

BSCS5002: Introduction to Natural Language Processing

Lecture 3 : Why is it hard to process natural language?

Parameswari Krishnamurthy



Language Technologies Research Centre
IIIT-Hyderabad

param.krishna@iiit.ac.in



Natural Language Processing Tasks

Core technologies

- Language modelling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic Role Labelling
- ...

Applications

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

Why NLP is hard?

1. Ambiguity at many levels:

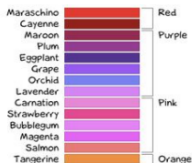
- ▶ Word senses: **bank** (finance or river?)
- ▶ Part of speech: **chair** (noun or verb?)
- ▶ Syntactic structure: **I saw a man with a telescope**
- ▶ Quantifier scope: **Every child loves some movie**
- ▶ Multiple: **I saw her duck**

⇒ NLP algorithms model ambiguity, and choose the correct analysis in context

2. Linguistic diversity

Linguistic Diversity: Semantics

Every language describes the world in a different way, for example, it depends on culture or historical conditions.



- Russian has relatively few names for colors; Japanese has hundreds
- Multiword expressions, e.g. **it's raining cats and dogs** or **wake up** and metaphors, e.g. **Love is a journey** are very different across languages

Sapir-Whorf Hypothesis:

the language we speak both affects and reflects our view of the world

Linguistic Diversity: Language Families

www.ethnologue.com

1. Niger–Congo (1,538 languages) (20.6%)
2. Austronesian (1,257 languages) (16.8%)
3. Trans–New Guinea (480 languages) (6.4%)
4. Sino-Tibetan (457 languages) (6.1%)
5. Indo-European (444 languages) (5.9%)
6. Australian (378 languages) (5.1%)
7. Afro-Asiatic (375 languages) (5.0%)
8. Nilo-Saharan (205 languages) (2.7%)
9. Oto-Manguean (177 languages) (2.4%)
10. Austroasiatic (169 languages) (2.3%)
11. Volta Congo (108 languages) (1.5%)
12. Tai–Kadai (95 languages) (1.3%)
13. Dravidian (85 languages) (1.1%)
14. Tupian (76 languages) (1.0%)

Why NLP is hard?

1. Ambiguity at many levels:

- ▶ Word senses: **bank** (finance or river?)
- ▶ Part of speech: **blue** (noun or verb?)
- ▶ Syntactic structure: **I saw a man with a telescope**
- ▶ Quantifier scope: **Every child loves some movie**
- ▶ Multiple: **I saw her duck**

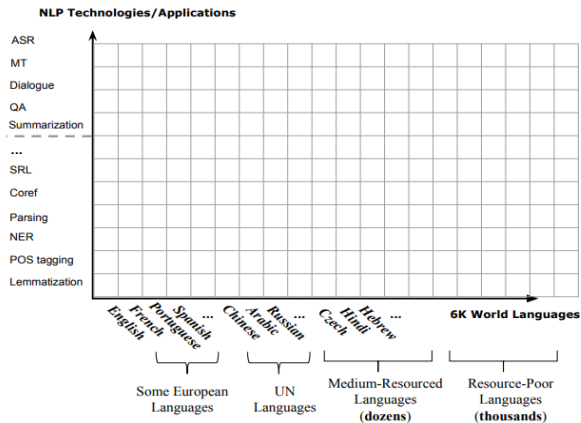
⇒ NLP algorithms model ambiguity, and choose the correct analysis in context

2. Linguistic diversity

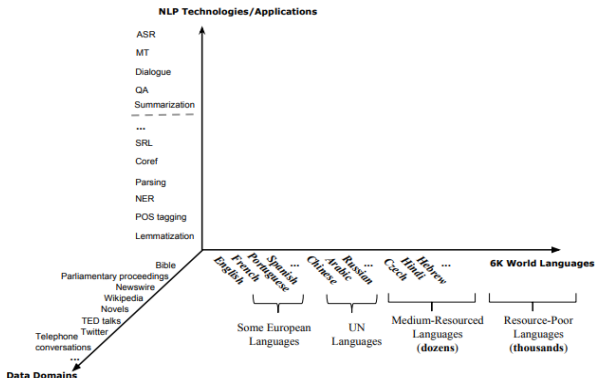
- ▶ 6–7K languages in the world, > 14 language families
- ▶ Languages diverge across all levels of linguistic structure
⇒ **no generic solution for a particular NLP task**
- ▶ **Most of the languages do not have sufficient resources to build statistical NLP models**

Low-resource languages – languages lacking large monolingual or parallel corpora and/or manually crafted linguistic resources sufficient for building statistical NLP applications

What NLP Technologies are Resource-Rich?



Low-Resource NLP is Not Only About Multilinguality




Why NLP is hard?


The man couldn't lift his son because he was so **weak**.  Who was weak?

The man couldn't lift his son because he was so **heavy**.  Who was heavy?

Mary and Sue are **sisters**.
Mary and Sue are **mothers**.  How are Mary and Sue related?

Joan made sure to thank Susan for all the help she had **received**.  Who had received help?

Joan made sure to thank Susan for all the help she had **given**.  Who had given help?

John **promised** Bill to leave so an hour later he left.
John **ordered** Bill to leave so an hour later he left.  Who left an hour later?

1. Ambiguity in Natural Language

- **Ambiguity:** A major challenge in NLP, where words, phrases or sentences can have multiple interpretations.
- **Impact on NLP:** Ambiguity makes it difficult for algorithms to correctly understand and process language.
- Types of ambiguity:

Lexical Ambiguity	Syntactic Ambiguity	Semantic Ambiguity	Pragmatic Ambiguity
<ul style="list-style-type: none">• Homonyms• Polysemy	<ul style="list-style-type: none">• Attachment• Coordination	<ul style="list-style-type: none">• Quantifier Scope• Anaphoric	<ul style="list-style-type: none">• Deictic• Speech Act• Irony/Sarcasm

A. lexical Ambiguity

Lexical Ambiguity: Occurs when a word has multiple meanings.

- **i. Homonyms:**

Words that sound alike or are spelled alike but have different meanings.

- *English Example:* **Bark** (the sound a dog makes vs. the outer covering of a tree)

- **ii. Polysemy:**

A single word with multiple related meanings.

- *English Example:* **Head** (the top part of a body vs. leader of an organization)

B. Syntactic Ambiguity

Syntactic Ambiguity: Occurs when a sentence can be parsed in more than one way due to its structure.

- **Attachment:**

Ambiguity arises when it is unclear which part of the sentence a modifier (e.g., a prepositional phrase) is associated with.

- *Example:* **He saw the man with the telescope.**

(Did he use a telescope to see the man or did he see a man who had a telescope?)

- **Coordination:**

Ambiguity occurs when it is unclear how conjunctions like "and" or "or" connect different parts of a sentence.

- *Example:* **Old men and women were admitted free.**

(Were only old men admitted free or both old men and all women?)

C. Semantic Ambiguity

Semantic Ambiguity: Occurs when a sentence or phrase can have multiple meanings due to the interpretation of words or phrases.

- **Quantifier Scope Ambiguity:**

Ambiguity arises when it's unclear how far a quantifier (like "all" or "some") applies within a sentence.

- *Example:* **All students read some books.**

(Does it mean each student reads the same books or different books?)

- **Anaphoric Ambiguity:**

Ambiguity occurs when a pronoun or a referring expression can be linked to multiple possible antecedents.

- *Example:* **John told Peter he would win.**

(Is "he" referring to John or Peter?)

D. Pragmatic Ambiguity

Pragmatic Ambiguity: Occurs when a sentence can be interpreted in multiple ways based on context, tone or conversational implications.

- **Deictic Ambiguity:**

Ambiguity arises when it is unclear to who or what a deictic expression (like "this," "that," "here," "there") refers.

- *Example: I'll meet you there.*
(Where is "there"?)

- **Speech Act Ambiguity:**

Ambiguity occurs when it is unclear what kind of speech act (e.g., request, statement, command) a sentence represents.

- *Example: Can you pass the salt?*
(Is this a question about ability or a polite request?)

D. Pragmatic Ambiguity

- **Irony/Sarcasm:**

Ambiguity arises when the intended meaning is the opposite of the literal meaning.

- *Example:* **Oh, great! Another meeting!**
(Is the speaker genuinely excited or are they being sarcastic?)

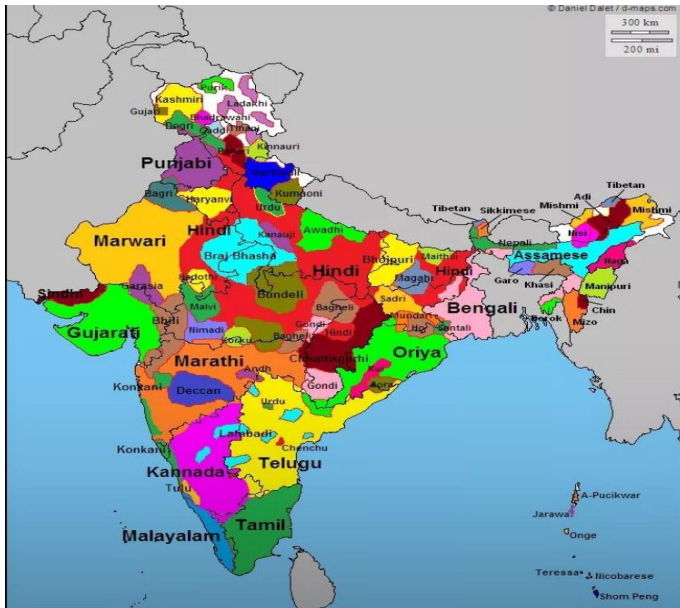
2. Challenges in Multilingual NLP

- Processing multiple languages poses unique challenges in NLP, especially in a multilingual country like India.
- Differences in grammar, syntax, word order, and vocabulary between languages make it difficult to develop unified models.

Indian Context:

- India has 22 official languages (as per 8th schedule), each with distinct linguistic characteristics.
- Many Indian languages, like Hindi, Telugu and Tamil, differ significantly from English and among themselves, making multilingual NLP particularly challenging.

Language Map of India



Multilingualism in use

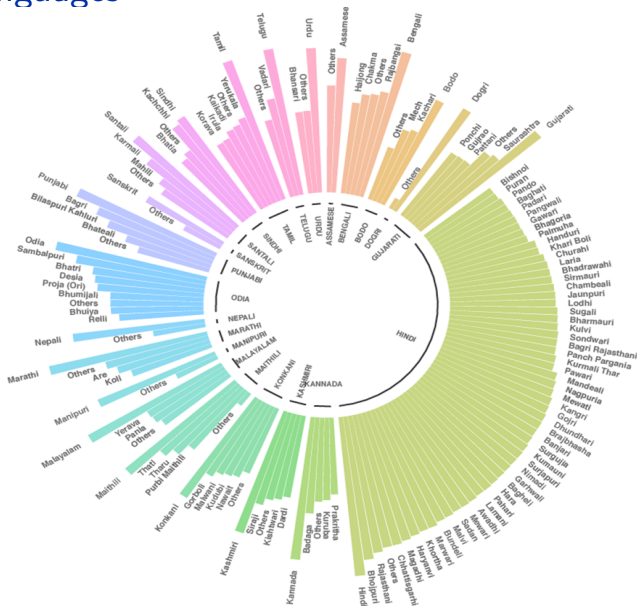


Tiruchirappalli International
Airport(Tamil-Hindi-English)



Imphal International Airport
(Meitei-Hindi-English)

Indian Languages



Most Diverse Indian State

NAGALAND- Languages spoken by the people

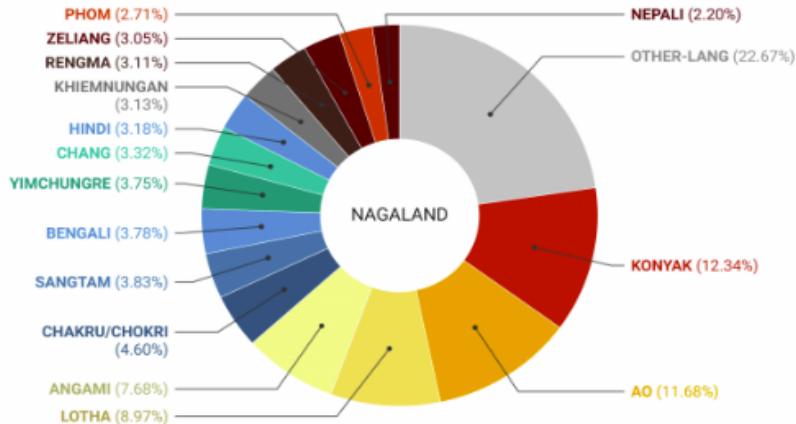


Chart: Shivakumar Jolad • Source: Census 2011 • Created with Datawrapper

Least Diverse Indian State

KERALA- Languages spoken by the people

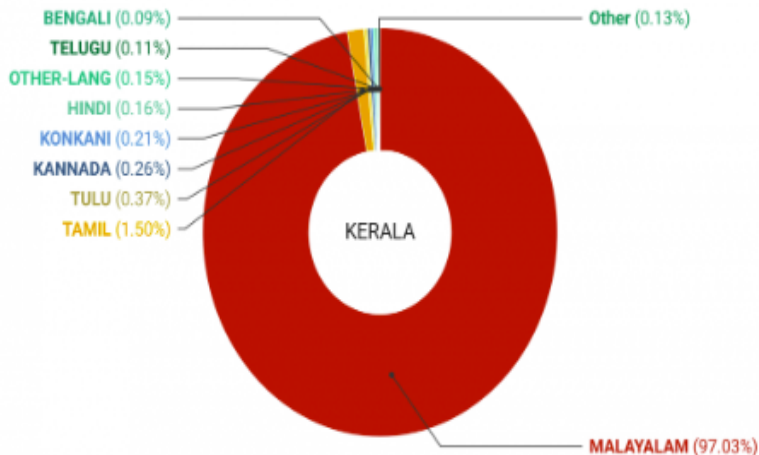


Chart: Shivakumar Jolad • Source: Census 2011 • Created with Datawrapper

Languages of India

Indo-Aryan Family: Assamese , Baigani, Banjari, Bengali ,, Bhatrī, Bhili, Bhunjia, Chakma, Chhattisgarhi, Dhanki, Dhodia, Dhundhari, Gadiali, Gamit/Gavti, Garasia/Girasia, Gojri/Gujjari, Gujarati , Hajong, Halbi, Harauti, Hindi , Jaunsari, Kachchi, Khotia, Kinnui, Kokni, Konkani , Kotwalia, Kudamamali, Thar, Lambani or Lamani , Laria, Magahi, Mahl, Marathi , Mavchi, Mewnri, Nagpuri, Naikadi, Nimari, Oriya , Rathi, Sarhodi, Shina, Tharu, Wagri, Warli.

Languages of India

Tibeto-Burman Family: Adi Ashing, Adi Bokar, Adi Bori, Adi Gallong, Adi Komkar, Adi Milang, Adi Minyong, Adi Padam, Adi Karko, Pailibo, Adi Pangi, Adi Pasi, Adi Ramo, Adi Shimong, Adi Tangam, Aimol, Anal, Angami, Ao, Apatani, Balti, Bangni/Dafla, Bawm, Bhotia, Biate, Bodo, Bugun, Chakhesang, Champa, Chang, Chiru, Chote, Chung, Dalu, Deori, Dokpa/Droskat, Duhlian-Twang, Gangte, Garo, Halam, Hmar, Hrusso/Aka, Hualngo, Kabui, Kachari, Kagati, Kak barak, Khamba, Khampa, Khiamngan, Koch, Koireng, Konyak, Kuki, Ladakhi, Lahauli, Lai Hawlh, Lakher/Mara, Lalung, Lamgang, Lepcha, Lisu, Lotha, Lushai/Mizo, Mag/Mogh, Mao, Maram, Maring, Memba, Mikir, Miri, Mishing, Mishmi, Monpa, Monsang, Moyon, Na, Naga, Sherdukpen, Nishi, Nocte, Paite, Pang, Phom, Pochury, Ralte, Rengma, Riang, Sajalong/Miju, Sangtam, Sema, Sherpa, Singpho, Sulung, Tagin, Tangsa, Thado, Tangkhul, Tibetan, Toto, Vaiphei, Wancho, Yim-chungre, Zakhring/Meyer, Zemi, Zou.

Languages of India

Dravidian Family: Dhurwa, Gadaba tribe , Gondi, Kadar tribe, Kannada, Kodagu, Kolami, Koraga, Kota, Koya/Koi, Kui, Kurukh, Kuvi, Malayalam, Malta, Maria, Naiki, Parji, Pengo, Tamil, Telugu , Toda, Tulu, Yerukula.

Austro-Asiatic Family: Asuri, Bhumij tribe, Birhor tribe, Birjia tribe, Bondo, Diday, Gutob, Ho, Juang, Kharia, Khasi, Kherwari, Korku, Korwa, Kurmi, Lodha, Mundari, Nicobarese, Santali, Saora/Savara, Shompen, Thar.

Andamanese Family: Andamanese Tribe, Jarawa tribe, Onge, Santinelese.

Indian Language and Data Resources

Language	Code	Pop. (M)	CC Size	
			(%)	Cat.
English	en	1,452	45.8786	H
Russian	ru	258	5.9692	H
German	de	134	5.8811	H
Chinese	zh	1,118	4.8747	H
Japanese	jp	125	4.7884	H
French	fr	274	4.7254	H
Spanish	es	548	4.4690	H
Italian	it	68	2.5712	H
Dutch	nl	30	2.0585	H
Polish	pl	45	1.6636	H
Portuguese	pt	257	1.1505	H
Vietnamese	vi	85	1.0299	H
Turkish	tr	88	0.8439	M
Indonesian	id	199	0.7991	M
Swedish	sv	13	0.6969	M
Arabic	ar	274	0.6658	M
Persian	fa	130	0.6582	M
Korean	ko	81	0.6498	M
Greek	el	13	0.5870	M
Thai	th	60	0.4143	M
Ukrainian	uk	33	0.3304	M
Bulgarian	bg	8	0.2900	M
Hindi	hi	602	0.1588	M

Indian Language and Data Resources

Bengali	bn	272	0.0930	L
Tamil	ta	86	0.0446	L
Urdu	ur	231	0.0274	L
Malayalam	ml	36	0.0222	L
Marathi	mr	99	0.0213	L
Telugu	te	95	0.0183	L
Gujarati	gu	62	0.0126	L
Burmese	my	33	0.0126	L
Kannada	kn	64	0.0122	L
Swahili	sw	71	0.0077	X
Punjabi	pa	113	0.0061	X
Kyrgyz	ky	5	0.0049	X
Odia	or	39	0.0044	X
Assamese	as	15	0.0025	X

Table 1: List of languages, language codes, numbers of first and second speakers, data ratios in the Common-Crawl corpus, and language categories. The languages are grouped into categories based on their data ratios in the CommonCrawl corpus: High Resource (H, $> 1\%$), Medium Resource (M, $> 0.1\%$), and Low Resource (L, $> 0.01\%$), and Extremely-Low Resource (X, $< 0.01\%$).

Challenges with Low-Resource Languages

- Many Indian languages have limited digital resources, making it difficult to train NLP models effectively.
- Lack of large annotated datasets and linguistic tools hinders the development of accurate NLP models for these languages.

Example:

- Languages like Telugu, Tamil, Bengali etc. despite having millions of speakers, are underrepresented in NLP resources compared to English or Hindi.
- This creates a disparity in the quality of NLP applications, such as machine translation, sentiment analysis, and speech recognition.

3. Language Variability Across Speakers

- Language varies significantly across different speakers based on a variety of factors.
- This variability can be observed in different **dialects, sociolects, and idiolects**.

A. Dialectal Variations:

- Dialects are distinct forms of a language spoken by specific groups, often distinguished by geographical regions. Each dialect can have unique vocabulary, grammar, and pronunciation.
- *Example:*
 - In American English, the word "**truck**" is commonly used, whereas in British (and Indian) English, the same vehicle is referred to as a "**lorry**".
 - The pronunciation of certain words, like "**tomato**" (*toh-MAH-toh* vs. *to-MAY-to*), also differs between regions.

B. Sociolects:

- Sociolects refer to variations in language used by specific social groups, often influenced by factors such as class, education, or occupation. The way language is used can signal social identity and status.
- *Example:*
 - In formal settings, people might say, "**Good morning, how are you?**" while in informal contexts, they might simply say, "**Hey, what's up?**"
 - Professional jargon is another example of sociolect, where specific terms are used within particular industries (e.g., legal jargon, medical terminology).

C. Idiolects:

- An idiolect is the unique form of language used by an individual. This includes personal preferences in word choice, pronunciation, and sentence structure. It reflects a person's background, experiences, and personality.
- *Example:*
 - One person might frequently use certain filler words like "**you know**" or "**like,**" while another avoids them.
 - Some individuals may have unique phrases or expressions that are distinctively their own.

Why NLP is hard?

- **Ambiguity:** Words and sentences can have multiple meanings depending on context.
 - Example: "I saw the man with the telescope."
- **Context and Semantics:** Understanding the context and meaning behind words and sentences requires deep comprehension.
 - Example: Sarcasm and irony can be difficult for machines to detect.
- **Variety of Languages:** Each language has its own rules, syntax and nuances.
 - NLP models need to be adaptable to multiple languages and dialects.

Why NLP is hard?

- **Evolution of Language:** Language is constantly changing with new slang, idioms and phrases emerging.
 - Keeping up with these changes is challenging for NLP systems.
- **Data Quality and Quantity:** High-quality, annotated data is necessary for training NLP models, but it is often scarce and expensive to obtain.
 - Large datasets are required for effective machine learning, but these can be difficult to compile.
- **Computational Complexity:** Advanced NLP models require significant computational resources for training and inference.
 - Deep learning models, such as transformers, are computationally intensive.
- **Cultural and Ethical Considerations:** NLP systems must be designed to avoid biases and respect cultural differences.
 - Ensuring fairness and reducing biases in NLP models is a significant challenge.