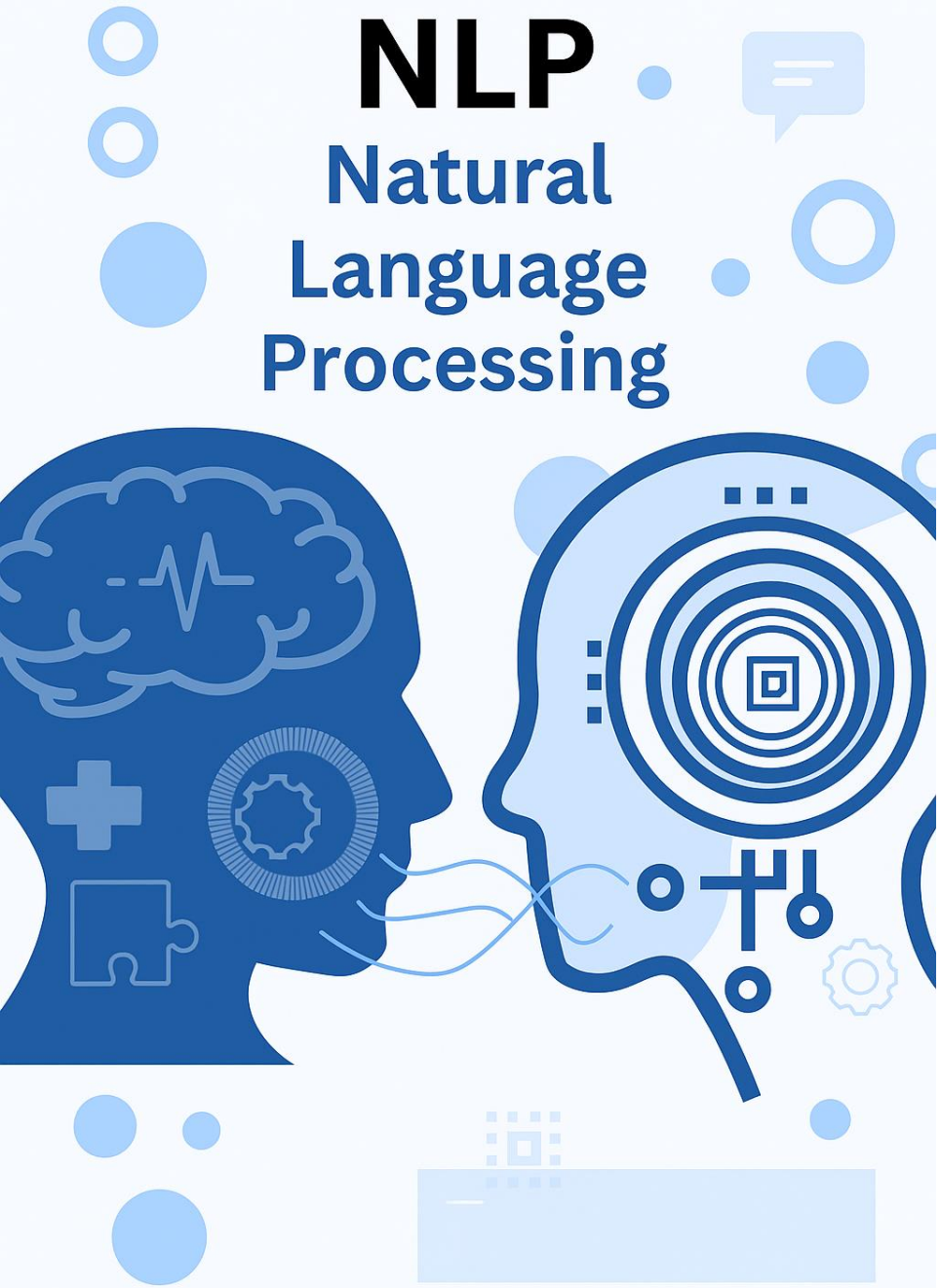


NATURAL LANGUAGE PROCESSING (NLP)

PMDS606L

MODULE 2

LECTURE 3

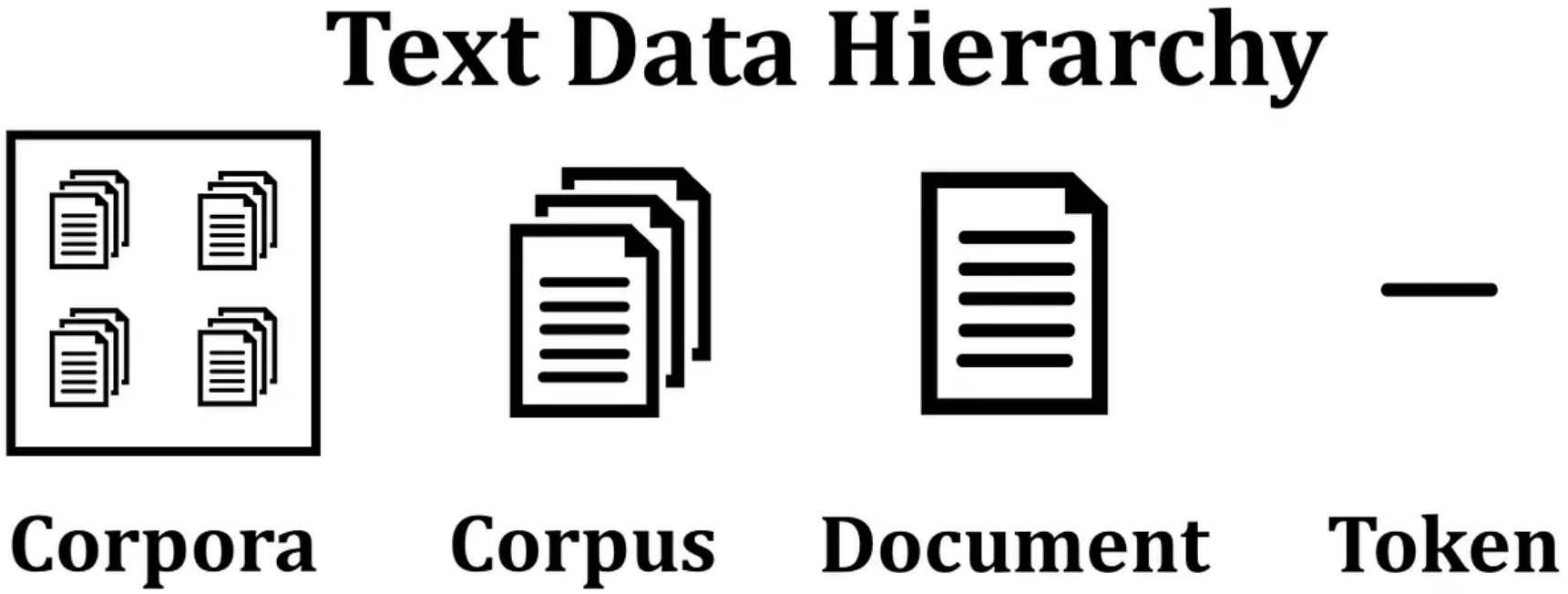
Dr. Kamanasish Bhattacharjee

Assistant Professor

Dept. of Analytics, SCOPE, VIT

CORPORA

- **Corpus** (plural: *corpora*): A structured set of texts (written or spoken) used for linguistic research, NLP modelling, or statistical analysis.



CORPORA CLASSIFICATIONS

Based on content – This determines what kind of data the corpus comprises

Based on size – This determines how large the corpus is

Based on structure – This determines the languages used and how they relate to each other

Based on purpose – This determines the specific use for the corpus

BAVL

CORPORA BASED ON CONTENT

- **Text corpora** are the most common type of corpora that contain texts from different sources.
- **Speech corpora** contain recordings of people speaking and verbatim audio transcriptions, and are often used to study how people speak a particular language or to develop speech recognition software.
- **Image corpora** contain images to develop computer vision algorithms, and usually, each image is tagged to allow for identification.
- **Video corpora** include videos and are used to create algorithms for tracking objects on video.

CORPORA BASED ON SIZE

- **Small corpora** typically comprise just a few texts and can be used for specific research tasks. For example, small corpora of medical texts might be used to study a specific disease.

- **Large corpora** are composed of hundreds or even millions of texts and are often used for general research tasks, such as studying the overall patterns of a language.

CORPORA BASED ON STRUCTURE

- **Monolingual corpora** are the most common corpora classification and contain texts only from a single language source.

- **Multilingual corpora**, simply put, contain more than one language. They can be classified further based on how the text was created and the relationship between both languages.

- **Parallel corpora** are made from two or more monolingual corpora where one corpus is the source and the second one will be a direct translation. In this type of corpora, both languages will be aligned to have matching segments at the paragraph or sentence level.

- **Comparable corpora** are made of two or more monolingual corpora built using the same principles and, therefore, offer similar results. However, as the text is not a translation of each other, they are not aligned.

CORPORA BASED ON PURPOSE

- **General corpora** contain various types of texts that can be utilized in different research fields, offering a baseline resource for general studies. The source can be written text or spoken language, along with transcriptions.

- **Specialized corpora** are designed for specific research goals containing a particular text type. These constraints can refer to a specific time frame or a particular subject, among other things.

- **Diachronic corpora** contain language data from different historical periods, and language experts use these to study the changes and development in a specific language.

- **Synchronic corpora** would be the opposite of diachronic corpora, and all texts must be compiled from the same period.

- **Monitor corpora** are diachronic and expandable. They are continuously updated to reflect the changes in language usage by incorporating new words and expressions.

- **National corpora** contain texts that represent language used in a specific country.

CORPORA BASED ON PURPOSE

- **Reference corpora,** in general terms, are used as the base of comparison with other corpora. However, these are expected to be large general corpora that offer comprehensive coverage, which the community of users can regard as the standard for the particular use case.

- **Learner corpora** include samples produced by non-native speakers of a language. This type of corpora allows researchers to compare the texts created by native speakers against those produced by language learners.

- **Developmental corpora** contain language data from monolingual speakers at different stages in their language development. These can track and understand first language acquisition and vocabulary development.

- **Raw corpora** provide no annotations or additional information and are given as originally collected.

- **Annotated corpora** contain texts annotated with information about their structure, content, or meaning. For example, a corpus of medical texts might be annotated with information about the diseases mentioned in each text. Annotated corpora are often used to develop computational linguistics applications, such as question-answering systems.

EXAMPLES FOR DIFFERENT CORPORA

Type	Description	Example
Monolingual	Texts in one language	Brown Corpus
Multilingual	Texts in multiple languages (not aligned)	Leipzig Corpora
Parallel	Texts aligned across languages (for translation)	Europarl, UN Corpus
Spoken	Transcriptions of speech (with timestamps, speaker labels, etc.)	Switchboard, TIMIT
Annotated	Contains linguistic metadata like POS, syntax, NER	Penn Treebank, CoNLL 2003
Domain-Specific	Medical, legal, financial, or social media texts	PubMed, Twitter Corpus
Diachronic	Spans multiple time periods (historical change analysis)	Corpus of Historical American English (COHA)

POPULAR NLP CORPORA

Name	Content Type	Used For
Brown Corpus	1M words of American English	POS tagging, lexical analysis
WordNet	Lexical database	Semantic similarity, ontology
SQuAD	QA pairs on Wikipedia	Question answering
IMDb/Twitter	Movie reviews or tweets	Sentiment analysis
CoNLL 2003	Newswire with NER labels	Named Entity Recognition

BUILDING A CORPUS

SIX-STEP PROCESS

1. Define the scope
2. Define the format
3. Organize the data
4. Use the right tools
5. Annotate the data
6. Analyze the data

BAVL

BUILDING A CORPUS

1. Define the scope

Decide what kind of corpus you want to create. As there are many different types of corpora, each type serves a specific purpose. Understanding exactly what kind of data you need is the first step to building an effective corpus for your project.

2. Define the format

Collect texts in whatever format your project requires. The text collection could be digital (e.g., websites or other digitally stored files) or physical (e.g., books or other printed documents). The collection stage could also require samples of spoken language that will need to be transcribed before the text can be used.

BUILDING A CORPUS

3. Organize the data

Organize your texts into a coherent structure. Doing this will make it easier to search and analyze the language data in your corpus later. Having your text divided into different categories or topics is a common first approach for text organization.

4. Use the right tools

Use a corpus-building tool or service to help create and manage your corpus. Many software options and platforms are available to help you collect or even generate new text for your project.

BUILDING A CORPUS

5. Annotate the data

Annotate your corpus with metadata. Tagging or annotations will describe the contents of each text and can be used to categorize and search the corpus for further implementation.

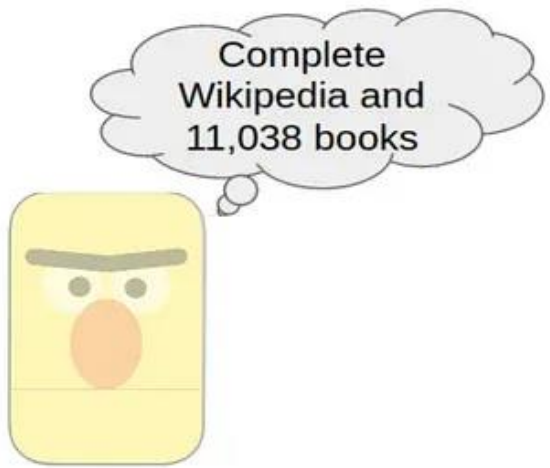
6. Analyze the data

Explore your corpus! Once you have built it, you can start to carry out all sorts of interesting analyses, such as looking at word frequencies or finding collocations and interesting language patterns.

Text corpus

(Self-supervised) Training

Pretrained LM



Adaptation

Tasks

Question Answering

Text Classification

Information Retrieval

⋮

APPLICATIONS OF CORPORA

NLP Task	Type of Corpus Used	Key Application
POS Tagging	Annotated (e.g., Brown, Penn Treebank)	Grammar tools, parsing
NER	Labeled (e.g., CoNLL-2003)	Entity detection in news, resumes
Sentiment Analysis	Opinion corpora (e.g., Twitter)	Review classification
Machine Translation	Parallel corpora (e.g., Europarl)	Cross-lingual applications
Text Classification	Topic-labeled corpora	Spam detection, document labeling
Language Modeling	Large unannotated corpora	GPT, BERT pretraining
Word Sense Disambiguation	Sense-annotated corpora (e.g., SemCor)	Semantic interpretation
Lexical Analysis	Balanced corpora (e.g., COCA, BNC)	Linguistic profiling, education
Information Retrieval	Web-scale corpora (e.g., ClueWeb)	Search systems