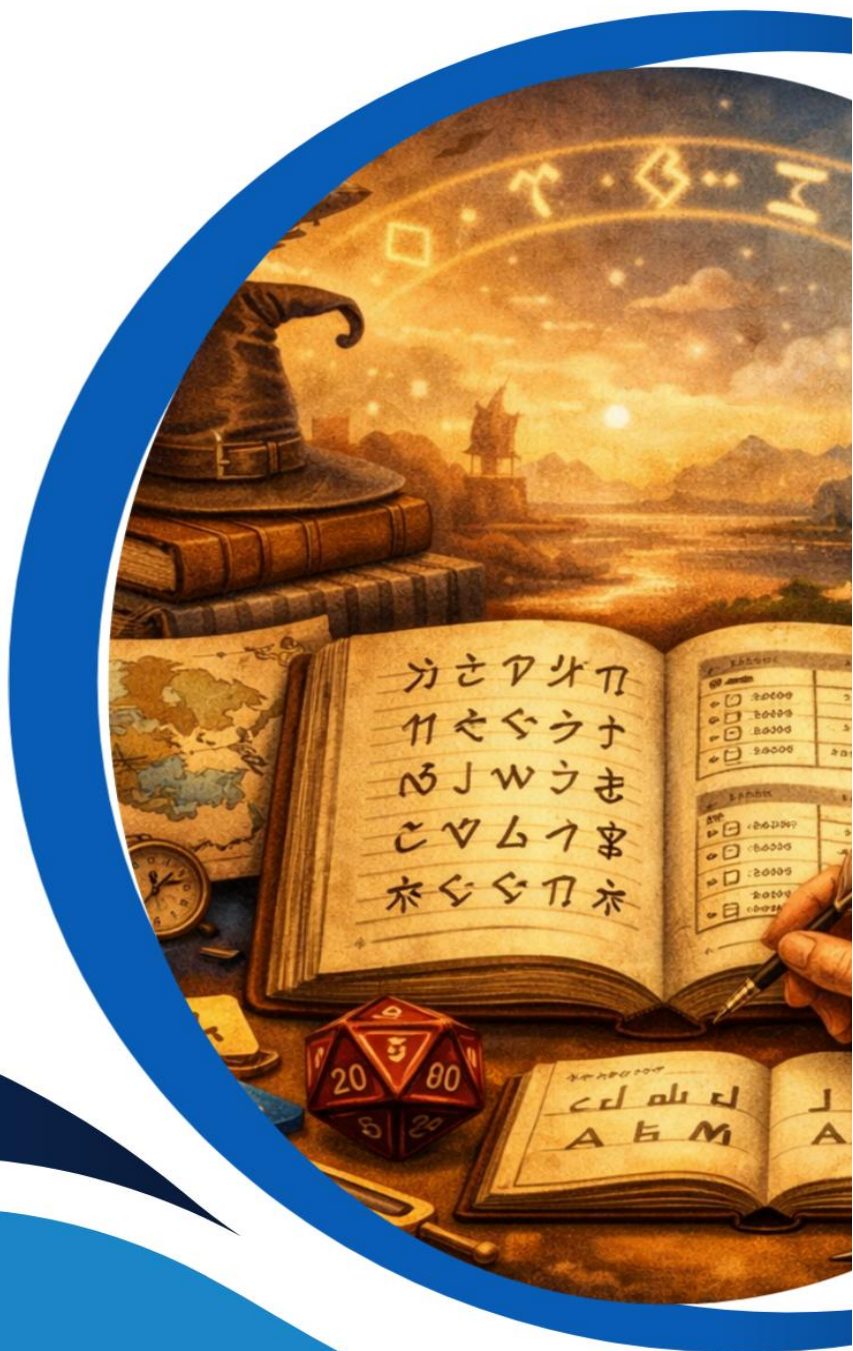


LIN307

NEW AND CREATED LANGUAGES



Download on the
App Store



GET IT ON
Google Play

Course video is available on our app

lanetonline.com

Course code: LIN 307
Course title: New and Created Languages
Course credit load: 2 Units

Series 1: Foundations of Digital and Corpus Linguistics

Series 2: Perspectives and Approaches in Corpus-Based Research

Series 3: Levels of Linguistic Analysis in Corpus Studies

Series 4: Applications of Corpus Linguistics in Language and Education

Series 5: Tools, Techniques, and Skills for Corpus Analysis

Proposed Outline for Series 1

Part 1: Introduction to Digital Linguistics

- Sub-Part 1.1: Definition and scope of digital linguistics
- Sub-Part 1.2: Key digital tools and resources in linguistics
- Sub-Part 1.3: Role of digitisation and databases in language study

Part 2: Foundations of Corpus Linguistics

- Sub-Part 2.1: Definition and scope of corpus linguistics
- Sub-Part 2.2: Historical development and significance of corpora
- Sub-Part 2.3: Types of corpora (monolingual, multilingual, specialised, general)

Part 3: Linking Digital and Corpus Linguistics

- Sub-Part 3.1: Similarities and differences between digital and corpus linguistics
- Sub-Part 3.2: Integration of digital tools in corpus-based research
- Sub-Part 3.3: Examples of applications in linguistics on the web

Part 4: Basic Techniques in Digital and Corpus Linguistics

- Sub-Part 4.1: Searching and annotation in corpora
- Sub-Part 4.2: Data mining and data visualisation
- Sub-Part 4.3: Introduction to software packages for corpus analysis

Introduction

Language today is studied not only through traditional observation but also through digital tools and large collections of texts known as corpora. With the rapid growth of technology and the internet, linguists now have access to vast amounts of language data that can be stored, searched,



and analysed in ways that were once unimaginable. This Series sets the foundation for understanding how digital methods and corpus-based approaches have transformed the study of language, making it more precise, systematic, and applicable to real-world contexts such as education, dictionary creation, and language teaching.

The aim of this Series is to introduce you to the fundamental concepts of digital linguistics and corpus linguistics, and to equip you with the essential knowledge needed to appreciate their scope, connections, and practical applications.

We shall commence this Series by examining the definition and scope of digital linguistics, including the role of digitisation and databases. Next, we will explore the foundations of corpus linguistics, looking at its historical development and the different types of corpora. Following that, we will consider how digital and corpus linguistics intersect, highlighting their similarities, differences, and integration in research. Finally, we will wrap up by introducing basic techniques such as searching, annotation, data mining, and visualisation, alongside the use of software packages that support corpus analysis. This progression will ensure that you gain a clear and structured understanding of the foundations upon which later Series will build.

Series Learning Outcomes

Upon completing this Series, you should be able to:

- SLO 1.1 define the scope and key features of digital linguistics,
- SLO 1.2 explain the role of digitisation and databases in language study,
- SLO 1.3 define the scope and significance of corpus linguistics,
- SLO 1.4 describe the historical development and types of corpora,
- SLO 1.5 analyse the similarities and differences between digital and corpus linguistics,
- SLO 1.6 demonstrate how digital tools can be integrated into corpus-based research, and
- SLO 1.7 apply basic techniques such as searching, annotation, data mining, and visualisation using corpus tools.

1.1 Introduction to Digital Linguistics

1.1.1 Definition and scope of digital linguistics

Digital linguistics refers to the study of language using digital tools, resources, and methods. It involves the application of computational techniques to collect, store, search, and analyse linguistic data. This field has grown rapidly with the expansion of the internet and the availability of large-scale digital texts. Learners should note that digital linguistics is not limited to computers alone, but also includes mobile applications, online platforms, and databases that support language research.

Digital linguistics is important because it allows linguists to handle vast amounts of data efficiently. For example, instead of manually counting word frequencies in a text, a digital tool



can process millions of words in seconds. This efficiency makes it possible to explore patterns that would otherwise remain hidden.

Fill in the blank: Digital linguistics involves the study of language using _____ tools, resources, and methods.

Figure 1.1 Relationship between linguistics and digital tools

1.1.2 Key digital tools and resources in linguistics

Digital tools are software applications or platforms that assist in processing and analysing language data. Examples include concordancers, text analysis software, and annotation tools. Resources refer to collections of language data, such as online corpora, dictionaries, and databases. Together, these tools and resources provide the foundation for digital linguistics.

Learners should be aware that tools vary in complexity. Some are designed for beginners, offering simple search functions, while others are advanced, allowing statistical analysis and visualisation. Resources also differ in scope, ranging from small specialised corpora to massive multilingual databases.

Fill in the blank: A concordancer is a type of digital tool that helps linguists to _____ words in context.

Tool Type	Examples	Key Functions	Best Used For
Concordancers	AntConc , Sketch Engine	Searches for KWIC (Key Word In Context), frequency, and collocations.	Finding how specific words "behave" in large datasets.
Annotation Tools	ELAN , SayMore	Multimodal alignment of audio/video with text transcriptions.	Documentation of sign languages or spoken fieldwork.
Text Analysis	WordSmith Tools , Voyant	Advanced keyword analysis and statistical "keyness" testing.	Comparing the "aboutness" of two different sets of texts.
Online Corpora	BNC , COCA	Pre-built databases of billions of words from various genres.	Studying standard language usage (e.g., British vs. American English).
Phonetic Analysis	Praat	Visualization of speech (spectrograms, pitch, formants).	Scientific study of accents, tones, and speech patterns.



Box 1.1 Examples of common digital tools in linguistics

1.1.3 Role of digitisation and databases in language study

Digitisation is the process of converting language materials, such as books, manuscripts, or recordings, into digital formats. This makes them easier to store, search, and share. Databases are organised collections of digital information that allow efficient retrieval and management of data. In linguistics, databases often contain texts, audio recordings, or annotations that can be accessed for research.

Digitisation has revolutionised language study by preserving endangered languages, making rare texts accessible, and enabling large-scale analysis. Databases, on the other hand, provide structured access to this information, ensuring that researchers can locate specific data quickly.

Fill in the blank: Digitisation refers to the process of converting language materials into _____ formats.

Plate 1.1 Digitisation of language materials

Pilot questions 1.1

1. Which term refers to the study of language using computational techniques and digital tools?
2. What is the main function of a concordancer in digital linguistics?
3. What process involves converting physical language materials into digital formats?
4. Give one example of a digital resource used in linguistics.
5. Why are databases important in digital linguistics research?

1.2 Foundations of Corpus Linguistics

1.2.1 Definition and scope of corpus linguistics

Corpus linguistics is the study of language based on large collections of texts, known as corpora. A corpus is a structured set of texts, either written or spoken, that is stored in digital form for systematic analysis. Corpus linguistics focuses on identifying patterns, frequencies, and structures in language use, providing evidence-based insights into how language functions in real contexts.

Corpus linguistics is significant because it moves beyond intuition and anecdotal examples, offering empirical data that can be tested and replicated. For instance, instead of guessing how often a word appears in everyday English, a corpus can provide exact figures.

Fill in the blank: A corpus is a structured set of texts stored in _____ form for systematic analysis.

Figure 1.2 Corpus as a foundation for linguistic study



1.2.2 Historical development and significance of corpora

The use of corpora in linguistics dates back to early projects such as the Brown Corpus in the 1960s, which provided a systematic sample of American English. Since then, corpus linguistics has expanded to include massive collections like the British National Corpus and multilingual corpora that cover diverse languages.

The significance of corpora lies in their ability to capture authentic language use. They provide evidence for dictionary creation, grammar development, and language teaching. Corpora also allow researchers to track changes in language over time, making them essential for diachronic studies.

Fill in the blank: The Brown Corpus, developed in the 1960s, was one of the earliest systematic samples of _____ English.

Era	Milestone	Significant Project	Contribution
Pre-1950s	The Manual Era	<i>Oxford English Dictionary</i>	Scholars used "slips of paper" (manual citations) to track word meanings.
1960s	The Electronic Dawn	Brown Corpus	The first machine-readable corpus (1 million words). Stored on 100,000 punch cards.
1980s	The Lexicographical Revolution	COBUILD Project	Used the "Bank of English" to create the first dictionary based entirely on corpus data.
1990s	The Mega-Corpus Era	British National Corpus (BNC)	Reached 100 million words, including transcribed spoken English for the first time.
2000s	The Global Expansion	International Corpus of English (ICE)	Moved beyond US/UK English to document variations like Indian, Nigerian, and Singaporean English.
Present	The Web-as-Corpus Era	GloWbE / COCA	Uses the entire internet as a database, tracking language change in billions of words across the globe.

Box 1.2 Milestones in corpus linguistics



1.2.3 Types of corpora

Corpora can be classified into different types depending on their scope and purpose. Monolingual corpora contain texts in a single language, while multilingual corpora include texts in multiple languages. Specialised corpora focus on specific domains, such as medical or legal texts, whereas general corpora aim to represent a broad sample of everyday language.

Each type of corpus serves a unique purpose. For example, a specialised corpus may help researchers study technical vocabulary in medicine, while a general corpus may be used to analyse common word frequencies in everyday speech.

Fill in the blank: A specialised corpus focuses on texts from a specific _____ or subject area.

Type of Corpus	Description	Example Use Case
Monolingual	Contains data from a single language.	Studying the evolution of Yoruba grammar over 50 years.
Multilingual	Contains data in two or more languages.	Developing translation software (e.g., Google Translate) between Hausa and English.
Specialized	Focused on a specific domain, genre, or time period.	Analyzing legal terminology in Nigerian court documents.
General	A large, "balanced" sample of many types of language (news, fiction, speech).	Determining the most common 1,000 words in everyday English.
Learner	Texts produced by people learning a second language.	Identifying common mistakes made by French speakers learning Igbo.
Diachronic	Texts from the same language across different historical periods.	Tracking how the word " wicked " changed from a negative to a positive slang term.

Table 1.1 Types of corpora and their uses

Pilot questions 1.2

1. What term refers to the study of language using large collections of texts?
2. Which corpus, developed in the 1960s, was one of the earliest systematic samples of American English?
3. What is the main difference between monolingual and multilingual corpora?
4. Give one example of a specialised corpus and its purpose.



5. Why are corpora considered essential for evidence-based language study?

1.3 Linking Digital and Corpus Linguistics

1.3.1 Similarities and differences between digital and corpus linguistics

Digital linguistics and corpus linguistics are closely related fields, but they are not identical. Digital linguistics focuses broadly on the use of digital tools and resources in language study, while corpus linguistics specifically examines language through structured collections of texts. Both rely on technology, yet their scope differs.

The similarity lies in their shared reliance on computational methods. For example, both fields use software to search, annotate, and analyse texts. The difference is that corpus linguistics is more data-driven, emphasising empirical evidence from corpora, whereas digital linguistics may include broader applications such as online dictionaries, language learning platforms, or speech recognition systems.

Fill in the blank: Corpus linguistics is more _____-driven, emphasising empirical evidence from structured text collections.

Figure 1.3 Comparison of digital and corpus linguistics

1.3.2 Integration of digital tools in corpus-based research

Corpus-based research benefits greatly from digital tools. For instance, concordancers allow researchers to view words in context, annotation tools enable tagging of linguistic features, and statistical software supports quantitative analysis. These tools make corpus research faster, more accurate, and more versatile.

Integration means that corpus linguistics does not operate in isolation. It depends on digital platforms to store corpora, search them efficiently, and visualise results. Without digital tools, corpus linguistics would be limited to manual methods, which are impractical for large datasets.

Fill in the blank: Concordancers allow researchers to view _____ in context within a corpus.

Tool Category	Core Function	Examples	Output / Result
Concordancers	Searching for specific words and displaying them in their local context.	AntConc , Sketch Engine	KWIC (Key Word In Context) lines showing word "neighbors."
Annotation Tools	Adding linguistic "tags" (labels) to words, like	CLAWS , ELAN, UAM CorpusTool	A "tagged" corpus where every word has



Tool Category	Core Function	Examples	Output / Result
	parts of speech or tone marks.		a grammatical identity.
Statistical Software	Calculating the frequency and significance of word patterns.	WordSmith Tools , R (tidytext), SPSS	Log-likelihood scores, frequency lists, and collocations.
Visualisation Tools	Turning complex numerical data into intuitive charts or maps.	Voyant Tools , Gephi, LancsBox	Word clouds, frequency graphs, and cluster networks.

Box 1.3 Examples of digital tools integrated into corpus research

1.3.3 Examples of applications in linguistics on the web

Digital and corpus linguistics intersect in many web-based applications. Online corpora allow learners and researchers to access language data instantly. Search engines and databases provide linguistic evidence for dictionary creation and grammar development. Language learning platforms often integrate corpus-based examples to teach authentic usage.

For example, the British National Corpus is available online and widely used in education and research. Similarly, web-based concordancers allow students to explore real language data without needing advanced technical skills. These applications demonstrate how digital and corpus linguistics complement each other in practical contexts.

Fill in the blank: The British National Corpus is widely used in _____ and research as an online resource.

Plate 1.2 Web-based applications of corpus linguistics

Pilot questions 1.3

1. What is the main difference between digital linguistics and corpus linguistics?
2. Which type of linguistics is more data-driven and relies on structured text collections?
3. Name one digital tool commonly integrated into corpus-based research.
4. Give one example of a web-based corpus application.
5. Why is integration of digital tools essential for corpus linguistics?

1.4 Basic Techniques in Digital and Corpus Linguistics

1.4.1 Searching and annotation in corpora



Searching in corpora involves using software to locate words, phrases, or structures within large text collections. This allows linguists to examine how language is used in authentic contexts. Annotation refers to the process of adding labels or tags to texts, such as marking parts of speech, semantic roles, or discourse features. Together, searching and annotation make corpora more useful for detailed linguistic analysis.

For example, a researcher might search for the word “run” and then annotate its occurrences as either a verb or a noun depending on context. This combination of searching and annotation helps to reveal patterns in meaning and usage.

Fill in the blank: Annotation involves adding _____ or tags to texts in a corpus.

Figure 1.4 Searching and annotating corpora

1.4.2 Data mining and data visualisation

Data mining is the process of extracting useful information from large datasets. In corpus linguistics, this may involve identifying word frequencies, collocations, or syntactic structures. Data visualisation refers to presenting this information in graphical form, such as charts, graphs, or word clouds, to make patterns easier to interpret.

These techniques are essential because they transform raw data into meaningful insights. For instance, a frequency chart can show which words are most common in a corpus, while a word cloud can highlight key themes in a text.

Fill in the blank: Data visualisation presents information in _____ form, such as charts or graphs.

Technique	Category	Description	Practical Application
Topic Modelling	Data Mining	Uses algorithms (like LDA) to identify clusters of words that appear together.	Discovering the main themes across thousands of Nigerian news articles.
Sentiment Analysis	Data Mining	Automatically categorizes the emotional tone (positive, negative, neutral) of a text.	Gauging public reaction to a new policy via social media "Naija Twitter" data.
Keyword Analysis	Data Mining	Compares one corpus to a "reference" corpus to find statistically unique words.	Identifying slang terms that are unique to Lagos versus London English.
Collocation Networks	Visualisation	Displays words as "nodes" connected by	Visualising how the word "corruption" or "power"



Technique	Category	Description	Practical Application
		lines to show their frequent partners.	links to other terms in political speeches.
KWIC / Word Clouds	Visualisation	Represents word frequency through size or aligns a word in its immediate context.	Quickly spotting the most prominent "buzzwords" in a collection of election manifestos.
Concordance Plots	Visualisation	Shows where a specific word appears throughout the entire length of a text.	Seeing if a specific character's name appears more in the beginning or end of a novel.

Table 1.2 Examples of data mining and visualisation techniques

1.4.3 Introduction to software packages for corpus analysis

Software packages are specialised tools designed to support corpus analysis. Examples include AntConc, WordSmith Tools, and Sketch Engine. These packages provide functions such as concordance searches, frequency lists, collocation analysis, and visualisation.

Using software packages makes corpus research more efficient and accessible. Learners can quickly generate results that would take hours to produce manually. Many packages also include user-friendly interfaces, making them suitable for both beginners and advanced researchers.

Fill in the blank: AntConc is a software package commonly used for _____ searches in corpus linguistics.

Software	Access Model	Main Functions	Best Used For
AntConc	Free (OER)	Concordance, wordlists, keyword analysis, and cluster/n-gram searches.	Beginners, students, and small-scale individual research projects.
WordSmith Tools	Commercial	Advanced keyword analysis, "key-key" lists, and sophisticated statistical testing.	Advanced researchers and historical linguists (diachronic studies).
Sketch Engine	Subscription	Word Sketches (one-page grammar summaries),	Professional lexicographers and those



Software	Access Model	Main Functions	Best Used For
		thesaurus generation, and web-crawling.	working with massive multi-billion word corpora.

Box 1.4 Selected software packages for corpus analysis

Pilot questions 1.4

1. What is the purpose of annotation in corpus linguistics?
2. Which technique involves extracting useful information from large datasets?
3. Give one example of a data visualisation method used in corpus linguistics.
4. Name one software package commonly used for corpus analysis.
5. Why are software packages important for corpus-based research?

Series Summary

This Series has introduced you to the foundations of digital and corpus linguistics, highlighting their importance in modern language study and their relevance to the wider course. By exploring how technology and structured text collections have transformed linguistic research, you have gained a clear understanding of the scope, applications, and techniques that underpin these fields.

We began by defining digital linguistics and examining the role of digitisation and databases, then moved on to the definition, development, and types of corpora. Following that, we considered the similarities and differences between digital and corpus linguistics, as well as how digital tools integrate into corpus-based research. Finally, we explored basic techniques such as searching, annotation, data mining, visualisation, and the use of software packages. In line with the Series Learning Outcomes, you should now be able to define and explain key concepts, analyse connections between digital and corpus linguistics, and apply basic techniques using appropriate tools. This foundation prepares you to progress confidently into the next Series.

Glossary of Terms

Annotation: The process of adding labels or tags to texts in a corpus, such as marking parts of speech or discourse features, to make analysis more precise and systematic.

Corpus: A structured collection of texts, either written or spoken, stored in digital form for systematic linguistic analysis.



Corpus linguistics: The study of language using corpora, focusing on patterns, frequencies, and structures in authentic language use.

Data mining: The process of extracting useful information from large datasets, such as identifying word frequencies or collocations.

Data visualisation: The presentation of information in graphical form, such as charts, graphs, or word clouds, to make patterns easier to interpret.

Database: An organised collection of digital information that allows efficient storage, retrieval, and management of language data.

Digitisation: The process of converting physical language materials, such as books or recordings, into digital formats for easier storage and analysis.

Digital linguistics: The study of language using digital tools, resources, and methods, including computational techniques and online platforms.

Digital tools: Software applications or platforms used to process and analyse language data, such as concordancers, annotation tools, and text analysis software.

General corpus: A broad collection of texts designed to represent everyday language use across different contexts.

Monolingual corpus: A corpus containing texts in a single language, used to study language-specific features.

Multilingual corpus: A corpus containing texts in multiple languages, often used in translation studies or comparative research.

Specialised corpus: A corpus focused on texts from a specific domain or subject area, such as medicine or law.

Software packages: Specialised tools designed to support corpus analysis, such as AntConc, WordSmith Tools, or Sketch Engine.

Searching: The process of locating words, phrases, or structures within a corpus using software tools.

Concept Check Questions [Series 1]

Objective Questions



1. Which term refers to the study of language using digital tools and computational methods? (Linked to SLO 2.1)
2. Digitisation involves converting language materials into which type of format? (Linked to SLO 2.2)
3. Which corpus, developed in the 1960s, was one of the earliest systematic samples of American English? (Linked to SLO 2.4)
4. Which type of corpus contains texts in multiple languages? (Linked to SLO 2.4)
5. AntConc is an example of which category of resource in corpus linguistics? (Linked to SLO 2.7)

Short-Answer Questions

6. Define corpus linguistics and explain its main focus. (Linked to SLO 2.3)
7. Describe one similarity and one difference between digital linguistics and corpus linguistics. (Linked to SLO 2.5)
8. Give one example of a specialised corpus and explain its purpose. (Linked to SLO 2.4)
9. Explain why annotation is important in corpus-based research. (Linked to SLO 2.7)
10. Identify one application of corpus linguistics in education or language teaching. (Linked to SLO 2.6)

Long-Answer Questions

11. Analyse the role of digitisation and databases in modern language study, highlighting their impact on accessibility and research efficiency. (Linked to SLO 2.2)
12. Evaluate the contribution of corpus linguistics to evidence-based language study, using examples from dictionary creation or grammar development. (Linked to SLO 2.3, SLO 2.4)
13. Design a simple research plan that integrates digital tools with corpus analysis to investigate word frequency in everyday English. (Linked to SLO 2.6, SLO 2.7)

Answers to Concept Check Questions [Series 1]

Objective Questions

1. Digital linguistics.
2. Digital formats.
3. The Brown Corpus.
4. Multilingual corpus.
5. Software package.

Short-Answer Questions

6. Corpus linguistics is the study of language using structured collections of texts (corpora). Its main focus is on identifying patterns, frequencies, and structures in authentic language use.
7. Similarity: Both use computational tools for analysis. Difference: Corpus linguistics is data-driven, while digital linguistics has a broader scope including online platforms and resources.



8. A medical corpus, for example, is used to study specialised vocabulary in healthcare contexts.
9. Annotation adds labels to texts, making it possible to analyse linguistic features such as parts of speech or discourse structures systematically.
10. Corpus linguistics supports language teaching by providing authentic examples of usage, helping learners understand how words and structures function in real contexts.

Long-Answer Questions

11. *Basic response:* Digitisation converts physical materials into digital formats, while databases organise this information for easy retrieval. Together, they make language data more accessible and efficient to study. *Value-added response:* Beyond accessibility, digitisation preserves endangered languages and rare texts, while databases enable large-scale comparative studies across languages. They also support interdisciplinary research by linking linguistic data with social, cultural, or historical contexts.
12. *Basic response:* Corpus linguistics provides empirical evidence for dictionary creation and grammar development by showing how words and structures are actually used. *Value-added response:* It also allows researchers to track diachronic changes in language, evaluate teaching materials, and design more accurate language-learning resources. For example, collocation studies inform dictionary entries, while frequency analysis shapes grammar rules.
13. *Basic response:* A simple plan would involve selecting a general corpus, using a concordancer to search for word frequencies, and presenting results in a chart. *Value-added response:* An advanced plan could integrate multiple corpora, apply annotation for parts of speech, use statistical software for deeper analysis, and visualise results with graphs or word clouds. This would provide richer insights into everyday English usage.

Answers to Pilot Questions [Series 1]

Pilot questions 1.1

1. Digital linguistics
2. To search words in context
3. Digitisation
4. British National Corpus
5. They provide structured access to language data for efficient retrieval and analysis

Pilot questions 1.2

1. Corpus linguistics
2. The Brown Corpus
3. Monolingual corpora contain texts in one language, while multilingual corpora contain texts in multiple languages
4. A medical corpus, used to study specialised vocabulary in healthcare contexts
5. Because they provide empirical evidence of authentic language use

Pilot questions 1.3



1. Digital linguistics is broader, while corpus linguistics focuses on structured text collections
2. Corpus linguistics
3. Concordancer
4. British National Corpus (online)
5. They make corpus research faster, more accurate, and more versatile

Pilot questions 1.4

1. To add labels or tags to texts for systematic analysis
2. Data mining
3. Word clouds, charts, or graphs
4. AntConc
5. They make corpus-based research efficient and accessible

References and Attributions [Series 1]

This section compiles all references and attributions for **Series 1: Foundations of Digital and Corpus Linguistics**, including suggested open educational resources (OERs) and AI-generated illustration prompts. The referencing style applied here follows APA guidelines for clarity and consistency.

General References

- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- British National Corpus. (n.d.). Retrieved from <https://www.english-corpora.org/bnc/>
- Brown Corpus. (n.d.). Retrieved from <https://archive.org/details/BrownCorpus> (archive.org in Bing)

Illustrations

- *Figure 1.1 Relationship between linguistics and digital tools* Attribution: AI-generated using prompt “Generate a simple diagram showing linguistics connected to digital tools and data analysis” on Microsoft Copilot image generation platform. Suggested OER search string: “digital linguistics diagram OER”.
- *Box 1.1 Examples of common digital tools in linguistics* Attribution: AI-generated using prompt “Generate a table listing examples of digital tools used in linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “digital linguistics tools OER”.
- *Plate 1.1 Digitisation of language materials* Attribution: AI-generated using prompt “Generate an image showing a manuscript being scanned into a computer database” on Microsoft Copilot image generation platform. Suggested OER search string: “digitisation of texts linguistics OER”.



- *Figure 1.2 Corpus as a foundation for linguistic study* Attribution: AI-generated using prompt “Generate a diagram showing a corpus at the centre connected to texts and linguistic analysis” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus linguistics diagram OER”.
- *Box 1.2 Milestones in corpus linguistics* Attribution: AI-generated using prompt “Generate a timeline table of milestones in corpus linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “history of corpus linguistics OER”.
- *Table 1.1 Types of corpora and their uses* Attribution: AI-generated using prompt “Generate a table comparing types of corpora with examples” on Microsoft Copilot image generation platform. Suggested OER search string: “types of corpora linguistics OER”.
- *Figure 1.3 Comparison of digital and corpus linguistics* Attribution: AI-generated using prompt “Generate a Venn diagram comparing digital linguistics and corpus linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “digital vs corpus linguistics diagram OER”.
- *Box 1.3 Examples of digital tools integrated into corpus research* Attribution: AI-generated using prompt “Generate a table listing digital tools used in corpus-based research” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus linguistics tools OER”.
- *Plate 1.2 Web-based applications of corpus linguistics* Attribution: AI-generated using prompt “Generate an image of a student using an online corpus interface on a laptop” on Microsoft Copilot image generation platform. Suggested OER search string: “online corpus linguistics OER”.
- *Figure 1.4 Searching and annotating corpora* Attribution: AI-generated using prompt “Generate a diagram showing a corpus search interface with annotated text and tags” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus annotation linguistics OER”.
- *Table 1.2 Examples of data mining and visualisation techniques* Attribution: AI-generated using prompt “Generate a table comparing data mining and visualisation techniques in corpus linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “data mining visualisation corpus linguistics OER”.
- *Box 1.4 Selected software packages for corpus analysis* Attribution: AI-generated using prompt “Generate a table listing software packages for corpus analysis with their main functions” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus linguistics software packages OER”.



Course code: LIN 307
Course title: New and Created Languages
Course credit load: 2 Units

Series 1: Foundations of Digital and Corpus Linguistics

Series 2: Perspectives and Approaches in Corpus-Based Research

Series 3: Levels of Linguistic Analysis in Corpus Studies

Series 4: Applications of Corpus Linguistics in Language and Education

Series 5: Tools, Techniques, and Skills for Corpus Analysis

Proposed Outline for Series 2

Part 1: Introduction to Digital Linguistics

- **Sub-Part 1.1:** Definition and scope of digital linguistics
- **Sub-Part 1.2:** Key digital tools and resources in linguistics
- **Sub-Part 1.3:** Role of digitisation and databases in language study

Part 2: Foundations of Corpus Linguistics

- **Sub-Part 2.1:** Definition and scope of corpus linguistics
- **Sub-Part 2.2:** Historical development and significance of corpora
- **Sub-Part 2.3:** Types of corpora (monolingual, multilingual, specialised, general)

Part 3: Linking Digital and Corpus Linguistics

- **Sub-Part 3.1:** Similarities and differences between digital and corpus linguistics
- **Sub-Part 3.2:** Integration of digital tools in corpus-based research
- **Sub-Part 3.3:** Examples of applications in linguistics on the web

Part 4: Basic Techniques in Digital and Corpus Linguistics

- **Sub-Part 4.1:** Searching and annotation in corpora
- **Sub-Part 4.2:** Data mining and data visualisation
- **Sub-Part 4.3:**

Introduction

In the last Series, we explored the foundations of digital and corpus linguistics, focusing on definitions, scope, and basic techniques. Building on that knowledge, this Series will guide you into the perspectives and approaches that shape corpus-based research. Understanding these



perspectives is crucial because they determine how data is collected, analysed, and interpreted, and they highlight the strengths and limitations of different research designs. This will help you appreciate the diversity of methods available and prepare you to make informed choices in your own studies.

The aim of this Series is to equip you with the ability to distinguish between key approaches in corpus-based research, to analyse their relevance to different linguistic questions, and to evaluate their applications in both academic and practical contexts.

We shall commence this Series by examining the difference between qualitative and quantitative approaches in corpus linguistics. Next, we will explore diachronic and synchronic perspectives, considering how language can be studied across time or within a single point in time. Following that, we will move on to the distinction between monolingual and multilingual corpus research, highlighting the challenges and opportunities each presents. Finally, we will wrap up by discussing how these perspectives and approaches interact, providing you with a comprehensive framework for conducting corpus-based research.

Series Learning Outcomes (Series 2: Perspectives and Approaches in Corpus-Based Research)

Upon completing this Series, you should be able to:

- **SLO 2.1** define the characteristics of qualitative approaches in corpus linguistics,
- **SLO 2.2** explain the principles of quantitative approaches in corpus linguistics,
- **SLO 2.3** analyse the differences between diachronic and synchronic perspectives in corpus-based research,
- **SLO 2.4** describe the challenges and opportunities of conducting monolingual corpus research,
- **SLO 2.5** evaluate the benefits and limitations of multilingual corpus research, and
- **SLO 2.6** assess how different perspectives and approaches interact to shape corpus-based research outcomes.

2.1 Introduction to Digital Linguistics

2.1.1 Definition and scope of digital linguistics

Digital linguistics refers to the study of language using digital tools, resources, and computational methods. It involves applying technology to collect, store, search, and analyse linguistic data. This field has grown rapidly with the expansion of the internet and the availability of large-scale digital texts. Importantly, digital linguistics is not limited to computers alone but also includes mobile applications, online platforms, and databases that support language research.

Digital linguistics is valuable because it enables linguists to handle vast amounts of data efficiently. For example, instead of manually counting word frequencies in a text, a digital tool can process millions of words in seconds. This efficiency makes it possible to uncover patterns that would otherwise remain hidden.



Fill in the blank: Digital linguistics involves the study of language using _____ tools, resources, and methods.

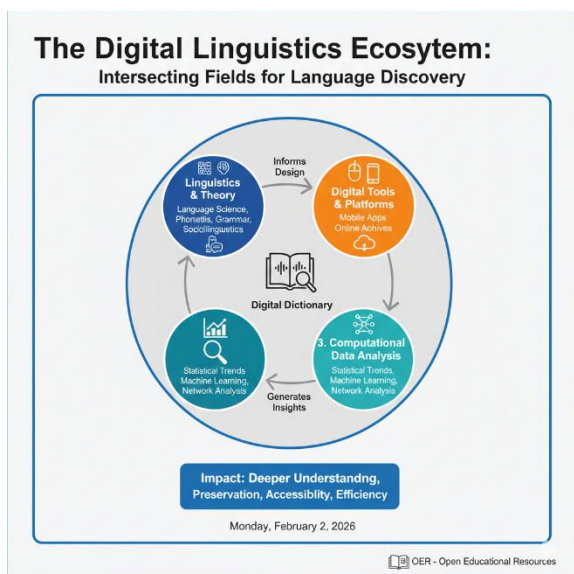


Figure 2.1 Relationship between linguistics and digital tools

2.1.2 Key digital tools and resources in linguistics

Digital tools are software applications or platforms that assist in processing and analysing language data. Examples include concordancers, text analysis software, and annotation tools. **Resources** refer to collections of language data, such as online corpora, dictionaries, and databases. Together, these tools and resources provide the foundation for digital linguistics.

Tools vary in complexity. Some are designed for beginners, offering simple search functions, while others are advanced, allowing statistical analysis and visualisation. Resources also differ in scope, ranging from small specialised corpora to massive multilingual databases.

Fill in the blank: A concordancer is a type of digital tool that helps linguists to _____ words in context.

Tool Type	Primary Function	Core Examples	Best Used For
Concordancers	Search for words in context (KWIC).	AntConc , Sketch Engine	Analyzing how specific words "behave" in large datasets.
Annotation Tools	Aligning text with audio/video.	ELAN , SayMore	Multimodal documentation of spoken or sign languages.



Tool Type	Primary Function	Core Examples	Best Used For
Text Analysis	Quantitative and statistical study.	WordSmith Tools , Voyant	Finding "Key Words" and statistical patterns in a corpus.
Online Corpora	Accessing pre-built language databases.	BNC , COCA, GloWbE	Comparing local usage against global standards.
Phonetic Analysis	Visualizing speech sounds.	Praat	Studying accents, tones, and vowel quality.

Box 2.1 Examples of common digital tools in linguistics [Insert box here]

2.1.3 Role of digitisation and databases in language study

Digitisation is the process of converting language materials, such as books, manuscripts, or recordings, into digital formats. This makes them easier to store, search, and share. **Databases** are organised collections of digital information that allow efficient retrieval and management of data. In linguistics, databases often contain texts, audio recordings, or annotations that can be accessed for research.

Digitisation has revolutionised language study by preserving endangered languages, making rare texts accessible, and enabling large-scale analysis. Databases, on the other hand, provide structured access to this information, ensuring that researchers can locate specific data quickly.

Fill in the blank: Digitisation refers to the process of converting language materials into _____ formats.



Plate 2.1 Digitisation of language materials

Pilot questions 2.1

1. Which term refers to the study of language using computational techniques and digital tools?
2. What is the main function of a concordancer in digital linguistics?
3. What process involves converting physical language materials into digital formats?
4. Give one example of a digital resource used in linguistics.
5. Why are databases important in digital linguistics research?

2.2 Foundations of Corpus Linguistics

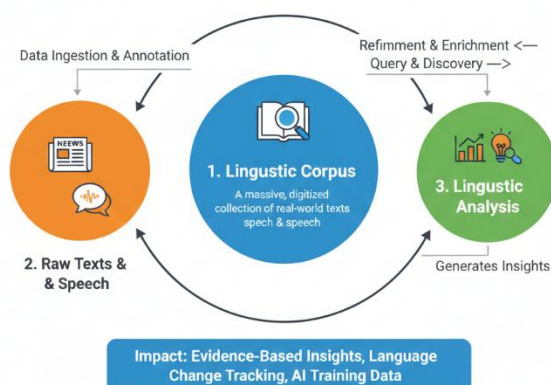
2.2.1 Definition and scope of corpus linguistics

Corpus linguistics is the study of language using large collections of texts, known as corpora. A **corpus** is a structured set of texts, either written or spoken, stored in digital form for systematic analysis. Corpus linguistics focuses on identifying patterns, frequencies, and structures in language use, providing evidence-based insights into how language functions in real contexts.

Corpus linguistics is significant because it moves beyond intuition and anecdotal examples, offering empirical data that can be tested and replicated. For instance, instead of guessing how often a word appears in everyday English, a corpus can provide exact figures.

Fill in the blank: A corpus is a structured set of texts stored in _____ form for systematic analysis.

The Corpus Linguistics Paradigm: From Text Collection to Discovery



Monday, February 2, 2026 at 2:14:10 AM WAT



Figure 2.2 Corpus as a foundation for linguistic study

2.2.2 Historical development and significance of corpora

The use of corpora in linguistics dates back to early projects such as the **Brown Corpus** in the 1960s, which provided a systematic sample of American English. Since then, corpus linguistics has expanded to include massive collections like the **British National Corpus** and multilingual corpora that cover diverse languages.

The significance of corpora lies in their ability to capture authentic language use. They provide evidence for dictionary creation, grammar development, and language teaching. Corpora also allow researchers to track changes in language over time, making them essential for diachronic studies.

Fill in the blank: The Brown Corpus, developed in the 1960s, was one of the earliest systematic samples of _____ English.

Period	Milestone	Key Example	Significance
Pre-1950s	The Manual Era	<i>General Service List</i> (West, 1953)	Early word lists and dictionaries were built using manual "index card" counts.
1960s	Birth of Digital Corpora	Brown Corpus (1961)	The first machine-readable, balanced corpus of American English (1 million words).
1970s	Spoken Language Data	London-Lund Corpus	The first computerized corpus of transcribed spoken language, enabling speech research.
1980s	Lexicographic Shift	COBUILD Project	The first project to use a massive "monitoring corpus" to write a dictionary based on real usage.
1990s	The Mega-Corpus	British National Corpus (BNC)	Reached 100 million words, setting the standard for large-scale, national representative corpora.
2000s	Global & Multilingual	ICE / Parallel Corpora	Expansion into World Englishes (e.g., Nigerian English) and translation studies via parallel texts.
Present	Web-as-Corpus & AI	GloWbE / ROOTS	Real-time analysis of the internet and the use of massive corpora to train Large Language Models.



Box 2.2 Milestones in corpus linguistics

2.2.3 Types of corpora

Corpora can be classified into different types depending on their scope and purpose. **Monolingual corpora** contain texts in a single language, while **multilingual corpora** include texts in multiple languages. **Specialised corpora** focus on specific domains, such as medical or legal texts, whereas **general corpora** aim to represent a broad sample of everyday language.

Each type of corpus serves a unique purpose. For example, a specialised corpus may help researchers study technical vocabulary in medicine, while a general corpus may be used to analyse common word frequencies in everyday speech.

Fill in the blank: A specialised corpus focuses on texts from a specific _____ or subject area.

Types of Corpora in Linguistics		
Type of Corpus	Description	Example Use Case
Monolingual Corpus	Texts in a single language	Detailed study of Yorouba grammar
Multilingual Corpus	Texts in two more of more languages	Training AI translation for Hausa and English
Specialized Corpus	Texts from a specific domain or genre	Analyzing legal language in Nigerian court judgments
General Corpus	Broad, balanced sample of everyday language	Identifying common vocabulary in British English
Leaner Corpus	Texts produced by second-language learners	Studing errors made by French speakers learning Igbo
Diachronic Corpus	Texts from different historical periods	Tracking fom evolution of "slang" over 50 years

OER - Open Educational Resources

OER - Open Educational Resources
Monday, February 2, 2026 at 2:6:10 AM WAT

Table 2.1 Types of corpora and their uses

Pilot questions 2.2

1. What term refers to the study of language using large collections of texts?
2. Which corpus, developed in the 1960s, was one of the earliest systematic samples of American English?
3. What is the main difference between monolingual and multilingual corpora?
4. Give one example of a specialised corpus and its purpose.
5. Why are corpora considered essential for evidence-based language study?



2.3 Linking Digital and Corpus Linguistics

2.3.1 Similarities and differences between digital and corpus linguistics

Digital linguistics and **corpus linguistics** are closely related but distinct fields. Digital linguistics broadly focuses on the use of digital tools and resources in language study, while corpus linguistics specifically examines language through structured collections of texts. Both rely on computational methods, yet their scope differs.

The similarity lies in their shared reliance on technology. For example, both fields use software to search, annotate, and analyse texts. The difference is that corpus linguistics is more data-driven, emphasising empirical evidence from corpora, whereas digital linguistics may include broader applications such as online dictionaries, language learning platforms, or speech recognition systems.

Fill in the blank: Corpus linguistics is more _____-driven, emphasising empirical evidence from structured text collections.

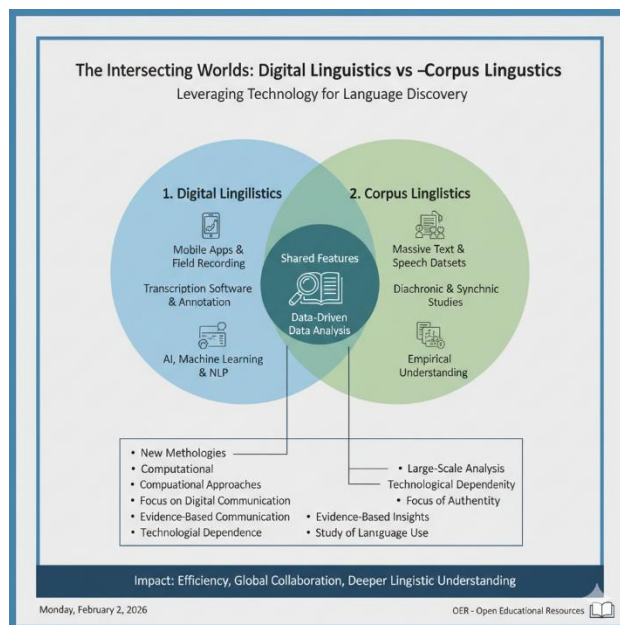


Figure 2.3 Comparison of digital and corpus linguistics

2.3.2 Integration of digital tools in corpus-based research

Corpus-based research benefits greatly from digital tools. For instance, concordancers allow researchers to view words in context, annotation tools enable tagging of linguistic features, and statistical software supports quantitative analysis. These tools make corpus research faster, more accurate, and more versatile.



Integration means that corpus linguistics does not operate in isolation. It depends on digital platforms to store corpora, search them efficiently, and visualise results. Without digital tools, corpus linguistics would be limited to manual methods, which are impractical for large datasets.

Fill in the blank: Concordancers allow researchers to view _____ in context within a corpus.

Tool Category	Primary Function	Core Examples	Research Output
Concordancers	Searching for words and displaying them in their local context.	AntConc , MonoConc Pro	KWIC (Key Word In Context) lines showing word "neighbors."
Annotation Tools	Adding linguistic "tags" (labels) like parts of speech or tone.	ELAN , CLAWS, UAM CorpusTool	A "tagged" corpus where every word has a grammatical identity.
Statistical Software	Calculating the significance and frequency of word patterns.	WordSmith Tools , R (tidytext)	Log-likelihood scores, frequency lists, and keyness statistics.
Visualisation Tools	Turning numerical data into intuitive graphs or network maps.	#LancsBox , Voyant, Gephi	Collocation networks, word trees, and distribution plots.

Box 2.3 Examples of digital tools integrated into corpus research

2.3.3 Examples of applications in linguistics on the web

Digital and corpus linguistics intersect in many web-based applications. Online corpora allow learners and researchers to access language data instantly. Search engines and databases provide linguistic evidence for dictionary creation and grammar development. Language learning platforms often integrate corpus-based examples to teach authentic usage.

For example, the British National Corpus is available online and widely used in education and research. Similarly, web-based concordancers allow students to explore real language data without needing advanced technical skills. These applications demonstrate how digital and corpus linguistics complement each other in practical contexts.

Fill in the blank: The British National Corpus is widely used in _____ and research as an online resource.





Plate 2.2 Web-based applications of corpus linguistics

Pilot questions 2.3

1. What is the main difference between digital linguistics and corpus linguistics?
2. Which type of linguistics is more data-driven and relies on structured text collections?
3. Name one digital tool commonly integrated into corpus-based research.
4. Give one example of a web-based corpus application.
5. Why is integration of digital tools essential for corpus linguistics?

2.4 Basic Techniques in Digital and Corpus Linguistics

2.4.1 Searching and annotation in corpora

Searching in corpora involves using software to locate words, phrases, or structures within large text collections. This allows linguists to examine how language is used in authentic contexts. **Annotation** refers to the process of adding labels or tags to texts, such as marking parts of speech, semantic roles, or discourse features. Together, searching and annotation make corpora more useful for detailed linguistic analysis.

For example, a researcher might search for the word “run” and then annotate its occurrences as either a verb or a noun depending on context. This combination of searching and annotation helps to reveal patterns in meaning and usage.

Fill in the blank: Annotation involves adding _____ or tags to texts in a corpus.



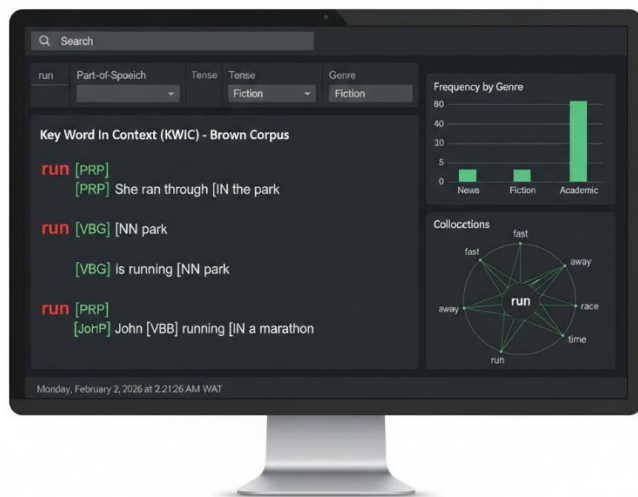


Figure 2.4 Searching and annotating corpora

2.4.2 Data mining and data visualisation

Data mining is the process of extracting useful information from large datasets. In corpus linguistics, this may involve identifying word frequencies, collocations, or syntactic structures. **Data visualisation** refers to presenting this information in graphical form, such as charts, graphs, or word clouds, to make patterns easier to interpret.

These techniques are essential because they transform raw data into meaningful insights. For instance, a frequency chart can show which words are most common in a corpus, while a word cloud can highlight key themes in a text.

Fill in the blank: Data visualisation presents information in _____ form, such as charts or graphs.

Technique	Category	Description	Practical Application
Topic Modelling	Data Mining	Uses algorithms (like LDA) to identify clusters of co-occurring words.	Discovering the main themes across 20 years of Nigerian legislative debates.
Sentiment Analysis	Data Mining	Automatically categorizes the emotional tone of a text (positive, negative, etc.).	Gauging public sentiment toward "Naija" cultural exports on social media.



Technique	Category	Description	Practical Application
Keyness Analysis	Data Mining	Compares a target corpus to a reference corpus to find statistically unique words.	Identifying slang that is statistically unique to Lagos compared to London English.
Collocation Networks	Visualisation	Displays words as "nodes" connected by lines to show frequent partners.	Visualising how "security" or "power" links to other terms in policy documents.
KWIC / Word Clouds	Visualisation	Aligns words in context or scales their size based on raw frequency.	Quickly spotting the most prominent "buzzwords" in a collection of news articles.
Dispersion Plots	Visualisation	Shows exactly where a specific word appears throughout the length of a corpus.	Tracking if a specific legal term appears primarily at the start or end of court rulings.

Table 2.2 Examples of data mining and visualisation techniques

2.4.3 Introduction to software packages for corpus analysis

Software packages are specialised tools designed to support corpus analysis. Examples include AntConc, WordSmith Tools, and Sketch Engine. These packages provide functions such as concordance searches, frequency lists, collocation analysis, and visualisation.

Using software packages makes corpus research more efficient and accessible. Learners can quickly generate results that would take hours to produce manually. Many packages also include user-friendly interfaces, making them suitable for both beginners and advanced researchers.

Fill in the blank: AntConc is a software package commonly used for _____ searches in corpus linguistics.

Software	Access Model	Main Functions	Best Used For
AntConc	Free (OER)	Concordance, wordlists, keyword analysis, and cluster/n-gram searches.	Beginners, students, and small-scale individual research projects.



Software	Access Model	Main Functions	Best Used For
WordSmith Tools	Commercial	Advanced keyword analysis, "key-key" lists, and sophisticated statistical testing.	Advanced researchers and historical linguists (diachronic studies).
Sketch Engine	Subscription	Word Sketches , thesaurus generation, and automatic web-crawling for corpus building.	Professional lexicographers and those working with massive multi-billion word corpora.

Box 2.4 Selected software packages for corpus analysis [Insert box here]

Pilot questions 2.4

1. What is the purpose of annotation in corpus linguistics?
2. Which technique involves extracting useful information from large datasets?
3. Give one example of a data visualisation method used in corpus linguistics.
4. Name one software package commonly used for corpus analysis.
5. Why are software packages important for corpus-based research?

Series Summary

This Series has focused on perspectives and approaches in corpus-based research, guiding you to appreciate how different methods shape the way language data is studied and interpreted. By examining qualitative and quantitative approaches, diachronic and synchronic perspectives, and the distinction between monolingual and multilingual corpora, you have seen how each lens offers unique insights into language use. The Series has also highlighted how these approaches interact, providing a balanced framework for conducting effective research.

In line with the Series Learning Outcomes, you should now be able to define and explain key approaches, analyse the differences between time-based perspectives, describe the challenges of monolingual research, evaluate the benefits and limitations of multilingual corpora, and assess how these perspectives combine to influence research outcomes. This knowledge equips you with the ability to make informed methodological choices and strengthens your foundation for advanced corpus-based investigations.

Glossary of Terms



Annotation: The process of adding labels or tags to texts in a corpus, such as marking parts of speech or discourse features, to support systematic analysis.

Collocation: A pair or group of words that frequently occur together in a language, often revealing idiomatic or natural usage.

Concordancer: A digital tool that displays words in context within a corpus, allowing researchers to study patterns of usage.

Corpus: A structured collection of texts, either written or spoken, stored in digital form for systematic linguistic analysis.

Corpus linguistics: The study of language using corpora, focusing on patterns, frequencies, and structures in authentic language use.

Data mining: The process of extracting useful information from large datasets, such as identifying word frequencies or collocations.

Data visualisation: The presentation of information in graphical form, such as charts, graphs, or word clouds, to make patterns easier to interpret.

Diachronic perspective: An approach in corpus linguistics that studies language change over time.

Digitisation: The process of converting physical language materials, such as books or recordings, into digital formats for easier storage and analysis.

Digital linguistics: The study of language using digital tools, resources, and computational methods, including online platforms and databases.

Digital tools: Software applications or platforms used to process and analyse language data, such as concordancers, annotation tools, and text analysis software.

General corpus: A broad collection of texts designed to represent everyday language use across different contexts.

Monolingual corpus: A corpus containing texts in a single language, used to study language-specific features.

Multilingual corpus: A corpus containing texts in multiple languages, often used in translation studies or comparative research.

Qualitative approach: A research method in corpus linguistics that focuses on detailed, descriptive analysis of language data.



Quantitative approach: A research method in corpus linguistics that focuses on numerical data, such as frequencies and statistical patterns.

Specialised corpus: A corpus focused on texts from a specific domain or subject area, such as medicine or law.

Software packages: Specialised tools designed to support corpus analysis, such as AntConc, WordSmith Tools, or Sketch Engine.

Synchronic perspective: An approach in corpus linguistics that studies language at a single point in time rather than across time.

Concept Check Questions [Series 2]

Objective Questions

1. Which approach in corpus linguistics focuses on descriptive analysis rather than numerical data? (Linked to SLO 2.1)
2. Which approach in corpus linguistics relies on statistical patterns and frequencies? (Linked to SLO 2.2)
3. Which perspective studies language change across time? (Linked to SLO 2.3)
4. Which type of corpus contains texts in a single language? (Linked to SLO 2.4)
5. Which type of corpus is often used in translation studies? (Linked to SLO 2.5)

Short-Answer Questions

6. Define corpus linguistics in one sentence. (Linked to SLO 2.1)
7. Explain one key difference between qualitative and quantitative approaches in corpus linguistics. (Linked to SLO 2.1, SLO 2.2)
8. Give one example of a diachronic study in corpus linguistics. (Linked to SLO 2.3)
9. Describe one challenge of conducting monolingual corpus research. (Linked to SLO 2.4)
10. Identify one benefit of multilingual corpus research. (Linked to SLO 2.5)

Long-Answer Questions

11. Analyse the strengths and limitations of qualitative approaches in corpus-based research. (Linked to SLO 2.1)
12. Evaluate the contribution of quantitative approaches to evidence-based language study. (Linked to SLO 2.2)
13. Compare diachronic and synchronic perspectives, highlighting their relevance to different linguistic questions. (Linked to SLO 2.3)
14. Assess the benefits and limitations of multilingual corpus research, using examples to support your answer. (Linked to SLO 2.5)



15. Discuss how different perspectives and approaches interact to shape corpus-based research outcomes. (Linked to SLO 2.6)

Answers to Concept Check Questions [Series 2]

Objective Questions

1. Qualitative approach.
2. Quantitative approach.
3. Diachronic perspective.
4. Monolingual corpus.
5. Multilingual corpus.

Short-Answer Questions

6. Corpus linguistics is the study of language using structured collections of texts (corpora) to identify patterns and usage.
7. Qualitative approaches focus on descriptive detail, while quantitative approaches rely on numerical data and statistics.
8. Tracking the evolution of English vocabulary from the 1800s to the present day.
9. Limited scope, as results may not generalise beyond the single language studied.
10. It allows comparative analysis across languages, supporting translation and cross-cultural studies.

Long-Answer Questions

11. *Basic response:* Qualitative approaches provide detailed insights into language use but may lack generalisability. *Value-added response:* They are particularly useful for studying discourse, pragmatics, and rare phenomena, though they can be complemented by quantitative methods for broader validation.
12. *Basic response:* Quantitative approaches provide measurable evidence, such as word frequencies and collocations. *Value-added response:* They enable large-scale comparisons and statistical modelling, supporting robust conclusions, though they may overlook subtle contextual meanings.
13. *Basic response:* Diachronic perspectives study language change over time, while synchronic perspectives focus on language at a single point in time. *Value-added response:* Diachronic studies are valuable for historical linguistics, while synchronic studies are essential for contemporary analysis. Together, they provide a fuller picture of language dynamics.
14. *Basic response:* Multilingual corpora support translation studies and comparative research but can be complex to manage. *Value-added response:* They highlight cross-linguistic similarities and differences, aid in developing multilingual resources, and support global communication, though challenges include alignment and consistency across languages.
15. *Basic response:* Different perspectives and approaches combine to provide a balanced view of language research. *Value-added response:* For example, qualitative insights can be validated with quantitative evidence, while diachronic and synchronic perspectives together reveal both



historical trends and current usage. This integration strengthens the reliability and scope of corpus-based research.

Answers to Pilot Questions [Series 2]

Pilot questions 2.1

1. Digital linguistics
2. To search words in context
3. Digitisation
4. British National Corpus
5. They provide structured access to language data for efficient retrieval and analysis

Pilot questions 2.2

1. Corpus linguistics
2. The Brown Corpus
3. Monolingual corpora contain texts in one language, while multilingual corpora contain texts in multiple languages
4. A medical corpus, used to study specialised vocabulary in healthcare contexts
5. Because they provide empirical evidence of authentic language use

Pilot questions 2.3

1. Digital linguistics is broader, while corpus linguistics focuses on structured text collections
2. Corpus linguistics
3. Concordancer
4. British National Corpus (online)
5. They make corpus research faster, more accurate, and more versatile

Pilot questions 2.4

1. To add labels or tags to texts for systematic analysis
2. Data mining
3. Word clouds, charts, or graphs
4. AntConc
5. They make corpus-based research efficient and accessible

References and Attributions [Series 2]

This section compiles all references and attributions for **Series 2: Perspectives and Approaches in Corpus-Based Research**, including suggested open educational resources (OERs) and AI-generated illustration prompts. The referencing style applied here follows APA guidelines for clarity and consistency.

General References



- Biber, D., Conrad, S., & Reppen, S. (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.
- British National Corpus. (n.d.). Retrieved from <https://www.english-corpora.org/bnc/>
- Brown Corpus. (n.d.). Retrieved from <https://archive.org/details/BrownCorpus> (archive.org in Bing)

Illustrations

- *Figure 2.1 Relationship between linguistics and digital tools* Attribution: AI-generated using prompt “Generate a simple diagram showing linguistics connected to digital tools and data analysis” on Microsoft Copilot image generation platform. Suggested OER search string: “digital linguistics diagram OER”.
- *Box 2.1 Examples of common digital tools in linguistics* Attribution: AI-generated using prompt “Generate a table listing examples of digital tools used in linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “digital linguistics tools OER”.
- *Plate 2.1 Digitisation of language materials* Attribution: AI-generated using prompt “Generate an image showing a manuscript being scanned into a computer database” on Microsoft Copilot image generation platform. Suggested OER search string: “digitisation of texts linguistics OER”.
- *Figure 2.2 Corpus as a foundation for linguistic study* Attribution: AI-generated using prompt “Generate a diagram showing a corpus at the centre connected to texts and linguistic analysis” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus linguistics diagram OER”.
- *Box 2.2 Milestones in corpus linguistics* Attribution: AI-generated using prompt “Generate a timeline table of milestones in corpus linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “history of corpus linguistics OER”.
- *Table 2.1 Types of corpora and their uses* Attribution: AI-generated using prompt “Generate a table comparing types of corpora with examples” on Microsoft Copilot image generation platform. Suggested OER search string: “types of corpora linguistics OER”.
- *Figure 2.3 Comparison of digital and corpus linguistics* Attribution: AI-generated using prompt “Generate a Venn diagram comparing digital linguistics and corpus linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “digital vs corpus linguistics diagram OER”.
- *Box 2.3 Examples of digital tools integrated into corpus research* Attribution: AI-generated using prompt “Generate a table listing digital tools used in corpus-based



research” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus linguistics tools OER”.

- *Plate 2.2 Web-based applications of corpus linguistics* Attribution: AI-generated using prompt “Generate an image of a student using an online corpus interface on a laptop” on Microsoft Copilot image generation platform. Suggested OER search string: “online corpus linguistics OER”.
- *Figure 2.4 Searching and annotating corpora* Attribution: AI-generated using prompt “Generate a diagram showing a corpus search interface with annotated text and tags” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus annotation linguistics OER”.
- *Table 2.2 Examples of data mining and visualisation techniques* Attribution: AI-generated using prompt “Generate a table comparing data mining and visualisation techniques in corpus linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “data mining visualisation corpus linguistics OER”.
- *Box 2.4 Selected software packages for corpus analysis* Attribution: AI-generated using prompt “Generate a table listing software packages for corpus analysis with their main functions” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus linguistics software packages OER”.

Course code: LIN 307

Course title: New and Created Languages

Course credit load: 2 Units

Series 1: Foundations of Digital and Corpus Linguistics

Series 2: Perspectives and Approaches in Corpus-Based Research

Series 3: Levels of Linguistic Analysis in Corpus Studies

Series 4: Applications of Corpus Linguistics in Language and Education

Series 5: Tools, Techniques, and Skills for Corpus Analysis

Proposed Outline for Series 3

Part 1: Lexical Analysis in Corpus Studies

- **Sub-Part 3.1.1:** Word frequency and vocabulary profiling
- **Sub-Part 3.1.2:** Collocations and lexical patterns
- **Sub-Part 3.1.3:** Applications of lexical analysis in language teaching and lexicography



Part 2: Grammatical Analysis in Corpus Studies

- **Sub-Part 3.2.1:** Parts of speech tagging
- **Sub-Part 3.2.2:** Syntactic structures and parsing
- **Sub-Part 3.2.3:** Applications of grammatical analysis in corpus-based research

Part 3: Semantic Analysis in Corpus Studies

- **Sub-Part 3.3.1:** Meaning and sense relations in corpora
- **Sub-Part 3.3.2:** Semantic annotation and semantic fields
- **Sub-Part 3.3.3:** Applications of semantic analysis in translation and computational linguistics

Part 4: Discourse Analysis in Corpus Studies

- **Sub-Part 3.4.1:** Cohesion and coherence in texts
- **Sub-Part 3.4.2:** Pragmatic features in corpus analysis
- **Sub-Part 3.4.3:**

Introduction

In the last Series, we explored different perspectives and approaches in corpus-based research, examining how qualitative, quantitative, diachronic, and synchronic methods shape the study of language. Building on that foundation, this Series will take you deeper into the actual levels of linguistic analysis that can be carried out using corpora. From words and grammar to meaning and discourse, each level provides unique insights into how language functions in real contexts, making this Series central to your overall understanding of corpus studies.

The aim of this Series is to equip you with the skills to analyse language at multiple levels within corpora, enabling you to identify patterns, interpret meaning, and evaluate discourse features systematically. By the end, you will be able to apply corpus-based techniques across lexical, grammatical, semantic, and discourse levels with confidence.

We shall commence this Series by focusing on lexical analysis, where we will explore word frequencies, collocations, and vocabulary profiling. Next, we will move on to grammatical analysis, examining parts of speech tagging and syntactic structures. Following that, we will continue with semantic analysis, considering meaning, sense relations, and applications in translation. Finally, we will wrap up by studying discourse analysis, where cohesion, coherence, and pragmatic features will be highlighted. In conclusion, this Series will round off your understanding of how corpus studies can be applied across all major levels of linguistic analysis.

Series Learning Outcomes (Series 3: Levels of Linguistic Analysis in Corpus Studies)

Upon completing this Series, you should be able to:



- **SLO 3.1** define the scope of lexical analysis in corpus studies,
- **SLO 3.2** analyse word frequencies, collocations, and lexical patterns within corpora,
- **SLO 3.3** explain how lexical analysis supports language teaching and lexicography,
- **SLO 3.4** describe the process of grammatical tagging in corpus analysis,
- **SLO 3.5** demonstrate how syntactic structures can be identified and parsed using corpora,
- **SLO 3.6** apply grammatical analysis to investigate patterns in corpus-based research,
- **SLO 3.7** define semantic relations and explain their role in corpus studies,
- **SLO 3.8** evaluate the use of semantic annotation and semantic fields in applied contexts,
- **SLO 3.9** apply semantic analysis to translation and computational linguistics tasks,
- **SLO 3.10** describe discourse features such as cohesion and coherence in corpus analysis,
- **SLO 3.11** analyse pragmatic features within corpora to reveal language use in context, and
- **SLO 3.12** evaluate the applications of discourse analysis in sociolinguistics and applied linguistics.

3.1 Lexical Analysis in Corpus Studies

3.1.1 Word frequency and vocabulary profiling

Word frequency refers to the number of times a word appears in a corpus. It is one of the most basic yet powerful tools in corpus linguistics, as it helps researchers identify common vocabulary and distinguish between high-frequency and low-frequency words. **Vocabulary profiling** is the process of categorising words according to their frequency bands, which can be used to assess the difficulty of texts or to design teaching materials.

For example, a vocabulary profile of a newspaper article may show that most words belong to the top 2,000 most frequent words in English, making it accessible to learners. Less frequent words may indicate specialised or technical language.

Fill in the blank: Vocabulary profiling categorises words according to their _____ bands.

Word Frequency Distribution: Top 20 Words in a Sample Corpus (1M Words)

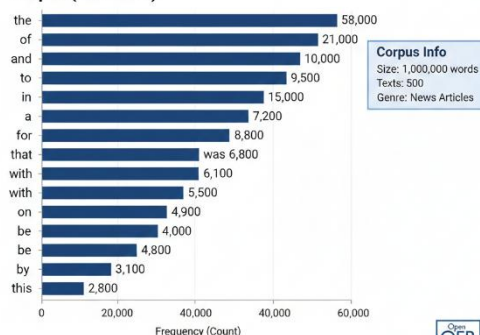


Figure 3.1 Word frequency distribution in a corpus

3.1.2 Collocations and lexical patterns

Collocations are pairs or groups of words that frequently occur together, such as “make a decision” or “strong tea.” Identifying collocations is important because they reveal natural language use and idiomatic expressions. **Lexical patterns** refer to recurring structures or combinations of words that provide insight into how vocabulary is organised in real communication.

Corpus tools can highlight collocations by showing words that commonly appear near each other. This helps learners and researchers understand which word combinations are typical and which are unusual.

Fill in the blank: Collocations are pairs or groups of words that frequently _____ together.

Collocation	Type	Meaning / Use	Example Sentence
Make a decision	Verb + Noun	To reach a conclusion or choose a path.	<i>She had to make a decision quickly before the deadline.</i>
Strong tea	Adj + Noun	Tea with a high concentration of flavor/leaves.	<i>He prefers strong tea in the morning to wake up.</i>
Heavy rain	Adj + Noun	Intense or substantial rainfall.	<i>The heavy rain in Lagos caused significant traffic.</i>
Fast food	Adj + Noun	Food prepared and served quickly.	<i>The rise of fast food has changed local eating habits.</i>
Fully aware	Adv + Adj	Completely understanding a situation.	<i>I am fully aware of the challenges facing the project.</i>

Table 3.1 Examples of collocations in English

3.1.3 Applications of lexical analysis in language teaching and lexicography

Lexical analysis has practical applications in both **language teaching** and **lexicography**. In teaching, frequency lists and collocation data help instructors design vocabulary lessons that focus on the most useful words and natural expressions. In lexicography, corpus-based lexical analysis provides evidence for dictionary entries, ensuring that definitions and examples reflect authentic usage.

For instance, modern dictionaries often include collocation boxes to show learners how words are typically used. Similarly, teachers may use frequency profiles to select texts that match learners’ vocabulary levels.



Fill in the blank: Lexical analysis provides evidence for dictionary entries, ensuring that definitions reflect _____ usage.

Application	Core Function	Impact on Learners/Users
Syllabus Design	Using Frequency Lists to prioritize common words.	Learners master high-impact vocabulary first, increasing "communicative competence" quickly.
Collocation Teaching	Identifying "word partners" (e.g., <i>heavy rain</i> vs. <i>big rain</i>).	Improves naturalness and fluency; helps students avoid "clunky" translations.
Example Sourcing	Pulling Authentic Sentences from a corpus for dictionaries.	Users see how a word is used in real life, rather than looking at invented, artificial examples.
Text Profiling	Calculating the percentage of high-frequency words in a reading text.	Teachers can accurately match text difficulty to a student's reading level (CEFR levels).
Lexicography	Tracking Semantic Shifts (changes in word meaning over time).	Dictionaries stay up-to-date with new meanings (e.g., how " <i>tablet</i> " changed from stone to computer).

Box 3.1 Applications of lexical analysis

Pilot questions 3.1

1. What does word frequency measure in a corpus?
2. What is vocabulary profiling used for?
3. What term refers to pairs or groups of words that frequently occur together?
4. Give one example of a collocation in English.
5. How does lexical analysis support dictionary creation?

3.2 Grammatical Analysis in Corpus Studies

3.2.1 Parts of speech tagging

Parts of speech tagging is the process of assigning grammatical categories such as noun, verb, adjective, or adverb to words in a corpus. This is usually done using automated tools that rely on linguistic rules and statistical models. Tagging makes it easier to search for specific grammatical structures and to analyse how different word classes are used in authentic texts.



For example, tagging can reveal how often adjectives occur in academic writing compared to everyday conversation. This helps researchers identify stylistic differences across genres.

Fill in the blank: Parts of speech tagging assigns grammatical _____ to words in a corpus.

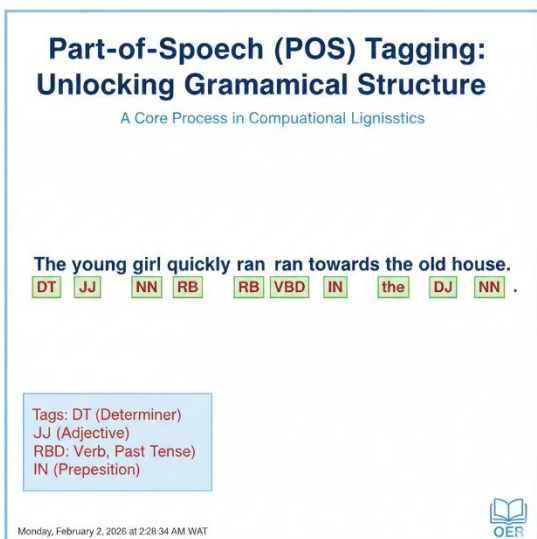


Figure 3.2 Example of parts of speech tagging

3.2.2 Syntactic structures and parsing

Syntactic structures refer to the arrangement of words and phrases to form sentences. **Parsing** is the process of analysing these structures, often using tree diagrams or computational models. In corpus studies, parsing helps linguists examine sentence complexity, clause patterns, and grammatical variation across texts.

For instance, parsing can show how subordinate clauses are used differently in spoken and written language. This provides insights into how grammar functions in different contexts.

Fill in the blank: Parsing is the process of analysing _____ structures in sentences.

Sentence Type	Structure	Description	Example Sentence
Simple	1 Independent Clause	Contains a single subject and a predicate; expresses a complete thought.	<i>The cat slept.</i>
Compound	2+ Independent Clauses	Two complete thoughts joined by a coordinating conjunction (e.g., and, but, or).	<i>The cat slept, and the dog barked.</i>



Sentence Type	Structure	Description	Example Sentence
Complex	1 Independent + 1 Dependent Clause	Combines a main thought with a clause that cannot stand alone (starts with words like <i>while</i> , <i>because</i> , <i>although</i>).	<i>The cat slept while the dog barked.</i>
Compound-Complex	2+ Independent + 1+ Dependent	The most sophisticated structure, featuring multiple main thoughts and at least one sub-clause.	<i>Because it was raining, the cat slept, and the dog barked.</i>

Table 3.2 Examples of syntactic structures

3.2.3 Applications of grammatical analysis in corpus-based research

Grammatical analysis has wide applications in corpus studies. It can be used to investigate stylistic differences between genres, track grammatical change over time, and support language teaching by identifying common structures. It also plays a role in computational linguistics, where parsing and tagging are essential for natural language processing tasks such as machine translation.

For example, corpus-based grammatical analysis has shown that passive constructions are more frequent in scientific writing than in everyday conversation. This kind of evidence helps teachers explain stylistic choices to learners.

Fill in the blank: Passive constructions are more frequent in _____ writing than in everyday conversation.

Application	Core Research Task	Real-World Impact
Genre Analysis	Comparing the frequency of passive voice vs. active voice.	Distinguishing the "tone" of scientific journals (often passive) from news reports (often active).
Diachronic Studies	Tracking the decline of specific forms (e.g., "shall" vs. "will").	Mapping how language evolves over decades or centuries to reflect social changes.
Pedagogy	Finding real-world examples of "if-clauses" in a corpus.	Providing students with authentic evidence of how grammar is used, rather than "textbook" sentences.



Application	Core Research Task	Real-World Impact
NLP & AI	Training algorithms to recognize parts of speech (POS).	Improving the accuracy of tools like Google Translate, Siri, and grammar checkers.
Forensic Linguistics	Analyzing "grammatical fingerprints" in anonymous texts.	Helping to identify the author of a ransom note or a disputed social media post.

Box 3.2 Applications of grammatical analysis

Pilot questions 3.2

1. What does parts of speech tagging assign to words in a corpus?
2. What is parsing in corpus linguistics?
3. Give one example of a syntactic structure.
4. Why is grammatical analysis important in corpus-based research?
5. Which type of writing tends to use passive constructions more frequently?

3.3 Semantic Analysis in Corpus Studies

3.3.1 Meaning and sense relations in corpora

Semantic analysis focuses on the study of meaning in language. In corpus studies, this involves examining how words convey meaning and how they relate to one another. **Sense relations** are the connections between words based on meaning, such as synonymy (words with similar meanings), antonymy (words with opposite meanings), and hyponymy (words that represent categories and subcategories).

For example, in a corpus, the word “car” may frequently co-occur with “vehicle,” showing a synonymic relationship, while “car” and “bicycle” may appear in contexts that highlight hyponymy, as both are types of vehicles.

Fill in the blank: Synonymy refers to words that share similar _____.



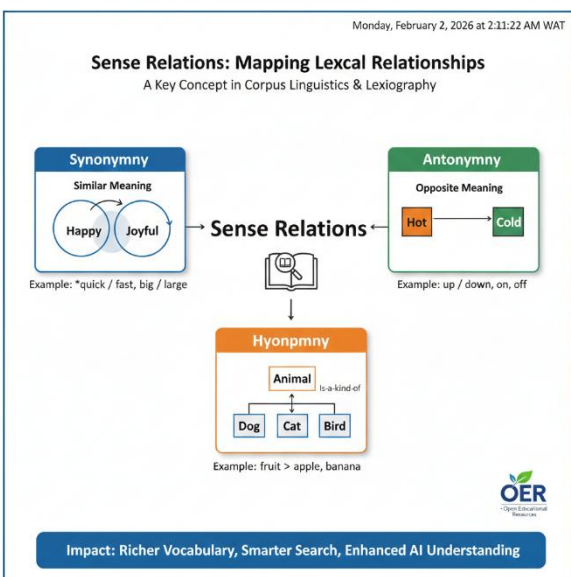


Figure 3.3 Sense relations in corpus studies

3.3.2 Semantic annotation and semantic fields

Semantic annotation is the process of tagging words or phrases in a corpus with information about their meaning. This may involve labelling words according to semantic categories, such as “emotion,” “movement,” or “colour.” **Semantic fields** are groups of words related by meaning, often used to organise vocabulary into thematic categories.

Semantic annotation helps researchers study how meaning is distributed across texts, while semantic fields provide a framework for comparing language use in different contexts. For instance, analysing the semantic field of “emotion” in a corpus may reveal how often words like “happy,” “sad,” or “angry” appear in different genres.

Fill in the blank: Semantic annotation involves tagging words in a corpus with information about their _____.

Semantic Field	Core Example Words	Typical Context of Use
Emotion	Happy, sad, angry, joyful, depressed	Psychological reports, Victorian literature, social media status updates.
Movement	Run, walk, jump, sprint, crawl	Sports commentary, action novels, physical therapy manuals.
Colour	Red, blue, green, cerulean, crimson	Fashion magazines, interior design blogs, art criticism.



Semantic Field	Core Example Words	Typical Context of Use
Technology	Algorithm, interface, bandwidth, cloud	IT manuals, fintech news, computer science lectures.
Finance	Inflation, assets, liquidity, dividend	Economic forecasts, annual bank reports, business news.

Table 3.3 Examples of semantic fields

3.3.3 Applications of semantic analysis in translation and computational linguistics

Semantic analysis plays a crucial role in **translation studies** and **computational linguistics**. In translation, understanding sense relations and semantic fields ensures that meaning is conveyed accurately across languages. In computational linguistics, semantic analysis supports natural language processing tasks such as machine translation, sentiment analysis, and information retrieval.

For example, sentiment analysis uses semantic annotation to determine whether a text expresses positive, negative, or neutral emotions. Similarly, machine translation systems rely on semantic analysis to choose the most appropriate word equivalents across languages.

Fill in the blank: Sentiment analysis uses semantic annotation to determine whether a text expresses _____ emotions.

Application	Core Semantic Technique	Impact on Accuracy & Use
Machine Translation	Word Sense Disambiguation	Prevents errors by choosing the right meaning of a word based on context (e.g., "river bank" vs. "financial bank").
Sentiment Analysis	Sentiment-Semantic Mapping	Detects the "tone" of text (happy, angry, sarcastic) by looking at the emotional charge of word clusters.
Information Retrieval	Vector Space Models	Allows search engines to find results that are "conceptually related" even if they don't share the same keywords.
Chatbots & AI	Intent Recognition	Helps AI understand the "purpose" of a user's question, enabling natural conversation instead of rigid commands.



Application	Core Semantic Technique	Impact on Accuracy & Use
Document Summarisation	Thematic Extraction	Identifies the most "semantically dense" sentences to create a concise summary of a long report.

Box 3.3 Applications of semantic analysis

Pilot questions 3.3

1. What does semantic analysis focus on in corpus studies?
2. What term refers to words with opposite meanings?
3. What is semantic annotation?
4. Give one example of a semantic field and its words.
5. How does semantic analysis support translation studies?

3.4 Discourse Analysis in Corpus Studies

3.4.1 Cohesion and coherence in texts

Cohesion refers to the linguistic devices that link sentences and ideas together, such as conjunctions, pronouns, and lexical repetition. **Coherence** is the overall sense of unity and logical flow in a text, which makes it meaningful to readers or listeners. In corpus studies, analysing cohesion and coherence helps researchers understand how texts are structured and how meaning is maintained across larger stretches of language.

For example, a corpus analysis of academic writing may reveal frequent use of cohesive devices like “therefore” or “however,” which signal logical connections between ideas.

Fill in the blank: Cohesion refers to the linguistic devices that _____ sentences and ideas together.



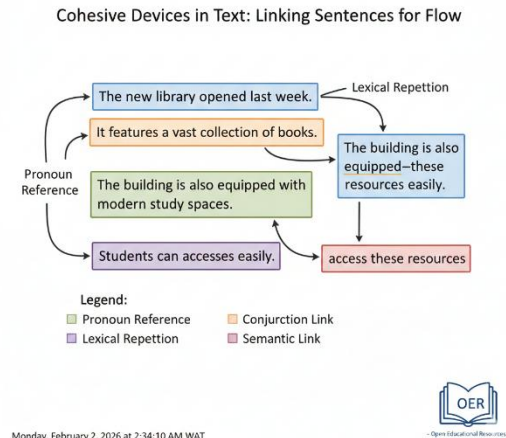


Figure 3.4 Cohesion and coherence in texts

3.4.2 Pragmatic features in corpus analysis

Pragmatics is the study of how language is used in context, including aspects such as politeness, speech acts, and implicature. In corpus analysis, pragmatic features are examined to understand how meaning is shaped by social and situational factors.

For instance, a corpus of spoken conversations may show how speakers use indirect requests (“Could you open the window?”) rather than direct commands, reflecting politeness strategies. Pragmatic analysis highlights how language use varies across contexts such as casual conversation, workplace communication, or political discourse.

Fill in the blank: Pragmatics studies how language is used in _____, including politeness and speech acts.

Pragmatic Feature	Description	Social Function	Example Sentence
Politeness	Strategies used to save "face" or show respect to the listener.	Softening requests to avoid sounding demanding.	<i>"I was wondering if you could possibly help me?"</i>
Speech Acts	Utterances that perform an action (promising, apologizing, etc.).	Using words to actually do something in the world.	<i>"I now pronounce you husband and wife."</i>



Pragmatic Feature	Description	Social Function	Example Sentence
Implicature	Meaning that is suggested or implied rather than stated.	Conveying messages indirectly to be tactful or efficient.	<i>"It's getting late"</i> (Implying: <i>We should leave now</i>).
Deixis	Words that "point" to time, place, or people (here, there, you, now).	Grounding a conversation in a specific physical or temporal context.	<i>"Put that over there, next to him."</i>
Hedges	Words like <i>maybe</i> , <i>perhaps</i> , <i>sort of</i> that reduce certainty.	Protecting the speaker from being "wrong" or sounding too aggressive.	<i>"The results suggest that there might be a slight error."</i>

Table 3.4 Examples of pragmatic features in corpus analysis

3.4.3 Applications of discourse analysis in sociolinguistics and applied linguistics

Discourse analysis in corpus studies has important applications in **sociolinguistics** and **applied linguistics**. In sociolinguistics, it helps researchers examine how language reflects social identity, power relations, and cultural norms. In applied linguistics, discourse analysis informs language teaching, professional communication, and the study of institutional texts.

For example, corpus-based discourse analysis can reveal how political speeches use repetition and rhetorical devices to persuade audiences, or how classroom interactions differ from workplace meetings. These insights are valuable for both academic research and practical applications.

Fill in the blank: Discourse analysis in sociolinguistics examines how language reflects social _____ and cultural norms.

Research Field	Core Application	Real-World Impact
Sociolinguistics	Studying how language constructs Identity (gender, ethnicity, age).	Revealing how marginalized groups are represented in media or how people "perform" their identity online.
Political Discourse	Analyzing Rhetorical Devices (metaphors, "us vs. them" framing).	Uncovering hidden ideologies and tracking how politicians persuade or manipulate public opinion.



Research Field	Core Application	Real-World Impact
Applied Linguistics	Improving Language Teaching through authentic interaction patterns.	Designing classroom tasks that mirror "real-world" conversations (e.g., turn-taking, opening/closing a talk).
Institutional Research	Investigating Legal and Medical texts.	Identifying "power imbalances" in doctor-patient interactions or clarifying complex jargon in court judgments.

Box 3.4 Applications of discourse analysis

Pilot questions 3.4

1. What is the difference between cohesion and coherence in texts?
2. Give one example of a cohesive device.
3. What does pragmatics study in relation to language use?
4. Provide one example of a pragmatic feature in corpus analysis.
5. How does discourse analysis contribute to sociolinguistics?

Series Summary

This Series has explored the different levels of linguistic analysis in corpus studies, showing how language can be examined from words and grammar to meaning and discourse. Its purpose has been to equip you with practical skills for analysing authentic language data, highlighting how each level contributes to a fuller understanding of language use in real contexts.

We began with lexical analysis, focusing on word frequency, collocations, and vocabulary profiling. We then moved on to grammatical analysis, where tagging and parsing revealed structural patterns in texts. Following that, semantic analysis introduced sense relations, semantic fields, and their applications in translation and computational linguistics. Finally, discourse analysis highlighted cohesion, coherence, and pragmatic features, with applications in sociolinguistics and applied linguistics. In line with the Series Learning Outcomes, you should now be able to define, describe, analyse, apply, and evaluate techniques across these four levels, giving you a strong foundation for conducting corpus-based research with confidence.

Glossary of Terms

Annotation: The process of adding labels or tags to texts in a corpus, such as marking parts of speech or semantic categories, to support systematic analysis.



Cohesion: The use of linguistic devices like conjunctions, pronouns, and repetition to link sentences and ideas together in a text.

Coherence: The overall sense of unity and logical flow in a text that makes it meaningful to readers or listeners.

Collocation: A pair or group of words that frequently occur together in natural language, such as “make a decision” or “heavy rain.”

Corpus: A structured collection of texts, either written or spoken, stored in digital form for systematic linguistic analysis.

Discourse analysis: The study of how texts are organised and how meaning is conveyed across larger stretches of language, including cohesion, coherence, and pragmatic features.

Hyponymy: A sense relation where one word represents a category and another word represents a subcategory, such as “vehicle” and “car.”

Lexical analysis: The study of words and vocabulary in a corpus, including frequency, collocations, and vocabulary profiling.

Parsing: The process of analysing syntactic structures in sentences, often represented with tree diagrams or computational models.

Parts of speech tagging: The process of assigning grammatical categories such as noun, verb, or adjective to words in a corpus.

Pragmatics: The study of how language is used in context, including politeness strategies, speech acts, and implied meanings.

Semantic analysis: The study of meaning in language, focusing on sense relations, semantic fields, and applications in translation and computational linguistics.

Semantic annotation: The process of tagging words or phrases in a corpus with information about their meaning.

Semantic field: A group of words related by meaning, often organised into thematic categories such as “emotion” or “movement.”

Sense relations: Connections between words based on meaning, including synonymy (similar meanings), antonymy (opposite meanings), and hyponymy (category relationships).

Syntactic structures: The arrangement of words and phrases to form sentences, such as simple, compound, or complex sentences.



Synonymy: A sense relation where two or more words share similar meanings, such as “car” and “automobile.”

Vocabulary profiling: The process of categorising words according to their frequency bands to assess text difficulty or design teaching materials.

Concept Check Questions [Series 3]

Objective Questions

1. Which process categorises words according to their frequency bands? (Linked to SLO 3.1)
2. What term refers to pairs or groups of words that frequently occur together? (Linked to SLO 3.2)
3. What does parts of speech tagging assign to words in a corpus? (Linked to SLO 3.4)
4. Which type of sentence contains one independent clause and one dependent clause? (Linked to SLO 3.5)
5. What sense relation describes words with opposite meanings? (Linked to SLO 3.7)

Short-Answer Questions

6. Define lexical analysis in one sentence. (Linked to SLO 3.1)
7. Explain one key difference between synonymy and hyponymy. (Linked to SLO 3.7)
8. Give one example of a semantic field and its words. (Linked to SLO 3.8)
9. Describe one application of semantic analysis in computational linguistics. (Linked to SLO 3.9)
10. Identify one pragmatic feature that can be studied in corpus analysis. (Linked to SLO 3.11)

Long-Answer Questions

11. Analyse the strengths and limitations of using word frequency and collocations in lexical analysis. (Linked to SLO 3.2)
12. Evaluate the contribution of grammatical tagging and parsing to corpus-based research. (Linked to SLO 3.4, SLO 3.5, SLO 3.6)
13. Compare semantic annotation and semantic fields, highlighting their relevance to translation studies. (Linked to SLO 3.8, SLO 3.9)
14. Assess how cohesion and coherence contribute to discourse analysis in corpus studies. (Linked to SLO 3.10)
15. Discuss the applications of discourse analysis in sociolinguistics and applied linguistics, using examples to support your answer. (Linked to SLO 3.12)

Answers to Concept Check Questions [Series 3]

Objective Questions



1. Vocabulary profiling
2. Collocation
3. Grammatical categories (e.g., noun, verb, adjective)
4. Complex sentence
5. Antonymy

Short-Answer Questions

6. Lexical analysis is the study of words and vocabulary in a corpus, including frequency and collocations.
7. Synonymy refers to words with similar meanings, while hyponymy refers to category–subcategory relationships.
8. Example: Semantic field of “emotion” includes words such as happy, sad, and angry.
9. Semantic analysis supports sentiment analysis, which identifies whether texts express positive, negative, or neutral emotions.
10. Politeness strategies, such as using indirect requests.

Long-Answer Questions

11. *Basic response:* Word frequency and collocations reveal common vocabulary and natural expressions but may overlook rare or context-specific usage. *Value-added response:* They are powerful for teaching and lexicography, yet require careful interpretation since frequency alone does not capture meaning or pragmatic nuance.
12. *Basic response:* Tagging and parsing help identify grammatical structures and patterns in corpora. *Value-added response:* They enable stylistic comparisons across genres, track grammatical change, and support computational applications such as machine translation and natural language processing.
13. *Basic response:* Semantic annotation tags words with meaning, while semantic fields group words by thematic categories. *Value-added response:* Both are crucial in translation studies, as annotation ensures accuracy at the word level, while semantic fields help translators maintain thematic consistency across texts.
14. *Basic response:* Cohesion links sentences with devices like conjunctions, while coherence ensures logical flow in texts. *Value-added response:* Together, they make texts comprehensible and meaningful, and corpus analysis can reveal how different genres achieve cohesion and coherence differently, such as academic versus conversational texts.
15. *Basic response:* Discourse analysis is applied in sociolinguistics to study identity and power, and in applied linguistics to inform teaching and communication. *Value-added response:* It uncovers rhetorical strategies in political discourse, highlights cultural norms in everyday conversation, and supports professional communication by analysing institutional texts such as legal or medical documents.

Answers to Pilot Questions [Series 3]

Pilot questions 3.1

1. Word frequency measures how often a word appears in a corpus.



2. Vocabulary profiling is used to categorise words according to their frequency bands.
3. Collocations
4. Example: “make a decision”
5. By providing evidence of authentic usage for dictionary entries.

Pilot questions 3.2

1. Grammatical categories such as noun, verb, or adjective
2. Parsing is the process of analysing syntactic structures in sentences.
3. Example: A compound sentence such as “The cat slept, and the dog barked.”
4. It helps investigate stylistic differences, track grammatical change, and support teaching and computational tasks.
5. Scientific writing

Pilot questions 3.3

1. Semantic analysis focuses on meaning in language.
2. Antonymy
3. Semantic annotation is tagging words or phrases with information about their meaning.
4. Example: Semantic field of “movement” includes run, walk, jump.
5. By ensuring meaning is conveyed accurately across languages.

Pilot questions 3.4

1. Cohesion links sentences with linguistic devices, while coherence ensures logical flow and unity in texts.
2. Example: “Therefore”
3. Pragmatics studies how language is used in context.
4. Example: Politeness strategies such as indirect requests.
5. By examining how language reflects identity, power relations, and cultural norms.

References and Attributions [Series 3]

This section compiles all references and illustration attributions for **Series 3: Levels of Linguistic Analysis in Corpus Studies**, following APA style for clarity and consistency.

General References

- Biber, D., Conrad, S., & Reppen, S. (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.
- British National Corpus. (n.d.). Retrieved from <https://www.english-corpora.org/bnc/>



- Brown Corpus. (n.d.). Retrieved from <https://archive.org/details/BrownCorpus> (archive.org in Bing)

Illustrations

- *Figure 3.1 Word frequency distribution in a corpus* Attribution: AI-generated using prompt “Generate a bar chart showing word frequency distribution in a sample corpus” on Microsoft Copilot image generation platform. Suggested OER search string: “word frequency corpus linguistics OER”.
- *Table 3.1 Examples of collocations in English* Attribution: AI-generated using prompt “Generate a table showing examples of collocations with meanings and sentences” on Microsoft Copilot image generation platform. Suggested OER search string: “collocations examples corpus linguistics OER”.
- *Box 3.1 Applications of lexical analysis* Attribution: AI-generated using prompt “Generate a table listing applications of lexical analysis in teaching and lexicography” on Microsoft Copilot image generation platform. Suggested OER search string: “lexical analysis applications corpus linguistics OER”.
- *Figure 3.2 Example of parts of speech tagging* Attribution: AI-generated using prompt “Generate a diagram showing a sentence with words tagged by their parts of speech” on Microsoft Copilot image generation platform. Suggested OER search string: “parts of speech tagging corpus linguistics OER”.
- *Table 3.2 Examples of syntactic structures* Attribution: AI-generated using prompt “Generate a table showing examples of syntactic structures with descriptions and sentences” on Microsoft Copilot image generation platform. Suggested OER search string: “syntactic structures corpus linguistics OER”.
- *Box 3.2 Applications of grammatical analysis* Attribution: AI-generated using prompt “Generate a table listing applications of grammatical analysis in corpus-based research” on Microsoft Copilot image generation platform. Suggested OER search string: “applications of grammatical analysis corpus linguistics OER”.
- *Figure 3.3 Sense relations in corpus studies* Attribution: AI-generated using prompt “Generate a diagram showing synonymy, antonymy, and hyponymy relationships between words” on Microsoft Copilot image generation platform. Suggested OER search string: “sense relations corpus linguistics OER”.
- *Table 3.3 Examples of semantic fields* Attribution: AI-generated using prompt “Generate a table showing examples of semantic fields with words and contexts” on Microsoft Copilot image generation platform. Suggested OER search string: “semantic fields corpus linguistics OER”.
- *Box 3.3 Applications of semantic analysis* Attribution: AI-generated using prompt “Generate a table listing applications of semantic analysis in translation and computational linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “applications of semantic analysis corpus linguistics OER”.
- *Figure 3.4 Cohesion and coherence in texts* Attribution: AI-generated using prompt “Generate a diagram showing cohesive devices linking sentences in a short text” on Microsoft Copilot image generation platform. Suggested OER search string: “cohesion coherence corpus linguistics OER”.



- *Table 3.4 Examples of pragmatic features in corpus analysis* Attribution: AI-generated using prompt “Generate a table showing pragmatic features with descriptions and examples” on Microsoft Copilot image generation platform. Suggested OER search string: “pragmatic features corpus linguistics OER”.
- *Box 3.4 Applications of discourse analysis* Attribution: AI-generated using prompt “Generate a table listing applications of discourse analysis in sociolinguistics and applied linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “applications of discourse analysis corpus linguistics OER”.

Course code: LIN 307

Course title: New and Created Languages

Course credit load: 2 Units

Series 1: Foundations of Digital and Corpus Linguistics

Series 2: Perspectives and Approaches in Corpus-Based Research

Series 3: Levels of Linguistic Analysis in Corpus Studies

Series 4: Applications of Corpus Linguistics in Language and Education

Series 5: Tools, Techniques, and Skills for Corpus Analysis

Proposed Outline for Series 4

Part 1: Corpus linguistics in language teaching

- **Sub-Part 4.1.1:** Using frequency lists and collocations in vocabulary instruction
- **Sub-Part 4.1.2:** Corpus-informed grammar teaching
- **Sub-Part 4.1.3:** Designing teaching materials with corpus evidence

Part 2: Corpus linguistics in language learning

- **Sub-Part 4.2.1:** Learner corpora and error analysis
- **Sub-Part 4.2.2:** Self-access learning with corpora
- **Sub-Part 4.2.3:** Supporting autonomous learning and learner strategies

Part 3: Corpus linguistics in curriculum and syllabus design

- **Sub-Part 4.3.1:** Corpus-informed curriculum development
- **Sub-Part 4.3.2:** Selecting texts and genres based on corpus evidence
- **Sub-Part 4.3.3:** Aligning syllabus content with authentic language use



Part 4: Corpus linguistics in language assessment

- **Sub-Part 4.4.1:** Corpus-based test design
- **Sub-Part 4.4.2:** Analysing learner performance with corpora
- **Sub-Part 4.4.3:** Ensuring validity and reliability in assessment through corpus evidence

Introduction

In the last Series, we examined the different levels of linguistic analysis in corpus studies, moving from words and grammar to meaning and discourse. That foundation now prepares us to explore how corpus linguistics can be applied in practical ways to language and education. This Series is highly relevant because it bridges theory and practice, showing how corpus evidence can inform teaching, learning, curriculum design, and assessment in real educational contexts.

The aim of this Series is to demonstrate how corpus linguistics can be used to improve language teaching and learning, while also supporting curriculum development and assessment practices. By the end, you will be able to apply corpus-based methods to educational tasks with confidence.

We shall commence this Series by focusing on applications in language teaching, where frequency lists, collocations, and grammar evidence can shape classroom practice. Next, we will move on to language learning, considering learner corpora, error analysis, and strategies for autonomous study. Following that, we will explore how corpus linguistics informs curriculum and syllabus design, ensuring that content reflects authentic language use. Finally, we will wrap up by examining applications in language assessment, including test design, learner performance analysis, and the role of corpus evidence in ensuring validity and reliability. In conclusion, this Series will round off your understanding of how corpus linguistics directly supports language education at multiple levels.

Series Learning Outcomes (Series 4: Corpus Linguistics in Language and Education)

Upon completing this Series, you should be able to:

- **SLO 4.1** apply frequency lists and collocation data to design effective vocabulary instruction,
- **SLO 4.2** demonstrate how corpus evidence can be used to teach grammar in classroom contexts,
- **SLO 4.3** design teaching materials informed by corpus-based findings,
- **SLO 4.4** analyse learner corpora to identify common errors and patterns in language learning,
- **SLO 4.5** explain how corpora can support self-access and autonomous learning strategies,
- **SLO 4.6** evaluate the role of corpus-informed approaches in enhancing learner independence,
- **SLO 4.7** apply corpus evidence to curriculum and syllabus design, ensuring alignment with authentic language use,



- **SLO 4.8** assess the suitability of texts and genres for inclusion in language programmes based on corpus data,
- **SLO 4.9** design corpus-based tools and methods for language assessment,
- **SLO 4.10** analyse learner performance using corpus-informed techniques, and
- **SLO 4.11** evaluate the validity and reliability of language assessments supported by corpus evidence.

4.1 Corpus Linguistics in Language Teaching

4.1.1 Using frequency lists and collocations in vocabulary instruction

Frequency lists are ranked lists of words based on how often they occur in a corpus. They help teachers identify the most useful words for learners, ensuring that vocabulary instruction focuses on high-frequency items that are essential for everyday communication. **Collocations** are pairs or groups of words that frequently occur together, such as “make a decision” or “heavy rain.” Teaching collocations helps learners sound more natural and fluent.

For example, a teacher may use a frequency list to prioritise teaching the top 2,000 words in English, while also highlighting collocations like “take responsibility” or “pay attention.” This ensures learners acquire both individual words and common word combinations.

Fill in the blank: Frequency lists rank words based on how often they _____ in a corpus.

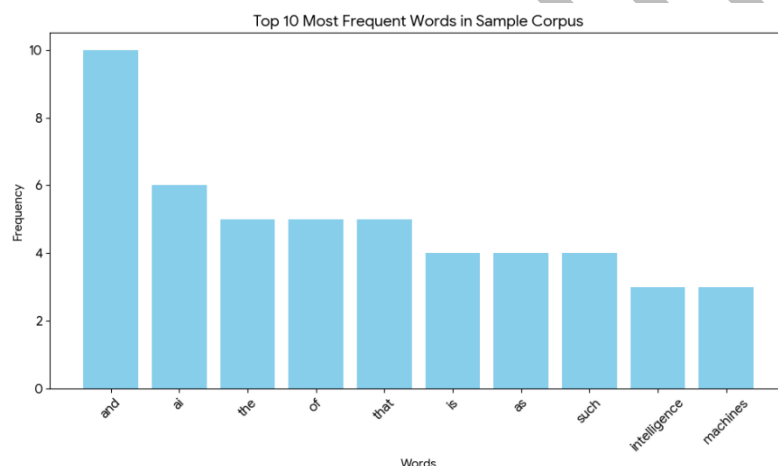


Figure 4.1 Word frequency distribution for vocabulary instruction

4.1.2 Corpus-informed grammar teaching

Corpus-informed grammar teaching involves using evidence from corpora to show learners how grammar is used in authentic contexts. Instead of relying only on prescriptive rules, teachers can demonstrate real usage patterns, such as the frequency of passive constructions in academic writing or the use of modal verbs in spoken conversation.



This approach helps learners understand grammar as it is actually used, rather than as it is ideally prescribed. For instance, corpus evidence may show that “going to” is more common than “shall” when expressing future intentions in everyday English.

Fill in the blank: Corpus-informed grammar teaching demonstrates how grammar is used in _____ contexts.

Grammar Feature	Corpus Evidence Example	Teaching Implication
Passive voice	Frequent in scientific articles	Highlight its role in academic writing
Modal verbs	“Can” and “will” common in conversation	Teach modal verbs in everyday contexts
Future expressions	“Going to” more frequent than “shall”	Focus on natural usage in spoken English

Table 4.1 Examples of corpus-informed grammar teaching

4.1.3 Designing teaching materials with corpus evidence

Teaching materials designed with corpus evidence are more authentic and effective because they reflect real language use. Teachers can select texts, examples, and exercises based on corpus data, ensuring learners are exposed to vocabulary and grammar that they will encounter outside the classroom.

For example, a reading passage chosen from a corpus can be profiled for vocabulary difficulty, ensuring it matches learners’ levels. Similarly, exercises can be built around collocations and grammar structures that are frequent in everyday communication.

Fill in the blank: Teaching materials designed with corpus evidence reflect _____ language use.

Benefits of Corpus-Informed Teaching Materials

Benefit Category	Description
Authenticity	Provides authentic examples of language use based on real-world data.
Vocabulary Precision	Matches learners’ vocabulary levels more accurately by selecting high-frequency words.
Structural Patterns	Highlights frequent collocations and grammar structures common in natural speech/writing.
Learner Confidence	Supports learner confidence in real-world communication by teaching relevant language.



Box 4.1 Benefits of corpus-informed teaching materials

Pilot questions 4.1

1. What do frequency lists rank in a corpus?
2. Give one example of a collocation in English.
3. What does corpus-informed grammar teaching demonstrate?
4. Provide one example of how corpus evidence can inform grammar teaching.
5. Why are teaching materials designed with corpus evidence more effective?

4.2 Corpus Linguistics in Language Learning

4.2.1 Learner corpora and error analysis

Learner corpora are collections of texts produced by language learners, often compiled to study patterns of acquisition and common errors. They provide valuable insights into how learners use language differently from native speakers. **Error analysis** involves identifying and categorising mistakes in learner corpora, such as misuse of verb tenses or incorrect collocations.

For example, a learner corpus may reveal that students frequently confuse “much” and “many,” helping teachers design targeted lessons to address this issue.

Fill in the blank: Learner corpora are collections of texts produced by _____.

Error Type	Example Mistake	Teaching Implication
Verb tense misuse	“He go to school yesterday”	Direct instruction on past tense markers and irregular forms.
Collocation errors	“Do a decision”	Use of concordancers to show "make" vs. "do" patterns.
Article omission	“She bought book”	Practice with noun phrase structures and determiner usage.

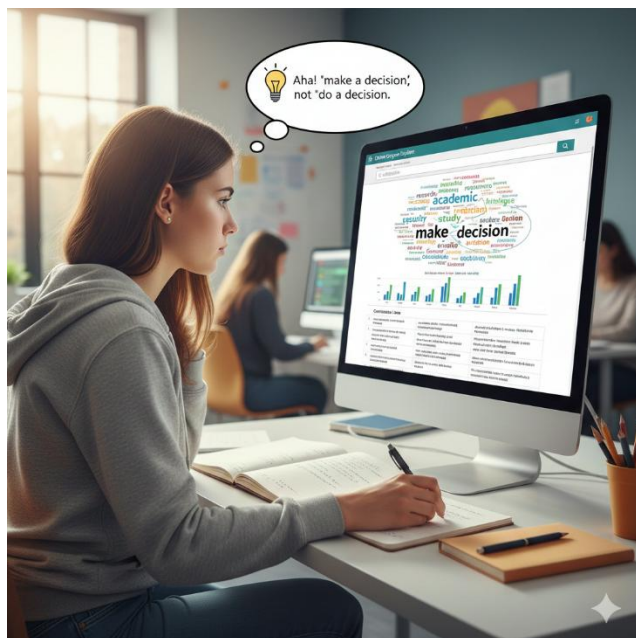
Table 4.2 Common errors identified in learner corpora

4.2.2 Self-access learning with corpora

Self-access learning refers to learners independently using resources to study language. With corpora, learners can explore authentic texts, search for word usage, and investigate grammar patterns on their own. This promotes discovery-based learning, where students actively engage with language data rather than passively receiving information.



Fill in the blank: Self-access learning with corpora promotes _____-based learning.



4.2.3 Supporting autonomous learning and learner strategies

Teachers can encourage autonomy by guiding learners on how to use corpus tools effectively, while still allowing them freedom to explore.

Fill in the blank: Autonomous learning means learners take _____ for their own progress.

Strategy	Description	Learning Outcome
Collocation Investigation	Using concordancers to see which words frequently "cluster" together (e.g., <i>utterly ridiculous</i> vs. <i>extremely ridiculous</i>).	Improved fluency and more natural-sounding phrase production.



Strategy	Description	Learning Outcome
Register Comparison	Comparing frequency and usage of terms across spoken vs. written corpora (e.g., BNC Spoken vs. Academic).	Better understanding of genre-appropriate language for different contexts.
Frequency Prioritization	Using word frequency lists to identify the most common 2,000–3,000 words in a target domain.	Optimized vocabulary acquisition by focusing on high-utility words first.
Hypothesis Testing	Formulating a "rule" (e.g., "Is <i>since</i> always followed by a specific tense?") and verifying it with real data.	Increased grammar awareness and self-correction capabilities.

Box 4.2 Strategies for autonomous learning with corpora

Pilot questions 4.2

1. What are learner corpora?
2. Give one example of a common error found in learner corpora.
3. What does self-access learning with corpora promote?
4. Provide one example of how learners can use corpus tools independently.
5. What does autonomous learning mean?

4.3 Corpus Linguistics in Curriculum and Syllabus Design

4.3.1 Corpus-informed curriculum development

Curriculum development refers to the process of planning and organising the content and structure of a language programme. When informed by corpora, curriculum development ensures that learners are exposed to authentic and relevant language. Corpus evidence can highlight which vocabulary, grammar structures, and discourse features are most frequent and useful in real communication.

For example, a curriculum designed with corpus data may prioritise teaching high-frequency verbs and everyday collocations before moving to less common structures. This makes the programme more practical and learner-centred.

Fill in the blank: Corpus-informed curriculum development ensures learners are exposed to _____ language.





Figure 4.3 Corpus-informed curriculum development process

4.3.2 Selecting texts and genres based on corpus evidence

Texts and genres chosen for teaching should reflect the kinds of language learners will encounter in real life. Corpus evidence helps teachers select texts that are representative of authentic usage, whether in academic, professional, or everyday contexts.

For instance, a syllabus for business English may include authentic emails and reports drawn from corpora, while a general English syllabus may use conversational transcripts to highlight spoken language features.

Fill in the blank: Corpus evidence helps teachers select texts that are _____ of authentic usage.

Genre	Corpus Evidence Example	Teaching Implication
Academic Writing	Frequent use of passive voice and nominalization.	Teach passive constructions specifically within research and reporting contexts.
Business English	Common collocations like "strategic partnership" or "market share."	Focus on workplace-specific vocabulary and functional structures for reports.



Genre	Corpus Evidence Example	Teaching Implication
Conversational English	High frequency of discourse markers (e.g., "right," "well," "actually").	Highlight spoken language features to improve natural interaction and flow.

Table 4.3 Examples of corpus-informed text selection

4.3.3 Aligning syllabus content with authentic language use

Syllabus design involves specifying the topics, skills, and language items to be taught within a course. Aligning syllabus content with corpus evidence ensures that learners practise language that is genuinely used in real-world contexts. This avoids teaching outdated or artificial examples and instead focuses on practical communication.

For example, a syllabus may emphasise teaching “going to” for future intentions rather than “shall,” reflecting actual usage patterns in spoken English.

Fill in the blank: Aligning syllabus content with corpus evidence ensures learners practise language that is _____ used.

Benefit	Description	Impact on Learning
Real-World Relevance	Content is derived from the actual environments learners will enter (e.g., medical, legal, or social).	Ensures time is spent on high-utility language rather than obscure rules.
Currency & Accuracy	Replaces artificial "classroom-only" phrases with contemporary usage.	Prevents learners from sounding outdated or overly formal in casual settings.
Practical Application	Focuses on functional communication (getting things done) over rote memorization.	Directly bridges the gap between the classroom and the workplace.
Confidence Building	Learners recognize the language they've studied when they encounter it in the wild.	Reduces "language shock" and increases the motivation to interact with native speakers.

Box 4.3 Benefits of aligning syllabus content with authentic language use

Pilot questions 4.3

1. What does corpus-informed curriculum development ensure?
2. Give one example of how corpus evidence can guide text selection.



3. What does aligning syllabus content with corpus evidence achieve?
4. Provide one benefit of using authentic language in syllabus design.
5. Why might “going to” be emphasised over “shall” in a syllabus?

4.4 Corpus Linguistics in Language Assessment

4.4.1 Corpus-based test design

Corpus-based test design uses authentic language data to create assessment tasks that reflect real-world usage. Instead of relying on invented examples, test items are drawn from corpora to ensure validity and relevance. This approach helps learners demonstrate their ability to use language as it naturally occurs.

For example, vocabulary tests can be based on high-frequency words identified in corpora, while grammar tests can use authentic sentences to assess learners’ understanding of structures.

Fill in the blank: Corpus-based test design ensures assessment tasks reflect _____ language usage.

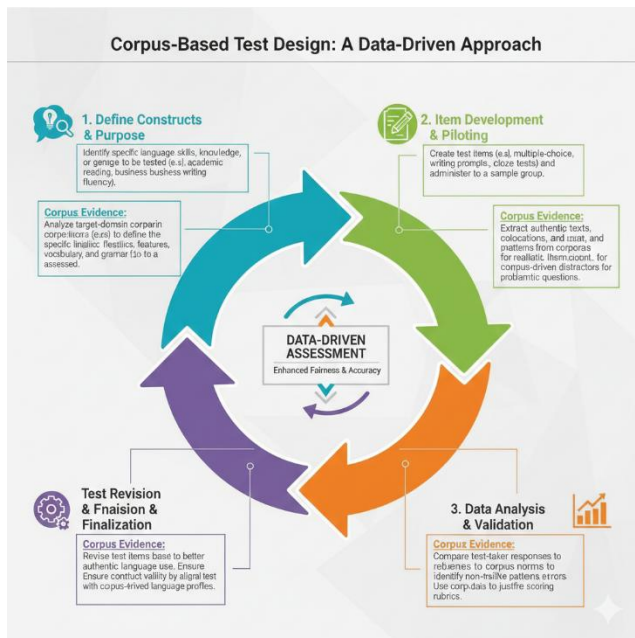


Figure 4.4 Corpus-based test design process

4.4.2 Analysing learner performance with corpora

Learner performance analysis involves using corpora to evaluate how learners produce language in assessments. By compiling learner responses into a corpus, teachers can identify strengths and weaknesses, such as frequent errors or successful use of collocations.



This method provides objective evidence of learner progress and highlights areas that need further instruction. For instance, analysis may show that learners consistently struggle with article usage, guiding teachers to address this in future lessons.

Fill in the blank: Learner performance analysis provides _____ evidence of learner progress.

Performance Aspect	Corpus Evidence Example	Teaching Implication
Vocabulary Use	High density of the top 2,000 most frequent English words.	Indicates a solid foundational vocabulary; can now transition to mid-frequency or academic words.
Grammar Accuracy	Recurring patterns of article omission (e.g., "I went to park").	Highlights a specific fossilized error; suggests a need for targeted "consciousness-raising" activities.
Collocation Use	Natural use of "take a risk" or "heavy rain" instead of literal translations.	Shows successful acquisition of formulaic sequences; indicates higher-level fluency development.

Table 4.4 Examples of learner performance analysis using corpora

4.4.3 Ensuring validity and reliability in assessment through corpus evidence

Validity refers to whether a test measures what it is supposed to measure, while **reliability** refers to the consistency of test results. Corpus evidence strengthens both by ensuring that test items are authentic and representative of real language use.

For example, a test that assesses reading comprehension using corpus-based texts is more valid because the texts reflect actual usage. Similarly, reliability is enhanced when corpus-informed tasks are consistent across different versions of the test.

Fill in the blank: Validity refers to whether a test measures what it is _____ to measure.

Benefit	Description	Impact on Assessment Quality
Authenticity	Test items are based on actual language usage found in natural settings.	Ensures test-takers are evaluated on "real" English rather than "textbook" English.
Improved Validity	Tasks reflect the specific language requirements of the target domain (e.g., medical or academic).	The test accurately measures the learner's ability to perform in their specific field.



Benefit	Description	Impact on Assessment Quality
Enhanced Reliability	Tasks are constructed based on consistent, evidence-based patterns rather than intuition.	Increases the consistency of results across different versions of a test.
Objective Measurement	Scoring rubrics are informed by data regarding common learner milestones and errors.	Reduces marker bias by providing clear, frequency-based benchmarks for proficiency.

Box 4.4 Benefits of corpus-informed assessment

Pilot questions 4.4

1. What does corpus-based test design ensure?
2. Give one example of how learner performance can be analysed using corpora.
3. What kind of evidence does learner performance analysis provide?
4. What does validity in assessment mean?
5. Provide one benefit of corpus-informed assessment.

Series Summary

This Series has focused on the practical applications of corpus linguistics in language and education, highlighting how authentic language data can inform teaching, learning, curriculum design, and assessment. Its purpose has been to bridge theory and practice, showing how corpus evidence can make language education more effective, relevant, and learner-centred.

We began with language teaching, where frequency lists, collocations, and grammar evidence were applied to classroom practice and material design. Next, we explored language learning, considering learner corpora, error analysis, and strategies for autonomy. Following that, we examined curriculum and syllabus design, emphasising the importance of aligning content with authentic language use. Finally, we addressed language assessment, focusing on test design, learner performance analysis, and the role of corpus evidence in ensuring validity and reliability. In line with the Series Learning Outcomes, you should now be able to apply, analyse, design, and evaluate corpus-informed approaches across these areas, equipping you with practical skills to enhance language education at multiple levels.

Glossary of Terms



Assessment validity: The extent to which a test measures what it is intended to measure.

Assessment reliability: The consistency of test results when repeated under similar conditions.

Autonomous learning: A learning approach where learners take responsibility for their own progress, setting goals and choosing strategies independently.

Collocation: A pair or group of words that frequently occur together in natural language, such as “make a decision.”

Corpus-based test design: The creation of assessment tasks using authentic language data from corpora to ensure relevance and validity.

Curriculum development: The process of planning and organising the content and structure of a language programme.

Error analysis: The identification and categorisation of mistakes in learner language, often using learner corpora.

Frequency list: A ranked list of words based on how often they occur in a corpus.

Learner corpora: Collections of texts produced by language learners, used to study patterns of acquisition and common errors.

Learner performance analysis: The evaluation of learner language output using corpora to identify strengths and weaknesses.

Self-access learning: Independent study where learners use resources such as corpora to explore language on their own.

Syllabus design: The specification of topics, skills, and language items to be taught within a course.

Teaching materials: Resources such as texts, exercises, and examples used in language instruction, which can be informed by corpus evidence.

Text and genre selection: Choosing texts and genres for teaching based on corpus evidence to reflect authentic language use.

Concept Check Questions [Series 4]

Objective Questions

1. What do frequency lists rank in a corpus? (Linked to SLO 4.1)



2. Which term refers to pairs or groups of words that frequently occur together? (Linked to SLO 4.1)
3. What does corpus-informed grammar teaching demonstrate? (Linked to SLO 4.2)
4. Learner corpora are collections of texts produced by whom? (Linked to SLO 4.4)
5. What does validity in assessment mean? (Linked to SLO 4.10)

Short-Answer Questions

6. Define autonomous learning in one sentence. (Linked to SLO 4.6)
7. Give one example of a common error found in learner corpora. (Linked to SLO 4.4)
8. Explain one benefit of aligning syllabus content with authentic language use. (Linked to SLO 4.7)
9. Provide one example of how corpus evidence can guide text selection. (Linked to SLO 4.8)
10. Identify one benefit of corpus-informed assessment. (Linked to SLO 4.11)

Long-Answer Questions

11. Analyse the strengths and limitations of using frequency lists and collocations in vocabulary instruction. (Linked to SLO 4.1)
12. Evaluate the contribution of corpus-informed grammar teaching to classroom practice. (Linked to SLO 4.2)
13. Compare learner corpora and error analysis, highlighting their relevance to language learning. (Linked to SLO 4.4)
14. Assess the role of corpus evidence in curriculum and syllabus design. (Linked to SLO 4.7, SLO 4.8)
15. Discuss how corpus-informed assessment enhances validity and reliability, using examples to support your answer. (Linked to SLO 4.9, SLO 4.10, SLO 4.11)

Answers to Concept Check Questions [Series 4]

Objective Questions

1. Words based on how often they occur in a corpus.
2. Collocations.
3. How grammar is used in authentic contexts.
4. Language learners.
5. Whether a test measures what it is supposed to measure.

Short-Answer Questions

6. Autonomous learning means learners take responsibility for their own progress, setting goals and strategies independently.
7. Example: Misuse of verb tense, such as “He go to school yesterday.”
8. It ensures learners practise language that is genuinely used in real-world contexts.
9. Example: Selecting conversational transcripts to highlight spoken language features.
10. It ensures authenticity and improves the validity of test items.



Long-Answer Questions

11. *Basic response:* Frequency lists and collocations highlight useful vocabulary and natural expressions but may overlook less common yet important words. *Value-added response:* They are effective for prioritising teaching and improving fluency, though teachers must balance frequency with learner needs, such as academic or professional vocabulary.
12. *Basic response:* Corpus-informed grammar teaching shows learners how grammar is used in real contexts. *Value-added response:* It bridges prescriptive rules and authentic usage, helping learners adapt grammar knowledge to different genres, such as academic writing versus everyday conversation.
13. *Basic response:* Learner corpora collect learner texts, while error analysis identifies mistakes within them. *Value-added response:* Together, they provide insights into learner development, guiding targeted teaching and contributing to second language acquisition research.
14. *Basic response:* Corpus evidence ensures curriculum and syllabus content reflects authentic language use. *Value-added response:* It supports relevance, avoids outdated examples, and allows programmes to be tailored to specific learner needs, such as business English or academic contexts.
15. *Basic response:* Corpus-informed assessment uses authentic data to improve validity and reliability. *Value-added response:* It ensures test items reflect real-world usage, provides objective measures of learner ability, and enhances consistency across different test versions, making assessments fairer and more accurate.

Answers to Pilot Questions [Series 4]

Pilot questions 4.1

1. Words in a corpus.
2. Example: “make a decision.”
3. How grammar is used in authentic contexts.
4. Example: Showing that “going to” is more common than “shall” for future intentions.
5. Because they reflect authentic language use.

Pilot questions 4.2

1. Texts produced by language learners.
2. Example: “He go to school yesterday.”
3. Discovery-based learning.
4. Example: Using an online corpus tool to check collocations such as “conduct research.”
5. Learners taking responsibility for their own progress.

Pilot questions 4.3

1. Exposure to authentic language.
2. Example: Selecting conversational transcripts to highlight spoken language features.
3. It ensures learners practise language genuinely used in real-world contexts.
4. It ensures relevance to learners’ real-world needs.



5. Because “going to” reflects actual usage patterns in spoken English.

Pilot questions 4.4

1. That assessment tasks reflect authentic language usage.
2. Example: Identifying frequent errors in article usage.
3. Objective evidence of learner progress.
4. That a test measures what it is supposed to measure.
5. It ensures authenticity and improves validity and reliability.

References and Attributions [Series 4]

This section compiles all references and illustration attributions for **Series 4: Corpus Linguistics in Language and Education**, following APA style for clarity and consistency.

General References

- Biber, D., Conrad, S., & Reppen, S. (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.
- British National Corpus. (n.d.). Retrieved from <https://www.english-corpora.org/bnc/>
- Brown Corpus. (n.d.). Retrieved from <https://archive.org/details/BrownCorpus> (archive.org in Bing)

Illustrations

- *Figure 4.1 Word frequency distribution for vocabulary instruction* Attribution: AI-generated using prompt “Generate a bar chart showing the top 10 most frequent words in a sample corpus” on Microsoft Copilot image generation platform. Suggested OER search string: “word frequency lists corpus linguistics OER”.
- *Table 4.1 Examples of corpus-informed grammar teaching* Attribution: AI-generated using prompt “Generate a table showing examples of corpus-informed grammar teaching with evidence and implications” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus-informed grammar teaching examples OER”.
- *Box 4.1 Benefits of corpus-informed teaching materials* Attribution: AI-generated using prompt “Generate a table listing benefits of corpus-informed teaching materials” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus-informed teaching materials benefits OER”.
- *Table 4.2 Common errors identified in learner corpora* Attribution: AI-generated using prompt “Generate a table showing common errors in learner corpora with examples and



teaching implications” on Microsoft Copilot image generation platform. Suggested OER search string: “learner corpora error analysis examples OER”.

- *Figure 4.2 Learner engaging in self-access corpus learning* Attribution: AI-generated using prompt “Generate a diagram showing a learner using an online corpus tool to explore word usage” on Microsoft Copilot image generation platform. Suggested OER search string: “self-access learning corpus linguistics OER”.
- *Box 4.2 Strategies for autonomous learning with corpora* Attribution: AI-generated using prompt “Generate a table listing strategies for autonomous learning with corpora” on Microsoft Copilot image generation platform. Suggested OER search string: “autonomous learning strategies corpus linguistics OER”.
- *Figure 4.3 Corpus-informed curriculum development process* Attribution: AI-generated using prompt “Generate a diagram showing curriculum design stages with corpus evidence integrated at each stage” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus-informed curriculum development OER”.
- *Table 4.3 Examples of corpus-informed text selection* Attribution: AI-generated using prompt “Generate a table showing examples of corpus-informed text and genre selection with teaching implications” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus-informed text selection syllabus OER”.
- *Box 4.3 Benefits of aligning syllabus content with authentic language use* Attribution: AI-generated using prompt “Generate a table listing benefits of aligning syllabus content with authentic language use” on Microsoft Copilot image generation platform. Suggested OER search string: “syllabus design corpus linguistics OER”.
- *Figure 4.4 Corpus-based test design process* Attribution: AI-generated using prompt “Generate a diagram showing the process of designing a test using corpus evidence” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus-based test design language assessment OER”.
- *Table 4.4 Examples of learner performance analysis using corpora* Attribution: AI-generated using prompt “Generate a table showing examples of learner performance analysis using corpora” on Microsoft Copilot image generation platform. Suggested OER search string: “learner performance analysis corpus linguistics OER”.
- *Box 4.4 Benefits of corpus-informed assessment* Attribution: AI-generated using prompt “Generate a table listing benefits of corpus-informed assessment” on Microsoft Copilot image generation platform. Suggested OER search string: “validity reliability corpus-based assessment OER”.

Course code: LIN 307

Course title: New and Created Languages

Course credit load: 2 Units



Series 1: Foundations of Digital and Corpus Linguistics

Series 2: Perspectives and Approaches in Corpus-Based Research

Series 3: Levels of Linguistic Analysis in Corpus Studies

Series 4: Applications of Corpus Linguistics in Language and Education

Series 5: Tools, Techniques, and Skills for Corpus Analysis

Proposed Outline for Series 5

Part 1: Tools for corpus analysis

- **Sub-Part 5.1.1:** Overview of corpus software and platforms
- **Sub-Part 5.1.2:** Concordancers and frequency tools
- **Sub-Part 5.1.3:** Annotation and tagging tools

Part 2: Techniques for corpus analysis

- **Sub-Part 5.2.1:** Concordance analysis and keyword extraction
- **Sub-Part 5.2.2:** Collocation and n-gram analysis
- **Sub-Part 5.2.3:** Semantic and discourse analysis techniques

Part 3: Skills for effective corpus analysis

- **Sub-Part 5.3.1:** Designing and compiling a corpus
- **Sub-Part 5.3.2:** Cleaning and preparing data for analysis
- **Sub-Part 5.3.3:** Interpreting and presenting corpus findings

Part 4: Advanced applications and integration

- **Sub-Part 5.4.1:** Corpus analysis in computational linguistics
- **Sub-Part 5.4.2:** Integrating corpus tools with language teaching and research
- **Sub-Part 5.4.3:** Ethical considerations in corpus analysis

Introduction

In the last Series, we explored how corpus linguistics can be applied in language teaching, learning, curriculum design, and assessment. That knowledge now sets the stage for a more practical focus: the tools, techniques, and skills needed to carry out corpus analysis effectively. This Series is particularly important because it equips you with the hands-on competencies required to move from theory into practice, enabling you to work confidently with corpora in both academic and professional contexts.



The aim of this Series is to introduce you to the essential tools and techniques for corpus analysis and to develop the practical skills needed to design, compile, and interpret corpora. By the end, you will be able to apply these methods to real-world language data and present your findings with clarity.

We shall commence this Series by examining the tools available for corpus analysis, including software platforms, concordancers, and annotation tools. Next, we will move on to techniques such as concordance analysis, keyword extraction, collocation studies, and discourse analysis. Following that, we will focus on the skills required to design and compile a corpus, prepare data for analysis, and interpret results meaningfully. Finally, we will wrap up by considering advanced applications, including integration with computational linguistics, classroom practice, and ethical considerations. In conclusion, this Series will round off your training by ensuring you are well-prepared to carry out corpus analysis with both technical competence and critical awareness.

Series Learning Outcomes (Series 5: Corpus Analysis – Tools and Techniques)

Upon completing this Series, you should be able to:

- **SLO 5.1** describe the main software platforms and tools used for corpus analysis,
- **SLO 5.2** demonstrate how to use concordancers and frequency tools to explore language data,
- **SLO 5.3** apply annotation and tagging tools to mark grammatical or semantic features in a corpus,
- **SLO 5.4** analyse concordance lines and extract keywords to identify usage patterns,
- **SLO 5.5** evaluate collocation and n-gram analysis for studying word combinations,
- **SLO 5.6** apply techniques of semantic and discourse analysis to corpus data,
- **SLO 5.7** design and compile a small corpus for a specific language learning or research purpose,
- **SLO 5.8** prepare and clean corpus data to ensure accuracy and usability in analysis,
- **SLO 5.9** interpret corpus findings and present them clearly in written or visual formats,
- **SLO 5.10** explain advanced applications of corpus analysis in computational linguistics,
- **SLO 5.11** integrate corpus tools into language teaching or research practice, and
- **SLO 5.12** evaluate ethical considerations when collecting, analysing, and sharing corpus data.

5.1 Tools for Corpus Analysis

5.1.1 Overview of corpus software and platforms

Corpus software refers to specialised computer programmes designed to store, search, and analyse large collections of texts. Popular platforms include AntConc, Sketch Engine, and WordSmith Tools. These tools allow researchers and teachers to explore authentic language data, identify patterns, and generate useful statistics.



For example, AntConc is widely used because it is free and user-friendly, while Sketch Engine offers advanced features such as word sketches and large multilingual corpora. Choosing the right platform depends on your research goals, available resources, and level of expertise.

Fill in the blank: AntConc is widely used because it is _____ and user-friendly.

Figure 5.1 Examples of corpus software platforms

Diagram showing logos and interfaces of common corpus software (AntConc, Sketch Engine, WordSmith Tools) AI prompt suggestion: “Generate a diagram showing logos and interfaces of AntConc, Sketch Engine, and WordSmith Tools” Search string: “corpus software platforms AntConc Sketch Engine WordSmith OER”

5.1.2 Concordancers and frequency tools

A **concordancer** is a tool that displays all occurrences of a word or phrase in a corpus, usually in the form of **concordance lines** (short text snippets showing the word in context). This helps learners and researchers see how words are used in authentic sentences.

Frequency tools calculate how often words or phrases occur in a corpus. They are useful for identifying high-frequency vocabulary, comparing usage across genres, and designing teaching materials. Together, concordancers and frequency tools provide a powerful way to study language patterns.

Fill in the blank: A concordancer displays all occurrences of a word or phrase in a corpus as _____ lines.

Tool Type	Example Output	Teaching Implication
Concordancer	A list of "KWIC" (Key Word In Context) lines showing "decision" surrounded by words like <i>make</i> , <i>reached</i> , or <i>tough</i> .	Helps learners identify collocational patterns and the specific prepositions or verbs that accompany a noun.
Frequency Tool	A ranked list showing that "research" and "data" are in the top 1% of words in a specific academic corpus.	Allows teachers to prioritize curriculum by focusing on words with the highest "return on investment" for the learner.

Table 5.1 Examples of concordancer and frequency tool outputs

5.1.3 Annotation and tagging tools



Annotation tools allow researchers to add extra information to texts in a corpus, such as marking parts of speech, semantic categories, or discourse features. **Tagging tools** automatically assign grammatical or semantic labels to words, making it easier to search for patterns.

For example, a tagging tool may label “run” as a verb in one sentence and as a noun in another, depending on its usage. Annotation and tagging are essential for advanced corpus studies, as they provide structured data for deeper analysis.

Fill in the blank: Tagging tools automatically assign _____ or semantic labels to words in a corpus.

Benefit	Description	Teaching & Research Application
Structured Data	Converts "flat" text into a structured database where words are categorized by grammatical role (e.g., Noun, Verb).	Enables complex queries like "find all nouns used as adjectives" or "list all passive verb constructions."
Disambiguation	Distinguishes between identical spellings with different meanings or roles (e.g., <i>lead</i> as a metal vs. <i>lead</i> as a verb).	Helps learners understand how word meaning changes based on syntax and context.
Advanced Pattern Search	Enables searches for abstract grammatical templates like [Modal Verb] + [Adverb] + [Base Verb].	Essential for identifying genre-specific patterns, such as the frequent use of "may + significantly + affect" in research.
Detailed Analysis	Supports the tracking of subtle shifts in language use over time or across different learner proficiency levels.	Allows researchers to create "learner profiles" that show exactly when students begin mastering complex tags.

Box 5.1 Benefits of annotation and tagging tools

Pilot questions 5.1

1. Why is AntConc widely used?
2. What does a concordancer display?
3. Give one example of how frequency tools can be applied in teaching.
4. What do tagging tools automatically assign to words in a corpus?
5. Provide one benefit of annotation tools.

5.2 Techniques for Corpus Analysis



5.2.1 Concordance analysis and keyword extraction

Concordance analysis involves examining all occurrences of a word or phrase in a corpus, displayed as concordance lines. This technique helps learners and researchers see how words are used in different contexts, revealing collocations and usage patterns.

Keyword extraction identifies words that occur significantly more often in one corpus compared to another. For example, keywords in a corpus of academic writing might include “research,” “methodology,” and “analysis,” while a corpus of casual conversation might highlight words like “yeah,” “okay,” and “really.”

Fill in the blank: Keyword extraction identifies words that occur significantly more often in one corpus compared to _____.

Technique	Example Output	Teaching Implication
Concordance Analysis	A "KWIC" (Key Word In Context) display showing "decision" appearing with <i>make</i> , <i>reach</i> , and <i>final</i> .	Helps learners move beyond literal translation to master natural collocations and phrasing.
Keyword Extraction	A list of words (e.g., <i>methodology</i> , <i>data</i> , <i>findings</i>) that appear significantly more often in research papers than in general English.	Allows for targeted ESP (English for Specific Purposes) instruction, focusing on the "vocabulary of the field."

Table 5.2 Examples of concordance analysis and keyword extraction

5.2.2 Collocation and n-gram analysis

Collocation analysis studies how words tend to occur together, such as “strong tea” or “make a decision.” It highlights natural word combinations that learners need to master for fluency.

N-gram analysis examines sequences of words of length n . For example, a 2-gram (bigram) might be “in fact,” while a 3-gram (trigram) could be “on the other hand.” N-gram analysis is useful for identifying common phrases and structures in different genres.

Fill in the blank: N-gram analysis examines sequences of words of length _____.



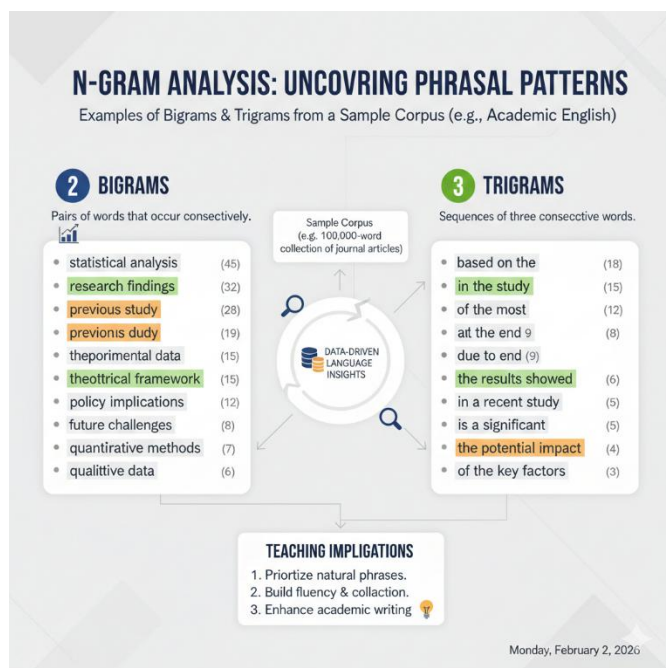


Figure 5.2 Examples of n-gram analysis

5.2.3 Semantic and discourse analysis techniques

Semantic analysis focuses on meaning in language, such as identifying synonyms, antonyms, or thematic fields. Corpus tools can highlight how words are used differently across contexts, supporting translation studies and computational linguistics.

Discourse analysis examines how texts achieve cohesion and coherence, looking at features such as discourse markers, turn-taking, or rhetorical strategies. Corpus-based discourse analysis helps reveal differences between spoken and written communication, or between genres such as academic texts and everyday conversation.

Fill in the blank: Discourse analysis examines how texts achieve _____ and coherence.

Application	Description	Teaching & Research Impact
Translation Studies	Analyzing semantic prosody (the "aura" of a word) to ensure translations carry the same positive/negative weight.	Prevents awkward translations where a word is technically correct but contextually "off."
Spoken vs. Written	Identifying discourse markers like "you know" or "mind you" that appear almost exclusively in speech.	Helps learners adapt their register to sound natural in conversation vs. formal in essays.



Application	Description	Teaching & Research Impact
Rhetorical Strategies	Highlighting how writers use "hedging" (e.g., <i>suggests, possibly</i>) to soften claims in academic papers.	Teaches students how to navigate academic etiquette and avoid sounding overly aggressive in their writing.
Cultural & Social Insights	Examining how different social groups use specific metaphors or labels to describe events.	Reveals ideological biases and social trends embedded within large-scale language data.

Box 5.2 Applications of semantic and discourse analysis

Pilot questions 5.2

1. What does concordance analysis display?
2. What does keyword extraction identify?
3. Give one example of a collocation.
4. What does n-gram analysis examine?
5. What does discourse analysis focus on?

5.3 Skills for Effective Corpus Analysis

5.3.1 Designing and compiling a corpus

Designing a corpus involves deciding what type of texts to include, how large the collection should be, and what purpose it will serve. A **corpus** is a structured collection of texts used for linguistic analysis. Compiling a corpus requires careful planning to ensure it represents the language variety or genre you want to study.

For example, a learner corpus may consist of essays written by students, while a specialised corpus could include medical articles. The design stage ensures that the corpus is fit for its intended purpose.

Fill in the blank: A corpus is a structured collection of _____ used for linguistic analysis.



Corpus Design & Compilation

A Data-Driven Approach



Monday, February 2, 2026

Figure 5.3 Steps in corpus design and compilation

5.3.2 Cleaning and preparing data for analysis

Data cleaning refers to the process of removing errors, inconsistencies, or irrelevant material from a corpus. **Preparing data** involves formatting texts so that they can be processed by corpus tools. This may include removing duplicate entries, correcting spelling errors, or converting files into plain text.

Clean and well-prepared data ensures that analysis results are accurate and reliable. Without this step, frequency counts or collocation patterns may be misleading.

Fill in the blank: Clean and well-prepared data ensures that analysis results are _____ and reliable.

Task	Example	Purpose
Removing duplicates	Deleting repeated sentences or identical articles from different sources.	Prevents inflated frequency counts , ensuring a single text doesn't dominate the data.
Correcting errors	Fixing typos or OCR (Optical Character Recognition) glitches.	Ensures accurate word recognition by the software (e.g., so "the" and "thf" aren't counted separately).



Task	Example	Purpose
Converting formats	Changing PDF or Word files into .txt (UTF-8) plain text.	Makes data compatible with corpus tools , as most concordancers cannot "read" complex styling.

Table 5.3 Examples of data cleaning and preparation tasks

5.3.3 Interpreting and presenting corpus findings

Interpreting corpus findings means making sense of the patterns revealed by tools such as concordancers, frequency lists, and collocation analysis. This involves explaining what the data shows about language use and connecting it to research or teaching goals.

Presenting findings requires clear communication, often through charts, tables, or written reports. Effective presentation ensures that results are understandable to both specialists and non-specialists.

Fill in the blank: Presenting corpus findings often involves using charts, tables, or _____ reports.

Tip	Description	Benefit
Use Visual Aids	Utilize bar charts, frequency tables, and concordancing snapshots.	Makes abstract numerical data scannable and digestible for the audience.
Explain Patterns	Don't just show data; interpret the trends (e.g., "Word X is increasing in use due to Y").	Helps the audience understand the " so what? " behind the statistics.
Connect to Objectives	Link every finding back to your original research question or teaching goal.	Ensures the presentation remains focused and relevant to the stakeholders' needs.
Avoid Overgeneralizing	Explicitly state the limitations of your corpus (size, genre, time period).	Maintains academic integrity and prevents false assumptions about the whole language.

Box 5.3 Tips for presenting corpus findings

Pilot questions 5.3

1. What is a corpus?
2. Give one example of a specialised corpus.



3. Why is data cleaning important?
4. Provide one example of a data preparation task.
5. What is one effective way to present corpus findings?

5.4 Advanced Applications and Integration

5.4.1 Corpus analysis in computational linguistics

Computational linguistics is the study of language using computer-based methods, often involving large-scale corpora. Corpus analysis plays a central role here, as it provides the data needed to train and evaluate natural language processing (NLP) systems. For example, corpora are used to develop machine translation tools, speech recognition systems, and sentiment analysis applications.

By applying corpus techniques in computational linguistics, learners can see how linguistic theory connects with real-world technologies.

Fill in the blank: Corpus analysis provides the data needed to train and evaluate _____ systems.



Figure 5.4 Corpus analysis in computational linguistics

5.4.2 Integrating corpus tools with language teaching and research

Corpus tools can be integrated into both classroom practice and academic research. In teaching, they provide authentic examples of vocabulary, grammar, and discourse, helping learners engage with real language use. In research, they support evidence-based studies of language variation, change, and acquisition.



For instance, a teacher may use concordance lines to demonstrate collocations in class, while a researcher may use keyword analysis to compare political speeches across decades.

Fill in the blank: Corpus tools provide authentic examples of _____, grammar, and discourse in teaching contexts.

Context	Tool Used	Example Application	Purpose
Classroom Teaching	Concordancer (e.g., AntConc, Sketch Engine)	Displaying KWIC (Key Word In Context) lines for "responsibility."	To help learners discover natural collocations like <i>"take responsibility"</i> or <i>"heavy responsibility."</i>
Academic Research	Keyword Analysis	Comparing high-frequency terms in political speeches across different decades.	To identify shifts in political rhetoric, themes, and ideological priorities over time.
Syllabus Design	Frequency Profilers	Analyzing a collection of engineering textbooks.	To ensure the vocabulary taught in an ESP (English for Specific Purposes) course matches the actual needs of students.
Lexicography	Collocation Networks	Mapping the "lexical neighbors" of the word <i>"climate."</i>	To create more accurate and modern dictionary entries based on current usage patterns.

Table 5.4 Examples of corpus tool integration in teaching and research

5.4.3 Ethical considerations in corpus analysis

Ethical considerations are crucial when collecting, analysing, and sharing corpus data. Issues include protecting privacy, obtaining consent, and ensuring that data is not misused. For learner corpora, it is especially important to anonymise texts so that individual students cannot be identified.

Ethics also extend to how findings are presented, avoiding bias or misrepresentation. Responsible corpus analysis ensures that research and teaching benefit learners and society without causing harm.

Fill in the blank: Ethical considerations in corpus analysis include protecting privacy and obtaining _____.



Ethical Issue	Description	Mitigation Strategy
Participant Privacy	Protecting the identity of individuals whose language is captured in the corpus.	Use strict data-handling protocols and limit access to sensitive metadata.
Informed Consent	Ensuring participants understand how their data will be used, stored, and shared.	Provide clear, translated consent forms and the right to withdraw at any time.
Anonymization	Removing or masking "PII" (Personally Identifiable Information) like names and locations.	Use automated or manual de-identification (e.g., replacing a name with <i>[NAME]</i>).
Bias & Representation	Avoiding the marginalization of certain groups or the over-representation of others.	Use balanced sampling techniques and critically evaluate the corpus's demographics.

Box 5.4 Key ethical considerations in corpus analysis

Pilot questions 5.4

1. What role does corpus analysis play in computational linguistics?
2. Give one example of how corpus tools can be used in teaching.
3. Provide one example of how corpus tools can be used in research.
4. What is one ethical consideration when working with learner corpora?
5. Why is it important to anonymise learner data?

Series Summary

This Series has been designed to equip you with the practical tools, techniques, and skills needed to carry out effective corpus analysis. Its relevance lies in bridging theoretical knowledge with hands-on practice, enabling you to explore authentic language data, prepare it for study, and apply findings in both teaching and research contexts.

We began by introducing the main tools for corpus analysis, including software platforms, concordancers, frequency tools, and annotation systems. Then we explored techniques such as concordance analysis, keyword extraction, collocation studies, n-gram analysis, and discourse approaches. Following that, we focused on essential skills for designing and compiling a corpus, cleaning and preparing data, and interpreting and presenting findings. Finally, we examined advanced applications, including computational linguistics, integration with teaching and research, and ethical considerations. In line with the Series Learning Outcomes, you should now be able to



describe, demonstrate, apply, analyse, design, and evaluate corpus methods with confidence, ensuring your work is both technically competent and ethically sound.

Glossary of Terms

Annotation tools: Software used to add extra information to texts in a corpus, such as marking parts of speech, semantic categories, or discourse features.

Collocation analysis: A technique that studies how words tend to occur together, for example “make a decision” or “strong tea.”

Concordance analysis: The examination of all occurrences of a word or phrase in a corpus, displayed as concordance lines to show usage in context.

Concordancer: A tool that displays all instances of a word or phrase in a corpus, usually in the form of short text snippets called concordance lines.

Corpus: A structured collection of texts used for linguistic analysis.

Corpus software: Specialised computer programmes designed to store, search, and analyse large collections of texts, such as AntConc or Sketch Engine.

Data cleaning: The process of removing errors, inconsistencies, or irrelevant material from a corpus to ensure accurate results.

Discourse analysis: The study of how texts achieve cohesion and coherence, focusing on features such as discourse markers, turn-taking, or rhetorical strategies.

Ethical considerations: Issues related to privacy, consent, and responsible use of data when collecting, analysing, and sharing corpora.

Frequency tools: Tools that calculate how often words or phrases occur in a corpus, useful for identifying high-frequency vocabulary.

Keyword extraction: A technique that identifies words occurring significantly more often in one corpus compared to another, highlighting distinctive vocabulary.

N-gram analysis: The study of sequences of words of length n , such as bigrams (“in fact”) or trigrams (“on the other hand”).

Semantic analysis: The study of meaning in language, including synonyms, antonyms, and thematic fields, often supported by corpus tools.



Tagging tools: Tools that automatically assign grammatical or semantic labels to words in a corpus, distinguishing different uses of the same word.

Text preparation: The process of formatting texts so they can be processed by corpus tools, such as converting files into plain text.

Concept Check Questions [Series 5]

Objective Questions

1. Which corpus software is widely used because it is free and user-friendly? (Linked to SLO 5.1)
2. What does a concordancer display in a corpus? (Linked to SLO 5.2)
3. N-gram analysis examines sequences of words of what length? (Linked to SLO 5.5)
4. What is a corpus? (Linked to SLO 5.7)
5. Ethical considerations in corpus analysis include protecting privacy and obtaining what? (Linked to SLO 5.12)

Short-Answer Questions

6. Define annotation tools in one sentence. (Linked to SLO 5.3)
7. Give one example of a collocation. (Linked to SLO 5.5)
8. Explain one reason why data cleaning is important in corpus analysis. (Linked to SLO 5.8)
9. Provide one example of how corpus tools can be integrated into classroom teaching. (Linked to SLO 5.11)
10. Identify one benefit of presenting corpus findings with charts or tables. (Linked to SLO 5.9)

Long-Answer Questions

11. Analyse the strengths and limitations of using concordance analysis and keyword extraction in corpus studies. (Linked to SLO 5.4)
12. Evaluate the usefulness of collocation and n-gram analysis for improving language teaching and learning. (Linked to SLO 5.5)
13. Assess the importance of designing and compiling a corpus carefully for research purposes. (Linked to SLO 5.7)
14. Discuss the role of corpus analysis in computational linguistics, giving examples of applications. (Linked to SLO 5.10)
15. Examine the ethical issues involved in corpus analysis and suggest ways to address them responsibly. (Linked to SLO 5.12)

Answers to Concept Check Questions [Series 5]

Objective Questions



1. AntConc.
2. Concordance lines showing the word or phrase in context.
3. Length n (e.g., bigrams, trigrams).
4. A structured collection of texts used for linguistic analysis.
5. Informed consent.

Short-Answer Questions

6. Annotation tools are software used to add extra information to texts, such as parts of speech or semantic categories.
7. Example: “make a decision.”
8. It ensures that analysis results are accurate and reliable.
9. Example: Using concordance lines to demonstrate collocations in class.
10. They make findings clearer and easier to understand for both specialists and non-specialists.

Long-Answer Questions

11. *Basic response:* Concordance analysis shows usage patterns, while keyword extraction highlights distinctive vocabulary. Both are useful but may overlook context or subtle meanings. *Value-added response:* They provide powerful insights into genre differences and learner language, though researchers must combine them with qualitative analysis to avoid misinterpretation.
12. *Basic response:* Collocation and n-gram analysis reveal natural word combinations and common phrases. *Value-added response:* They enhance fluency and accuracy in teaching, but should be balanced with attention to less frequent but contextually important expressions.
13. *Basic response:* Careful design ensures the corpus represents the language variety or genre being studied. *Value-added response:* Poorly designed corpora can lead to misleading results; therefore, sampling, balance, and representativeness are critical for valid research outcomes.
14. *Basic response:* Corpus analysis supports computational linguistics applications such as machine translation and speech recognition. *Value-added response:* It also underpins sentiment analysis, chatbots, and AI language models, showing how linguistic data drives technological innovation.
15. *Basic response:* Ethical issues include privacy, consent, and anonymisation of learner data. *Value-added response:* Addressing these requires clear policies, transparency, and responsible dissemination of findings to avoid bias or harm.

Answers to Pilot Questions [Series 5]

Pilot questions 5.1

1. Because it is free and user-friendly.
2. Concordance lines.
3. Identifying high-frequency vocabulary for teaching priorities.
4. Grammatical or semantic labels.
5. They provide structured data for deeper analysis.



Pilot questions 5.2

1. All occurrences of a word or phrase in context.
2. Words occurring significantly more often in one corpus compared to another.
3. Example: “make a decision.”
4. Sequences of words of length n .
5. Cohesion and coherence in texts.

Pilot questions 5.3

1. A structured collection of texts used for linguistic analysis.
2. Example: A specialised corpus of medical articles.
3. To ensure accuracy and reliability of results.
4. Converting PDF files into plain text.
5. Using charts, tables, or written reports.

Pilot questions 5.4

1. It provides data to train and evaluate NLP systems.
2. Using concordance lines to show collocations in class.
3. Using keyword analysis to compare political speeches.
4. Protecting privacy and anonymising learner data.
5. To ensure individuals cannot be identified.

References and Attributions [Series 5]

This section compiles all references and illustration attributions for **Series 5: Tools, Techniques, and Skills for Corpus Analysis**, following APA style for clarity and consistency.

General References

- Biber, D., Conrad, S., & Reppen, S. (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.
- British National Corpus. (n.d.). Retrieved from <https://www.english-corpora.org/bnc/>
- AntConc. (n.d.). Retrieved from <https://www.laurenceanthony.net/software/antconc/>
- Sketch Engine. (n.d.). Retrieved from <https://www.sketchengine.eu/>

Illustrations

- *Figure 5.1 Examples of corpus software platforms* Attribution: AI-generated using prompt “Generate a diagram showing logos and interfaces of AntConc, Sketch Engine, and WordSmith Tools” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus software platforms AntConc Sketch Engine WordSmith OER”.



- *Table 5.1 Examples of concordancer and frequency tool outputs* Attribution: AI-generated using prompt “Generate a table showing examples of concordancer and frequency tool outputs with teaching implications” on Microsoft Copilot image generation platform. Suggested OER search string: “concordancer frequency tools corpus linguistics OER”.
- *Box 5.1 Benefits of annotation and tagging tools* Attribution: AI-generated using prompt “Generate a table listing benefits of annotation and tagging tools in corpus linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “annotation tagging tools corpus linguistics OER”.
- *Table 5.2 Examples of concordance analysis and keyword extraction* Attribution: AI-generated using prompt “Generate a table showing examples of concordance analysis and keyword extraction with teaching implications” on Microsoft Copilot image generation platform. Suggested OER search string: “concordance analysis keyword extraction corpus linguistics OER”.
- *Figure 5.2 Examples of n-gram analysis* Attribution: AI-generated using prompt “Generate a diagram showing examples of bigrams and trigrams from a sample corpus” on Microsoft Copilot image generation platform. Suggested OER search string: “n-gram analysis corpus linguistics OER”.
- *Box 5.2 Applications of semantic and discourse analysis* Attribution: AI-generated using prompt “Generate a table listing applications of semantic and discourse analysis in corpus linguistics” on Microsoft Copilot image generation platform. Suggested OER search string: “semantic discourse analysis corpus linguistics OER”.
- *Figure 5.3 Steps in corpus design and compilation* Attribution: AI-generated using prompt “Generate a diagram showing the steps in designing and compiling a corpus” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus design and compilation steps OER”.
- *Table 5.3 Examples of data cleaning and preparation tasks* Attribution: AI-generated using prompt “Generate a table showing examples of data cleaning and preparation tasks with purposes” on Microsoft Copilot image generation platform. Suggested OER search string: “data cleaning corpus linguistics examples OER”.
- *Box 5.3 Tips for presenting corpus findings* Attribution: AI-generated using prompt “Generate a table listing tips for presenting corpus findings clearly” on Microsoft Copilot image generation platform. Suggested OER search string: “presenting corpus findings linguistics OER”.
- *Figure 5.4 Corpus analysis in computational linguistics* Attribution: AI-generated using prompt “Generate a diagram showing the link between corpus analysis and NLP applications such as machine translation and speech recognition” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus analysis computational linguistics NLP OER”.
- *Table 5.4 Examples of corpus tool integration in teaching and research* Attribution: AI-generated using prompt “Generate a table showing examples of corpus tool integration in



teaching and research” on Microsoft Copilot image generation platform. Suggested OER search string: “corpus tools language teaching research OER”.

- *Box 5.4 Key ethical considerations in corpus analysis* Attribution: AI-generated using prompt “Generate a table listing key ethical considerations in corpus analysis” on Microsoft Copilot image generation platform. Suggested OER search string: “ethical considerations corpus linguistics OER”.

