“......Good Morning!” said Bilbo, and he meant it. The sun was shining, and the grass was very green. But Gandalf looked at him from under long bushy eyebrows that stuck out further than the brim of his shady hat.

“What do you mean?” he said. Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?

“All of them at once”, said Bilbo. ......

— The Hobbit, J. R. R. Tolkien (1892 - 1973)
Using a language to represent or convey information comes “naturally” to a human, but the use of human language, actually, is far from trivial. The term computational linguistics refers to the inter-disciplinary field at the intersection of linguistics, phonetics, computer science, cognitive science, artificial intelligence and formal logic, frequently assisted by statistical techniques [7]. Analysing or processing linguistic phenomena necessitates viewing natural language text in a “structured way”. The structured approach to the study of language has given birth to three main areas of linguistics — the study of the arrangement of the basic units of the language following basic rules, the study of the meaning that each word bears, and the study of the context that may change the meaning of a word. Each area or sub-area of language processing (for instance, stemming — finding the root, stem, or base form from an inflected word; identification of grammatical category of a word) needs analysis of linguistic phenomena in a computational environment supported by knowledge of traditional grammar.

A number of approaches to solve specific problems have been proposed and tried. Researchers have tried to address general issues as well as language specific ones. This thesis starts out with a fundamental question—

“Are language related techniques developed for one language equally applicable to other languages? Should new techniques be developed that are more appropriate for the individual characteristics of a specific language?”

The problems we tackle in this work have been amply discussed in the context of a language like English. However, there are many language specific issues that need to be addressed. In addition, we believe that there are easier and more efficient ways to handle issues better than existing techniques for a specific language. From the perspective of the language we are interested in, the problems investigated in this thesis have not been studied. Our experiments attempt to answer the question given above in the context of Assamese. We develop a computational linguistic model of Assamese. In some situations, we extend the proposed techniques to neighboring languages to generalize our approach.

Though Assamese is one of the scheduled national languages of India, little computational work has been done so far for this language. Developing a language-processing tool is not a straightforward process due to the nature of natural languages and the manner in which language is represented at various levels. There are ambiguities at each phase of representation and processing. Since computational linguistics is new and there are substantial variations in terms of rules or dialects, there arises the problem of standardization such as the standardization of morphological classes and the standardization
of a PoS tagset. Overcoming all these barriers requires a lot of effort. This research investigates methodologies to deal with the problems of stemming, PoS tagging, identification of multi-word units and parsing of Assamese text.

1.1 Objective

The primary objective of this thesis is to develop an approach for parsing Assamese text. Early in our research, we realized that the basic software modules and resources required to parse texts are not adequately available for Assamese. Therefore, our work covers the prior tasks essential for parsing, too. We define three interrelated sub-objectives of our work.

- **Module: 1** - Finding the root or stem of a word,
- **Module: 2** - Finding the grammatical category, and
- **Module: 3** - Finding sentence structures

To perform our computational experiments, the basic need is a corpus of Assamese. For our experiments, we use four different collections of texts — (a) Assamese Pratidin corpus\(^1\); (b) EMILLE corpus\(^2\); (c) Wikipedia corpus\(^3\); (d) Tezu Assamese corpus\(^4\). Except the Assamese Pratidin corpus, others are encoded in UTF-8 format. We have developed a converter to convert ASCII based normalised text of the Assamese Pratidin corpus to UTF-8.

1.2 The target language

In this section, we provide a brief relevant linguistic background of our target language, Assamese. For comparative study and to generalize our proposed approach, we also work with Bengali, Bishnupriya Manipuri and Bodo, in some places. The Assamese language is a member of the Indo-Aryan language family and spoken in the north eastern part of India and its neighbouring regions. The word formation process of Assamese includes

---

\(^1\)Developed by U. Sharma, Tezpur University.
\(^2\)http://lancs.ac.uk/fass/projects/corpus/emille
\(^3\)Collected article from http://as.wikipedia.org
\(^4\)Collection of online news and blog articles
inflection, derivation, coinage, clipping, back formation, translation and transliteration. Being an inflectional and agglutinative language, it is morphologically rich. Though it is a relatively free word order language, the dominant word order is subject-object-verb (SOV). In spoken form, Assamese has a number of dialects and in text, a standard form of the language is generally written using the “Assamese script”. When we mention Assamese throughout this thesis we primarily mean standard Assamese as opposed to other dialectal variations.

Like most Indian languages, Assamese has been studied infrequently in the global context. It still lacks even a single balanced corpus or basic language processing modules like stemmers and morphological analysers. Before we started our work, we found a few reported computational efforts in Assamese. Among them [8, 9, 10, 11, 12] are the main. Sharma et al. [8] describe an approach to extract stems from affix evidence. An extended version [10, 12] presents an unsupervised approach to learn morphology using corpus text. A spell checker for ASCII based Assamese text is reported by Das et al. [9]. Considering just this set of reported work as our base, we started our experiments to design the three main language processing tools, namely a stemmer, a PoS tagger and a parser that are discussed at length in subsequent chapters.

For transliteration of examples in the thesis, we use an in-house transliteration scheme and also provide representation using the International Phonetics Alphabets (IPA). The general syntax used in our examples is - word_in_non-roman_script (IPA : meaning).

1.3 Contributions of this thesis

The most important contributions of our work are given below.

- **Efficient stemming:** We propose a hybrid stemming approach for Assamese and apply our method to some neighbouring languages, viz., Bengali, Bishnupriya Manipuri, and Bodo.

- **PoS tagger:** We design an Assamese-specific hierarchical tagset and experiment with three PoS tagging methods for Assamese. The first approach is the rule-based approach to classify noun and verb from raw text. After that, we use a dictionary with the rule-based approach to increase the PoS tagging accuracy. Lastly, we
experiment with an HMM based approach to classify Assamese text. We obtain 87-90% precision using the HMM based tagger.

- **Extraction of MWU**: We label each sentence with standard IOB (Inside Outside Beginning) [1] tags and employ Yamcha [2], a supervised Support Vector Machine based tool to identify and classify MWU from the annotated corpus.

- **Dependency parsing model**: We explore three dependency parsing models for Assamese, viz. Link grammar [3] parsing, Malt parsing [4] and MST [5] parsing. We develop the rule-base and a dictionary for link grammar parser. We also compare the performance of these three parsing models. We have developed a repository to store the parsed sentences.

### 1.4 Outline of the thesis

This thesis is organised around four main parts.

**Chapter 2** describes the preparation and development of the Assamese corpora used throughout the experiments. For our experiments, we use four different collections of raw text, viz., the EMILLE corpus, the Assamese Pratidin corpus, the Wikipedia corpus and the Tezu Assamese corpus.

**Chapter 3** presents a detailed description of stemming in the context of Assamese text. We analyse previous work on stemming and, we extend our experiments to three other Indian languages to generalise our proposed approach. We examine stemming accuracies with three different approaches and analyse the outcomes. First, we design a rule-based approach to remove suffixes from words. To reduce over-stemming and under-stemming errors, we introduce a dictionary of frequent words. We observe that for these languages a dominant portion of suffixes are 1-letter long and these suffixes cause ambiguity during suffix stripping. Finally, we introduce an HMM based hybrid approach to classify the mis-matching of the last character with the single letter suffix set. For each word, stemming is performed by computing the most probable path in the four defined HMM states.

**Chapter 4** describes our work on development of Assamese PoS tagger. We study the state-of-the-art in PoS tagging of Assamese. We develop a hierarchical tagset specifically for Assamese, and introduce three approaches for automatic tagging. The first approach describes the identification of noun and verb, since these two are open
class categories like in other Indian languages. The second approach is a dictionary based model that uses a word list and rule-base to determine the grammatical category and the third approach is a Hidden Markov Model based approach. We also describe our approach to identification and processing of multi-word units. We present our experimental results in identification of reduplicatives, compound nouns and compound verbs.

In Chapter 5, we describe dependency parsing and its state-of-the-art for Indian languages. We develop a Link grammar and parser for Assamese. We also explore the Malt parser and MST parser for Assamese and discuss evaluation methods. Finally, we discuss the architecture of Tezu-TreeBank, the TreeBank for Assamese.

In Chapter 6, we conclude this thesis by summarising our major achievements and discussing possible future work.