

## **Large Language Models (LLM) and the Role of linguists in the World of AI**

Vasu Renganathan

University of Pennsylvania

vasur@sas.upenn.edu

(<http://www.sas.upenn.edu/~vasur/project.html>)

### **Introduction:**

Large Language Models (LLMs) and correspondingly building vector databases in the recent time produced appealing advances in enabling machines to behave like humans. Linguists can play a vital role in training LLMs by using many linguistic theories such as lexical semantics, semiotics and so on to identify and account for complexities in meanings of words and expressions so a much more robust use of this research can be achieved. For example, the Tamil word "paṭi" has multiple meanings, such as "to read", "to settle", "step", "a measuring container" and so on. Not only the homonymous words but also extended meanings of expressions as in *vayiru erikīratu* to mean both 'burning sensation in stomach' as well as 'feeling hurt mentally due to unusual attitudes of others' would also pose problems for machines in parallel to how humans understand the natural language. A linguist could help an LLM to be trained with such nuances of expressions so a live interpretation can be achieved by machines. Such efforts would allow the machine to interpret dialogues in a more meaningful manner than now. Prediction and probabilities within complex linguistic structures need to be accounted for more comprehensively than before in order to design plausible and life-like machine-human interactions. Linguistic theories have traditionally focused more on structures than on the bidirectional predictability of words or even sentences, which is what systems like BERT and GPT attempt to do (Cf. Noir 2020). An attempt is made here to demonstrate using a robot as to how the recent developments of NLP can be implemented and tested for Tamil.

### **Recent Advances in NLP tasks and development of a Robot:**

With an extensive research and plausible outcome from many open-source projects such as text to speech, machine translation, Wiki-resources and speech to text, it is now quite possible to integrate them into a mini-robot type of machines and converse to them in a natural way. This project aims at a similar effort with a robot and attempts to converse in Tamil with it. The activities include commanding the robot to move around such as forward, backward, circle around and so on with commands in Tamil, like *முன்னால போங்க, பின்னால வாங்க, சுத்துங்க*, respectively. This system also can be trained with particular set of movements and use later to perform them using commands in Tamil. For example, navigating from one place to another can be recorded in sequence and link it to commands such as *சமையலறைக்கு போங்க, சமையலறைக்கு போயிட்டு வாங்க* and so on so forth, so these commands will call the routine that was trained earlier. The robot that was built earlier and is demonstrated in <http://robot.tamilnlp.com> extensively uses a voice recognition card called EasyVR and its speech to text capability is restricted to a single person. Whereas, the current project employs Google's text to speech and speech to text APIs extensively and attempts to process Tamil voice in a natural way independent of the speaker. The issue with this project, however, lies in the way the parsing of Tamil sentences is done in a meaningful manner, so the machine can respond with its text to speech API. What is crucial is the processing task of the text obtained between the two processes namely speech to text and text to speech. The speech to text

API attempts to convert most of the spoken Tamil expressions into corresponding literary text. What is of utmost importance is that one needs to find the input and output of this API with a set of data and analyze the commonalities, so the processing task can be done efficiently. For example, the output நீங்கள் போய்க்கொண்டு இருக்கிறீர்களா, நீங்கள் நாளைக்கு என்னுடைய வீட்டுக்கு உங்களுடைய புத்தகத்தை எடுத்துக் கொண்டு வருவீர்களா are interpreted from the corresponding literary form of the speech. However, the spoken form of the output நீங்க நாளைக்கு என்னோட வீட்டுக்கு உங்க புத்தகத்தை எடுத்துட்டு வருவீங்களா requires special attention. The word நாளைக்கி is understood/interpreted as நாளைக்கு, எடுத்துக்கிட்டு is interpreted as எடுத்துட்டு. When these words are uttered separately, படிச்சுக்கிட்டு, எடுத்துக்கிட்டு are produced as output with “க்க” in them. Strangely, when these words are included as part of a sentence the output produces only எடுத்துட்டு as in உங்க புத்தகத்தை எடுத்துட்டு வருவீங்களா. Obviously, this kind of analysis needs to be done with a large set of expressions and a common feature of the output from this application needs to be arrived at, so further processing of the output from speech to text API can be possible. Particularly, this kind of study is needed for applications like question/answering systems, database retrieval systems and so on.

### **Using the Wiki resources in Tamil:**

When interacting with the Wiki resources, an attempt is made to parse sentences with the word “பத்தி” and the API to Wiki resources is called. Following are some of the interactions and results from this device:

**Input:** ஒரு பத்தி சொல்லுங்க:

**Output:** ஒரு கல் ஒரு கண்ணாடி, உதயநிதி ஸ்டாலின் நடிக்க ராஜேஷ் இயக்கத்தில் வெளிவந்த தமிழ் காதல் - நகைச்சுவைத் திரைப்படம். இத்திரைப்படத்தில் தயாரிப்பாளர், உதயநிதி ஸ்டாலின், கதாநாயகனாக அறிமுகமானார்.

**Input:** குமரகுரு கல்லூரி பத்தி சொல்லுங்க

**Output:** குமரகுரு தொழில்நுட்பக் கல்லூரி கோவையில் அமைந்துள்ள ஒரு தனியார் பொறியியல் கல்லூரி. 1984 ஆம் ஆண்டு தொடங்கப்பட்ட இக்கல்லூரி சத்தி வணிக குழுமத்தின் இராமானந்த அடிகளார் அறக்கட்டளையினால் நிர்வகிக்கப்படுகிறது.

Although the second query and result seem to be desired, the first query and result requires attention as the text from the query and result do not have any direct link in any of the phrases within the output. However, the input ஒரு கல் கண்ணாடி பத்தி சொல்லுங்க does indeed produce the same output as above. Except for such minor issues, accessing API resources from Tamil Wiki offers a good result for the names of towns, institutions, projects and so on.

**Input:** மயிலாடுதுறை பற்றி சொல்லுங்க:

Output: மயிலாடுதுறை (Mayiladuthurai) (முன்பு மாயவரம் என்று அழைக்கப்பட்டது) இந்தியாவில், தமிழ்நாடு மாநிலத்தில் மயிலாடுதுறை மாவட்டத்தில் உள்ள நிர்வாகத் தலைமையிடமும், சிறப்பு நிலை நகராட்சியும் ஆகும். மயில்கள் ஆடும் துறை என்பதால் மயிலாடுதுறை என அழைக்கப்படுகிறது.

Input: மதுரை திட்டம் பற்றி சொல்லுங்க:

மதுரை தமிழ் இலக்கிய மின் தொகுப்புத் திட்டம் (Project Madurai) என்பது தமிழ் இலக்கியங்களை, இணையத்தில் இலவசமாக வெளியிடும், ஒரு திறந்த, தன்னார்வ, உலகளாவிய முயற்சியாகும். 1998 ஆம் ஆண்டு, பொங்கல் தினத்தன்று, தொடங்கப்பட்ட இத்திட்டம், இன்றளவில் தொடர்ந்து இயங்கி வருகிறது.

### Translation resources:

Google's translation resources are incorporated in this device and the commands such as ஹிந்திலெ சொல்லுங்க, மலயாளத்துல சொல்லுங்க, கன்னடால சொல்லுங்க etc., followed by any sentence in Tamil is translated to the corresponding target languages. Following are some of the outcome from this device:

ஹிந்தில சொல்லுங்க  
हिंदी में कहें

hindee mein kahen

நீங்களும் உங்க தம்பியும் நாளைக்கு என்னோட வீட்டுக்கு சாப்பிட வரீங்களா  
क्या तुम और तुम्हारा भाई कल रात के खाने के लिए मेरे घर आओगे?

kya tum aur tumhaara bhaee kal raat ke khaane ke lie mere ghar aaoge?

தமிழ் நல்லா பேச வருமா

क्या आप अच्छी तमिल बोल सकते हैं?

kya aap achchhee tamil bol sakate hain?

நீங்க எல்லாரும் இந்தியாவுல எந்தெந்த இடத்துக்கு போய் இருக்கீங்கன்னு  
ஒன்னு விடாம சொல்றீங்களா

क्या आप मुझे बता सकते हैं कि आप सभी भारत में कहां गये हैं?

kya aap mujhe bata sakate hain ki aap sabhee bhaarat mein kahaan gaye hain?

கனடால சொல்லுங்க

ಕೆನಡಿಯನ್ ಎಂಡು ಹೇಳಿ

Kenadiyan endu hēḷi

நீங்க நாளைக்கு என்னோட வீட்டுக்கு வரீங்களா

ನಾಳೆ ನನ್ನ ಮನೆಗೆ ಬರುತ್ತೀಯಾ?

Nāḷe nanna manege baruttīyā?

நீங்க காலையில் சாப்பாடு சாப்பிட்டீங்களா

ನೀವು ಬೆಳಿಗ್ಗೆ ತಿಂಡಿತ್ತೀರಾ?

Nīvu beḷigge tindiddīrā?

The “speech to text” system converts input voice into Tamil text and the translation API translates the respective text to corresponding target language and outputs in the respective script. Correspondingly, the output text is both recorded in the database and converted to corresponding voice using the “text to speech” application. Each of the target language text as given here are uttered by the text to speech engine API correspondingly by exploiting all of the resources.

### **Accessing online text databases:**

It's quite possible to access online literary databases and convert them to voice as well as text with these resources. Following example shows how this device is used to access online Thirukkural databases:

Input: குறள் 52:

Output in both text as well as in voice: 52. பொருளாட்சி போற்றாதார்க்கு இல்லை அருளாட்சி ஆங்கில்லை ஊன்தின் பவர்க்கு. பொருள் ஆட்சி போற்றார்தார்க்கு இல்லைபொருளால் பயன் கோடல் அதனைப் பாதுகாவாதார்க்கு இல்லை; ஆங்கு அருள் ஆட்சி ஊன் தின்பவர்களுக்கு இல்லைஅது போல அருளாள் பயன் கோடல் ஊன் தின்பவர்களுக்கு இல்லை. விளக்கம் (பொருட்பயன் இழத்தற்குக் காரணம் காவாமை போல, அருட்பயன் இழத்தற்கு ஊன் தின்னல் காரணம் என்பதாயிற்று. ஊன் தின்றாராயினும் உயிர்கட்கு ஒரு தீங்கும் நினையாதார்க்கு அருள் ஆள்தற்கு இழுக்கு இல்லை என்பாரை மறுத்து, அஃது உண்டு என்பது இவை இரண்டு பாட்டானும் கூறப்பட்டது.)

### **System resources and converting them to Tamil speech:**

Besides the above online resources, it is also quite possible to interact with the system resources such as time and date and have this device record in text as well as speak it in Tamil as shown below:

Input: இப்ப மணி என்ன

Output: இப்பொழுது அமெரிக்கக் கிழக்கு நேரம் மாலை எட்டு மணி ஒன்பது நிமிடம் ஐம்பத்து ஒன்று வினாடி

Input: இன்னைக்கு என்ன தேதி

Output: இன்றைக்கு இரண்டாயிரத்து இருபத்து மூன்றுஆம் வருடம் பத்துஆம் மாதம் பதின் ஒன்றுஆம் நாள்

In order to convert Tamil numbers to corresponding Tamil text the resource from [http://robot.tamilnlp.com/py/convert\\_tamil\\_number.py](http://robot.tamilnlp.com/py/convert_tamil_number.py) is extensively used as part of this device.

### **Conclusion:**

With this initial analysis with a robot on the speech to text and text to speech resources that are available online, an attempt is made to test the performances of the robot (cf. Renganathan 2022) in a multiple number of ways. What is yet to be attempted, but in the process of being developed, is the process of analyzing output text from speech to text resources in a meaningful way to build some of the NLP tasks such as question-answering systems, man-machine interactions along the line of natural conversations and so on. Attempting to decipher the correct interpretation of commands involving ambiguous words would be a challenging task. As already mentioned, such tasks can be accomplished only when the training is made with extensive database containing all possible bi-directional predictable expressions. Capturing the nuances of expressions involving homonymous words, semantically extended phrases etc., are to be accounted for in a precise manner possible so further advances can be made. Such projects would mainly explore the intersection between the theoretical knowledge of linguistics and the linguistic performances related to the recent advances of AI particularly in the context of building LLM and development of vector databases. Obviously, as one can see that the linguistic performances of AI models such as Bard, Chat-GPT have made enormous successes mostly without the application of much of the knowledge from theoretical linguistics, but the outcome of these models, as has been cited in this work, requires proper application of linguistic theories further so a desired and most plausible outcome can be arrived at.

#### References:

- Noir, Nicole 2020. A dummy's guide to Bert. (<https://medium.com/swlh/bert-139acce0592d>).
- Renganathan, Vasu. 2021. Paper presented at the International Conference on Tamil Computing, TIC2021. “என் பேரு தமிழு (en pēru tamiḷu): A Speech Recognition Robot for Tamil” - [http://uttamam.org/papers/21\\_32.pdf](http://uttamam.org/papers/21_32.pdf) - (<http://robot.tamilnlp.com>).
- Renganathan, Vasu 2016. Computational Approaches to Tamil Linguistics. Cre-A, Chennai.