19CSE453 – Natural Language Processing

Linguistics Resources

By

Ms. Kavitha C.R.

Dept. of Computer Science and Engineering

Amrita School of Computing, Bengaluru

Amrita Vishwa Vidyapeetham

Linguistics Resources

Linguistics is the **scientific study of human language**, meaning that it is a comprehensive, systematic, objective, and precise study of language

Linguist Resources means, **resources for the purpose of assisting linguists in their fieldwork, study and research**.
These resources include textbooks, electronic and printed reference materials, corpora, dictionary, ontology, software for linguistic data management and analysis.

Examples of Language Resources are **written and spoken corpora, computational lexica, terminology databases, speech collection, etc**.

Basic software tools are also important for the acquisition, preparation, collection, management, customization and use of these Language Resources and other resources.

Linguistics Resources

Major Areas:

Morphology - the study of the formation of words.
Syntax - the study of the formation of sentences.
Semantics - the study of meaning.
Pragmatics - the study of language use

Linguistic Tools:

Computational linguistic tools are **programs that perform operations on linguistic data**.
i.e. analysis, transformations or other tasks that add to or change language data, or that assist people in performing such tasks.

Basic Components

Linguists have identified five basic components across languages

phonology, **(study of the patterns of sounds in a language and across languages)**
morphology,
syntax,
semantics, and
Pragmatics

Introduction to Corpus

Corpus linguistics is **the study of a language as that language is expressed in its** text corpus (plural corpora), its body of "real world" text.
These results can be used to explore the relationships between that subject language and other languages which have undergone a similar analysis.

What is Corpus Linguistics?

A **corpus** is a collection of naturally-occurring language text, chosen to characterize a state of variety of a language. (J. Sinclair, 1991)

Corpus Linguistics is the study of language/linguistic phenomena through the analysis of data obtained from a corpus.

Corpus

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.

Elements in balanced corpus

•A balanced corpus covers a wide range of text categories which are supposed to be representative of the language (variety) under consideration.

•The proportions of different kinds of text it contains should correspond with informed and intuitive judgements.

•There is no scientific measure for balance – just best estimation.

•In other words, we can say that the accepted balance is determined by its intended uses only.

Sampling

Sampling

Another important element of corpus design is sampling.
Corpus representativeness and balance is very closely associated with sampling.
That is why we can say that sampling is inescapable in corpus building.

While obtaining a representative sample, we need to consider the following −

- **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 all 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.

With the advancement in technology, the corpus size also increases.
The following table of comparison will help you understand how the corpus size works −

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

TreeBank Corpus

It may be defined as linguistically parsed text corpus that annotates syntactic or semantic sentence structure.

The term ‘treebank’, represents 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.

Part of Speech Tags

- Uses:
  - Text cleaning
  - Feature engineering tasks
  - Word sense disambiguation

Sentence1 : Please book my flight for NewYork;
Sentence 2: I like to read a book on NewYork

Types of TreeBank Corpus

Semantic and Syntactic Treebanks are the two most common types of Treebanks in linguistics.

Semantic Treebanks

These Treebanks use a formal representation of sentence's semantic structure. They vary in the depth of their semantic representation.

Examples: Robot Commands Treebank, Geoquery, Groningen Meaning Bank, RoboCup Corpus

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.

Examples: **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

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.

- In Corpus Linguistics

In case of Corpus linguistics, the best use of Treebanks is to study syntactic phenomena.

- In Theoretical Linguistics and Psycholinguistics

The best use of Treebanks in theoretical and psycholinguistics is interaction evidence.

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.

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.

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 classes by refinement and addition of subclasses for achieving syntactic and semantic coherence among class members.

VerbNet(VN) class

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.

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.

WordNet

WordNet 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.

NLTK Python Tutorial (Natural Language Toolkit)

https://data-flair.training/blogs/nltk-python-tutorial/

Please refer this which will help to do the projects

https://www.nltk.org/book/

Managing Linguistic Data

Structured collections of annotated linguistic data are essential in most areas of NLP, however,

- How do we design a new language resource and ensure that its coverage, balance, and documentation support a wide range of uses?
- When existing data is in the wrong format for some analysis tool, how can we convert it to a suitable format?
- What is a good way to document the existence of a resource we have created so that others can easily find it?

The Life-Cycle of a Corpus

Corpora are not born fully-formed, but involve careful preparation and input from many people over an extended period.

Raw data needs to be collected, cleaned up, documented, and stored in a systematic structure.

Various layers of annotation might be applied, some requiring specialized knowledge of the morphology or syntax of the language.

Success at this stage depends on creating an efficient workflow involving appropriate tools and format converters.
Quality control procedures to find inconsistencies in the annotations, and to ensure the highest possible level of inter-annotator agreement.
Because of the scale and complexity of the task, large corpora may take years to prepare, and involve tens or hundreds of person-years of effort.

Three Corpus Creation Scenarios

1. the design unfolds over in the course of the creator's explorations. This is the pattern typical of traditional "field linguistics," in which material from elicitation sessions is analyzed as it is gathered, with tomorrow's elicitation often based on questions that arise in analyzing today's. The resulting corpus is then used during subsequent years of research, and may serve as an archival resource indefinitely.

2. Experimental research where a body of carefully-designed material is collected from a range of human subjects, then analyzed to evaluate a hypothesis or develop a technology. It has become common for such databases to be shared and re-used within a laboratory or company, and often to be published more widely. Corpora of this type are the basis of the "common task" method of research management, which over the past two decades has become the norm in government-funded research programs in language technology.

3. There are efforts to gather a "reference corpus" for a particular language, such as the *American National Corpus* (ANC) and the *British National Corpus* (BNC). Here the goal has been to produce a comprehensive record of the many forms, styles and uses of a language. Apart from the sheer challenge of scale, there is a heavy reliance on automatic annotation tools together with post-editing to fix any errors. However, we can write programs to locate and repair the errors, and also to analyze the corpus for balance.

Quality Control

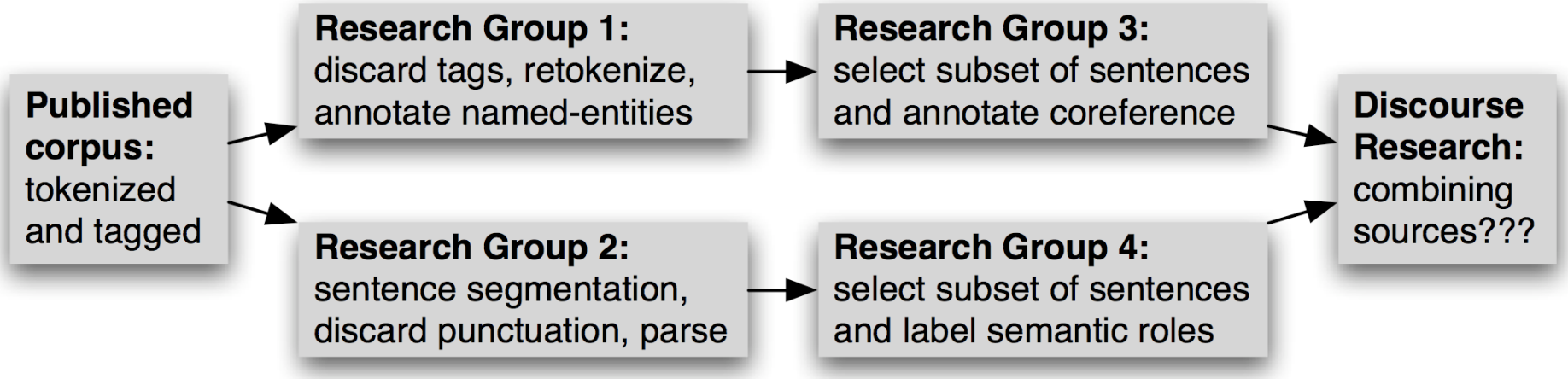
Good tools for automatic and manual preparation of data are essential.
The creation of a high-quality corpus depends on documentation, training, and workflow.

Annotation guidelines define the task and document the markup conventions. They may be regularly updated to cover difficult cases, along with new rules that are devised to achieve more consistent annotations.

Annotators need to be trained in the procedures, including methods for resolving cases not covered in the guidelines. A workflow needs to be established, possibly with supporting software, to keep track of which files have been initialized, annotated, validated, manually checked, and so on.

There may be multiple layers of annotation, provided by different specialists.

Curation vs Evolution



Acquiring Data

1. Obtaining Data from the Web
2. Obtaining Data from Word Processor Files
3. Obtaining Data from Spreadsheets and Databases
4. Converting Data Formats
5. Deciding Which Layers of Annotation to Include
6. Standards and Tools
7. Special Considerations when Working with Rare Languages

Resource management with XML

The Extensible Markup Language (XML) provides a framework for designing domain-specific markup languages.

It is sometimes used for representing annotated text and for lexical resources.
Unlike HTML with its predefined tags, XML permits us to make up our own tags.
Unlike a database, XML permits us to create data without first specifying its structure, and it permits us to have optional and repeatable elements.

Using XML for Linguistic Structures

1. flexibility and extensibility, XML is a natural choice for representing linguistic structures.
2. **well formed**
3. XML permits us to repeat elements
4. to link our lexicon to some external resource, such as WordNet, using external identifiers.

XML to represent many kinds of linguistic information. However, the flexibility comes at a price. Each time we introduce a complication, such as by permitting an element to be optional or repeated, make more work for any program that accesses the data. Make it more difficult to check the validity of the data, or to interrogate the data using one of the XML query languages.

GATE, General Architecture for Text Engineering

GATE includes an information extraction system called **ANNIE** (**A Nearly-New Information Extraction System**) which is a set of modules comprising a tokenizer, a gazetteer, a sentence splitter, a part of speech tagger, a named entities transducer and a coreference tagger. ANNIE can be used as-is to provide basic information extraction functionality, or provide a starting point for more specific tasks.

Languages currently handled in GATE include
English, Chinese, Arabic, Bulgarian, French, German, Hindi, Italian, Cebuano, Romanian, Russian, Danish.

Plugins are included for machine learning with Weka, RASP, MAXENT, SVM Light, as well as a LIBSVM integration and an in-house perceptron implementation, for managing ontologies like WordNet, for querying search engines like Google or Yahoo, for part of speech tagging with Brill or TreeTagger, and many more.
Many external plugins are also available, for handling e.g. tweets

GATE accepts input in various formats, such as
TXT, HTML, XML, Doc, PDF documents, and
Java Serial, PostgreSQL, Lucene, Oracle Databases with help of RDBMS storage over JDBC.

Thank You