CORPUS AND ITS ROLE IN NLP

Thenarasu Sakkan, Lecturer in LDC- II, CII, Mysore.

INTRODUCTION
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 Tamil literature, the body of English literature 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 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). To summarize, Corpus is a collection of texts. Spoken or Written which has been designed and compiled based on a set of pre-defined criteria (Rao, GU & Thenarasu, S 2007).

2. WHY DO WE NEED CORPUS?
   The corpus has paved way too many new areas of language research which were unknown to us even a few decades ago. Language corpora, and 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). Intuition alone is not enough.

1. Is arambam-starting always replaceable by totakkam-beginning?
2. Is -ninal-think of vs cinti-think about?
Corpus. It may be used to produce reference materials for language learning or translation, and it is often used as baseline in comparison with more specialized corpora. Because of this secondary function, it is also sometimes called a reference corpus. Well-known general corpora include the British National Corpus (100 million words).

Comparative Corpora

Two or more corpora in different languages (e.g., Tamil and English) 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, conversation, and so on. Comparable corpora of different languages can be used by translators and by learners to identify differences and equivalencies in each language. The ICE corpora (International Corpus of English) are comparable corpora of 1 million words each of different varieties of English.

Parallel Corpora

Two (or more) corpora in different languages containing text that have been translated from one language into 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.

4. CORPUS MANAGEMENT

Corpora, as a kind of empirical data, play a crucial role in current NLP and linguistics. While the size of the corpora has increased from three million to a few more million for Indian languages (cf. CII-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 vast amount of data is undeniably complicated. There is a need for corpus management system which can be used to deal with extremely large corpora and is able to provide a platform for computing a wide range of lexical statistics. As Rychnovsky (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:

i) Text preparation – conversion from various formats, encodings, etc.,

ii) Metadata management – integration of the information about the source of data, authors, topics, genre, etc.

iii) Tokenization – language – dependent determination of the elementary unit accessed, usually a word;

iv) Corpus annotation – potentially ambiguous, manual and automatic tagging on morphological, syntactic, semantic and pragmatic levels

v) Efficient corpus storage – the storage requirements of the indexes needed for querying should be minimized as should the time required for their creation;

vi) Concordance – retrieving language data matching the user’s query;

vii) 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;

viii) The language – especially text preparation, tokenization and corpus annotation;

ix) The platform (efficient storage and retrieval of corpus data as well as demanding computation present a challenging task for a platform in dependent implementation).

To meet all these requirements, people develop corpus management tool 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 is deals with design and development of system that can be employed to manage corpora, especially, extremely large ones with millions 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 is major and sub-categories of texts, source, and 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 sub-category or the texts from a particular period, etc.

5. WHO WANT TO USE CORPUS?

Lexicographers, language teachers and learners, translators, language engineers and NLP researchers, grammar and vocabulary leaner, 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 collocation and grammatical patterns (Adam Kilgariff & Michael Rundell, 2006). For language teachers/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 corpus. For linguists, corpus is a useful resource to pursuing linguistic research based on real rather than contrived. For translator corpus is a training resource, parallel and in addition to a dictionary and a thesaurus. The study of parallel texts which are 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.

6. 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 areas of Computational Linguistics of 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, and so on. A study of any written corpus may enhance our knowledge especially in the real of standardization of that language used in the written domain.

The quality and the development of many of 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. Linguistics.