Applied NLP

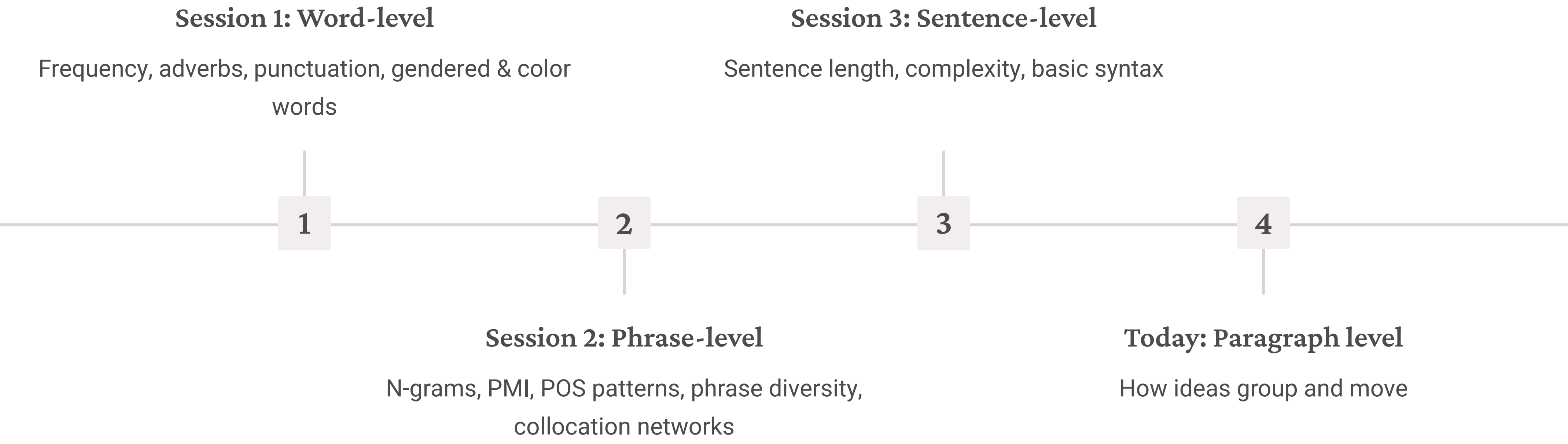
Session 4

Lecturer: Narges Chinichian

Winter Semester 2025-2026

Course Progress

From Words to Whole Texts



1

Session 1: Word-level

Frequency, adverbs, punctuation, gendered & color words

2

Session 2: Phrase-level

N-grams, PMI, POS patterns, phrase diversity, collocation networks

3

Session 3: Sentence-level

Sentence length, complexity, basic syntax

4

Today: Paragraph level

How ideas group and move

Today's Measures

1 Measure 1 – Paragraph Coherence

2 Measure 2 – Topic Drift Between Paragraphs

3 Measure 3 – Discourse Marker Density

4 Measure 4 – Paragraph–Summary Similarity

5 Measure 5 – Paragraph Function Classification

Paragraphs as Units of Thought

Writing Tradition

A paragraph ≈ one **main idea** in many writing traditions

In literature, this is not always strict, but still:

- Paragraphs cluster **sentences** that "belong together"
- Authors use paragraph breaks for **scene changes**, **focus shifts**, **time jumps**

For LLMs and agents

Paragraphs are natural **chunks** for retrieval and context windows

Measure 1

Paragraph Semantic Coherence

Measure 1: Intuition

Question: How much do the sentences inside a paragraph "talk about the same thing"?

If sentences in a paragraph share many content words → **high coherence**

If they talk about different things → **low coherence**

Our method for quantifying semantic coherence uses an embedding-based approach:

1. Use **sentence embeddings** (e.g., MiniLM) for each sentence in the paragraph
2. Compute a **paragraph centroid embedding** by averaging all sentence embeddings
3. Calculate the **cosine similarity** between each sentence embedding and the paragraph centroid
4. Average these similarities to obtain the **final coherence score** for the paragraph

Measure 1: Interpretation

Higher scores

- Sentences are semantically similar to paragraph centroid
- Focused idea with consistent semantic content
- Sentences reinforce the same concept
- Often found in reflective, descriptive, or emotional paragraphs

Lower scores

- Many semantic shifts within the paragraph
- Mixed topics or concepts
- Could indicate digressions, fast-moving scenes, or complex narrative transitions

Measure 2

Topic Drift Between Paragraphs

Measure 2: Intuition

Question: How strongly is one paragraph connected to the next one?

Smooth Transition

If two paragraphs are very similar → **smooth transition**

Topic Jump

If they are very different → **topic jump**, new scene, new idea

We use paragraph embeddings to capture semantic topic representation.

This involves computing one embedding for each paragraph and then measuring the cosine similarity between the embeddings of consecutive paragraphs.

Measure 2: Interpretation

High similarity (close to 1)

- Continuous narrative flow
- Smooth thematic transitions
- Stable topic focus
- Gradual development of ideas

Low similarity (close to 0)

- Scene changes or breaks
- Time skips in narrative
- Character perspective changes
- Sudden structural breaks in the text

LLM Connection: Embedding-based versions of this are used to:

- Detect **section boundaries**
- Split long documents for **RAG** and long-context reasoning

Measure 3

Discourse Marker Density

Measure 3: Intuition

Discourse markers are "signposting words" like:

- however, therefore, meanwhile, suddenly, although
- for example, in contrast, at the same time, finally, after all

We count how many such markers appear in each paragraph.

They tell us:

- How the author links ideas
- When they show contrast, cause, elaboration, example

Measure 3: Interpretation

High density paragraphs:

- Often more "**analytical**" or **logical**, with explicit structure
- Might appear in essays, arguments, explanations

Low density paragraphs:

- Flow more implicitly, often **narrative** or **dialogue-heavy**
- Structure is in events, not in explicit connectors

LLM Connection:

LLM-based discourse models look at such markers (plus implicit patterns) to:

- Recognize **contrast**, **cause**, **example**
- Improve summarization, reasoning, and planning

Measure 4

Paragraph–Summary Similarity

Measure 4: Intuition

Question: How easily can a paragraph be compressed into a shorter summary?

Steps:

- Take the longest sentence as a naïve extractive summary
- Embed both paragraph and summary using MiniLM
- Compute cosine similarity between embeddings = compressibility score

This gives a rough "compressibility" score based on the semantic similarity of embeddings.

Measure 4: Interpretation

High similarity:

- Summary shares most important words with the paragraph
- Paragraph is relatively straightforward / focused

Low similarity:

- Long paragraph with many side details, metaphors, or multiple ideas
- Harder to compress without losing information

LLM Connection:

- LLMs constantly perform **compression** to fit long contexts into limited windows
- Paragraphs with low compressibility are "expensive" for LLMs
- Good RAG systems often compress text into dense summaries or vectors

Measure 5

Paragraph Function Classification

Measure 5: Intuition

Different paragraphs **play different roles** in a text:

Dialogue

Action

(movement, events)

Description

(setting, appearance)

Internal monologue

(thoughts, feelings)

Other / mixed

We assign each paragraph a functional label using embeddings based on these categories.

Method:

1. Create short prototype texts for each label
2. Embed prototypes using MiniLM
3. Embed each paragraph
4. Assign label = prototype with highest cosine similarity

Measure 5: Interpretation

Distribution of paragraph types:

- How much of the book is **dialogue** vs **description**?
- Does one book have more **action** scenes?

Structural rhythm:

- Alternation of **dialogue** and **action**
- Long stretches of **description** or **internal monologue**

> ### LLM Connection:
>
> Agentic systems often classify segments as:
>
> - "Instruction", "context", "observation", "plan", "reflection"…
>
> This is the **literary analogue**: we label each chunk by its role in the narrative.

Summary of Key Learnings

Paragraph Coherence

Measured semantic consistency within paragraphs using sentence embeddings and cosine similarity.

Topic Drift

Assessed connections between consecutive paragraphs via embedded similarity, identifying smooth transitions or abrupt jumps.

Discourse Markers

Examined how 'signposting words' influence text structure, indicating analytical vs. narrative styles.

Summary Similarity

Gauged paragraph compressibility by comparing embeddings of paragraphs and their extractive summaries.

Paragraph Function

Classified paragraphs by their narrative roles (dialogue, action, description) using prototype embeddings.

These measures provide a comprehensive toolkit for analyzing text at the paragraph level, enhancing our understanding of narrative structure and coherence.