COSI-230B: Natural Language Annotation for Machine Learning

Lecture 4: Corpus Selection & Data Sourcing

Jin Zhao

Brandeis University
Computational Linguistics Program

January 28, 2026

Today's Agenda

1. **Corpus Selection Criteria**
   - The MAMA framework for corpus evaluation
2. **Evaluating & Historic Corpora**
   - Sampling, representativeness, balance, reliability
   - Brown Corpus, Switchboard Corpus
3. **Sampling Strategies**
   - How to select data for annotation
4. **Data Licensing, Ethics & Synthetic Data**
   - Licenses, ethical considerations, decontamination

What is a Corpus?

Definition

A **corpus** is a collection of machine-readable texts (or other media) that have been produced in a natural communicative setting.

Key characteristics:

- **Machine-readable** — can be processed by computers
- **Natural** — reflects real language use (not artificial)
- **Collected systematically** — according to some criteria
- May or may not be **annotated**

Examples:

- News articles, social media posts, scientific papers
- Transcribed speech, dialogue, conversations
- Audio recordings, images with captions, videos

Why Corpus Selection Matters

Your corpus determines:

- What phenomena you can study
- How well your model will generalize
- Whether your results are meaningful
- How much work annotation will require

Common mistakes:

- Choosing data that doesn't contain the phenomenon
- Selecting biased or unrepresentative samples
- Ignoring licensing restrictions
- Underestimating annotation difficulty

Key Principle

Spend time choosing the right corpus *before* you start annotating!

The MAMA Framework for Corpus Evaluation

MAMA = Four criteria for evaluating corpora

M – Medium What type of text/media is it?

A – Annotation What annotations exist or are needed?

M – Multimodality Does it include multiple modalities?

A – Availability Can you access and use it legally?

(Not to be confused with the MAMA annotation cycle!)

Let's examine each criterion...

M = Medium

What type of text or media is it?

Consider:

- **Genre:** News, social media, academic, conversational?
- **Domain:** Medical, legal, technical, general?
- **Register:** Formal, informal, mixed?
- **Language:** Monolingual, multilingual, code-switching?

Why it matters:

- Different genres have different linguistic properties
- Models trained on one genre may not transfer to another
- Annotation guidelines may need adaptation

Example: NER on news vs. Twitter requires different approaches (informal language, hashtags, abbreviations)

A = Annotation

What annotations exist or are needed?

Existing annotations:

- Does the corpus already have annotations you can use?
- Are they of sufficient quality?
- Do they match your task requirements?

Needed annotations:

- What new annotations do you need to add?
- How complex is the annotation task?
- How much does the phenomenon occur in the data?

Example: If you want to annotate sarcasm, you need data where sarcasm actually occurs — most news articles won't work!

M = Multimodality

Does the corpus include multiple modalities?

Modalities:

- Text only
- Text + images (social media posts, news with photos)
- Text + audio (transcribed speech)
- Text + video (subtitles, video descriptions)
- Text + structured data (tables, code)

Considerations:

- Do you need multimodal data for your task?
- Does meaning depend on multiple modalities?
- Can you handle the additional complexity?

Example: Sentiment in memes often requires understanding both image and text

A = Availability

Can you access and use the data legally?

Key questions:

- Is the data publicly available?
- What license does it have?
- Can you redistribute your annotations?
- Are there privacy concerns?
- Is institutional approval needed (IRB)?

Common issues:

- Social media data often has restrictive terms of service
- Medical/legal data has privacy requirements
- Some datasets require licensing agreements (LDC)
- Web-scraped data may have unclear provenance

Corpus Evaluation Checklist

Criterion	Questions to Ask
Medium	Does the genre/domain match your needs? Is the language appropriate?
Annotation	Does the phenomenon occur frequently enough? Can you annotate it reliably?
Multimodality	Do you need multiple modalities? Can you handle the complexity?
Availability	Can you legally use and share the data? Can you access it?

For your semester project: Use this checklist when selecting your dataset!

Evaluating Corpora: Key Questions

For each corpus, consider:

Sampling: How was the subset selected from the broader domain?

- What techniques were used?
- Is the selection systematic or ad-hoc?

Representativeness: Does the corpus contain the full range of variation?

- Are all relevant phenomena represented?

Balance: Are different categories/genres proportionally represented?

- Is any category over/under-represented?

Reliability: How consistent and accurate are the annotations?

- What quality control was used?

Pre-1960: Corpus Linguistics

The "Empiricists"

- Empirical and statistical methods were popular starting in the 1950s
- Early machine translation attempts drove interest
- Key figures: Claude Shannon, Yehoshua Bar-Hillel

> *"All of us were convinced that speech, in English or any other language, was a Markov process."*

Limitation: Computers were too limited to handle large corpora

1960–1990s: Rules and Rigor

Using corpora to devise hard-coded predictive rules

- Finite state transducers (FSTs)
- "If-else" models
- Focus on high-level snapshots of language
- Rule-based systems dominated NLP

Key insight: Corpora were used to *discover* rules, not to train statistical models

Brown Corpus (1964)

"Standard Corpus of Present-Day American English"

- **Size:** ~1,000,000 tokens
- **Creators:** Henry Kučera and W. Nelson Francis (Brown University)
- **First** million-word digitized corpus
- Annotators were also the primary researchers

The Data:

> *"1,014,312 words of running text of edited English prose printed in the United States during the calendar year 1961."*

Legacy: Has been extended many times with additional annotation layers

Brown Corpus: Evaluation

Applying our evaluation criteria:

Sampling: Systematic selection from 15 genres of 1961 publications

Balance: Explicitly designed to cover multiple genres

- Press, fiction, academic, etc.

Representativeness: Limited to edited American English prose

- No speech, no informal writing

Reliability: Limited documentation of annotation process

- Small team, expert annotators

Switchboard Corpus (1992)

Telephone conversations between unknown participants

- One of the first corpora built "from the ground up"
- Originally collected by Texas Instruments (1990–1991)
- Funded by DARPA
- Initially focused on speech recognition

Collection method:

- 543 participants (302 male, 241 female)
- Two-sided calls handled by the "Robotoperator"
- Automated system gave callers recorded prompts
- Topics assigned to encourage natural conversation

Switchboard Corpus: Annotation

Primary task: Audio transcriptions with timestamp alignment

Annotation process:

- Approximately half transcribed by court reporters
- Other half by temporary transcribers at Texas Instruments
- All transcriptions reviewed and corrected by QC transcribers

Legacy:

- Extended many times with additional annotation layers
- Dialogue acts, syntax, disfluencies
- Still widely used for speech and dialogue research

Switchboard Corpus: Evaluation

Applying our evaluation criteria:

Sampling: Designed collection with controlled topics

- Participants recruited systematically

Balance: Gender balance, varied topics

- But limited demographic diversity

Representativeness: Spontaneous telephone speech

- May not generalize to other speech contexts

Reliability: Professional transcribers + quality control

- Well-documented annotation guidelines

Lessons from Historic Corpora

What we learn from Brown and Switchboard:

1. **Design matters:** Explicit sampling criteria lead to better corpora
2. **Documentation is crucial:** Well-documented corpora remain useful for decades
3. **Corpora evolve:** Successful corpora get extended with new annotations
4. **Limitations persist:** Even classic corpora have known biases
5. **Context matters:** Understanding *why* a corpus was built helps you use it appropriately

Other Historic Corpora

Additional foundational corpora:

British National Corpus (1994): 100M words of British English

- 90% written texts, 10% transcribed speech
- Designed as balanced, representative sample

Penn Treebank (1992–1996): 1M words of Wall Street Journal

- Parsed into syntax trees
- Standard benchmark for parsing and language modeling

Legacy: These corpora established practices in corpus collection (balance, annotation) that influenced all subsequent work.

Modern Corpora for Language Modeling

The rise of large-scale corpora (2000s–2020s):

Dataset	**Size**	**Domain**	**License**
English Gigaword (2003)	~1.8B words	Newswire	LDC (restricted)
One Billion Word (2013)	~0.8B words	News	Open
BooksCorpus (2015)	~984M words	Fiction books	Copyright issues
Wikipedia	~9B+ words	Encyclopedia	CC BY-SA (open)
WikiText-103 (2016)	~103M tokens	Wikipedia	CC BY-SA

Key insight: Corpora grew from millions to billions of words as neural models demanded more data.

Web-Scale Corpora

Modern LLM training data:

OpenWebText (2019): ~40 GB from Reddit-linked web pages

- Open recreation of GPT-2's training data

C4 (2019): ~750 GB from Common Crawl

- Used to train T5; extensive cleaning/filtering

The Pile (2020): ~825 GB from 22 diverse sources

- Books, Wikipedia, arXiv, GitHub, etc.
- MIT License (open)

RedPajama (2023): ~1.2 trillion tokens

- Open reproduction of LLaMA training data

Multilingual Parallel Corpora

Corpora for machine translation:

Dataset	**Languages**	**Size**	**Domain**
Canadian Hansard	En–Fr	~85M words each	Parliament
Europarl (2005)	21 EU langs	~50M words/lang	EU Parliament
UN Corpus (2016)	6 UN langs	~11M sentences	UN documents
OpenSubtitles	60+ langs	50M+ pairs	Movie subtitles
ParaCrawl (2018)	24+ langs	Millions of pairs	Web-crawled

Impact: These corpora enabled the rise of statistical and neural machine translation.

The Evolution of NLP Corpora

Historical trend:

- **1960s–1990s:** Million-word parsed corpora (Brown, PTB)
- **2000s–2010s:** Billion-word web corpora (Gigaword, Wikipedia)
- **2020s:** Trillion-token multilingual datasets (RedPajama, mC4)

Licensing evolution:

- Early corpora: Often restricted (LDC subscription)
- Modern trend: More open datasets (CC licenses, MIT)
- Ongoing issues: Copyright (books), privacy (web data)

Key Takeaway

Open corpora have democratized NLP research, but quality and ethical concerns remain.

Why Sampling Matters

You usually can't annotate everything.

Constraints:

- Time: Annotation takes 2–10x longer than reading
- Budget: Human annotation costs \$0.10–\$10+ per item
- Annotators: Limited availability of qualified people

Goal of sampling:

- Get a **representative** subset of the data
- Ensure sufficient examples of all categories
- Avoid systematic biases
- Make annotation feasible

Sampling Methods

Random sampling:

- Every item has equal probability of selection
- Simple but may miss rare phenomena

Stratified sampling:

- Divide data into strata (e.g., by source, date, length)
- Sample from each stratum proportionally
- Ensures representation of subgroups

Targeted sampling:

- Use keywords or heuristics to find relevant examples
- Good for rare phenomena
- Risk of bias if heuristics are incomplete

Active learning:

- Model selects most informative examples
- Efficient but requires initial model

Sampling for Your Project

Practical recommendations:

1. **Estimate target size:** How many examples do you need?
   - Classification: 100–500 per class minimum
   - Sequence labeling: 500–2000 sentences
   - Complex tasks: depends on phenomenon frequency
2. **Check phenomenon frequency:**
   - Use grep/search to estimate occurrence rate
   - If rare, use targeted sampling
3. **Ensure diversity:**
   - Sample from different sources/time periods
   - Avoid over-representation of outliers

Why Licensing Matters

Annotation is expensive. You want to:

- Share your annotated data with others
- Publish research using the data
- Use it in commercial applications (maybe)

But you must respect:

- Original data creator's rights
- Privacy of individuals in the data
- Terms of service (for web data)
- Institutional requirements

Important

Check the license *before* you start annotating!

Common Open Licenses

License	Attribution	ShareAlike	Commercial
CC0 (Public Domain)	No	No	Yes
CC-BY	Yes	No	Yes
CC-BY-SA	Yes	Yes	Yes
CC-BY-NC	Yes	No	No
CC-BY-NC-SA	Yes	Yes	No
MIT / Apache 2.0	Yes	No	Yes

Key terms:

- **Attribution (BY):** Must credit the creator
- **ShareAlike (SA):** Derivatives must use same license
- **NonCommercial (NC):** Cannot use for commercial purposes

Licensing Considerations for Annotation

When you annotate data:

- Your *annotations* are a new creative work
- You can license annotations separately from original data
- But you must comply with original data's license

Common approaches:

- **Release annotations only:** Provide offsets/IDs, not original text
- **Provide reconstruction script:** Users download original + your annotations
- **Request permission:** Contact data owner for redistribution rights

Problem: What if the original data changes or disappears?

- Social media posts get deleted
- Websites change their content
- Your annotations become orphaned

Ethical Considerations

Privacy:

- Does the data contain personal information?
- Can individuals be identified?
- Did people consent to their data being used?

Bias:

- Is the data representative of the population?
- Are certain groups over/under-represented?
- Could your annotations perpetuate biases?

Harm:

- Could your dataset be misused?
- Are annotators exposed to harmful content?
- What safeguards are needed?

Synthetic Data Generation

New option: Generate data with LLMs

Approaches:

- **Direct generation:** Ask LLM to generate examples
- **Paraphrasing:** Transform existing examples
- **Augmentation:** Create variations of real data
- **Seed-based:** Generate from prompts/templates

Example prompt:
"Generate 10 examples of customer complaints about late delivery. Include varied emotions and phrasing."

Pros and Cons of Synthetic Data

Advantages:

- Fast and cheap to generate
- No licensing issues
- Control over distribution
- Can target rare phenomena
- Privacy-preserving

Disadvantages:

- May not reflect real language
- LLM biases transfer
- Limited diversity
- Quality varies
- Still needs validation

Best Practice

Use synthetic data to **supplement** real data, not replace it entirely.

Data Contamination: The Problem

What is data contamination?

- Your test data appears in the LLM's training data
- Model has "seen the answers" before
- Evaluation results are inflated and invalid

Why it matters:

- LLMs are trained on massive web crawls
- Popular datasets are likely in training data
- You can't trust evaluation on contaminated data

Critical Issue

If you're evaluating LLMs, contamination invalidates your results!

Decontamination Strategies

How to avoid/detect contamination:

1. **Use recent data:**
   - Collect data after LLM training cutoff
   - Check model documentation for training dates
2. **Create novel examples:**
   - Generate new test cases
   - Use private/unpublished data
3. **Check for overlap:**
   - N-gram overlap with known training data
   - Membership inference tests
4. **Use held-out data:**
   - Never publish your test set
   - Use hidden evaluation servers

Contamination in Practice

Known contaminated benchmarks:

- Many standard NLP benchmarks (GLUE, SuperGLUE, SQuAD)
- Popular datasets on Hugging Face
- Anything widely used before 2021–2023

For your projects:

- If evaluating LLMs, document potential contamination
- Consider using recent or private data for test sets
- Be transparent about limitations

Research frontier:

- Detecting contamination is an active research area
- No perfect solutions yet
- Best practice: multiple evaluation strategies

Key Takeaways

1. **MAMA framework:** Medium, Annotation, Multimodality, Availability
2. **Sampling matters:** Random, stratified, targeted, or active
3. **Check licenses:** CC-BY, CC-BY-NC, etc. — know what you can do
4. **Consider ethics:** Privacy, bias, potential harm
5. **Synthetic data:** Useful but has limitations
6. **Decontamination:** Critical for LLM evaluation

Questions?

jinzhao@brandeis.edu

Office Hours: Wed 1–3pm (Volen 109)

MOODLE for announcements