Few Examples of Corpus

1. TreeBank Corpus

It may be defined as linguistically parsed text corpus that annotates syntactic or semantic sentence structure. Geoffrey Leech coined the term ‘treebank’, which represents that the most common way of representing the grammatical analysis is by means of a tree structure. Generally, Treebanks are created on the top of a corpus, which has already been annotated with part-of-speech tags.

Types of TreeBank Corpus

Semantic and Syntactic Treebanks are the two most common types of Treebanks in linguistics. Let us now learn more about these types −

A) Semantic Treebanks

These Treebanks use a formal representation of sentence’s semantic structure. They vary in the depth of their semantic representation. Robot Commands Treebank, Geoquery, Groningen Meaning Bank, RoboCup Corpus are some of the examples of Semantic Treebanks.

B) Syntactic Treebanks

Opposite to the semantic Treebanks, inputs to the Syntactic Treebank systems are expressions of the formal language obtained from the conversion of parsed Treebank data. The outputs of such systems are predicate logic based meaning representation. Various syntactic Treebanks in different languages have been created so far. For example, **Penn Arabic Treebank, Columbia Arabic Treebank** are syntactic Treebanks created in Arabia language. **Sininca** syntactic Treebank created in Chinese language. **Lucy, Susane** and **BLLIP WSJ** syntactic corpus created in English language.

Applications of TreeBank Corpus

Followings are some of the applications of TreeBanks −

1) In Computational Linguistics

If we talk about Computational Linguistic then the best use of TreeBanks is to engineer state-of-the-art natural language processing systems such as part-of-speech taggers, parsers, semantic analyzers and machine translation systems.

2) In Corpus Linguistics

2. PropBank Corpus

PropBank more specifically called “Proposition Bank” is a corpus, which is annotated with verbal propositions and their arguments. The corpus is a verb-oriented resource; the annotations here are more closely related to the syntactic level. Martha Palmer et al., Department of Linguistic, University of Colorado Boulder developed it. We can use the term PropBank as a common noun referring to any corpus that has been annotated with propositions and their arguments.

In Natural Language Processing (NLP), the PropBank project has played a very significant role. It helps in semantic role labeling.

3. VerbNet(VN)

VerbNet(VN) is the hierarchical domain-independent and largest lexical resource present in English that incorporates both semantic as well as syntactic information about its contents. VN is a broad-coverage verb lexicon having mappings to other lexical resources such as WordNet, Xtag and FrameNet. It is organized into verb classes extending Levin classes by refinement and addition of subclasses for achieving syntactic and semantic coherence among class members.

Each VerbNet (VN) class contains −

1. A set of syntactic descriptions or syntactic frames

For depicting the possible surface realizations of the argument structure for constructions such as transitive, intransitive, prepositional phrases, resultatives, and a large set of diathesis alternations.

2. A set of semantic descriptions such as animate, human, organization

For constraining, the types of thematic roles allowed by the arguments, and further restrictions may be imposed. This will help in indicating the syntactic nature of the constituent likely to be associated with the thematic role.

4. WordNet

WordNet, created by Princeton is a lexical database for English language. It is the part of the NLTK corpus. In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called **Synsets.** All the synsets are linked with the help of conceptual-semantic and lexical relations. Its structure makes it very useful for natural language processing (NLP).

In information systems, WordNet is used for various purposes like word-sense disambiguation, information retrieval, automatic text classification and machine translation. One of the most important uses of WordNet is to find out the similarity

among words. For this task, various algorithms have been implemented in various packages like Similarity in Perl, NLTK in Python and ADW in Java.