Appendix A

Definition of Terms:

Commonly used definitions of term in the text of thesis are described below:

Annotation: An annotation is metadata (e.g., a comment, explanation, and presentational markup) attached to text, image, or other data. It means to make a corpus more powerful, an enrichment of the original corpus.

Computational Linguistics: Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective.

Collocation: A sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components.

Corpus: A corpus or text corpus is a large and structured set of texts. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.

Knowledge Representation: Different schemes for representing knowledge have been advanced, including predicate calculus, scripts and frames, and production systems.

Language: A system of signs used for communication between human beings. Natural or ordinary languages include English, Bengali, Hindi, Urdu, German, and any other used by a community of humans.

Lexical Item: A lexical item is a single word, a part of a word, or a chain of words that forms the basic elements of a language's lexicon.

Lexicon: A lexicon is a language's inventory of lexemes. The word "lexicon" derives from the Greek _____ (*lexicon*), neuter of _____ (*lexikos*) meaning "of or for words".

Linguistics: Linguistics is the scientific study of language. There are broadly three aspects to the study, which include language form, language meaning, and language in context.

Morphosyntactic: The study of grammatical categories or linguistic units that have both morphological and syntactic properties

Morpheme: A meaningful linguistic unit consisting of a word, such as *man,* or a word element, such as *-ed* in *walked,* that cannot be divided into smaller meaningful parts.

Morphology: A branch of linguistics that studies and describes patterns of word formation, including inflection, derivation, and compounding of a language.

Multiword Expression: It is made up of the combination of two or more than two words in which most of the time words lose their individual meaning and form a new resultant meaning.

Part of speech tagging: In corpus linguistics, part of speech tagging also called grammatical tagging or word-category disambiguation is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context.

Tagset: The set of tags used for annotation in a particular language in a particular corpus.

Tagged Corpus: The text corpus in which all the lexical items are annotated with its proper part of speech tag is known as tagged corpus. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistics rules within a specific language territory.