IMPERIAL

Introduction to Natural Language Processing

Shamsuddeen Muhammad
Advanced Research Fellow, Imperial College London
https://shmuhammadd.github.io/
18/03/2025

Who is Shamsuddeen?

- Bs CS at Bayero University, Kano Nigeria

BAYERO UNIVERSITY KANO

- MSc CS, University of Manchester, UK

MANCHESTER 1824
The University of Manchester

- PhD Machine Learning, University of Porto

MAPi DOCTORAL PROGRAMME IN COMPUTER SCIENCE

- Advanced Research Fellow/Google DeepMind Fellow, Imperial College London

- Senior Lecturer, Bayero University

IMPERIAL

2nd in the world

1st in the UK and Euro

QS World University Rankings 2025

https://shmuhammadd.github.io/

What do I work on?

- **Computational Social Science:** Sentiment Analysis, Hate Speech Detection, Emotion Analysis

- **Low-Resource Languages**: Dataset Creation, Data Annotation, LLM Evaluation, and Adaptation

- **Multilingual NLP**: Enhancing Models for Underrepresented Languages and Cultures

Tutors

Tutors: Jeff, Miriam and Ranya

About you?

AIMS | African Institute for Mathematical Sciences CAMEROON

About the Course !

- Natural language processing (NLP) is the field of working with language to automatically perform a variety of tasks, instead of or in collaboration with people.

- NLP can focus on **the Generation (NLG**) and/or **Understanding (NLU)** of natural language.

- Recently, large language models (LLMs) like ChatGPT have gotten the attention of the general public, but they have also greatly changed the landscape of modern NLP research.

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

Learning Objective

By the end of the course, you will be able to…

- *a solid theoretical and practical understanding of contemporary machine learning models designed for processing human language and the underlying computational properties of NLP models.*
- *The student will have learned how underlying linguistic phenomena, that is, the linguistic features, can be modelled and automatically learned from data using deep learning techniques.*

Course Content

- **Week 1: Foundations of NLP & Classical Approaches**
- **Week 2: Transformers, Transfer Learning & Fine-Tuning**
- **Week 3: Advanced Topics, Knowledge Graphs & Ethical NLP**

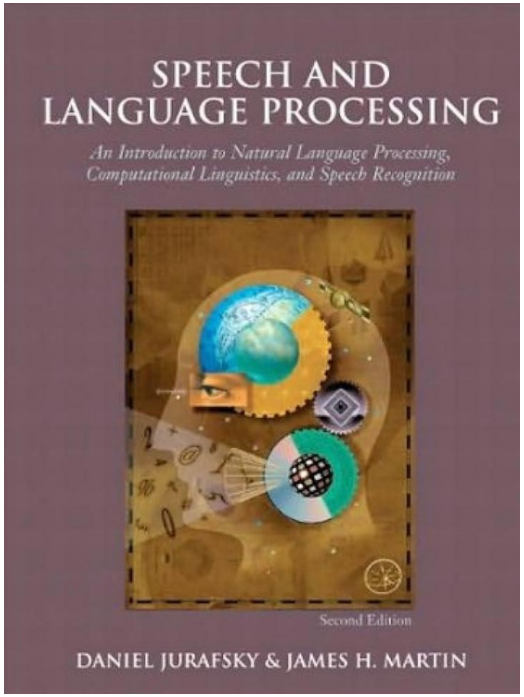
We will be discussing state-of-the-art papers about large language models.

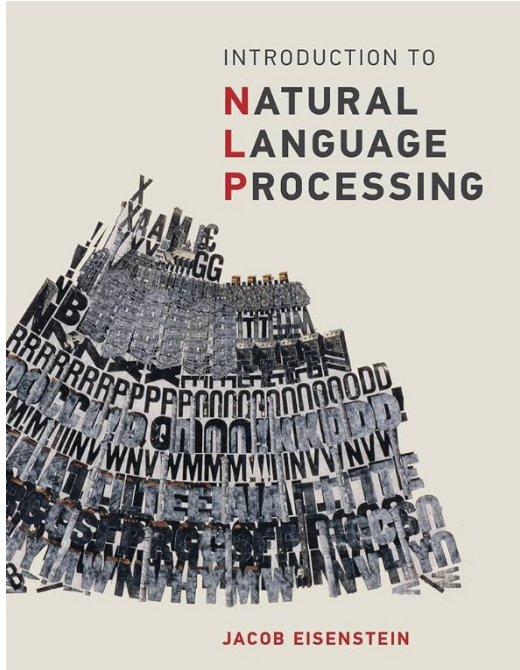
Pre-requisites

- Machine Learning
- Python Programming
- Deep learning

Materials

- **Course page:** https://github.com/shmuhammadd/aims-nlp-course
- Assignment descriptions





Books

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

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

Natural Language Processing, Jacob Eisenstein https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

A Primer on Neural Network Models for Natural Language Processing, Yoav Goldberghttp://u.cs.biu.ac.il/~yogo/nnlp.pdf

Conference and Journal

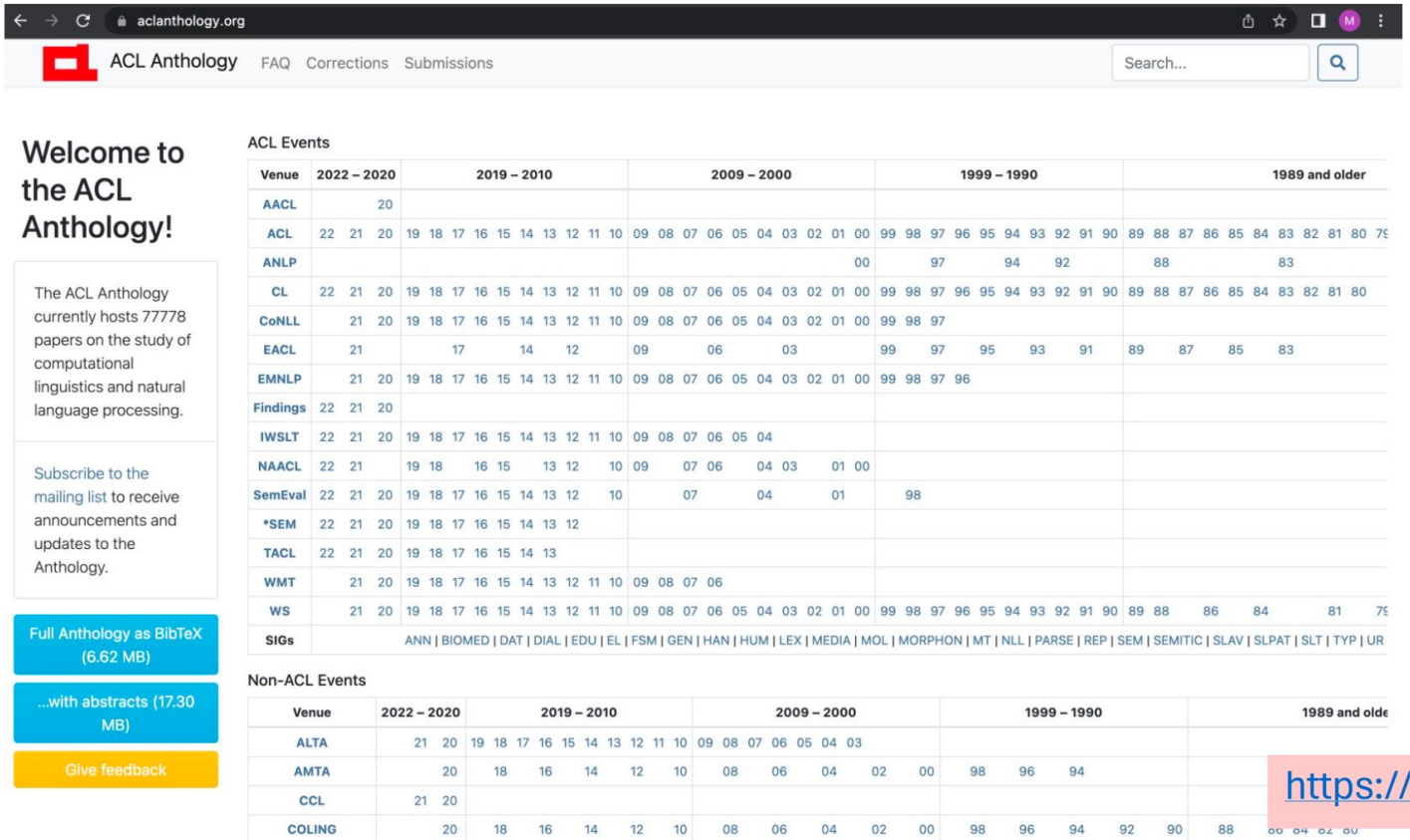
Journals

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



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

Office Hours

By appointment: s.muhammad@imperial.ac.uk

What is Natural Language Processing (NLP)?

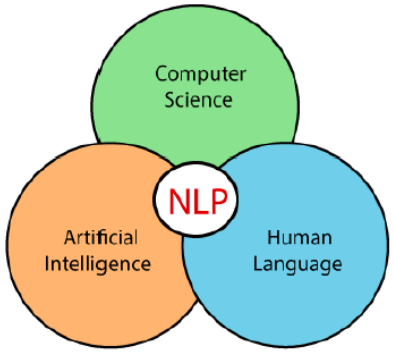
Natural Language Processing

What is a Natural Language?

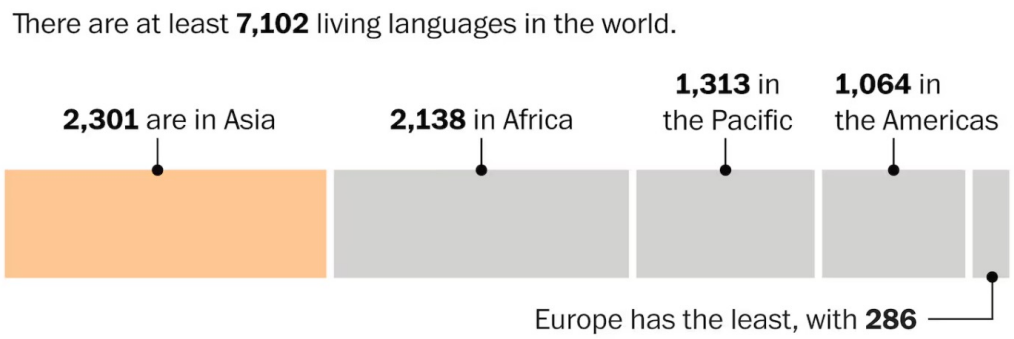
Any language that has evolved naturally in humans through use and repetition without conscious planning or pre-meditation.

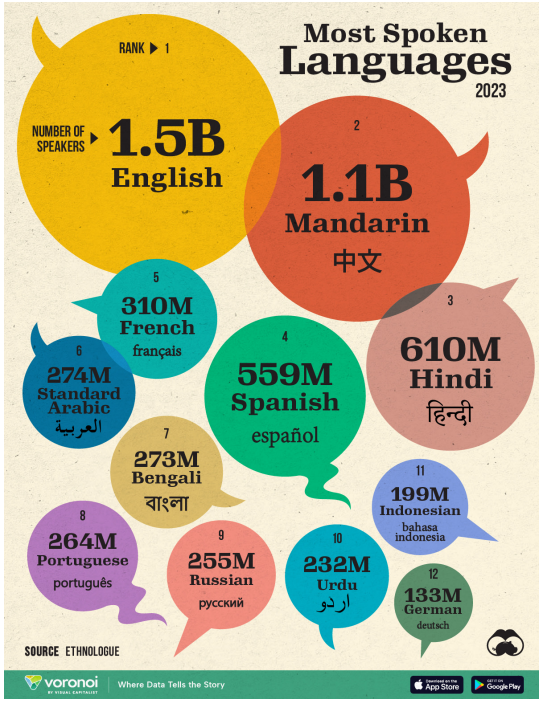
What is a Natural Language Processing?

A field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages.

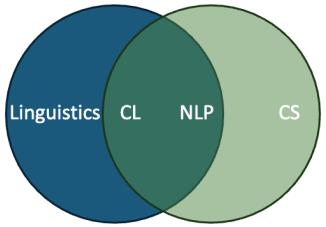


Why NLP is interesting?





NLP vs Computational Linguistics?



The computational **study** of language

Computational Linguistics

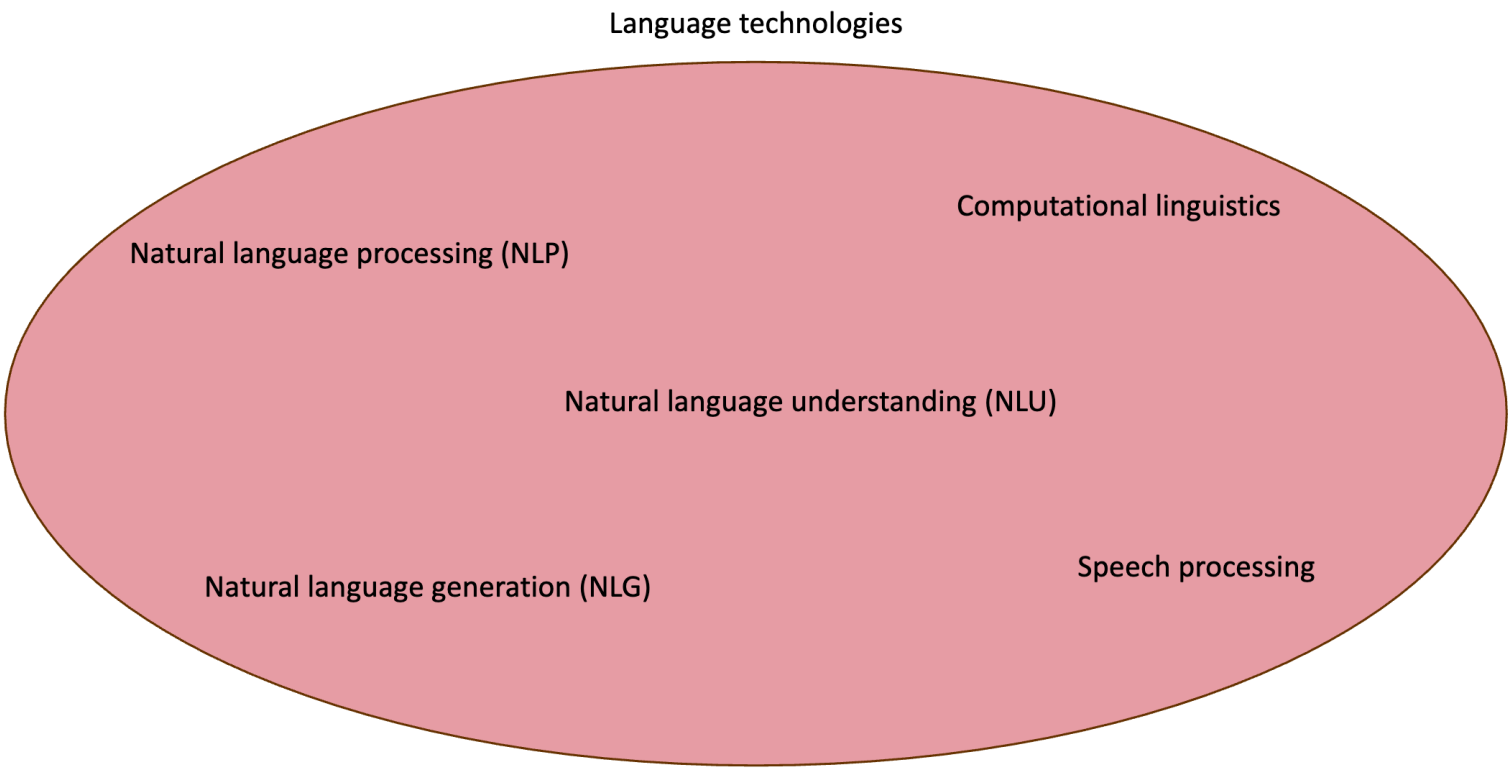
≈

Natural Language Processing

The computational **use** of language

Association for Computational Linguistics

Language Technologies



Natural Language Processing and Computational Linguistics

Natural language processing (NLP) develops methods for solving practical problems involving language [Johnson, 2014].

- Automatic speech recognition.
- Machine translation.
- Information extraction from documents.

Computational linguistics (CL) studies the computational processes underlying (human) language.

- How do we understand language?
- How do we produce language?
- How do we learn language?

Similar methods and models are used in NLP and CL.

Computational Linguistics vs NLP

- Most of the conferences and journals that host natural language processing research bear the name "computational linguistics" (e.g., ACL, NACL).

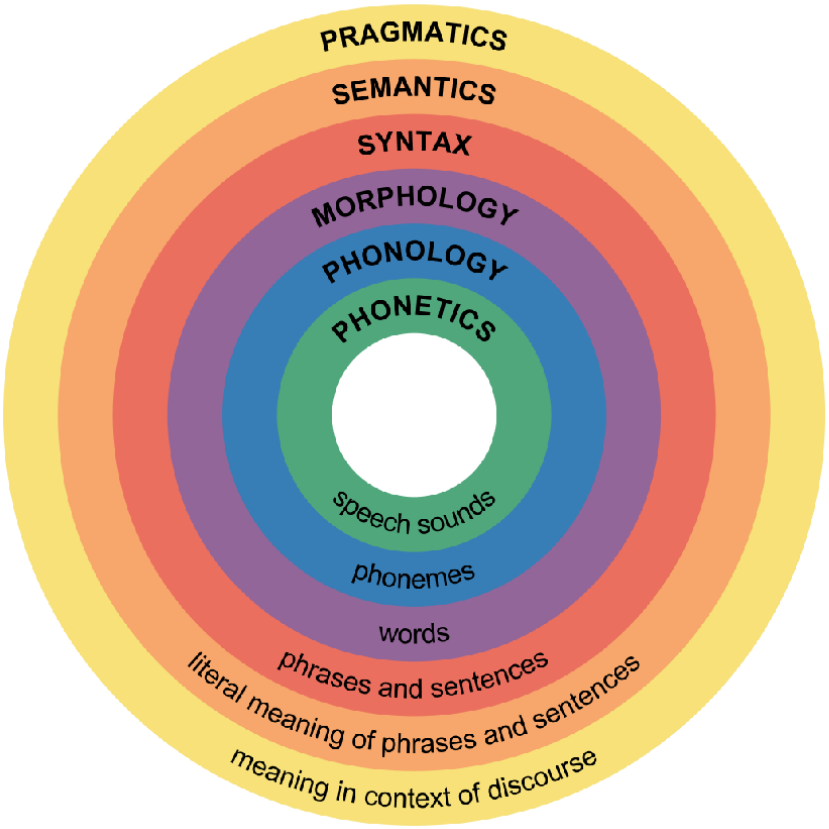
- NLP and CL may be thought of as essentially synonymous.

- **CL: 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).

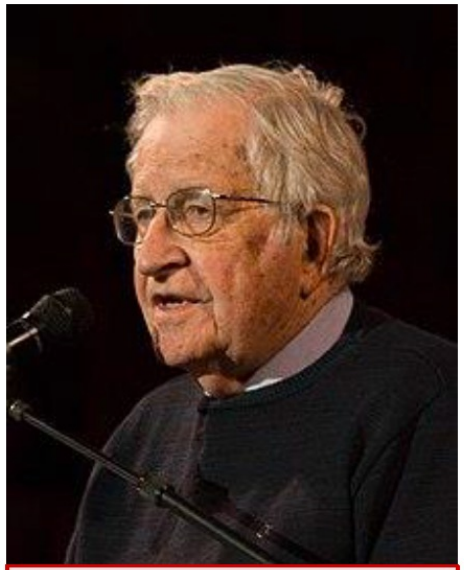
- **NLP:** NLP focuses on solving well-defined tasks involving human language (e.g., translation, query answering, holding conversations).

Linguistics

The study of language



Natural Language Processing



The father of modern linguistics

In 1957, **Noam Chomsky**'s **Syntactic Structures** revolutionized Linguistics with '**universal grammar**', a rule-based system of syntactic structures

He is a laureate professor of linguistics at University of Arizona and an institute professor emeritus at MIT.

Where does the data come from?

- Corpus (plural: corpora): Literally a “body” of text
- Languages with few corpora are called “low-resource languages”
- We can collect corpora in a few different ways: Curation: data tagged & organized by experts
- Internet: data “scraped” from open-access sources (Wikipedia, Reddit)
- Or data collected with permission from closed sources (Facebook, texts) – more rare

Data creation in Africa

Lacuna Fund
Our voice on data

Lanfrica catalogues and links African language resources in order to mitigate the difficulty encountered in discovering African works.

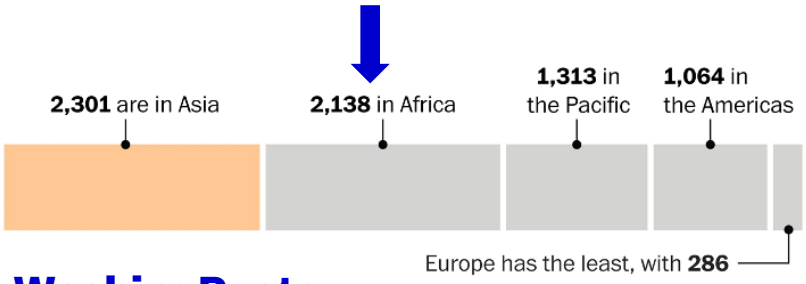
Browse Records | Join our Slack

Languages in the World

7,151 known languages (Ethnologue, 2022)

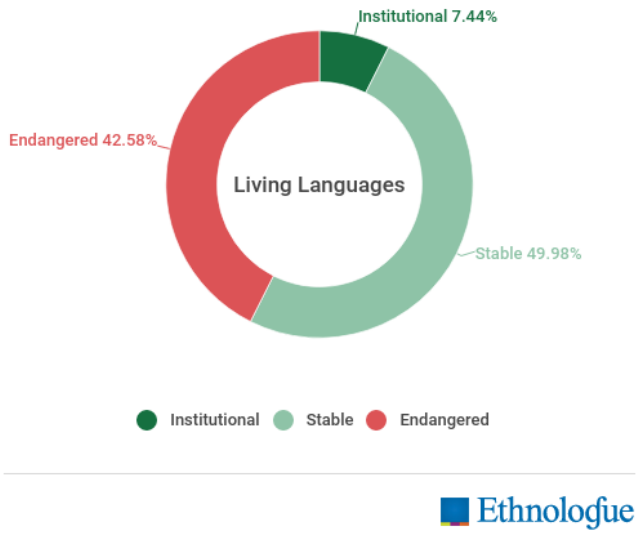
Around **400** have more than **1M speakers**

Around **1,200 languages** have more than **100k**



WashingPosts

- Only **7.4%** are **institutional**
  - Often used by **governments**, **schools**, and **mass media**
  - **Even less are supported by language technologies.**



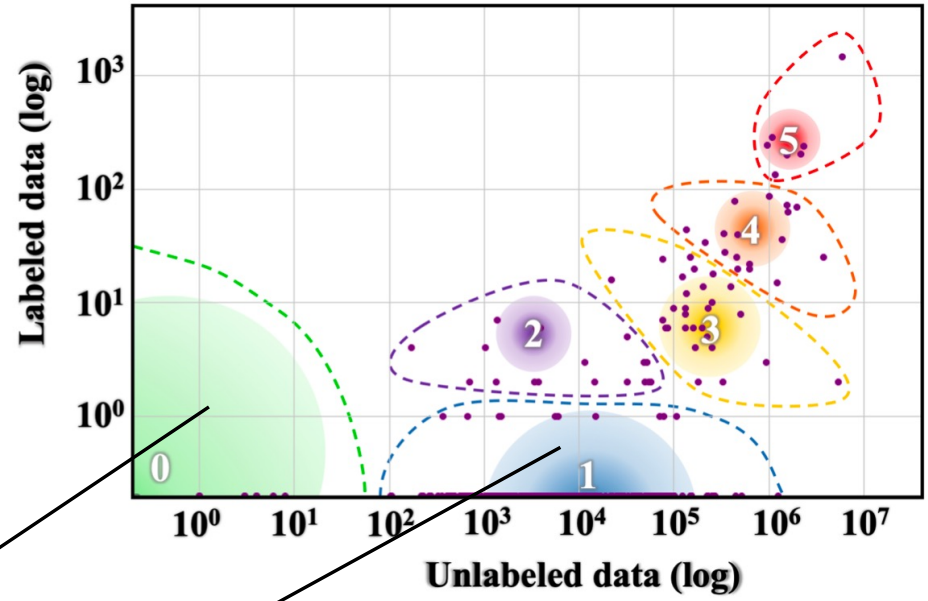
Under-resourced languages: Labelled+Unlabelled data

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

- Unlabelled corpora
- Labelled corpora

categorization of languages based on the amount

of NLP resources available for each language



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

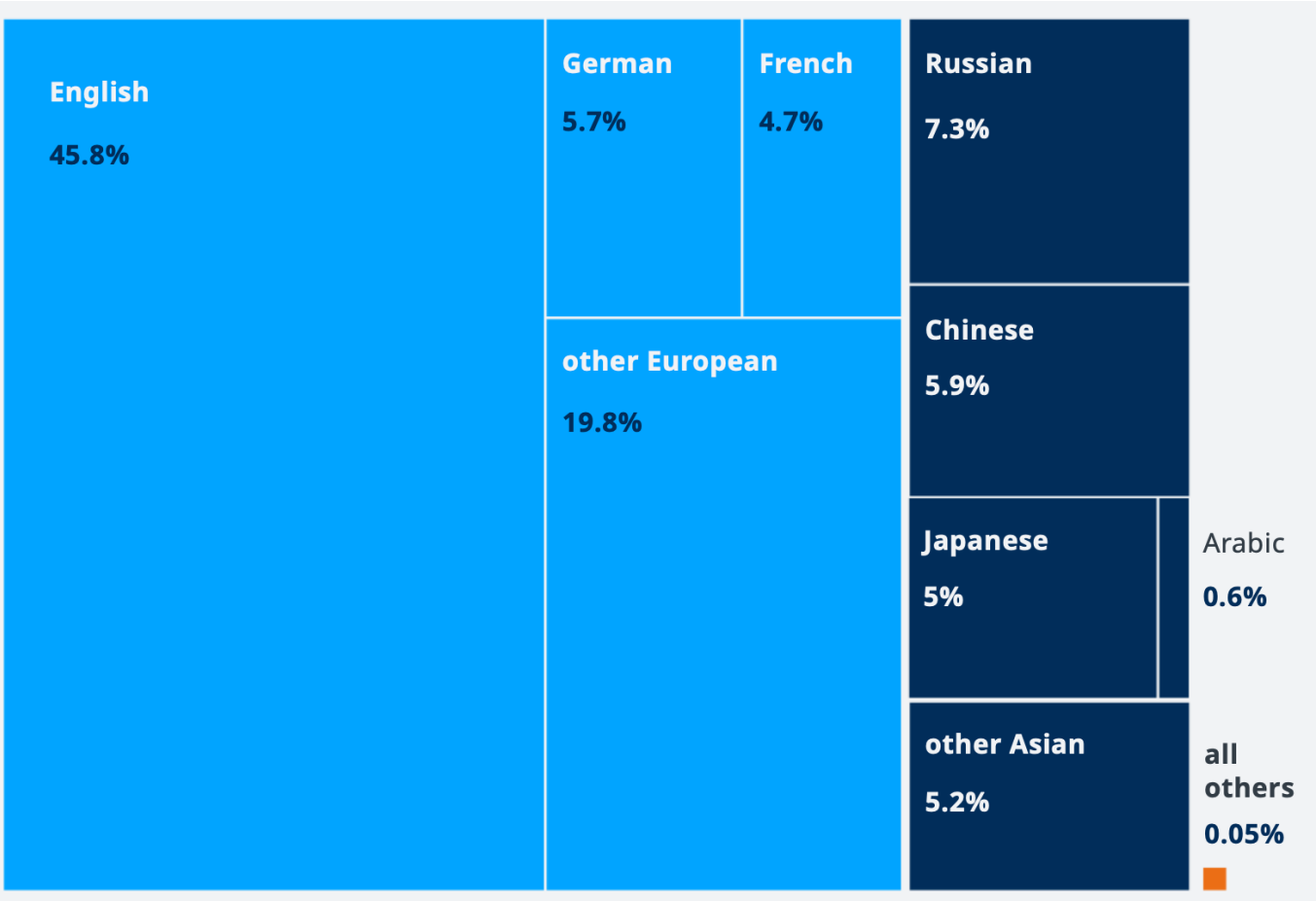
- Languages with almost no NLP resources, where some digitized text may exist, but annotated corpora and NLP tools are scarce.
- Examples: Wolof, Aymara, Tigrinya

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

Lack of Publicly Available Dataset

languages in the Common Crawl internet archive



Source: Common Crawl | More Info: github.com/dw-data/ai-languages

30%

World languages are African (Ethnologue)

0.05%

Why is NLP Hard?

Why is NLP Hard?

Ambiguity

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

Lexical Ambiguity

Lexical Ambiguity: Words have multiple meanings depending on the context.

Example:

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

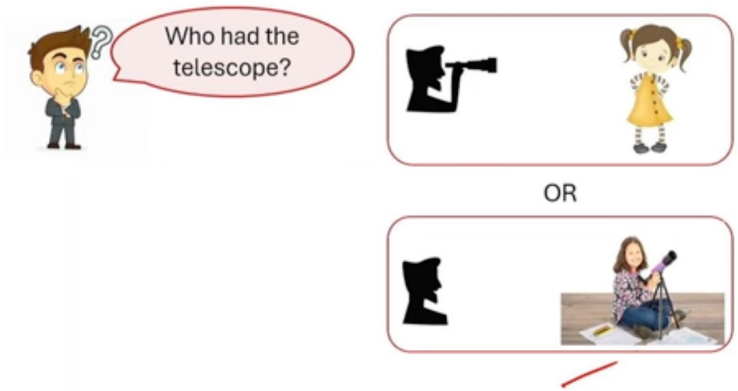
Syntactic Ambiguity

A sentence has multiple possible structures, leading to different interpretations.

Example:

I saw a girl with the telescope.

- Did I use a telescope to see the girl?
- Or does the girl have a telescope?



Ambiguity in Language

- I ate food with Spoon

- I ate rice with curd

- I ate rice with Muhammad

Semantic Ambiguity

The meaning of a sentence is unclear due to contextual dependency.

Example: The chicken is ready to eat.

- Is the chicken going to eat something?
- Or is the chicken cooked and ready to be eaten?

Pragmatic Ambiguity

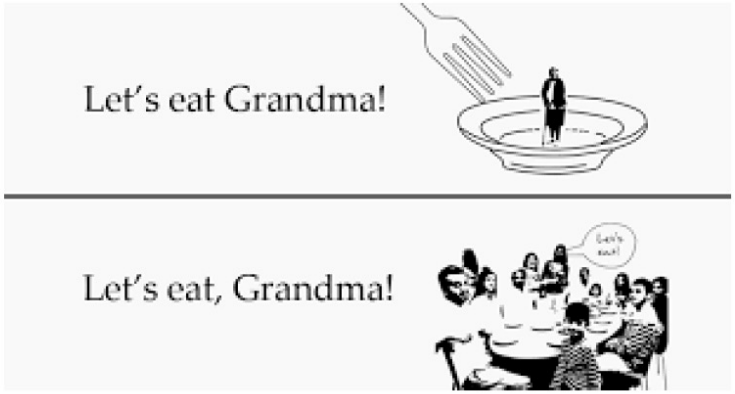
Meaning is influenced by context, intent, and social norms.

Example:

Can you pass the salt?

- A literal interpretation: "Are you physically capable of passing the salt?"
- A pragmatic interpretation: This is a polite request, meaning "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.

Why NLP is difficult?

Non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either ♥

Segmentation Issues

the New York-New Haven Railroad

the New York-New Haven Railroad

Idioms / Multiword

dark horse
get cold feet
lose face
throw in the towel
Khana-wana (Echo)

Neologisms

unfriend
Retweet
bromance

World Knowledge

Mary and Juhi are sisters.
Mary and Juhi are mothers.

Tricky Entity Names

Where is *A Bug's Life* playing ...
Let It Be was recorded ...
... a mutation on the *for* gene ...

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 in a language.

Types of Morphology:

Inflectional Morphology – Changes a word's form to express tense, number, or gender without changing its core meaning.

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

Derivational Morphology – Creates new words by adding prefixes or suffixes, sometimes changing the word's part of speech.

Example: *happy → unhappy* (negation), *teach → teacher* (verb → noun)

Why is Morphology Important in NLP?

Morphology helps computers understand how words are built,

- Helps in **lemmatization and stemming** (reducing words to their root form).
- Improves **machine translation** and **speech recognition**.
- Crucial for **handling complex languages** with rich word structures like **Turkish, Arabic, and Swahili**.

Morphology

- English, Chinese, etc. are commonly referred as *morphologically-poor* language.
- Hindi, Turkish, Hungarian, etc. are termed as *morphologically-rich* language.

English	Hindi	Linguistic property
I will go.	मैं जाऊँगा।	Different morphological forms of word 'will go' in Hindi
We will go.	हम जाएंगे।	
You will go.	तुम जाओगे।	
He will go.	वह जाएगा।	
She will go.	वह जाएगी।	

Word and Token

- A **word** is the smallest meaningful unit of language that can stand alone.
- In NLP, a word is typically defined as a unit of meaning, which can be a single morpheme or a combination of morphemes.

Example:

- Sentence: ***I love NLP research.***
- Words: I, love, NLP, research.
- Here, each word is a meaningful unit.

Word and Token

- A **token** is a unit of text obtained after tokenization, which is the process of breaking a text into individual components (words, subwords, or even characters).
- A token may or may not correspond directly to a word.

Tokenization *is the process of breaking down text into smaller units, called* ***tokens****, which can be words, subwords, or characters.*

It is a fundamental preprocessing step in ***Natural Language Processing (NLP)*** *used for tasks such as machine translation, text classification, and language modeling.*

Types of Tokenization

Word Tokenization: The text is split into individual words based on whitespace and punctuation.
Example:

- **Input:** ***I love NLP!***
- **Tokens: ["I", "love", "NLP", "!"]**

Subword Tokenization (used in BPE, WordPiece, SentencePiece): Words are broken into smaller meaningful units, especially for handling out-of-vocabulary (OOV) words.

Example (Byte Pair Encoding - BPE):

- **Input:** ***unhappiness***
- **Tokens: ["un", "happiness"]**

Example (WordPiece, as used in BERT):

- **Input:** ***playing***
- **Tokens: ["play", "##ing"]**

Types of Tokenization

Character Tokenization

- Each character is treated as a separate token, useful for handling unknown words in languages with complex morphology (e.g., Chinese, Japanese).

Example:

1. **Input:** *Hello*
2. **Tokens:** ["H", "e", "l", "l", "o"]

Sentence Tokenization

The text is split into sentences instead of words, useful in summarization **or translation tasks.**

Example:

1. **Input:** *I love NLP. It is exciting!*
2. **Tokens:** ["I love NLP.", "It is exciting!"]

Word and Token

Ram, Shyam, and Mohan are playing.	⇒	[Ram] [,] [Shyam] [,] [and] [Mohan] [are] [playing] [.]
21,53,010 COVID cases in India.	⇒	[21] [,] [53] [,] [010] [COVID] [cases] [in] [India] [.]
		[21,53,010] [COVID] [cases] [in] [India] [.] ✅
Check this out...https://www.abc.com	⇒	[Check] [this] [out] [.] [.] [.] [https] [:] [/] [/] [www] [.] [abc] [.] [com]
		[Check] [this] [out] [...] [https://www.abc.com] ✅
#GreatDayEver	⇒	[#] [Great] [Day] [Ever]

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.
- In NLP, semantic analysis helps machines understand, interpret, and generate human language correctly by considering context and relationships between words.

Lexical Semantics (Word Meaning)

Studies the meaning of individual words and their relationships (synonyms, antonyms, homonyms).

Example:

bank" (financial institution) vs. *"bank"* (side of a river) → **Lexical Ambiguity**

"big" ≈ "large" (**synonym**), *"happy" ≠ "sad"* (**antonym**)

Used in Word Sense Disambiguation (WSD), sentiment analysis, and information retrieval.

Compositional Semantics (Sentence Meaning)

Compositional Semantics (Sentence Meaning)

Studies how word meanings combine to form sentence meaning.

Example:

"John loves Mary." → Meaning depends on word order.
"Mary loves John." ≠ *"John loves Mary."* (Different subject-object relationship)

Used in machine translation, question answering, and text comprehension.

Pragmatic Semantics (Context & Intent)

Meaning depends on **context, tone, and implied intent**.

Example:

“Can you pass the salt?”

- ▪ **Literal meaning:** Asking about ability.
- ▪ **Pragmatic meaning:** A request to pass the salt.

*Used in **chatbots, sarcasm detection, and human-computer interaction.***

Task we want to solve in NLP?

Basic NLP Tasks

Tokenization: Splitting text into words (word tokenization) or sentences (sentence tokenization).
Example:

Input: "NLP is amazing!"
Output: ['NLP', 'is', 'amazing', '!’]

Stopword Removal: Eliminating common words like "the," "is," "in," etc. Helps in reducing noise in text analysis.

Example:

Input: "This is a great NLP tutorial.”
Output: ['great', 'NLP', 'tutorial']

Basic NLP Tasks

Stemming reduces words to their base or **stem form**, often by chopping off suffixes. However, it may not always result in a meaningful word.

Stemmed Words:

- **cats** → cat
- **were** → were (unchanged)
- **running** → run
- **gardens** → garden
- **happily** → happi (incorrect stem)
- **injured** → injur

- Some words (like *happily → happi*) lose meaning due to aggressive chopping.
- It does not consider the proper root word, just removes common suffixes.

Lemmatization

Lemmatization converts words into their **dictionary base form (lemma)**, using linguistic knowledge.

Lemmatized Words:

- **cats** → cat
- **were** → be (correct lemma)
- **running** → run
- **gardens** → garden
- **happily** → happy
- **injured** → injure

- More accurate and meaningful base forms.
- Proper grammatical transformations (e.g., *were → be*, *happily → happy*).
- Context-aware, ensuring the correct dictionary form.

Stemming vs Lemmatization?

Use Stemming when speed is important and minor errors are acceptable (e.g., search engines).

Use Lemmatization when **accuracy matters** (e.g., NLP applications like chatbots, text analysis).

Part-of-Speech (POS) Tagging

Assigns grammatical categories (noun, verb, adjective) to each word.

Example:

Input: "The dog runs fast."

Output: [('The', 'DET'), ('dog', 'NOUN'), ('runs', 'VERB'), ('fast', 'ADV')]

Named Entity Recognition (NER)

Identifies names of people, organizations, locations, etc.

Example:

- **Input**: "Elon Musk is the CEO of Tesla."
- **Output**: [('Elon Musk', 'PERSON'), ('Tesla', 'ORG')]

Sentiment Analysis

Determines if a text is positive, negative, or neutral.

Used in social media monitoring, customer reviews, etc.

Example:

- **Input**: "I love this movie!"
- **Output**: Positive Sentiment

Text Classification

Categorizes text into predefined labels (e.g., spam vs. not spam).

Example:

Input: "Congratulations! You have won a prize!"

Output: Spam

Machine Translation

Translates text from one language to another (e.g., Google Translate).

Example:

Input: "Hello, how are you?" (English)

Output: "Hola, ¿cómo estás?" (Spanish)

Text Summarization

Generates short summaries of long documents.

Types:

- **Extractive (selects key sentences)**
- **Abstractive (generates new sentences)**

Example:

- **Input**: A 500-word article
- **Output**: A 50-word summary

Question Answering (QA)

Extracts answers from text based on a question.

Example:

- **Passage**: "J.K. Rowling is the author of the Harry Potter series.”
- **Question**: "Who wrote Harry Potter?"
- **Output**: "J.K. Rowling"

Text Generation

Generates human-like text (e.g., ChatGPT, GPT-4, BERT).

Example:

Input: "Once upon a time..."

Output: "there was a young girl who dreamed of exploring the world."

Speech Recognition & Speech-to-Text

Converts spoken language into text (e.g., Siri, Google Assistant).

Example:

- **Input (Voice Command)**: "Play my favorite song."
- **Output**: Text: "Play my favorite song."

Fake News Detection

Identifies whether a news article is fake or real.

Example:

- **Input**: "COVID-19 vaccines contain microchips!"
- **Output**: Fake News

Multimodal NLP

Combines NLP with images and videos (e.g., caption generation for images).

Example:

- **Input**: (Image of a dog running)
- **Output**: "A dog is running in the park."

Text-to-SQL

Converts natural language queries into SQL queries.

Example:

Input: "Show me the sales of the last quarter."

Output: SELECT SUM(sales) FROM revenue WHERE quarter = 'Q4';

Code Generation & Understanding

AI-powered code assistants (e.g., GitHub Copilot).

Example:

- **Input**: "Write a Python function for Fibonacci sequence."
- **Output**: A Python code snippet.

IMPERIAL

Q and A