# Lecture 7: Text as data

## LING-351 Language Technology and LLMs

Instructor: Hakyung Sung

September 16, 2025

*Acknowledgment: These course slides are based on materials by Lelia Glass @ Georgia Tech (Course: Language & Computers)

# Table of contents

# Review

# CALL

- CALL stands for Computer-Assisted Language Learning

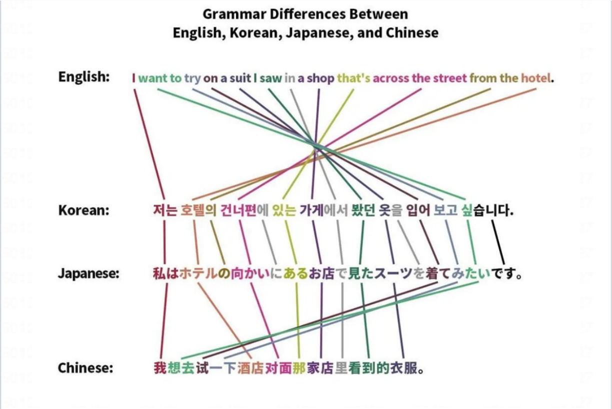
# Language Learning

L1 vs. L2: - what is the similarity? - what is the difference?

# Language transfer

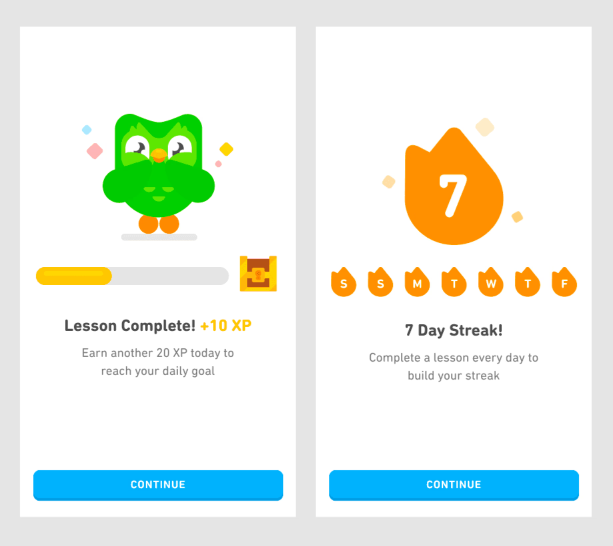
L1 vs. L2: - positive transfer - negative transfer - typology

# Language typology




*https://wals.info/*

# Motivation: Gamification and reinforcement




Sourced from 2024 Duolingo language report : *https://nudgenow.com/blogs/duolingo-gamification-strategy*

# Self-determination theory (Deci & Ryan)

- Motivation increases when three needs are met:

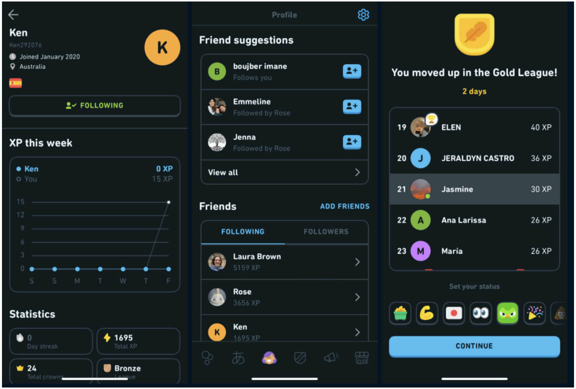
# Self-determination theory (Deci & Ryan)

- Motivation increases when three needs are met:
    - Autonomy: choice over pace and goals

# Self-determination theory (Deci & Ryan)

- Motivation increases when three needs are met:
  - Autonomy: choice over pace and goals
  - Competence: sense of progress (points, levels, streaks)

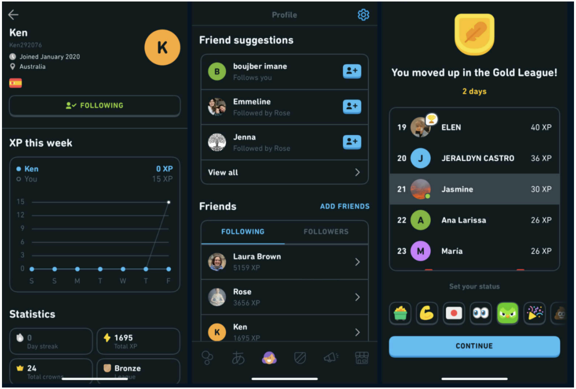
# Social comparison




Sourced from 2024 Duolingo language report : *https://raw.studio/blog/how-duolingo-utilises-gamification/*

- Leaderboards and friend lists create social competition

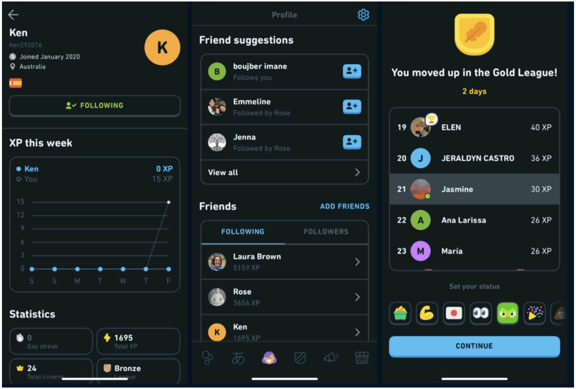
# Social comparison




Sourced from 2024 Duolingo language report : *https://raw.studio/blog/how-duolingo-utilises-gamification/*

- Leaderboards and friend lists create social competition
- Similar to SNS: recognition and belonging motivate persistence

# Social comparison




Sourced from 2024 Duolingo language report : *https://raw.studio/blog/how-duolingo-utilises-gamification/*

- Leaderboards and friend lists create social competition
- Similar to SNS: recognition and belonging motivate persistence
- Learners compare progress and are encouraged to “keep up”

# Attitude, motivation, and growth mindset

- Other factors:

# Attitude, motivation, and growth mindset

- Other factors:
    - Intrinsic/extrinsic motivation

# Attitude, motivation, and growth mindset

- Other factors:
    - Intrinsic/extrinsic motivation
    - Positive attitudes toward the target language and culture

# Attitude, motivation, and growth mindset

- Other factors:
    - Intrinsic/extrinsic motivation
    - Positive attitudes toward the target language and culture
    - Willingness to take risks and make mistakes

# Coming back to CALL

# Coming back to CALL

At the intersection of language learning and educational technology, the field of Computer-Assisted Language Learning (CALL) develops tools to support and enhance second language acquisition.

# What do we mean by CALL?

- **Broad sense**: Refers to the many ways computers intersect with education and society in language learning.
- Examples: multimedia textbooks, online dictionaries, digital writing tools, consuming media, and connecting socially with L2 speakers.

# What do we mean by CALL?

- **Narrow sense**: Describes instructional tools that deliver sequenced exercises, provide feedback on responses, and are often used in *language assessment contexts*.

# Example: *Fill-in-the-blank*

The detective lives _____ Baker Street.

- Free-text input?
  - One correct answer: *on*?
  - What if the learner enters: *near, by, at*? How to respond?
  - How to grade correctness or give feedback automatically?

# Example: *Fill-in-the-blank*

The detective lives _____ Baker Street.

- Free-text input?
  - One correct answer: *on*?
  - What if the learner enters: *near, by, at*? How to respond?
  - How to grade correctness or give feedback automatically?
- Multiple choice alternative:
  - *At, On, In, With*
  - How are these distractors chosen?
  - What feedback should be given for a wrong choice?

# Example: *Fill-in-the-blank*

The detective lives _____ Baker Street.

- Free-text input?
    - One correct answer: *on*?
    - What if the learner enters: *near, by, at*? How to respond?
    - How to grade correctness or give feedback automatically?
- Multiple choice alternative:
    - *At, On, In, With*
    - How are these distractors chosen?
    - What feedback should be given for a wrong choice?
- Sequencing:
    - How to ensure the question is not too hard or too easy?

# Motivating intelligent CALL

- Pre-specifying all possible answers and feedback is:
    - Laborious
    - Brittle to unexpected input

# Motivating intelligent CALL

- Pre-specifying all possible answers and feedback is:
    - Laborious
    - Brittle to unexpected input
- Instead, we want an *intelligent tutoring system* (ITS) that can reason about:
    - **Language** – grammar, usage, semantics
    - **The learner** – skill level, past errors, learning history

# Motivating intelligent CALL

- Pre-specifying all possible answers and feedback is:
    - Laborious
    - Brittle to unexpected input
- Instead, we want an *intelligent tutoring system* (ITS) that can reason about:
    - **Language** – grammar, usage, semantics
    - **The learner** – skill level, past errors, learning history
- Recently, ITS are also used in domains like:
    - Math
    - Computer science (teaching coding skills)

# Motivating intelligent CALL

- Pre-specifying all possible answers and feedback is:
  - Laborious
  - Brittle to unexpected input
- Instead, we want an *intelligent tutoring system* (ITS) that can reason about:
  - **Language** – grammar, usage, semantics
  - **The learner** – skill level, past errors, learning history
- Recently, ITS are also used in domains like:
  - Math
  - Computer science (teaching coding skills)
- These domains are often **more constrained**, which may make:
  - Feedback and hints easier to automate
  - Learner modeling more reliable

## CALL design considerations

- How to write the task?

# CALL design considerations

- How to write the task?
- How to grade it automatically?

# CALL design considerations

- How to write the task?
- How to grade it automatically?
- How hard is it to implement?

# CALL design considerations

- How to write the task?
- How to grade it automatically?
- How hard is it to implement?
- How fun or useful is it for learners?

# CALL design considerations

- How to write the task?
- How to grade it automatically?
- How hard is it to implement?
- How fun or useful is it for learners?
- Trade-offs:

# CALL design considerations

- How to write the task?
- How to grade it automatically?
- How hard is it to implement?
- How fun or useful is it for learners?
- Trade-offs:
    - Multiple choice: easy to grade, but limited expression

# CALL design considerations

- How to write the task?
- How to grade it automatically?
- How hard is it to implement?
- How fun or useful is it for learners?
- Trade-offs:
  - Multiple choice: easy to grade, but limited expression
  - Free-text: richer data, but harder to parse and evaluate

*https://www.tandfonline.com/action/showAxaArticles?journalCode=ncal20*

# Lesson plan

# Lesson plan

- ~~Review: CALL~~

Key idea:

# Lesson plan

- ~~Review: CALL~~
- Text as data: Two different approaches

Key idea:

# Lesson plan

- ~~Review: CALL~~
- Text as data: Two different approaches
- Questions with answers in text

Key idea:

# Lesson plan

- ~~Review: CALL~~
- Text as data: Two different approaches
- Questions with answers in text
- Good data for the data-driven approach

Key idea:

# Lesson plan

- ~~Review: CALL~~
- Text as data: Two different approaches
- Questions with answers in text
- Good data for the data-driven approach

Key idea:

# Lesson plan

- ~~Review: CALL~~
- Text as data: Two different approaches
- Questions with answers in text
- Good data for the data-driven approach

Key idea: Language technology is not only for answering linguistic questions—it can also address a wide range of issues using text data.

# Intro

# Language technology always built on linguistic theory?

So far, our tour of language technology has incorporated a great deal of linguistic representations and theories:

So far, our tour of language technology has incorporated a great deal of linguistic representations and theories:

- Different encodings (e.g., alphabetic, syllabic, logographic)

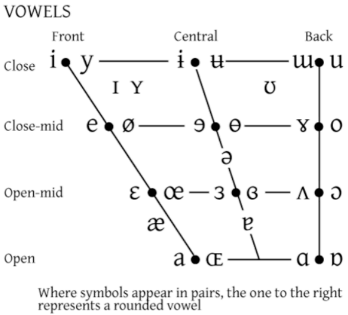
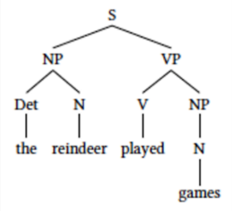


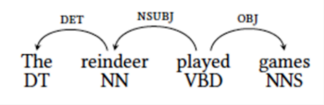

Figure 1.2: International Phonetic Alphabet of vowels (https://commons.wikimedia.org/wiki/File:Ipa-chart-vowels.png)

# Language technology always built on linguistic theory?

So far, our tour of language technology has incorporated a great deal of linguistic representations and theories:

- Theories of grammar for writers' aids



# Language technology always built on linguistic theory?

So far, our tour of language technology has incorporated a great deal of linguistic representations and theories:

- L1 and L2 learning research for CALL

# Language technology always built on linguistic theory?

So far, our tour of language technology has incorporated a great deal of linguistic representations and theories:

- L1 and L2 learning research for CALL

You might be getting the impression that language technology always builds on constructs from linguistics.

# Letting the data speak

But there's also much to be gained from letting language data speak for itself—**independent of any particular theory.**

# Letting the data speak

But there's also much to be gained from letting language data speak for itself—**independent of any particular theory.**

Every day, people generate **billions of words** in electronic text:

- Social media, emails, journalism

# Letting the data speak

But there's also much to be gained from letting language data speak for itself—**independent of any particular theory.**

Every day, people generate **billions of words** in electronic text:

- Social media, emails, journalism
- Schoolwork, scholarly papers, lawsuits

# Letting the data speak

But there's also much to be gained from letting language data speak for itself—**independent of any particular theory.**

Every day, people generate **billions of words** in electronic text:

- Social media, emails, journalism
- Schoolwork, scholarly papers, lawsuits
- Wikipedia, books, TV scripts, reviews

# Letting the data speak

But there's also much to be gained from letting language data speak for itself—**independent of any particular theory.**

Every day, people generate **billions of words** in electronic text:

- Social media, emails, journalism
- Schoolwork, scholarly papers, lawsuits
- Wikipedia, books, TV scripts, reviews
- Doctors' notes, and more

# Letting the data speak

But there's also much to be gained from letting language data speak for itself—**independent of any particular theory.**

Every day, people generate **billions of words** in electronic text:

- Social media, emails, journalism
- Schoolwork, scholarly papers, lawsuits
- Wikipedia, books, TV scripts, reviews
- Doctors' notes, and more

This vast amount of text reflects patterns in:

# Letting the data speak

But there’s also much to be gained from letting language data speak for itself—**independent of any particular theory.**

Every day, people generate **billions of words** in electronic text:

- Social media, emails, journalism
- Schoolwork, scholarly papers, lawsuits
- Wikipedia, books, TV scripts, reviews
- Doctors’ notes, and more

This vast amount of text reflects patterns in:

- society, education, law, economics, health science, etc.

# What is Text as Data?

**Text as data** is a cross-disciplinary endeavor to:

- Extract information from large-scale corpora

*This marks a pivot from:*

# What is Text as Data?

**Text as data** is a cross-disciplinary endeavor to:

- Extract information from large-scale corpora
- Utilize it for a wide variety of purposes

*This marks a pivot from:*

# What is Text as Data?

**Text as data** is a cross-disciplinary endeavor to:

- Extract information from large-scale corpora
- Utilize it for a wide variety of purposes

*This marks a pivot from:*

- Top-down, **knowledge-driven** applications

# What is Text as Data?

**Text as data** is a cross-disciplinary endeavor to:

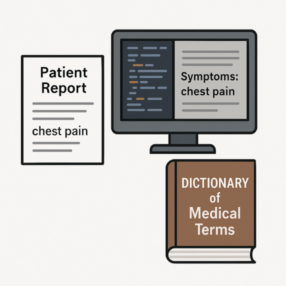
- Extract information from large-scale corpora
- Utilize it for a wide variety of purposes

*This marks a pivot from:*

- Top-down, **knowledge-driven** applications
- Bottom-up, **data-driven** language technology

# Example 1

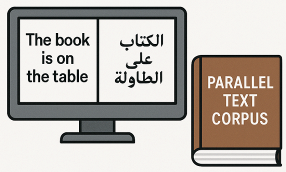
I'm trying to build a system that uses a hand-crafted dictionary of medical terms and explicit grammar rules to extract symptoms from patient reports.




Q. Knowledge-driven or data-driven? Why?

# Example 2

I'm trying to build a translation tool trained on millions of parallel sentences between English and Arabic, and produces translations by predicting word sequences statistically.




Q. Knowledge-driven or data-driven? Why?

# Example 3

The research team recently built a chatbot which follows *if–then* rules written by linguists and domain experts, where every possible user input is matched against a predefined template.

Q. Knowledge-driven or data-driven? Why?

# Example 4

We now use an algorithm that detects fake news by training on large corpora of labeled real vs. fake news articles, learning which word patterns correlate with each label.

Q. Knowledge-driven or data-driven? Why?

# Example 5

A researcher builds a sentiment analyzer by creating a list of positive/negative words and assigning scores manually.

Q. Knowledge-driven or data-driven?

# Example 6

A system identifies people's names in text using a neural named entity recognition (NER) model trained on millions of labeled sentences.

Q. Knowledge-driven or data-driven?

# Example 7

In a low-resource language project, linguists encode morphology and syntax rules by hand because there isn't enough digital text to train a model.

Q. Knowledge-driven or data-driven?

## Key considerations when using Text

When analyzing text data, consider:

- **Corpus size** — How big is it? Is it sufficient?

# Key considerations when using Text

When analyzing text data, consider:

- **Corpus size** — How big is it? Is it sufficient?
- **Corpus type** — What kind of text is included?

# Key considerations when using Text

When analyzing text data, consider:

- **Corpus size** — How big is it? Is it sufficient?
- **Corpus type** — What kind of text is included?
- **Annotations** — POS tags, dependency relations, NER, etc.

# Key considerations when using Text

When analyzing text data, consider:

- **Corpus size** — How big is it? Is it sufficient?
- **Corpus type** — What kind of text is included?
- **Annotations** — POS tags, dependency relations, NER, etc.
- **Metadata** — Date, location, author, audience

# Key considerations when using Text

When analyzing text data, consider:

- **Corpus size** — How big is it? Is it sufficient?
- **Corpus type** — What kind of text is included?
- **Annotations** — POS tags, dependency relations, NER, etc.
- **Metadata** — Date, location, author, audience
- **Methods** — How is information extracted?

# Key considerations when using Text

When analyzing text data, consider:

- **Corpus size** — How big is it? Is it sufficient?
- **Corpus type** — What kind of text is included?
- **Annotations** — POS tags, dependency relations, NER, etc.
- **Metadata** — Date, location, author, audience
- **Methods** — How is information extracted?
- **Statistics** — What inference methods are applied?

# Questions with answers in text

# Touring text-as-data Questions

We begin with a tour of real-world questions answered using text across various fields:

- Digital Humanities

# Touring text-as-data Questions

We begin with a tour of real-world questions answered using text across various fields:

- Digital Humanities
- Computational Social Science

# Touring text-as-data Questions

We begin with a tour of real-world questions answered using text across various fields:

- Digital Humanities
- Computational Social Science
- Author Profiling

# Touring text-as-data Questions

We begin with a tour of real-world questions answered using text across various fields:

- Digital Humanities
- Computational Social Science
- Author Profiling
- Corpus Linguistics

# 1. Digital humanities

Digital Humanities = the study of literature, culture, and history using digital tools.

**Digital Humanities** = the study of literature, culture, and history using digital tools. Examples:

- Mapping places mentioned in novels

# 1. Digital humanities

**Digital Humanities** = the study of literature, culture, and history using digital tools. Examples:

- Mapping places mentioned in novels
- Visualizing character interactions

# 1. Digital humanities

Digital Humanities = the study of literature, culture, and history using digital tools. Examples:

- Mapping places mentioned in novels
- Visualizing character interactions
- Tracking genre trends across time

# 1. Digital humanities

Digital Humanities = the study of literature, culture, and history using digital tools. Examples:

- Mapping places mentioned in novels
- Visualizing character interactions
- Tracking genre trends across time
- Mining digital archives for silenced voices

# Examples from the textbook

Followings are example studies from the textbook.

# Example 1: Distant Reading (Moretti, 2013)

Challenge: A scholar can only read a tiny fraction of all books ever published.

# Example 1: Distant Reading (Moretti, 2013)

**Challenge:** A scholar can only read a tiny fraction of all books ever published.

**Solution:** *Distant Reading* — using "graphs, maps, and trees" to gain a macro-view of literature:

- Map of places in novels

# Example 1: Distant Reading (Moretti, 2013)

**Challenge:** A scholar can only read a tiny fraction of all books ever published.

**Solution:** *Distant Reading* — using “graphs, maps, and trees” to gain a macro-view of literature:

- Map of places in novels
- Graph of character interactions in plays

# Example 1: Distant Reading (Moretti, 2013)

**Challenge:** A scholar can only read a tiny fraction of all books ever published.

**Solution:** *Distant Reading* — using “graphs, maps, and trees” to gain a macro-view of literature:

- Map of places in novels
- Graph of character interactions in plays
- Genre publication trends over time

# Example 1: Distant Reading (Moretti, 2013)

**Challenge:** A scholar can only read a tiny fraction of all books ever published.

**Solution:** *Distant Reading* — using “graphs, maps, and trees” to gain a macro-view of literature:

- Map of places in novels
- Graph of character interactions in plays
- Genre publication trends over time
- Library acquisition records

## Example 2: Gender in Fiction (Underwood et al., 2018)

Goal: Analyze how gender is represented in fiction from 1800s to 2000s. Methodologically:

# Example 2: Gender in Fiction (Underwood et al., 2018)

Goal: Analyze how gender is represented in fiction from 1800s to 2000s. Methodologically:

- Used NLP tools to extract characters and associated words

Some findings:

# Example 2: Gender in Fiction (Underwood et al., 2018)

Goal: Analyze how gender is represented in fiction from 1800s to 2000s. Methodologically:

- Used NLP tools to extract characters and associated words
- Focused on stereotypical associations: *"smile/laugh" (women), "grin/chuckle" (men)*

Some findings:

# Example 2: Gender in Fiction (Underwood et al., 2018)

Goal: Analyze how gender is represented in fiction from 1800s to 2000s. Methodologically:

- Used NLP tools to extract characters and associated words
- Focused on stereotypical associations: *"smile/laugh" (women), "grin/chuckle" (men)*

Some findings:

- Gender prediction from word use becomes harder over time

# Example 2: Gender in Fiction (Underwood et al., 2018)

Goal: Analyze how gender is represented in fiction from 1800s to 2000s. Methodologically:

- Used NLP tools to extract characters and associated words
- Focused on stereotypical associations: *"smile/laugh" (women), "grin/chuckle" (men)*

Some findings:

- Gender prediction from word use becomes harder over time
- Decline of women authors during mid-1800s to mid-1900s

# Example 2: Gender in Fiction (Underwood et al., 2018)

**Goal:** Analyze how gender is represented in fiction from 1800s to 2000s. Methodologically:

- Used NLP tools to extract characters and associated words
- Focused on stereotypical associations: *"smile/laugh" (women), "grin/chuckle" (men)*

Some findings:

- Gender prediction from word use becomes harder over time
- Decline of women authors during mid-1800s to mid-1900s
- Women authors depict men and women equally; men tend to depict more male characters

# 2. Computational social science

Computational Social Science (CSS) = Use of corpora to study social science questions.

- Q. *What is social science?*

# 2. Computational social science

Computational Social Science (CSS) = Use of corpora to study social science questions.

- Q. *What is social science?*
- Media analysis: Topic coverage across news outlets
- Network analysis: Spread of ideas on social media
- Community behavior: Conformity and uniqueness in forums
- Online harm: Trolling, misinformation, fake news

## Example 1: Politeness and Power (Danescu-Niculescu-Mizil et al., 2013)

Goal: Investigate how computational methods be used to identify and model linguistic aspects of politeness.

# Example 1: Politeness and Power (Danescu-Niculescu-Mizil et al., 2013)

Goal: Investigate how computational methods be used to identify and model linguistic aspects of politeness.

Data: Requests from Stack Exchange and Wikipedia talk pages

# Example 1: Politeness and Power (Danescu-Niculescu-Mizil et al., 2013)

Goal: Investigate how computational methods be used to identify and model linguistic aspects of politeness.

Data: Requests from Stack Exchange and Wikipedia talk pages

Method:

- Human ratings of politeness via Amazon Mechanical Turk
- Trained a classifier to predict politeness

# Example 1: Politeness and Power (Danescu-Niculescu-Mizil et al., 2013)

Goal: Investigate how computational methods be used to identify and model linguistic aspects of politeness.

Data: Requests from Stack Exchange and Wikipedia talk pages

Method:

- Human ratings of politeness via Amazon Mechanical Turk
- Trained a classifier to predict politeness

A snippet of the findings:

- Wikipedia editors become **less polite** after gaining power

# Example 2: Political Framing on Twitter (Demszky et al., 2019)

Goal: Investigate how political identity shapes framing differences in public discourse on Twitter.

# Example 2: Political Framing on Twitter (Demszky et al., 2019)

Goal: Investigate how political identity shapes framing differences in public discourse on Twitter.

**Data:** Tweets about U.S. mass shootings

# Example 2: Political Framing on Twitter (Demszky et al., 2019)

Goal: Investigate how political identity shapes framing differences in public discourse on Twitter.

Data: Tweets about U.S. mass shootings

Method:

- Classified users by party (based on followed politicians)

Findings:

# Example 2: Political Framing on Twitter (Demszky et al., 2019)

Goal: Investigate how political identity shapes framing differences in public discourse on Twitter.

Data: Tweets about U.S. mass shootings

Method:

- Classified users by party (based on followed politicians)
- Analyzed word choice and topics

Findings:

# Example 2: Political Framing on Twitter (Demszky et al., 2019)

Goal: Investigate how political identity shapes framing differences in public discourse on Twitter.

Data: Tweets about U.S. mass shootings

Method:

- Classified users by party (based on followed politicians)
- Analyzed word choice and topics

Findings:

- *Republicans*: *"crazy"* (white shooters), *"terrorist"* (shooters of color)

# Example 2: Political Framing on Twitter (Demszky et al., 2019)

Goal: Investigate how political identity shapes framing differences in public discourse on Twitter.

Data: Tweets about U.S. mass shootings

Method:

- Classified users by party (based on followed politicians)
- Analyzed word choice and topics

Findings:

- *Republicans*: *"crazy"* (white shooters), *"terrorist"* (shooters of color)
- *Democrats*: reversed pattern, and more likely to mention gun laws

# Example 3: Menu Language and Social Class (Jurafsky et al., 2018)

**Goal:** Examine how the language of restaurant menus reflects and constructs social class, by analyzing how menu wording varies across different types and price levels of restaurants.

# Example 3: Menu Language and Social Class (Jurafsky et al., 2018)

Goal: Examine how the language of restaurant menus reflects and constructs social class, by analyzing how menu wording varies across different types and price levels of restaurants.

Data: Restaurant menus (with type and price from Yelp)

# Example 3: Menu Language and Social Class (Jurafsky et al., 2018)

Goal: Examine how the language of restaurant menus reflects and constructs social class, by analyzing how menu wording varies across different types and price levels of restaurants.

Data: Restaurant menus (with type and price from Yelp)

Findings:

- **Cheap menus:** more words, traditional authenticity (e.g., Grandma's recipe)

# Example 3: Menu Language and Social Class (Jurafsky et al., 2018)

Goal: Examine how the language of restaurant menus reflects and constructs social class, by analyzing how menu wording varies across different types and price levels of restaurants.

Data: Restaurant menus (with type and price from Yelp)

Findings:

- **Cheap menus:** more words, traditional authenticity (e.g., Grandma's recipe)
- **Expensive menus:** fewer words, natural authenticity (e.g., wild-caught salmon)

# Example 3: Menu Language and Social Class (Jurafsky et al., 2018)

**Goal:** Examine how the language of restaurant menus reflects and constructs social class, by analyzing how menu wording varies across different types and price levels of restaurants.

**Data:** Restaurant menus (with type and price from Yelp)

**Findings:**

- **Cheap menus:** more words, traditional authenticity (e.g., Grandma's recipe)
- **Expensive menus:** fewer words, natural authenticity (e.g., wild-caught salmon)
- Quality on cheap menus is described; on expensive menus, it's implied

**Author profiling** = Inferring characteristics of a writer from their text.

- Gender, age, personality, political ideology

# 3. Author profiling and identification

Author profiling = Inferring characteristics of a writer from their text.

- Gender, age, personality, political ideology
- Mental health, native language

**Goal:** Use language to identify or describe authors in criminal cases

**Technique:** Analyze phrasing, spelling, vocabulary, etc.

Even when the author is known, we can quantify their style:

- Frequent n-grams (e.g., bigrams like “of course”)

# 3-2. Stylometry and literary voice

Even when the author is known, we can quantify their style:

- Frequent n-grams (e.g., bigrams like “of course”)
- Common word classes: verbs, adjectives, nouns

# 3-2. Stylometry and literary voice

Even when the author is known, we can quantify their style:

- Frequent n-grams (e.g., bigrams like “of course”)
- Common word classes: verbs, adjectives, nouns
- Lexical diversity (diverse words? limited words?)

# 3-2. Stylometry and literary voice

Even when the author is known, we can quantify their style:

- Frequent n-grams (e.g., bigrams like “of course”)
- Common word classes: verbs, adjectives, nouns
- Lexical diversity (diverse words? limited words?)
- Average sentence length and starter words

**Corpus linguistics** = Studying *language itself* through large collections of real-world text (corpora)

- Annotated corpora (e.g., POS tags, syntax trees)
- Used to test linguistic theories and train computational models
- Enables empirical observation on how different language users actually produced their languages

## Example: L2-English Speakers' Production (Sung & Kyle, 2025)

Goal: Investigate whether more proficient L2-English speakers produce

- more diverse grammatical structures
- more diverse verbs within the same grammatical forms

# Example: L2-English Speakers' Production (Sung & Kyle, 2025)

Goal: Investigate whether more proficient L2-English speakers produce

- more diverse grammatical structures
- more diverse verbs within the same grammatical forms

Example:

- **Less proficient:**

# Example: L2-English Speakers' Production (Sung & Kyle, 2025)

Goal: Investigate whether more proficient L2-English speakers produce

- more diverse grammatical structures
- more diverse verbs within the same grammatical forms

Example:

- **Less proficient:**
- "I made a cake." (Subject–Verb–Object)

# Example: L2-English Speakers' Production (Sung & Kyle, 2025)

Goal: Investigate whether more proficient L2-English speakers produce

- more diverse grammatical structures
- more diverse verbs within the same grammatical forms

Example:

- **Less proficient:**
- "I made a cake." (Subject–Verb–Object)
- "Mom made a meal."

# Example: L2-English Speakers' Production (Sung & Kyle, 2025)

Goal: Investigate whether more proficient L2-English speakers produce

- more diverse grammatical structures
- more diverse verbs within the same grammatical forms

Example:

- **Less proficient:**
- "I made a cake." (Subject–Verb–Object)
- "Mom made a meal."
- **More proficient:**

# Example: L2-English Speakers' Production (Sung & Kyle, 2025)

Goal: Investigate whether more proficient L2-English speakers produce

- more diverse grammatical structures
- more diverse verbs within the same grammatical forms

Example:

- **Less proficient:**
- "I made a cake." (Subject–Verb–Object)
- "Mom made a meal."
- **More proficient:**
- "I made him a cake." (Double-object construction)

# Example: L2-English Speakers' Production (Sung & Kyle, 2025)

Goal: Investigate whether more proficient L2-English speakers produce

- more diverse grammatical structures
- more diverse verbs within the same grammatical forms

Example:

- **Less proficient:**
- "I made a cake." (Subject–Verb–Object)
- "Mom made a meal."
- **More proficient:**
- "I made him a cake." (Double-object construction)
- "Mom cooked a meal for us." (Prepositional dative)

# Example: L2-English Speakers' Production (Sung & Kyle, 2025)

Goal: Investigate whether more proficient L2-English speakers produce

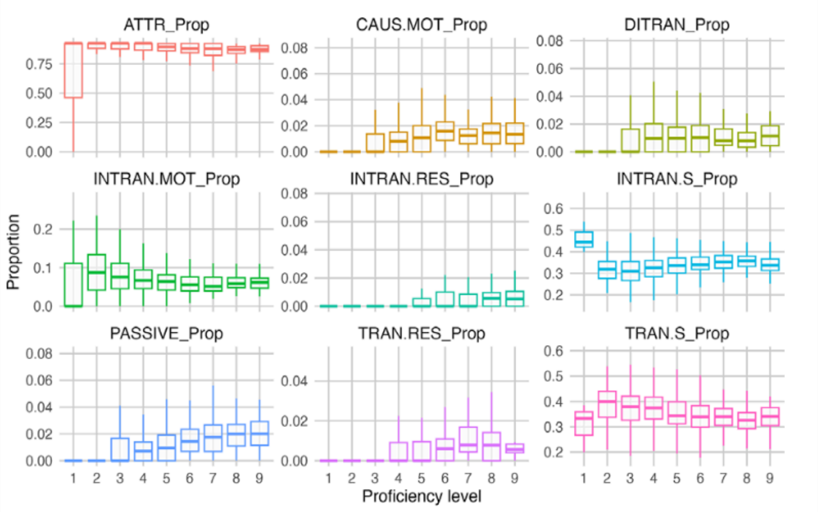
- more diverse grammatical structures
- more diverse verbs within the same grammatical forms

Example:

- **Less proficient:**
- "I made a cake." (Subject–Verb–Object)
- "Mom made a meal."
- **More proficient:**
- "I made him a cake." (Double-object construction)
- "Mom cooked a meal for us." (Prepositional dative)

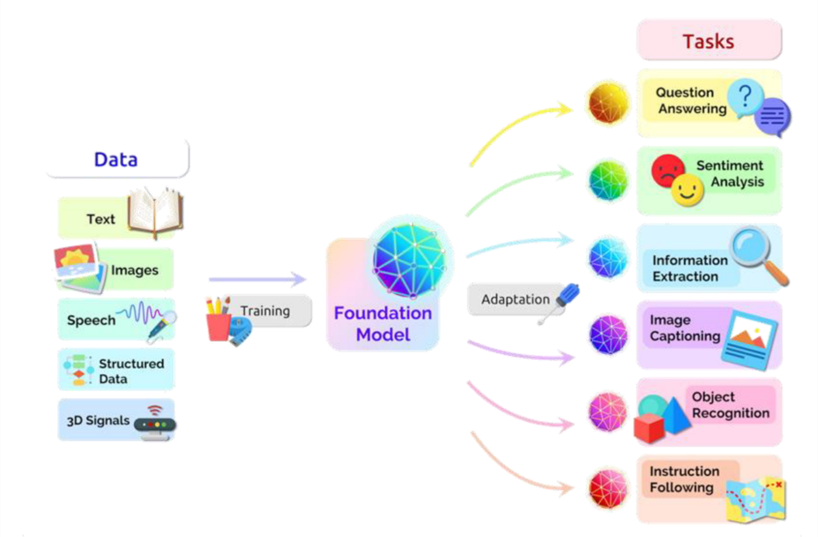
*This might sound obvious, but it has been challenging to measure in large datasets—previous studies have often remained descriptive rather than computational.*

# Example: L2-English Speakers' Production (Sung & Kyle, 2025)



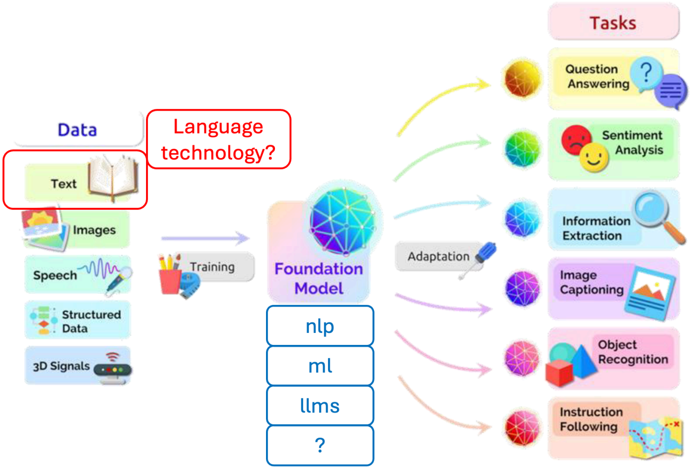
# Good data for training

# Logistics of the data-driven approach

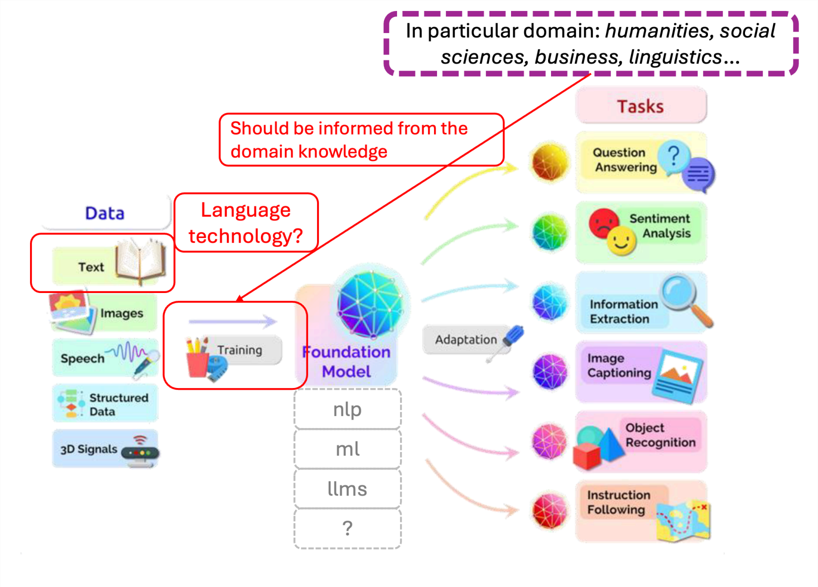


# Logistics of the data-driven approach

In particular domain: *humanities, social sciences, business, linguistics*...



# Logistics of the data-driven approach



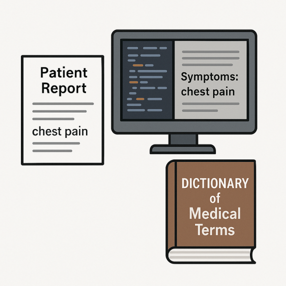
# Recall: Example 1

I’ve used a system that uses a hand-crafted dictionary of medical terms and explicit grammar rules to extract symptoms from patient reports.

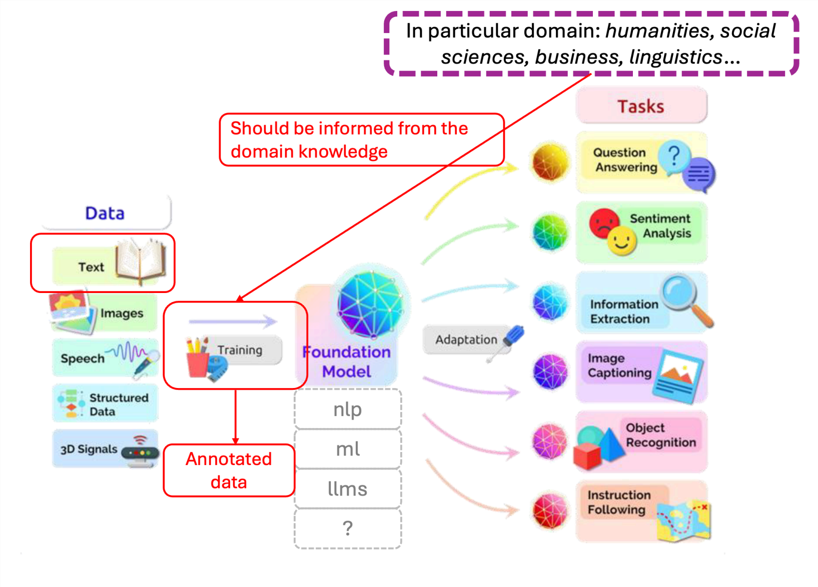
# Recall: Example 1

I’ve used a system that uses a hand-crafted dictionary of medical terms and explicit grammar rules to extract symptoms from patient reports.

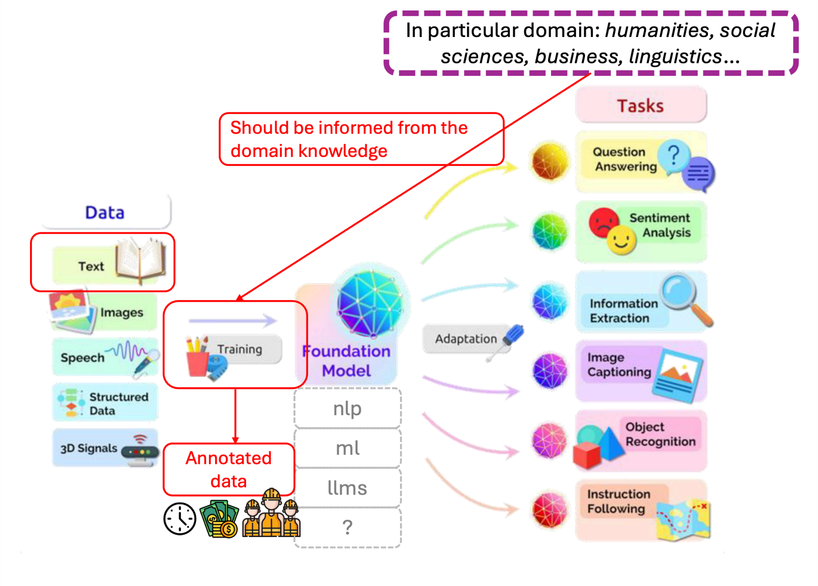
But, now, I know how to train LLMs to automatically extract medical terms from the patient reports! What do I need first?



# Logistics of the data-driven approach: Annotation



# Logistics of the data-driven approach: Annotation



# Annotation sources (in practice)

- **Expert annotation**: an expert of the target domain (e.g., a doctor-medical research)
- **Crowdsourced annotation:** via Mechanical Turk, Prolific
- **Automated tools:** taggers/parsers trained on some annotated data

# Wrap-up

# Wrap-up

- Text as data: Two different approaches

# Wrap-up

- Text as data: Two different approaches
- Questions with answers in text

# Wrap-up

- Text as data: Two different approaches
- Questions with answers in text
- Good data for the data-driven approach

# Wrap-up

- Text as data: Two different approaches
- Questions with answers in text
- Good data for the data-driven approach

Key idea: Language technology is not only for answering linguistic questions—it can also address a wide range of issues using text data.

# Wrap-up

- Text as data: Two different approaches
- Questions with answers in text
- Good data for the data-driven approach

Key idea: Language technology is not only for answering linguistic questions—it can also address a wide range of issues using text data. To do this effectively, we first need to understand how the field approaches text as data.

# Brainstorm your research interests

| | | | | |
|---|---|---|---|---|
| 4 | 9/16 | Text as data | [LC] Ch. 4.1-4.3 | |
| | 9/18 | Python tutorial 3 | | |
| 5 | 9/23 | Word vectors | [LC] Ch. 4.4 | |
| | 9/25 | Python tutorial 4 | | Exercise 3 |
| 6 | 9/30 | Text classification | [LC] Ch. 5 | |
| | 10/2 | Python tutorial 5 | | Student presentation topics submission |

# Brainstorm your research interests

| | | | | |
|---|---|---|---|---|
| 10 | 10/28 | Paper presentation (Papers 1, 2) | | |
| | 10/30 | Paper presentation (3, 4) | | |
| 11 | 11/4 | Paper presentation (5, 6) | | |
| | 11/6 | Paper presentation (7, 8) | | |
| 12 | 11/11 | Paper presentation (9) | | |
| | 11/13 | Paper presentation (10, 11) | | Assignment 1 |
| 13 | 11/18 | Paper presentation (12, 13) | | |
| | 11/20 | Paper presentation (14, 15) | | |
| 14 | 11/25 | Paper presentation (16, 17) | | |
| | 11/27 | **Thanksgiving break (No class)** | | |
| 15 | 12/2 | Paper presentation (18) | | |
| | 12/4 | Final wrap-up | | Assignment 2 |

# What needs to be decided (By October 2nd)

1. Review the sample papers on the course website (*https://hksung.github.io/Fall25_LING351/materials/*)

# What needs to be decided (By October 2nd)

1. Review the sample papers on the course website (*https://hksung.github.io/Fall25_LING351/materials/*)
2. Add your names to the shared sheet (*https://docs.google.com/spreadsheets/d/1on8icHoXUsj74m1UNEHk8CycHEAmVH1nRsUatpn9xYc/edit?usp=sharing*) - *First come first served*

# What needs to be decided (By October 2nd)

1. Review the sample papers on the course website (*https://hksung.github.io/Fall25_LING351/materials/*)
2. Add your names to the shared sheet (*https://docs.google.com/spreadsheets/d/1on8icHoXUsj74m1UNEHk8CycHEAmVH1nRsUatpn9xYc/edit?usp=sharing*) - *First come first served*
3. You may also choose articles beyond this list (e.g., CALL), but please check with me first

# What needs to be decided (By October 2nd)

1. Review the sample papers on the course website (*`https://hksung.github.io/Fall25_LING351/materials/`*)
2. Add your names to the shared sheet (*`https://docs.google.com/spreadsheets/d/1on8icHoXUsj74m1UNEHk8CycHEAmVH1nRsUatpn9xYc/edit?usp=sharing`*) - *First come first served*
3. You may also choose articles beyond this list (e.g., CALL), but please check with me first
4. I will form groups of 2–3 people based on your selections