La-Net

# LIN309

## COMPUTATIONAL LINGUISTICS I

**Course code: LIN 309**
**Course title: Computational Linguistics I**
**Course credit load: 2 Units**

**Series 1:** Foundations of Computational Linguistics

**Series 2:** Natural Language Processing Problems and Perspectives

**Series 3:** Corpus Linguistics Essentials

**Series 4:** Probability Calculus in Language Analysis

**Series 5:** N-Grams and Language Modeling

**Proposed Outline for Series 1**

**Part .1 Concepts and Scope of Computational Linguistics**

- 1.1.1 Definition and nature of computational linguistics
- 1.1.2 Historical development and relevance to language studies
- 1.1.3 Key applications in modern contexts

**Part .2 Core Terminologies and Principles**

- 1.2.1 Essential vocabulary in computational linguistics
- 1.2.2 Principles guiding computational approaches to language
- 1.2.3 Relationship between linguistics and computer science

**Part .3 Computational Linguistics as a Discipline**

- 1.3.1 Interdisciplinary connections (linguistics, computer science, statistics)
- 1.3.2 Role in natural language processing (NLP)
- 1.3.3 Emerging trends and perspectives

**Part .4 Practical Orientation and Learning Pathways**

- 1.4.1 Skills required for computational linguistics
- 1.4.2 Typical problems addressed in the field
- 1.4.3 Preparing for advanced topics in subsequent Series

**Introduction**

Language is one of the most powerful tools humans possess, yet making computers process and interpret it requires a special field of study known as computational linguistics. This Series sets the foundation for your journey into the discipline, helping you to appreciate how linguistics and computer science come together to solve language-related problems. By engaging with this Series, you will be well prepared to tackle the more advanced topics that follow in the course.

The aim of this Series is to introduce you to the fundamental concepts, terminologies, and guiding principles of computational linguistics, and to equip you with the ability to explain its scope, relevance, and interdisciplinary nature.

We shall commence this Series by exploring the concepts and scope of computational linguistics, including its definition, nature, and historical development. Next, we will move on to examine the core terminologies and principles that underpin the field. Following that, we will consider computational linguistics as a discipline, highlighting its interdisciplinary connections and role in natural language processing. Finally, we will wrap up by discussing the practical orientation of the field, focusing on the skills required, typical problems addressed, and pathways for further study.

**Series Learning Outcomes**

Upon completing this Series, you should be able to:

- **SLO 1.1** define the concept and scope of computational linguistics,
- **SLO 1.2** explain the historical development and relevance of computational linguistics to language studies,
- **SLO 1.3** identify and describe core terminologies and guiding principles in computational linguistics,
- **SLO 1.4** analyse the interdisciplinary connections of computational linguistics with linguistics, computer science, and statistics,
- **SLO 1.5** evaluate the role of computational linguistics in natural language processing, and
- **SLO 1.6** demonstrate awareness of practical skills, typical problems, and pathways for further study in computational linguistics.

## 1.1 Concepts and Scope of Computational Linguistics

### 1.1.1 Definition and nature of computational linguistics

**Computational linguistics** is the scientific study of language using computational methods. It involves applying algorithms, models, and computer systems to analyse, process, and generate human language. This field bridges the gap between linguistics, which studies the structure and use of language, and computer science, which provides the tools and techniques to handle large-scale language data.

Computational linguistics is not limited to programming or linguistics alone. It is an interdisciplinary domain that combines insights from mathematics, statistics, psychology, and artificial intelligence. Learners should note that computational linguistics is both theoretical, in its attempt to model language, and practical, in its application to real-world problems such as machine translation and speech recognition.

Fill in the blank: Computational linguistics is the scientific study of language using __________ methods.

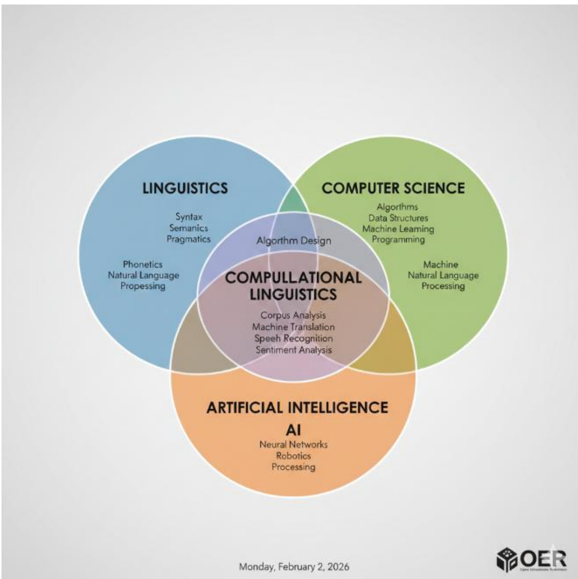



Figure 1.1 Relationship between linguistics, computer science, and computational linguistics

### 1.1.2 Historical development and relevance to language studies

The origins of computational linguistics can be traced back to the mid-twentieth century when researchers began exploring how computers could process human language. Early efforts focused on **machine translation**, which is the automatic conversion of text from one language to another. Although these early systems were limited, they laid the foundation for modern natural language processing.

Over time, computational linguistics expanded to include speech recognition, text analysis, and dialogue systems. Today, it plays a crucial role in everyday technologies such as search engines, digital assistants, and grammar checkers. Its relevance to language studies lies in its ability to provide empirical methods for analysing large corpora of text, thereby offering insights that traditional linguistic approaches may not easily uncover.

Fill in the blank: The earliest application of computational linguistics was in the field of __________ translation.

Plate 1.1 Early computer systems for language processing

### 1.1.3 Key applications in modern contexts

Modern computational linguistics has a wide range of applications. For example, **speech recognition** refers to the ability of computers to convert spoken language into text. **Text-to-speech synthesis** allows computers to generate spoken language from written text. **Information retrieval** enables search engines to locate relevant documents quickly, while **sentiment analysis** helps businesses evaluate customer opinions expressed in text.

These applications demonstrate how computational linguistics impacts daily life, from using mobile phones to interacting with intelligent assistants. Learners should recognise that the scope of computational linguistics is not confined to academic research but extends to practical tools that shape communication and information access.

Fill in the blank: Sentiment analysis is used to evaluate __________ expressed in text.

| Application | Description | Real-World Example |
|---|---|---|
| **Speech Recognition** | Converting spoken language into text by identifying phonetic patterns and linguistic context. | Voice assistants like **Siri** or **Alexa**; automated captioning. |
| **Text-to-Speech (TTS)** | Synthesizing artificial human speech from written text using prosody and intonation rules. | **Screen readers** for the visually impaired; GPS navigation voices. |
| **Machine Translation** | Automatically translating text or speech from one language to another using neural networks. | **Google Translate**; real-time translation during video calls. |
| **Sentiment Analysis** | Identifying and extracting subjective information (emotions, opinions) from text. | Analyzing **Amazon product reviews** or social media feedback for brand sentiment. |

| Application | Description | Real-World Example |
|---|---|---|
| **Information Retrieval** | Modeling how to find and extract relevant information from large-scale databases or the web. | **Search engines** like Google; academic database indexing. |

Box 1.1 Selected applications of computational linguistics

### Pilot questions 1.1

1. Define computational linguistics in your own words.
2. Identify two disciplines that contribute to computational linguistics.
3. What was the earliest application of computational linguistics?
4. Mention two modern applications of computational linguistics.
5. Explain why computational linguistics is relevant to language studies.

## 1.2 Core Terminologies and Principles

### 1.2.1 Essential vocabulary in computational linguistics

**Terminology** is central to mastering computational linguistics. Some of the most important terms include:

- **Algorithm**: A step-by-step procedure for solving a problem or performing a task.
- **Corpus**: A large collection of texts used for linguistic analysis.
- **Parsing**: The process of analysing the grammatical structure of a sentence.
- **Tokenisation**: The division of text into smaller units such as words or sentences.

These terms will recur throughout the course, so it is important to be comfortable with them.

Fill in the blank: The process of dividing text into smaller units such as words or sentences is called ___________.

| Term | Definition | Role in the Process |
|---|---|---|
| **Corpus** | A large, structured collection of machine-readable texts used for linguistic analysis. | Provides the **raw data** and real-world evidence for training models. |
| **Tokenization** | The process of breaking down a stream of text into smaller units, such as words or phrases (tokens). | The **first step** in processing; it turns a sentence into a list of individual items. |

| Term | Definition | Role in the Process |
|---|---|---|
| **Parsing** | Analyzing a string of tokens to determine its grammatical structure (e.g., identifying subjects and verbs). | Assigns **syntactic meaning**; it helps the computer understand "who did what to whom." |
| **Algorithm** | A set of defined rules or instructions used by a computer to solve a specific problem or perform a task. | The **engine** that drives the analysis, from simple sorting to complex neural network predictions. |

Box 1.2 Selected computational linguistics terms

### 1.2.2 Principles guiding computational approaches to language

Computational linguistics is guided by several principles. One principle is **formalisation**, which means representing language phenomena in precise, rule-based or mathematical terms. Another is **empiricism**, which stresses the importance of using real-world data to test theories. A third principle is **automation**, which involves designing systems that can process language with minimal human intervention.

These principles ensure that computational linguistics remains both scientific and practical.

Fill in the blank: The principle that stresses the importance of using real-world data to test theories is called __________.

### 1.2.3 Relationship between linguistics and computer science

Computational linguistics exists at the intersection of **linguistics** and **computer science**. Linguistics provides the theoretical understanding of language, while computer science contributes the computational tools and methods. Together, they enable the development of systems that can analyse, interpret, and generate human language.

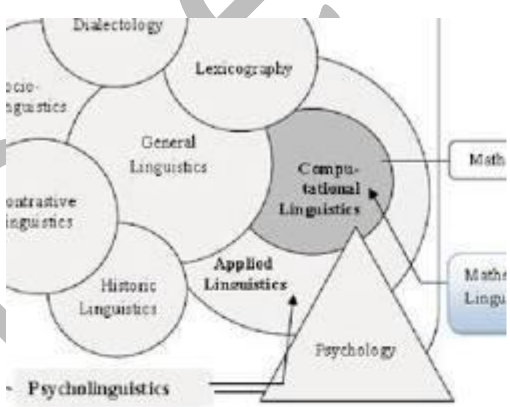


Figure 1.2 Intersection of linguistics and computer science

### Pilot questions 1.2

1. Define the term algorithm in computational linguistics.
2. What is a corpus used for?
3. State the principle that involves representing language phenomena in precise terms.
4. Explain the role of empiricism in computational linguistics.
5. How do linguistics and computer science contribute to computational linguistics?

## 1.3 Computational Linguistics as a Discipline

### 1.3.1 Interdisciplinary connections (linguistics, computer science, statistics)

**Interdisciplinary connections** refer to the way computational linguistics draws upon multiple fields. From **linguistics**, it gains theories of grammar, semantics, and pragmatics. From **computer science**, it adopts algorithms, programming, and data structures. From **statistics**, it borrows methods for analysing large datasets and modelling probabilities. Together, these disciplines provide the tools and frameworks needed to study language computationally.

Fill in the blank: Computational linguistics borrows methods for analysing large datasets and modelling probabilities from __________.

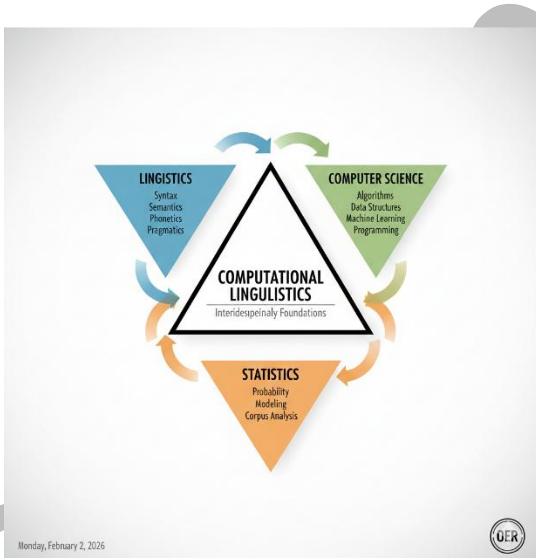



Figure 1.3 Interdisciplinary foundations of computational linguistics

### 1.3.2 Role in natural language processing (NLP)

**Natural language processing (NLP)** is the branch of artificial intelligence concerned with enabling computers to understand, interpret, and generate human language. Computational linguistics provides the theoretical and methodological backbone for NLP. For example, parsing techniques from linguistics are used in syntactic analysis, while statistical models are applied in machine translation.

Fill in the blank: The branch of artificial intelligence concerned with enabling computers to understand human language is called __________.

### 1.3.3 Emerging trends and perspectives

Emerging trends in computational linguistics include the use of **deep learning**, which involves training neural networks to process language data, and **multilingual NLP**, which aims to create systems that can handle multiple languages effectively. Another perspective is the ethical dimension, where researchers consider issues such as bias in language models and responsible use of language technologies.

| Trend | Description | Key Impact |
|---|---|---|
| **Deep Learning & World Models** | Systems that move beyond predicting the next word to creating internal simulations of cause-and-effect and physical reality. | Enables **grounded reasoning**, allowing AI to understand the physical and social logic behind a sentence. |
| **Multilingual & Polyglot Equity** | A shift from English-centric models to "cross-lingual transfer" that supports low-resource and endangered languages. | Promotes **global accessibility**, ensuring linguistic technology serves diverse cultures rather than just a few dominant ones. |
| **Ethical & Emotional Intelligence** | Integration of emotional recognition and bias-correction layers (like SHAP/LIME) into standard workflows. | Creates **fairer, empathetic systems** that can support mental health or act as unbiased sounding boards. |

Box 1.3 Emerging trends in computational linguistics

### Pilot questions 1.3

1. Name three disciplines that contribute to computational linguistics.
2. What role does computational linguistics play in NLP?
3. Define NLP in your own words.
4. Mention one emerging trend in computational linguistics.
5. Why is ethics considered important in computational linguistics?

## 1.4 Practical Orientation and Learning Pathways

### 1.4.1 Skills required for computational linguistics

To succeed in computational linguistics, learners need a combination of skills. These include **linguistic analysis**, which involves understanding grammar and meaning, **programming skills**, which allow you to implement algorithms, and **statistical reasoning**, which helps in interpreting data. Developing these skills ensures that learners can engage with both theoretical and applied aspects of the field.

Fill in the blank: The ability to implement algorithms in computational linguistics requires strong __________ skills.

### 1.4.2 Typical problems addressed in the field

Computational linguistics addresses problems such as **machine translation**, **speech recognition**, **information retrieval**, and **text summarisation**. Each of these problems requires different methods, but they all share the goal of enabling computers to process human language effectively.

| Problem | Explanation | Computational Challenge |
|---|---|---|
| **Machine Translation (MT)** | Automatically converting a text or speech from a "source" language to a "target" language. | Handling **syntactic differences** (word order) and **polysemy** (words with multiple meanings). |
| **Speech Recognition (ASR)** | Converting acoustic signals (sound waves) into machine-readable text. | Managing **accents, dialects**, and background noise that distort the input signal. |
| **Information Retrieval (IR)** | Searching and extracting relevant documents or data from a massive, unstructured repository. | Determining **relevance** beyond simple keyword matching using semantic understanding. |
| **Text Summarization** | Producing a concise version of a longer text while retaining its key information and meaning. | Choosing between **extraction** (copying sentences) and **abstraction** (generating new text). |

Table 1.1 Typical problems in computational linguistics

### 1.4.3 Preparing for advanced topics in subsequent Series

This Series lays the groundwork for more advanced topics. Learners will later explore natural language processing problems, corpus linguistics, probability calculus, and language modelling. By mastering the foundations, you will be better prepared to analyse complex systems and design solutions in later Series.

Fill in the blank: The next Series after Foundations of Computational Linguistics will focus on __________ problems and perspectives.

### Pilot questions 1.4

1. List three skills required for computational linguistics.
2. Identify two typical problems addressed in computational linguistics.
3. What is the goal shared by problems such as machine translation and speech recognition?
4. How does this Series prepare learners for subsequent topics?
5. Which skill helps in interpreting data in computational linguistics?

## *Series Summary*

This Series has introduced you to the foundations of computational linguistics, setting the stage for your study of how computers can process and analyse human language. Its purpose has been to familiarise you with the scope, principles, and practical orientation of the discipline, while highlighting its relevance to both language studies and everyday applications.

We began by defining computational linguistics and examining its scope and historical development, then moved on to explore essential terminologies and guiding principles. Following that, we considered computational linguistics as a discipline, focusing on its interdisciplinary connections and role in natural language processing, before concluding with the skills required, typical problems addressed, and pathways for further study. By engaging with these Parts, you should now be able to define the concept and scope of computational linguistics, explain its historical relevance, describe key terms and principles, analyse its interdisciplinary nature, evaluate its role in NLP, and demonstrate awareness of practical skills and problems, as outlined in the Series Learning Outcomes.

## *Glossary of Terms*

- **Algorithm**: A step-by-step procedure or set of rules used by computers to solve problems or perform tasks.
- **Corpus**: A large, structured collection of texts that is used for linguistic analysis and research.
- **Computational linguistics**: The scientific study of language using computational methods, combining linguistics, computer science, and statistics to analyse and process human language.
- **Empiricism**: A principle in computational linguistics that emphasises the use of real-world data to test theories and models.
- **Formalisation**: The process of representing language phenomena in precise, rule-based, or mathematical terms.
- **Machine translation**: The automatic conversion of text from one language into another using computational methods.

- **Natural language processing (NLP)**: A branch of artificial intelligence concerned with enabling computers to understand, interpret, and generate human language.
- **Parsing**: The process of analysing the grammatical structure of a sentence to determine how words relate to each other.
- **Sentiment analysis**: A computational method used to evaluate opinions, emotions, or attitudes expressed in text.
- **Speech recognition**: The ability of computers to convert spoken language into written text.
- **Statistics**: A discipline that provides methods for analysing large datasets and modelling probabilities, often applied in computational linguistics.
- **Tokenisation**: The process of dividing text into smaller units such as words, phrases, or sentences for analysis.

## *Concept Check Questions [Series 1]*

### Objective questions

1. Computational linguistics is the scientific study of language using __________ methods. (Linked to SLO 2.1)
2. The earliest application of computational linguistics was in the field of __________ translation. (Linked to SLO 2.2)
3. The process of dividing text into smaller units such as words or sentences is called __________. (Linked to SLO 2.3)
4. The branch of artificial intelligence concerned with enabling computers to understand human language is called __________. (Linked to SLO 2.5)
5. Which principle in computational linguistics emphasises the use of real-world data to test theories? (Linked to SLO 2.3)

### Short-answer questions

6. Define the term **algorithm** in the context of computational linguistics. (Linked to SLO 2.3)
7. Explain why computational linguistics is considered interdisciplinary. (Linked to SLO 2.4)
8. Identify two typical problems addressed in computational linguistics. (Linked to SLO 2.6)
9. Describe the role of computational linguistics in natural language processing. (Linked to SLO 2.5)
10. State one emerging trend in computational linguistics and explain its significance. (Linked to SLO 2.4)

### Long-answer questions

11. Analyse the historical development of computational linguistics and evaluate its relevance to language studies today. (Linked to SLO 2.2)
12. Discuss the skills required for computational linguistics and demonstrate how they prepare learners for advanced topics in the field. (Linked to SLO 2.6)

## *Answers to Concept Check Questions [Series 1]*

**Objective questions**

1. Computational methods.
2. Machine translation.
3. Tokenisation.
4. Natural language processing (NLP).
5. Empiricism.

**Short-answer questions**

6. An algorithm is a step-by-step procedure or set of rules used by computers to solve problems or perform tasks.
7. Computational linguistics is interdisciplinary because it combines theories from linguistics, computational tools from computer science, and statistical methods for data analysis.
8. Examples include machine translation and speech recognition.
9. Computational linguistics provides the theoretical and methodological backbone for NLP, enabling computers to analyse, interpret, and generate human language.
10. Deep learning is an emerging trend that uses neural networks to process language data, making NLP systems more accurate and versatile.

**Long-answer questions**

11. *Basic response*: Computational linguistics began in the mid-twentieth century with early machine translation projects. Over time, it expanded to include speech recognition, text analysis, and dialogue systems. Its relevance today lies in its ability to support technologies such as search engines, digital assistants, and grammar checkers. *Value-added response*: Beyond these applications, computational linguistics has transformed language studies by enabling large-scale corpus analysis and empirical testing of linguistic theories. It also raises important ethical questions about bias in language models and the responsible use of language technologies.
12. *Basic response*: Skills required include linguistic analysis, programming, and statistical reasoning. These skills prepare learners to engage with both theoretical and applied aspects of computational linguistics. *Value-added response*: Outstanding learners may also develop expertise in machine learning, data visualisation, and multilingual NLP. These advanced skills not only prepare them for subsequent Series in this course but also position them to contribute to cutting-edge research and industry applications.

*Answers to Pilot Questions [Series 1]*

**Part 1.1 Concepts and Scope of Computational Linguistics**

1. Computational linguistics is the scientific study of language using computational methods.
2. Linguistics and computer science.
3. Machine translation.
4. Speech recognition and sentiment analysis.
5. It provides empirical methods for analysing large corpora and supports practical applications in language technology.

### Part 1.2 Core Terminologies and Principles

1. An algorithm is a step-by-step procedure used to solve problems or perform tasks.
2. A corpus is used for linguistic analysis and research.
3. Formalisation.
4. Empiricism ensures theories are tested against real-world data.
5. Linguistics contributes theories of language, while computer science provides computational tools and methods.

### Part 1.3 Computational Linguistics as a Discipline

1. Linguistics, computer science, and statistics.
2. It provides the theoretical and methodological backbone for NLP.
3. Natural language processing is the branch of AI concerned with enabling computers to understand and generate human language.
4. Deep learning.
5. Ethics is important to address bias in language models and ensure responsible use of language technologies.

### Part 1.4 Practical Orientation and Learning Pathways

1. Linguistic analysis, programming skills, and statistical reasoning.
2. Machine translation and text summarisation.
3. They all aim to enable computers to process human language effectively.
4. It introduces foundational concepts, preparing learners for advanced topics such as corpus linguistics and language modelling.
5. Statistical reasoning.

## *References and Attributions [Series 1]*

This section compiles all references and attributions for Series 1: *Foundations of Computational Linguistics*. Entries include suggested open educational resources (OERs) and notes on AI-generated illustration prompts. The referencing style follows APA (7th edition).

### General References


- Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Draft manuscript. Retrieved from https://web.stanford.edu/~jurafsky/slp3/
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.


### Illustrations

- *Figure 1.1 Relationship between linguistics, computer science, and computational linguistics*

- Attribution: Suggested OER diagram from Wikimedia Commons (search string: "computational linguistics interdisciplinary diagram OER").
- AI prompt: "Generate a Venn diagram showing overlap between linguistics, computer science, and computational linguistics."

- *Plate 1.1 Early computer systems for language processing*
    - Attribution: Suggested OER image from Wikimedia Commons (search string: "early machine translation history OER image").
    - AI prompt: "Generate a black-and-white illustration of early computer systems used for language translation."
- *Box 1.1 Selected applications of computational linguistics*
    - Attribution: Suggested OER table resource (search string: "applications of computational linguistics OER table").
    - AI prompt: "Generate a table listing five applications of computational linguistics with short descriptions."
- *Box 1.2 Selected computational linguistics terms*
    - Attribution: Suggested OER glossary resource (search string: "computational linguistics glossary OER").
    - AI prompt: "Generate a table listing four key computational linguistics terms with definitions."
- *Figure 1.2 Intersection of linguistics and computer science*
    - Attribution: Suggested OER diagram from Wikimedia Commons (search string: "computational linguistics interdisciplinary diagram OER").
    - AI prompt: "Generate a diagram showing overlap between linguistics and computer science with computational linguistics at the centre."
- *Figure 1.3 Interdisciplinary foundations of computational linguistics*
    - Attribution: Suggested OER diagram (search string: "computational linguistics interdisciplinary foundations OER diagram").
    - AI prompt: "Generate a triangular diagram showing linguistics, computer science, and statistics feeding into computational linguistics."
- *Box 1.3 Emerging trends in computational linguistics*
    - Attribution: Suggested OER resource (search string: "emerging trends computational linguistics OER").
    - AI prompt: "Generate a table listing three emerging trends in computational linguistics with short notes."
- *Table 1.1 Typical problems in computational linguistics*

- Attribution: Suggested OER table resource (search string: “computational linguistics problems OER table”).
- AI prompt: “Generate a table listing four typical problems in computational linguistics with explanations.”

**Course code: LIN 309**
**Course title: Computational Linguistics I**
**Course credit load: 2 Units**

**Series 1:** Foundations of Computational Linguistics

**Series 2:** Natural Language Processing Problems and Perspectives

**Series 3:** Corpus Linguistics Essentials

**Series 4:** Probability Calculus in Language Analysis

**Series 5:** N-Grams and Language Modeling

**Proposed Outline for Series 2**

**Part .1 Introduction to Natural Language Processing Problems**

- 2.1.1 Definition and scope of NLP problems
- 2.1.2 Categories of NLP tasks (syntactic, semantic, pragmatic)
- 2.1.3 Common challenges in NLP

**Part .2 Core Problems in NLP**

- 2.2.1 Machine translation challenges
- 2.2.2 Speech recognition and synthesis issues
- 2.2.3 Information retrieval and text summarisation problems

**Part .3 Perspectives in NLP Development**

- 2.3.1 Rule-based approaches
- 2.3.2 Statistical and probabilistic approaches
- 2.3.3 Machine learning and deep learning perspectives

**Part .4 Practical Implications and Future Directions**

- 2.4.1 Applications in everyday technology
- 2.4.2 Ethical considerations in NLP

- 2.4.3 Future trends and research pathways

**Introduction**

In the last Series, we explored the foundations of computational linguistics, focusing on its concepts, scope, and interdisciplinary nature. Building on that knowledge, we now turn to natural language processing, which is one of the most practical and visible applications of computational linguistics. This Series is important because it helps you to recognise the challenges computers face when dealing with human language and to appreciate the different perspectives that have shaped solutions in this field.

The aim of this Series is to examine the major problems encountered in natural language processing and to present the perspectives and approaches that have been developed to address them. By doing so, the Series intends to equip you with the ability to analyse these challenges and evaluate the methods used to overcome them.

We shall commence this Series by introducing the scope of natural language processing problems, including their categories and common challenges. Next, we will move on to core problems such as machine translation, speech recognition, and information retrieval. Following that, we will consider perspectives in NLP development, ranging from rule-based approaches to statistical and machine learning methods. Finally, we will wrap up by discussing practical implications, ethical considerations, and future directions in NLP research and applications.

**Series Learning Outcomes**

Upon completing this Series, you should be able to:

- **SLO 2.1** define the scope of natural language processing problems and categorise them into syntactic, semantic, and pragmatic tasks,
- **SLO 2.2** explain common challenges encountered in natural language processing,
- **SLO 2.3** analyse core problems such as machine translation, speech recognition, and information retrieval,
- **SLO 2.4** evaluate different perspectives in NLP development, including rule-based, statistical, and machine learning approaches, and
- **SLO 2.5** assess the practical implications of NLP applications, including ethical considerations and future research directions.

## 2.1 Introduction to Natural Language Processing Problems

### 2.1.1 Definition and scope of NLP problems

**Natural language processing (NLP)** is a branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. The scope of NLP problems is broad, ranging from recognising spoken words to analysing written text. These problems arise because human language is complex, ambiguous, and context-dependent, making it difficult for machines to process in the same way humans do.

Learners should note that NLP problems are not confined to one area of language but span multiple levels, including syntax, semantics, and pragmatics. Each level presents unique challenges that require specialised approaches.

Fill in the blank: Natural language processing is a branch of __________ that enables computers to work with human language.

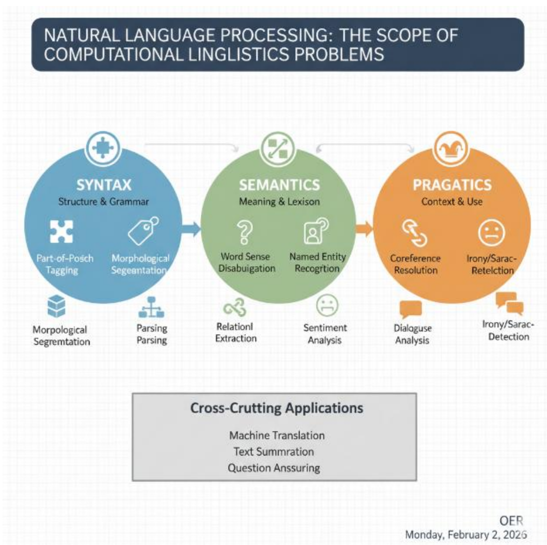



Figure 2.1 Scope of NLP problems

### 2.1.2 Categories of NLP tasks (syntactic, semantic, pragmatic)

NLP tasks can be grouped into three broad categories. **Syntactic tasks** deal with the structure of language, such as parsing sentences and identifying parts of speech. **Semantic tasks** focus on meaning, including word sense disambiguation and semantic role labelling. **Pragmatic tasks** consider context and usage, such as understanding implied meaning or managing dialogue in conversation systems.

These categories help us organise the wide range of NLP problems and provide a framework for studying them systematically.

Fill in the blank: Tasks that deal with the structure of language, such as parsing, are called __________ tasks.

| Category | Definition | Key Tasks | Example Application |
|---|---|---|---|
| **Syntactic** | Focuses on the **structure** and grammar of a sentence, regardless of its meaning. | Part-of-Speech (POS) Tagging, Lemmatization, Parsing. | Identifying that in *"The cat sat,"* "cat" is the **noun/subject**. |
| **Semantic** | Focuses on the **literal meaning** of words and how they combine into sentences. | Named Entity Recognition (NER), Word Sense Disambiguation. | Distinguishing between "bank" as a **financial institution** vs. a **river edge**. |
| **Pragmatic** | Focuses on the **intended meaning** based on context, social cues, and world knowledge. | Coreference Resolution, Sentiment Analysis, Sarcasm Detection. | Understanding that *"Do you have a watch?"* is a **request for the time**, not just a yes/no question. |

Table 2.1 Categories of NLP tasks

**2.1.3 Common challenges in NLP**

Several challenges make NLP particularly difficult. One challenge is **ambiguity**, where words or sentences can have multiple meanings. Another is **context-dependence**, since meaning often relies on surrounding text or situation. A third challenge is **resource limitation**, as many languages lack large corpora or annotated datasets for training computational models.

These challenges highlight why NLP remains an active area of research and why solutions often require combining linguistic theory with computational techniques.

Fill in the blank: When a word or sentence has multiple possible meanings, this is referred to as __________.

| Challenge | Description | Why It's Hard for Machines |
|---|---|---|
| **Ambiguity** | When a word, phrase, or sentence has multiple possible meanings. | A machine must decide if *"bank"* is a place for money or a river's edge without the "common sense" humans have. |
| **Context-Dependence** | Meaning that changes based on surrounding sentences or the physical situation. | Machines struggle with **anaphora** (e.g., in *"The trophy didn't fit in the* |

| Challenge | Description | Why It's Hard for Machines |
|---|---|---|
| | | *suitcase because it was too large,"* what does "it" refer to?). |
| **Resource Limitation** | The lack of large-scale, high-quality digital data for most of the world's 7,000+ languages. | Most progress is in "high-resource" languages like English; "low-resource" languages lack the data needed to train robust models. |

Box 2.1 Common challenges in NLP

### Pilot questions 2.1

1. Define natural language processing in your own words.
2. Identify the three broad categories of NLP tasks.
3. Give one example of a syntactic task in NLP.
4. What is ambiguity in the context of NLP?
5. Mention one reason why NLP problems remain difficult to solve.

## 2.2 Core Problems in NLP

### 2.2.1 Machine translation challenges

**Machine translation** refers to the automatic conversion of text from one language into another. While it has advanced significantly, challenges remain. One major issue is **structural differences** between languages, such as word order or grammatical rules. Another is **idiomatic expressions**, which often do not translate literally. A further challenge is **context sensitivity**, since the meaning of a word or phrase may change depending on its usage.

Fill in the blank: When a phrase cannot be translated literally because its meaning depends on usage, it is called an __________ expression.

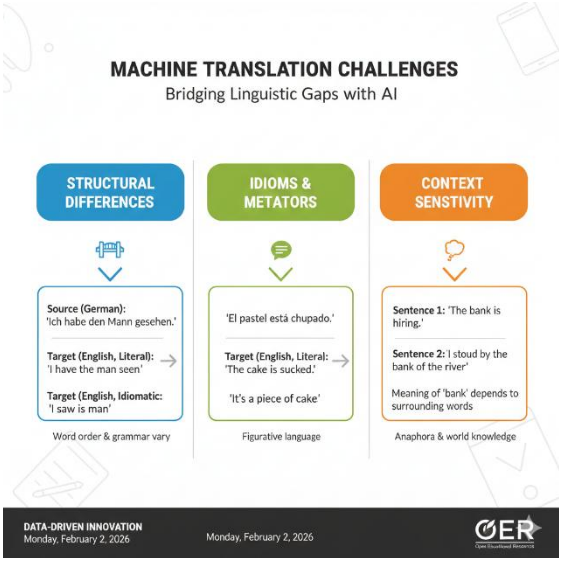



Figure 2.2 Challenges in machine translation

### 2.2.2 Speech recognition and synthesis issues

**Speech recognition** is the process of converting spoken language into text, while **speech synthesis** refers to generating spoken language from text. Both face challenges. Recognition systems struggle with **accent variation**, **background noise**, and **homophones** (words that sound alike but have different meanings). Synthesis systems, on the other hand, must produce speech that sounds natural, with appropriate **intonation** and **prosody**.

Fill in the blank: Words that sound alike but have different meanings are called __________.


Plate 2.1 Speech recognition in practice

### 2.2.3 Information retrieval and text summarisation problems

**Information retrieval** involves finding relevant documents or data from large collections, such as search engines. The main challenge is **relevance ranking**, ensuring that the most useful results appear first. **Text summarisation** aims to condense large texts into shorter versions while retaining meaning. Its challenges include **preserving accuracy**, **avoiding bias**, and **capturing key ideas** without oversimplification.

Fill in the blank: The main challenge in information retrieval is ensuring that the most useful results appear __________.

| Problem | Description | Computational Hurdle |
|---|---|---|
| **Relevance Ranking** | Determining which documents or sentences best satisfy a user's specific information need. | Moving beyond keyword matching to **semantic intent** (e.g., knowing "Java" means the code, not the coffee). |
| **Preserving Accuracy** | Ensuring that condensed text (summaries) does not introduce "hallucinations" or false info. | Maintaining **factual consistency** when the model rephrases or connects distant ideas. |
| **Capturing Key Ideas** | Identifying the "salience" or core message of a text while ignoring redundant or trivial details. | Distinguishing between **supporting evidence** and the **primary thesis** of a document. |
| **Coherence & Flow** | Linking extracted sentences or generated ideas into a logical, readable narrative. | Solving **coreference issues** (e.g., ensuring "he" or "it" refers to the correct person/thing in the summary). |

Box 2.2 Common problems in information retrieval and text summarisation

Pilot questions 2.2

1. What is machine translation and what are its main challenges?
2. Give one example of a difficulty faced in speech recognition.
3. Define speech synthesis and state one of its challenges.
4. What is the main challenge in information retrieval?
5. Mention one difficulty in text summarisation.

## 2.3 Perspectives in NLP Development

### 2.3.1 Rule-based approaches

**Rule-based approaches** rely on manually crafted linguistic rules to process language. These systems use grammars, dictionaries, and explicit instructions to analyse or generate text. While they can be precise for well-defined tasks, they struggle with flexibility and scalability, especially when dealing with the variability of natural language.

Fill in the blank: Systems that rely on manually crafted linguistic rules are called __________ approaches.

| Component | Description | Function |
|---|---|---|
| **Lexicon (Dictionary)** | A comprehensive list of words with their categories (noun, verb, etc.). | Provides the basic building blocks for the system. |
| **Grammar Rules** | "If-Then" statements (e.g., *If Noun + Verb, then valid sentence*). | Defines the structural relationships between words. |
| **Parser** | The engine that matches the input text against the grammar rules. | Creates a structural representation (like a parse tree) of the input. |
| **Text Analysis** | The final output where the system assigns meaning based on the rules. | Delivers the structured result (e.g., "This is a question about weather"). |

Figure 2.3 Rule-based approach in NLP

**2.3.2 Statistical and probabilistic approaches**

**Statistical approaches** use large datasets and probability models to make predictions about language. For example, they can estimate the likelihood of a word appearing in a given context. These methods are more adaptable than rule-based systems because they learn patterns from data rather than relying solely on fixed rules. However, they require substantial amounts of annotated data to perform well.

Fill in the blank: Approaches that estimate the likelihood of words appearing in context are called __________ approaches.

| Feature | Rule-Based NLP | Statistical NLP |
|---|---|---|
| **Flexibility** | **Low:** Rigid and struggles with slang, typos, or evolving language. | **High:** Adaptable to diverse datasets and handles variation well. |
| **Data Requirements** | **Minimal:** Requires expert linguistic knowledge rather than large corpora. | **Extensive:** Needs massive amounts of annotated data to "learn" patterns. |
| **Accuracy (Precision)** | **High in Narrow Domains:** Extremely precise when rules are well-defined. | **High Overall:** Better at capturing semantic nuances and broad intent. |

| Feature | Rule-Based NLP | Statistical NLP |
|---|---|---|
| **Transparency** | **White-Box:** Decisions are easy to trace back to a specific human-written rule. | **Black-Box:** Often difficult to understand *why* a model made a specific prediction. |
| **Maintenance** | **Manual:** Rules must be updated by hand as the language changes. | **Automatic:** The model can be "re-trained" on new data to stay current. |

Table 2.2 Features of statistical and probabilistic approaches

### 2.3.3 Machine learning and deep learning perspectives

**Machine learning approaches** involve training algorithms to learn from data and improve performance over time. **Deep learning**, a subset of machine learning, uses neural networks with multiple layers to capture complex patterns in language. These perspectives have revolutionised NLP, enabling breakthroughs in machine translation, sentiment analysis, and conversational agents.

Fill in the blank: Deep learning uses __________ networks with multiple layers to capture complex language patterns.

| Approach | Core Philosophy | Key Characteristics |
|---|---|---|
| **Rule-Based** | **Hand-coded Logic:** Relies on human-written grammar rules and vast dictionaries. | Highly transparent and predictable, but "brittle"—it breaks easily when it encounters slang or typos. |
| **Statistical** | **Probability-Based:** Uses mathematical models to predict the most likely meaning based on frequency in a corpus. | More flexible than rules; handles variation better but requires large amounts of manually annotated data. |
| **Deep Learning** | **Neural Networks:** Uses multi-layered architectures (like Transformers) to learn abstract representations of language. | Exceptional at capturing context and nuance; requires massive data and compute power, but often lacks transparency. |

Box 2.3 Perspectives in NLP development

## Pilot questions 2.3

1. What are rule-based approaches in NLP?
2. State one limitation of rule-based systems.

3. What do statistical approaches rely on to make predictions?
4. Define deep learning in the context of NLP.
5. Mention one breakthrough enabled by deep learning in NLP.

## 2.4 Practical Implications and Future Directions

### 2.4.1 Applications in everyday technology

Natural language processing is embedded in many technologies we use daily. Examples include **digital assistants** such as Siri or Alexa, which rely on speech recognition and dialogue management, and **search engines**, which use information retrieval techniques to provide relevant results. Other applications include **chatbots** for customer service and **grammar checkers** that support writing tasks. These examples show how NLP has moved beyond research labs into tools that shape communication and productivity.

Fill in the blank: Digital assistants such as Siri or Alexa rely on __________ recognition and dialogue management.

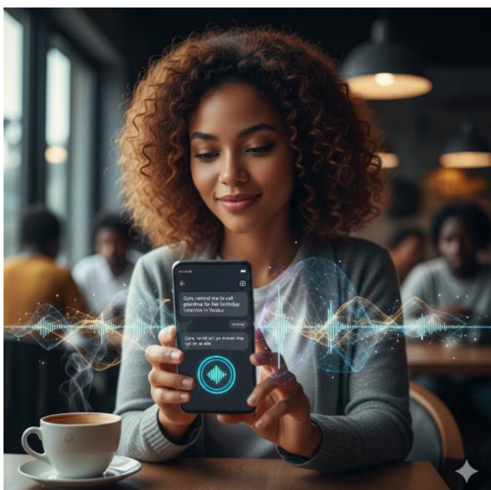
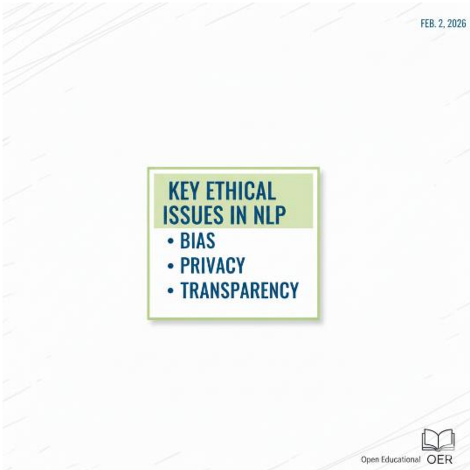


Plate 2.2 Everyday applications of NLP

### 2.4.2 Ethical considerations in NLP

As NLP systems become more widespread, ethical issues must be addressed. One concern is **bias**, where models trained on biased data may reproduce or amplify stereotypes. Another is **privacy**, since NLP often involves processing sensitive personal information. A further consideration is **transparency**, ensuring that users understand how NLP systems make decisions. Addressing these ethical issues is essential for building trust and ensuring responsible use of language technologies.

Fill in the blank: When NLP models reproduce stereotypes because of biased training data, this is referred to as __________.




Box 2.4 Ethical considerations in NLP

### 2.4.3 Future trends and research pathways

The future of NLP is shaped by emerging trends such as **multilingual NLP**, which aims to build systems that can handle multiple languages effectively, and **low-resource NLP**, which focuses on developing tools for languages with limited data. Another pathway is **explainable AI**, which seeks to make NLP models more transparent and interpretable. These directions highlight the ongoing evolution of NLP and the opportunities for learners to contribute to advancing the field.

Fill in the blank: Developing NLP tools for languages with limited data is known as __________ NLP.


Figure 2.4 Future directions in NLP

## Pilot questions 2.4

1. Give two examples of everyday applications of NLP.
2. What ethical issue arises when NLP models reproduce stereotypes?
3. State one privacy concern in NLP.
4. What is explainable AI in the context of NLP?
5. Mention one future trend in NLP research.

## *Series Summary*

This Series has examined the problems and perspectives of natural language processing, building on the foundations of computational linguistics. Its purpose has been to highlight the challenges computers face when dealing with human language and to present the approaches that have been developed to address them. By engaging with this Series, you have seen how NLP connects theory with practical applications that shape everyday technologies.

We began by introducing the scope of NLP problems, their categories, and common challenges. Then we explored core problems such as machine translation, speech recognition, and information retrieval. Following that, we considered perspectives in NLP development, including rule-based, statistical, and deep learning approaches. Finally, we discussed practical implications, ethical considerations, and future directions. By working through these Parts, you should now be able to define the scope of NLP problems, explain their challenges, analyse core issues, evaluate different approaches, and assess practical and ethical implications, as outlined in the Series Learning Outcomes.

## *Glossary of Terms*

- **Ambiguity**: A situation where a word, phrase, or sentence has more than one possible meaning.
- **Deep learning**: A subset of machine learning that uses neural networks with multiple layers to capture complex language patterns.
- **Homophones**: Words that sound alike but have different meanings, such as "flower" and "flour."
- **Information retrieval**: The process of finding relevant documents or data from large collections, often used in search engines.
- **Machine translation**: The automatic conversion of text from one language into another using computational methods.
- **Natural language processing (NLP)**: A branch of artificial intelligence concerned with enabling computers to understand, interpret, and generate human language.
- **Pragmatic tasks**: NLP tasks that deal with context and usage, such as understanding implied meaning or managing dialogue.
- **Prosody**: The rhythm, stress, and intonation patterns in speech that make it sound natural.
- **Rule-based approaches**: NLP methods that rely on manually crafted linguistic rules, grammars, and dictionaries to process language.

- **Semantic tasks**: NLP tasks that focus on meaning, such as word sense disambiguation and semantic role labelling.
- **Speech recognition**: The process of converting spoken language into written text.
- **Speech synthesis**: The process of generating spoken language from written text.
- **Statistical approaches**: NLP methods that use probability models and large datasets to predict language patterns.
- **Syntactic tasks**: NLP tasks that deal with the structure of language, such as parsing sentences and identifying parts of speech.
- **Text summarisation**: The process of condensing a large text into a shorter version while retaining its main ideas and meaning.
- **Transparency**: An ethical principle in NLP that ensures users understand how systems make decisions.

## *Concept Check Questions [Series 2]*

**Objective questions**

1. Natural language processing is a branch of __________ concerned with enabling computers to work with human language. (Linked to SLO 2.1)
2. Tasks that deal with the structure of language, such as parsing, are called __________ tasks. (Linked to SLO 2.1)
3. The automatic conversion of text from one language into another is known as __________. (Linked to SLO 2.3)
4. Words that sound alike but have different meanings are called __________. (Linked to SLO 2.3)
5. Developing NLP tools for languages with limited data is referred to as __________ NLP. (Linked to SLO 2.5)

**Short-answer questions**

6. Define ambiguity in the context of NLP. (Linked to SLO 2.2)
7. Explain one challenge faced in machine translation. (Linked to SLO 2.3)
8. State one limitation of rule-based approaches in NLP. (Linked to SLO 2.4)
9. Identify two ethical considerations in NLP. (Linked to SLO 2.5)
10. Describe one everyday application of NLP and explain its relevance. (Linked to SLO 2.5)

**Long-answer questions**

11. Analyse the differences between rule-based, statistical, and deep learning approaches in NLP. (Linked to SLO 2.4)
12. Evaluate the practical implications of NLP in everyday technology and discuss future research directions. (Linked to SLO 2.5)

## *Answers to Concept Check Questions [Series 2]*

**Objective questions**

1. Artificial intelligence.
2. Syntactic tasks.
3. Machine translation.
4. Homophones.
5. Low-resource NLP.

**Short-answer questions**

6. Ambiguity occurs when a word, phrase, or sentence has more than one possible meaning.
7. One challenge in machine translation is handling idiomatic expressions that cannot be translated literally.
8. Rule-based approaches are limited because they struggle with flexibility and scalability when dealing with diverse language use.
9. Bias and privacy are two key ethical considerations in NLP.
10. Digital assistants such as Siri or Alexa are everyday applications of NLP, relevant because they enable speech recognition and dialogue management for user interaction.

**Long-answer questions**

11. *Basic response*: Rule-based approaches rely on manually crafted linguistic rules, statistical approaches use probability models and large datasets, while deep learning employs neural networks to capture complex patterns. *Value-added response*: Beyond these distinctions, rule-based systems are precise but rigid, statistical approaches adapt better but depend heavily on annotated data, and deep learning has revolutionised NLP by enabling breakthroughs in machine translation and conversational agents. However, deep learning also raises concerns about transparency and bias.
12. *Basic response*: NLP has practical implications in technologies such as search engines, grammar checkers, and chatbots. Future directions include multilingual NLP and low-resource NLP. *Value-added response*: In addition to these applications, NLP impacts society by shaping communication, productivity, and access to information. Future research pathways such as explainable AI aim to make systems more transparent, while ethical considerations like fairness and privacy will guide responsible innovation in the field.

## *Answers to Pilot Questions [Series 2]*

**Part 2.1 Introduction to Natural Language Processing Problems**

1. Natural language processing is the branch of artificial intelligence concerned with enabling computers to work with human language.
2. Syntactic, semantic, and pragmatic tasks.
3. Parsing sentences.
4. Ambiguity.

5. Because human language is complex, context-dependent, and often lacks sufficient resources for computational modelling.

**Part 2.2 Core Problems in NLP**

1. Machine translation is the automatic conversion of text from one language into another. Its main challenges include structural differences between languages, idiomatic expressions, and context sensitivity.
2. Accent variation.
3. Speech synthesis is the process of generating spoken language from text. One challenge is producing natural-sounding speech with appropriate intonation and prosody.
4. Relevance ranking.
5. Preserving accuracy and capturing key ideas without oversimplification.

**Part 2.3 Perspectives in NLP Development**

1. Rule-based approaches are systems that rely on manually crafted linguistic rules, grammars, and dictionaries to process language.
2. They struggle with flexibility and scalability when dealing with diverse language use.
3. Statistical approaches rely on probability models and large datasets to make predictions about language.
4. Deep learning is a subset of machine learning that uses neural networks with multiple layers to capture complex language patterns.
5. Breakthroughs include machine translation, sentiment analysis, and conversational agents.

**Part 2.4 Practical Implications and Future Directions**

1. Digital assistants such as Siri or Alexa, and grammar checkers.
2. Bias.
3. Processing sensitive personal information without adequate safeguards.
4. Explainable AI refers to methods that make NLP models more transparent and interpretable.
5. Multilingual NLP, low-resource NLP, and explainable AI.

### *References and Attributions [Series 2]*

This section compiles all references and attributions for Series 2: *Natural Language Processing Problems and Perspectives*. Entries include suggested open educational resources (OERs) and notes on AI-generated illustration prompts. The referencing style follows APA (7th edition).

**General References**


- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.
- Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Draft manuscript. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

**Illustrations**

- *Figure 2.1 Scope of NLP problems*
  - Attribution: Suggested OER diagram (search string: "scope of NLP problems syntax semantics pragmatics OER diagram").
  - AI prompt: "Generate a diagram showing NLP problems across syntax, semantics, and pragmatics."
- *Table 2.1 Categories of NLP tasks*
  - Attribution: Suggested OER table (search string: "categories of NLP tasks syntactic semantic pragmatic OER table").
  - AI prompt: "Generate a table listing syntactic, semantic, and pragmatic NLP tasks with examples."
- *Box 2.1 Common challenges in NLP*
  - Attribution: Suggested OER resource (search string: "common challenges in NLP ambiguity context dependence OER").
  - AI prompt: "Generate a box listing three common challenges in NLP with explanations."
- *Figure 2.2 Challenges in machine translation*
  - Attribution: Suggested OER diagram (search string: "machine translation challenges OER diagram").
  - AI prompt: "Generate a diagram showing challenges in machine translation including structural differences, idioms, and context sensitivity."
- *Plate 2.1 Speech recognition in practice*
  - Attribution: Suggested OER image (search string: "speech recognition illustration OER image").
  - AI prompt: "Generate an illustration of a person using speech recognition software on a mobile device."
- *Box 2.2 Common problems in information retrieval and text summarisation*
  - Attribution: Suggested OER resource (search string: "information retrieval text summarisation challenges OER").
  - AI prompt: "Generate a box listing common problems in information retrieval and text summarisation with explanations."
- *Figure 2.3 Rule-based approach in NLP*

- Attribution: Suggested OER diagram (search string: “rule-based NLP approach diagram OER”).
- AI prompt: “Generate a diagram showing rule-based NLP with grammar rules feeding into text analysis.”

- *Table 2.2 Features of statistical and probabilistic approaches*
  - Attribution: Suggested OER table (search string: “statistical vs rule-based NLP comparison OER table”).
  - AI prompt: “Generate a table comparing rule-based and statistical NLP approaches with features and limitations.”
- *Box 2.3 Perspectives in NLP development*
  - Attribution: Suggested OER resource (search string: “NLP development approaches OER summary”).
  - AI prompt: “Generate a box listing three perspectives in NLP development: rule-based, statistical, and deep learning.”
- *Plate 2.2 Everyday applications of NLP*
  - Attribution: Suggested OER image (search string: “digital assistant NLP application OER image”).
  - AI prompt: “Generate an illustration of a person using a smartphone digital assistant powered by NLP.”
- *Box 2.4 Ethical considerations in NLP*
  - Attribution: Suggested OER resource (search string: “ethical issues in NLP bias privacy transparency OER”).
  - AI prompt: “Generate a box listing three ethical considerations in NLP: bias, privacy, and transparency.”
- *Figure 2.4 Future directions in NLP*
  - Attribution: Suggested OER diagram (search string: “future trends in NLP multilingual low-resource explainable AI OER diagram”).
  -

**Course code: LIN 309**
**Course title: Computational Linguistics I**
**Course credit load: 2 Units**

**Series 1:** Foundations of Computational Linguistics

**Series 2:** Natural Language Processing Problems and Perspectives

**Series 3:** Corpus Linguistics Essentials

**Series 4:** Probability Calculus in Language Analysis

**Series 5:** N-Grams and Language Modeling

**Proposed Outline for Series 3**

**Part .1 Introduction to Corpus Linguistics**

- 3.1.1 Definition and scope of corpus linguistics
- 3.1.2 Historical development and relevance
- 3.1.3 Key principles and terminology

**Part .2 Types of Corpora and Their Applications**

- 3.2.1 General corpora vs specialised corpora
- 3.2.2 Written vs spoken corpora
- 3.2.3 Applications in language teaching, research, and NLP

**Part .3 Tools and Methods in Corpus Linguistics**

- 3.3.1 Corpus compilation and annotation
- 3.3.2 Concordancing and frequency analysis
- 3.3.3 Collocation and keyword analysis

**Part .4 Practical Perspectives and Future Directions**

- 3.4.1 Case studies of corpus-based research
- 3.4.2 Ethical considerations in corpus use
- 3.4.3 Emerging trends and digital resources

**Introduction**

In the last Series, we explored natural language processing problems and perspectives, focusing on the challenges computers face when dealing with human language and the approaches used to address them. We now turn to corpus linguistics, which provides the empirical foundation for much of this work. By studying large collections of texts, corpus linguistics allows us to observe language in use, test linguistic theories, and develop practical applications. This Series is important because it equips you with the essential concepts and tools needed to work with corpora, a skill that underpins both linguistic research and computational applications.

The aim of this Series is to introduce you to the essentials of corpus linguistics and to enable you to apply its principles, methods, and tools in analysing language data. By the end of the Series, you should be able to recognise different types of corpora, use basic corpus tools, and appreciate the role of corpora in both research and practice.

We shall commence this Series by defining corpus linguistics, tracing its historical development, and clarifying its scope and key terminology. Next, we will examine the different types of corpora, distinguishing between general and specialised collections, as well as written and spoken corpora, while considering their applications. Following that, we will move on to tools and methods, including corpus compilation, annotation, concordancing, and frequency analysis. Finally, we will wrap up by exploring practical perspectives, ethical considerations, and future directions, rounding off the Series with case studies and emerging trends in corpus-based research.

**Series Learning Outcomes**

Upon completing this Series, you should be able to:

- **SLO 3.1** define corpus linguistics and explain its scope, historical development, and key terminology,
- **SLO 3.2** distinguish between different types of corpora, including general, specialised, written, and spoken collections,
- **SLO 3.3** analyse practical applications of corpora in language teaching, linguistic research, and natural language processing,
- **SLO 3.4** demonstrate the use of basic corpus tools such as concordancing, frequency analysis, and collocation analysis, and
- **SLO 3.5** evaluate ethical considerations and emerging trends in corpus linguistics, supported by case studies of corpus-based research.

## 3.1 Introduction to Corpus Linguistics

### 3.1.1 Definition and scope of corpus linguistics

**Corpus linguistics** is the study of language through large collections of texts, known as corpora. A **corpus** is a structured set of texts, either written or spoken, that is stored in electronic form for systematic analysis. Corpus linguistics provides empirical evidence about how language is actually used, rather than relying solely on intuition or prescriptive rules. Its scope includes exploring grammar, vocabulary, discourse, and variation across different contexts.

Fill in the blank: A structured set of texts stored in electronic form for systematic analysis is called a __________.

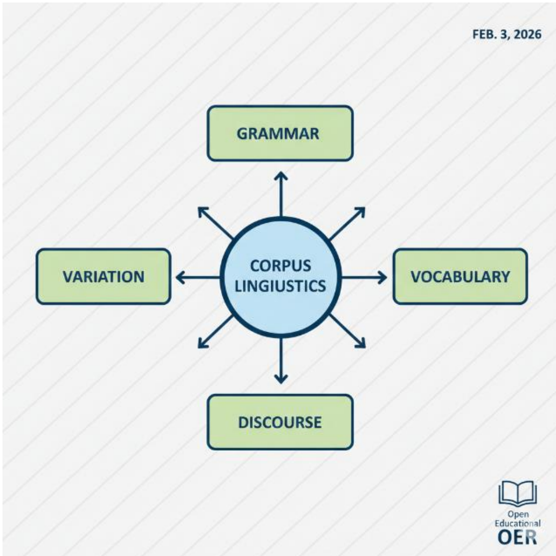



Figure 3.1 Scope of corpus linguistics

### 3.1.2 Historical development and relevance

Corpus linguistics has its roots in early linguistic studies, where scholars manually collected examples of language use. With the advent of computers, the field expanded rapidly, allowing researchers to store and analyse millions of words efficiently. Early projects such as the Brown Corpus in the 1960s marked a turning point, providing systematic data for linguistic research. Today, corpus linguistics is relevant not only to linguists but also to educators, translators, and developers of natural language processing systems.

Fill in the blank: The Brown Corpus, developed in the 1960s, was one of the first systematic collections of __________.


Plate 3.1 Early computer-based corpus analysis

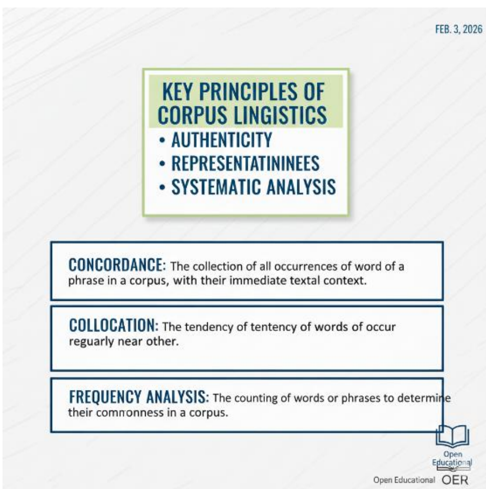
### 3.1.3 Key principles and terminology

Corpus linguistics is guided by several principles. One is **authenticity**, meaning corpora should contain real language data rather than artificially constructed examples. Another is **representativeness**, ensuring that the corpus reflects the variety of language use in a given context. A third principle is **systematic analysis**, which involves using tools to examine patterns such as frequency, collocation, and concordance.

Key terms include:

- **Concordance**: A list of occurrences of a word or phrase in context.
- **Collocation**: Words that frequently occur together.
- **Frequency analysis**: Counting how often words or structures appear in a corpus.

Fill in the blank: When a corpus reflects the variety of language use in a given context, it is said to have __________.




Box 3.1 Key principles and terminology in corpus linguistics

Pilot questions 3.1

1. Define corpus linguistics in your own words.
2. What is a corpus?
3. Name one early project that marked a turning point in corpus linguistics.
4. State two principles that guide corpus linguistics.
5. What is meant by collocation in corpus analysis?

## 3.2 Types of Corpora and Their Applications

### 3.2.1 General corpora vs specialised corpora

A **general corpus** is a large collection of texts designed to represent a broad sample of a language, often across different genres and contexts. Examples include newspapers, novels, academic writing, and spoken transcripts. In contrast, a **specialised corpus** focuses on a particular domain, such as medical texts, legal documents, or classroom interactions. General corpora are useful for studying overall language patterns, while specialised corpora provide insights into specific fields.

Fill in the blank: A corpus that focuses on a particular domain, such as medicine or law, is called a __________ corpus.

| Feature | General (Reference) Corpora | Specialised Corpora |
|---|---|---|
| **Purpose** | To represent a language as a whole (the "norm"). | To study specific genres, domains, or time periods. |
| **Size** | Large (often hundreds of millions to billions of words). | Smaller (focused on reaching "saturation" of technical terms). |
| **Composition** | Balanced mix of spoken, written, formal, and informal texts. | Narrowly defined (e.g., only medical papers, only legal contracts). |
| **Representativeness** | Aiming for a "slice" of total language use. | Aiming for "exhaustiveness" within a specific niche. |
| **Primary Examples** | British National Corpus (BNC), COCA. | Michigan Corpus of Academic Spoken English (MICASE). |
| **Applications** | Dictionary making, general grammar research, NLP training. | ESP (English for Specific Purposes), Terminology extraction, Translation. |

Table 3.2 Comparison of general and specialised corpora

### 3.2.2 Written vs spoken corpora

**Written corpora** consist of texts such as books, articles, and reports, while **spoken corpora** are collections of transcribed speech from conversations, interviews, or broadcasts. Written corpora are easier to compile because texts are readily available, but spoken corpora are valuable for studying everyday language use, including features like hesitation, intonation, and informal expressions. Together, they provide a fuller picture of how language operates in different modes.

Fill in the blank: Collections of transcribed speech from conversations or interviews are known as __________ corpora.


Plate 3.2 Collecting spoken corpus data

### 3.2.3 Applications in language teaching, research, and NLP

Corpora have wide-ranging applications. In **language teaching**, they help teachers design materials based on authentic usage rather than intuition. In **linguistic research**, corpora provide evidence for theories of grammar, vocabulary, and discourse. In **natural language processing (NLP)**, corpora serve as training data for systems such as machine translation, speech recognition, and text summarisation. These applications demonstrate the versatility of corpus linguistics in both academic and practical contexts.

Fill in the blank: In language teaching, corpora are valuable because they provide examples of ___________ usage.

| Domain | Primary Goal | Example Corpus |
|---|---|---|
| **Education** | Authentic Input | BNC (British National Corpus) |
| **Research** | Pattern Discovery | COCA (Corpus of Contemporary American English) |
| **NLP** | Model Performance | Europarl (Multilingual Parallel Corpus) |

Box 3.2 Applications of corpora

Pilot questions 3.2

1. What is the difference between a general corpus and a specialised corpus?
2. Give one example of a specialised corpus.
3. What is a spoken corpus?
4. State one advantage of using written corpora.
5. Mention one application of corpora in natural language processing.

## 3.3 Tools and Methods in Corpus Linguistics

### 3.3.1 Corpus compilation and annotation

**Corpus compilation** refers to the process of collecting texts to form a corpus. This involves selecting sources, ensuring representativeness, and storing the texts in electronic form. **Annotation** is the process of adding linguistic information to the corpus, such as part-of-speech tags, syntactic structures, or semantic roles. Together, compilation and annotation make corpora more useful for systematic analysis.

Fill in the blank: Adding linguistic information such as part-of-speech tags to a corpus is called __________.

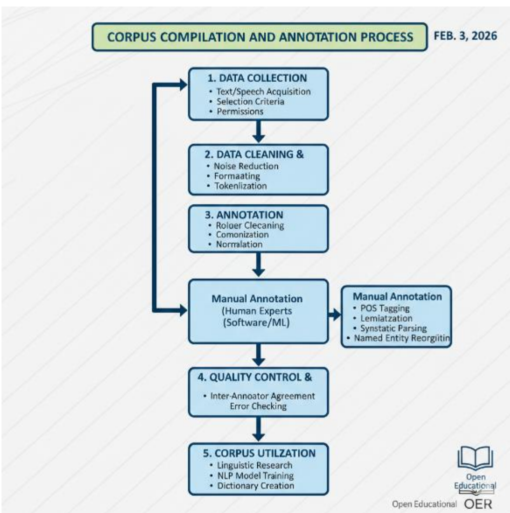



Figure 3.3 Corpus compilation and annotation process

### 3.3.2 Concordancing and frequency analysis

A **concordance** is a list of occurrences of a word or phrase in its context, allowing researchers to study how it is used. **Frequency analysis** involves counting how often words or structures appear in a corpus. These methods help identify patterns in language use, such as common collocations or stylistic tendencies.

Fill in the blank: A list of occurrences of a word or phrase in its context is called a __________.

| Word (Node) | Frequency (per million) | Sample Concordance Lines (KWIC) | Linguistic Insight |
|---|---|---|---|
| **Bank** | 452 | ...deposited the check at the **bank** this morning... | Primarily functions as a financial noun in this corpus. |
| | | ...sat quietly on the grassy **bank** of the river... | Secondary "geographic" sense identified through context. |
| **Address** | 215 | ...please write your mailing **address** on the form... | Used as a noun referring to a physical location. |
| | | ...the president is set to **address** the nation... | Used as a verb meaning "to speak to." |
| **Light** | 890 | ...the morning **light** filtered through the... | High frequency due to use as noun, adjective, and verb. |
| | | ...he carried a **light** suitcase for the trip... | Shows collocations with items of physical weight. |

Table 3.3 Examples of concordance and frequency analysis

### 3.3.3 Collocation and keyword analysis

**Collocation** refers to words that frequently occur together, such as “make a decision” or “strong tea.” Identifying collocations helps reveal natural patterns of word usage. **Keyword analysis** highlights words that occur significantly more often in one corpus compared to another, making it useful for comparing texts across genres or domains. These methods provide deeper insights into meaning and discourse.

Fill in the blank: Words that frequently occur together, such as “make a decision,” are known as __________.

| Feature | Collocation | Keyword Analysis |
|---|---|---|
| **Focus** | Relationship between *two or more words*. | Relationship between *a word and a specific text*. |
| **Main Question** | "What words usually hang out with this word?" | "What words make this specific text unique?" |

| Feature | Collocation | Keyword Analysis |
|---|---|---|
| **Linguistic Goal** | Understanding phraseology and naturalness. | Discovering the central themes/topics of a text. |

Box 3.3 Collocation and keyword analysis

### Pilot questions 3.3

1. What is corpus compilation?
2. Define annotation in corpus linguistics.
3. What is a concordance?
4. State one purpose of frequency analysis.
5. Give one example of collocation.
6. What is keyword analysis used for?

## 3.4 Practical Perspectives and Future Directions

### 3.4.1 Case studies of corpus-based research

Corpus linguistics is not just theoretical; it has been applied in many practical studies. For example, researchers have used corpora to investigate how academic writing differs from everyday language, or how medical professionals communicate with patients. These case studies show how corpus methods provide evidence-based insights into language use across different domains.

Fill in the blank: Corpus-based studies provide __________ evidence about how language is used in real contexts.

| Context | Primary Dataset | Key Metric Analyzed |
|---|---|---|
| **Academic** | Student vs. Expert Papers | Use of Hedging & Formal Connectors |
| **Medical** | Consultations & Transcripts | Jargon usage vs. Back-channeling |
| **Classroom** | Recorded Lesson Audio | Lexical Bundles & Signposting |

Box 3.4 Examples of corpus-based research

### 3.4.2 Ethical considerations in corpus use

Working with corpora raises important ethical issues. One concern is **privacy**, especially when corpora include personal communications such as emails or social media posts. Another is **consent**, ensuring that participants whose language is recorded have agreed to its use. A further issue is **bias**, since corpora may over-represent certain groups or genres, leading to skewed results. Ethical awareness is essential for responsible corpus research.

Fill in the blank: When corpora include personal communications, researchers must pay attention to issues of __________.

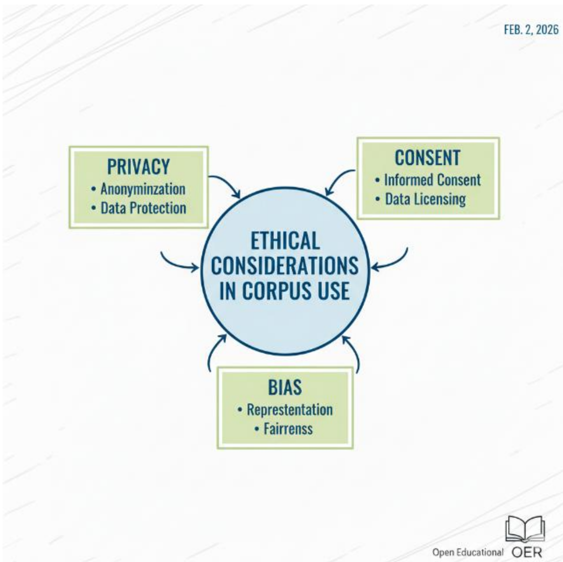



Figure 3.4 Ethical considerations in corpus use

### 3.4.3 Emerging trends and digital resources

Corpus linguistics continues to evolve with new technologies. **Multimodal corpora** now include not only text but also audio and video, allowing analysis of speech, gesture, and visual communication. **Web-based corpora** provide access to vast amounts of online text, while **learner corpora** focus on language produced by students, helping improve teaching methods. These emerging trends highlight the growing importance of digital resources in corpus linguistics.

Fill in the blank: Corpora that include text, audio, and video for studying communication are called __________ corpora.


Plate 3.4 Multimodal corpus analysis

### Pilot questions 3.4

1. Give one example of a corpus-based research study.
2. What ethical issue arises when corpora include personal communications?
3. State one reason why consent is important in corpus research.
4. What are multimodal corpora?
5. Mention one emerging trend in corpus linguistics.

### *Series Summary*

This Series has introduced you to the essentials of corpus linguistics, highlighting its role as a foundation for both linguistic research and practical applications such as language teaching and natural language processing. The purpose has been to show how corpora provide empirical evidence of language use, enabling more accurate analysis and informed decision-making in academic and professional contexts.

We began by defining corpus linguistics, tracing its historical development, and clarifying its scope and key principles. Then we examined different types of corpora, distinguishing between general and specialised collections, as well as written and spoken corpora, and considered their applications. Following that, we explored tools and methods such as compilation, annotation, concordancing, frequency analysis, collocation, and keyword analysis. Finally, we discussed practical perspectives, ethical considerations, and emerging trends including multimodal and web-based corpora. By completing this Series, you should now be able to define corpus linguistics and its terminology, distinguish between types of corpora, analyse their applications, demonstrate the use of basic corpus tools, and evaluate ethical issues and future directions, as outlined in the Series Learning Outcomes.

## *Glossary of Terms*

- **Annotation**: The process of adding linguistic information to a corpus, such as part-of-speech tags or syntactic structures.
- **Collocation**: Words that frequently occur together in natural language, for example "make a decision" or "strong tea."
- **Concordance**: A list of occurrences of a word or phrase in its context, showing how it is used across different texts.
- **Corpus**: A structured collection of texts, either written or spoken, stored in electronic form for systematic analysis.
- **Corpus compilation**: The process of collecting and organising texts to form a corpus, ensuring representativeness and authenticity.
- **Corpus linguistics**: The study of language through large collections of texts (corpora), providing empirical evidence of how language is used.
- **Frequency analysis**: Counting how often words or structures appear in a corpus to identify patterns of usage.
- **General corpus**: A large collection of texts designed to represent a broad sample of a language across different genres and contexts.
- **Keyword analysis**: Identifying words that occur significantly more often in one corpus compared to another, useful for comparing texts.
- **Multimodal corpus**: A corpus that includes text, audio, and video, allowing analysis of speech, gesture, and visual communication.
- **Representativeness**: A principle in corpus linguistics that ensures a corpus reflects the variety of language use in a given context.
- **Specialised corpus**: A corpus focused on a particular domain, such as medicine, law, or classroom interaction.
- **Spoken corpus**: A collection of transcribed speech from conversations, interviews, or broadcasts, used to study everyday language use.
- **Systematic analysis**: The principle of examining language patterns in a structured way using corpus tools such as concordancing and frequency counts.
- **Written corpus**: A collection of texts such as books, articles, and reports, used to study language in written form.

## *Concept Check Questions [Series 2]*

**Objective questions**

1. Natural language processing is a branch of __________ concerned with enabling computers to work with human language. (Linked to SLO 2.1)
2. Tasks that deal with the structure of language, such as parsing, are called __________ tasks. (Linked to SLO 2.1)

3. The automatic conversion of text from one language into another is known as __________. (Linked to SLO 2.3)
4. Words that sound alike but have different meanings are called __________. (Linked to SLO 2.3)
5. Developing NLP tools for languages with limited data is referred to as __________ NLP. (Linked to SLO 2.5)

**Short-answer questions**

6. Define ambiguity in the context of NLP. (Linked to SLO 2.2)
7. Explain one challenge faced in machine translation. (Linked to SLO 2.3)
8. State one limitation of rule-based approaches in NLP. (Linked to SLO 2.4)
9. Identify two ethical considerations in NLP. (Linked to SLO 2.5)
10. Describe one everyday application of NLP and explain its relevance. (Linked to SLO 2.5)

**Long-answer questions**

11. Analyse the differences between rule-based, statistical, and deep learning approaches in NLP. (Linked to SLO 2.4)
12. Evaluate the practical implications of NLP in everyday technology and discuss future research directions. (Linked to SLO 2.5)

### *Answers to Concept Check Questions [Series 2]*

**Objective questions**

1. Artificial intelligence.
2. Syntactic tasks.
3. Machine translation.
4. Homophones.
5. Low-resource NLP.

**Short-answer questions**

6. Ambiguity occurs when a word, phrase, or sentence has more than one possible meaning.
7. One challenge in machine translation is handling idiomatic expressions that cannot be translated literally.
8. Rule-based approaches are limited because they struggle with flexibility and scalability when dealing with diverse language use.
9. Bias and privacy are two key ethical considerations in NLP.
10. Digital assistants such as Siri or Alexa are everyday applications of NLP, relevant because they enable speech recognition and dialogue management for user interaction.

**Long-answer questions**

11. *Basic response*: Rule-based approaches rely on manually crafted linguistic rules, statistical approaches use probability models and large datasets, while deep learning employs neural networks to capture complex patterns. *Value-added response*: Beyond these distinctions, rule-based systems are precise but rigid, statistical approaches adapt better but depend heavily on annotated data, and deep learning has revolutionised NLP by enabling breakthroughs in machine translation and conversational agents. However, deep learning also raises concerns about transparency and bias.

12. *Basic response*: NLP has practical implications in technologies such as search engines, grammar checkers, and chatbots. Future directions include multilingual NLP and low-resource NLP. *Value-added response*: In addition to these applications, NLP impacts society by shaping communication, productivity, and access to information. Future research pathways such as explainable AI aim to make systems more transparent, while ethical considerations like fairness and privacy will guide responsible innovation in the field.

## *Answers to Pilot Questions [Series 3]*

### Part 3.1 Introduction to Corpus Linguistics

1. Corpus linguistics is the study of language through large collections of texts, known as corpora.
2. A corpus is a structured set of texts stored in electronic form for systematic analysis.
3. The Brown Corpus.
4. Authenticity and representativeness.
5. Collocation refers to words that frequently occur together.

### Part 3.2 Types of Corpora and Their Applications

1. A general corpus represents a broad sample of a language, while a specialised corpus focuses on a specific domain.
2. An example is a medical corpus.
3. A spoken corpus is a collection of transcribed speech from conversations, interviews, or broadcasts.
4. Written corpora are easier to compile because texts are readily available.
5. Corpora are used in NLP as training data for systems such as machine translation.

### Part 3.3 Tools and Methods in Corpus Linguistics

1. Corpus compilation is the process of collecting and organising texts to form a corpus.
2. Annotation is adding linguistic information such as part-of-speech tags to a corpus.
3. A concordance is a list of occurrences of a word or phrase in its context.
4. Frequency analysis helps identify patterns in language use by counting word occurrences.
5. An example of collocation is "make a decision."
6. Keyword analysis is used to compare texts across genres or domains by highlighting distinctive words.

### Part 3.4 Practical Perspectives and Future Directions

1. One example is a study of academic writing compared with everyday language.
2. Privacy.
3. Consent ensures that participants agree to their language being recorded and analysed.
4. Multimodal corpora.
5. Emerging trends include learner corpora and web-based corpora.

### *References and Attributions [Series 3]*

This section compiles all references and attributions for Series 3: *Corpus Linguistics Essentials*. Entries include suggested open educational resources (OERs) and notes on AI-generated illustration prompts. The referencing style follows APA (7th edition).

**General References**


- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.


**Illustrations**

- *Figure 3.1 Scope of corpus linguistics*
    - Attribution: Suggested OER diagram (search string: "scope of corpus linguistics diagram OER").
    - AI prompt: "Generate a diagram showing corpus linguistics connected to grammar, vocabulary, discourse, and variation."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Plate 3.1 Early computer-based corpus analysis*
    - Attribution: Suggested OER image (search string: "Brown Corpus history corpus linguistics OER image").
    - AI prompt: "Generate an illustration of early computer systems used for corpus analysis in the 1960s."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Box 3.1 Key principles and terminology in corpus linguistics*
    - Attribution: Suggested OER resource (search string: "corpus linguistics principles concordance collocation frequency OER summary").
    - AI prompt: "Generate a box listing key principles of corpus linguistics and definitions of concordance, collocation, and frequency analysis."
    - Tool/platform: Microsoft Copilot Image Creator.

- *Table 3.2 Comparison of general and specialised corpora*
    - Attribution: Suggested OER table (search string: "general vs specialised corpora comparison OER table").
    - AI prompt: "Generate a table comparing general and specialised corpora with examples and applications."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Plate 3.2 Collecting spoken corpus data*
    - Attribution: Suggested OER image (search string: "spoken corpus linguistics recording OER image").
    - AI prompt: "Generate an illustration of a researcher recording spoken language for a spoken corpus."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Box 3.2 Applications of corpora*
    - Attribution: Suggested OER resource (search string: "applications of corpora language teaching research NLP OER summary").
    - AI prompt: "Generate a box listing applications of corpora in language teaching, linguistic research, and NLP."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Figure 3.3 Corpus compilation and annotation process*
    - Attribution: Suggested OER diagram (search string: "corpus compilation annotation process OER diagram").
    - AI prompt: "Generate a diagram showing the steps of corpus compilation and annotation."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Table 3.3 Examples of concordance and frequency analysis*
    - Attribution: Suggested OER table (search string: "concordance frequency analysis corpus linguistics OER table").
    - AI prompt: "Generate a table showing examples of concordance lines and frequency counts for selected words."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Box 3.3 Collocation and keyword analysis*
    - Attribution: Suggested OER resource (search string: "collocation keyword analysis corpus linguistics OER summary").

    - AI prompt: “Generate a box listing definitions and examples of collocation and keyword analysis.”
    - Tool/platform: Microsoft Copilot Image Creator.
- *Box 3.4 Examples of corpus-based research*
    - Attribution: Suggested OER resource (search string: “corpus-based research case studies linguistics OER summary”).
    - AI prompt: “Generate a box listing examples of corpus-based research in academic, medical, and classroom contexts.”
    - Tool/platform: Microsoft Copilot Image Creator.
- *Figure 3.4 Ethical considerations in corpus use*
    - Attribution: Suggested OER diagram (search string: “ethical issues in corpus linguistics privacy consent bias OER diagram”).
    - AI prompt: “Generate a diagram showing ethical considerations in corpus use: privacy, consent, and bias.”
    - Tool/platform: Microsoft Copilot Image Creator.
- *Plate 3.4 Multimodal corpus analysis*
    - Attribution: Suggested OER image (search string: “multimodal corpus linguistics analysis OER image”).
    - AI prompt: “Generate an illustration of a researcher analysing multimodal corpus data with text, audio, and video.”
    - Tool/platform: Microsoft Copilot Image Creator.

**Course code: LIN 309**
**Course title: Computational Linguistics I**
**Course credit load: 2 Units**

**Series 1:** Foundations of Computational Linguistics

**Series 2:** Natural Language Processing Problems and Perspectives

**Series 3:** Corpus Linguistics Essentials

**Series 4:** Probability Calculus in Language Analysis

**Series 5:** N-Grams and Language Modeling

**Proposed Outline for Series 4**

**Part .1 Fundamentals of Probability in Linguistics**

- 4.1.1 Definition of probability and its relevance to language
- 4.1.2 Key concepts: random variables, events, and distributions
- 4.1.3 Examples of probability in everyday language use

**Part .2 Probability Models in Language Analysis**

- 4.2.1 N-gram models and their applications
- 4.2.2 Conditional probability and language prediction
- 4.2.3 Bayesian approaches in linguistic analysis

**Part .3 Applications of Probability Calculus in NLP**

- 4.3.1 Speech recognition and probabilistic modelling
- 4.3.2 Machine translation and probability-based methods
- 4.3.3 Text classification and sentiment analysis

**Part .4 Challenges and Future Directions**

- 4.4.1 Limitations of probabilistic models in language
- 4.4.2 Combining probability with deep learning
- 4.4.3 Emerging trends in probabilistic language analysis

## Introduction

In the last Series, we explored corpus linguistics and saw how large collections of texts can provide empirical evidence about language use. Building on that foundation, we now turn to probability calculus, which offers a mathematical framework for analysing and predicting language patterns. Probability is central to many modern applications of language technology, from speech recognition to machine translation, and understanding its role will help you appreciate how uncertainty in language can be modelled and managed.

The aim of this Series is to introduce you to the principles of probability as applied to language analysis and to equip you with the skills to interpret, apply, and evaluate probabilistic models in linguistic and computational contexts.

We shall commence this Series by examining the fundamentals of probability in linguistics, including definitions, key concepts, and everyday examples. Next, we will explore probability models such as n-gram models, conditional probability, and Bayesian approaches. Following that, we will move on to practical applications in natural language processing, including speech recognition, machine translation, and text classification. Finally, we will wrap up by considering the challenges and future directions of probabilistic language analysis, including its limitations, integration with deep learning, and emerging trends.

**Series Learning Outcomes**

Upon completing this Series, you should be able to:

- **SLO 4.1** define probability and explain its relevance to language analysis,
- **SLO 4.2** describe key concepts such as random variables, events, and distributions in linguistic contexts,
- **SLO 4.3** apply n-gram models to predict language patterns,
- **SLO 4.4** demonstrate the use of conditional probability in language prediction tasks,
- **SLO 4.5** analyse Bayesian approaches in linguistic analysis,
- **SLO 4.6** evaluate probabilistic methods in speech recognition, machine translation, and text classification, and
- **SLO 4.7** assess the limitations of probabilistic models and discuss emerging trends in probabilistic language analysis.

## 4.1 Fundamentals of Probability in Linguistics

### 4.1.1 Definition of probability and its relevance to language

**Probability** is the measure of how likely an event is to occur, expressed as a number between 0 and 1. In linguistics, probability helps us model uncertainty in language use. For example, when predicting the next word in a sentence, probability allows us to estimate which word is most likely to appear based on previous patterns. This makes probability central to applications such as speech recognition, machine translation, and text prediction.

Fill in the blank: Probability is expressed as a number between __________ and __________.

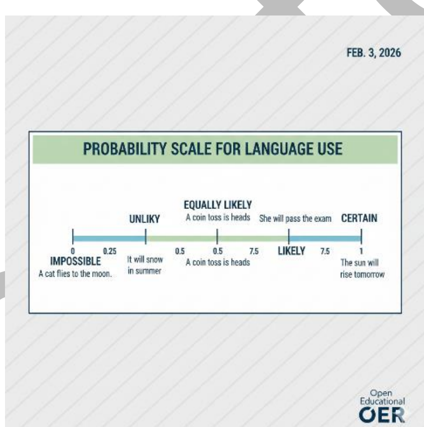



Figure 4.1 Probability scale in language analysis

### 4.1.2 Key concepts: random variables, events, and distributions

A **random variable** is a quantity whose value depends on chance, such as the next word in a sentence. An **event** is a specific outcome, like the occurrence of the word “the” in a text. A **distribution** describes how probabilities are spread across possible outcomes. In language analysis, distributions help us understand which words or structures are more frequent and which are rare.

Fill in the blank: A specific outcome, such as the occurrence of the word “the,” is called an __________.

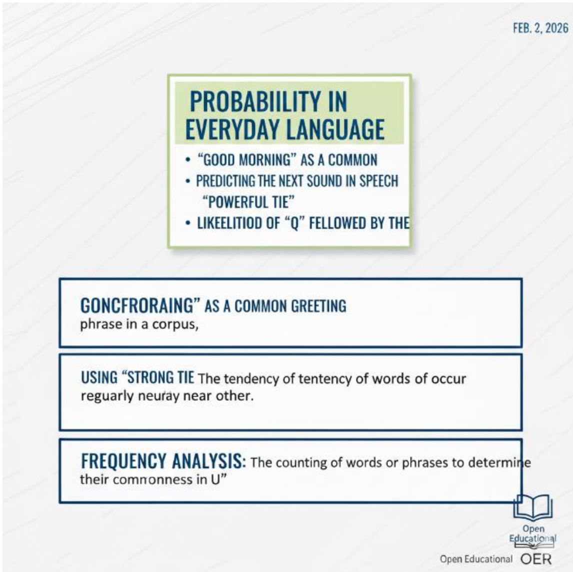
| Concept | Mathematical Definition | Linguistic Translation | Example (English) |
|---|---|---|---|
| **Random Variable** | A function that assigns a numerical value to each outcome in a sample space. | A specific linguistic feature we are observing or measuring. | **$X$** = The number of syllables in a randomly chosen word from a book. |
| **Event** | A subset of outcomes from the sample space (what actually happens). | A specific occurrence of a linguistic variable during speech or writing. | **$E$** = A speaker uses the word "Soda" instead of "Pop" or "Coke." |
| **Distribution** | A function that provides the likelihood of occurrence of different possible outcomes. | The pattern or frequency of how linguistic features are spread across a dataset. | **Zipf's Law:** The distribution showing that the most frequent word (the) occurs twice as often as the second most frequent. |

Table 4.1 Key probability concepts in linguistics

**4.1.3 Examples of probability in everyday language use**

Probability is evident in everyday communication. For instance, some words are more likely to follow others, such as “good” often followed by “morning.” Similarly, certain sounds are more probable in specific contexts, which helps listeners anticipate what comes next in speech. These everyday examples show how probability underpins both spoken and written language.

Fill in the blank: The phrase “good __________” illustrates how probability influences word prediction in everyday language.




Box 4.1 Everyday examples of probability in language

### Pilot questions 4.1

1. Define probability in your own words.
2. What is the range of values for probability?
3. What is a random variable in linguistics?
4. Give one example of an event in language analysis.
5. What does a distribution show in probability?
6. Provide one everyday example of probability in language use.

## 4.2 Probability Models in Language Analysis

### 4.2.1 N-gram models and their applications

An **n-gram model** is a probabilistic model that predicts the likelihood of a word based on the previous *n-1* words. For example, a bigram model (n=2) considers the probability of a word given the one before it, while a trigram model (n=3) considers the two preceding words. These models are widely used in text prediction, spelling correction, and speech recognition because they capture local patterns in language.

Fill in the blank: A model that predicts the likelihood of a word based on the previous word is called a __________ model.

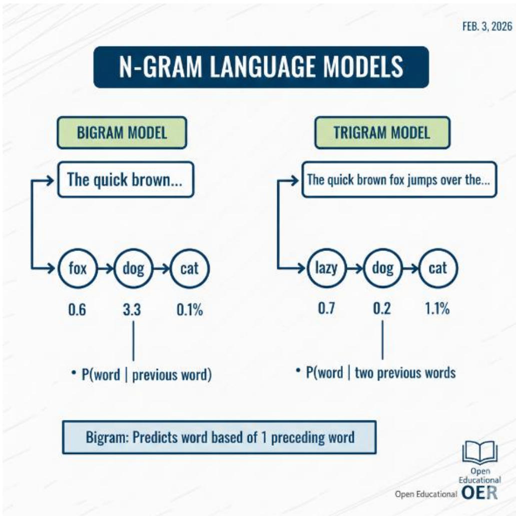



Figure 4.2 N-gram models in language prediction

### 4.2.2 Conditional probability and language prediction

**Conditional probability** is the probability of an event occurring given that another event has already occurred. In language analysis, this means estimating the likelihood of a word appearing given the words before it. For instance, the probability of "morning" given "good" is much higher than the probability of "banana" given "good." Conditional probability is the mathematical foundation of predictive text systems.

Fill in the blank: The probability of "morning" given "good" is an example of __________ probability.

| Word A (wn−1) | Word B (wn) | Notation | Est. Probability | Why? |
|---|---|---|---|---|
| **Good** | **Morning** | $P(\text{morning} \mid \text{good})$ | **0.120** | A very common social formula/greeting. |
| **Good** | **Luck** | $P(\text{luck} \mid \text{good})$ | **0.045** | Common, but "Good" has many other partners. |
| **Natural** | **Language** | $P(\text{language} \mid \text{natural})$ | **0.082** | A strong technical collocation. |
| **The** | **Table** | $P(\text{table} \mid \text{the})$ | **0.002** | Low, because "The" can be followed by almost any noun. |

| Word A (wn−1) | Word B (wn) | Notation | Est. Probability | Why? |
|---|---|---|---|---|
| **Of** | **The** | $P(\text{the} \mid \text{of})$ | **0.190** | Very high; "of the" is one of the most frequent bigrams. |
| **San** | **Francisco** | $P(\text{francisco} \mid \text{san})$ | **0.650** | Extremely high; "San" is almost always part of a specific name. |

Table 4.2 Examples of conditional probability in language

### 4.2.3 Bayesian approaches in linguistic analysis

A **Bayesian approach** applies Bayes' theorem to update probabilities based on new evidence. In linguistics, Bayesian models are used for tasks such as spam detection, where the probability of a message being spam is updated as more features (like suspicious words) are observed. Bayesian methods are powerful because they combine prior knowledge with observed data, making them flexible and adaptive.

Fill in the blank: Bayesian approaches update probabilities based on new __________.

| Feature | Rule-Based (Old) | Bayesian (Modern/2026) |
|---|---|---|
| **Logic** | "If word = 'Cash', then Spam." | "How likely is 'Cash' to be Spam in *this* context?" |
| **Flexibility** | Static; easy for spammers to trick. | Adaptive; learns from the user's specific habits. |
| **Accuracy** | High false-positives (good mail blocked). | Very low error rate due to weighted evidence. |

Box 4.2 Bayesian approaches in language analysis

## Pilot questions 4.2

1. What is an n-gram model?
2. Give one example of how n-gram models are applied.
3. Define conditional probability in your own words.
4. Provide one example of conditional probability in language.
5. What does a Bayesian approach do with probabilities?
6. Mention one application of Bayesian methods in language analysis.

## 4.3 Applications of Probability Calculus in NLP

### 4.3.1 Speech recognition and probabilistic modelling

Speech recognition systems rely heavily on probability to interpret spoken language. When you say a word, the system considers multiple possible interpretations of the sound and uses **probabilistic modelling** to select the most likely match. For example, if the input sound could be “there” or “their,” the system uses contextual probabilities to decide which word fits best. This approach helps manage uncertainty in speech signals and improves accuracy.

Fill in the blank: Speech recognition systems use __________ modelling to select the most likely interpretation of sounds.

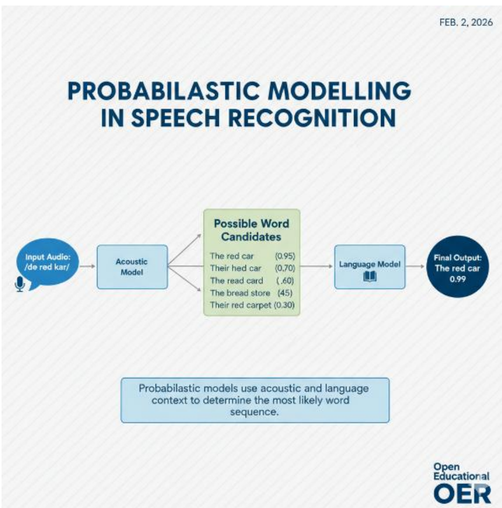



Figure 4.3 Probabilistic modelling in speech recognition

### 4.3.2 Machine translation and probability-based methods

Machine translation systems use probability to predict the most likely translation of a sentence. By analysing large bilingual corpora, they calculate the probability of word or phrase correspondences. For instance, the English word “house” is more likely to be translated as “casa” in Spanish than “hogar” in certain contexts. Probability-based methods allow translations to be more natural and context-sensitive compared to simple word-for-word substitution.

Fill in the blank: Machine translation systems calculate the probability of word or phrase __________ across languages.

| English Word (e) | Spanish Candidate (f) | Probability P(f\|e) | Contextual Usage |
|---|---|---|---|
| **Bank** | **Banco** | **0.85** | The standard financial or seating institution. |
| | **Orilla** | **0.12** | Specifically for a river bank (geographic context). |
| | **Ribera** | **0.03** | More poetic or specific technical river term. |
| **Light** | **Luz** | **0.60** | Used as a noun (e.g., "The light is on"). |
| | **Ligero** | **0.25** | Used as an adjective for weight (e.g., "A light box"). |
| | **Claro** | **0.15** | Referring to color or brightness (e.g., "Light blue"). |
| **Paper** | **Papel** | **0.75** | General material or a physical document. |
| | **Artículo** | **0.15** | Specifically an academic or journalistic "paper." |
| | **Periódico** | **0.10** | When "paper" is used as shorthand for newspaper. |

Table 4.3 Probability-based translation examples

**4.3.3 Text classification and sentiment analysis**

Text classification tasks, such as identifying spam emails or categorising news articles, often use probability to assign texts to categories. **Sentiment analysis** applies probability to determine whether a text expresses positive, negative, or neutral sentiment. For example, the phrase "I love this product" has a high probability of being classified as positive. These applications show how probability calculus enables machines to make informed decisions about text meaning and function.

Fill in the blank: The phrase "I love this product" has a high probability of being classified as __________ sentiment.

| Application | Key Feature (Likelihood) | Target Output (Class) |
|---|---|---|
| **Spam Detection** | Financial/Urgency keywords | Spam vs. Ham (Legitimate) |
| **Topic Categorization** | Domain-specific terminology | Politics, Science, Art, etc. |
| **Sentiment Analysis** | Adjectives & Emotional tokens | Positive, Negative, Neutral |

Box 4.3 Probability in text classification and sentiment analysis

### Pilot questions 4.3

1. How does probabilistic modelling improve speech recognition?
2. What role does probability play in machine translation?
3. Give one example of a probability-based translation.
4. What is text classification?
5. How is probability used in sentiment analysis?

## 4.4 Challenges and Future Directions

### 4.4.1 Limitations of probabilistic models in language

Probabilistic models, while powerful, have certain limitations. They often rely heavily on large amounts of data, which means they may struggle with languages or domains where data is scarce. Another limitation is that they usually capture surface-level patterns rather than deeper meanings, so they may misinterpret context or nuance. For example, an n-gram model might predict the next word correctly in a simple sentence but fail in complex or creative language use.

Fill in the blank: Probabilistic models often struggle in contexts where __________ is scarce.

| Feature | Probabilistic Model | Human Brain |
|---|---|---|
| **Learning Source** | Billions of text tokens | Real-world experience & sensory input |
| **Logic Type** | Mathematical frequency | Reasoning, empathy, and intent |
| **New Situations** | Struggles with "Out-of-Distribution" data | Can generalize from a single example |
| **Output Goal** | Maximizing likelihood | Communicating a specific meaning |

Box 4.4 Limitations of probabilistic models

### 4.4.2 Combining probability with deep learning

Recent advances in language technology combine probability with **deep learning** methods. Deep learning models, such as neural networks, can capture complex patterns in language, while probability provides a framework for managing uncertainty. Together, they create hybrid systems that are more accurate and robust. For instance, modern machine translation systems integrate probabilistic models with deep learning to produce fluent and context-sensitive translations.

Fill in the blank: Modern machine translation systems often combine probabilistic models with __________ learning.

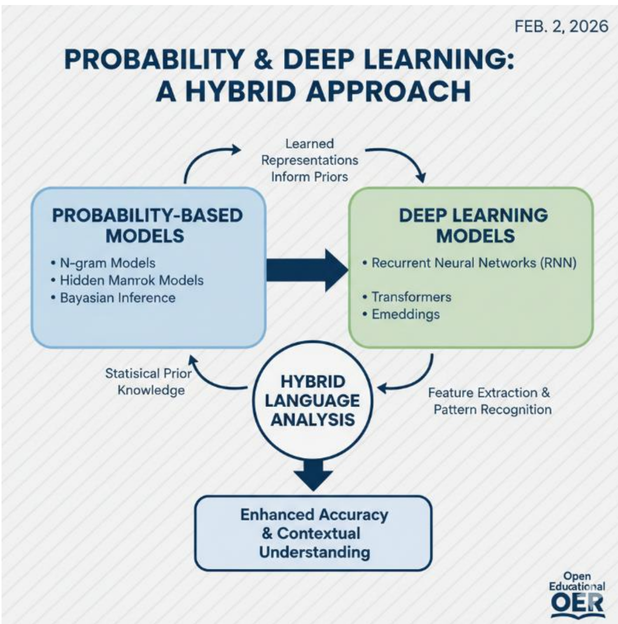



Figure 4.4 Combining probability with deep learning

### 4.4.3 Emerging trends in probabilistic language analysis

The future of probabilistic language analysis includes exciting trends. **Multimodal models** are being developed to analyse not only text but also speech, images, and video. **Low-resource language modelling** is another trend, focusing on applying probability to languages with limited data. Additionally, **explainable AI** is gaining importance, aiming to make probabilistic decisions more transparent and understandable to users. These trends show how probability calculus continues to evolve in response to new challenges and opportunities.

Fill in the blank: Probabilistic models that analyse text, speech, images, and video are called __________ models.


Plate 4.4 Emerging trends in probabilistic language analysis

### Pilot questions 4.4

1. State one limitation of probabilistic models in language analysis.
2. Why do probabilistic models struggle with low-resource languages?
3. How does combining probability with deep learning improve language analysis?
4. Give one example of a hybrid application in NLP.
5. What are multimodal models?
6. Mention one emerging trend in probabilistic language analysis.

### *Series Summary*

This Series has introduced you to the role of probability calculus in language analysis, showing how mathematical models can help us manage uncertainty and make predictions about linguistic behaviour. Its purpose has been to highlight the relevance of probability to both theoretical linguistics and practical applications in natural language processing, thereby strengthening your ability to connect abstract concepts with real-world language technologies.

We began by exploring the fundamentals of probability in linguistics, including definitions, random variables, events, and distributions. Then we examined probability models such as n-gram models, conditional probability, and Bayesian approaches. Following that, we considered practical applications in speech recognition, machine translation, and text classification. Finally, we discussed the challenges of probabilistic models, their integration with deep learning, and emerging trends such as multimodal analysis and explainable AI. By completing this Series, you should now be able to define probability in linguistic contexts, apply n-gram and conditional models, analyse Bayesian approaches, evaluate probabilistic applications in NLP, and assess both the limitations and future directions of probabilistic language analysis, as outlined in the Series Learning Outcomes.

## *Glossary of Terms*

- **Bayesian approach**: A method that uses Bayes' theorem to update probabilities when new evidence is introduced, often applied in tasks such as spam detection.
- **Conditional probability**: The probability of an event occurring given that another event has already happened, for example the likelihood of "morning" following "good."
- **Distribution**: A description of how probabilities are spread across possible outcomes, showing which words or structures are more frequent and which are rare.
- **Event**: A specific outcome in probability, such as the occurrence of a particular word in a text.
- **Multimodal model**: A probabilistic model that analyses multiple types of data such as text, speech, images, and video.
- **N-gram model**: A probabilistic model that predicts the likelihood of a word based on the previous *n-1* words, commonly used in text prediction and speech recognition.
- **Probability**: A measure of how likely an event is to occur, expressed as a number between 0 (impossible) and 1 (certain).
- **Probabilistic modelling**: The use of probability to manage uncertainty and select the most likely outcome, often applied in speech recognition and text analysis.
- **Random variable**: A quantity whose value depends on chance, such as the next word in a sentence.
- **Sentiment analysis**: A text classification method that uses probability to determine whether a text expresses positive, negative, or neutral sentiment.
- **Text classification**: The process of assigning texts to categories, such as spam or non-spam, often using probability to guide decisions.

## *Concept Check Questions [Series 4]*

**Objective questions**

1. Probability is expressed as a number between __________ and __________. (Linked to SLO 4.1)
2. A specific outcome in probability, such as the occurrence of a word, is called an __________. (Linked to SLO 4.2)
3. A model that predicts the likelihood of a word based on the previous word is called a __________ model. (Linked to SLO 4.3)
4. The probability of "morning" given "good" is an example of __________ probability. (Linked to SLO 4.4)
5. Bayesian approaches update probabilities based on new __________. (Linked to SLO 4.5)

**Short-answer questions**

6. Define probability in the context of language analysis. (Linked to SLO 4.1)
7. Explain what a random variable represents in linguistics. (Linked to SLO 4.2)
8. Describe one practical application of n-gram models. (Linked to SLO 4.3)

9. Give one example of how conditional probability is used in predictive text. (Linked to SLO 4.4)
10. State one limitation of probabilistic models in language analysis. (Linked to SLO 4.7)

**Long-answer questions**

11. Analyse the differences between n-gram models, conditional probability, and Bayesian approaches in language analysis. (Linked to SLOs 4.3, 4.4, 4.5)
12. Evaluate the role of probability calculus in natural language processing applications such as speech recognition, machine translation, and sentiment analysis. (Linked to SLO 4.6)
13. Assess the limitations of probabilistic models and discuss how emerging trends such as deep learning and multimodal analysis address these challenges. (Linked to SLO 4.7)

*Answers to Concept Check Questions [Series 4]*

**Objective questions**

1. 0 and 1.
2. Event.
3. Bigram model.
4. Conditional probability.
5. Evidence.

**Short-answer questions**

6. Probability is the measure of how likely an event is to occur, expressed between 0 and 1.
7. A random variable represents a possible linguistic outcome, such as the next word in a sentence.
8. N-gram models are used in text prediction, for example predicting the next word in a search query.
9. Predictive text uses conditional probability to suggest "morning" after typing "good."
10. One limitation is that probabilistic models depend heavily on large amounts of data.

**Long-answer questions**

11. *Basic response*: N-gram models predict words based on preceding context, conditional probability estimates likelihoods given prior words, and Bayesian approaches update probabilities with new evidence. *Value-added response*: N-gram models are simple and effective for local patterns but limited in scope. Conditional probability provides a mathematical foundation for predictive systems but can be data-intensive. Bayesian approaches are more flexible, combining prior knowledge with observed data, making them suitable for adaptive tasks such as spam detection.
12. *Basic response*: Probability calculus supports speech recognition by selecting the most likely word, machine translation by predicting natural correspondences, and sentiment analysis by classifying text as positive, negative, or neutral. *Value-added response*: Beyond these, probability enables systems to handle uncertainty in noisy data, improve contextual accuracy, and scale across

multilingual applications. It underpins many modern NLP systems, making them more robust and user-friendly.

13. *Basic response*: Probabilistic models are limited by data dependence and shallow contextual understanding. Emerging trends such as deep learning and multimodal analysis help overcome these challenges. *Value-added response*: Deep learning integrates probability with neural networks to capture complex patterns, while multimodal models extend analysis beyond text to include speech, images, and video. Explainable AI further enhances transparency, ensuring probabilistic decisions are interpretable and ethically sound.

## *Answers to Pilot Questions [Series 4]*

### Part 4.1 Fundamentals of Probability in Linguistics

1. Probability is the measure of how likely an event is to occur, expressed between 0 and 1.
2. 0 and 1.
3. A random variable is a quantity whose value depends on chance, such as the next word in a sentence.
4. An event.
5. A distribution shows how probabilities are spread across possible outcomes.
6. “Good morning” is an everyday example of probability in language use.

### Part 4.2 Probability Models in Language Analysis

1. An n-gram model predicts the likelihood of a word based on the previous *n-1* words.
2. They are applied in text prediction, spelling correction, and speech recognition.
3. Conditional probability is the probability of an event occurring given that another event has already occurred.
4. The probability of “morning” given “good.”
5. A Bayesian approach updates probabilities with new evidence.
6. Spam detection.

### Part 4.3 Applications of Probability Calculus in NLP

1. Probabilistic modelling improves speech recognition by selecting the most likely interpretation of sounds.
2. Probability helps machine translation predict the most natural and context-sensitive translation.
3. The English word “house” is most likely translated as “casa” in Spanish.
4. Text classification is the process of assigning texts to categories such as spam or non-spam.
5. Probability is used in sentiment analysis to classify text as positive, negative, or neutral.

### Part 4.4 Challenges and Future Directions

1. One limitation is that probabilistic models depend heavily on large amounts of data.
2. They struggle with low-resource languages because of limited data availability.

3. Combining probability with deep learning improves language analysis by capturing complex patterns and managing uncertainty.
4. Hybrid applications include modern machine translation systems that integrate probabilistic models with deep learning.
5. Multimodal models analyse text, speech, images, and video.
6. Emerging trends include multimodal analysis, low-resource language modelling, and explainable AI.

### *References and Attributions [Series 4]*

This section compiles all references and attributions for *Series 4: Probability Calculus in Language Analysis*. Entries include suggested open educational resources (OERs) and notes on AI-generated illustration prompts. The referencing style follows APA (7th edition).

**General References**


- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
- Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Draft manuscript, Stanford University.
- Norvig, P. (1992). *Paradigms of artificial intelligence programming: Case studies in Common Lisp*. Morgan Kaufmann.


**Illustrations**

- *Figure 4.1 Probability scale in language analysis*
  - Attribution: Suggested OER diagram (search string: "probability scale diagram linguistics OER").
  - AI prompt: "Generate a diagram showing probability as a scale from 0 to 1 with examples from language use."
  - Tool/platform: Microsoft Copilot Image Creator.
- *Table 4.1 Key probability concepts in linguistics*
  - Attribution: Suggested OER table (search string: "random variable event distribution linguistics OER table").
  - AI prompt: "Generate a table defining random variable, event, and distribution with examples from language."
  - Tool/platform: Microsoft Copilot Image Creator.
- *Box 4.1 Everyday examples of probability in language*
  - Attribution: Suggested OER resource (search string: "probability examples in everyday language linguistics OER summary").

- AI prompt: “Generate a box listing everyday examples of probability in language such as collocations and sound prediction.”
- Tool/platform: Microsoft Copilot Image Creator.

- *Figure 4.2 N-gram models in language prediction*
  - Attribution: Suggested OER diagram (search string: “n-gram model bigram trigram linguistics OER diagram”).
  - AI prompt: “Generate a diagram showing bigram and trigram models predicting words based on preceding context.”
  - Tool/platform: Microsoft Copilot Image Creator.
- *Table 4.2 Examples of conditional probability in language*
  - Attribution: Suggested OER table (search string: “conditional probability language prediction OER table”).
  - AI prompt: “Generate a table showing examples of conditional probability in language with word pairs.”
  - Tool/platform: Microsoft Copilot Image Creator.
- *Box 4.2 Bayesian approaches in language analysis*
  - Attribution: Suggested OER resource (search string: “Bayesian approaches language analysis Bayes theorem OER summary”).
  - AI prompt: “Generate a box summarising Bayes’ theorem and its application in language analysis.”
  - Tool/platform: Microsoft Copilot Image Creator.
- *Figure 4.3 Probabilistic modelling in speech recognition*
  - Attribution: Suggested OER diagram (search string: “probabilistic modelling speech recognition linguistics OER diagram”).
  - AI prompt: “Generate a diagram showing probabilistic modelling narrowing down possible word interpretations in speech recognition.”
  - Tool/platform: Microsoft Copilot Image Creator.
- *Table 4.3 Probability-based translation examples*
  - Attribution: Suggested OER table (search string: “probability-based machine translation examples OER table”).
  - AI prompt: “Generate a table showing examples of English words with their most probable translations in Spanish.”
  - Tool/platform: Microsoft Copilot Image Creator.
- *Box 4.3 Probability in text classification and sentiment analysis*

- Attribution: Suggested OER resource (search string: “probability text classification sentiment analysis OER summary”).
- AI prompt: “Generate a box summarising probability applications in spam detection, topic categorisation, and sentiment analysis.”
- Tool/platform: Microsoft Copilot Image Creator.

- *Box 4.4 Limitations of probabilistic models*
  - Attribution: Suggested OER resource (search string: “limitations of probabilistic models language analysis OER summary”).
  - AI prompt: “Generate a box listing limitations of probabilistic models in language analysis.”
  - Tool/platform: Microsoft Copilot Image Creator.
- *Figure 4.4 Combining probability with deep learning*
  - Attribution: Suggested OER diagram (search string: “probability deep learning hybrid models language OER diagram”).
  - AI prompt: “Generate a diagram showing probability and deep learning working together in language analysis.”
  - Tool/platform: Microsoft Copilot Image Creator.
- *Plate 4.4 Emerging trends in probabilistic language analysis*
  - Attribution: Suggested OER image (search string: “multimodal probabilistic language analysis OER image”).
  - AI prompt: “Generate an illustration of a researcher working with multimodal data including text, speech, and images.”
  - Tool/platform: Microsoft Copilot Image Creator.

**Course code: LIN 309**
**Course title: Computational Linguistics I**
**Course credit load: 2 Units**

**Series 1:** Foundations of Computational Linguistics

**Series 2:** Natural Language Processing Problems and Perspectives

**Series 3:** Corpus Linguistics Essentials

**Series 4:** Probability Calculus in Language Analysis

**Series 5:** N-Grams and Language Modeling

**Proposed Outline for Series 5**

**Part .1 Introduction to N-grams**

- 5.1.1 Definition of n-grams and their role in language analysis
- 5.1.2 Types of n-grams: unigram, bigram, trigram, and beyond
- 5.1.3 Advantages and limitations of n-gram models

**Part .2 Building and Applying N-gram Models**

- 5.2.1 Constructing n-gram models from corpora
- 5.2.2 Calculating probabilities with n-grams
- 5.2.3 Applications in predictive text and speech recognition

**Part .3 Smoothing and Advanced Techniques**

- 5.3.1 The problem of sparse data in n-gram models
- 5.3.2 Smoothing methods: Laplace, Good-Turing, and back-off
- 5.3.3 Evaluating n-gram models

**Part .4 N-grams in Modern Language Modelling**

- 5.4.1 Transition from n-gram models to neural language models
- 5.4.2 Hybrid approaches combining n-grams and deep learning
- 5.4.3 Future directions in language modelling

**Introduction**

In the last Series, we explored probability calculus in language analysis and saw how mathematical models can help us manage uncertainty and predict linguistic behaviour. Building on that foundation, we now turn to **n-grams and language modelling**, which provide one of the most practical ways of applying probability to real language data. N-gram models are widely used in predictive text, speech recognition, and machine translation, making them an essential tool for understanding how computers process and generate language.

The aim of this Series is to introduce you to the principles and applications of n-gram models, to equip you with the skills to construct, apply, and evaluate these models, and to connect them with modern approaches to language modelling.

We shall commence this Series by defining n-grams and examining their types, advantages, and limitations. Next, we will move on to building and applying n-gram models, including how they are constructed from corpora and used in predictive systems. Following that, we will consider smoothing techniques and advanced methods that address the problem of sparse data and improve model performance. Finally, we will wrap up by exploring how n-grams fit into modern language modelling, including their transition to neural approaches, hybrid systems, and future directions in the field.

**Series Learning Outcomes**

Upon completing this Series, you should be able to:

- **SLO 5.1** define n-grams and explain their role in language modelling,
- **SLO 5.2** describe the different types of n-grams, including unigrams, bigrams, and trigrams,
- **SLO 5.3** demonstrate how to construct n-gram models from corpora,
- **SLO 5.4** calculate probabilities using n-gram models,
- **SLO 5.5** apply n-gram models in tasks such as predictive text and speech recognition,
- **SLO 5.6** analyse the problem of sparse data in n-gram models,
- **SLO 5.7** evaluate smoothing techniques such as Laplace, Good-Turing, and back-off methods,
- **SLO 5.8** assess the performance of n-gram models using appropriate evaluation measures, and
- **SLO 5.9** compare traditional n-gram models with modern neural language models and discuss future directions.

## 5.1 Introduction to N-grams

### 5.1.1 Definition of n-grams and their role in language analysis

An **n-gram** is a sequence of *n* items, usually words or characters, taken from a given text or speech sample. For example, a bigram is a sequence of two words, while a trigram is a sequence of three. N-grams are fundamental in language analysis because they allow us to capture local patterns and predict what comes next in a sentence. They are widely used in applications such as predictive text, search engines, and speech recognition.

Fill in the blank: A sequence of three words taken from a text is called a __________.

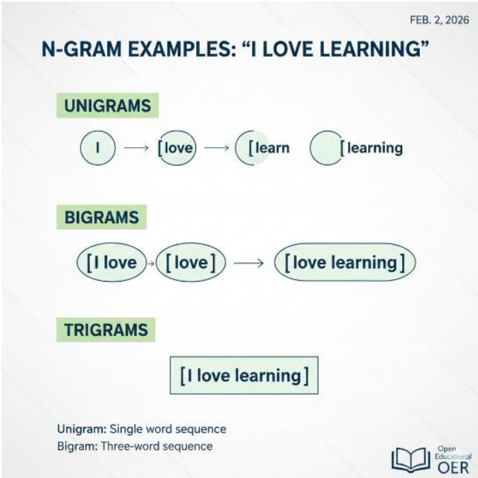



Figure 5.1 Examples of n-grams in a sentence

### 5.1.2 Types of n-grams: unigram, bigram, trigram, and beyond

A **unigram** is a single word treated as an independent unit. A **bigram** considers two consecutive words, while a **trigram** looks at three. Higher-order n-grams, such as four-grams or five-grams, capture longer contexts but require more data to be effective. Each type of n-gram balances simplicity and predictive power, with unigrams being easy to compute but less informative, and higher-order n-grams offering richer context but facing challenges like data sparsity.

Fill in the blank: A single word treated as an independent unit is called a __________.

| Type | Size (n) | Example (from "The quick brown fox") | Key Features & Applications |
|---|---|---|---|
| **Unigram** | 1 | "The", "quick", "brown", "fox" | **Features:** Ignores word order. High frequency.<br><br>**Uses:** Word clouds, simple sentiment analysis. |
| **Bigram** | 2 | "The quick", "quick brown", "brown fox" | **Features:** Captures basic local context (adjective-noun pairs). |

| Type | Size (n) | Example (from "The quick brown fox") | Key Features & Applications |
|---|---|---|---|
| | | | **Uses:** Autocomplete, simple spelling correction. |
| **Trigram** | 3 | "The quick brown", "quick brown fox" | **Features:** Captures phrase-level meaning.<br>**Uses:** Statistical Machine Translation, spam filtering. |
| **4-gram** | 4 | "The quick brown fox" | **Features:** Captures specific idiomatic expressions.<br>**Uses:** Plagiarism detection, high-quality text generation. |
| **Higher-order** | 5+ | "The quick brown fox jumped" | **Features:** Very specific but rare. Leads to "Sparsity."<br>**Uses:** Specialized technical domain modeling. |

Table 5.1 Types of n-grams and their characteristics

### 5.1.3 Advantages and limitations of n-gram models

N-gram models are simple, intuitive, and effective for capturing local word patterns. They are easy to build from text corpora and form the basis of many natural language processing systems.

However, they also have limitations. One major issue is **data sparsity**, which occurs when higher-order n-grams are rare or absent in the training data. Another limitation is that n-grams focus only on local context and may fail to capture deeper meaning or long-distance dependencies in language.

Fill in the blank: The problem that arises when higher-order n-grams are rare or absent in training data is called __________.

| Type | Size (n) | Example (from "The quick brown fox") | Key Features & Applications |
|---|---|---|---|
| **Unigram** | 1 | "The", "quick", "brown", "fox" | **Features:** Ignores word order. High frequency.<br><br>**Uses:** Word clouds, simple sentiment analysis. |
| **Bigram** | 2 | "The quick", "quick brown", "brown fox" | **Features:** Captures basic local context (adjective-noun pairs).<br><br>**Uses:** Autocomplete, simple spelling correction. |
| **Trigram** | 3 | "The quick brown", "quick brown fox" | **Features:** Captures phrase-level meaning.<br><br>**Uses:** Statistical Machine Translation, spam filtering. |
| **4-gram** | 4 | "The quick brown fox" | **Features:** Captures specific idiomatic expressions. |

| Type | Size (n) | Example (from "The quick brown fox") | Key Features & Applications |
|---|---|---|---|
| | | | **Uses:** Plagiarism detection, high-quality text generation. |
| **Higher-order** | 5+ | "The quick brown fox jumped" | **Features:** Very specific but rare. Leads to "Sparsity."<br><br>**Uses:** Specialized technical domain modeling. |

Box 5.1 Advantages and limitations of n-gram models

Pilot questions 5.1

1. What is an n-gram?
2. Give one example of a trigram.
3. Define a unigram in your own words.
4. What is the main advantage of n-gram models?
5. What is data sparsity?
6. State one limitation of n-gram models.

## 5.2 Building and Applying N-gram Models

### 5.2.1 Constructing n-gram models from corpora

To build an **n-gram model**, we start with a **corpus**, which is a large collection of texts. The corpus is segmented into sequences of words, and counts are taken of how often each n-gram occurs. These counts are then converted into probabilities, showing how likely a word or sequence is to appear. For example, in a corpus of news articles, the bigram "breaking news" may occur frequently, giving it a high probability compared to less common pairs.

Fill in the blank: A large collection of texts used to build an n-gram model is called a __________.

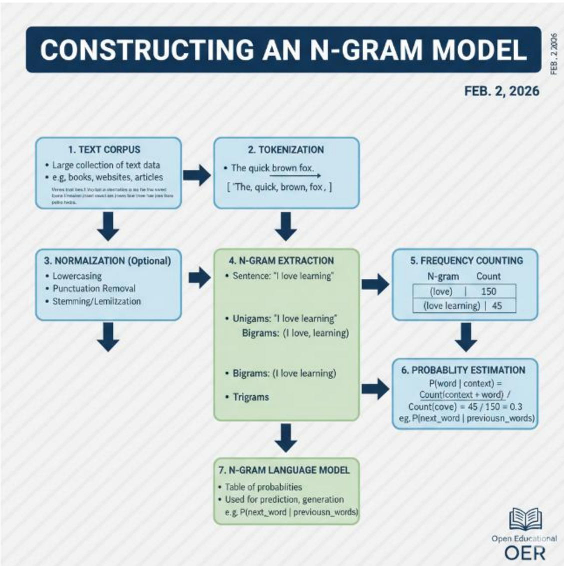
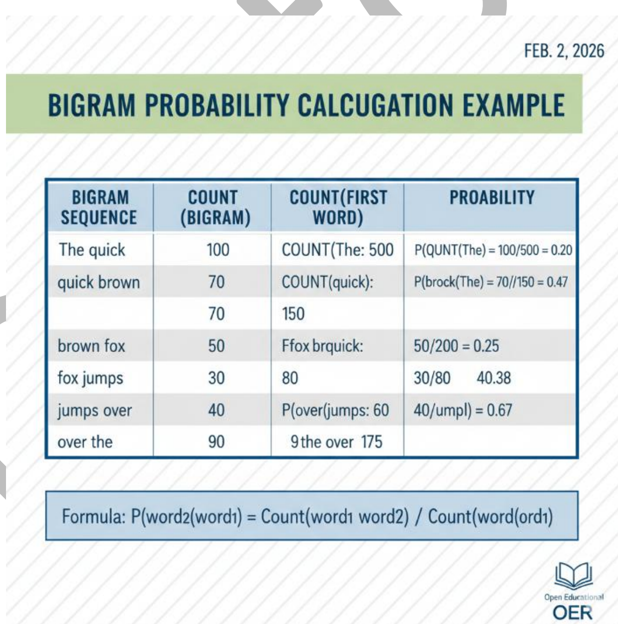



Figure 5.2 Constructing an n-gram model from a corpus

### 5.2.2 Calculating probabilities with n-grams

Once counts are obtained, probabilities are calculated by dividing the frequency of an n-gram by the total number of possible sequences. This gives us the likelihood of a word appearing in a given context. For example, the probability of "morning" following "good" is calculated by dividing the number of times "good morning" occurs by the total occurrences of "good" in the corpus. This simple calculation makes n-gram models powerful tools for predicting language patterns.

Fill in the blank: The probability of "morning" given "good" is calculated by dividing the frequency of "good morning" by the total occurrences of __________.


FEB. 2, 2026

BIGRAM PROBABILITY CALCUGATION EXAMPLE

| BIGRAM SEQUENCE | COUNT (BIGRAM) | COUNT(FIRST WORD) | PROABILITY |
|---|---|---|---|
| The quick | 100 | COUNT(The: 500 | P(QUNT(The) = 100/500 = 0.20 |
| quick brown | 70 | COUNT(quick): | P(brock(The) = 70//150 = 0.47 |
| | 70 | 150 | |
| brown fox | 50 | Ffox brquick: | 50/200 = 0.25 |
| fox jumps | 30 | 80 | 30/80 40.38 |
| jumps over | 40 | P(over(jumps: 60 | 40/umpl) = 0.67 |
| over the | 90 | 9 the over 175 | |

Formula: P(word2(word1) = Count(word1 word2) / Count(word(ord1)

Open Educational
OER

Table 5.2 Calculating probabilities with n-grams

### 5.2.3 Applications in predictive text and speech recognition

N-gram models are widely applied in predictive text systems, such as those on mobile phones, where they suggest the next word based on previous input. They are also used in speech recognition, helping systems choose the most probable word sequence from audio signals. For instance, if the sound could be interpreted as "there" or "their," the model uses probabilities from the corpus to decide which fits best in the sentence. These applications show how n-gram models bridge theory and practice in language technology.

Fill in the blank: Predictive text systems use n-gram models to suggest the __________ word based on previous input.

| Feature | N-Gram Model | Neural Model (LLM) |
|---|---|---|
| **Speed** | Near-instant (Low latency) | Slower (High latency) |
| **Power Use** | Minimal (Edge/Mobile) | High (Server/GPU) |
| **Context** | Short (Last 2-4 words) | Long (Thousands of words) |
| **Role** | Immediate Prediction/ASR | Complex Reasoning/Chat |

Box 5.2 Applications of n-gram models

Pilot questions 5.2

1. What is a corpus in the context of n-gram models?
2. How are n-gram counts converted into probabilities?
3. Give one example of calculating a bigram probability.
4. State one application of n-gram models in predictive text.
5. How do n-gram models assist in speech recognition?

## 5.3 Smoothing and Advanced Techniques

### 5.3.1 The problem of sparse data in n-gram models

One of the main challenges in working with n-gram models is **data sparsity**, which occurs when certain word sequences are rare or absent in the training corpus. For example, a trigram like "bright red balloon" may not appear often enough to provide a reliable probability estimate. Sparse data leads to zero probabilities for unseen sequences, which can cause models to fail in predicting reasonable outputs.

Fill in the blank: When certain word sequences are rare or absent in the training corpus, the problem is called __________.

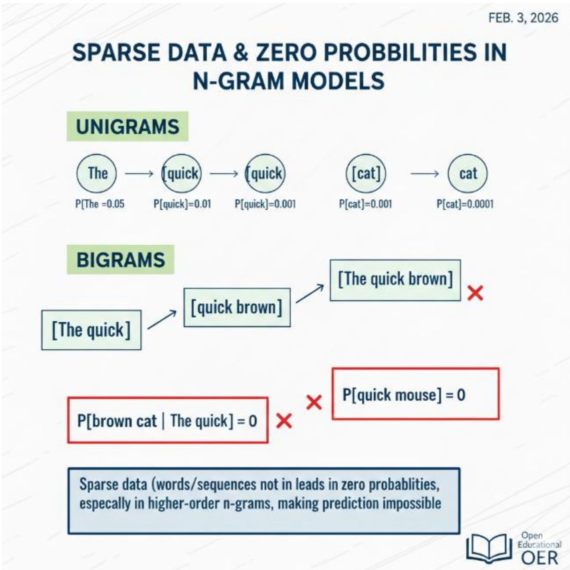



Figure 5.3 Sparse data problem in n-gram models

**5.3.2 Smoothing methods: Laplace, Good-Turing, and back-off**

**Smoothing** refers to techniques used to adjust probability estimates so that unseen sequences are given small but non-zero probabilities. **Laplace smoothing** adds one to all counts to avoid zeros. **Good-Turing smoothing** redistributes probability mass from frequent events to rare or unseen ones. **Back-off methods** use lower-order n-grams when higher-order ones are unavailable. These techniques ensure that models remain robust even when data is limited.

Fill in the blank: The smoothing method that adds one to all counts to avoid zero probabilities is called __________ smoothing.

| Smoothing Method | Core Definition | How it Works | Example / Context |
|---|---|---|---|
| **Laplace (Add-One)** | The simplest method; assumes every possible n-gram occurs at least once. | Add 1 to all n-gram counts. The denominator increases by the vocabulary size ($V$). | **P(apple \| green)**: If "green apple" appeared 0 times, it now appears 1 time. |
| **Good-Turing** | Estimates the probability of unseen items based on items seen exactly once. | Uses the count of things seen $r+1$ times to estimate the probability | If you see 10 words only once, you use that "10" to guess how |

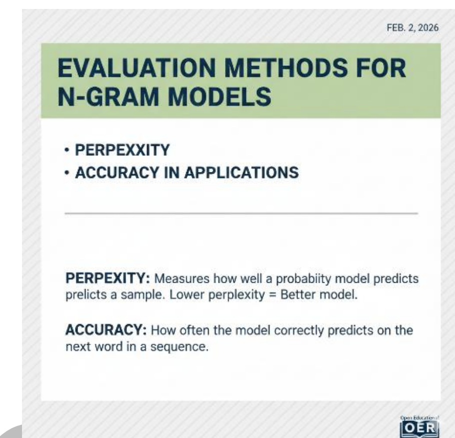
| Smoothing Method | Core Definition | How it Works | Example / Context |
|---|---|---|---|
| | | of things seen $r$ times. | many words you *haven't* seen. |
| **Back-off (Katz)** | A hierarchical approach that "drops down" to simpler models when data is missing. | If a **trigram** (3-word) count is 0, the model looks at the **bigram** (2-word) probability instead. | Predicting "morning" after "very good". If "very good morning" is unknown, check "good morning." |

Table 5.3 Common smoothing methods in n-gram models

### 5.3.3 Evaluating n-gram models

To ensure n-gram models are effective, they must be evaluated using measures such as **perplexity**, which indicates how well a model predicts a sample of text. Lower perplexity means better predictive performance. Evaluation also considers accuracy in real applications, such as predictive text or speech recognition. By testing models on unseen data, we can determine whether they generalise well beyond the training corpus.

Fill in the blank: The measure that indicates how well a model predicts a sample of text is called __________.




Box 5.3 Evaluation of n-gram models

## Pilot questions 5.3

1. What is data sparsity in n-gram models?
2. State one consequence of sparse data.
3. What does Laplace smoothing do to counts?
4. How does Good-Turing smoothing handle rare events?

5. What is the purpose of back-off methods?
6. What does perplexity measure in n-gram models?

## 5.4 N-grams in Modern Language Modelling

### 5.4.1 Transition from n-gram models to neural language models

While **n-gram models** have been foundational in natural language processing, they are limited in capturing long-distance dependencies and deeper meaning. Modern systems increasingly rely on **neural language models**, which use artificial neural networks to learn complex patterns in text. These models can represent words in continuous vector spaces, allowing them to capture semantic relationships beyond local word sequences. This transition marks a shift from simple probability counts to more sophisticated learning approaches.

Fill in the blank: Modern language models use __________ networks to capture complex patterns in text.

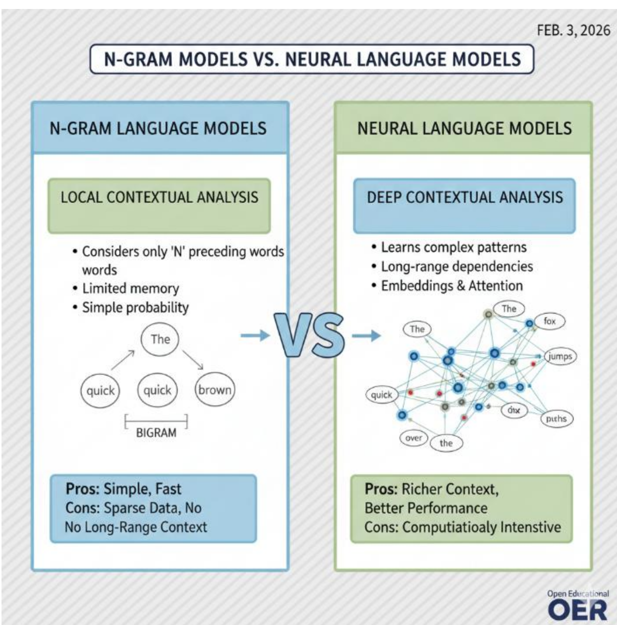



Figure 5.4 Transition from n-gram to neural language models

### 5.4.2 Hybrid approaches combining n-grams and deep learning

In practice, some systems combine the strengths of n-gram models with deep learning. For example, n-grams may be used for fast initial predictions, while neural models refine the output for accuracy and fluency. This hybrid approach ensures efficiency while maintaining high-quality results. It demonstrates how traditional methods remain relevant when integrated with modern techniques.

Fill in the blank: Hybrid approaches combine the strengths of n-gram models with __________ learning.

| Feature | N-Gram Contribution | Deep Learning Contribution |
|---|---|---|
| **Speed** | Instantaneous lookup. | Slower, parallel processing. |
| **Context** | Local (next 2–3 words). | Global (entire document). |
| **Interpretability** | "Glass Box" (Easy to audit). | "Black Box" (Hard to audit). |
| **Data Needs** | Works with small data. | Needs massive data. |
| **Hybrid Benefit** | **High efficiency on edge devices with near-human accuracy.** | |

Box 5.4 Hybrid approaches in language modelling

### 5.4.3 Future directions in language modelling

The future of language modelling is moving towards **multimodal systems**, which integrate text, speech, and images, and **explainable AI**, which makes model decisions more transparent. While neural models dominate current research, n-grams still play a role in lightweight applications and as baselines for comparison. Learners should appreciate that language modelling is a dynamic field, with ongoing innovations that build on the foundations of probability and n-gram analysis.

Fill in the blank: Models that integrate text, speech, and images are called __________ systems.

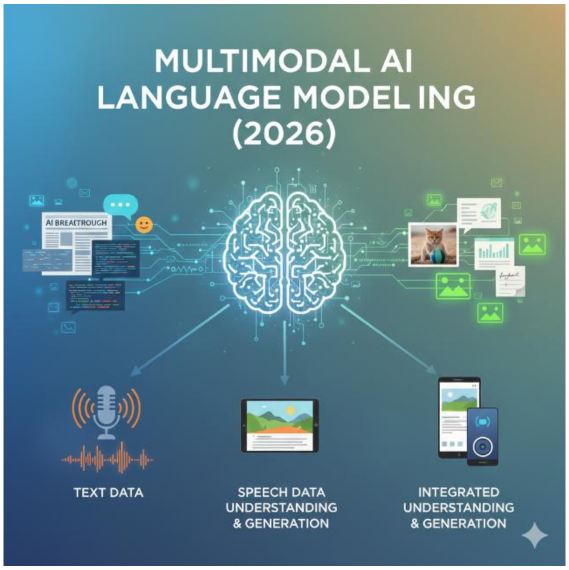



Plate 5.4 Future directions in language modelling

## Pilot questions 5.4

1. What is the main limitation of n-gram models compared to neural models?
2. How do neural language models represent words?
3. Give one example of a hybrid approach in language modelling.
4. What are multimodal systems?
5. Mention one emerging trend in language modelling.

## *Series Summary*

This Series has focused on the role of n-gram models in language analysis, showing how they provide a practical way of applying probability to real text data. The purpose has been to help you understand how sequences of words can be used to predict language patterns and support applications such as predictive text and speech recognition.

We began by introducing the concept of n-grams, their types, and their advantages and limitations. Then we explored how n-gram models are constructed from corpora, how probabilities are calculated, and how these models are applied in everyday technologies. Following that, we examined the problem of sparse data and studied smoothing techniques such as Laplace, Good-Turing, and back-off, alongside methods of evaluating model performance. Finally, we considered the transition from traditional n-gram models to neural language models, hybrid approaches, and future directions in language modelling. By completing this Series, you should now be able to define and describe n-grams, construct and apply n-gram models, analyse data sparsity, evaluate smoothing methods, assess model performance, and compare n-gram approaches with modern neural models, as outlined in the Series Learning Outcomes.

## *Glossary of Terms*

- **Back-off method**: A smoothing technique that uses lower-order n-grams when higher-order ones are unavailable, ensuring predictions are still possible.
- **Bigram**: An n-gram consisting of two consecutive words, such as “good morning.”
- **Corpus**: A large collection of texts used as data for building and testing language models.
- **Data sparsity**: A problem in n-gram models where certain word sequences are rare or absent in the training data, leading to zero probabilities.
- **Good-Turing smoothing**: A method that redistributes probability from frequent events to rare or unseen ones, helping models handle sparse data.
- **Hybrid approach**: A language modelling method that combines traditional n-gram models with modern deep learning techniques for efficiency and accuracy.
- **Laplace smoothing**: A simple smoothing technique that adds one to all counts to avoid zero probabilities for unseen sequences.
- **Multimodal system**: A model that integrates multiple types of input such as text, speech, and images.
- **Neural language model**: A modern approach to language modelling that uses artificial neural networks to capture complex patterns and long-distance dependencies in text.
- **N-gram**: A sequence of *n* items (usually words) taken from a text or speech sample, used to predict language patterns.
- **Perplexity**: A measure of how well a language model predicts a sample of text, with lower values indicating better performance.
- **Predictive text**: A technology that uses n-gram models to suggest the next word based on previous input, commonly found in mobile keyboards.
- **Speech recognition**: A technology that converts spoken language into text, often supported by n-gram models to select the most probable word sequence.
- **Trigram**: An n-gram consisting of three consecutive words, such as “I love learning.”
- **Unigram**: An n-gram consisting of a single word treated as an independent unit.

## *Concept Check Questions [Series 5]*

**Objective questions**

1. An n-gram consisting of a single word is called a __________. (Linked to SLO 5.2)
2. The smoothing method that adds one to all counts to avoid zero probabilities is called __________ smoothing. (Linked to SLO 5.7)
3. The measure used to evaluate how well a language model predicts text is called __________. (Linked to SLO 5.8)
4. A large collection of texts used to build n-gram models is called a __________. (Linked to SLO 5.3)
5. Modern language models use __________ networks to capture complex patterns in text. (Linked to SLO 5.9)

**Short-answer questions**

6. Define an n-gram in the context of language modelling. (Linked to SLO 5.1)
7. Give one example of a trigram. (Linked to SLO 5.2)
8. Explain how probabilities are calculated in bigram models. (Linked to SLO 5.4)
9. State one application of n-gram models in predictive text. (Linked to SLO 5.5)
10. Describe the problem of data sparsity in n-gram models. (Linked to SLO 5.6)

**Long-answer questions**

11. Analyse the differences between Laplace, Good-Turing, and back-off smoothing methods. (Linked to SLO 5.7)
12. Evaluate the usefulness of perplexity as a measure of n-gram model performance. (Linked to SLO 5.8)
13. Compare traditional n-gram models with neural language models, highlighting their strengths and limitations. (Linked to SLO 5.9)

*Answers to Concept Check Questions [Series 5]*

**Objective questions**

1. Unigram.
2. Laplace smoothing.
3. Perplexity.
4. Corpus.
5. Neural.

**Short-answer questions**

6. An n-gram is a sequence of *n* items, usually words, taken from a text or speech sample.
7. Example: "I love learning."
8. By dividing the frequency of a bigram (e.g., "good morning") by the total occurrences of the first word ("good").
9. Predictive text uses n-gram models to suggest the next word based on previous input.
10. Data sparsity occurs when certain word sequences are rare or absent in the training corpus, leading to zero probabilities.

**Long-answer questions**

11. *Basic response*: Laplace smoothing adds one to all counts, Good-Turing redistributes probability to rare events, and back-off uses lower-order n-grams when higher-order ones are unavailable. *Value-added response*: Laplace is simple but can distort probabilities, Good-Turing is more accurate for rare events, and back-off ensures flexibility by falling back to simpler models. Together, they address different aspects of sparse data.
12. *Basic response*: Perplexity measures how well a model predicts text, with lower values indicating better performance. *Value-added response*: While perplexity is useful for comparing models, it does not always reflect real-world performance in applications like speech recognition. Combining perplexity with task-specific accuracy provides a fuller evaluation.

13. *Basic response*: N-gram models are simple, easy to build, and effective for local patterns, but they struggle with long-distance dependencies. Neural models capture deeper meaning and context but require more data and computational resources. *Value-added response*: N-gram models remain useful in lightweight applications and as baselines, while neural models dominate modern NLP tasks. Hybrid approaches combine efficiency with accuracy, and future directions include multimodal and explainable AI systems.

## *Answers to Pilot Questions [Series 5]*

### Part 5.1 Introduction to N-grams

1. An n-gram is a sequence of *n* items, usually words, taken from a text or speech sample.
2. Example: "bright red balloon."
3. A unigram is a single word treated as an independent unit.
4. Their main advantage is simplicity and effectiveness in capturing local word patterns.
5. Data sparsity.
6. They cannot capture long-distance dependencies and often struggle with rare word sequences.

### Part 5.2 Building and Applying N-gram Models

1. A corpus.
2. By dividing the frequency of an n-gram by the total number of possible sequences.
3. Probability of "morning" given "good" = frequency of "good morning" ÷ total occurrences of "good."
4. They suggest the next word based on previous input.
5. They help choose the most probable word sequence from audio signals.

### Part 5.3 Smoothing and Advanced Techniques

1. Data sparsity.
2. It leads to zero probabilities for unseen sequences.
3. It adds one to all counts to avoid zero probabilities.
4. It redistributes probability mass from frequent events to rare or unseen ones.
5. They use lower-order n-grams when higher-order ones are unavailable.
6. Perplexity.

### Part 5.4 N-grams in Modern Language Modelling

1. They struggle with long-distance dependencies and deeper meaning.
2. Neural language models represent words in continuous vector spaces.
3. Using n-grams for initial predictions and neural models for refinement.
4. Systems that integrate text, speech, and images.
5. Emerging trends include multimodal systems and explainable AI.

## *References and Attributions [Series 5]*

This section compiles all references and attributions for *Series 5: N-Grams and Language Modelling*. Entries include suggested open educational resources (OERs) and notes on AI-generated illustration prompts. The referencing style follows APA (7th edition).

**General References**


- Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Draft manuscript, Stanford University.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
- Norvig, P. (1992). *Paradigms of artificial intelligence programming: Case studies in Common Lisp*. Morgan Kaufmann.


**Illustrations**

- *Figure 5.1 Examples of n-grams in a sentence*
  - Attribution: Suggested OER diagram (search string: "n-gram examples unigram bigram trigram linguistics OER diagram").
  - AI prompt: "Generate a diagram showing unigram, bigram, and trigram sequences from the sentence 'I love learning'."
  - Tool/platform: Microsoft Copilot Image Creator.
- *Table 5.1 Types of n-grams and their characteristics*
  - Attribution: Suggested OER table (search string: "types of n-grams examples linguistics OER table").
  - AI prompt: "Generate a table listing types of n-grams (unigram, bigram, trigram, higher-order) with examples and features."
  - Tool/platform: Microsoft Copilot Image Creator.
- *Box 5.1 Advantages and limitations of n-gram models*
  - Attribution: Suggested OER resource (search string: "advantages limitations n-gram models linguistics OER summary").
  - AI prompt: "Generate a box summarising advantages and limitations of n-gram models in language analysis."
  - Tool/platform: Microsoft Copilot Image Creator.
- *Figure 5.2 Constructing an n-gram model from a corpus*
  - Attribution: Suggested OER diagram (search string: "constructing n-gram model corpus linguistics OER diagram").
  - AI prompt: "Generate a diagram showing the process of constructing an n-gram model from a text corpus."
  - Tool/platform: Microsoft Copilot Image Creator.
- *Table 5.2 Calculating probabilities with n-grams*
  - Attribution: Suggested OER table (search string: "bigram probability calculation example linguistics OER table").
  - AI prompt: "Generate a table showing sample calculations of bigram probabilities using word counts."

    - Tool/platform: Microsoft Copilot Image Creator.
- *Box 5.2 Applications of n-gram models*
    - Attribution: Suggested OER resource (search string: "applications of n-gram models predictive text speech recognition OER summary").
    - AI prompt: "Generate a box summarising applications of n-gram models in predictive text and speech recognition."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Figure 5.3 Sparse data problem in n-gram models*
    - Attribution: Suggested OER diagram (search string: "sparse data n-gram models linguistics OER diagram").
    - AI prompt: "Generate a diagram showing sparse data leading to zero probabilities in higher-order n-gram models."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Table 5.3 Common smoothing methods in n-gram models*
    - Attribution: Suggested OER table (search string: "smoothing methods Laplace Good-Turing back-off n-gram models OER table").
    - AI prompt: "Generate a table comparing Laplace, Good-Turing, and back-off smoothing methods with definitions and examples."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Box 5.3 Evaluation of n-gram models*
    - Attribution: Suggested OER resource (search string: "evaluation of n-gram models perplexity linguistics OER summary").
    - AI prompt: "Generate a box summarising evaluation methods for n-gram models including perplexity and accuracy."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Figure 5.4 Transition from n-gram to neural language models*
    - Attribution: Suggested OER diagram (search string: "n-gram vs neural language models linguistics OER diagram").
    - AI prompt: "Generate a diagram comparing n-gram models with neural language models, showing local versus deep contextual analysis."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Box 5.4 Hybrid approaches in language modelling*
    - Attribution: Suggested OER resource (search string: "hybrid approaches n-gram deep learning language modelling OER summary").
    - AI prompt: "Generate a box summarising hybrid approaches combining n-gram models and deep learning."
    - Tool/platform: Microsoft Copilot Image Creator.
- *Plate 5.4 Future directions in language modelling*
    - Attribution: Suggested OER image (search string: "multimodal language modelling text speech image OER illustration").
    - AI prompt: "Generate an illustration showing multimodal language modelling with text, speech, and image inputs."
    - Tool/platform: Microsoft Copilot Image Creator.