APPENDIX

Appendix A

Definitions of Terms

This appendix lists the terms frequently used in this thesis.

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

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.

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.

Morph syntactic: 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.

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.

Tag set: 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.

Stemming: Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived)words to their word stem, base or root form generally a written word form.

Chunking: Chunking is a natural language processing (NLP) task that focuses on dividing a text into syntactically correlated non-overlapping and non exhaustive groups of words, i.e., a word can only be a member of one chunk and not all words are in chunks

Named Entity: Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify elements in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.