# BELA: Bot for English Language Acquisition

**Muskan Mahajan**
Sri Venkateshwar International School
Sector 18, Dwarka
New Delhi
muskanmahajan2004@gmail

## Abstract

In this paper, we introduce a conversational agent (chatbot) for Hindi-speaking youth called BELA—Bot for English Language Acquisition. Developed for the young underprivileged students at an Indian non-profit[1], the agent supports both Hindi and Hinglish (code-switched Hindi and English, written primarily with English orthography) utterances. BELA has two interaction modes: a question-answering mode for classic English language learning tasks like word meanings, translations, reading passage comprehensions, etc., and an open-domain dialogue system mode to allow users to practice their language skills.

We present a high-level overview of the design of BELA, including the implementation details and the preliminary results of our early prototype. We also report the challenges in creating an English-language learning chatbot for a largely Hindi-speaking population.

## 1 Introduction

Our paper introduces 'BELA', Bot for English Language Acquisition, an application of conversational agents (chatbots) in the domain of English language learning. BELA is developed for young underprivileged students at an Indian non-profit called Udayan Care. We were motivated to develop BELA for the students at Udayan Care because we observed a lack of volunteer support by the non-profit's English language mentors, leading to a halt in the mentees' second-language acquisition. Therefore, BELA is intended to emulate an English language mentor for the Udayan Care students, and support the non-profit's volunteers by reducing their workload.

Our conversational agent has two interaction modes: a retrieval mode to facilitate question-answering on classic English tasks like word meanings, translations, reading passage comprehensions, etc. (called the Tutor Bot), and a generative mode to facilitate open-domain chit-chat on general topics like the movies, songs, food, and environment (called the Buddy Bot).

Three tenets have governed the design of BELA:

1. **Support for Hindi utterances**: BELA is developed for a learner population which communicates largely in Hindi and Hinglish language (Hafiz, 2021). BELA's natural language understanding pipeline uses a language identifier, an Indic-language transliterator and a translator to support Hindi and Hinglish utterances.

2. **Reliability of answers to learners' queries**: BELA's responses to thesaurus/meaning-related queries are generated using tested translation and thesaurus APIs.[2]

3. **Graceful failure**: BELA's dialogue management system routes user utterances unrelated to language learning to the generative Buddy Bot.

Some challenges to developing BELA were the lack of data for intent classification and dialogue management, and a lack of a database of reading passages and English videos levelled by learner-proficiency level. Our paper discusses how we overcame these challenges.

**Organization**: The rest of the paper is organized as follows: We begin with a high-level overview of the Tutor Bot and the Buddy Bot, the two interaction modes of our conversational agent (Section 2); We next discuss the natural language understanding and dialogue management strategy of our conversational agent (Section 3); Further, we discuss in detail the first prototype implementation of the agent (Section 4); We next present related

[1] https://udayancare.org/

[2] https://developer.oxforddictionaries.com/



*Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 142 - 148

work (Section 5), and close with concluding remarks (Section 6)

## 2 Interaction Modes

Our conversational agent has two interaction modes: an English language question-answering mode called the Tutor Bot, and a general chit-chat mode called the Buddy Bot.

### 2.1 Tutor Bot

Tutor Bot is a retrieval-based response generator that provides answers classic English language learning queries. Some of these tasks as identified by us after a detailed survey with the Udayan Care mentees were: getting reading recommendations, word meanings, word antonyms/synonyms, 'word of the day,' English video recommendations, phrase pronunciations, writing prompts, phrase translations, grammatical/spelling corrections, and advice on the four core English skills (reading, writing, speaking, listening).

Every user utterance routed to the Tutor Bot is classified into one of these ten tasks, termed user intents, by the intent classifier. Further, the utterance is routed to a helper function corresponding to the identified intent. The helper function generates the required response. The design of these helper functions is described in Section 4.

### 2.2 Buddy Bot

Buddy Bot is a neural response generator that performs chit-chat on the following topics: movies, music, food, and environment. This interaction mode aims to help language learners learn new phrases, prepare the learners for conversations in real-life settings, and also help improve user adherence to the bot.

Buddy Bot uses the text completion endpoint of OpenAI's GPT-3 to generate a response based on the current user utterance and past conversations. The prompt design for the GPT-3 text completion model is discussed in great detail in Section 4.

## 3 Natural Language Understanding & Dialogue Management

The natural language understanding and dialogue management system of our agent is simple and intuitive.

### 3.1 Natural Language Understanding

The user utterance is first routed to a language identifier; BELA uses the XLM-RoBERTa Transformer model[3] from HuggingFace for language detection. If the detected language is Hindi, it is run through a Python API for transliteration[4]. The transliterated text, which is in the Devnagari script, is passed through a Transformer-based Machine Translator from Salesken.ai[5].

The final output is an English query that is routed to the dialogue management system, discussed below.

### 3.2 Dialogue Management

Firstly, the user utterance is routed to the mode classifier of the dialogue management system to classify the query as being related to English learning (for eg: asking for the translation of a sentence) or not (for eg: asking for an opinion on a movie actor).

If the query is unrelated to English learning, it is routed to Buddy Bot. If the query is related to English learning, it is routed to the Tutor Bot. Here, the query is classified into one of ten intents discussed in Section 2. The following section discusses the mode classifier and intent classifier in greater detail.

#### 3.2.1 Mode classifier

The mode classifier is a binary classifier to predict whether a user utterance is related to English learning. To classify the user utterance, we use the output from a BERT encoder as the input to a linear classification layer trained with a crossentropy loss function.

The classifier dataset consists of utterances that are related to English-language learning (positive examples), and general utterances (negative examples). The positive examples were taken from the dataset created for the English-query intent classifier. The general utterances are sampled from user discussions on the following subreddits[6]: r/Food, r/Movies, r/MovieDetails, r/MusicSuggestions, r/AskReddit, r/AskScience, r/Politics, r/AskSocialScience, and r/AskGames.

The training data information is shown in Table 1. And the evaluation results are shown in Table 2.

[3]https://huggingface.co/papluca/xlm-roberta-base-language-detection

[4]https://pypi.org/project/google-transliteration-api/

[5]https://huggingface.co/salesken/translation-hi-en

[6]a subreddit is a forum dedicated to a specific topic on the website Reddit.

#### 3.2.2 English query intent classifier

The query intent classifier is a multi-class classifier to predict the nature of the user's English learning query. The user utterance is classified into one of the following ten intents: *getReadRecommendations, getWordMeaning, getSynonymAntonym, getWordOfTheDay, getPrononciation, getVideoRecommendations, getTranslation, getWritingPrompts, getCorrection*, and *getAdvice*. This classifier uses the output from a BERT encoder as the input to a linear classification layer trained with a crossentropy loss function.

To train our classifier, we created a dataset of utterances and the corresponding intent/query label. Since the training data size is of utmost importance for text classification tasks, we have used text augmentation techniques like back translation, and paraphrase generation using Parrot Paraphraser (Damodaran, 2021). We have also included utterances with spelling mistakes in our dataset to make the classifier robust to the common spelling mistakes made by the language learner.

The training data information is shown in Table 3. And the evaluation results are shown in Table 4.

| | |
|---|---|
| Train | 3040 |
| Validation | 380 |
| Test | 380 |

Table 1: Mode Classifier Data

| | |
|---|---|
| Train accuracy | 0.998 |
| Test accuracy | 0.987 |

Table 2: Mode Classifier Evaluation Results

## 4 BELA Prototype Implementation

### 4.1 Tutor Bot Implementation

In the previous section, we discussed that the user utterance/query classified by the mode classifier as related to English learning is routed to the Tutor Bot. Here, the query is classified into one of ten intents by the intent classifier. In the following section, we discuss the helper function related to each user intent of the Tutor Bot, and the datasets used to create them.

#### 4.1.1 Helper-function Datasets

1. **CEFR level predictor dataset**

   This is a dataset[7] provided by Adam Montgomerie to predict the Common European Framework of Reference for Languages (CEFR) level of a blob of text, a measure of English text complexity for an English as Second Language (ESL) learner. The dataset contains 1500 example texts split over the 6 CEFR levels. The texts are a mixture of dialogues, stories, articles, and other formats. (Montgomerie, 2021)

   We used these passages for training a TFIDF-based CEFR level predictor which achieves 27.6% more accuracy than the baseline described by Montgomerie (Table 5).

2. **CEFR levelled reading passages**

   We scraped reading passages from an ESL website[8] with free reading exercises and saved them to a file called passages.csv. Subsequently, we passed these passages through the CEFR-predictor trained by us; and stored the passage-CEFR label pairs in a file called cefr-levelled-passages.csv.

   We use these passages for the 'Reading recommendation' helper function discussed in Section 4.1.2.

3. **CEFR levelled word list** We created a list of words and their corresponding CEFR label and stored it in cefr-levelled-words.csv. The list was scraped from English Vocabulary profile[9], a website with information about words and phrases used by learners at each CEFR level.

   We use this list of words for the 'Word of the day' helper function discussed in Section 4.1.2.

4. **CEFR levelled videos**

   We used the TED – Ultimate Dataset[10] from Kaggle to retrieve a set of educational English-language videos, their titles, URLs, descriptions and transcripts. Then, we found the CEFR level of each video using the CEFR level predictor on the video transcripts. The

[7] https://github.com/AMontgomerie/CEFR-English-Level-Predictor/tree/main/data

[8] https://www.myenglishpages.com/english/

[9] https://www.englishprofile.org/wordlists

[10] https://www.kaggle.com/datasets/miguelcorraljr/ted-ultimate-dataset

| | |
|---|---|
| Train | 1520 |
| Validation | 190 |
| Test | 190 |

Table 3: Intent Classifier Data

| | |
|---|---|
| Train accuracy | 0.997 |
| Test accuracy | 0.995 |

Table 4: Intent Classifier Evaluation Results

video links, descriptions and their CEFR labels are stored in cefr-levelled-tedtalks.csv.

We use these videos for the 'Video recommendation' helper function discussed in Section 4.1.2.

#### 4.1.2 Helper functions

1. **Reading Recommendation helper function**

   This function prompts the user with four questions[11] to assess their CEFR level, i.e. their English proficiency level. The CEFR level is stored in the chatbot state for other helper functions.

   After determining the CEFR level, the function retrieves a reading passage of the same CEFR level from cefr-levelled-passages.csv. This passage is also accompanied by three multiple-choice questions (MCQs) on the passage to facilitate top and bottom-up processing of the text. (British Council, 2001) The MCQs are generated by OpenAI's GPT-3 fed with a prompt shown in Figure 4.

2. **Word Meaning helper function**

   This function performs the role of identifying the word whose meaning is to be found using regex extraction techniques and providing its definition and examples using the Oxford Thesaurus API.

3. **Synonym-Antonym helper function**

   This function performs the role of identifying the word whose synonym/antonym is to be found using regex extraction techniques and providing it using the Oxford API.

4. **Word of the Day helper function**

   The function provides the user with a new word daily accompanied by its definition and usage to help augment the user's vocabulary. The new word given is in line with the user's CEFR level. The CEFR level is either inferred from the chatbot state or assessed using the four question-survey discussed earlier. The words are retrieved from cefr-levelled-words.csv

5. **Video Recommendation helper function**

   The function provides the user with a TedTalk link, accompanied by its description. The video is in line with the user's CEFR level and provides the learner with an opportunity for language immersion. The video and the description are retrieved from cefr-levelled-tedtalks.csv

6. **Pronunciation helper function**

   This function uses Google Translate's Text-to-Speech API[12] with the language parameter set to 'English' and tld (top-level domain) parameter set to "co.in" to provide pronunciations of English words/phrases in an Indian accent.

7. **Writing prompts helper function**

   This function provides the user with a letter, essay or miscellaneous writing prompt based on their request, to help them improve their writing skills. The prompts are retrieved from writingPrompts.csv.

8. **Grammar/Spelling helper function**

   This function performs the two-pronged role of identifying the phrase/word to be corrected using regex extraction techniques and correcting it using 'Gramformer'. Gramformer[13] is a framework that uses Transformers to detect and correct grammatical errors in natural language text.

9. **Translator helper function**

[11] Based on https://bit.ly/CEFR_Self_assessment

| | |
|---|---|
| Baseline Accuracy | 41.8% |
| Train Accuracy | 72.2% |
| Test Accuracy | 69.4% |

Table 5: CEFR Predictor Evaluation Results

[12] https://pypi.org/project/gTTS/

[13] https://github.com/PrithivirajDamodaran/Gramformer

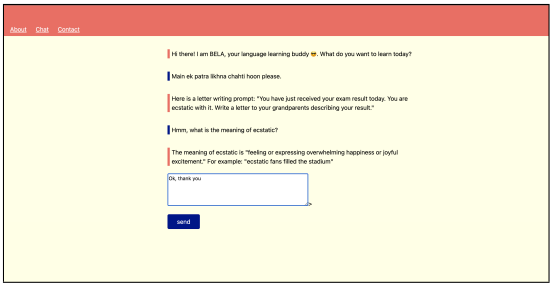


Figure 1: An interaction between the user and the Tutor Bot. Here, the user makes a query in the Hinglish language which translates to "I want to write a letter today."

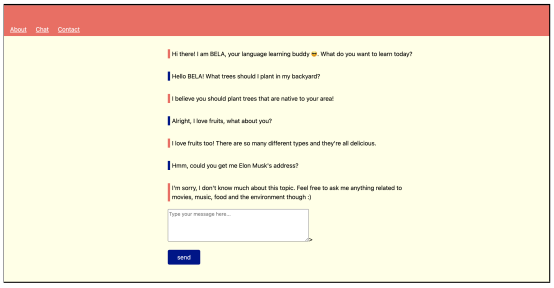


Figure 2: An interaction between the user and the Buddy Bot. Here, the agent politely nudges the user to a relevant topic if they discuss anything beyond movies, music, food and environment.

This function identifies the word/phrase whose translation is to be found from the user utterances using regex extraction techniques; identifies the language of the phrase; and returns a translation. The function uses the XLM-RoBERTa[14] transformer model from HuggingFace for language detection and a Transformer-based Machine Translator from Salesken.ai[15] for Hindi-English translation.

10. **Advice helper function**

    This function provides the user with two pieces of advice for improving each of these English language skills: Listening, Speaking, Reading, and Writing (LSRW). These pieces are taken from credible research focused on LSRW skill acquisition for ESL learners. (Gomathi, 2014)

### 4.2 Buddy Bot Implementation

Buddy Bot is a neural response generator that performs chit-chat on the following topics: movies, music, food and environment. This interaction mode aims to help language learners learn new phrases, prepare the learners for conversations in real-life settings, and also help improve user adherence to the bot.

Buddy Bot uses the text-completion endpoint of OpenAI's GPT-3 to generate a response based on the current user utterance and past conversation. The text completion model is 'programmed' using a prompt (Figure 3) that provides instruction on how the BuddyBot should function. The prompt gives the text completion model an identity: a "chit-chat bot that talks to users on the topics of movies, music, food and the environment." Before responding to the user, the bot also performs a topic-relevance check- is the user utterance related to one of the four topics? This behaviour was injected into the model by providing two examples to the GPT-3 prompt. If the user-utterance is not related to one of the four topics, the Buddy Bot politely nudges the user to it.

We limited the scope of conversations of the Buddy Bot to just four topics to prevent the extraction of sensitive data, including personally identifiable information (PII) — names, phone numbers, addresses, etc., through training data extraction attacks. (Carlini et al., 2020)

## 5 Related Work

### 5.1 Hindi and Hinglish Conversational Agents

Indian telecom companies like Haptik (Haptik.AI, 2021b) and AmplifyReach have developed multilingual chatbots that support Hindi and Hinglish languages. However, these bots are dedicated to the domain of customer service and use proprietary software (Haptik.AI, 2021a) for multilingual natural language understanding.

### 5.2 Using Dialogue Systems for Learning

Li et al. (2022) developed an online language learning tool to provide learners with conversational experience by using dialog systems as conversation practice partners. The conversational agent simulated a human resource professional interviewing users as potential job candidates; the researchers also explored making the system more adaptive to user profile information by using reinforcement learning algorithms.

In another work, Ruan et al. (2021) created 'EnglishBot', which used Automatic Speech Recognition to converse with students interactively on

[14] https://huggingface.co/papluca/xlm-roberta-base-language-detection

[15] https://huggingface.co/salesken/translation-hi-en

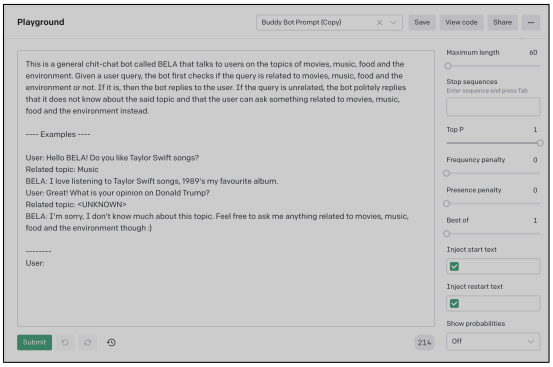


Figure 3: GPT-3 prompt for the Buddy Bot

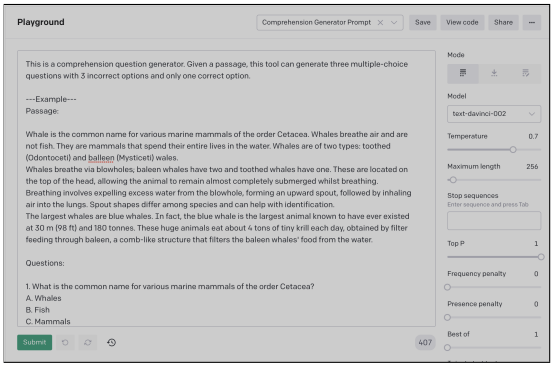


Figure 4: GPT-3 prompt to generate multiple-choice questions (MCQs) of the reading passages.

college-related topics and provided adaptive feedback.

## 6 Conclusion

BELA is our first step toward making personalised second-language acquisition more accessible to Hindi-speaking learners. Our future work would focus on increasing the range of English learning tasks that BELA can assist with, improving the Hinglish language understanding pipeline and making the dialogue management system more robust to failure.

## Limitations

BELA's Tutor Bot can only cater to limited English language learning tasks. Therefore, our future work will focus on adding more skills to the Tutor Bot, including the ability to paraphrase passages, make edits to passages, provide exercises based on grammar topics, etc.

BELA's natural language understanding pipeline tends to translate the named entities in the Hinglish queries. For example, here is a query in the Hinglish language: "Translate *mujhe jio ka sim chahiye* to English." This query literally means "Translate *I want a Jio sim*," where Jio is the name of a telecom company. However, the NLU Pipeline infers Jio as the hindi verb meaning life and outputs the response "I want a live sim."

India also has regional variations of the Hinglish language. As we get more people to use BELA, we aim to use the user messages to improve BELA's natural language understanding pipeline.

Finally, while GPT-3 used in the Buddy Bot provides detailed and context-aware responses to general chit-chat queries, the presence of a pay-wall to the GPT-3 API limits the scalability of the Buddy Bot.

## Ethics Statement

In today's globalised economy, English fluency has become important to facilitate communication and improve a person's job prospects. BELA is our first step toward making personalised English-language acquisition more accessible for the young students at Udayan Care. However there are a few ethical challenges to deploying BELA, especially the Buddy Bot interaction mode:

1. **GPT-3 and Toxicity**: The Buddy Bot, which is based on GPT-3, a large-language model, can have the tendency to generate offensive text. Therefore, we have to anticipate and plan for text-generation mishaps either by adding more safeguards to the text generation prompts, or by fine-tuning the Buddy Bot on more examples to make it robust to adversarial user input.

2. **Fine-tuning GPT-3 on Indic-language data**: We need to fine tune the Buddy Bot on Indic-language dialog datasets to allow it to support languages like Hindi and Hinglish. This is a challenge because dialog generation data for low-resource languages is scarce.

## References

The British Council. 2001. Top down.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. *CoRR*, abs/2012.07805.

Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for nlu.

B.S. Gomathi. 2014. Enriching the skills of rural students with effective methods of teaching english language using lsrw skills. In *International Journal of Education and Information Studies*, pages 65–69.

Jasmin Hafiz. 2021. https://www.milestoneloc.com/guide-to-hinglish-language/.

Haptik.AI. 2021a. Linguist pro - building multilingual chatbots for business.

Haptik.AI. 2021b. The next big thing for multilingual chatbots: Hinglish.

Yu Li, Chun-Yen Chen, Dian Yu, Sam Davidson, Ryan Hou, Xun Yuan, Yinghua Tan, Derek Pham, and Zhou Yu. 2022. Using chatbots to teach languages. In *Proceedings of the Ninth ACM Conference on Learning @ Scale*, L@S '22, page 451–455, New York, NY, USA. Association for Computing Machinery.

Adam Montgomerie. 2021. Attempting to predict the cefr level of english texts.

Sherry Ruan, Liwei Jiang, Qianyao Xu, Zhiyuan Liu, Glenn M Davis, Emma Brunskill, and James A. Landay. 2021. Englishbot: An ai-powered conversational system for second language learning. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 434–444, New York, NY, USA. Association for Computing Machinery.

# A Appendix

## A.1 Helper function datasets

| | |
|---|---|
| cefr-levelled-passages.csv | Link |
| cefr-levelled-words.csv | Link |
| cefr-levelled-tedtalks.csv | Link |
| writingPrompts.csv | Link |

Table 6: Helper function datasets

## A.2 Dialogue Management classifier datasets

| | |
|---|---|
| Mode Classifier Dataset | Link |
| Intent Classifier Dataset | Link |

Table 7: Dialogue Management classifier datasets

## A.3 GPT-3 Prompts

| | |
|---|---|
| Reading Comprehension MCQ | Link |
| BuddyBot | Link |

Table 8: GPT-3 Prompts