

8.19 Issues in Developing Corpus for Malayalam from Web as Source

S. A. Shanavas, University of Kerala

Abstract— This paper is about developing corpora for Malayalam language from the web as the source. It discusses the issues faced when attempting to prepare text corpus from the web. Creating corpora from web or web as corpus is a new and emerging area. Source for this corpus development is Malayalam language texts in the net. Earlier it was developed by using keyboard, OCR, etc. Now it is towards the web. The quantity of texts in the web is enormous. The two possibilities with respect to web corpus are the web as the source for making corpus and considering the web itself as the corpus. Here the reference is for the former one. The advantages with web as source are significant as it include the selection of texts from various domains, less spelling errors, etc. There are some serious issues to be considered in developing such corpora; the problems relate to encoding, spilt of words, space alignments, non-uniformity or non-standardisation of fonts, text in different formats, copyright issues, selection of texts without banner and advertisements, popup files, etc.

I. INTRODUCTION

ANY collection of more than one text can be called a corpus, but the term "corpus" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition. The following list describes the four main characteristics of the modern corpus:

- Sampling and representativeness
- Finite size
- Machine-readable form
- Updating possible

The concept of carrying out research on written or spoken texts is not restricted to corpus linguistics. Indeed, individual texts are often used for many kinds of literary and linguistic analysis - the stylistic analysis of a poem, or a conversation analysis of a TV talk show, etc. However, the notion of a *corpus* as the basis for a form of empirical linguistics is different from the examination of single texts in several fundamental ways.

A large collection of texts for any type of research is said to be corpus and for centuries people have been collecting manuscripts, books and newspapers for analysis of a very laborious nature. Thankfully, as technological advances make the computerized storage and access of large quantities of information easier, so the construction and use of text corpora continue to increase, and the potential for research has widened considerably. A corpus can also be thought of as a

collection of texts gathered according to particular principles for some particular purpose. A corpus is valuable because together its component texts allow statements to be made about language as a whole. Corpus linguistics provides the methodology to extract meaning from texts.

Corpus Linguistics (CL) has now emerged as one among the major areas in language technology and linguistic study. Computational Linguistics (CL), Corpus Lexicography (CL) and Corpus Linguistics (CL) are the three 'CL's important to the Modern Language Technologists. Once it was rejected by relativists and generative grammarians like Chomsky and others who argued that corpus would not be helpful for linguistic analysis. Realising the importance and uses of corpus in wide areas of corpus development has now become the focus of attention. Taking as its starting point the fact that language is not a mirror of reality but let us share what we know, believe and think about reality, it focuses on language as a social phenomenon, and makes visible the attitudes and beliefs expressed by the members of a discourse community. As the corpus construction plays an important role in language technology, researchers are building their own corpora, solving problems independently, and producing project specific systems, which cannot easily be re-used. The process of corpus building is a new research area, which lacks standardization and appropriate tools. Corpora, however, have changed the way in which linguists can look at a language.

The importance of corpora to linguistic study can be appreciated from the definitions given to it. It is interesting to note that the first definition suggests that corpora can help in studying language, while the second goes further and implies that categorical statements can be made on the basis of a well-defined corpus. This is the most basic type of linguistic corpus annotation - the aim being to assign to each lexical unit in the text a code indicating its part of speech. Part-of-speech annotation is useful because it increases the specificity of data retrieval from corpora, and also forms an essential foundation for further forms of analysis (such as syntactic parsing and semantic field annotation). Part-of-speech annotation also allows us to distinguish between homographs.

Language corpora contain large amount of textual information in machine-readable form of whole texts or large continuous genre of text extracts from different fields. The texts can be whole books, newspapers,

S.A. Shanavas is with Technology and Resource Center for Malayalam, Department of Linguistics, University of Kerala, Kariavattom, Thiruvananthapuram, Kerala (e-mail: drsasha2002@yahoo.com)

journals, speeches etc, or consist of extracts of varying length. A corpus may contain single texts in single language (*monolingual corpus*) or text data in multiple languages (*multilingual corpus*).

Attempts of corpus creation are going on at different parts of the world and Corpus linguistics is emerging as a separate discipline. English being the language of Science and Technology takes a major share in the net text. English language also leads in the development of corpus of millions of words. The important source of creation of corpora may be from e-texts or the web. Some English corpora in use are British National Corpus (BNC), American National Corpus (ANC), European Corpus Initiative Multilingual Corpus I (ECI/MCI), EMILLE/CIIL, LOB, Brown Corpus, etc. The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources designed to represent a wide cross-section of British English.

The first corpus in Indian languages is the *Kolhapur Corpus of Indian English (KCIE)* developed by S.V Shastri (1988) and his colleagues at Shivaji University, Kolhapur. It contains approximately one million words of Indian English. The corpora contain large amount of texts in machine-readable form of whole texts or collection of different genre of text from different fields. The Central Institute of Indian Languages (CIIL, Mysore) has initiated the project on development of corpus for Indian languages in the late eighties. CIIL has developed corpora for different Indian languages. The corpus for each language is of 3 million lexical items which has been developed by different agencies. A corpus consisting of 2,350,000 words is made available for Malayalam. Any other major attempt is not reported so far for the language. They are neither widely representative nor balanced, nor even large with database. Also there has never been any attempt to make error free or enlarge that corpus with regular augmentation of new text samples from various text types and sources.

The Malayalam corpora developed so far are either by using keyboard or OCR method. Now it is the turn of web. There are so many advantages with web as a source than the earlier time. Text quantity is immense; it gets representation from almost all domains, time required is less, less errors, and so on.

II. USES OF CORPORA

The study of Linguistic phenomena through large collections of machine-readable texts of corpora is known as Corpus Linguistics. It is the study and analysis of data obtained from a corpus. Corpora are multifunctional resources. Corpora have uses in

Linguistics, Natural Language Technology and in Natural Language Processing. One can retrieve appropriate information from corpora to employ in various linguistic studies such as lexicography design, writing grammars, semantic study of words, pragmatic analysis of texts, sociolinguistic study, discourse analysis, etc. (Leech and Fligelstone 1992: 129). Teachers, students, and researchers in a variety of fields (e.g. languages, business studies, law, medicine and engineering) use corpora for teaching materials, class room exercises, and making dissertations, etc. Software developers, engineers and programmers use corpora to develop reference tools, natural language processing applications etc. Corpora allow researchers not only to count categories in traditional approaches to language but also to observe categories and phenomena that have not been noticed before.

Corpus based approach makes it possible to identify and analyze complex patterns of language use, allowing the storage and analysis of a larger database of natural language than could be dealt with by hand. Corpus based studies attempt to categorize registers, dialects, styles or individual literary works in terms of their linguistic association patterns.

In Lexical Studies: Empirical data has been used in lexicography long before the discipline of corpus linguistics was invented. Samuel Johnson, for example, illustrated his dictionary with examples from literature and in the 19th Century the Oxford Dictionary used citation slips to study and illustrate word usage. A linguist who has access to a corpus, or other (non-representative) collection of machine readable text can call up all the examples of a word or phrase from many millions of words of text in a few seconds. Dictionaries can be produced and revised much more quickly than before, thus providing up-to-date information about language. Also, definitions can be more complete and precise since a larger number of natural examples are examined.

III. CORPORA AND GRAMMAR

Grammatical (or syntactic) studies have, along with lexical studies, been the most frequent types of research which have used corpora. Corpora make a useful tool for syntactical research because of:

- The potential for the representative quantification of a whole language variety.
- Their role as empirical data for the testing of hypotheses derived from grammatical theory.

Many smaller-scale studies of grammar using corpora have included quantitative data analysis (Schmied's 1993). There is now a greater interest in the more systematic study of grammatical frequency - for

example, Oostdijk and de Haan (1994a) are aiming to analyse the frequency of the various English clause types.

Since the 1950s the rational-theory based/ empiricist-descriptive division in linguistics has often meant that these two approaches have been viewed as separate and in competition with each other. However, there is a group of researchers who have used corpora in order to *test* essentially rationalist grammatical theory, rather than use it for pure description or the inductive generation of theory. The formal grammar is first devised by reference to introspective techniques and to existing accounts of the grammar of the language. The grammar is then loaded into a computer parser and is run over a corpus to test how far it accounts for the data in the corpus. The grammar is then modified to take account of those analyses which it missed or got wrong.

IV. CORPORA AND SEMANTICS

The main contribution that corpus linguistics has made to semantics is by helping to establish an approach to semantics which is objective, and takes account of indeterminacy and gradience. Mindt (1991) demonstrates how a corpus can be used in order to provide objective criteria for assigning meanings to linguistic terms. Mindt points out that frequently in semantics, meanings of terms are described by reference to the linguist's own intuitions - the rationalist approach that we mentioned in the section on Corpora and Grammar. Mindt argues that semantic distinctions are associated in texts with characteristic observable contexts - syntactic, morphological and prosodic - and by considering the environments of the linguistic entities an empirical objective indicator for a particular semantic distinction can be arrived at.

Another role of corpora in semantics has been in establishing more firmly the notions of **fuzzy categories** and gradience. In theoretical linguistics categories are usually seen as being hard and fast - either an item belongs to a category or it does not. However, psychological work on categorization suggests that cognitive categories are not usually "hard and fast" but instead have fuzzy boundaries. So it is not so much a question of whether an item belongs to one category or the other, but how often it falls into one category as opposed to the other one. By looking empirically at natural language in corpora it is clear that this "fuzzy" model accounts better for the data: clear-cut boundaries do not exist; instead there are gradients of membership, which are connected with *frequency* of inclusion.

V. CORPORA IN THE TEACHING OF LANGUAGES AND LINGUISTICS

Resources and practices in the teaching of languages and linguistics tend to reflect the division between the empirical and rationalist approaches. Many textbooks contain only invented examples and their descriptions are based upon intuition or second-hand accounts. Other books, however, are explicitly empirical and use examples and descriptions from corpora or other sources of real life language data.

Corpus examples are important in language learning as they expose students to the kinds of sentences that they will encounter when using the language in real life situations. Students who are taught with traditional syntax textbooks, which contain sentences such as- "Steve puts his money in the bank" are often unable to analyse more complex sentences such as- "The government has welcomed a report by an Australian royal commission on the effects of Britain's atomic bomb testing program in the Australian desert in the fifties and early sixties."

Apart from being a source of empirical teaching data, corpora can be used to look critically at existing language teaching materials. Kennedy (1987a, 1987b) has looked at ways of expressing quantification and frequency in ESL (English as a second language) textbooks. Holmes (1988) has examined ways of expressing doubt and certainty in ESL textbooks, while Mindt (1992) has looked at future time expressions in German textbooks of English. These studies have similar methodologies - they analyse the relevant constructions or vocabularies, both in the sample text books and in Standard English corpora and then they compare their findings between the two sets. Most studies found that there were considerable differences between what textbooks are teaching and how native speakers actually use language as evidenced in the corpora. Some textbooks gloss over important aspects of usage, or foreground less frequent stylistic choices at the expense of more common ones. The general conclusion from these studies is that non-empirically based teaching materials can be misleading and that corpus studies should be used to inform the production of material so that the more common choices of usage are given more attention than those which are less common.

Corpora have also been used in the teaching of linguistics. Kirk (1994) requires his students to base their projects on corpus data which they must analyse in the light of a model such as Brown and Levinson's politeness theory or Grice's co-operative principle. In taking this approach, Kirk is using corpora not only as a

way of teaching students about variation in English but also to introduce them to the main features of a corpus-based approach to linguistic analysis.

A further application of corpora in this field is their role in computer-assisted language learning. Recent work at Lancaster University has looked at the role of corpus-based computer software for teaching undergraduates the rudiments of grammatical analysis (McEnery and Wilson 1993). This software - Cytor - reads in an annotated corpus (either part-of-speech tagged or parsed) one sentence at a time, hides the annotation, and asks the student to annotate the sentence by him/herself. Students can call up help in the form of the list of tag mnemonics, a frequency lexicon or concordances of examples. McEnery, Baker and Wilson (1995) carried out an experiment over the course of a term to determine how effective Cytor was at teaching part-of-speech learning by comparing two groups of students - one who were taught with Cytor, and another, who were taught via traditional lecturer-based methods. In general the computer-taught students performed better than the human-taught students throughout the term.

VI. TEXT ENCODING AND ANNOTATION

In order to make the corpora more useful for doing linguistic research, they are often subjected to a process known as annotation. An example for annotating a corpus is part-of-speech tagging, or POS-tagging, in which information about each word's part of speech (verb, noun, adjective, etc.) are added to the corpus in the form of tags. If corpora is said to be unannotated it appears in its existing raw state of plain text, whereas annotated corpora has been enhanced with various types of linguistic information. Unsurprisingly, the utility of the corpus is increased when it has been annotated, making it no longer a body of text where linguistic information is implicitly present, but one which may be considered a repository of linguistic information. The implicit information has been made explicit through the process of concrete annotation.

For example, the form "pooyi 'go (past)" contains the implicit part-of-speech information "past tense verb" but it is only retrieved in normal reading by recourse to our pre-existing knowledge of the grammar of Malayalam. However, in an annotated corpus the form "pooyi 'go'" might appear as "pooyi 'go VZ", with the code VZ indicating that it is a past tense (Z) form of a lexical verb (VV). Such annotation makes it quicker and easier to retrieve and analyse information about the language contained in the corpus.

VII. FORMATS OF ANNOTATION

Currently, there are no widely agreed standards of representing information in texts; in the past, many different approaches have been adopted, some are more lasting than others. One long-standing annotation practice is known as COCOA references. COCOA was an early computer program used for extracting indexes of words in context from machine readable texts. Its conventions were carried forward into several other programs, notably the OCP (Oxford Concordance Program). The Longman-Lancaster corpus and the Helsinki corpus have also used COCOA references. Very simply, a COCOA reference consists of a balanced set of angled brackets (<>) which contains two entities:

- A code which stands for a particular variable name.
- A string or set of strings, which are the instantiations of that variable.

For example, the code "A" could be used to refer to the variable "author" and the string would stand for the author's name. Thus COCOA references which indicate the author of a passage of text would look like the following:

```
<A CHARLES DICKENS>
<A WOLFGANG VON GOETHE>
<A HOMER>
```

COCOA references only represent an informal trend for encoding specific types of textual information, e.g. authors, dates, and titles. Current trends are moving towards more formalised international standards of encoding. The flagship of this current trend is the Text Encoding Initiative (TEI), a project sponsored by the Association for Computational Linguistics, the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities. Its aim is to provide standardised implementations for machine-readable text interchange. The TEI uses a form of document markup known as SGML (Standard Generalised Markup Language). SGML has the following advantages:

- Clarity
- Simplicity
- Formally rigorous
- Already recognised as an international standard

In the TEI, each text (or document) consists of two parts - a header and the text itself. The header contains information such as the following:

- Author, title and date
- The edition or publisher used in creating the machine-readable text
- Information about the encoding practices adopted.

VIII. RECENT DEVELOPMENTS IN WEB AS CORPUS

Web or World Wide Web is the new resource identified for corpus collection. There are some major initiatives in this direction. WebCorp is the pioneer in this field. Creating corpora from web or web as corpus is a new and emerging area. Here the materials/data are immense, easily accessible and up-to-date representation of all types of texts and selection from all domains. The web is immense, free and available by a mouse click, hundreds and millions of words of texts can be used for all manner of language research. Language scientists and technologists are turning to web as source of language data, because it is the only language source for the type of texts one is interested. It can be stored from the web as source web as corpus.

WebCorp is a suite of tools which allows access to the World Wide Web as a corpus - a large collection of texts from which facts about the language can be extracted. The WebCorp interface is similar to the interfaces provided by standard search engines. You enter a word or phrase, choose options from the menus provided and then press the 'Submit' button. WebCorp works 'on top of' the search engine of your choice, taking the list of URLs returned by that search engine and extracting concordance lines from each of those pages - examples of your chosen word or phrase in context. All of the concordance lines are presented on a single results page, with links to the sites from which they came.

Despite the fact that a growing body of work has shown that the World Wide Web is a mine of language data of unprecedented richness and ease of access (Kilgariff and Grefenstette, 2003), many fundamental issues about the viability and exploitation of the Web as a linguistic corpus are just starting to be tackled, ranging from Web frequency distributions and registers to efficient handling of massive data sets and copyright. Research on the Web as corpus is currently at a very exciting stage: increasing evidences point to the enormous potential of Internet as a source of linguistic data, but it is still far from a working, fully-fledged linguists' search engine. The issues are discussed so far

and have to be discussed while developing web corpus. Describe web corpus collection projects or modules for one part of the processes (crawling, filtering, language-id, tokenising, lemmatising, POS-tagging, indexing), exploring characteristics of Web data, from a linguistics/NLP perspective use crawled Web data for NLP purposes.

The web text is immense, free and available just by a mouse-click. It contains hundreds of billions of words of text and can be used for all manner of linguistic research. Language scientists and technologists are increasingly turning to the web as a source of language data, because it is so big, because it is the only available

source for the type of language one is looking to and it is free and instantly available. Web is the source from where more language data at different texts can be collected. The source for the development of the corpus is primarily from the net and the e-text available in the net.

IX. MALAYALAM CORPUS - FROM WEB

Malayalam is one of the important Indian languages, which have got more web publication than any other regional languages. Malayalam produced hundreds of websites not only on the language and the literature but also in Malayalam scripts itself. Earlier texts were prepared in different codes – ASCII, ISCII, etc. and with different text editors. In web different codes with dynamic font or in separate font with downloading facility were used. But today because of the Unicode system large number of texts are prepared in it and uploaded in its original script.

Almost all leading dailies in the language have their online editions and about a hundred journals or weeklies got publication through the web. It can also be noticed that the dailies go for four to five editions or updating in a day. Most of the dailies and weeklies keep their archives for quite some time so that a user can access without much difficulty. Kerala Govt. sites, portals, commercial portals, sites belonging to educational institutions, Govt. publications, blogs and personal files/sites etc. make the web source extensive. Source for corpus making can also be the digital texts prepared by different agencies either to publish or for other purposes.

The practical issues when attempting to prepare a corpus for Malayalam can be divided into two. One is related to text, encoding issues, font and code, text format, etc. i.e., technical issues and the second one is language specific, i.e. related to spelling, spacing, borrowed lexical items, etc.

The steps involved in building a corpus are selection of texts, data entry, data validation and a set of tools for management and retrieval of data. The text can be collected from different sources such as from web (Malayalam sites), newspapers, journals, magazines, medical reports, autobiography, government reports, diaries, notices etc. While collecting texts of Malayalam from web, categorization of them is an important and necessary part of documentation. The varieties of language we use in different situations referred as registers. (e.g., Medicine, literature, astrology) are categorized into different levels and the documents

copied from sites are systematically added to our corpus under the specified register. This saves time, data duplication, memory space etc. Data collection will also be easy. Finally it can be traced that all Malayalam documents are not incorporated from the web sites or from all particular domain.

The primary hurdle is that different agencies are using different softwares and codes for developing e-texts. Non-standardization of fonts or non-uniformity in spelling system is another key issue. Perhaps one text itself may put/see in different spelling system. The text from different sources show in different fonts and spelling systems is an important issue for a corpus developer. Those texts collected from different sources, which are in different fonts and spelling systems, need to be carefully handled. One remedy for this is that the texts with different fonts have to be converted to a single or uniform font in Unicode system.

Lammatiation or identification of lexical items from compounds and complex constructions is another vital issue to be carefully looked into. Parts of speech identification for noun and verb also require additional attention. Space between lexical items, as in that of languages like English, Hindi, etc. cannot always been applied for Malayalam text. Position of lexical items with grammatical function expects that of finite verb, is another significant issue.

Following can be listed as some key issues with respect to text corpus:

1. Non standardization of spelling – difference in spelling system
2. Font differences, including that of half consonants (chillu)
3. Non spacing between lexical items
4. Sandhi changes
5. Borrowed lexical items
6. Part of speech differences
7. Position class of lexical items in different context

The issues related to web are:

1. Problems relates with display
2. Banners / advertising
3. Pop up screens
4. Images and text as images
5. Different formats (HTML, PDF, XML, etc.)
6. Encoding issues

The concern for the time being is a general monolingual text corpus of the language containing data from all subject domains, texts, genres, types, and walks of life. This corpus will be properly analysed to be used for word level tagging, word count, letter count, frequency count, spell checker development, etc.

Here is a text taken from an Internet edition of a newspaper. Fig.1 is gibberish display of the text. If we go to some site and try to get the display of it, we get a rubbish display.

©Ojj¼ dj£ÈJw

Same Malayalam text in a different font in language script (*MLW-TTKarthika*)

[illegible]

1 വേലയിൽ വിളയുന്ന വിദ്യാഭ്യാസം (മഹാത്മാഗാന്ധി - 1869-1948, രാഷ്ട്രപിതാവ് ആദർശങ്ങൾ കർമ്മഫലത്തിലെത്തിച്ച സമാൻ. ജീവിതത്തിന്റെ എല്ലാ മേഖലകളെയും സമഗ്രമായി വീക്ഷിക്കുന്നു 'എന്റെ സത്യാന്വേഷണ പരീക്ഷണങ്ങൾ' അദ്ദേഹത്തിന്റെ ആത്മകഥയാണ്. ഗാന്ധിജിയുടെ ചിന്തകളും ദർശനങ്ങളും സമാഹരിച്ച് മലയാളത്തിൽ ഏഴു ഭാഗങ്ങളായി 'ഗാന്ധിസാഹിത്യം' എന്നപേരിൽ ഒരു ഗ്രന്ഥപരമ്പര പ്രസിദ്ധീകരിച്ചിട്ടുണ്ട്. ഒരു നവസാഹിത്യത്തിലൊരപ്തം

Preference will be given to projects where Web data are downloaded and processed directly, rather than via search engine interfaces. The required data can be collected from the Net or e-texts, may be by manual selection and store or automatically through a search

program to store. Malayalam works are being developed and printed in abroad also. Thousands of literary and other works published and printed regularly have a literature of world standards. Malayalam produces more than any other regional language in the country. Cyber Malayalam is more than in print, and that include text types like books, newspapers, journals, blogs, portals, etc.

A model of data collected and saved in *.doc format.

<Site address: <http://www.malayalamresourcecentre.org/Mrc/Karshikam/rubber/rubber1.html>, Root: Kaarshikam/rubber, Author: C-DAC, Date: 03/05/07, Remark:

The agriculture of Kerala describing about rubber, cardamom, cashew nut, clove, coconut, paddy, tea, and pepper.>

ഭാരതത്തിലെ റബ്ബർതോട്ടവൃവസായം ശതാബ്ദിയിലെത്തി നില്ക്കുകയാണിന്ന്. പത്തുലക്ഷത്തോളം വരുന്ന റബ്ബർ കർഷകരിൽ എണ്ണപത്തേഴുശതമാനത്തോളം ചെറുകിട കർഷകർ-ഭാരതത്തിലെ റബ്ബർ കൃഷി മേഖലയുടെ മാത്രം പ്രത്യേകതയാണിത്. ആഗോള വല്കരണത്തിന്റെ ഭാഗമായി സംജാതമാകുന്ന പരിതസ്ഥിതികൾ തരണം ചെയ്യുന്നതിനും എന്ന്നെന്നും റബ്ബർതോട്ട വൃവസായത്തെ സാധാരണ കർഷകന്റെ ആശങ്കകേന്ദ്രമായി നില നിർത്തുന്നതിനും റബ്ബർകൃഷിയുടെ ശാസ്ത്രീയത പ്രയോഗിക തലത്തിൽ എത്തിച്ചേരുന്നതിനും ഉല്പാദനചെലവു ചുരുക്കിയും ഉല്പാദനക്ഷമത വർദ്ധിപ്പിച്ചും ഉല്പന്നത്തിന്റെ ഗുണ മേന്മ ഉറപ്പുവരുത്തിയും മാത്രമേ കൃഷിയിടങ്ങളെ ആദായകരമാക്കാൻ സാധ്യമാകുകയുള്ളൂ. ഭാരതത്തിലെ റബ്ബർ കൃഷി ഇന്ന് ഒരു പരീക്ഷണഘട്ടത്തിലെത്തിയിരിക്കുകയാണ്. മറ്റു നാണുവിളകളുമായി തട്ടിച്ചുനോക്കിയാൽ ഉല്പാദനക്ഷമതയിലും ആദായത്തിലും ഉന്നതമായ സ്ഥാനം റബ്ബർകൃഷിക്കുണ്ടെങ്കിലും ഉല്പാദനഗുണനിലവാരവർദ്ധനവിലൂടെ വരുമാനമുയർത്താൻ ഇനിയും സാധ്യത ധാരാളമുള്ള കൃഷിയാണ് റബ്ബർ. പ്രധാന റബ്ബർ ഉല്പാദക രാഷ്ട്രങ്ങളുമായി തട്ടിച്ചുനോക്കിയാൽ ഉല്പാദനക്ഷമതയിൽ നാം ഒന്നാം സ്ഥാനത്താണ്. ഭാരതത്തിൽ റബ്ബറിന്റെ ഉല്പാദനക്ഷമത 1567 കിലോഗ്രാമാണ്. എന്നാൽ 3500 കിലോഗ്രാംവരെ ഉല്പാദനക്ഷമതയുള്ള തോട്ടങ്ങൾ ഭാരതത്തിൽ ധാരാളമുണ്ടെന്നതാണ് വസ്തുത. ഈ വസ്തുത കണക്കിലെടുത്താൽ ഇന്ത്യയിലെ ഉല്പാദനക്ഷമത 1567 കിലോഗ്രാമിൽ നിന്നും 2500 - 3000 കിലോഗ്രാമായി വർദ്ധിപ്പിക്കാൻ സാധിക്കുമെന്നത് നിസ്തർക്കമായ സംഗതിയാണ്.

(MLW-TTKarthika)

XI. ISSUES IN THE DESIGN PHASE OF MALAYALAM CORPUS

First stage in Malayalam corpus building is the collection of data from more than one source or from one genre. So the corpus builder should have a clear view about the categories and subcategories of the texts, which are to be collected.

- Non standard / different fonts, Text in different format(HTML, PDF, etc.)
- Banners, popup screens, advertisements
- Text as images
- Spelling errors
- Alignment issues

Classification of texts belonging to different domains is one of the main issues related to the classification of texts. Subsections of texts classified by date, subject-matter, region, age-group, sex, type, style, register, etc. are also to be considered seriously.

Some of the Malayalam sites that can be used for data collection:

Online newspapers

1. www.keralakaumudi.com
2. <http://www.manoramaonline.com/cgi-bin/MMOnline.dll/portal/ep/home.do?newUser=yes>
3. <http://www.thejasonline.com/java-thejason/index.jsp>
4. <http://www.deshabhimani.com>

5. <http://www.deepika.com>
6. <http://www.keralaonline.com/>
7. <http://www.mathrubhumi.com/>
8. <http://www.mangalam.com/>
9. <http://sify.com/malayalam/>

Periodicals

1. <http://www.vellinakshatram.com/934/index.asp>
2. <http://www.vellinakshatram.com/kala/1655/index.asp>
3. <http://www.vellinakshatram.com/fire/168/index.asp>
4. <http://www.manoramaonline.com/cgi-bin/MMOnline.dll/portal/ep/home.do?tabId=5>
5. <http://www.jaalakam.com/>
6. <http://www.weblokam.com/>
7. <http://www.malayalamvarikha.com/>
8. <http://www.vellinakshatram.com/muhurtham/57/index.asp>
9. <http://www.vellinakshatram.com/ayurarogyam/58/index.asp>
10. <http://www.vellinakshatram.com/snehitha/77/index.asp>

Other useful sites

1. <http://www.tapioca.in> (Malayalam search engine)
2. <http://www.puzha.com/>
3. <http://www.chintha.com/malayalam/blogroll.php>

XII. CONCLUSION

A corpus is not simply collection of texts; rather a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus, therefore, depends upon what it is meant to represent. A corpus can be developed from web as a source. There are a lot of advantages in web as a source. Similarly a number of issues are to be solved for creating corpus from the web.

Some of the notable advantages of web as a source for corpus development are:

- New and emerging area
- Text/ data is immense, free and available by the click of a mouse
- Easily accessible
- Up-to-date representation
- Selection from all domains
- Can be stored directly from the web with or without modification
- Web can also be treated as corpus
- The web text in different formats need to be unified
- Hundreds and millions of words, texts
- Language scientists and technologists are more interested
- Because it is the only language source for the type of texts one is interested
- Access to web corpus is ubiquitous and size exceeding all previous corpuses.

The major issues that are creating hurdles for creating corpus from the web are:

1. Non standardization of spelling – difference in spelling system
2. Font differences, including that of half consonants (chillu)
3. Non spacing between lexical items
4. Sandhi changes
5. Borrowed lexical items
6. Part of speech differences
7. Position class of lexical items in different context

The issues related to web are:

1. Problems relates with display
2. Banners / advertising
3. Popup screens
4. Images and text as images
5. Different formats (HTML, PDF, XML, etc.)
6. Encoding issues

Malayalam corpus can be developed from the web as a basic resource for language analysis and research. Some of the above issues can be solved easily and some others require much research and analytical study. This is high time for us to turn our attention towards web as corpus or web as source for corpus, and of course, web is the real resource or mine for such development of our linguistic research and description.

REFERENCES

- [1] Biber, D., S. Conrad and R. Reppen 1998. "Corpus Linguistics: Investigating Language Structure and Use" Cambridge: Cambridge University Press.
- [2] Bharati Akshar, Chaitanya Vineet, Sangal Rajeev 1996 Natural Language Processing- A Paninian Perspective, New Delhi : Printice - Hall
- [3] Dash, N.S., and B.B. Chaudhuri. 2000. "The process of designing a multidisciplinary monolingual sample corpus." International Journal of Corpus Linguistics. 5(2): 179-197.
- [4] Dash, N. S., "Language Corpora: Present Indian Need." Indian Linguistics.
- [5] Dash, N.S. "Utilization of Corpora in compilation of a monolingual general dictionary in electronic form"
- [6] Kilgariff, Adam "Web as Corpus" Computational Linguistics Vol. V No. N
- [7] Orasan Constantin, Ramesh Krishnamurthy, UK "An Open Architecture for the Construction and Administration of Corpora"
- [8] Ruslan Mitkove, The Oxford Handbook of Computational Linguistics, Oxford: Oxford Press, 2001.
- [9] Report of the Committee on Malayalam Character Encoding and Keyboard Layout Standerdisation, Govt. of Kerala, 18th December 2001.
- [10] Volk Martin "Using the Web as Corpus for Linguistic Research", article from web.
- [11] www.webcorp.org
- [12] www.kilgariff.co.uk

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.