CHAPTER IV
CORPUS FOR MMD

Introduction

This chapter will discuss about the corpus in computational as well as general term using FSA and RE, and its requirements while building a multilingual dictionary i.e. MMD. As the Manipuri language has no corpus based on Meitei Mayek script we need to develop one parallel mini corpus, based on few principles of how corpus can be built using the best features and best characteristics. This chapter will discuss about the different types of corpus and also will discuss how to choose the best corpus suited for MMD and its main advantages. It will also discuss the methods and methodology to develop the corpus for MMD and analyze regarding the applications areas of corpus in NLP and computational linguistics areas.

4.1 Corpus

A Corpus (the plural is usually Corpora) is a collection of text, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database. Most Corpora contain information in addition to the text that make them up, such as information about the texts themselves, parts-of-speech, tags for each word, and parsing information. Corpus building mainly has two stages, design and implementation, but these cannot be completely separated, for reasons, which are largely practical. One is the cost i.e. putting the text into electronic format. In computational and Finite Automata theory a Corpus can be defined as a finite-state automaton M that have five tuples

Where

$M= (Q, \Sigma, \delta, q_0, F)$

Q = A finite set of States.

Σ = A finite Input Alphabets of Manipuri i.e ꯀ ꯁ ꯂ ꯃ ..ꯑ

q_0 = A set of Final Accepting states, which is subset of Q.

δ = Transition Function, which is a total function from $\delta (Q \times \Sigma)$ *to* Q

$\delta:(Q \times \Sigma)$ ----> Q

δ is defined for any q in Q and δ in Σ and is equal to some state q in Q, would be $q'=q$.

Transition diagrams similar to those used for representing finite-state transducers can also be used to represent finite-state automata. The only difference is that in the case of finite-state automata, an edge that corresponds to a transition rule (p, , p) is labelled by the string. The diagrams below show how input can be accepted by the FSA and how output comes out from the machine.

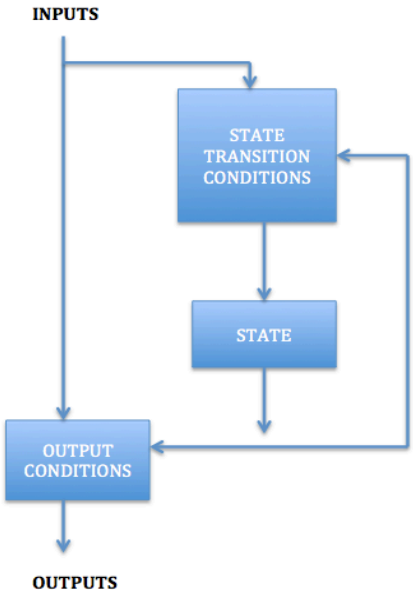

Figure 25: FSA Diagram of Corpus

4.1.1 Regular Expression in Corpus

1. Union

Union of two Automata A and B accepts all strings of A and all string of B. Denoted by A U B

Example: ಶಕೆಯ U ಶಖೆ= ಶಕೆಯಶಖ (chaam + tara = chaamtara)

2. Concatenation

The concatenation of two automata A and B accepts all string that are concatenation of two string, the first one being accepted by A and second by B. denoted by A in B. obtained by connecting all the final state of A to the initial state of B using E transition.

Example: ಛಬೆಲ + ಛ = ಛಬೆಲಛ (Thawan + tha = thawantha)

3. Iteration

Closure of an Automaton A accepts to generate the concatenation of any number of its string and the empty string E. It is denoted by A* A= {E, A, AuA, AA, AAAu...}. It can also be obtained by listing the final state of A to its initial state using e transition.

Example: ꯏꯟꯁꯪ * ꯊꯣꯡꯕ= {ꯏꯟꯁꯪ, ꯀꯥꯡꯍꯧ, ꯏꯔꯣꯝꯕ...}

(Enshang thongba ={enshang, kaanghou, eromba...})

4. Intersection

The intersection of two automatons A∩ B or the intersection of two words is accepted by corpus. One such example is given below in Manipuri words.

Example. ꯑꯍꯤꯡ ꯇꯨꯝꯗꯕ ꯌꯥꯔꯦꯛꯄ ∩ ꯑꯍꯤꯡ ꯌꯥꯔꯦꯛꯄ = ꯑꯍꯤꯡ ꯌꯥꯔꯦꯛꯄ

Ahing tumdab ∩ ahing yaarekpa= ahing yaarekpa

5. Difference

The corpus will accept the difference of words or the automaton of A-B, which means the string, accepted by A but not by B.

Example: ꯏꯁꯤꯡ - ꯃꯥꯏꯁꯤꯡ = ꯏꯁꯤꯡ

eshing – maishing = eshing

It is also true that Corpus software presents the researcher with language in a form that is not normally encountered and that this can highlight pattern that often goes unnoticed. Corpus linguistics have also led to a reassessment of what language is like. A corpus can exist only in raw text format i.e. plain text with no additional information. A corpus is (**C**) apable (**O**) f (**R**) epresenting (**P**) otentially (**U**) nlimited (**S**) elections of texts, it may also be defined as compatible to computer, operational in research and application, representative of the source language, process abled by both man and machine, unlimited in amount of data, and systematic in formation and representation.

A corpus is a remarkable thing, not so much because it is a collection of language text, but because of the properties that it acquires if it is well-designed and carefully constructed. The term corpus, derived from Latin, usually refers to a body of texts (collection of linguistics data) either in written or spoken form (transcribed recorded speech). It is a representative sample of different varieties of language preserved in machine-readable form, which can be used as a starting point of linguistic description or a means for verifying hypotheses about a language (Crystal, 1980).

According to Sinclair (1996) 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 language. Similarly, Mcenery Tony Y Wilson, Andrew (Mcenery, Tony Y Wilson, Andrew, 1996) is of the opinion that the term corpus should refer to:

(i) (loosely) anybody of text:

(ii) (most commonly) a body machine readable text

So whatever may be the point of discount regarding representational parameters, we can assume that a corpus, designed methodically, should have the following characteristics features.

1. It should be large in size containing a healthy amount of language data.
2. It should be authentic and reliable in representation of language.
3. It should consist of structured collection of text specifically compiled.
4. It can be either a simple plain text or a text with annotation.
5. It should be user-friendly for easy handling.
6. It should be properly and systematically documented.
7. It should have authentic referential value.

4.2 Motivation

Parallel Corpora are being used for a large range of applications in Natural Language Processing. Now-a-days, most typically monolingual research areas are benefiting from parallel corpora usage. To help in working with big amounts of parallel corpora we propose a client-server approach to store them, and make it accessible for query.

While developing MMD Corpus we had the following points in mind:

1. Be able to query for concordances, both simple and parallel: search for a word or pattern, in the source or target language, or in both at the same time.
2. Be able to query Probabilistic Translation Dictionaries This way we can relate words from corpora translation units.

3. Be able to query more than one corpus at the same time, and with different languages. Also, to be able to query for corpora meta- information like involved languages, number of translation units and other.

4. Support big corpus, more than one million-translation unit. Most studies are statistics, and the result's precision highly depends on the corpus size.

5. Fast for interactive and batch tasks, which lead us to a double architecture:

    1. Reduce loading time for indexes and dictionaries when using them interactively, like in a web-based application (we do not want the user to wait a long time for the answer or for the web server timeouts). With this in mind, there is a server, which loads all the information just once reducing the load time.

    2. Reduce the overhead time for the communication. For batch processes, which query repeatedly the same corpus, it is better to load the corpus indexes and query them in memory. In these cases, the load time overhead is too small compared to the overall time of the process.

6. Easy to distribute work:

    1. For big corpora we can split them in small chunks and make them available for making querying from different servers.

    2. Different applications or users are able to query the same corpus in the same server, thus reducing the need for replication. For instance, for the implementation of distributed translation memories (Simo˜es, Guinovart, and Almeida, 2004).

7. Be an open-source and free tool. While there are some applications to manage corpora like Corpus Query Processor (Ko¨nig, 1999) or SARA for the BNC corpus (Dodd, 1997) they are not freely available. SARA is commercial software and CQP is just available for research with license limitations. There are some web-based tools like TransSearch (RALI Laboratory, 2006) and COMPARA (Frankenberg- Garcia and Santos, 2003) to query corpora. The first

is paid. COMPARA is freely available but uses as backend CQP. To be an open- source tool is important for other researchers to be able to enhance the program in accordance with their needs, to compare times, and others.

8. Prepare a simple API to write server clients with few lines.

9. Develop a set of real web-clients to test and validate the tool, and to test the API.

10. Develop multilayer support: add levels of information related to corpora words. The base layer includes the words or tokens that constitute the corpus. Other layers add information like lemmatization or parts-of-speech.

4.3 Basic Principles of Corpus Building

The principles, which follow for building a corpus of Manipuri language using Meitei Mayek, are listed below:

1. The contents of a corpus should be selected regardless of the language they contain, but according to their communicative function in the community in which they arise.
2. Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.
3. Only those components of corpora, which have been designed to be independently contrastive, should be contrasted.
4. Criteria for determining the structure of a corpus should be small in number, clearly separate from each other and efficient as a group in delineating a corpus that is representative of the language or variety under examination.
5. Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in application.
6. Samples of language for a corpus should wherever possible be of entire document or transcription of complete speech events, or should get as close to this target as possible. This means that sample will differ substantially in size.

7. The design and composition of a corpus should be documented fully with information about the content and arguments in justification of the decision taken.
8. The corpus builder should retain, as target motion, representativeness and balance. While these are not precisely defined and attainable goals, they must be used to guide the design of a corpus and the selection of its components.
9. Any control of subject matter in a corpus should be imposed by the use of external and not internal criteria.
10. A corpus should aim for homogeneity in its component while maintaining adequate coverage, and rogue text should be avoided.

4.4 Methodology

Choice of text types should reflect the need of users. So, we carried out a survey of language teachers and language engineers to get their opinions on the texts that might be of use to them (Al-Sulaiti & Atwell, 2003). The choice of text types was based partially on a survey conducted at the University of Leeds to find out the most frequently translated text types for the purpose of compiling a multilingual corpus for machine translation evaluation (Elliot et al, 2003). In this regard we are following the guidelines used in other existing corpora such as the BNC. The texts belong to the following major categories:

Written: Fiction – Arts – Science – Business – Miscellaneous

Spoken: TV – Radio – Conversation

4.5 Characteristic Features of MMD Corpus

A corpus here, Parallel Corpora must assume to have certain characteristics attached to it. MMD (Multilingual Manipuri Dictionary) is following the basic characteristics, which all general corpuses have and which consist of the following features:

1. Quantity: It means that a corpus should be very big in size containing large amount of language data either in spoken or written form. Size is virtually the sum of its components, which constitute its body.

2. Quality: It means ‘authenticity’. All materials should be gathered from genuine normal use of speech and writing.

3. Representation: It should include language samples from a broad range of materials. It should be balanced in all disciplines and subjects to represent maximum number of linguistic features of a language. Future analysis and study devised on it will need verification and authentication of information from the corpus representing the language.

4. Simplicity: It implies that corpus should contain simple plain texts so that the user can expect an unbroken string of characters without any additional linguistic information marked-up within the text. A simple plain text is virtually opposed to any kind of annotation with various linguistics and non-linguistic information.

5. Retrievability: Information, examples, and references should be easily retrievable from corpus by the end-users. This pays attention to the preserving techniques of the data in the electronic format in the computer for the users. The present technology has made it possible for the users to generate corpus in a PC and preserve it in such a way that one can easily retrieve the data when required.

6. Verifiability: Corpus should be subject to any kind of empirical analysis and verification. Anybody can use the database for any kind of investigation either to verify or refute the observations made by intuitive estimation. This advantage puts corpus linguistics a step ahead of generative model of language study.

7. Augmentability: It should increase regularly. This will put corpus 'at par' to register linguistic changes noted within a language in course of time. Over time, by regular addition of new linguistic data, corpus will attain a historical dimension to preserve information for diachronic linguistic studies, and to display linguistic clues to arrest changes in life and society.

8. Documentability: Full information of components should be kept separately from the component itself. It is better to keep the documentation of texts in a separate place from the texts themselves, and to include only a minimal header that contains a reference to the documentation. For corpus management, this practice allows effective separation of plain texts from annotation with only a small amount of programming effort**.**

4.6 Domain Types

Different Domain types and its sub-domain are listed below for MMD (Multilingual Manipuri Dictionary). Eighteen (18) different Domains are categorized and each domain has sub categories.

a) **Geography :** in this domain historical geography, graphic representation of earth, ancient world, Europe, Asia, Africa, North America, South America, others areas and miscellaneous are the sub-categories.

b) **Literature:** in this domain history and development, fiction, poetry, drama, essays, letters, satire and humour, literary figure and popular characters, Indian literature in English, Indian literature in Indian language, folk literature and miscellaneous are included.

c) **Society and community:** community welfare service and association, marriage, relationship and kinship, housing and household equipment, housekeeping and décor, child caring, area planning, landscape architecture, transportation and traffic, water feature, public structure, residential and related building, language, communication and miscellaneous are sub-categories.

d) **Law:** International law, criminal law, cyber law, private law, religion and law, personalities, constitutional and administrative law, military tax trade industrial law, social, labour, welfare and related law, civil procedure and court, law(status), regulations cases, law of specific, jurisdiction and area and miscellaneous are sub types of law.

e) **National Security and Defence:** history, personalities, police and internal security, Indian defence service, warfare methods and equipment's, water technology, military law, International relations and miscellaneous.

f) **Religion:** festival, rituals and practices, gods, religious leaders saints and reformers, religious text, history and development in religions, mythology, pilgrimages, religious art and architecture, ancient religious, spirituality and miscellaneous are sub types of domain religion.

g) **Philosophy:** sub categories are movements, philosophers, theories and school of thought, writing scripture and miscellaneous.

h) **History:** monuments, wars, civilizations, museums, historical figures, archaeology, colonisations, Indian history, world history and historical and archaeologist are sub types of history.

i) **Economy:** different types of economy are employment, domestic and foreign trade, Industries, e-commerce, miscellaneous, economic policy and law, banking, personalities/economist, history, labour economics, financial economics, cooperatives, socialism and related system, public finance, International economic, production, macroeconomics and related topics and home economics

j) **Sports:** sports and entertainment, sport events, indoor and outdoor games, milestones and records, sportsperson, traditional games and miscellaneous are sub types of sports.

k) **Health:** blood, heart and circulation, bones joints and muscles, brain and nerves, digestive system, kidneys and urinary system, lungs and breathings, oral and dental, skin hairs and nails, female reproductive system, male reproductive system, life style, paediatrics, medical events and camps and miscellaneous are categorised as sub types of health.

l) **Society and educations:** equality discriminations and reservations, urban and rural education, education in India and abroad, career guidelines, academic, researchers, miscellaneous, elementary education, secondary education, higher education, special educations, Government regulations, control support and education institutions and management are the sub types of Society and educations.

m) **Tourism:** pilgrimage, ecotourism, heritage tourism, adventure tourism, mass travel and tour, treasure tour, medical tourism, nautical tourism, culinary tourism, disaster tourism, dark tourism, space tourism, shopping tourism, war tourism , general description and miscellaneous are tourism sub types.

n) **Art and Culture:** classical performing arts, crafts and handicrafts, fine arts, costume and personal appearance, cuisines, miscellaneous, history of art forms, personalities, folk tradition, customs, photography, recreational activities, graphic and print making and textile art are different sub types of art and culture.

o) **Science and Technology:** sub categories are mathematics, engineering and allied operation, Botany and zoology, Bio-science and life science, Earth science and climate, personalities, para science, discoveries and inventions, miscellaneous, history, personalities, Astrology, natural science, natural history, physics, Chemistry, allied science, palaeontology and paleozoology, technology, architecture, ecology, paranormal science, neuroscience and Psychology and other science.

p) **Politics:** Constitutions, Fundamental right , justice, Governance, Democracy, miscellaneous, civil and political movements, policies, political leaders, bills and ordinance, elections, history, diplomacy, central government and local government are sub types of politics.

q) **Agriculture:** kinds of farming, agricultural economics, agricultural marketing, agriculture machinery and equipment, agricultural research, grains and land management, crop production, Indian agri- events, Indian role in world agriculture, agriculture policy, agricultural organisation in India, agriculture and technology, weather and agriculture, personality related to agriculture, agriculture and its uses and miscellaneous are the sub types of agriculture.

r) **Entertainment:** Film: history and development, film: scripts, film: review, film: news, film: personalities, performing art: news, performing arts: personality, media: general, media: history and development, media: news, media: personality, literature: history and development, literature: story, literature: personality, miscellaneous, culture: general , culture: festival, culture: lifestyle, culture: cuisines and multimedia, animation and graphics are the sub types of entertainment.

4.6.1 Methods of Data Input

Data from electronic sources: In this process texts from newspapers, journals, magazines, books etc. are included if these are found in electronic form.

Data from the websites: This includes texts from web pages, web sites, and home pages, not for target language.

Data from e-mails: Electronic typewriting, e-mails, etc. are also used as source of data not for Manipuri language.

Machine reading of text: It converts printed texts into machine-readable form by way of optical character recognition (OCR) system. Using this method, printed materials are quickly entered into a corpus.

Manual data input: It is done through typing texts in computer. This is the best means for data collection from hand-written materials, transcriptions of spoken texts, and old manuscripts. The method is applying for MMD corpus. The process of data input is indirectly based on the method of text sampling. We can use two pages after every ten pages from a book. This makes a corpus best representative of data stored in a physical text. For instance, if a book has several chapters, each chapter containing different subject matters written by different writers, then the text samples collected in this process from all chapters are properly represented. Each text file should have a Header which contains metadata – the physical information about the texts such as genre of the text (e.g., literature, science, commerce, technology, engineering, etc.), type of text (e.g., literature, story, travelogue, humor, etc.), sub-type of text (e.g., fiction, historical, social, biographical, science fiction, etc.), name of book, name of the author(s), name of the editor(s), year of publication, edition number, name of the publisher, place of publication, number of pages taken for input, etc. This information is required for maintaining records and dissolving copyright problems. It is also advantageous to keep detailed records of the materials so that the texts are identified on grounds other than those, which are selected as formatives of corpus. Information whether the text is a piece of fiction or non-fiction, book, journal or newspaper, formal or informal etc. are useful for both linguistic and non-linguistic studies.

4.6.2 Meta Data for Domains

Corpora Source

1. Books: must include the Name of Books, Name of Author/Editor, Name of Chapter/article, Page number, Name of Publisher, Year of Publishing, Place of Publishing.

2**. Magazine:** Name of Magazine, Name of Articles, Page number, Name of Author, Date/ Volume of Magazine, Place of issue.

3. Movie: Name of Movie, name of Director, date of Release, additional Information.

4. Newspaper: page no, name of author, date of issue, place of Issue.

5. Web Source: name of website, name of article, name of Author, date posted, website URL, date retrieved.

At the time of input of text, the original text of the physical source must be kept unchanged. After a paragraph is entered, one blank line should be given before a new paragraph starts. When texts are collected in a random sampling manner, a unique mark or flag is needed to be posted at the beginning of a new sample of text.

4.7 Types of Corpus

a) **Sample corpus:** A fixed sample of text, often used as a reference corpus for comparing.

b) **Monitor Corpus:** A corpus which develops and is added to or filtered depending on the researchers needs**.**

c) **Mini-Corpus:** a small corpus (e.g. to be compared with a reference corpus)

d) **Multilingual Corpora:** Corpus in a variety of language.

e) **Comparable Corpus:** Text in two language or two language varieties but not matched up.

f) **Parallel Corpus:** Text is translation of each other, e.g. Manipuri Hansard, Corpus of Version of Plato, and Bible.

g) **Translation Corpus:** two or more set of text classified as either originals or transition, the purpose being to identify features of translation (Manchester: Baker)

h) **Diachronic Corpus:** Helsinki, LOBA V. FLOB

i) **Learner Corpus:** texts are written by language learner.

4.8 Corpus Types with Regards to User

Types of corpus classified according to the user of target user are given below in table no. 14.

Table 14: Type of corpus users and their needs with regard to the type of corpus

Target users	Corpus
Descriptive linguists	General, written, and speech corpus
NLP and LT people	General, monitor, parallel, spoken, aligned corpus
Speech technology people	Speech corpus and spoken corpus
Lexicographers and terminologists	General, monitor, specialized, reference, opportunistic corpus
Dialogue researchers	Speech, spoken, annotated, specialized corpus
Sociolinguistics	General, written, speech, monitor corpus
Psycholinguistics	Specialized, speech, written corpus
Historians	Literary, diachronic corpus
Social scientists	General, speech, written and special corpus
Comparative linguists	Bilingual, multilingual, parallel, comparable corpus
MT specialists	Bilingual, multilingual, parallel, comparable, annotated corpus
Information retrieval specialists	General, monitor, and annotated corpus
Tagging, processing and parsing specialists	Annotated, monitor, written, spoken, general corpus
Core-grammar designer	Comparable, bilingual, and general corpus
Word-Sense disambiguation worker	Annotated, monitor, written, spoken, general corpus
Teachers and students	Learner, monitor, and general corpus
Linguists	All types of corpus

4.9 Types of corpus - types of tool

As pointed out by the Text Encoding Initiative (TEI26, 1993), the term language corpus is used to mean a number of rather different things. However, for TEI, the only distinguishing feature of a corpus that really matters is that its components

have been selected or structured according to some conscious set of design criteria. A similar corpus definition is given by Atkins & al. (Atkins & al, 1992) that is partly building on earlier work by Quémada, distinguish four types of machine-readable text collection:

1. ***archive***: a repository of readable electronic texts not linked in any coordinated way;

2. ***electronic text library*** *(ETL)*: a collection of electronic texts in standardized format with certain conventions relating to content etc., but without rigorous selection constraints;

3. ***corpus*****:** a subset of an ETL, built according to explicit design criteria for a specific purpose; and

4. ***subcorpus***: a subset of a corpus, either a static component of a complex corpus or a dynamic selection from a corpus during on-line analysis.

The tools used for pre-processing and the tools used for analysis must be compatible, by which it is meant, among other things, that the analysis tools must conform to the format of the pre-processed texts and be able to make use of the complementary information of the annotations. Useful guidelines for defining such formats can be found in the SGML standard as it is utilized by the text encoding Initiative (TEI P2, 1992).

4.10 Parallel Corpora

A *parallel corpus* is a collection of texts, each of which is translated into one or more other languages than the original. The simplest case is where two languages only are involved: one of the corpora is an exact translation of the other. Some parallel corpora, however, exist in several languages. Also, the direction of the translation need not be constant, so that some texts in a parallel corpus may have been translated from language A to language B and the other way around. The direction of the translation may not even be known.

Parallel corpora are made in the business of communication in multilingual societies, such as the United Nations, NATO, The EU and officially bilingual countries such as Canada.

4.10.1 Advantages of Parallel Corpora

Besides cost efficiency, one principal advantage of the proposed technique is that it helps to further diminish the role of human intuition. Accordingly, in this approach, neither source language nor target language LUs is extracted directly by lexicographers from the corpus. Instead, LUs are determined by their contexts both in the SL and in the TL corpus and their translational equivalents provided by the parallel sentences.

Moreover, the method ranks the translation candidates according to how likely they are, based on automatically determined translational probabilities. This in turn renders it possible to determine which sense of a given lemma is the most frequently used, provided that distinct translations are available. Thus, representative corpora guarantee not only that the most important source lemmas which will be included in the dictionary as in traditional corpus based lexicography but also the translations of their most relevant senses. The third great advantage of the proposed technique is that all the relevant natural contexts can be provided both for the source and for the target language (Dash, 2008). The contexts of the source language and the target language words could be exploited for multiple purposes.

First, they can be of great help in determining which translation variants should be used, thus enabling lexicographers to find the most appropriate translation on the one hand, and to describe the use of the target language expression in grammatical or collocation terms, on the other. Hence, the great amount of easily accessible natural contexts facilitates the creation of encoding dictionaries (Dash, 2008).

Secondly, different sub-senses of a headword can be characterized manually based on the retrieved contexts. Accordingly, dictionaries relying on such information can provide positive evidence for the user, that all of these sub-senses are translated with the same lemma into the target language.

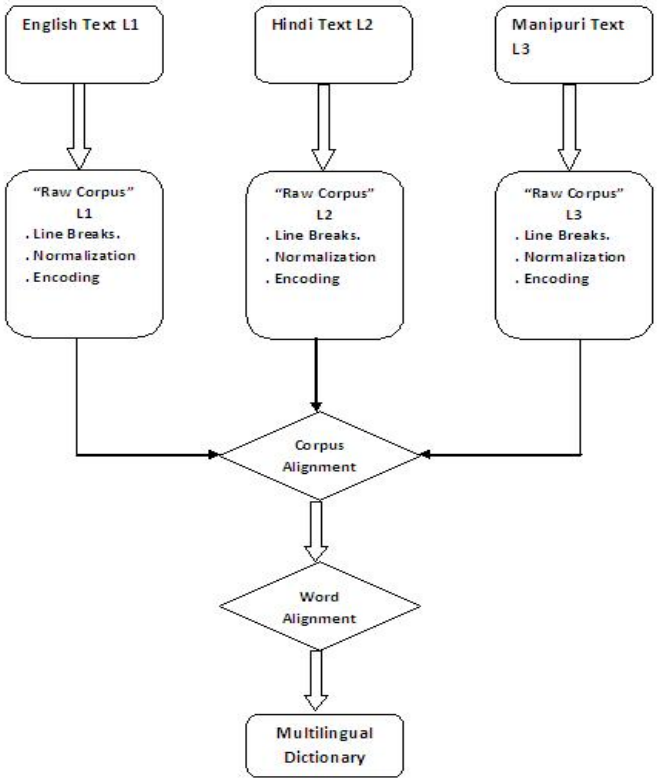

Figure 26: Workflow of parallel corpus of MMD

4.11 Corpus Size

Here we will discuss about the size of the corpus. How big a corpus will be? This implies that size is an important issue in corpus generation for MMD. It is concerned with total number of words (tokens) and different words (types) to be taken into a corpus. It also involves the decision of how many categories we like keep in the corpus, how many samples of texts we need put into each category, and how many words we shall keep in each sample. Although the question of size affects validity and reliability of a corpus, it is stressed that any corpus, however big, is nothing more than a minuscule sample of all speech and writing varieties produced by the users of a language.

In early days of corpus generation, when computer technology for procuring language data was not much advanced, it was considered that a corpus containing one million words is large enough to represent a language or variety. But by the middle of 1980s, computer technology went through a vast change with unprecedented growth of its storage, processing, and accessing abilities that have been instrumental in changing

the concept regarding the size of a corpus. Now it is believed that the bigger the size of corpus the more it is faithful in representing the language under consideration. With advanced computer technology we can generate corpus of very large size containing hundreds of million of words.

4.12 Corpus Schematic Diagram

Raw text was collected from hard copies with different domain and thereafter converting raw to wx-notation, it was stored in database and the database was uploaded on website for queries and text processing. The schematic diagram is given below:

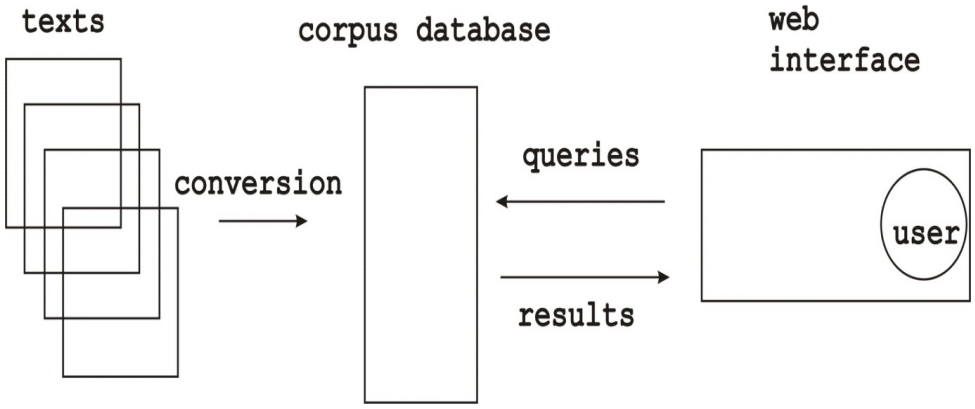

Figure 27: Schematic Diagram of corpus processing

4.13 CQL (Corpus Query Language)

CQL (Corpus Query Language) was developed by the Corpora and Lexicons groups, University of Stattgart. It is a language for building complex queries using RE (Regular expression), Attributes and Values. Using CQL many text processing can be done for different purposes in Corpus.

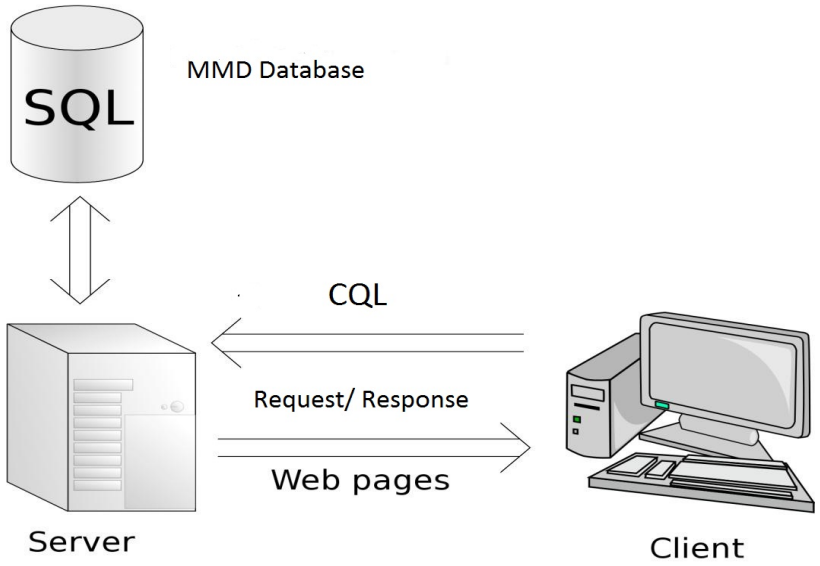

Figure 28: Process Diagram of web-based corpus.

4.14 Corpus Cleaning/Sanitation

The corpora have to be cleaned from such unintended errors as typos, wrong splits, foreign characters, which may have been introduced while keying the text i.e. in the process of digitization. For example, some of these corrections include, removal of '-', '~', '_' etc. which are introduced to break words at the end of lines while keying in the text. Sometimes the conversion of corpus text from one standard format to another may have introduced viz. alt, control characters (^C, ^M, ^Z etc.) are also removed. The resulting text is free from all such errors. Finally, the entire Indian Language corpora shall be converted to case sensitive roman notations in wx- scheme.

Certain other non-alpha-numeric characters that have introduced the unwanted, the process of corpus sanitation begin i.e. text editing, after the text are produced in electronics form. Generally, five types of error may occur at the time of manual data (Dash, 2008):

1. Omission of character,
2. Addition of Character,
3. Repetition of Character,
4. Substitution of character, and
5. Transposition of character

To remove spelling errors, we need to thoroughly check the corpus and compare it with the actual physical data source, and do manual correction. Care has to be taken to ensure that the spelling of words used in the corpus must resemble with the spelling of words used in the source texts. Also, it has to be checked if words are changed, repeated or omitted, punctuation marks are properly used, lines are properly maintained, and separate paragraphs are made for each other.

Besides error correction, we have to verify the omission of foreign words, quotations, dialectal forms, etc. after generation of a corpus.

4.14.1 Wildcards in Corpus

We can fine-tune a search by using any of the following wildcards. On the **Edit** menu, click **Find** or **Replace**. Some of methods for cleaning corpus are given under table with commands.

Table 15: Wildcards in corpus sanitation

To Find	Wildcard	Examples
Any single character	?	s?t finds "sat" and "set" ꯋ?ꯟ finds "ꯋꯤꯟ" and "ꯋꯣꯟ"
Any string of characters	*	s*d finds "sad" and "started". ꯁ*ꯗ finds "ꯁꯤꯗ" and "ꯁꯋꯨꯗ".
One of the specified characters	[]	w[ion finds "win" and "won" ꯋ[ꯤ ꯣ] ꯟ finds "ꯋꯤꯟ" and "ꯋꯣꯟ"
Any single character in this range	[-]	[r-t]ight finds "right" and "sight". [ꯁ-ꯇ]ꯤꯟ finds "ꯁ ꯤꯟ" and "ꯇ ꯤꯟ".
Any single character except the characters inside the brackets	[!]	m [!a]st finds "mist" and "most", but not "mast". ꯋꯤ[!ꯁ]ꯟ finds "ꯋꯤꯗꯟ" and "ꯋꯤꯌꯟ"
Any single character except characters in the range inside the brackets	[!x-z]	t[!a-m]ck finds "tock" and "tuck", but not "tack" or "tick". ꯋꯤ[!ꯗ-ꯇ]ꯟ finds "ꯋꯤꯗꯟ" and "ꯋꯤꯌꯟ"
Exactly *n* occurrences of the previous character or expression	{n}	fe{2}d finds "feed" but not "fed" ꯋ{2}ꯛ finds "ꯋꯦꯛ" but not "ꯋꯦꯣꯛ"

At least *n* occurrences of the previous character or expression	{n,}	fe{1,}d finds "fed" and "feed". ꯐꯤ{1,}ꯗfinds "ꯐꯤꯗ" and "ꯐꯤꯏꯗ".
From *n* to *m* occurrences of the previous character or expression	{n,m}	10{1,3} finds "10", "100", and "1000" ꯱꯰{1,3} finds "꯱꯰", "꯱꯰꯰", and "꯱꯰꯰꯰"
One or more occurrences of the previous character or expression	@	lo@t finds "lot" and "loot". ꯂ@ꯑ`ꯁꯗ finds "ꯂꯑ`ꯁꯗ" and "ꯂꯑꯑ`ꯁꯗ".
The beginning of a word	<	<(inter) finds "interesting" and "intercept", but not "splintered" <(ꯐ`ꯁ) finds "ꯐ`ꯁꯄꯤꯏ" and "ꯐ`ꯁꯆꯤꯏꯗ"
The end of a word	>	(ꯃꯤ)> finds "ꯃꯤ" and "ꯑꯃꯤ"

4.14.2 Codes That Work in The Find What Box Only (When Wildcards are off)

Table 16:Table codes when wildcards are off

To specify	Type
Any character	^?
Any digit	^#
Any letter	^$
Footnote mark	^f
Endnote mark	^e
Field	^d
Section break	^b
White space (any combination of regular and nonbreaking spaces, and tab characters)	^w

4.15 Schematic Diagram of Corpus Processing

The given schematic diagram represents how the corpus is processed for developing MMD corpus from different sources.

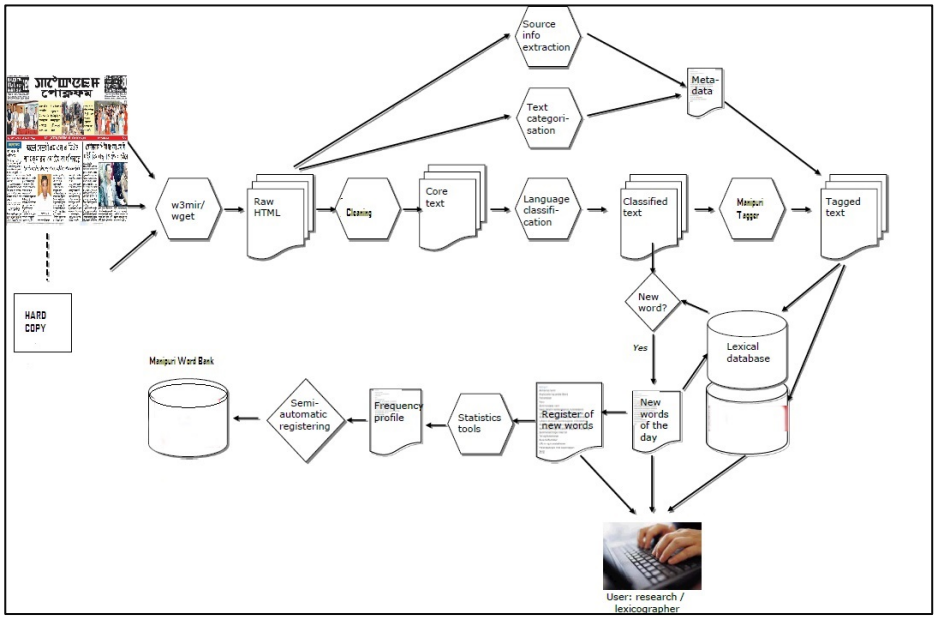

Figure 29: Schematic diagram of Corpus processing

4.16 CQL(Corpus Query Language) Result

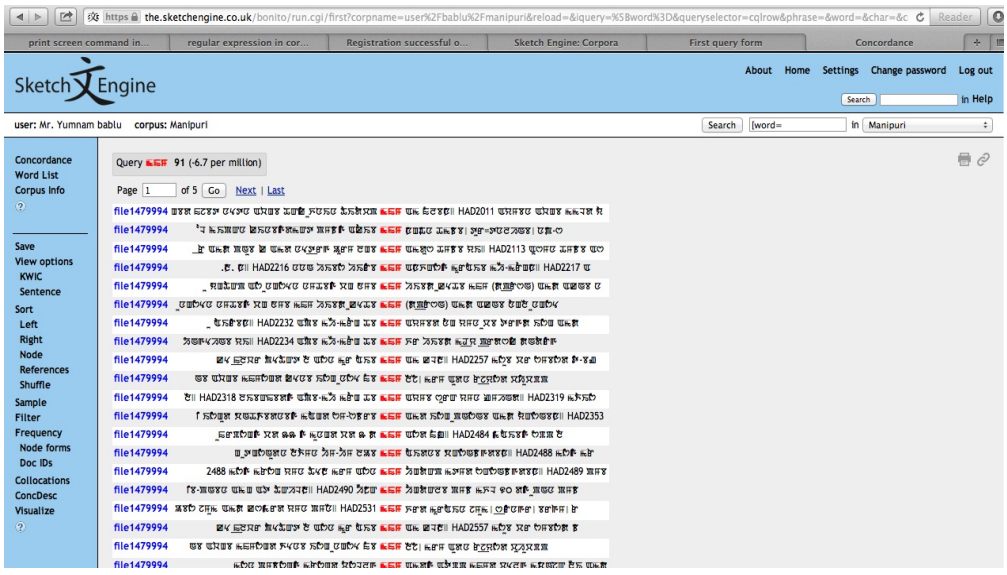

Figure 30: Output page of CQL

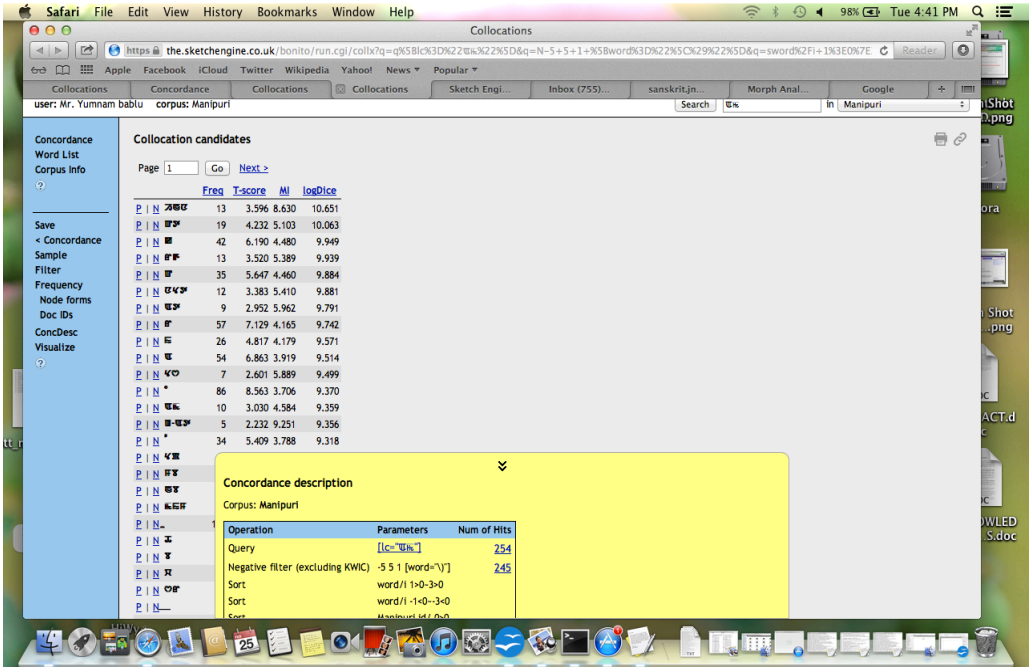

Figure 31: Frequency of word in CQL

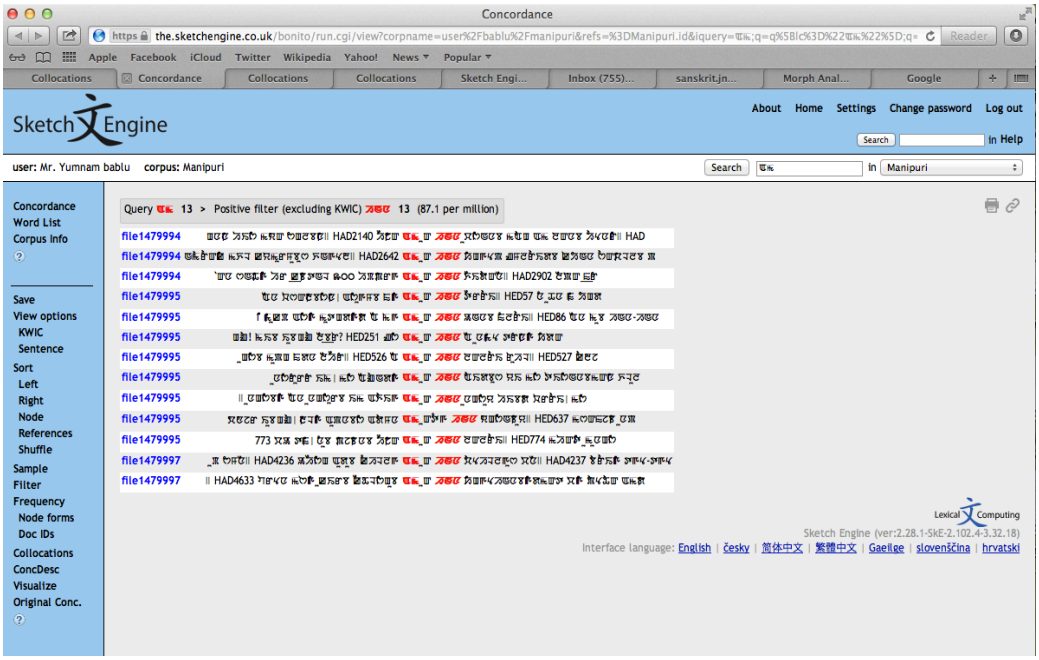

Figure 32: Word occurrence of CQL

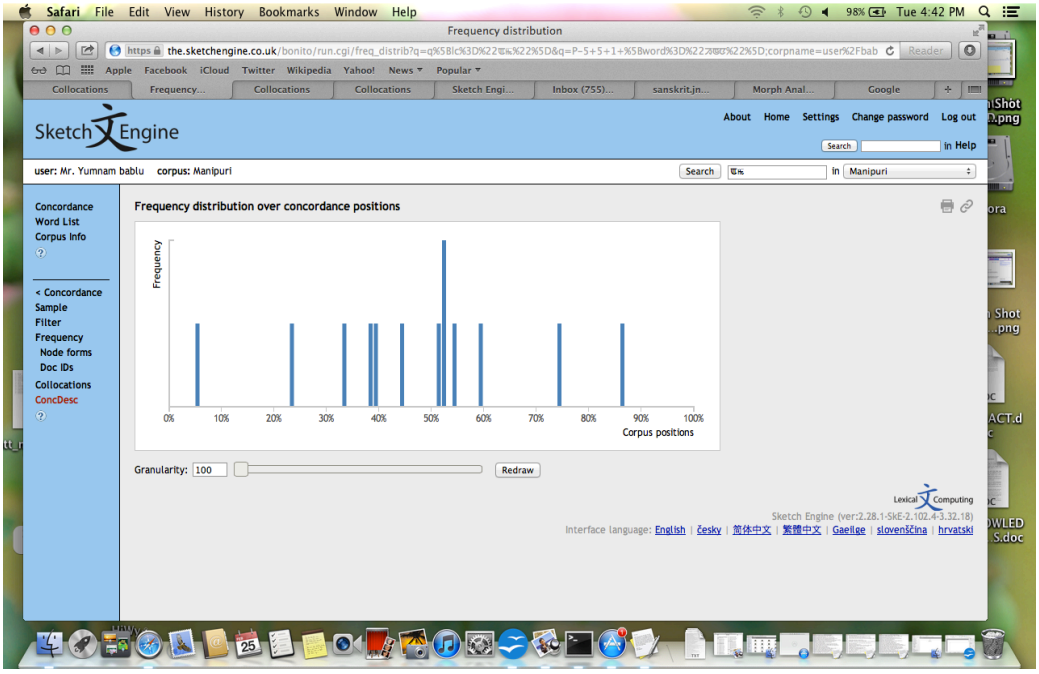

Figure 33: Frequency graph of Corpus processing

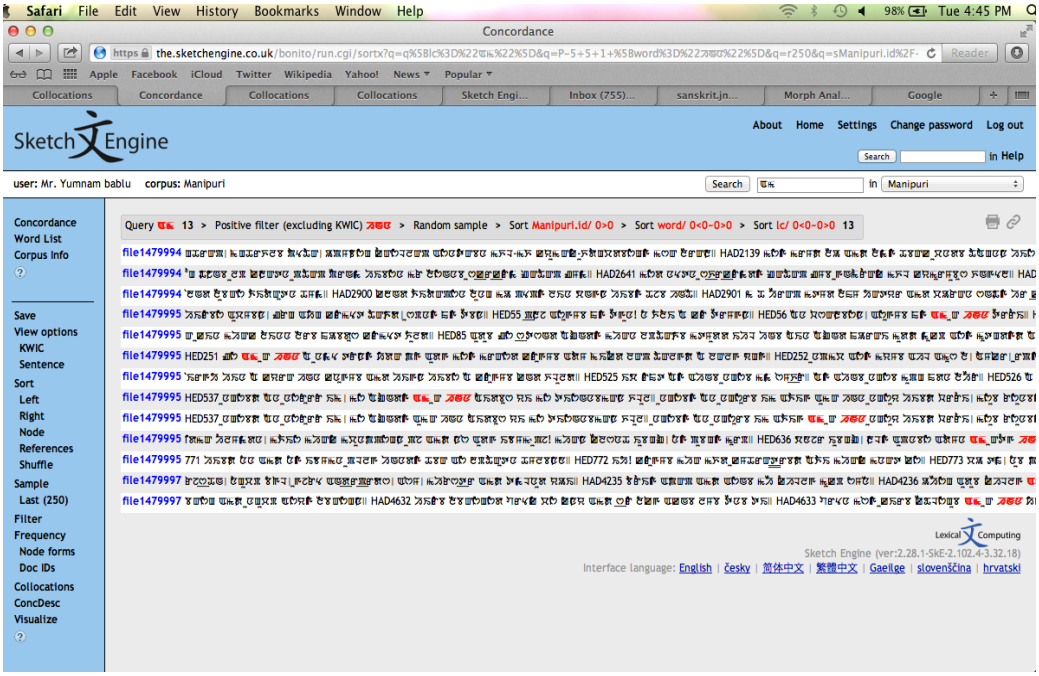

Figure 34: Concordance of word in Corpus processing

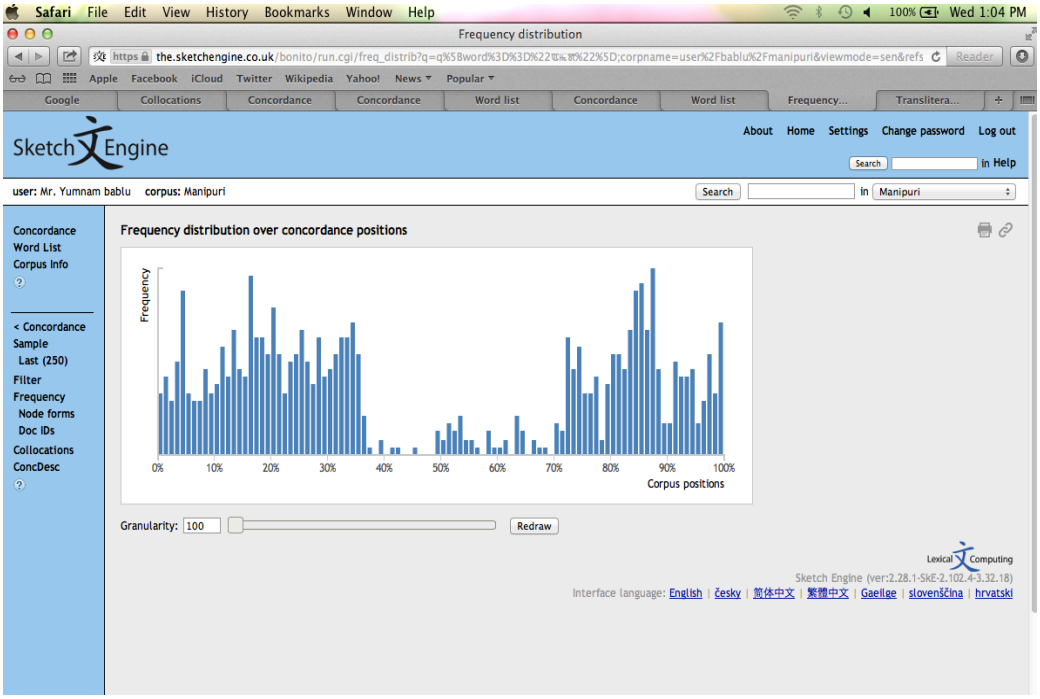

Figure 35: Frequency distribution of Corpus processing

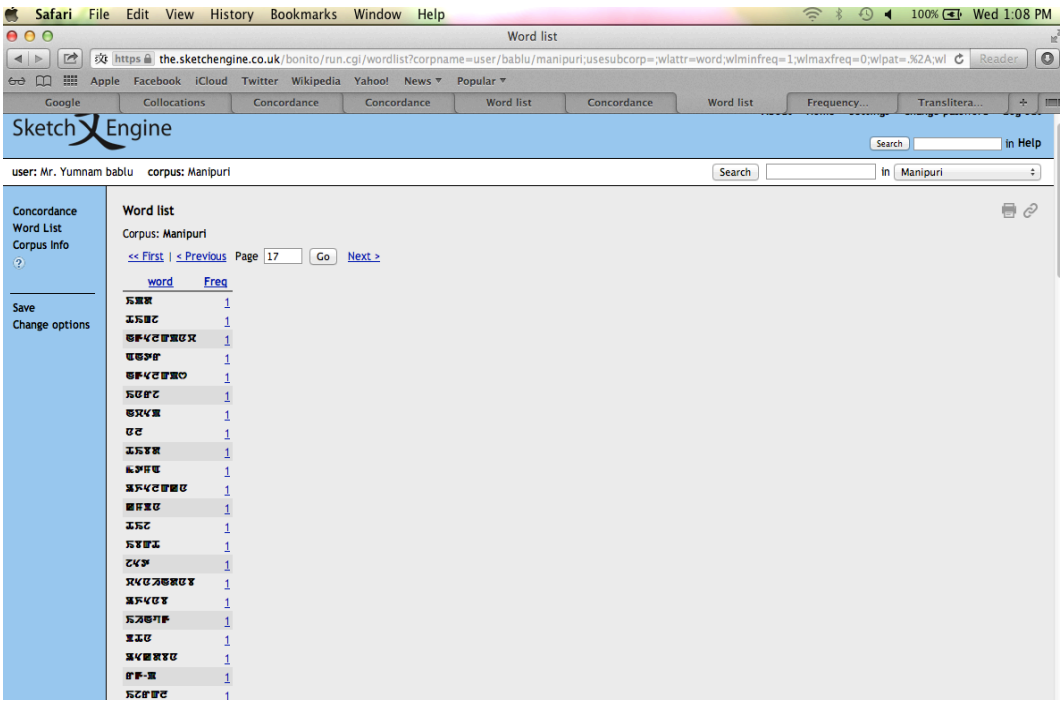

Figure 36: Word list of Corpus processing

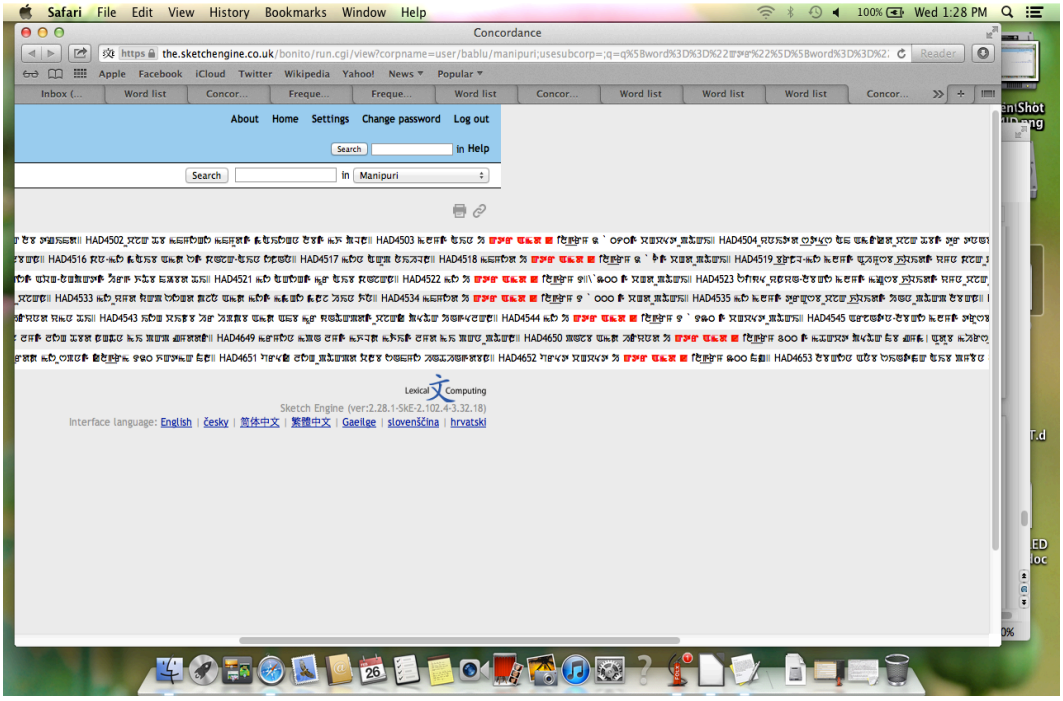

Figure 37: Word Concordance of Corpus processing

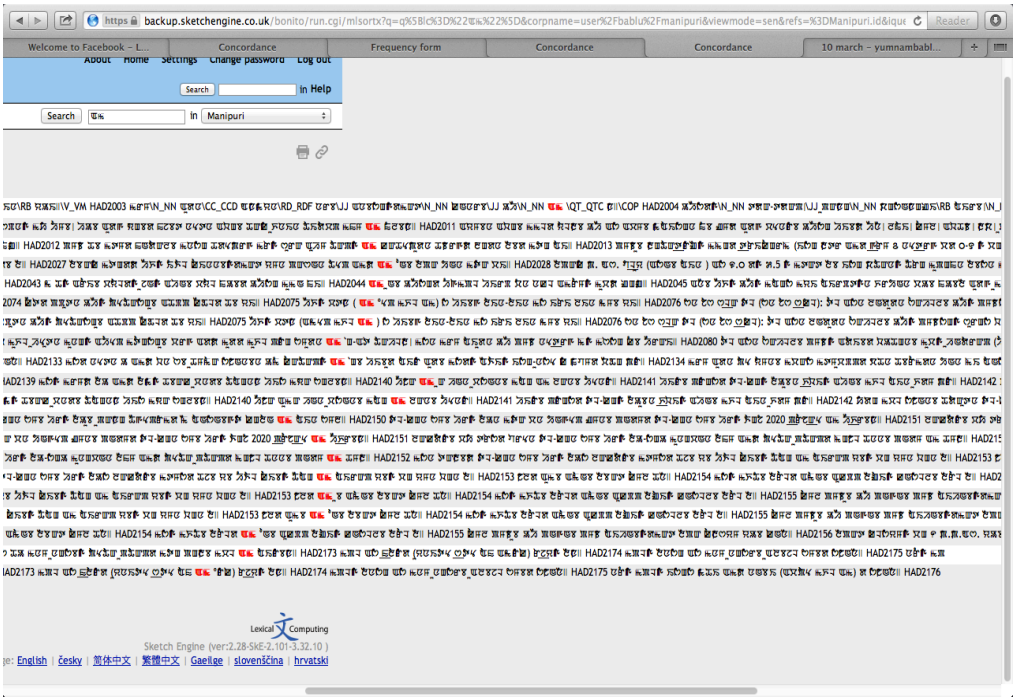

Figure 38: Occurrence of particular word

4.17 Conversion of Raw Text to WX- Schema

The raw data or texts are converted in WX notation i.e. in ASCII format for text processing by using a Perl programming language.

4.17.1 Raw Text Corpus

[illegible]

4.17.2 Text Converted in WX – Schema/Notation

itihAsata voVkVKiba acOba vOxoVkVsifVgi manufVxa mAleVmV asigi lamV pumVnamakVta capV mAnVnan maauaoViba amani hAina lOnabaxi Para_AnVsa_ laEbAkVta voVkVkiba miyAmVgi yAaoVlVgi ihO asini| hAiriba vOxoVkV asixa laEbAkV asixa laEba nifVvO pAnVbagi catVnabi amasufV lOyAmV cAbasifVna sAsanVgi cupVli pAibagi catVnabi axu maru PamVna muvaatVlaga manifV-maKA tamVba; amaga amagagi marakVta KeVnVnaba laEtaba amasufV marupV-mapAfV aoVibagi vAKalVloVnVxa sAgatVpa miyAmVna masAgi laEbAkV fAkVcagaxabani hAibagi catVnabigi maKAxa para_jAsifVna sAsanV tOba laEbAkV ama lifVKatVlamVmi|

4.18 Frequency

Frequency counting are done for counting unique words for text processing, corpus representativeness, word frequency and range, treatment of word families, treatment of idioms and fixed expression, range of information and various other criteria.

Word frequency are counted by using

```
chomp(@words = <STDIN> ) ;

for each $word(@words)

{

$count{$word} = $count{$word

} + 1;

} for each $word (keys %count)
```

```
{

print "$count{$word} $word\n"; #print "$word was seen $count{$word} times\n";

}
```

Frequencies Count

Takes a file with words and the corresponding frequencies # and outputs syllables with their frequencies. while ($line = <>)

```
{    chomp($line); ($frq, $_) = split (/\s+/, $line); while ($_ ne '')  {

# Do until the line ends.

if (/[^aieou]*[aiueo]+/i)

{        #space+Consonant (cluster)+vowel

$syllable = $&; $_ = $';

print $frq, ' ', $syllable, "\n"; }

elsif(/[^aiueo]*[aiueo]+[^aiueo]*/i) { # Conso cluster + vowel $syllable = $&; $_ =
$'; print $frq, ' ', $syllable, "\n";

} else { last;

# print STDERR "Some non-syllable at the end: $_\n";

last; }
```

Table 17: Table of Word Frequency Count

No of Frequency	Words
15	ꯑꯁꯇꯗꯄꯤ
15	ꯂꯣꯈꯪꯊꯄꯤ
18	ꯆꯦꯁꯍꯤꯄꯤ
23	ꯑꯅꯦꯌꯗꯇ
352	ꯍꯣꯌ
308	ꯑꯗꯨꯒ
319	ꯌꯋꯌ
320	ꯑꯂꯤꯒ
324	ꯁꯨꯃꯤ
328	ꯆꯗꯊ
664	ꯑꯗꯨ
683	ꯅꯦꯌꯇ
886	ꯑꯍꯂꯨꯝ

4.19 Annotation for POS Tagging

For POS tagging in Manipuri words we follow the Annotations guidelines of BIS (Bureau of Indian standard) which is given in the table below:

Table 18: Annotation Table

POS	ANNOTATIONS
Noun_Common	N_NN
Noun_Proper	N_NNP
Noun_Nloc	N_NST
Pronoun_Personal	PR_PRP
Pronoun_Reflexive	PR_PRF
Pronoun_Relative	PR_PRL
Pronoun_Reciprocal	PR_PRC
Pronoun_Wh-word	PR_PRQ
Demonstrative_Deictic	DM_DMD
Demonstrative_Relative	DM_DMR
Demonstrative_Wh-word	DM_DMQ
Verb_Main	V_VM
Verb_Finite	V_VM_VF
Verb_Non-Finite	V_VM_VNF
Verb_Infinite	V_VM_VINF
Verb_Gerund	V_VM_VNG
Noun_Verbal	N_NNV
Verb_Auxiliary	V_VAUX
Adjective	JJ
Adverb	RB
Postposition	PSP
Conjunction_Co-ordinator	CC_CCD
Conjunction_Subordinator	CC_CCS
Conjunction_Quotative	CC_CCS_UT
Particles_Default	RP_RPD
Particles_Classifier	RP_CL
Particles_InterJection	RP_INJ
Particles_Intensifier	RP_INTF
Particles_Negation	RP_NEG
QuanifiersGeneral	QT_QTF
QuanifiersCardinals	QT_QTC
QuanifiersOrdinals	QT_QTO
Residuals_ForeignWord	RD_RDF
Residuals_Symbol	RD_SYM
Residuals_Punctuation	RD_PUNC
ResidualsUnknown	RD_UNK
Residuals_Echowords	RD_ECH
Copula	COP
Cunjunc verb	V_CV

Compound verb	V_CV_C
Conjunc verb with space	V_CV_SPC
Pronoun_indefinite	PR_PRI

4.19.1 Annotation Process

Part of Speech tagging (POS) for MMD corpus is done by manually following the BIS tagset, which is a standard tagset and all Indian languages are following this guidelines and the process is done manually by linguistics person and for Manipuri language we use to add few extra tagset like Copula, postposition etc. and the process of working for tagging is in such a way as is given by the diagram below in figure (39).

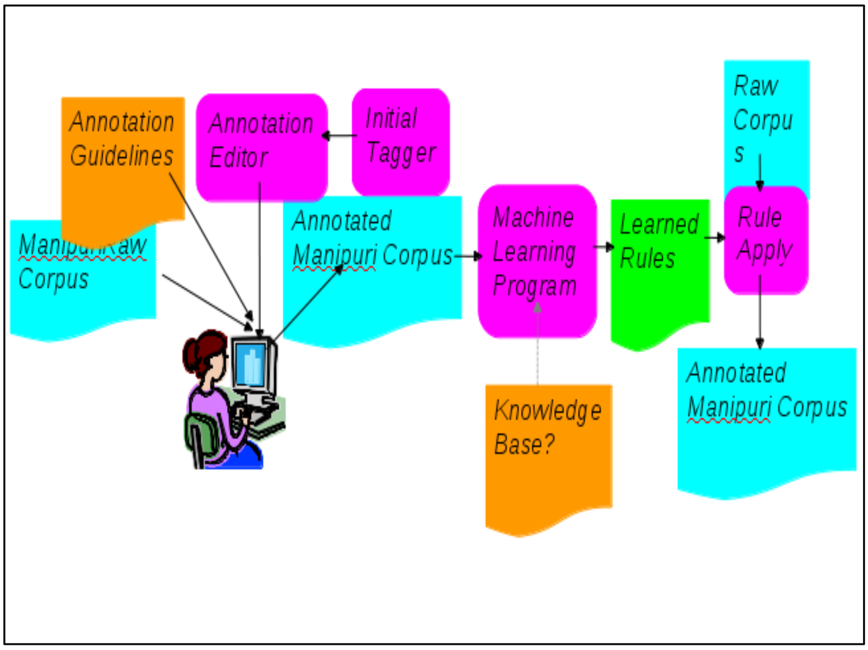

Figure 39: Annotation process for POS tagging

4.19.2 Tagged Corpus

[illegible]\RD_RDF [illegible]\CC_CCS_UT [illegible]\N_NN [illegible]\DM_DMD [illegible]\VA-VM_VNF [illegible]\N_NN [illegible]\N_NN [illegible]\N_NN [illegible]\RP_INTF [illegible]\JJ [illegible]\N_NN [illegible]\RB [illegible]\N_NNV [illegible]\RP_INTF [illegible]\JJ [illegible]\QT_QTC [illegible]\N_NN [illegible]\JJ [illegible]\N_NN [illegible]||\COP

[illegible]\N_NN [illegible]\RD_RDF [illegible]\CC_CCD [illegible]/RD_RDF [illegible]\DM_DMD [illegible]\N_NN [illegible]\RB [illegible]||\V_VM

ꯃꯐꯝ\N_NNꯑꯗꯨꯅ\CC_CCDꯑꯣꯅꯤꯃꯤꯔꯦꯅ\RD_RDF ꯅꯦꯐꯕ\JJ ꯑꯅꯦꯕꯂꯤꯗꯤꯄꯤꯇꯃꯗꯁ\N_NN ꯊꯦꯋꯅꯐꯕ\JJ ꯈꯔꯨ\N_NN ꯑꯃ\QT_QTC ꯅꯤ꯫\COP

ꯈꯔꯨꯂꯤꯇꯄꯤ\N_NN ꯁꯇꯣꯗ-ꯁꯇꯣꯗꯒ꯬\JJ ꯒꯨꯗꯅꯤꯗ\N_NN ꯌꯤꯗꯂꯤꯋꯅꯤꯗꯀꯦꯍ\RB ꯑꯣꯍꯐꯕ꯬\N_NNV ꯃꯅꯝ\N_NN ꯂꯨꯗꯂꯤꯕ\JJ ꯑꯃꯂꯨꯗ\CC_CCD ꯔꯦꯈꯐꯕ\JJ ꯅꯣꯝ\RD_RDF ꯑꯃꯇꯤ\CC_CCD ꯑꯦꯍꯂꯗꯐꯤꯝꯂꯨ\N_NN ꯂꯣꯝꯅꯪ꯫\V_VM

4.19.3 CQL Tagged Corpus

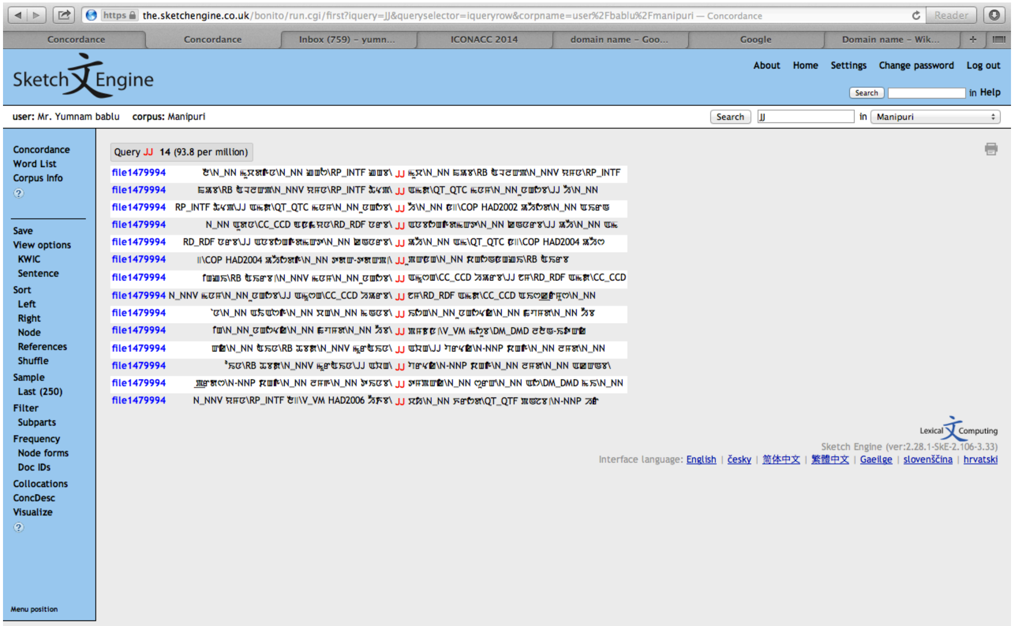

Figure: 40 CQL for POS in Manipuri Corpus.

4.20 Hardware and Software

For developing the MMD corpus we used a Personal Computer (PC) with a GIST or, software, namely the BBEdit, a monitor, one conventional computer key-board, a multilingual printer, and some CDs. Text files are developed with the help of the TC installed in the PC. This allows the display of various Indian scripts and Manipuri script on computer screen in a very convenient manner. The code for various characters used for the Indian scripts are standardized by the (BIS) *Bureau of Indian Standards, Govt. of India.* With installation of this card inside a PC, one can access almost the entire range of text-oriented application packages. One can also input and retrieve data in Indian languages with the help of this card. The software also provides a choice between two operational display modes on the monitor: one in conventional English mode, and the other in Indian multilingual mode

4.21 Application areas of Corpus

The purpose of a MMD corpus is not merely to gather a big file of different texts and store it on the computer, but also to prepare the texts and put them in a certain format so that they can be used by search tools and the results of the search can be displayed in a way that is meaningful and useful to the linguist, teacher and learner especially at the advanced level. For example, scholars, teachers and learners can explore the use of a word in different types of texts to see how frequently this word is used, how many meanings it has, what syntactic environment it occurs in, whether the word has the same frequency of occurrence in all types of texts. Teachers can identify the most frequent words and select them as a basis for their material. There is also the study of syntactic structures and analyzing the distribution of competing structures. For example, the uses of verb–subject vs. subject–verb word order in Manipuri: which word order is more preferred in children's stories, interviews, and scientific documents? The MMD corpus is to be annotated with XML mark-up which includes information about the text, author, and source; this gives the opportunity to conduct empirical analyses which control extra-linguistic factors (such as age, sex, region, social class, and education level) and examine the accompanying linguistic variations. This would make retrieval of useful information qualitatively and quantitatively much richer and easier to handle.

There are a number of areas where language corpus is directly used as in language description, study of syntax, phonetics and phonology, prosody, intonation, morphology, lexicology, semantics, lexicography, discourse, pragmatics, language teaching, language planning, sociolinguistics, psycholinguistics, semiotics, cognitive linguistics, computational linguistics to mention a few. In fact, there is hardly any area of linguistics where corpus has not found its utility. This has been possible due to great possibilities offered by computer in collecting, storing, and processing natural language databases. The availability of computers and machine-readable corpora has made it possible to get data quickly and easily and also to have this data presented in a format suitable for analysis.

1. Corpus as Knowledge Resource: like developing multilingual libraries, designing course books for language teaching, compiling monolingual dictionaries (printed and electronic), developing bilingual dictionaries (printed and electronic),

multilingual dictionaries (printed and electronic), monolingual thesaurus (printed and electronic version), various reference materials (printed and electronic version), developing machine readable dictionaries (MRDs), developing multilingual lexical resources, electronic dictionary (easily portable, can be duplicated as many copies as needed, can be modified easily for newer versions, can be customized according to need of users, can be read and accessed easily, more durable than printed dictionary, etc.).

2. Corpus In Language Technology: like word processing, spelling checking, text editing, morphological processing, sentence parsing, frequency counting, item-search, text summarization, text annotation, information retrieval, concordance, word sense disambiguation, WordNet (synset), semantic web, Semantic Net, Parts-of-Speech Tagging, Local Word Grouping, etc.

3. Corpus For Translation Support Systems: like language resource access systems, Machine translation systems, multilingual information access systems, and cross-language information retrieval systems, etc.

4. Corpus For Human-Machine Interface Systems: used for OCR, voice recognition, text-to-speech, e-learning, on-line teaching, e-text preparation, question-answering, computer-assisted language education, computer-aided instruction, e-governance, etc.

5. Corpus In Speech Technology: Speech corpus technology is used to develop general framework for speech technology, phonetic, lexical, pronunciation variability in dialectal versions, automatic speech recognition, automatic speech synthesis, automatic speech processing, speaker identification, repairing speech disorders, and forensic linguistics, etc.

6. Corpus In Mainstream Linguistics: is used for following mainstream linguistics: language description, lexicography, paribhasa formation, grammar writing, semantic study, language learning, dialect study, sociolinguistics, psycholinguistics, stylistics, bilingual dictionary, extraction, translation equivalents, generation of terminology databank, lexical selection restriction, dissolving lexical ambiguity, grammatical mapping, semiotics, pragmatic and discourse study, etc.

4.22 Conclusion

The chapter provided an extensive survey how MMD corpus can be developed, mainly based on parallel corpora and mono corpora for Manipuri raw data from hard copy of books, newspaper and other hard copy sources. To our knowledge, our system is the first one aimed at building semantic lexicons from raw text using semantic knowledge. Our corpus-based approach is designed to support fast semantic lexicon construction. A user only needs to supply a representative text corpus and a small set of seed words for each target category. We envisaged that not only this corpus fills a gap in the general field of corpus linguistics but it would also have a role in providing authentic material for teaching Manipuri as a foreign language, developing tools that serve the spread of the use of Manipuri, and encouraging wide scale research into investigating linguistic phenomena based on a large, varied dataset. The initial version of the Corpus of Contemporary as an extension to this work plans to set up infrastructure and prototype sampler corpus for the International Corpus of Manipuri in Meitei Mayek. Building semantic lexicons will always be a subjective process, and the quality of a semantic lexicon is highly dependent on the task for which it will be used. But there is no question that semantic knowledge is essential for many problems in natural language processing. This can provide a knowledge management environment for computer-supported collaborative work including discussion and authoring of standards, and tools for collation, mark-up, lexicon-grammatical analysis, exploration and dissemination of the International Corpus of Manipuri in Meitei Mayek script. This will establish Manipuri at the center of international development and exploitation of Manipuri corpus linguistics and language.

The chapter has described the work of compiling a corpus to support the building of the MMD, its challenges, and the solutions adopted for using this corpus for lexicographical tasks, NLP task and many more Applications.