**CHAPTER 2**

**LITERATURE SURVEY**

There are various approaches to develop the Shallow Parser. In this chapter some of these approaches are described with their working methodologies and advantages and limitations of each approach. Section 2.1 gives a brief introduction to the various terms used in parsing. Section 2.2 contains the detailed description of the various approaches of the Shallow parser. Section 2.3 gives an overview to the morphology of the “Hindi” language.

**2.1 Parsing**

In linguistics and computer science, **parsing**, or, more formally, **syntactic analysis**, is the process of analyzing a text, made of a sequence of tokens (for example, words), to determine its grammatical structure with respect to a given (more or less) formal grammar. Parsing can also be used as a linguistic term, for instance when discussing how phrases are divided up in garden path sentences.

Parsing is also an earlier term for the diagramming of sentences of natural languages, and is still used for the diagramming of inflected languages, such as the Romance languages or Latin. The term parsing comes from Latin *pars* (*ōrātiōnis*), meaning part (of speech).

Parsing is a common term used in psycholinguistics when describing language comprehension. In this context, parsing refers to the way that human beings, rather than computers, analyze a sentence or phrase (in spoken language or text) "in terms of grammatical constituents, identifying the parts of speech, syntactic relations, etc." This term is especially common when discussing what linguistic cues help speakers to parse garden-path sentences.

Parsing in natural languages can be classified at two abstraction levels namely, deep parsing and Shallow parsing.

**Shallow Parsing**

In this technique, we get hierarchical and grammatical information while preserving robustness and efficiency of the processing. In this perspective, we make use of a grammar represented in the Property Grammar formalism.

**Deep Parsing**

Deep analysis is directly based on property grammars. It consists, for a given sentence, in building all the possible subsets of juxtaposed elements that can describe a syntactic category. A subset is positively characterized if it satisfies the constraints of a grammar. These subsets are called edges; they describe a segment of the sentence between two positions.

**Word**: Word is defined as a smallest thought unit vocally expressible composed of one or more sounds combined in one or more syllables. A word is a minimum free form consisting of one or more morphemes. There are many ways to combine morphemes to create words. Four of these methods are common and play important roles in speech and language processing: Inflection, Derivation, Compounding and Cliticization.

**Inflection:** Inflection is the combination of a word stem with a grammatical morpheme, usually resulting in a word of the same class as the original stem, and usually filling some syntactic function, e.g. plural of nouns.

Table(singular)

Table+s(plural)

The meaning of the resulting word is easily predictable. Inflectional morphemes modify a word’s tense, number, aspect and so on.

**Derivation:** Derivation is the combination of a word stem with a grammatical morpheme, usually resulting in a word of a different class, often with a meaning hard to predict exactly.

**Compounding:** Compounding is the joining of two or more base forms to form a new word. For instance, two nouns “car” and “driver” can be fused to create “car-driver’. Such frequent root-root fusions are very common in written Hindi. Semantic interpretation of compound words is even more difficult than with derivates. Almost any syntactic relationship may hold between the components of a compound.

**Phrase:** In everyday speech, a **phrase** may refer to any group of words. In linguistics, a phrase is a group of words (or sometimes a single word) that form a constituent and so function as a single unit in the syntax of a sentence. A phrase is lower on the grammatical hierarchy than a clause.

Most phrases have an important word defining the type and linguistic features of the phrase. This word is the head of the phrase and gives its name to the phrase category. The heads in the following phrases are in bold:

too **slowly** - Adverb phrase (AdvP)

very **happy** - Adjective phrase (AP)

the massive **dinosaur** - Noun phrase (NP)

**at** lunch - [Preposition phrase](http://en.wikipedia.org/wiki/Preposition_phrase) (PP)

**watch** TV - [Verb phrase](http://en.wikipedia.org/wiki/Verb_phrase) (VP)

The head can be distinguished from its *dependents* (the rest of the phrase other than the head) because the head of the phrase determines many of the grammatical features of the phrase as a whole. The examples just given show the five most commonly acknowledged types of phrases. Further phrase types can be assumed, although doing so is not common. For instance one might acknowledge subordinator phrases:

**before** that happened - Subordinator phrase (SP)

This "phrase" is more commonly classified as a full subordinate [clause](http://en.wikipedia.org/wiki/Clause) and therefore many grammars would not label it as a phrase. If one follows the reasoning of heads and dependents, however, then subordinate clauses should indeed qualify as phrases. Most theories of syntax see most if not all phrases as having a head. Sometimes, however, non-headed phrases are acknowledged. If a phrase lacks a head, it is known as [exocentric](http://en.wikipedia.org/wiki/Exocentric), whereas phrases with heads are [endocentric](http://en.wikipedia.org/wiki/Endocentric).

**Head:** In linguistics, the head of a [phrase](http://en.wikipedia.org/wiki/Phrase) is the word that determines the [syntactic](http://en.wikipedia.org/wiki/Syntax) type of that phrase or analogously the [stem](http://en.wikipedia.org/wiki/Word_stem) that determines the semantic category of a [compound](http://en.wikipedia.org/wiki/Compound_%28linguistics%29) of which it is a part. The other elements modify the head and are therefore the head's *dependents*. Headed phrases and compounds are [endocentric](http://en.wikipedia.org/wiki/Endocentric), whereas [exocentric](http://en.wikipedia.org/wiki/Exocentric) ("headless") phrases and compounds (if they exist) lack a clear head. Heads are crucial to establishing the direction of [branching](http://en.wikipedia.org/wiki/Branching_%28linguistics%29). Head-initial phrases are right-branching, head-final phrases are left-branching, and head-medial phrases combine left- and right-branching.

Examine the following expressions:

big red **dog**

The word dog is the **head** of big red dog, since it determines that the phrase is a [noun phrase](http://en.wikipedia.org/wiki/Noun_phrase), not an adjective phrase. Because the adjectives big and red modify this head noun, they are its *dependents*.

**2.2 Approaches of Shallow Parser**

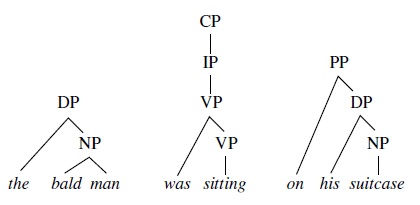
Various NLP research groups have developed different approaches and algorithm for shallow parsing. Some of the algorithms are language dependent and some of them are language independent. A brief survey of various approaches involved in Shallow parsing includes the following:

1. Parsing By Chunks
2. Machine Learning Approach for Shallow Parser
3. Memory based Shallow Parsing
4. HMM based Shallow Parsing
5. Text chunking based on generalized version of the Winnow Algorithm

**2.2.1 Parsing By Chunks**

This approach was given by Steven P. Abney who used intuitive reading as the basis of parsing. He realized the fact that we read a sentence one chunk at a time. These chunks correspond in some way to prosodic patterns. A simple context-free grammar is quite adequate to describe the structure of chunks. Gee and Grosjean did not assign syntactic structure to chunks. To remedy these, he assumed that chunk has syntactic structure and defined them in terms of major heads. Major heads are all content words except those that appear between a function word f and the content word that f selects.

The parse tree segments associated with some sample chunks are illustrated in (2).



Chunker implemented was a non-deterministic version of an LR parser employing a best-first search. Grammar used in the published implementation can be found in the references section.

Since shallow parsers have to deal with the entire Natural Language, they need thousands of rules. This makes building shallow parsers a labor-intensive task. This was a big limitation of Abney’s approach which used hand-crafted cascaded FST (Finite State Transducers).

**2.2.2 Machine Learning Approach**

This approach was given by Ramshaw and Marcus in their paper published in 1995. It formulates NP-chunking as a tagging task. The approach targets higher level of chunk structure using Brill's transformation-based learning mechanism, in which a sequence of transformational rules is learned from a corpus; this sequence iteratively improves upon a baseline model for some interpretive feature of the text.

In this study, training and test sets marked with two different types of chunk structure were derived algorithmically from the parsed data in the Penn Treebank corpus of Wall Street Journal text (Marcus et al., 1994). The source texts were then run through Brill's part-of-speech tagger (Brill, 1993c), and, as a baseline heuristic, chunk structure tags were assigned to each word based on its part-of-speech tag. Rules were then automatically learned that updated these chunk structure tags based on neighbouring words and their part-of-speech and chunk tags.