Applied NLP

Session 3

Lecturer: Narges Chinichian

Winter Semester 2025-2026

Recap

S1: words as data

freq, adverbs, punctuation, gender, color

S2: phrases & collocations

n-grams, PMI, POS patterns, phrase diversity, networks

Today: sentences

who writes long, complex, dialog-heavy prose?

Why sentences?

Sentences are the main carrier of *propositions*

For example, "The cat sat on the mat" vs "Democracy requires informed citizens who actively participate in civic discourse."

Style often shows up in sentence shape (long periodic vs short paratactic)

"He ran. She followed. They stopped." vs "Having considered the matter at length, and weighing all possible consequences, he finally decided."

Readability and target audience are sentence-level phenomena

"See Spot run. Run, Spot, run!" vs "The epistemological implications of poststructuralist hermeneutics necessitate a reconsideration of traditional paradigms."

The 5 measures today

1. Sentence length & distribution
2. Readability indices
3. Sentence embeddings (LLMs as semantic encoders)
4. Clause density / subordination
5. Sentence types & dialogue ratio

Measure 1: Sentence length

Definition

words per sentence, chars per sentence

Output

mean, median, std, histogram

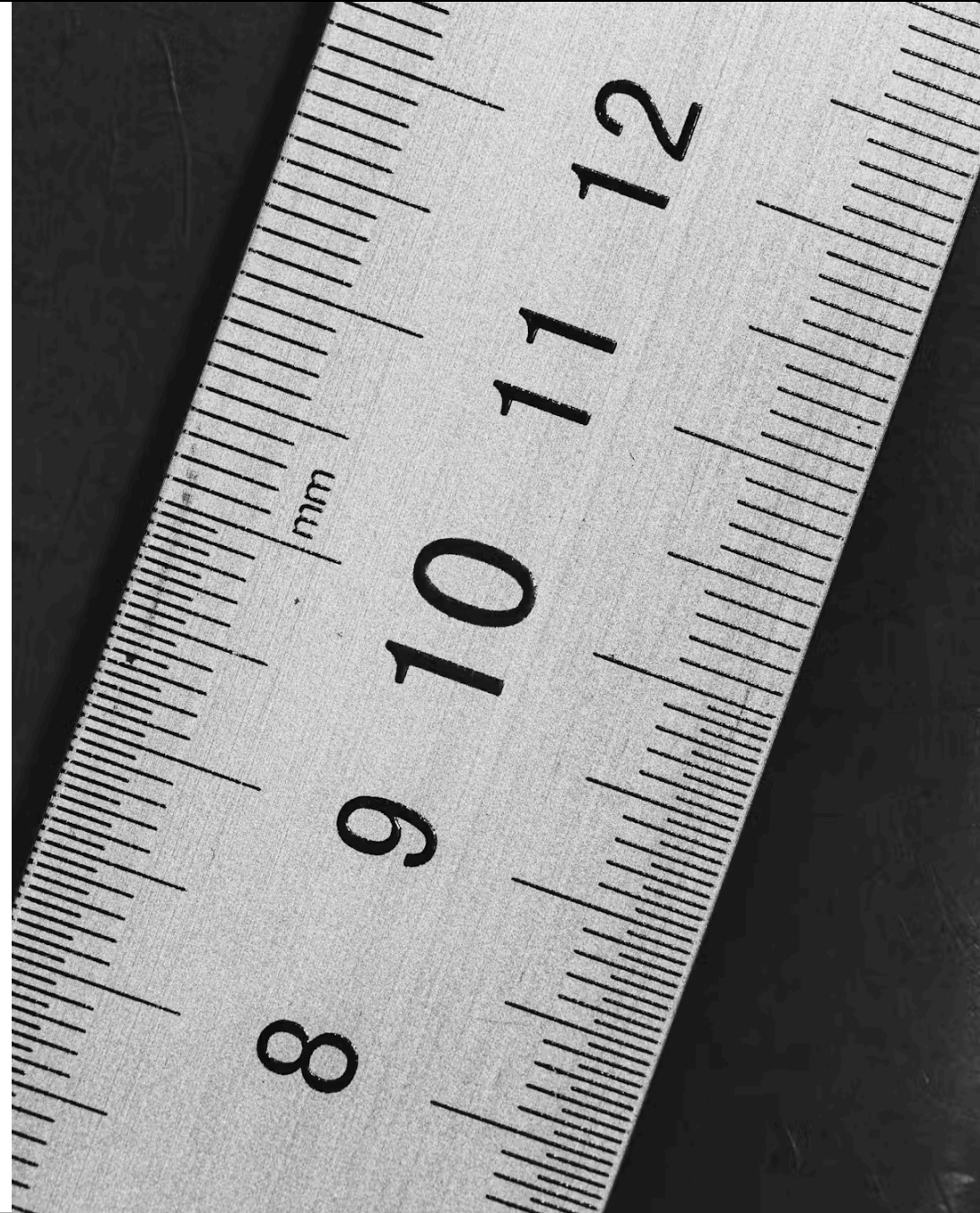
Interpretation

shorter → faster, oral

longer → reflective, academic, 19th c. novels

Task

apply to your author, compare 2 works



Measure 1: Caveats

Important considerations when measuring sentence length

Naive splitting on . ? ! vs NLP sentence splitter

Dialogue can break the pattern

Translations may shorten sentences → note language!

Measure 2: Readability

Example formulas

Flesch Reading Ease:

$$206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

Where ASL = Average Sentence Length (words per sentence) and ASW = Average Syllables per Word. Score ranges:

- 90-100: very easy
- 80-89: easy
- 70-79: fairly easy
- 60-69: standard
- 50-59: fairly difficult
- 30-49: difficult
- 0-29: very difficult

Flesch-Kincaid Grade Level:

$$(0.39 \times ASL) + (11.8 \times ASW) - 15.59$$

This formula gives the US grade level needed to understand the text.

Inputs

sentences, words, syllables

What it tells us

how hard is this text for an average reader?

Note: designed for English

Measure 2: Use in literary NLP

Compare original vs translation

Compare chapter 1 vs chapter 10

Find "difficult" passages automatically

Measure 3: Sentence Embeddings

- Modern NLP uses pretrained transformers to embed entire sentences into vectors
- Each sentence becomes a point in a “meaning space”
- Similar sentences → closer together
- We can visualize or cluster to study style, tone, or content

Measure 3: Example: Sentence Embeddings in Action

Sentence	Top-2 Nearest Neighbors (by cosine similarity)	Similarity score
Alice looked at the cat.	*The cat stared back.*	0.88
	She watched the rabbit run away.	0.82
It was a very strange day.	*Everything felt unusual.*	0.91
	Nothing seemed normal anymore.	0.87
He opened the door and left.	*She stepped outside.*	0.84

What this shows

1 Semantic Grouping

The model groups sentences primarily by their underlying meaning, rather than relying solely on shared keywords or surface-level lexical overlap.

2 Beyond Exact Matches

Sentences like "Alice looked at the cat" and "The cat stared back" are recognized as highly similar, even though their exact wording is different, demonstrating the model's understanding of context and narrative flow.

3 Rich Relational Capture

Sentence embeddings effectively capture complex semantic and stylistic relationships between texts, a capability that traditional methods focusing on word counts or basic clause structures often miss.

Measure 4: Clause density

1

Long ≠ complex

2

Count subordinators

(that, because, when, dass, weil, obwohl…)

3

Clauses per sentence or per 10 tokens

4

Good for some languages

Measure 4: Interpretation

High clause density

→ academic, reflective, explanatory

Low

→ dialogic, journalistic, action scenes

Measure 5: Sentence types & dialogue ratio

- Count: declaratives, interrogatives, exclamatives
- Count sentences in quotes → dialogue share
- Fiction often alternates: narration → dialogue → narration
- Compare narrator-heavy vs dialogue-heavy authors

Your task

1	**Fork today's repo**
2	**Run at least 3 of the 5 measures on your own corpus** correct the issues when they arise.
3	**Add visualizations** (hist, bar chart)

Next session:

After sentence level → paragraph / discourse (coherence, topic shifts)