IMPERIAL

Natural Language Processing
and
Large Language Models

Shamsuddeen Muhammad
Google DeepMind Academic Fellow,
Imperial College London
https://shmuhammadd.github.io/

About Me

- BSc CS at Bayero University, Kano Nigeria
- MSc CS, University of Manchester, UK
- PhD Machine Learning, University of Porto
- Senior Lecturer, Bayero University

- **Google DeepMind Fellow**, Imperial College London

About the Course !

- Natural language processing (NLP) is the field of working with **language to automatically** perform a variety of tasks.

- Recently, **large language models (LLMs)** like ChatGPT have changed the landscape of **modern NLP research.**

- This course will show you **both old & new techniques** that are used today and will give you a basic understanding of **why & how** we do NLP.

Prerequisite

- Python
- Machine Learning, Deep Learning
- Comfort with probability, linear algebra, and mathematical notation
- Foundational understanding of PyTorch, or familiarity with other deep learning frameworks like TensorFlow, will be beneficial.
- Willingness to learn

Course Topics

Part 1 — Natural Language Processing

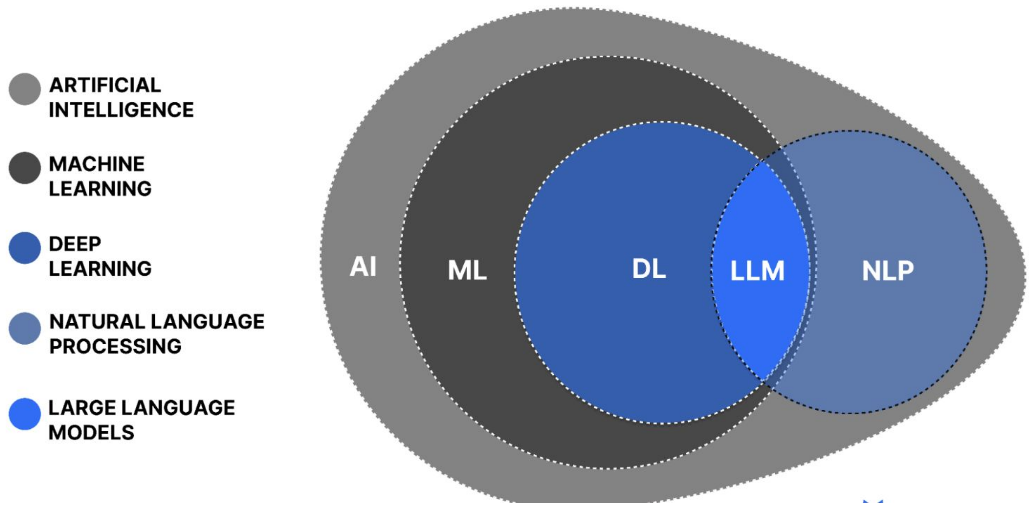
- Introduction to NLP and LLMs
- How Language Modelling Started (N-grams)
- Text Classification
- Word Vectors
- Sequence Modelling
- Attention

Part 2 — Large Language Models

- Introduction to Transformers
- Pretraining
- Post-training
- Model Compression
- Benchmarking and Evaluation

About the Course !

- **Two parts:**
  - **Part I: NLP** applies a combination of rule-based systems and machine learning to process text and speech efficiently.
  - **Part II: LLMs**, rely on deep learning for language (knowledge) comprehension

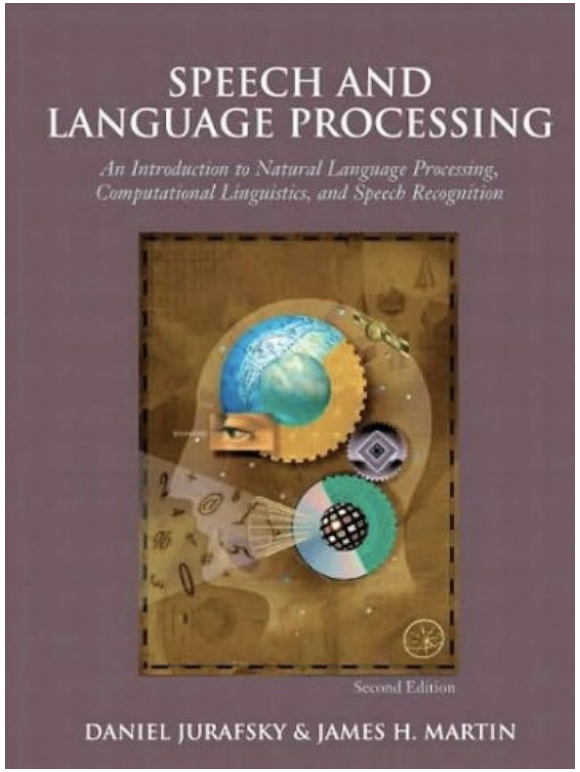


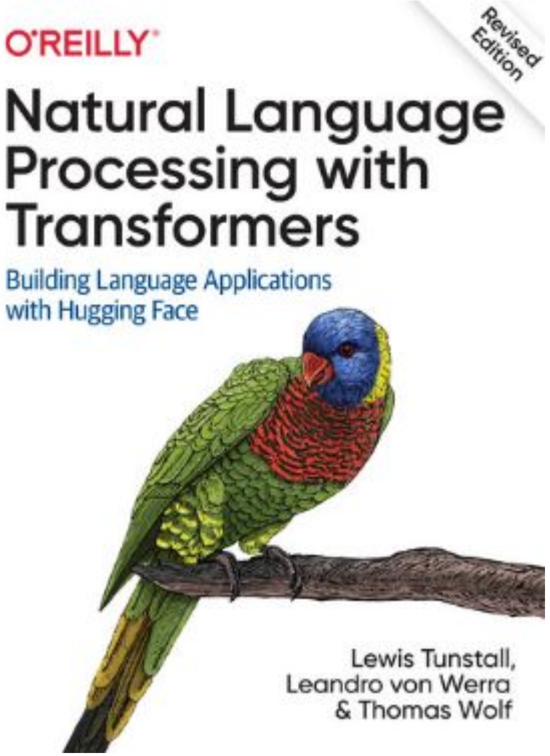
About the Course !

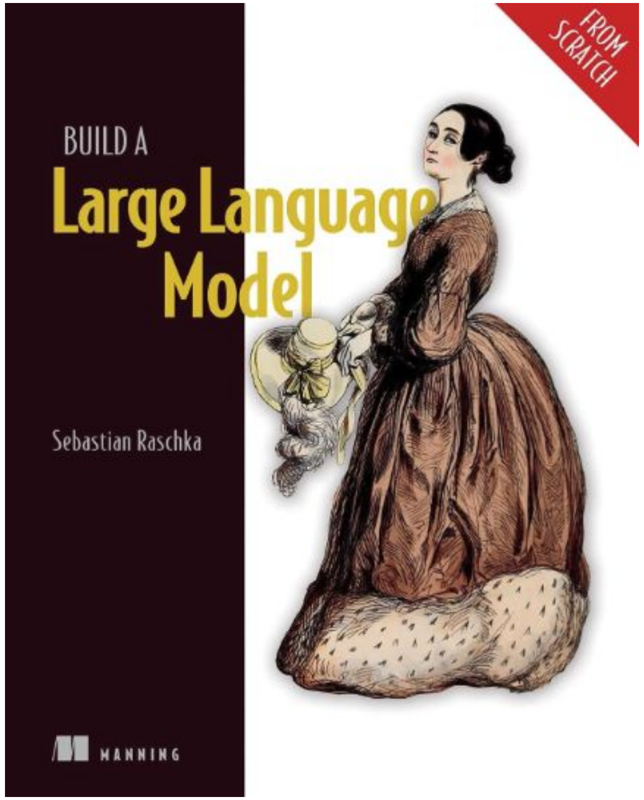
NLP vs LLMs

Aspect	Natural Language Processing (NLP)	Large Language Models (LLMs)
Data Requirement	Structured, labeled data	Large-scale, unstructured datasets
Computational Power	Low to moderate; can run on local machines	High-performance GPUs and cloud-based processing
Primary Use Cases	Sentiment analysis, translation, speech recognition, text classification	Conversational AI, content creation, coding assistance, document summarization
Flexibility	Task-specific and specialized	Adaptable across domains and capable of handling diverse queries
Cost	Lower infrastructure demands; more cost-effective	High due to extensive computational and storage requirements
Scalability	Easily scalable for structured applications	Requires significant cloud-based resources to scale effectively

Reference Books







Reference Books NLP

1. Speech and Language Processing, Dan Jurafsky and James H. Martin
   https://web.stanford.edu/~jurafsky/slp3/

2. Foundations of Statistical Natural Language Processing, Chris Manning and Hinrich Schütze

3. Build a Large Language Model (From Scratch): https://github.com/rasbt/LLMs-from-scratch

4. Hands-On Large Language Models: https://github.com/HandsOnLLM/Hands-On-Large-Language-Models

Other Sources

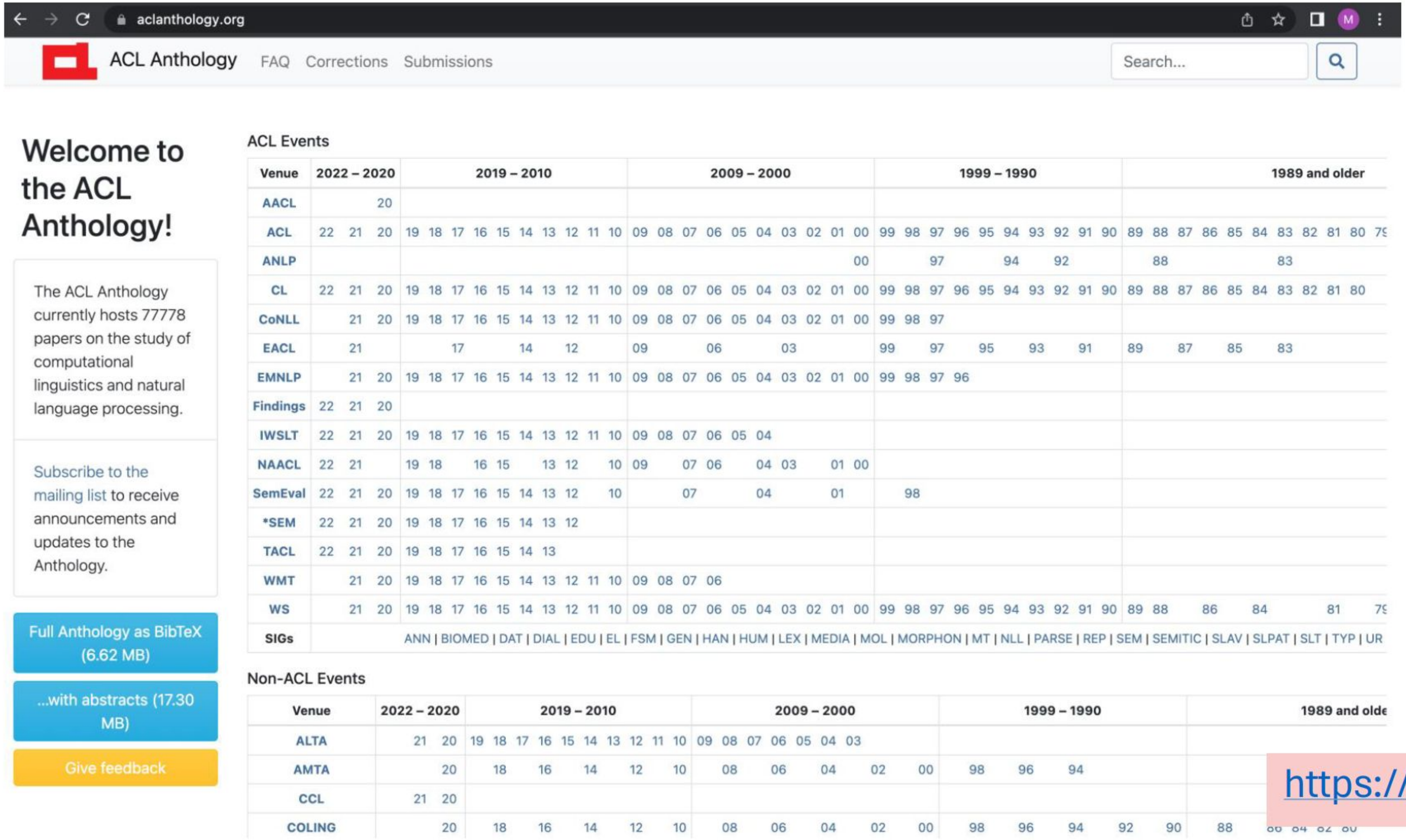
Journals

Computational Linguistics, Natural Language Engineering, TACL, JMLR, TMLR, etc

Conferences

ACL, EMNLP, NAACL, COLING, AAAI, IJCNLP, ICML, NeurIPS, ICLR, WWW, KDD, SIGIR, etc

ACL Anthology



ACL Events

Venue	2022 – 2020	2019 – 2010	2009 – 2000	1999 – 1990	1989 and older
AACL	20				
ACL	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 7
ANLP			00	97 94 92	88 83
CL	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80
CoNLL	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97	
EACL	21	17 14 12	09 06 03	99 97 95 93 91	89 87 85 83
EMNLP	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96	
Findings	22 21 20				
IWSLT	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04		
NAACL	22 21	19 18 16 15 13 12 10	09 07 06 04 03 01 00		
SemEval	22 21 20	19 18 17 16 15 14 13 12 10	07 04 01	98	
*SEM	22 21 20	19 18 17 16 15 14 13 12			
TACL	22 21 20	19 18 17 16 15 14 13			
WMT	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06		
WS	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 86 84 81 7
SIGs	ANN | BIOMED | DAT | DIAL | EDU | EL | FSM | GEN | HAN | HUM | LEX | MEDIA | MOL | MORPHON | MT | NLL | PARSE | REP | SEM | SEMITIC | SLAV | SLPAT | SLT | TYP | UR				

Non-ACL Events

Venue	2022 – 2020	2019 – 2010	2009 – 2000	1999 – 1990	1989 and olde
ALTA	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03		
AMTA	20	18 16 14 12 10	08 06 04 02 00	98 96 94	
CCL	21 20				
COLING	20	18 16 14 12 10	08 06 04 02 00	98 96 94 92 90	88

https://aclanthology.org/

ArXiv

arXiv.org > cs > cs.CL

Computation and Language

Authors and titles for recent submissions

- Wed, 19 Aug 2020
- Tue, 18 Aug 2020
- Mon, 17 Aug 2020
- Fri, 14 Aug 2020
- Thu, 13 Aug 2020

[total of 84 entries: **1-25** | 26-50 | 51-75 | 76-84]
[showing 25 entries per page: fewer | more | all]

Wed, 19 Aug 2020

[1] **arXiv:2008.07905** [**pdf**, **other**]
Glancing Transformer for Non-Autoregressive Neural Machine Translation
Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, Lei Li
Comments: 11 pages, 3 figures, 4 tables
Subjects: **Computation and Language (cs.CL)**

[2] **arXiv:2008.07880** [**pdf**, **other**]
COVID-SEE: Scientific Evidence Explorer for COVID-19 Related Research
Karin Verspoor, Simon Šuster, Yulia Otmakhova, Shevon Mendis, Zenan Zhai, Biaoyan Fang, Jey Han Lau, Timothy Bal
Comments: COVID-SEE is available at this http URL
Subjects: **Computation and Language (cs.CL)**; Information Retrieval (cs.IR)

[3] **arXiv:2008.07772** [**pdf**, **other**]
Very Deep Transformers for Neural Machine Translation
Xiaodong Liu, Kevin Duh, Liyuan Liu, Jianfeng Gao
Comments: 6 pages, 3 figures and 3 tables
Subjects: **Computation and Language (cs.CL)**

[4] **arXiv:2008.07723** [**pdf**, **other**]
NASE: Learning Knowledge Graph Embedding for Link Prediction via Neural Architecture Search
Xiaoyu Kou, Bingfeng Luo, Huang Hu, Yan Zhang
Comments: Accepted by CIKM 2020, short paper
Subjects: **Computation and Language (cs.CL)**

https://arxiv.org/list/cs.CL/recent

Acknowledgments

- Advanced NLP, Graham Neubig http://www.phontron.com/class/anlp2022/
- Advanced NLP, Mohit Iyyer https://people.cs.umass.edu/~miyyer/cs685/
- NLP with Deep Learning, Chris Manning, http://web.stanford.edu/class/cs224n/
- Understanding Large Language Models, Danqi Chen https://www.cs.princeton.edu/courses/archive/fall22/cos597G/
- Natural Language Processing, Greg Durrett https://www.cs.utexas.edu/~gdurrett/courses/online-course/materials.html
- Large Language Models: https://stanford-cs324.github.io/winter2022/
- Natural Language Processing at UMBC, https://laramartin.net/NLP-class/
- Computational Ethics in NLP, https://demo.clab.cs.cmu.edu/ethical_nlp/
- Self-supervised models, CS 601.471/671: Self-supervised Models (jhu.edu)
- WING.NUS Large Language Models, https://wing-nus.github.io/cs6101

What is Natural Language Processing (NLP)?

What is Natural Language Processing (NLP) ?

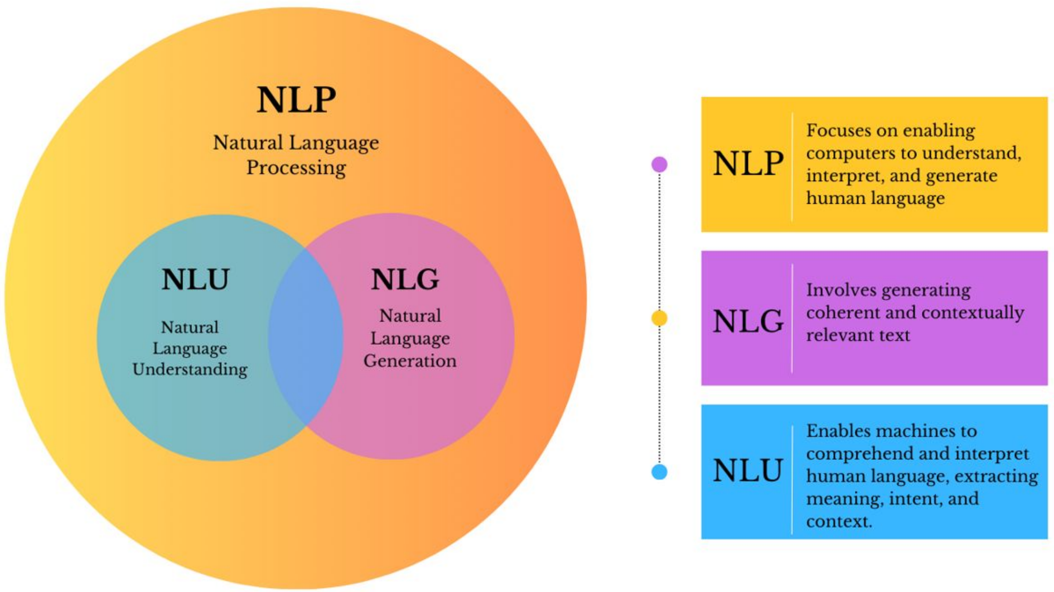
- **Natural Language Processing (NLP)** is a field of artificial intelligence focused on enabling machines to **understand**, **interpret**, and **generate** human language.

- **NLP** combines methods from linguistics, machine learning, and computer science to build systems that work with **text** and **speech** at scale.

- NLP includes two major subfields:

    - **NLU – Natural Language Understanding:** extracting meaning, intent, entities, and structure.

    - **NLG – Natural Language Generation:** producing coherent, context-appropriate text or speech.



https://geekflare.com/blog/natural-language-understanding/

NLU vs NLG

🧠 NLU
Natural Language Understanding

📊 Sentiment Analysis

INPUT TEXT:
"This movie was absolutely amazing! I loved every minute of it."

⬇️

ANALYSIS:
Sentiment: Positive ✅
Confidence: 95%
Emotion: Joy, Excitement

❓ Question Answering

CONTEXT:
"Albert Einstein was born in 1879 in Germany. He developed the theory of relativity."

QUESTION:
"When was Einstein born?"

⬇️

ANSWER:
1879

✍️ NLG
Natural Language Generation

🌐 Machine Translation

SOURCE (English):
"Hello, how are you today?"

⬇️

TRANSLATION (Spanish):
"Hola, ¿cómo estás hoy?"

📝 Summarization

ORIGINAL TEXT:
"Climate change is one of the most pressing issues of our time. Rising global temperatures are causing ice caps to melt, sea levels to rise, and extreme weather events to become more frequent. Scientists warn that without immediate action, the consequences could be catastrophic for future generations."

⬇️

SUMMARY:
"Climate change threatens the planet with melting ice caps and rising sea levels, requiring immediate action to prevent catastrophic consequences."

NLU vs NLG

NLU Tasks

- **Sentiment Analysis** – Determining emotional tone (positive/negative/neutral)
- **Question Answering** – Extracting answers from context
- **Named Entity Recognition** – Identifying people, places, organizations
- **Intent Classification** – Understanding user's goal or intention
- **Text Classification** – Categorizing documents into topics
- **Information Extraction** – Extracting structured data from text
- **Relation Extraction** – Finding relationships between entities

NLG Tasks

- **Machine Translation** – Converting text from one language to another
- **Summarization** – Creating concise summaries of documents
- **Text Generation** – Creating coherent original text
- **Paraphrasing** – Rewriting text with same meaning
- **Report Generation** – Creating reports from structured data
- **Image Captioning** – Generating descriptions for images
- **Dialogue Generation** – Creating conversational responses

Hybrid Tasks

- **Conversational AI** – Chatbots combining understanding & generation
- **Semantic Search** – Understanding queries + retrieving relevant results
- **Reading Comprehension** – Understanding + answering questions
- **Speech Recognition** – Converting speech to text
- **Text-to-Speech** – Converting text to spoken words
- **Document Parsing** – Extracting structure and content
- **Virtual Assistants** – Complete NLU + NLG pipeline

NLU: Understanding Input · **NLG: Generating Output** · **Hybrid: Both NLU + NLG**

Natural Language Processing and Computational Linguistics

Natural Language Processing
(NLP)

Goal: Build practical applications that solve real-world problems involving human language.

Focus: "How can we make computers DO things with language?"

Real-World Applications:

- **Speech Recognition:** Converting voice to text (Siri, Alexa)
- **Machine Translation:** Google Translate, DeepL
- **Information Extraction:** Pulling key facts from documents automatically
- **Chatbots:** Customer service bots, virtual assistants
- **Spam Detection:** Filtering unwanted emails

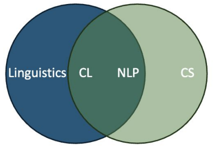
Computational Linguistics
(CL)

Goal: Understand how human language works using computational methods and models.

Focus: "How does language actually WORK in the human mind and brain?"

Research Questions:

- **? How do we understand language?**
  Example: How do children learn grammar rules without being explicitly taught?
- **? How do we produce language?**
  Example: How does the brain decide which words to use in which order?
- **? How do we learn language?**
  Example: What makes some languages easier to learn than others?
- **? What are universal language structures?**
  Example: Are there grammar rules common to all human languages?



The computational **study** of language

Computational Linguistics

≈

Natural Language Processing

The computational **use** of language

Association for Computational Linguistics

Both fields work with human language using computers, but they have different goals and perspectives!

Natural Language Processing and Computational Linguistics

- Most of the **conferences** and **journals** that host natural language processing research bear the name “computational linguistics” (e.g., A**CL**, N**ACL**).

- NLP and CL may be thought of as essentially synonymous.

- While there is substantial overlap, there is an important difference in focus
  - CL is essentially linguistics supported by computational methods (similar to computational biology, and computational astronomy)
  - NLP focuses on solving well-defined tasks involving human language (e.g., translation, query answering, holding conversations).

Natural Language Processing and Computational Linguistics

- Most of the **conferences** and **journals** that host natural language processing research bear the name “computational linguistics” (e.g., A**CL**, N**ACL**).

- NLP and CL may be thought of as essentially synonymous.

Google Scholar

Top publications

Categories > Engineering & Computer Science > **Computational Linguistics**

	Publication	h5-index	h5-median
1.	Meeting of the Association for Computational Linguistics (ACL)	236	387
2.	Conference on Empirical Methods in Natural Language Processing (EMNLP)	218	323
3.	Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)	126	202
4.	Transactions of the Association for Computational Linguistics	96	204
5.	International Conference on Computational Linguistics (COLING)	81	122
6.	Conference of the European Chapter of the Association for Computational Linguistics (EACL)	77	128
7.	International Conference on Language Resources and Evaluation (LREC)	68	108
8.	Computer Speech & Language	47	84
9.	IEEE Spoken Language Technology Workshop (SLT)	44	71
10.	Computational Linguistics	41	109
11.	International Joint Conference on Natural Language Processing (IJCNLP)	41	77
12.	International Workshop on Semantic Evaluation	40	72
13.	Workshop on Machine Translation	39	80
14.	ACM Transactions on Asian and Low-Resource Language Information Processing	38	56
15.	Language Resources and Evaluation	36	66
16.	Conference on Empirical Methods in Natural Language Processing: System Demonstrations	35	80
17.	Arabic Natural Language Processing Workshop	30	54
18.	Natural Language Engineering	30	54
19.	Conference on Computational Natural Language Learning (CoNLL)	30	47
20.	Biomedical Natural Language Processing	29	49

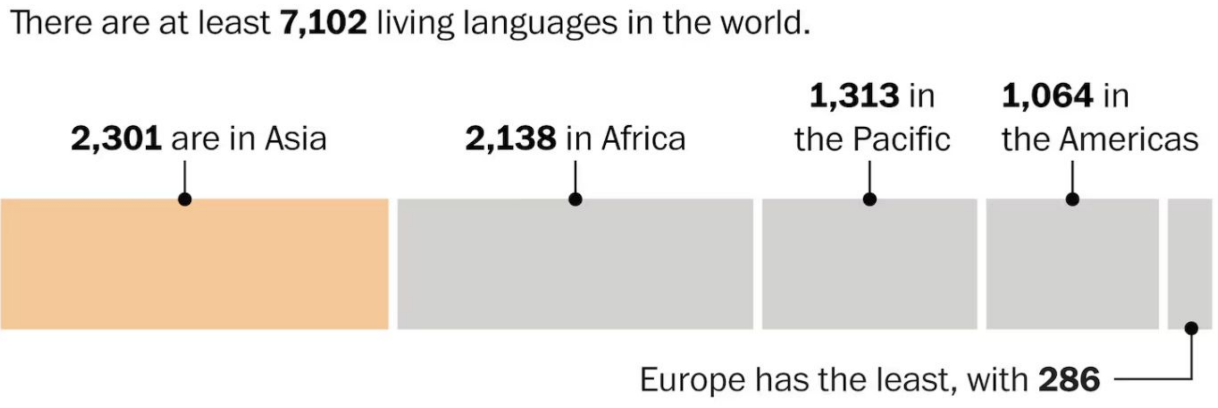
Dates and citation counts are estimated and are determined automatically by a computer program.

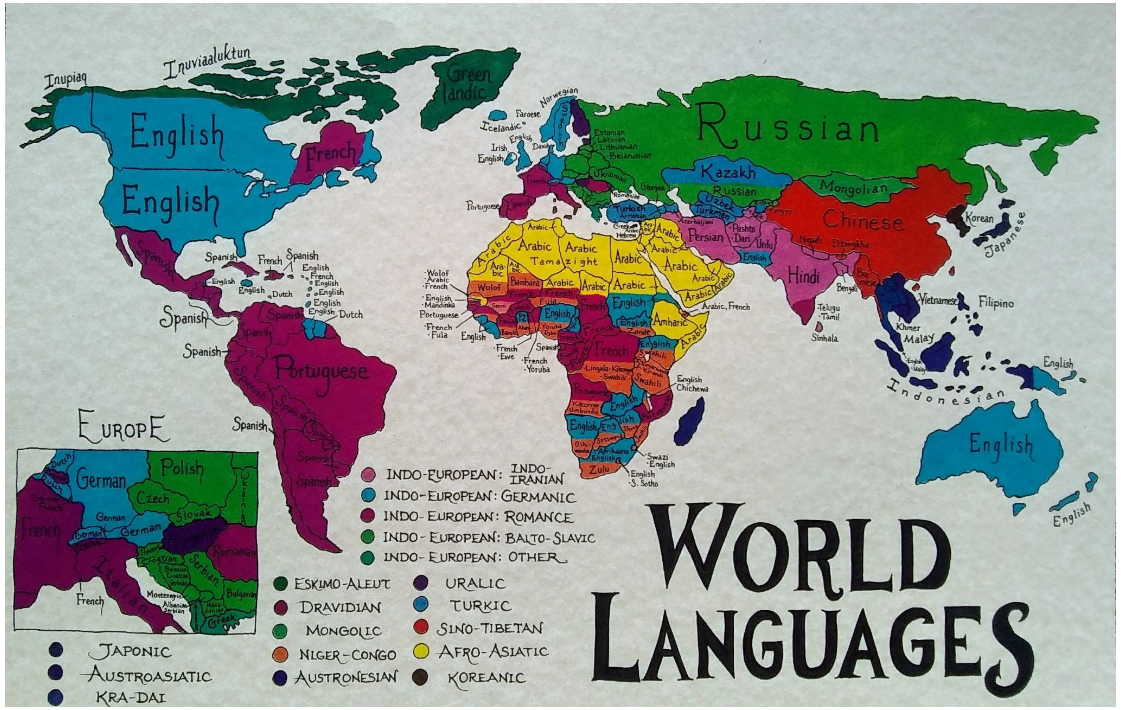
Natural Language Processing and Computational Linguistics

While there is substantial overlap, there is an important difference in focus

- **CL** is essentially linguistics supported by computational methods (similar to computational biology, and computational astronomy)
- **NLP** focuses on solving well-defined tasks involving human language (e.g., translation, query answering, holding conversations).

Why is NLP interesting?





NLP powers a broad range of applications: Machine translation, information extraction, question answering, summarization, sentiment analysis, speech technologies, code generation, document retrieval, fact checking, etc.

The field is also rapidly evolving with the rise of large language models, offering new research challenges

Before Building NLP Systems…

- We need **large, high-quality text data**.
    - But the world's **7,000+ languages** are not equally represented in digital form.
    - This leads to major gaps: many languages have **little or no available corpus**, making them "**low-resource languages (e.g., African Languages)**."
    - we must understand **how it is collected**, and **why it matters** for multilingual NLP.

Where does the data come from?

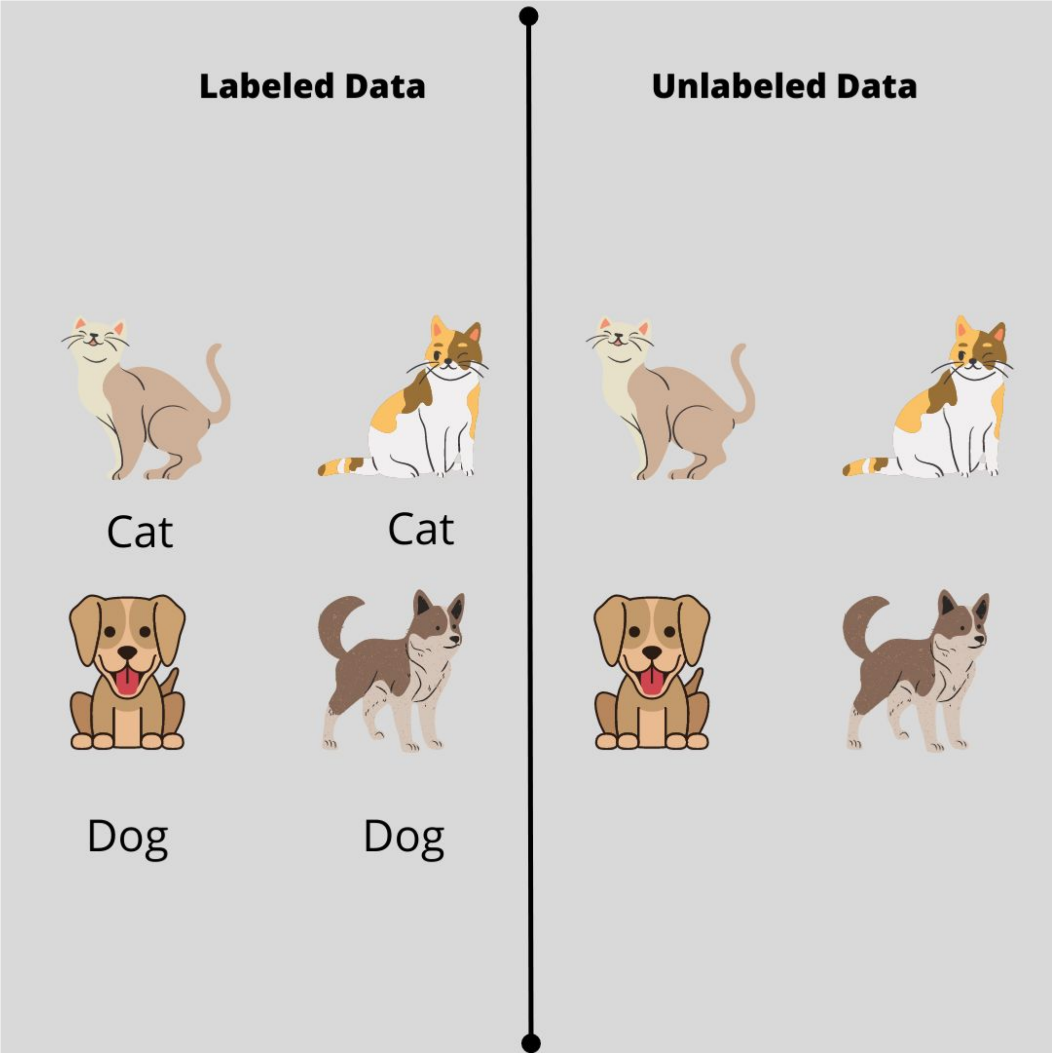
Corpus and Low-Resource Languages

- **Corpus** (plural: *corpora*): a structured collection of text used for training or evaluating NLP models.
- Languages with limited corpora are known as **low-resource languages**.

How Corpora Are Collected

- **Expert-curated data:** manually tagged and organized by linguists or annotators.
- **Open-web data:** collected from freely accessible sources (e.g., Wikipedia, blogs, forums).
- **Permission-based data:** obtained from closed platforms (e.g., messaging apps, social media) with explicit consent, less common but often higher quality.

Corpora



Unlabelled corpora

- An **unlabelled corpus** is a collection of text (or speech) data where each instance has no explicit human

- Typically you only have the raw content (and possibly metadata like date/source).

Examples

  - A dump of news articles, tweets, Wikipedia pages, web crawl text.
  - Audio recordings with no transcripts.
  - Sentences without tags such as sentiment, topic, entity spans, etc.

Labelled corpora

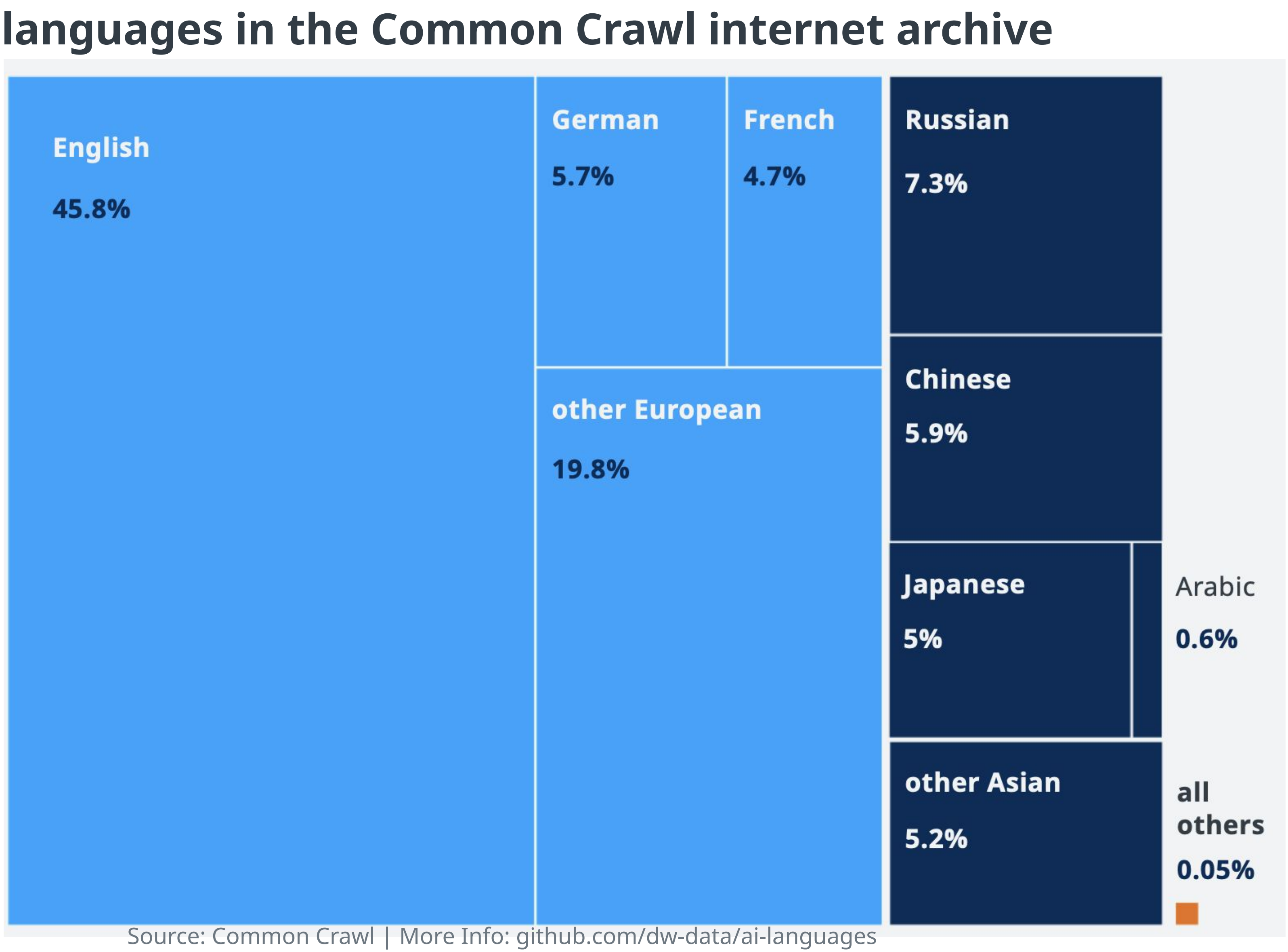
A **labelled corpus** is a collection of data where each instance is paired with annotations (“labels”) that encode desired information for a specific task.

Labels can be categorical, numeric, structured spans, trees, etc.

Examples

- Sentences labelled with sentiment: positive / negative / neutral.
- Token-level named entity labels: PER/ORG/LOC spans.
- Parallel corpora for machine translation: (source sentence, target sentence) pairs.

African Languages are low-resource



30%

World languages are African (Ethnologue)

0.05%

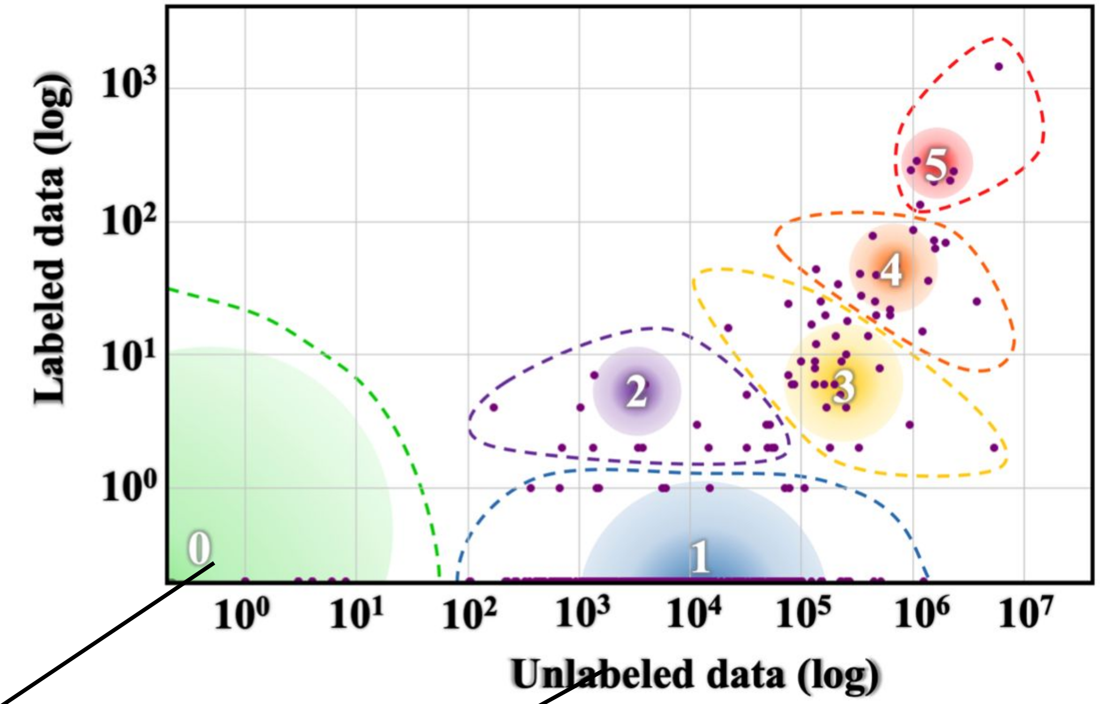
Under-resourced languages: Labelled+Unlabelled data

Six-class categorization of languages based on Joshi et al (2020)

categorization of languages based on the amount

of NLP resources available for each language

- Unlabelled corpora
- Labelled corpora



No unlabelled
texts
80% of languages

Few texts

Six-class categorization

Highly Resourced (HRL)- Winners

- Languages with extensive NLP resources, including large-scale corpora, pre-trained models, and strong computational tools.
- Examples: English, Chinese, Spanish, French

Moderately Resourced (MRL) - Moderate

- Languages with reasonable NLP resources, but still lacking in some areas such as large-scale pre-trained models.
- Examples: Dutch, Russian, Korean

Somewhat Resourced (SRL) - Hopefuls

- Languages with limited but growing NLP resources, including some datasets and a few pre-trained models.
- Examples: Swahili, Finnish, Turkish

Six-class categorization

Low Resourced (LRL) - Scraping By

- Languages with very limited annotated data, small corpora, and minimal computational resources.
- Examples: Hausa, Tamil, Uzbek

Extremely Low Resourced (XLRL) - Left Behind

- Languages with almost no NLP resources, where some digitized text may exist, but annotated corpora and NLP tools are scarce.
- Examples: Wolof, Aymara, Tigrinya

Unsupervised (UL) - The Rest

- Languages with virtually no digital footprint, requiring unsupervised or few-shot learning techniques for any NLP progress.
- Examples: Many indigenous and endangered languages like Chadic languages, Amazonian languages

Why is NLP hard?

Ambiguity

NLP is challenging due to the inherent complexities of human language.

Lexical Ambiguity

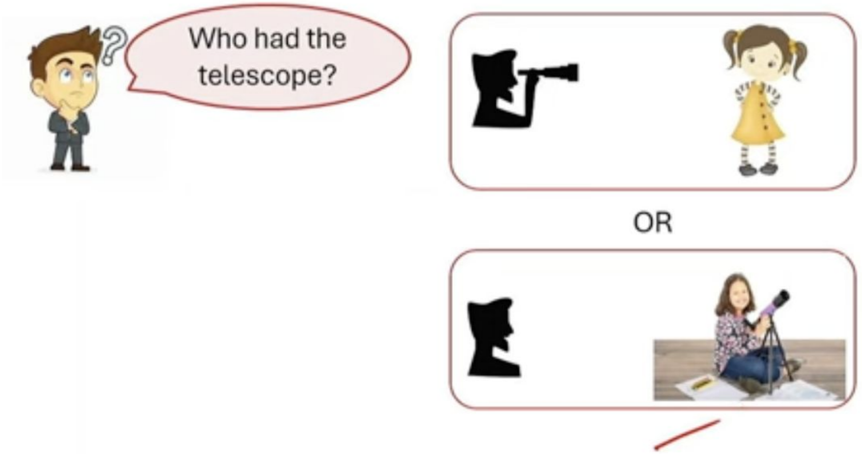
A single word can have multiple meanings, and the intended meaning depends on the context.

Example: The bank is closed today. (Does "bank" refer to a financial institution or the side of a river?)

Syntactic Ambiguity

A sentence can have more than one valid structure, leading to different interpretations.

Example: I saw a girl with the telescope.



Possible meanings:

- I used a telescope to see the girl.
- The girl is the one holding the telescope.

Ambiguity in Language

- I ate food with Spoon

  "with" = tool used to eat

- I ate rice with curd

  "with" = ingredient served together

- I ate rice with Muhammad

  "with" = person I ate together with

Semantic Ambiguity

A sentence can have more than one meaning because its interpretation depends on context.

Example: : “The chicken is ready to eat.”

Possible interpretations:

- The chicken is **going to eat** something
- The chicken is **cooked and ready to be eaten**.

Pragmatic Ambiguity

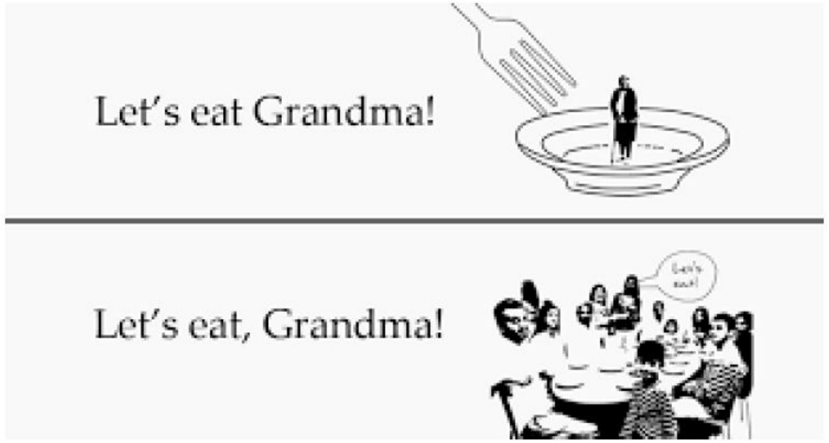
Meaning depends on context, social norms, and the speaker's intention, not just the words themselves.

Example: *"Can you pass the salt?"*

Two interpretations:

- **Literal:** "Are you able to pass the salt?" (asking about ability)
- **Pragmatic:** A polite way to say **"Please pass the salt."**

Ambiguity in Punctuation



A woman without her man is nothing.

A woman, without her man, is nothing.

A woman: without her, man is nothing.

Punctuation is powerful.

How NLP Overcomes Language Challenges

- Human language is full of challenges: lexical, syntactic, semantic, and pragmatic ambiguities.
- These difficulties vary across languages and are even more complex in **low-resource** contexts. But despite these challenges:
    - NLP provides tools to interpret meaning.
    - resolve ambiguity using context and large datasets,
    - and build systems that can understand and generate language with high accuracy.

In this course, we will explore how modern NLP, especially transformers and LLMs, addresses these challenges and how these methods can be applied to many languages, including those with limited resources.

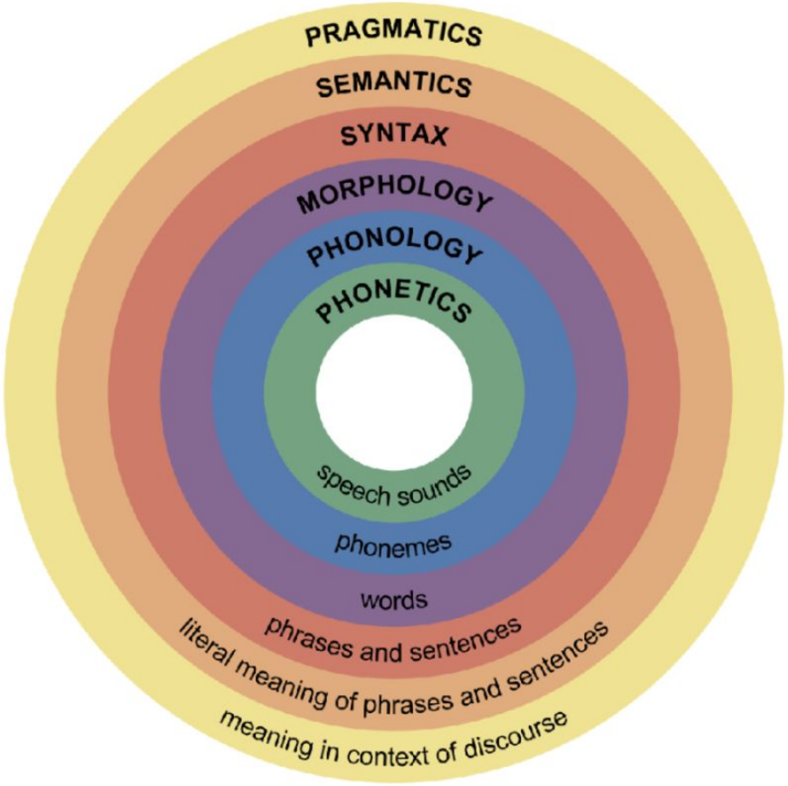
Break

NLP Layers

- Understanding the semantics is a non-trivial task.
- Needs to performs a series of incremental tasks to achieve this.
- NLP happens in layers.

Pragmatics & Discourse	*Study of semantics in context.*
Semantics	*Meaning of the sentence.*
Parsing	*Syntactic structure of the sentence.*
Chunking	*Grouping of meaningful phrases.*
Part of speech tagging	*Grammatical classes.*
Morphology	*Study of word structure.*

Increasing Complexity Of Processing



Morphology

Morphology is the study of how words are formed and structured.

It looks at the smallest meaningful units (called **morphemes**) like prefixes, suffixes, and root words.

Different languages handle morphology differently -

some use very little word modification (morphologically poor), while others heavily modify words (morphologically rich).

Morphology helps computers understand how words are built and there two types/

Inflectional Morphology and Derivational Morphology

Morphology

Morphology is the part of linguistics that studies word structure, how words are built from smaller meaningful pieces.

Those smaller pieces are called **morphemes**: the smallest units that carry meaning or grammatical function.

- **Root / stem:** the core meaning part
  - play in play-ed, play-er, re-play
- **Prefix:** comes before the root
  - un- in un-happy, re- in re-write
- **Suffix:** comes after the root
  - -ed in walk-ed, -s in cat-s, -er in teach-er

A quick on-board demo:

- un + help + ful + ness → unhelpfulness
  (each piece contributes something)

Morphologically “poor” Vs Morphologically “rich”

- **Morphologically poor** : words usually don’t change much. Grammar is often shown with separate words and word order.
  - Example idea (English-like): “will go”, “did go”, “more beautiful” (grammar partly outside the word).
- **Morphologically rich**: words change a lot, often adding prefixes/suffixes (or other changes) to pack tense, person, number, case, etc. inside one word.

Examples of languages

- More morphology-poor: Mandarin Chinese, Vietnamese, Thai.
- More morphology-rich: Turkish, Finnish, Swahili, Hungarian.

Inflectional Morphology

Changes the form of a word (tense, number, gender) without changing its core meaning.

Examples: walk → walked (past tense), cat → cats (plural)

Why this matters for NLP:

- **Tokenization & vocabulary size:** Inflected forms multiply the number of word types models must handle.
- **Lemmatization:** Systems need to group different forms into the same base word (*walk*).
- **POS tagging & parsing:** Inflections signal tense, plurality, agreement, etc.
- **Low-resource languages:** Richly inflected languages (e.g., Amharic, Arabic, Hausa) create data sparsity.

Derivational Morphology

Derivation creates **new words** with related meanings:

- *happy* → *unhappy*
- *teach* → *teacher*
- *kind* → *kindness*

Why this matters for NLP:

- **Word embeddings:** Models must learn that *teach* and *teacher* are related but not identical.
- **Sentiment analysis:** Prefixes like *un-* or *dis-* change polarity.
- **Text classification:** Derivational patterns signal topic or domain (*biology, biological, biologist*).

Why morphology matters in NLP

- If a system treats *connect, connected, connecting, connection as unrelated tokens,* it wastes data.

- **Morphology** lets you link them via shared morphemes (connect + **-ed**, **-ing**, **-ion**).

Part-of-Speech Tagging (POS)

- **Part of Speech (PoS)** refers to the grammatical category of words in a sentence based on their function and meaning.

- **PoS** tagging is essential in NLP for understanding sentence structure and meaning.

Grammatical class of the word.

He	ate	an	apple	.
PRP	VBD	DT	NN	.

PoS disambiguation:

- A word can belong to different grammatical classes.

He	went	to	the	***park***	in	a	car	.
PRP	VBD	TO	DT	***NN***	IN	DT	NN	.

They	went	to	***park***	the	car	in	the	shed	.
PRP	VBD	TO	***VB***	DT	NN	IN	DT	NN	.

Tags

PRP: Personal Pronoun
VBD: Verb, Past
DT: Determiner
NN: Noun, Singular, Mass
TO: *to*
IN: Preposition

- 45 tags in Penn Treebank tagset
- 146 tags in C7

Chunking

Chunking is the process of grouping words into meaningful phrases based on their Part of Speech (PoS) tags.

It helps in identifying syntactic structures like noun phrases (NP), verb phrases (VP), and prepositional phrases (PP).

Example of Chunking

Consider the sentence:

"The quick brown fox jumps over the lazy dog."

Chunking

Step 1: PoS Tagging

The (DT) quick (JJ) brown (JJ) fox (NN) jumps (VBZ) over (IN) the (DT) lazy (JJ) dog (NN)

Step 2: Chunking (Noun Phrases & Verb Phrases)

[NP The quick brown fox] [VP jumps] [PP over] [NP the lazy dog]

Here, the **Noun Phrases (NP)** and **Verb Phrases (VP)** are extracted.

Semantics

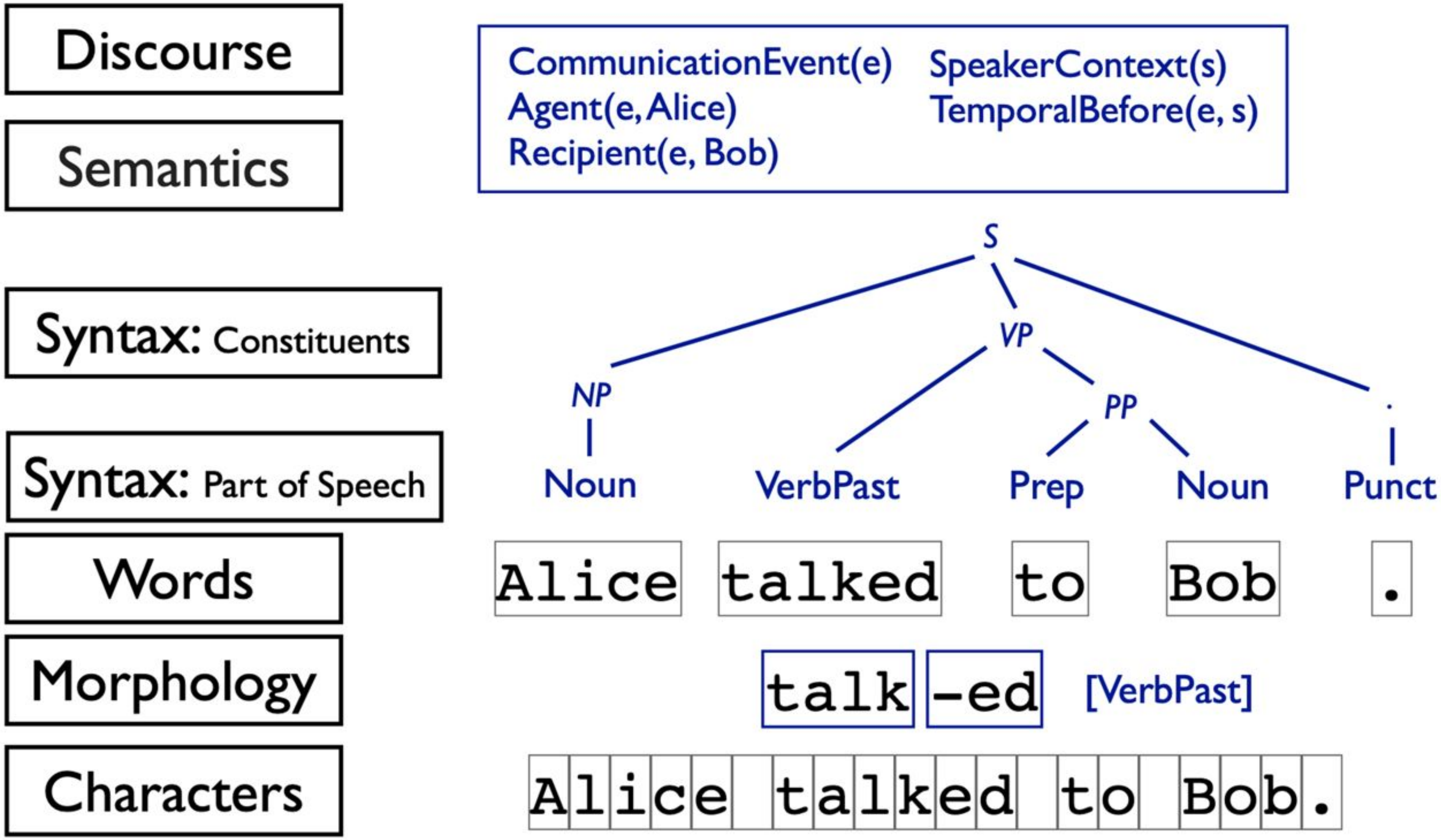
- **Semantics** is the study of meaning in language—how words, phrases, and sentences convey meaning.
- Semantic analysis helps NLP systems:
  - understand what a sentence means, not just what words it contains,
  - interpret words based on context (“bank” = money vs. riverbank),
  - capture relationships between words (who did what, to whom),
  - generate meaningful and coherent text.

Pragmatics & Discourse

Meaning in context: speaker intent, implications, and cross-sentence coherence.

Definition	Example

• Pragmatics: how context changes interpretation (e.g., requests, implicatures).
• Discourse: relations across sentences (e.g., coreference, coherence).
	`Pragmatics:`
	`“Can you pass the salt?” → polite request (not a yes/no question)`
	
	`Discourse / coreference:`
	`“Alice dropped the glass. It shattered.”`
	`It → the glass`



Discourse
Semantics
Syntax: Constituents
Syntax: Part of Speech
Words
Morphology
Characters
CommunicationEvent(e)
Agent(e, Alice)
Recipient(e, Bob)
SpeakerContext(s)
TemporalBefore(e, s)
S
VP
NP
PP
.
Noun
VerbPast
Prep
Noun
Punct
Alice
talked
to
Bob
.
talk
-ed
[VerbPast]
Alice talked to Bob.

Task we want to solve in NLP?

NLP Tasks

Understanding Tasks

Text Classification
Sentiment analysis, topic classification

Named Entity Recognition
Finding people, places, organizations

Part-of-Speech Tagging
Identifying grammatical roles

Dependency Parsing
Understanding sentence structure

Question Answering
Answering natural-language queries

Generation Tasks

Machine Translation
Converting text between languages

Summarization
Producing concise summaries

Text Generation
Writing coherent sentences or documents

Dialogue Systems / Chatbots
Human-like conversation

Paraphrasing
Rewriting text with same meaning

NLP Tasks Overview

NLP Tasks

Speech & Multimodal Tasks

Speech Recognition
Speech → text conversion

Text-to-Speech
Text → speech conversion

Vision–Language Tasks
Image captioning, Visual Question Answering (VQA)

Low-Level / Core Tasks

Tokenization and Segmentation
Breaking text into words, sentences, or characters

Lemmatization and Stemming
Reducing words to their base form

Morphological Analysis
Analyzing word structure and inflections

Coreference Resolution
Determining who/what pronouns refer to

NLP Tasks Overview

IMPERIAL

Q and A