What's the difference between "few” and “a few”

Native speaker intuition is unreliable

-provides no information on frequency of occurrence

-“head” => body part

Is this the most used sense?

Help answering questions of usage easily

-More than one character is/are

-Worth to do / worth doing

According to Leech (1992) corpus is a more powerful methodology from the point of view of the scientific method, as it is open to objective verification of results. Svartvik's (1966) pointed out that quantitative data is of use to linguistics. The study of passivisation used quantitative data extracted from corpora. Elsewhere, all successful approaches to automated part-of-speech analysis reply on quantitative data from corpora. The proof of the pudding is in the eating. What Abercrombie (1963) says that corpus research is time-consuming, expensive and error-prone are no longer applicable thanks to the development of powerful computers and software which is able to perform complex calculations in seconds, without error.

3. WHAT ARE THE TYPES OF CORPORA AVAILABLE?

There are many types of corpora, which can be used for different kinds of analyses (cf. Kennedy 1998). Some (not necessarily mutually exclusive) examples of corpus types are as below.

3. 1 WRITTEN VS SPOKEN

One of the major distinctions between different types of corpora is whether they comprise spoken or written data. This is an extremely important distinction because written language generally tends to be far easier to process than spoken language, as it does not contain **fillers**, **hesitations**, **false starts** or ungrammatical constructs. When creating a spoken corpus, one also needs to think about whether an orthographic representation of the text will be sufficient, whether the corpus should be represented in phonetic transcription, or whether it should support annotation on various different levels.

3. 2 SPECIALIZED CORPUS

A corpus of texts of a particular type, such as newspaper editorials, geography textbooks, academic articles in a particular subject, lectures, casual conversations, essays written by students etc. It aims to be representative of a given type of text. It is used to investigate a particular type of language. Researchers often collect their own specialized corpora to reflect the kind of language they want to investigate. There is no limit to the degree of specialization involved, but the parameters are set to limit the kind of texts included. For example, a corpus might be restricted to a time frame, consisting of texts from a particular century, or to a social setting, etc. Some well-known specialized corpora include the 5 million words Cambridge and Nottingham Corpus of Discourse in English (CANCODE) (informal registers of British English) and the Michigan Corpus of Academic Spoken English (MICAS) (spoken registers in a US academic setting).

3. 3 GENERAL CORPUS

A corpus of texts of many types, it may include written and spoken language, or both and may include texts produced in one country or many. It is unlikely to be representative of a particular 'whole', but will include as wide a spread of texts as possible. A general corpus is usually much larger than a specialized corpus. It may be used to produce reference materials for language learning or translation, and it is often used as a base-line in comparison with more specialized corpora. Because of this second function it is also sometimes called a **reference corpus**. Well-known general corpora include the British National Corpus (100 million words) and the Bank of English (400 million words in January 2001), both of which comprise a range of sub-corpora from different sources. Much earlier general corpora were the LOB corpus, consisting of written British English, and the Brown corpus, consisting of written of American English, both compiled in the 1960s and comprising 1 million words each.

3. 4 COMPARABLE CORPORA

Two (or more) corpora in different languages (e.g. English and Hindi) or in different varieties of a language (e.g. Indian English and Canadian English). They are designed along the same lines, for example they will contain the same proportions of newspaper texts, novels, causal conversation, and so on. Comparable corpora of varieties of the same language can be used to compare those varieties. Comparable corpora of different languages can be used by translators and by learners to identify differences and equivalences in each language. The ICE corpora (International Corpus of English) are comparable corpora of 1 million words each of different varieties of English.

3. 5 PARALLEL CORPORA

Two (or more) corpora in different languages, each containing text that have been translated from one language into to another (e.g. a novel in English that has been translated into Tamil, and one in Tamil that has been translated into English) or texts that have been produced simultaneously in two or more languages. They can be used by translators and by learners to find potential equivalent expressions in each language and to investigate differences between languages. Some parallel corpora include European Union regulations, which are published in all the official languages of the European Union.

3. 6 LEARNER CORPUS

A collection of texts – essays, for example – produced by learners of a language. The purpose of this corpus is to identify in what respects learners differ from each other and from the language of native speakers, for which a comparable corpus of native-speaker texts is required. There are a number of learner corpora around the world, of which the best known is the International Corpus of Learner English (ICLE). This is in fact a collection of corpora of 20,000 words each, one comprising essays written by learners of English from a particular language background (French, Swedish, German, etc). There is a comparable corpus of essays written by native speakers of English: the Louvain Corpus of Native English Essays (LOCNESS).

3. 7 PEDAGOGIC CORPUS

A corpus consisting of all the quantum of language a learner has been exposed to. For most learners, their pedagogic corpus does not exist in physical form. If a teacher or researcher does decide to collect a pedagogic corpus, it can consist of all the course books, readers etc a learner has used, plus any tapes etc they have heard. The term 'pedagogic corpus' is used by Willis (1993). A pedagogic corpus can be used to collect together for the learner all instances of a word or phrase they have come across in different contexts, for the purpose of raising awareness. It can also be compared with a corpus of naturally occurring English to check that the learner is being presented with language that is natural-sounding and useful.

3.8 HISTORICAL OR DIACHRONIC CORPUS

A corpus of texts from different periods of time. It is used to trace the development of aspects of a language over time. Perhaps, the best-known historical corpus of English is the Helsinki Corpus, which consists of texts from 700 to 1700 and comprises 1.5 million words.

3.9 MONITOR CORPUS

A corpus designed to track current changes in a language. A monitor corpus is added annually, monthly or even daily, so it rapidly increases in size. However, the proportion of text types remains constant, so that each year (or month or day) is directly comparable with every other.

3. 10 NATIVE VS LEARNER, NATIVE VS TRANSLATED

The native vs learner and native vs translated corpora are the corpora which have been used in contrastive studies of the learner language corpora (Granger 1996, 1998). When matched with comparable native-speaker texts, a learner language corpus provides the basis for revealing the characteristics of the learner language, e.g. identifying interference from the mother tongue. A comparison of texts produced by learners with different mother-tongue backgrounds makes it possible to reveal general characteristics of the learner language, in much the same way as corpora of translated texts in different

languages may be used to identify general characteristics of translated texts. Granger (1993) pointed out that contrastive and learner corpora are closely interrelated. He provides the basis for describing the relationships between languages and formulating hypotheses about learning problems. The letter can be used to identify characteristics of learner language, which may in their turn be related to a contrastive description.

3. 11 PLAIN TEXT VS ANNOTATED TEXT

Perfectly plain, no information about text (usually, not even edition), can this at all be considered a corpus? Marked up for formatting attributes: e.g. page breaks, paragraphs, font sizes, italics, etc. Annotated with identifying information, e.g. edition date, author, genre, register, etc. Annotated text is for part of speech, syntactic structure, discourse information, etc.

There has been an increasing need for annotated corpora of Indian languages lately by researchers addressing different issues in linguistics and natural language processing. Linguists are using annotated corpora to study a number of linguistic phenomena. Hardt, D. (1992) uses tagged corpora for a study on VP ellipsis. Niv, M. (1993) uses a syntactically annotated corpus to develop a theory about how humans resolve syntactic ambiguity in parsing.

3. 12 MONOLINGUAL VS MULTILINGUAL

A monolingual corpus is an equally valuable resource, though usually for different purposes. As monolingual corpora are generally larger and, in some cases, may be considered representative, they are able to offer information about more or less standard language use on the basis of quantitative data. Moreover, a monolingual corpus can be an important source of translation equivalents for specific expressions, technical terms, or recent borrowings, naturally requiring different search strategies. Unlike the dictionary, a concordance leaves it to the user to work out how an expression is used from the data. This typically calls for more in-depth processing than does consulting a dictionary, thereby increasing the probability of learning. In more general terms, by drawing attention to the different ways expressions are typically used and with

what frequencies, corpora can make learners more sensitive to issues of phraseology, register, and frequency, which are poorly documented by other tools (Aston, 1999). The CIIL corpora are examples of monolingual corpora such as Tamil, Telugu, Kannada and Malayalam. And English, Hindi, Punjabi by Indian Institute of Technology, New Delhi, and Marathi, Gujarati by Deccan College, Pune, and Oriya, Bangla, Assamese by Indian Institute of Applied Language Sciences, Bhubaneswar, and Sanskrit Sampurnanand Sanskrit University, Varanasi, and Urdu, Sindhi, Kashmiri Aligarh Muslim University Aligarh, each languages carries 3 million word of monolingual corpus.

3. 13 MULTILINGUAL CORPORA

Not all corpora are monolingual, and an increasing amount of work is being carried out on the building of multilingual corpora, which contain texts of several different languages. First we must make a distinction between two types of multilingual corpora: the first can really be described as small collections of individual monolingual corpora in the sense that the same procedures and categories are used for each language, but each contains completely different texts in those several languages. For example, the Aarthus corpus of Danish, French and English contract law consists of a set of three monolingual law corpora, which does not comprise translations of the same texts.

The second type of multilingual corpora (and the one which receives the most attention) is parallel corpora. This refers to corpora which hold the same texts in more than one language. The parallel corpus dates back to mediaeval times when "polyglot bibles" were produced which contained the biblical texts side by side in Hebrew, Latin and Greek etc.

A parallel corpus is not immediately user-friendly. For the corpus to be useful it is necessary to identify which sentences in the sub-corpora are translations of each other, and which words are translations of each other. A corpus which shows these identifications is known as an aligned corpus as it makes an explicit link between the elements which are mutual translations of each other. For example, in a corpus the Tamil sentence "*oru puttakam mēcai mītu uḷḷatu*" and "One book is on the table" might be aligned to one another. At a different level, specific words might be aligned, e.g. "*oru*"

with "one". This is not always a simple process, however, as often one word in one language might be equal to two words more than two in another language.

At present there are few cases of annotated parallel corpora, and those which exist tend to be bilingual rather than multilingual. However, two EU-funded projects (CRATER and MULTEXT) are aiming to produce genuinely multilingual parallel corpora. The Canadian Hansard corpus is annotated, and contains parallel texts in French and English, but it only covers a restricted range of text types (proceedings of the Canadian Parliament). However, this is an area of growth, and the situation is likely to change dramatically in the near future.

4. SOME MAJOR CORPORA

As corpus building is an activity that takes time and costs money, readers may wish to use ready-made corpora to carry out their work. However, as a corpus is always designed for a particular purpose, the usefulness of a ready-made corpus must be judged with regard to the purpose to which a user intends to put it. There are thousands of corpora in the world, but most of them are created for specific research projects and are thus not publicly available. While abundant corpus resources for languages other than English are also available now (Taylor & Francis, 2006). Some major English corpora are Brown (1964, American English written 1 million), LOB (1980, British English, written 1 million), London-Lund (1990, British English, spoken 1million), Helsinki (1993, diachronic English), Cobuild, Bank of English (more than 500 million), Penn Tree Bank (1992, syntactically annotated), British National Corpus (1994, 100 million) etc.

5. CORPUS MANAGEMENT

Corpora, as a kind of empirical data, plays a crucial role in current NLP and linguistics. While the size of the corpora has been increased from three million to some few more million for Indian languages (cf. CIIL-EMILLE Corpora for Indian Languages). In case of English, corpora has increased from several million to hundreds of millions (cf. Brown Corpus), the management of such a vast amount of data is undeniably

complicated. So there is a need for corpus management system which can able to deal with extremely large corpora and is able to provide a platform for computing a wide range of lexical statistics. As Rychly, (2000) points out that an ideal general-purpose corpus management tool should embrace the complete life cycle of a corpus. For text data, it should enable:

- text preparation – conversion from various formats, encodings, etc.;
- metadata management – integration of the information about the source of data, authors, topics, genre, etc.
- tokenization – language-dependent determination of the elementary unit accessed, usually a word;
- corpus annotation – potentially ambiguous, manual and automatic tagging on morphological, syntactic, semantic and pragmatic levels
- efficient corpus storage – the storage requirements of the indexes needed for querying should be minimized as should the time required for their creation;
- concordancing – retrieving language data matching the user's query;
- computation of statistics – searching for typical patterns in data, frequency distribution of various features, co-occurrence statistics, etc.

Moreover, the ideal corpus management tool should implement all these tasks independent of:

- the language - especially text preparation, tokenization and corpus annotation;
- the platform (efficient storage and retrieval of corpus data as well as demanding computation present a challenging task for a platform independent implementation).

To meet all of these requirements, people develop corpus management tools to handle and implement all these criteria and provide an appropriate platform for integrating the language and annotation-dependent tasks carried out by external tools. It deals with design and development of systems that can be employed to manage corpora, especially,

extremely large ones with millions and billions of words, and enables the efficient evaluation of complex queries and the computation of advanced statistics. The representative information that is stored within the body of the corpus are major and sub categories of texts, source, date of origin, authorship and publishers. Using the corpus management tool one can also retrieve these texts selectively. For example, one can extract all the texts grouped under a particular subcategory or the texts from a particular period, etc.

6. INDIAN LANGUAGES CORPORA

For the first time the texts of Indian languages are made available in machine readable form through the project 'Development of Corpora of text of Indian Languages' started in 1991 by the Department of Electronics (DoE), Govt. of India. The corpora development project for the 15 scheduled languages has been undertaken by six different centres. Later languages, newly added to the 8th schedule have also been included for building corpus. The objective, size of the corpora, coordination between centres, etc. have been discussed elaborately by Annamalai (1994). The Central Institute of Indian Languages, Mysore has taken up the corpora development work for Kannada, Malayalam, Tamil and Telugu Languages (Ganesan & Raja, 2004). And English, Hindi, Punjabi by Indian Institute of Technology, New Delhi, and Marathi, Gujarati by Deccan College, Pune, and Oriya, Bangla, Assamese by Indian Institute of Applied Language Sciences, Bhubaneswar, and Sanskrit Sampurnanand Sanskrit University, Varanasi, and Urdu, Sindhi, Kashmiri Aligarh Muslim University Aligarh, each languages carries 3 million word corpus (Dash N. S. 2003 and 2007). The developed Indian languages corpora for fifteen languages are being centrally maintained at Central Institute of Indian Languages. These corpora can be used for education and research purpose.

6. 1 LDC-IL, CIIL CORPUS

Recently, with so much work being done on the analysis of some Indian languages corpora at LDC-IL (2009), it is seen as essential to annotate a corpus with the results of the research. Obviously, this can act as a bootstrap for an increasingly detailed and

accurate analysis at the same linguistic level or for the next level of research. We can build a hierarchy of analyses from POS tagging, parsing, semantic tagging to discourse analysis.

In natural language processing, various researchers have been using annotated corpora to train the stochastic part of speech taggers and parsers (Church, K. 1988). Annotated corpora are being used as the gold standard by which different parsers can be objectively compared (1991). At present, researchers are limited by the existing annotated corpora of Indian languages and the descriptions provided in those corpora. A system that could automatically annotate a corpus in any Indian language with little human labor required would greatly enhance progress that could be made by researchers using corpora in their work. Even if an adequate annotation accuracy level cannot be obtained using automated procedures, an automated annotator could still be used to bootstrap the process of manually annotating a corpus. In (Marcus, M et., al. 1993), it is shown that manually correcting the output of an automated tagger results in greater speed and accuracy than manually annotating from scratch. A number of researchers in corpus-based computational linguistics believe that the size of available annotated corpora is the current limiting factor in creating accurate corpus-trained natural language processing systems. If this is the case, the cycle of automatically annotating a corpus, manually correcting it and retraining the automatic annotator on the larger corpus could provide a fast mechanism for providing very large annotated corpora.

[To know more about Indian languages corpora and their development one can visit LDC-IL website for details http://www.ldcil.org/CorporaCreationInIndianLanguages.html]

7. AREAS OF USING CORPORA

Lexicographers, language teachers and learners, translators, language engineers and NLP researchers, grammar and vocabulary learner, and examination, business and general English course books have all benefited from the information in the corpus. We

no longer have to rely heavily on intuition to know what people say or write; instead we can see what hundreds of different speakers or writers have actually said or written. So, materials developed with a corpus are more authentic and can illustrate language as it is really used.

Corpus is of best possible use to lexicographers if it is loaded into a corpus query tool which supports them in finding collocational and grammatical patterns. To that end the corpus must be grammatically analyzed. While suitable tools were available for English (Adam Kilgarriff & Micheal Rundell, 2006). For Language teacher/learners, corpus is used for syllabus design, materials development, and classroom activities. The syllabus organizes the teacher's decisions regarding the focus of a class with respect to the students' needs. Frequency and register information could be quite helpful in course planning choices. The development of materials often relies on a developer's intuitive sense of what students need to learn. With the help of a corpus, a materials developer could create exercises based on real examples which provide students with an opportunity to discover features of language use (Barlow 2002). For many NLP applications rely on the availability of large amounts of corpus. For linguists, corpus is a useful resource to pursue linguistic research based on real rather than contrived. For translators the corpus is a training resource, parallel and in addition to a dictionary and a thesaurus. The study of parallel texts enables translators to see how similar meanings were expressed in the texts serving similar functions. For language engineers, corpus has become an important resource to develop any applications.

8. WHAT ONE CAN DO WITH CORPUS?

Corpora makes use of implicit knowledge explicitly in the form of knowledge base. We can model intelligent behavior by making use of raw quantitative data to generate some statistical models of natural language behavior. Corpora provide raw data for the approaches of language modelling (whether mathematical or statistical models involving quantification) while analyzing the corpora capturing linguistic behavior.

Using corpus one can develop software tools for language processing, bigram, trigram, n-gram and concordance/KWIC, collocation, keywords, word frequency, syllable frequency, character frequency. There are some softwares or programmes which are developed and freely available for public use one can make use of this.

9. CORPUS IN NATURAL LANGUAGE PROCESSING

Natural Language Processing is unthinkable without involving corpora. Corpora are essential ingredients of every aspect of natural language processing. Corpora are extremely relevant in the construction of viable natural language processing systems for a wide range of tasks. Language modelling involves a variety of uses of corpus in the areas such as parts of speech analysis, computational lexicography, morphological analyzer, parsing, word sense disambiguation etc.

10. ANNOTATED CORPORA

Computational processing of language may use corpora of different kinds. Often such corpora are enriched with explicitly labelled parts of speech of words. Assignment of POS labels to the elements of corpora is called automatic *tagging*. The task of tagging corpora is usually done manually or automatically. Automatic assignment of such task by computers is called *tagging*. Parts of speech taggers are important in corpus linguistics. POS taggers eschew linguistic rules in assigning parts of speech to corpora. Alternatively, one may use statistical *heuristics* for assigning tags, on the basis of contextually dependent guess work. Today Parts of Speech information is introduced into the texts as the primary step in the development of annotated corpora with a high degree of automation.

Corpus building projects provide corpora, which are automatically annotated with parts of speech. By using computers to annotate, parse and calculate the probabilities of items of the corpora, the tedious task of manual corpus building and analysis is avoided to save time and cost. Another important exercise in the development of annotated corpora involves the introduction of labelling syntactic structures in the text. Again one may follow manual or automatic techniques.

Automated parsing may involve already POS tagged texts. Parsing involves the identification of grouping syntactically dependent elements and labelling them with appropriate syntactic tags. Many natural language processing applications regularly use corpora annotated with syntactic parsing.

10. 1 CORPUS VS PARTS OF SPEECH

Part of speech (POS) tagging is the process of labelling annotation of syntactic categories for each word in a corpus. In other words, POS tagging is the process of labelling a part of speech or other lexical category to each and every word in a sentence. Tagging is also known as the primary phase in the assignment of structure to a text.

Example, The boy loves the girl.

The\DET boy\NC loves\MV the\DET girl\NC .\PUN

For some time, part-of-speech tagging was considered an inseparable part of natural language processing, because of certain cases of which the correct part-of-speech could not be decided without understanding the semantics or even the pragmatics of the context. This is extremely expensive, especially because analysing the higher levels is much harder when multiple part-of-speech possibilities ought to be considered for each word.

It is impossible to think over research on Part-of-Speech tagging without corpus. Corpus has been used for innumerable studies about part-of-speech, and also has inspired the development of similar "tagged" corpora in many other languages. However, by this time it has been superseded by larger corpora such as the 100 million words British National Corpus. For a corpus, there is a necessity for POS tagging so that the resources can be utilised in natural language parsing, machine translation etc.

10. 2 CORPUS VS MORPHOLOGICAL ANALYZER

Morphological analysis is the process of segmenting words into morphemes and analysing the word formation. It is a primary step for various types of text analysis of

any language. Morphological analyzer takes a word as an input and produces the root, and its grammatical features as the output (Rao 2006).

For example,

Input: oxen (English)

Output: {root = ox, category = n, number = plural}

It is used as one of the components in machine translation etc. Corpus is used in the development of morphological analyser for enhancing its coverage. The partially developed tagged corpus can be used to improve morphological analyzer, and hereafter, once we have analysed one chunk of data, we allow feedback to be incorporated into the morphological analyzer for the next, chunk. For this purpose, we first divide the original text into many chunks. In addition, using bootstrapping methods one can improve the coverage of Morphological Analyzer.

10. 3 CORPUS VS MACHINE TRANSLATION

Machine Translation (MT) is a task with multiple components, each of which can be very challenging. MT makes it possible and easier to collect knowledge in other languages, as well as to distribute knowledge to other languages. Parallel corpus has been used as an aligned pair of any two languages in MT. The aligned data comes from a corpus.

Using corpus approach to machine translation usually begins with a bilingual training corpus. This approach is to extract from the corpus generalised statistical knowledge that can be applied to new, unseen test sentences. A different approach is to simply memorise the bilingual corpus. This is called translation memory, and it provides excellent translation quality in the case of a "hid" (i.e., a test sentence to be translated has actually been pre-observed in the memorised corpus).

MT systems may use a large monolingual corpus to improve the accuracy of translated words, phrases/sentences. The MT system may produce alternative

translations and use the large monolingual corpus to (re)rank the alternative translations. The MT system may receive an input text segment in a source language, compare alternate translation for the said input text string in a target language and record a number of occurrences of the alternate translations in the large monolingual corpus. The MT system may then re-rank the alternate translations based, at least in part, on the number of occurrences of each translation in the corpus.

10. 4 CORPUS VS WORD SENSE DISAMBIGUATION

Word sense disambiguation is defined as the task of finding the correct sense of a word in a context. This is crucial for applications like Machine Translation and Information Extraction. Sense-tagged corpus is still not large enough to create the building of a wide coverage, high accuracy WSD program that can significantly outperform the most-frequent-sense classifier over all content words encountered in an arbitrarily chosen unrestricted text. The amount of human annotation effort needed can be considered as an upper bound on the manual effort needed to construct the necessary sense-tagged corpus to achieve wide coverage WSD. It may turn out that we can achieve our goal with much less annotation effort.

Large text corpora and the computational resources to handle them have recently become available to computational linguists. In order to apply to multi-million word corpora, natural language processing techniques must be efficient and domain-independent; for this reason, coarse or partial analyses are becoming more attractive. For example, coarse syntactic interpretation, such as partial parsing (de Marcken 1990) (McDonald 1990) and automatic collocation generation (Smadja & McKeown 1990) (Chouek 1988) are being explored.

10.5 Bootstrapping from Bilingual Corpora

The time required for hand-labelling the training sentences is prohibitive, but there is a way it might be automated. Recently several researchers (e.g. (Brown et al. 1991), (Dagan et al. 1991)) have suggested using bilingual aligned corpora in the lexical disambiguation task, (the term “aligned” indicates that within the bilingual database,

sentences that are translations of one another are grouped together). Although most of the discussion is in terms of choosing words for translating one language to another, it is also suggested that, because the words that have more than one sense in language A have only one sense or a different set of senses in language B, by using a most frequently occurring bilingual dictionary, the correct sense of a word in language A can be determined by comparing it with its translation in B. This disambiguation method is of limited applicability, of course, because it requires a bilingual corpus. However, a corpus of translated text could be used to bootstrap catchword, providing it with initial training instances (sentences containing a target homograph tagged with its sense), thereby eliminating the hand-labelling step.

Corpus-based methods are called ``supervised'' when they learn from previously sense-annotated data, and therefore they usually require a large amount of human intervention to annotate the training data Ng (1997). Although several attempts have been made for example, Leacock (1998), Mihalcea (1999), Cuadros (2004), the knowledge acquisition bottleneck (too many languages, too many words, too many senses, too many examples per sense) is still an open problem that poses serious challenges to the supervised learning approach for WSD.

11. SEMANTIC ANNOTATION

Yet another kind of enrichment of corpora involves annotation of texts with semantic labelling. To understand text and analyse it appropriately it is essential that the relevant text may be semantically tagged. NLP applications such as Information Retrieval and Machine Translation may use semantic labelling as a part of sense disambiguation exercise.

12. CONCLUSION

It has been observed that there are many corpus linguists who are more interested in Computational aspects than Linguistics. They have used corpora for research in the area of Computational Linguistics or Natural Language Processing (NLP) (Mayer, 2002). A statistical study of certain aspects of any written corpus following various computational

techniques serve a number of purpose in building NLP applications such as Morphological Analyzers, Generators, Machine Translation, Text Generation, Dictionary tools, Lemmatization, Speech application, Grammar Checking, Spell checking and the retrieval of documents. A study of any written corpus may enhance our knowledge especially in the realm of standardisation of that language used in the written domain.

The quality and the development of many NLP applications rely on the availability of large amounts of textual data today. Many applications use statistical algorithms that are trained on electronic corpora. Machine Translation is a case in point. With the arrival of fast computers and large amounts of machine-readable texts in the 1970s, it has become possible to start using corpus-based computational techniques for translation purposes. Today, parallel corpora and various alignment techniques in Machine Translation and Machine aided Translation are increasingly used.

Corpus is a basic resource for empirical linguistic research. Corpora are readily available objective material for analysis, since the samples are representative of language in use, easily accessible, natural and rich resources. It is easy to use for analysis since it is often enriched by annotation.

ACKNOWLEDGEMENTS

I would like to thank my guide Prof. G. Uma Maheswar Rao for his guidance and support. I would like to extend a special thank to Dr. Natarajapillai who has corrected first draft of this paper and Dr. Dr. Niladri Sekhar Das for his comments and suggestions. A special thanks to Dr. P. Matthew anna for his continuous and generous support suggestion and my LDC-IL team.

REFERENCES

- Aroonmanakun, W. (2006). Corpus Linguistics. Bangkok: Chulalongkorn University Press.

- Biber, Douglas, Conrad, Susan, & Reppen, Randi. (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge: CUP.
- Baker, Paul, Andrew Hardie & Tony McEnery. (2006). A Glossary of Corpus Linguistics. Edinburgh: Edinburgh University Press.
- Black, E. et al. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In Proceedings of fourth DARPA Speech and Natural Language Workshop, pages 306-311.
- Brown, P. F., S. A. D. Pietra, V. J. D. Pietra & R. L. Mercer (1991). Word sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264-270.
- Choueka, Y. (1988). Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases. *Proceedings of the RIAO*, pages 609-623.
- Church, K. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In Proceedings of the Second Conference on Applied Natural Language Processing. ACL.
- Cuadros, M., Atserias, J., Castillo, M., Rigau, G. 2004. Automatic acquisition of sense examples using ex-retriever In IBERAMIA Workshop on Lexical Resources and The Web for Word Sense Disambiguation. Puebla, Mexico.
- Dagan, I., A. Itai. & U. Schwall (1991). Two languages are more informative than one. In *Proceedings of the 29th Annual meeting of the Association for Computational Linguistics*, pages 130-137.
- Dash, Niladri Sekhar (2001) "Application of corpus in language teaching and research". Journal of Applied Linguistic Research. Vol. 2. No. 2. Pp. 69-82.

- Dash, Niladri Sekhar (2003) "Corpus linguistics in India: present scenario and future direction". Indian Linguistics. Vol. 64. No. 1-4. Pp. 85-113.
- Dash, Niladri Sekhar (2007) "Some techniques used for processing Bengali corpus to meet new demands of linguistics and language technology". SKASE Journal of Theoretical Linguistics. Vol. 4. No. 2. Pp. 13-33.
- de Marcken, C. G. (1990). Parsing the lob corpus. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 243-252.
- Fillmore, Charles J. (1992). ""Corpus linguistics" or "Computer-aided armchair linguistics"". In Svartvik, Jan (ed.) Directions in Corpus Linguistics. Berlin: de Gruyter. 35-60.
- Ganesan, M. Raja, S. (2004). Morpheme and Parts-of-Speech tagging of Tamil corpus. Published in Lecture Compendium: SIMPLE'04 – Symposium on Indian Morphology, Phonology & Language Engineering. Shyama Printing Works, Prembazar, Kharagpur.
- Granger, S. (1993). The International Corpus of Learner English. In: Aarts J./de Haan P./Oostdijk N. (eds.), English Language Corpora: Design, Analysis and Exploitation. Amsterdam: Rodopi, 57-69.
- D. Hardt. (1992). An algorithm for VP ellipsis. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.
- Leackock, Chodorow, Miller Leackock and Miller, G. (1998). Using corpus statistics and WordNet relations for sense identification Computational Linguistics. Special Issue on WSD, 24(1).
- Leech, G., Myers, G. & Thomas, J. (eds.) (1995). Spoken English on Computer. London: Longman.

- Leech, G. (1992). "Corpora and theories of linguistic performance": in Svartvik, J. (ed), Directions in corpus linguistics: proceedings of Nobel symposium 82, Berlin and New York, Mouton de Gruyter, 125-148.
- Kennedy, Graeme. (1998). An introduction to Corpus Linguistics. London & New York: Longman.
- Kilgarriff, Michael Rundell & Elaine Uí Dhonnchadha. (2006). Efficient corpus development for lexicography: building the New Corpus for Ireland. Language Resources and Evaluation Journal 40 (2): 127-152.
- McDonald, D. D. (1990). Robust partial-parsing through incremental, multi-level processing: rationales and biases. In P.S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research in Text Analysis, Information Extraction, and Retrieval,* pages 61-65. GE Research & Development Centre, TR 90CRD198.
- Marcus, M. et. al. (1993). Building a large annotated corpus of English: the Penn Treebank. To appear in Computational linguistics.
- McEnery, Tony & Andrew Wilson. (2001). Corpus Linguistics. Edinburgh: Edinburgh UP.
- Meyer, Charles F. (2002). English Corpus Linguistics: An Introduction. Cambridge: CUP.
- Mihalcea, R. Moldovan, D. (999) A Method for word sense disambiguation of unrestricted text In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic, ACL'99, 152-158 Maryland, USA.
- Ng, H. (1997). Exemplar-Base Word Sense Disambiguation: Some Recent Improvements. *In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, EMNLP.

- Niv, M. (1993). Resolution of syntactic ambiguity: the case of new subjects. In Proceedings of the 15th Annual Meeting of the Cognitive Science Society.
- Roa, Uma Maheswar G. & Thennarasu, S. (2007). Corpus Linguistics PGDCAIL-421. Centre for Distance Education, University of Hyderabad. Hyderabad.
- Rychly, P. (2000). Corpus Managers and their effective implementation. PhD Thesis, Faculty of Informatics, Masaryk University.
- Smadja, F. A. & K. R. McKeown (1990). Automatically extracting and representing collocations for language generation. *In Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 252-259.
- Schabes, Y. et. al. (1993). Parsing the Wall Street Journal with the inside-outside algorithm. In Proceedings of the 1993 European ACL, Uterich, The Netherlands.
- Scherer, Carmen. (2006). Korpuslinguistik. Heidelberg: Winter.
- Susan Hunston. (2002). Corpora in Applied linguistics. Cambridge, Cambridge University Press, United Kingdom.
- Teubert, Wolfgang & Anna Cermáková. (2007). Corpus linguistics: A short introduction. London. Continuum.
- Krieger, Daniel. (1999). Corpus Linguistics: What It Is and How It Can Be Applied to Teaching: Siebold University of Nagasaki (Nagasaki, Japan)

Developing Linguistic Corpora: a Guide to Good Practice

- (http://www.ahds.ac.uk/creating/guides/linguisticcorpora)
- ftp://ftp.cordis.europa.eu/pub/tmr/docs/soclangcorp980480.pdf

Expert Advisory Group on Language Engineering Standards

- (http://www.ilc.cnr.it/EAGLES96/home.html)
- http://ell.phil.tu-chemnitz.de/compPhil/corpus.html

Citation: Sakkan, Thennarasu (2010). Corpus and its Role in NLP , HindiTech: A Blind Double Peer Reviewed Bilingual Web-Research Journal, 1 (3), 24-48. URL: https://hinditech.in/corpus-and-its-role-in-nlp/