Project Guidelines

Intro to Linguistics - I

Akshit Kumar

September 2, 2023

As per the course guidelines, the team project should focus on developing a strong understanding of linguistic analysis and applying fundamental linguistic concepts to real-life problems in different languages. The project guidelines document consists of the required specifications for a valid project and list of suggested tasks.

Specifications

Source Language The project should be conducted in an Indian language.

Switching The chosen language should be clearly stated at the outset. Switching to a different language during the project will not be permissible.

Typesetting It is recommended that the project be typed and submitted as a .pdf file, you are encouraged to explore typesetting systems like LaTeX. Proficiency in using IPA symbols is important.

Data The same dataset must be used for the duration of the course project.

Data

- Collecting data is a fundamental exercise in linguistic analysis and is crucial for the course project.
- The data does not necessarily have to be in text form. Recorded voice samples and videos are equally valid, and perhaps even better, as they can involve more complex analyses (such as analysing the speaker's tone).
- The data can be sourced from any appropriate origin, provided the original source is either attached or publicly accessible on the internet, with proper citation.
- A few source ideas
  - movie dialogues
  - scenes from a TV show
  - recorded interviews
  - news articles
  - political speeches
  - chapters from books

Suggested Tasks

Phonetic transcription

This task involves transcribing the data into text for further analysis.

1. The text should first be written down in the script typically used for the chosen language. (e.g Devanagari for Hindi)
2. The text should then be transcribed into the roman alphabet
3. Finally, the text should be transcribed using the International Phonetic Alphabet (IPA).

- The first two subtasks may vary in difficulty depending on the format chosen and the marks will be allotted on the basis of effort invested. (For example, a student transcribing a movie scene is investing more effort into task 1 compared to someone using a news article)
- There are no restrictions on the method of phonetic transcription, you can either transcribe manually – by yourselves or write your own program in your preferred language.
- A tutorial will be organised to teach simple text processing tools like regex and common bash commands like awk, sed, grep.
- You are restricted to writing your own functions and using inbuilt libraries.

The rest of the tasks are summarised ahead.

Syllabification

- Select a few sentences from the text and syllabify each word in the sentence.
- This task will help downstream with tasks involving analysis of morphemes.
- Can be done manually or by writing a subroutine to process the language syntax.
- A minimum number of (TBD) manually analysed sentences will be required before being allowed to use a script.

Morphological typology

- Analyse the morphological structure of the chosen language.
- Highlight root/free morphemes
- Identify allomorphs and perform some analysis
- Perform categorisation (e.g., isolating, agglutinative, inflectional)

Syntax analysis

- Investigate the sentence structure of the language
- Identify basic word order
- Analyse distribution of various constituents

Phase structure and dependency grammar

- Create a phase structure and a de- pendency tree for all sentences in the dataset.
- Followed it with a phase structure and dependency grammar of the language chosen taking help from the phrase structure trees written for the previous task.
- Writing a simple phase structure/dependency grammar parser/generator with linguistic reasoning to why you'd choose a particular model.
- A tutorial based on CFGs will be organised.

Language family and historical analysis

- Present the linguistic features of the language family chosen by citing examples from the dataset
- Trace the historic development of the language by taking language features from the dataset and analysing their origin and influences from other languages.

Any combination of the above tasks can be chosen to be worked upon. The exact amount will be decided and announced, it is recommended that you start with your most preffered task as soon as possible.