Introduction to Corpus

INTRODUCTION

- A corpus is a large and structured set of machine-readable texts that have been produced in a natural communicative setting.
- Its plural is corpora.
- They can be derived in different ways like text that was originally electronic, transcripts of spoken language and optical character recognition, etc.

Elements of Corpus Design

- Language is infinite but a corpus has to be finite in size.
- For the corpus to be finite in size, we need to sample and proportionally include a wide range of text types to ensure a good corpus design.
- Some important elements for corpus design
  - Corpus Representativeness
    - ✓ Corpus Balance
    - ✓ Sampling
  - Corpus Size

Corpus Representativeness

- Representativeness is a defining feature of corpus design.
- According to Leech (1991), “A corpus is thought to be representative of the language variety, it is supposed to represent if the findings based on its contents can be generalized to the said language variety”.
- According to Biber (1993), “Representativeness refers to the extent to which a sample includes the full range of variability in a population”.
- Representativeness of a corpus are determined by the following two factors:
  - Balance – The range of genre include in a corpus.
  - Sampling – How the chunks for each genre are selected

Corpus Balance

- Important element of corpus design is corpus balance.
  - The range of genre included in a corpus.
- Representativeness of a general corpus depends upon how balanced the corpus is.
- A balanced corpus covers a wide range of text categories, which are supposed to be representatives of the language.
- We do not have any reliable scientific measure for balance but the best estimation and intuition works in this concern.
- So the accepted balance is determined by its intended uses only.

Sampling

- Corpus representativeness and balance is very closely associated with sampling.
- According to Biber(1993),
  - "Some of the first considerations in constructing a corpus concern the overall design.
  - For example, the kinds of texts included, the number of texts, the selection of particular texts, the selection of text samples from within texts, and the length of text samples.
  - Each of these involves a sampling decision, either conscious or not."
- While obtaining a representative sample, we need to consider
  - Sampling unit:
    - ✓It refers to the unit which requires a sample.
    - ✓For example, for written text, a sampling unit may be a newspaper, journal or a book.
  - Sampling frame:
    - ✓The list of al sampling units is called a sampling frame.
  - Population:
    - ✓ It may be referred as the assembly of all sampling units.
    - ✓It is defined in terms of language production, language reception or language as a product.

Corpus Size

- Another important element of corpus design is its size. How large the corpus should be?
- There is no specific answer to this question.
- The size of the corpus depends upon the purpose for which it is intended as well as on some practical considerations as follows:
  - Kind of query anticipated from the user.
  - The methodology used by the users to study the data.
  - Availability of the source of data.

Comparison

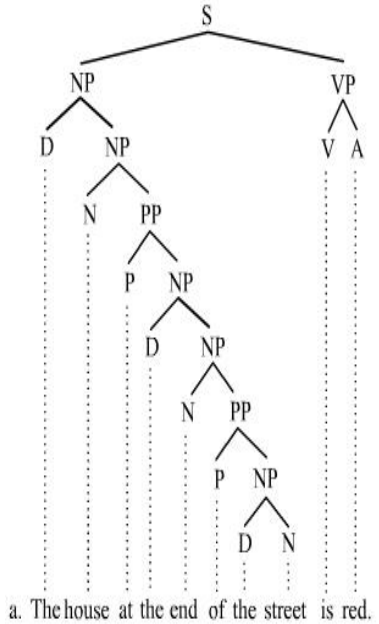
Year	Name of the Corpus	Size (in words)
1960s-70s	Brown and LOB	1 Million words
1980s	The Birmingham corpora	20 Million words
1990s	The British National corpus	100 Million words
Early 21st century	The Bank of English corpus	650 Million words

Examples of corpus

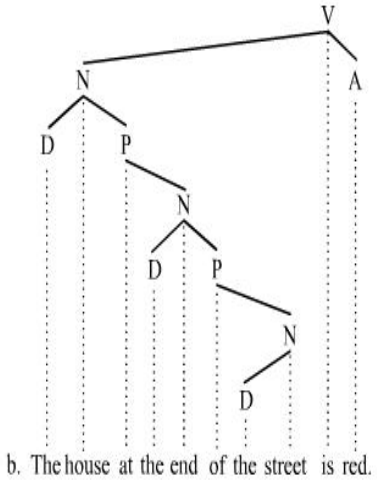
1. TreeBank Corpus
2. PropBank Corpus
3. VerbNet(VN)
4. WordNet

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.



Constituency structure

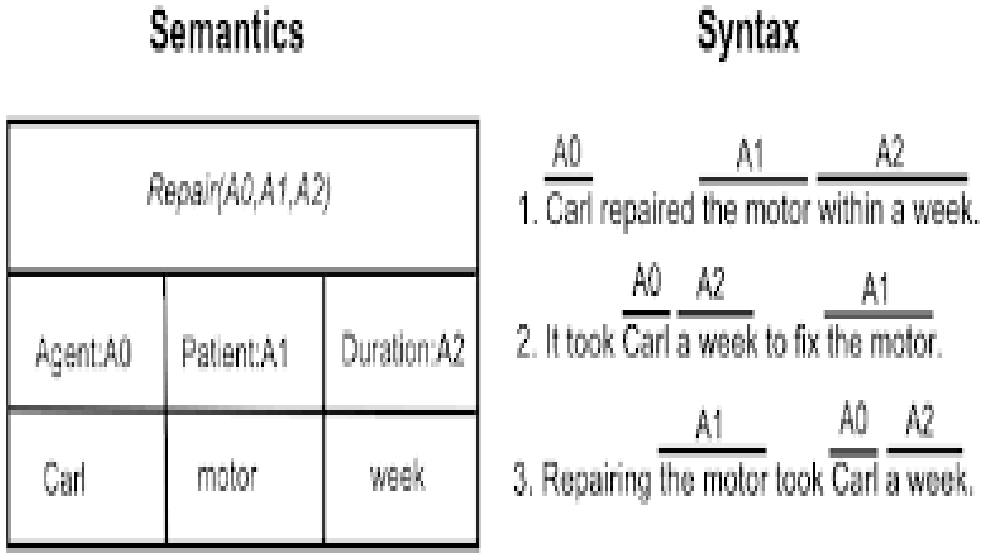


Dependency structure

Types of TreeBank Corpus

1. 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.



2. Syntactic Treebanks

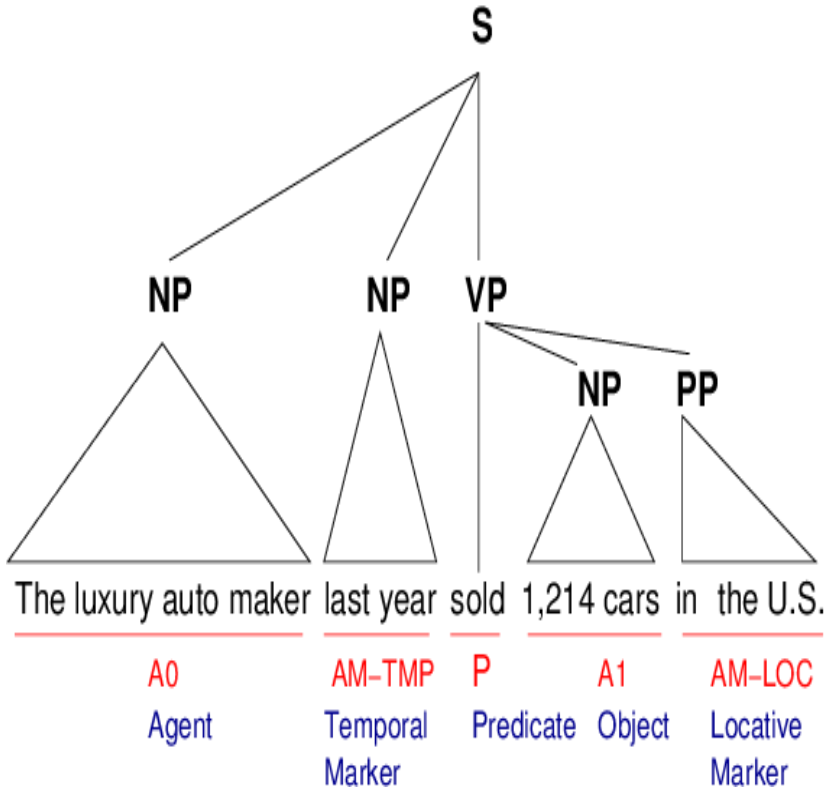
- Opposite to the semantic Treebanks, inputs to the Syntactic Treebank systems are expressions of the formal language obtained from the conversion of parsed Treebank data.
- The outputs of such systems are predicate logic based meaning representation. Various syntactic Treebanks in different languages 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

- In Computational Linguistics
  - To engineer state-of-the-art natural language processing systems such as part-of-speech taggers, parsers, semantic analyzers and machine translation systems.
- In Corpus Linguistics
  - To study syntactic phenomena.
- In Theoretical Linguistics and Psycholinguistics
  - Interaction evidence.

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.



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.

Frames

REF KEY

NP V NP

EXAMPLE "Carol cut the bread."

SYNTAX Agent V Patient

SEMANTICS cause(Agent, E) manner(during(E), Motion, Agent) contact(during(E), Instrument, Patient) degradation_material_integrity(result(E), Patient)

NP V NP PP.instrument

EXAMPLE "Carol cut the bread with a knife."

SYNTAX Agent V Patient {with} Instrument

SEMANTICS cause(Agent, E) manner(during(E), Motion, Agent) contact(during(E), Instrument, Patient) degradation_material_integrity(result(E), Patient) use(during(E), Agent, Instrument)

NP V PP

EXAMPLE "Carol cut at the bread."

SYNTAX Agent V (at) Patient

SEMANTICS cause(Agent, E) manner(during(E), Motion, Agent) contact(during(E), Instrument, Patient)

NP V PP PP

EXAMPLE "Carol cut at the bread with a knife."

SYNTAX Agent V (at) Patient {with} Instrument

SEMANTICS cause(Agent, E) manner(during(E), Motion, Agent) contact(during(E), Instrument, Patient) use(during(E), Agent, Instrument)

NP V ADVP-Middle

EXAMPLE "The bread cuts easily."

SYNTAX Patient V ADV

SEMANTICS property(Patient, Prop) Adv(Prop)

NP.instrument V NP

EXAMPLE "The knife cut the bread."

SYNTAX Instrument V Patient

SEMANTICS contact(during(E), Instrument, Patient) degradation_material_integrity(result(E), Patient)

Each VerbNet (VN) class contains:

- 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.
- 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

Types of text corpora

- **Monolingual corpus**
  - contains texts in one language only

- **Parallel corpus, multilingual corpus**
  - A parallel corpus consists of two or more monolingual corpora. The corpora are the translations of each other

- **Comparable corpus**
  - A comparable corpus is one corpus in a set of two or more monolingual corpora, typically each in a different language, built according to the same principles.

- **Diachronic corpus**
  - A diachronic corpus is a corpus containing texts from different periods and is used to study the development or change in language

- **Synchronic corpus**
  - The opposite is a **synchronic corpus** whose texts come from the same point of time.