British National Corpus (BNC)

Topic-wise study notes | Linguistics and Natural Language Processing

Composition, spoken versus written text, corpus size and annotation, and use in linguistic and NLP research. Includes definitions, characteristics, architectures, worked examples, and applications.

1. Definition

The **British National Corpus**, commonly called BNC1994, is a large electronic collection of sampled British English from the late twentieth century. It contains written texts and orthographic transcripts of speech, selected from many sources [1].

Corpus: a systematically collected body of language data that can be searched and analysed. A corpus is evidence of observed language use, rather than a dictionary or a list of grammar rules.

Core characteristics

• **General-purpose:** supports many linguistic questions.

• **Sampled:** contains selected passages and recordings rather than all British English.

• **Broad coverage:** includes multiple written genres and spoken settings.

• **British English:** its findings are tied to its variety and period.

• **Annotated:** includes linguistic labels, structure, and metadata.

• **Finite historical resource:** the original corpus does not continuously update itself.

Scope and versions

The project was built during 1991-1994 [2]. These notes focus on the original BNC, including its XML representation. **BNC2014 is a separate later project**; do not transfer BNC1994 size and composition figures to it.

Learning objectives

Explain the corpus composition, compare speech and writing, interpret annotation and metadata, and design a reproducible corpus study with normalized counts.

2. Composition of the BNC

Definition

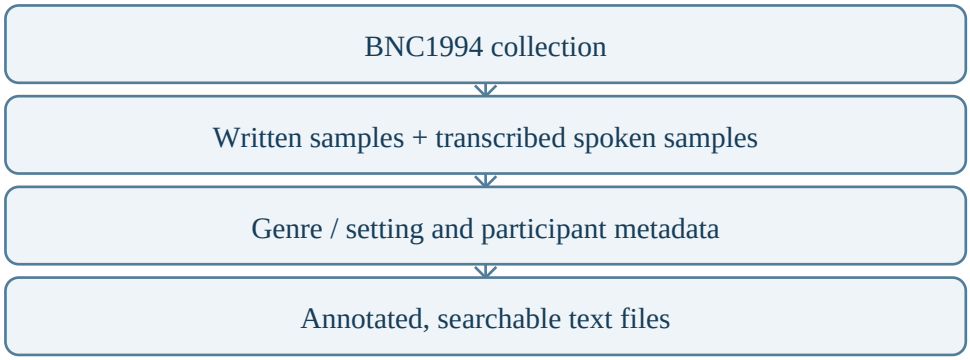
Corpus composition describes the kinds of language material included and their sampling proportions. The BNC is conventionally described as approximately 90% written and 10% spoken, by word volume [3].

Component	Approximate share	Typical material
Written	90%	Fiction, newspapers, academic and other prose, letters, essays, and other documents.
Spoken	10%	Everyday conversations and speech from varied institutional or public settings.

Characteristics of composition

- Written material dominates the total word count.
- Written selection considers factors such as domain, medium, and time.
- Spoken selection includes demographic conversation and context-governed speech.
- Sampling breadth aims to reduce dependence on a single genre.
- Balance does not mean equal amounts from every category.

Conceptual architecture



The component box combines two branches conceptually; this diagram is a resource structure, not a neural network architecture.

Step-by-step example: choose a subcorpus

- **1:** Ask how a word is used in newspaper language.
- **2:** Filter texts using written-news metadata.
- **3:** Search within the selected texts.
- **4:** Inspect contexts and compute frequencies.
- **5:** Report that findings concern this subcorpus, not the entire language.

Applications

Composition enables genre comparison, register studies, representative sampling exercises, and careful selection of data for NLP experiments.

3. Spoken versus Written Text

Definitions

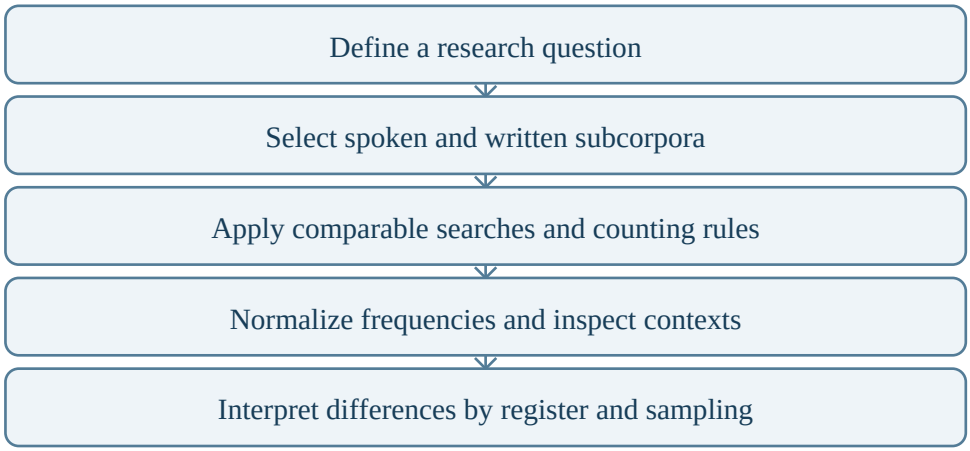
Written text: language originally produced in writing.
Spoken text: language produced through speech and represented in the corpus by transcription. Orthographic transcription primarily records words using conventional spelling; it is not a complete phonetic record.

Typical characteristics

Aspect	Spoken material	Written material
Production	Often immediate and interactive.	Often planned or edited.
Structure	May include fragments, repairs, and repetitions.	Often has organized sentences and paragraphs.
Context	Often depends on speakers and the situation.	Often provides more explicit context for readers.
Corpus structure	Speaker turns and utterances.	Paragraphs, headings, and other document structure.
Metadata	May include speaker and recording information.	May include source, genre, author, and publication information.

These are tendencies, not absolute rules: a formal speech may be planned, while an informal letter may be conversational.

Architecture for comparison



Step-by-step example

To compare I think: (1) select the two components; (2) search the phrase; (3) inspect concordances; (4) count occurrences and words; (5) calculate frequencies per million words; (6) interpret the discourse function in each setting.

Applications

Useful for discourse-marker studies, register comparison, language teaching, and testing how NLP models handle conversational versus edited language.

4. Corpus Size and Annotation

Corpus size: definition and figures

Corpus size is the amount of language data, commonly measured in words or tokens. BNC1994 is usually called a **100-million-word corpus** and comprises over 4,000 text files [1,4]. Exact counts vary with edition and counting units.

The XML reference guide reports **96,986,707 orthographic words** and **98,363,783 POS-tagged w-units** [5]. A word and an annotation unit are not necessarily identical, so reports should state what was counted.

Annotation: definition

Annotation adds structured information to language data. In the BNC, this includes word-class information, text structure, and descriptive metadata. The XML edition represents these using TEI-based encoding [2,6].

Layer	What it supplies
Token and sentence structure	Units for search, counting, and alignment.
POS annotation	Word-class labels produced using CLAWS tagging.
Headword information	Normalized lexical forms in the XML representation.
Document structure	Paragraphs, headings, or spoken utterance structure.
Metadata	Information about sources, text categories, speakers, and recording contexts where available.

Characteristics and cautions

• Machine-readable labels support more precise queries.

• Metadata enables principled subcorpus selection.

• Automatic linguistic annotation can contain errors.

• The standard corpus is not universally annotated with full parse trees, semantic roles, or coreference links.

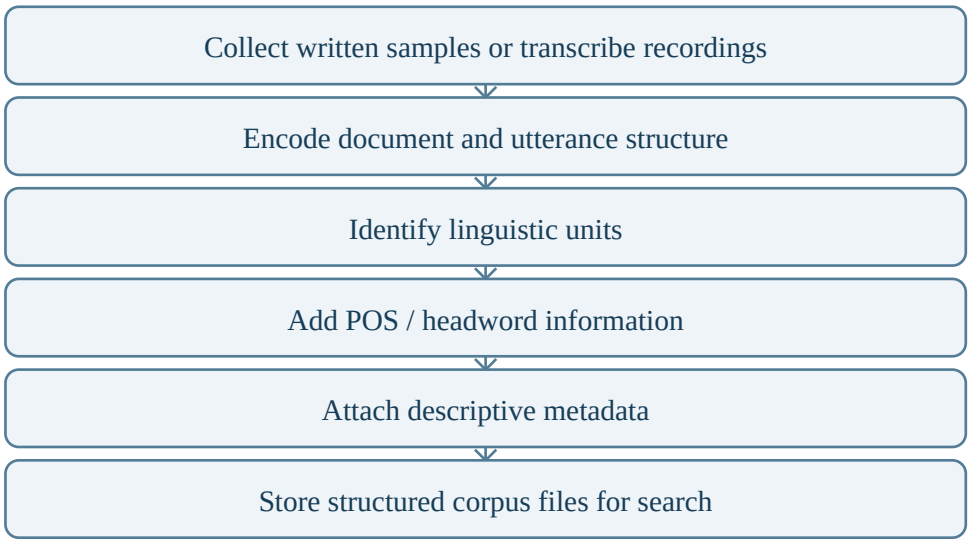
• Tagsets differ: BNC tags should not be treated as Penn Treebank tags.

Applications

Annotated data supports POS-based searches, lemma frequencies, register studies, and evaluation of language-processing tools with suitable validation.

4.1 Annotation Architecture and Example

Conceptual annotation workflow



Step-by-step teaching example

Sentence: “The students studied carefully.” This is an invented example, not a retrieved BNC quotation.

Token	Example C5 tag	Headword
The	AT0: article	the
students	NN2: plural noun	student
studied	VVD: past-tense lexical verb	study
carefully	AV0: general adverb	carefully

• **1:** Segment the sentence into tokens.

• **2:** Assign a word-class label to each token.

• **3:** Store headword information where provided.

• **4:** Link tokens to their sentence and text identifier.

• **5:** Link the text to genre and source metadata.

• **6:** Search all NN2 tokens or all forms with headword study.

What a headword search changes

An exact word search for studied finds that form. A headword search for study can group multiple inflected forms. Researchers should state whether their query uses surface forms, headwords, tags, or combinations.

A corpus query tool may use its own query syntax or tag mapping. The table illustrates annotation concepts and is not a promised XML serialization or tool command.

5. Linguistic and NLP Research

Definition

Corpus-based research investigates language using systematically collected examples and counts. Linguistic research studies usage patterns; NLP research uses the data to develop or assess computational methods.

Research characteristics

• Uses attested language rather than intuition alone.

• Combines quantitative counts with qualitative context inspection.

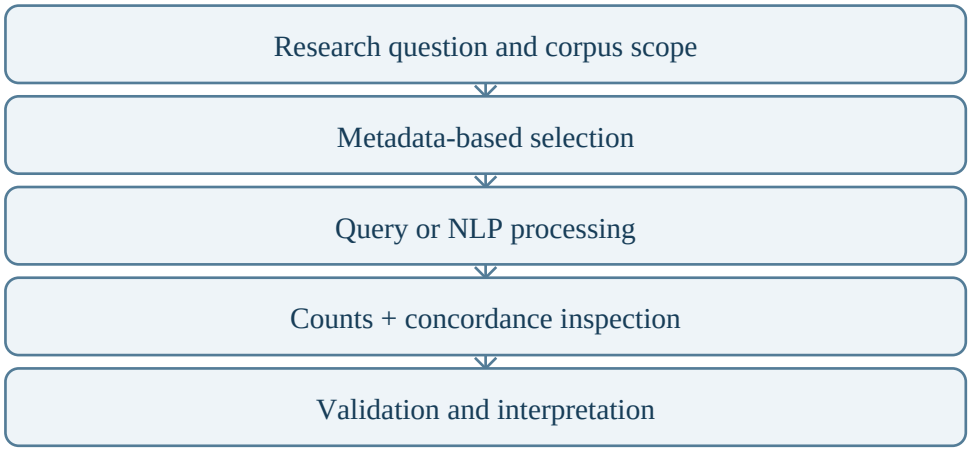
• Requires explicit corpus scope, query, and counting rules.

• Benefits from metadata and linguistic annotation.

• Needs caution about historical coverage and sampling bias.

Application	Illustrative research question
Lexicography	Which meanings and constructions occur with a word?
Collocation	Which words tend to occur near strong or powerful?
Grammar	How do grammatical constructions differ across genres?
Sociolinguistics	How do sampled speaker groups differ, when metadata supports comparison?
Language teaching	Which vocabulary and expressions occur in relevant registers?
POS tagging / lemmatization	How does a model handle selected corpus passages?
Language modeling	What lexical or n-gram patterns appear in this historical sample?
Domain / register robustness	How does performance differ across spoken and written material?

Research architecture



For evaluation, automatic BNC tags are useful reference material but not infallible gold labels. Manually check a suitable sample when correctness matters. Avoid training and testing on overlapping texts.

6. Worked Research Example

Question

Is I think more frequent in a selected spoken subcorpus than in a selected written subcorpus? All counts below are **invented for calculation practice**, not actual BNC findings.

Step 1: Define the query

Search consecutive word tokens I + think, ignoring case. Specify how punctuation, utterance boundaries, and tokenization are treated.

Step 2: Select and count

Subcorpus	Words	Matches
Spoken sample	1,000,000	800
Written sample	4,000,000	400

Step 3: Normalize

Frequency per million = (matches / total words) × 1,000,000.

• Spoken: (800 / 1,000,000) × 1,000,000 = **800 per million**.

• Written: (400 / 4,000,000) × 1,000,000 = **100 per million**.

• Rate ratio: 800 / 100 = **8** in this invented example.

Step 4: Inspect contexts

Read a sample of matching lines. Distinguish literal reports of thinking from stance expressions, such as “I think we should leave”. Check transcription and query artefacts.

Step 5: Interpret carefully

The invented samples show a higher normalized rate in speech. A real study should examine text-level dispersion, speaker effects, and uncertainty before generalizing. Raw counts alone are misleading when subcorpora differ in size.

KWIC: key word in context

A concordance displays each match with nearby language. For example, “Well, **I think** we should leave now.” This invented line illustrates how context reveals usage; it is not a BNC quotation.

7. Limitations, Revision, and References

Limitations

• Historical coverage does not fully represent present-day vocabulary or online communication.

• British English findings do not automatically generalize to all English varieties.

• Approximately 90/10 composition reflects design and practical constraints, not equal representation.

• Transcripts omit some acoustic and interactional detail.

• Annotation errors and uneven metadata require checking.

• A pattern may be concentrated in a few texts; inspect dispersion.

Practice questions and answer hints

1. Define BNC. A sampled electronic corpus of late-twentieth-century British English.
2. What is its approximate composition? 90% written, 10% transcribed spoken language.
3. Why normalize counts? Subcorpora may contain different numbers of words.
4. What does POS annotation provide? Grammatical word-class labels.
5. Is BNC1994 the same as BNC2014? No, they are distinct resources.
6. Is a corpus automatically representative of every language use? No; representativeness depends on design and the question asked.

Source references

[1] Oxford BNC project homepage.
[2] About the British National Corpus.
[3] BNC products and composition.
[4] BNC XML edition.
[5] Reference guide: corpus design and counts.
[6] BNC XML user reference guide.

Architectures and worked examples are teaching illustrations. Numerical research results are explicitly hypothetical. Consult the selected corpus edition and tool documentation for exact encodings, counts, and query syntax.