TRANSLITRATION FROM HINGLISH TO HINDI (SULEKH)

by

VALLABHANI SIVAMANI RAMA KRISHNA 421269

MAYANK KUMAR CAM   421218

Under the guidance of

Dr. HIMABINDU K

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY ANDHRA PRADESH

TADEPALLIGUDEM-534101, INDIA

APRIL 2023

leaving second page

TRANSLITERATION FROM HINGLISH TO HINDI

Thesis submitted to
National Institute of Technology Andhra Pradesh
for the award of the degree

of

Bachelor of Technology

by

VALLABHANI SIVAMANI RAMA KRISHNA 421269
MAYANK KUMAR CAM   421218

Under the guidance of

Dr. HIMABINDU K

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY ANDHRA PRADESH
TADEPALLIGUDEM-534101, INDIA
APRIL 2023

DECLARATION

I declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Vallabhani Sivamani Rama Krishna   Mayank Kumar Cam

421269   421218

Date:_____   Date:_____

CERTIFICATE

It is certified that the work contained in the thesis titled "**Transliteration from Hinglish to Hindi**" by "Vallabhani Sivamani Rama Krishna, bearing Roll No: 421269" and "Mayank Kumar Cam,bearing Roll No: 421218" has been carried out under our supervision and that this work has not been submitted elsewhere for a degree.

Signature
Dr. Himabindu K
DCSE
N.I.T. Andhra Pradesh
April, 2023

Abstract

" Transliteration from Hinglish to Hindi " is a Sentence Transcription Management System that aims to automate and simplify the process of Transcripting Hinglish to Hindi correctly. The system offers features like ask sentence for user, answer sentence to annotator, verify sentence to verifier, confirm sentence verified by verifier to confirmer and admin to appoint someone as verifier or confirmer and to see feedbacks. It is a web-based application built using HTML, CSS, JavaScript, Bootstrap, AI Bharat ai tool to easily transcript and Django with an SQL database. The system will improve the quality and correctness of the Hindi sentence, reduce the workload of the annotators because ai tool use, trustable because of three level of confirmation. " Transliteration from Hinglish to Hindi " will provide a better platform than existing to write Hindi correctly from Hinglish.

TABLE OF CONTENTS

Contents

1 Introduction

Our mini project, Transliteration from Hinglish to Hindi, is a web-based appilcation .

Transliteration is the process of converting text from one script to another. It involves representing the sounds of one language using the characters of another language. Transliteration from Hinglish to Hindi is a challenging task due to the inherent complexities of the two languages. Hinglish, as the name suggests, is a hybrid of Hindi and English, which is widely used in India, especially in informal settings. However, writing Hinglish in the Devanagari script, which is used for Hindi, poses significant challenges due to the lack of standardized rules and variations in pronunciation and usage.

To address this problem ,we have developed a web-based application that simplifies the process of transliterating Hinglish script to Hindi. The system utilizes machine learning tool to generate multiple responses for every word in the sentence along with a custom hindi text box where the users can enter hindi word if he things the 10 options are not close enough , which are then verified and confirmed by a team of annotators, verifiers, and confirmers. This three-step verification process ensures the accuracy and consistency of the transliterated text, making it more reliable than existing systems.

The system is designed to be user-friendly and accessible to a wide range of users. Users can log in to the portal and request to have Hinglish sentences transcribed into Hindi. The portal includes dashboards for users, annotators, verifiers, and confirmers, and task assigner making it easy to track the progress of each transcription request.

The system is built using a combination of AI Bharat AI tool, Django, HTML, CSS, and JavaScript, which allows for scalability and flexibility. The tool can handle large volumes of data and adapt to new language variations, making it ideal for organizations that deal with multilingual content on a regular basis.

Overall, our project aims to make the process of transliterating Hinglish to Hindi more accurate, efficient, and accessible to users. The system offers several advantages over existing solutions, including improved accuracy, scalability.

2 EXISTING SYSTEM AND DISADVANTAGES

- Many existing systems rely on simple phonetic conversions, which can lead to inaccuracies in the transliteration.

- Machine tools may not always be accurate in transcribing sentences, especially when it comes to mixed-language sentences or those with complex grammar.

- Some existing machine tools may have a limited vocabulary and may not recognize certain words or phrases, leading to incorrect transcription.

- Existing systems do not allow users to customize the transliteration process based on their preferences or specific needs.

- The manual process of transliteration used by many existing systems can be time-consuming and inefficient.

- Machine tools may not be able to understand the context in which a word or phrase is being used, leading to errors in transcription.

- Machine tools may not be able to pick up on nuances in language, such as sarcasm or irony, leading to incorrect transcription.

- Machine tools lack the human touch and may not be able to interpret tone, mood or emotion in a sentence, leading to errors or misinterpretations.

3   PROPOSED SYSTEM AND ADVANTAGES

- Accuracy: The use of a machine learning model ensures greater accuracy in transliteration. Users tend to answer more accurately when provided with options.

- Efficiency: The system can transcribe mixed-language sentences quickly and efficiently, saving time and effort for users, and reducing computation power usage.

- Enhanced user experience: By providing an easy-to-use platform for transcribing mixed-language sentences, the website can enhance the user experience and make it easier for users to work with content in multiple languages.

- Scalability: The use of a machine learning model and user input to transcribe sentences allows for scalability and the ability to handle large volumes of content in multiple languages.

4 TECHNOLOGY STACK

Front-end development:

- HTML

- CSS

- Bootstrap 5.3

- JavaScript

Back-end development:

- Python Django

Database language:

- Postgres SQL (Structured Query Language)

- Django Models

5 INSIGHTS INTO OUR PORTAL

1. User-friendly interface: Our portal has a clean and intuitive interface that is easy to navigate, making it user-friendly for people with different levels of technical expertise.

2. Secure login: We have implemented a secure login system that ensures the privacy and security of our users' data. Users can log in using their unique usernames and passwords, which are encrypted to protect them from unauthorized access.

3. Dashboard: We have developed dashboards for different user roles, including annotators, verifiers, confirmers, and task assigners. These dashboards display the relevant tasks and information for each role, enabling users to manage their tasks efficiently.

4. Custom input: Our portal allows users to provide custom input and suggest alternative transliterations, enabling our system to learn and improve over time.

5. Verification process: We have implemented a three-step verification process that ensures the accuracy and quality of our transliterations. Annotators provide initial answers, verifiers check and verify the answers, and confirmers finalize the answers after they have been verified by the verifiers.

6. Data management: Our portal also includes data management tools that enable users to manage and organize the data efficiently. Users can create and manage datasets, assign tasks, and track progress, ensuring the smooth and efficient functioning of the system.

7. Robust security: Our portal employs robust security measures to protect user data and ensure the integrity of the system. User passwords are encrypted and stored securely, and access to sensitive data is restricted to authorized personnel only.

6 FUNCTIONALITY OF THE PORTAL

6.1 Admin

The admin portal is a crucial component of any website, as it allows website administrators to manage and maintain the website's content, users, and other aspects. In our website, the admin portal provides additional functionality by allowing the admin to promote users on the basis of their ratings.

This feature is particularly useful for websites that rely on user-generated content where the quality and accuracy of user contributions is important. With this functionality, the admin can reward active and helpful users by promoting them to higher levels, granting them additional privileges and responsibilities. This, in turn, can encourage other users to contribute more and engage with the website, leading to a more vibrant and active community.

In our website the users cannot decide by themselves what role the want we by default make them as annotators later if we like their performance the admins will promote them. if a non admin user logs in he cannot see the any role assigning option in our website again if he tries to go to the link by entering path of the role assignment it will again verify his eligibility if he is not eligible, he will be redirected to dashboard page. In this way we are taking utmost care to make our website secure.

Overall, the admin portal provides a powerful set of tools that allow the admin to maintain the website's functionality and user experience

Admin dashboard:

6.2 User

The user portal in our website provides users with a range of functionalities that make it easy for them to get their sentences transcribed accurately and conveniently. Firstly, users can ask any sentence of any length on the portal or upload a file for transcription. This feature allows users to submit their content quickly and easily, whether it's a short sentence or a longer document, and get it transcribed without any hassle. The system is designed to accept XCEL sheets and the users can select the type of transcription they need - Hinglish to Hindi or vice versa.

Additionally, users can view their previous submissions and check whether they have been transcribed or not. This feature provides greater transparency and accountability in the transcription process, allowing users to keep track of their content and ensure that they receive accurate and timely transcriptions.

Moreover, users can see all confirmed questions and can search the sentence they want in our custom search box. These users doesn't have any access to our website other than these parts all accesses are restricted for them .

Overall, the user portal in our website is designed to provide a user-friendly and efficient experience for users who need to get their content transcribed. With features such as the ability to submit any sentence or file, view previous submissions, users can get high-quality transcriptions that meet their needs while also helping to improve the quality of the service.

User dashboard:
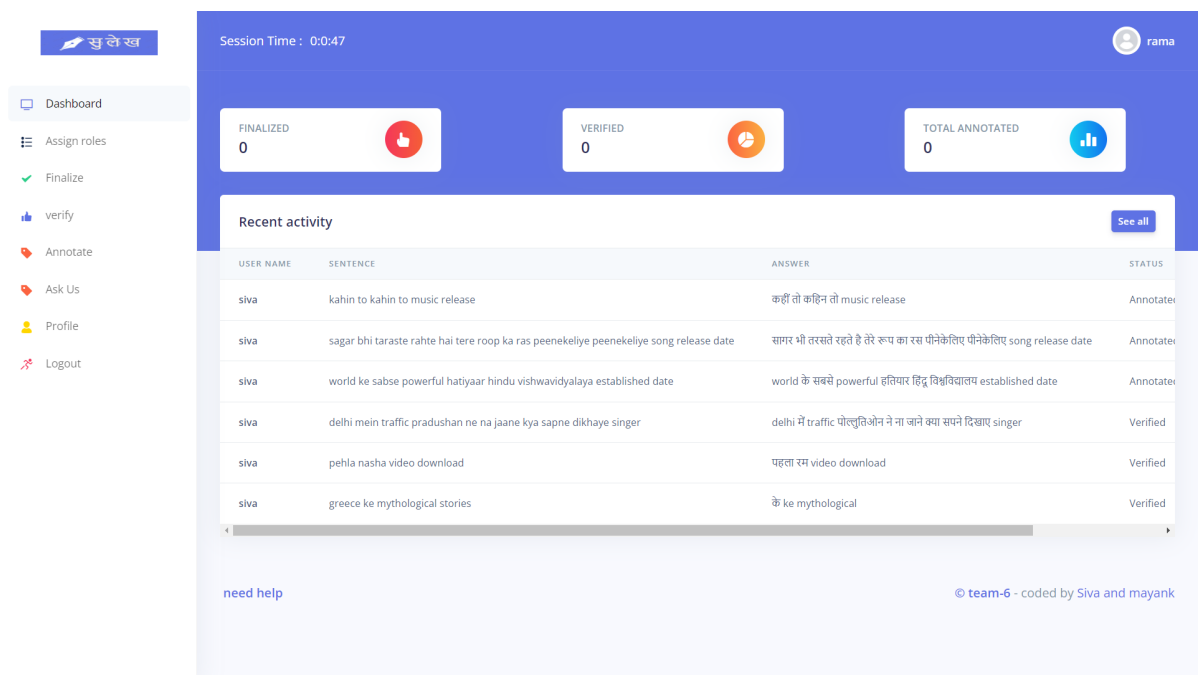
Screenshot 1

सु ले ख

Session Time : 0:3:23   rama

- Dashboard
- Assign roles
- Finalize
- verify
- Annotate
- Contribute
- Profile
- Logout

Enter dataset name :

Choose file :   Choose file   No file chosen

submit

Recent activity   See all

DATASET	FROM	TO

Screenshot 2

सु ले ख

Session Time : 0:0:26   siva

- Dashboard
- Assign roles
- Finalize
- verify
- Annotate
- Contribute
- Profile
- Logout

Search

SENTENCE	ANSWER
pal pal dil ke paas tum rehti ho song composer	पाल पाल दिल के पास तुम रहती हो song composer
US open mein indian khiladi bharat ka national sport	US open में indian खिलाड़ी भारत का national sport
crazy kiya re singer	crazy किया रे singer
amitabh bachchan ki latest movie	amitabh bachchan की latest movie
puri jagannadh yatra latest news	puri jagannadh यात्रा latest news
pehla nasha video download	पहला नशा video download
delhi mein traffic pradushan ne na jaane kya sapne dikhaye singer	delhi में traffic प्रदूषण ने ना जाने क्य सपने दिखाए singer
world ke sabse powerful hatiyaar hindu vishwavidyalaya established date	world के सबसे powerful हतियार हिंदु विश्वविद्यालय established date
sagar bhi taraste rahte hai tere roop ka ras peenekeliye peenekeliye song release date	सागर भी तरसते रहते है तेरे रूप का रस पीनेकेलिए पीनेकेलिए song release date
ab mujhe raat din tumhara hi khayal hai video download	अब मुझे रात दिन तुम्हरा ही खयल है video download
oh jaanam song composer	ओह जानम song composer

6.3 Annotator portal

The annotator portal in our website is designed to provide annotators with an intuitive and efficient way to transliterate sentences requested by users. Annotators can log into their account and access the sentences that are available for transliteration. They can then read through the sentence and provide an accurate transliteration of the Hinglish or Hindi sentence requested by the user. After completing the transliteration, annotators can tag the sentence as" annotated" and submit it for review. Once a question is submitted no annotator can annotate it again it can only be verified.

Concurrency is the ability of multiple users to access and modify the same data at the same time. In our website, we maintain concurrency by using a lock tag for every sentence in the dataset. When a user selects a sentence to annotate, the sentence is locked and marked as "locked" in the database. This prevents other users from accessing the same sentence and modifying it simultaneously, which could lead to data inconsistencies and errors.

While the sentence is locked, the user who locked it can modify the sentence and add tags or meanings. Once they submit or skip the sentence, it will be unlocked and marked as "unlocked" in the database, allowing other users to access and modify it. By using the lock tag, we ensure that each sentence is annotated by only one user at a time, avoiding conflicts and preserving data integrity. This mechanism enables multiple users to work on the same dataset simultaneously without affecting the quality and accuracy of the annotations.

Furthermore, the annotator portal provides a profile portal for the annotator, where they can view their ratings and scores. The ratings and scores are based on the accuracy and reliability of the transcriptions provided by the annotator. This feature provides an incentive for annotators to provide high-quality transcriptions, as they can see their progress and strive to improve their ratings and scores over time. The profile portal also allows annotators to view their history of submissions, including the sentences they have annotated and any feedback or comments they have received from users or admins.

Overall, the annotator portal provides a simple and effective way for annotators to transcribe sentences accurately and efficiently. With the added feature of ratings and scores, annotators are motivated to provide the best possible service, which helps to ensure the quality and reliability of the transcriptions provided on the website Annotator dashboard:

6.4 Verifier portal

The verifier portal in our website is designed to provide verifiers with an easy and effective way to verify the sentences that have been tagged as "annotated" by annotators. Verifiers can log into their account and access the sentences that are available for verification. They can then read through the sentence and review the transcription provided by the annotator.

If the verifier is satisfied with the transcription, they can mark the sentence as "verified". However, if they are not happy with the transcription, they have the option to change the transcription and mark the sentence as "verified". This feature ensures that the transcriptions provided on the website are accurate and reliable, and that any errors or inaccuracies are corrected in a timely manner.

Similar to annotators, verifiers can also access and modify the same data at the same time. To maintain concurrency in our website, we use a lock tag for every sentence in the dataset. When a verifier selects a sentence to verify, the sentence is locked and marked as "locked" in the database. This prevents other users, including annotators and other verifiers, from accessing and modifying the same sentence simultaneously, which could lead to data inconsistencies and errors. While the sentence is locked, the verifier can modify the existing tags and meanings, and add new ones if necessary. Once they submit or skip the sentence, it will be unlocked and marked as "unlocked" in the database, allowing other users to access and modify it.

By using the lock tag, we ensure that each sentence is verified by only one verifier at a time, avoiding conflicts and preserving data integrity. This mechanism enables multiple verifiers to work on the same dataset simultaneously without affecting the quality and accuracy of the annotations.

Overall, the verifier portal plays a critical role in ensuring the accuracy and reliability of the transcriptions provided on the website. With the ability to verify and make corrections to the transcriptions, verifiers can help to improve the quality of the service provided to users.

Verifier dashboard:

6.5   Confirmer portal

The confirmer portal in our website is designed to provide confirmers with the ability to confirm the sentences that have been verified by the verifier. The confirmed sentences are considered to be the final version of the transcription and are made available to the users. Confirmer can log into their account and access the sentences that are available for confirmation. They can then review the verified transcriptions provided by the verifier.

If the confirmer is satisfied with the verified transcription, they can mark the sentence as "confirmed" and make it available to the users. However, if they are not happy with the verified transcription, he can modify it.

This feature ensures that any inaccuracies or errors in the transcription are corrected before the sentence is confirmed and made available to the users.

Similar to annotators and verifiers, confirmers can also access and modify the same data at the same time. To maintain concurrency in our website, we use a lock tag for every sentence in the dataset. When a confirmer selects a sentence to confirm, the sentence is locked and marked as "locked" in the database. This prevents other users, including annotators, verifiers, and other confirmers, from accessing and modifying the same sentence simultaneously, which could lead to data inconsistencies and errors.

While the sentence is locked, the confirmer can review the existing tags and meanings, and modify them if necessary. Once they confirm or skip the sentence, it will be unlocked and marked as "unlocked" in the database, allowing other users to access and modify it.

By using the lock tag, we ensure that each sentence is confirmed by only one confirmer at a time, avoiding conflicts and preserving data integrity. This mechanism enables multiple confirmers to work on the same dataset simultaneously without affecting the quality and accuracy of the annotations.

Overall, the confirmer portal plays a crucial role in ensuring the accuracy and reliability of the transcriptions provided on the website. With the ability to confirm and make corrections to the verified transcriptions, confirmers can help to improve the quality of the service provided to users. Additionally, the ability to send sentences back for re-annotation ensures that any inaccuracies or errors are corrected before they are made available to the users.

Confirmer dashboard:

19

7 INSIGHTS INTO THE DATABASE USED

Entities:

- Admin
- User
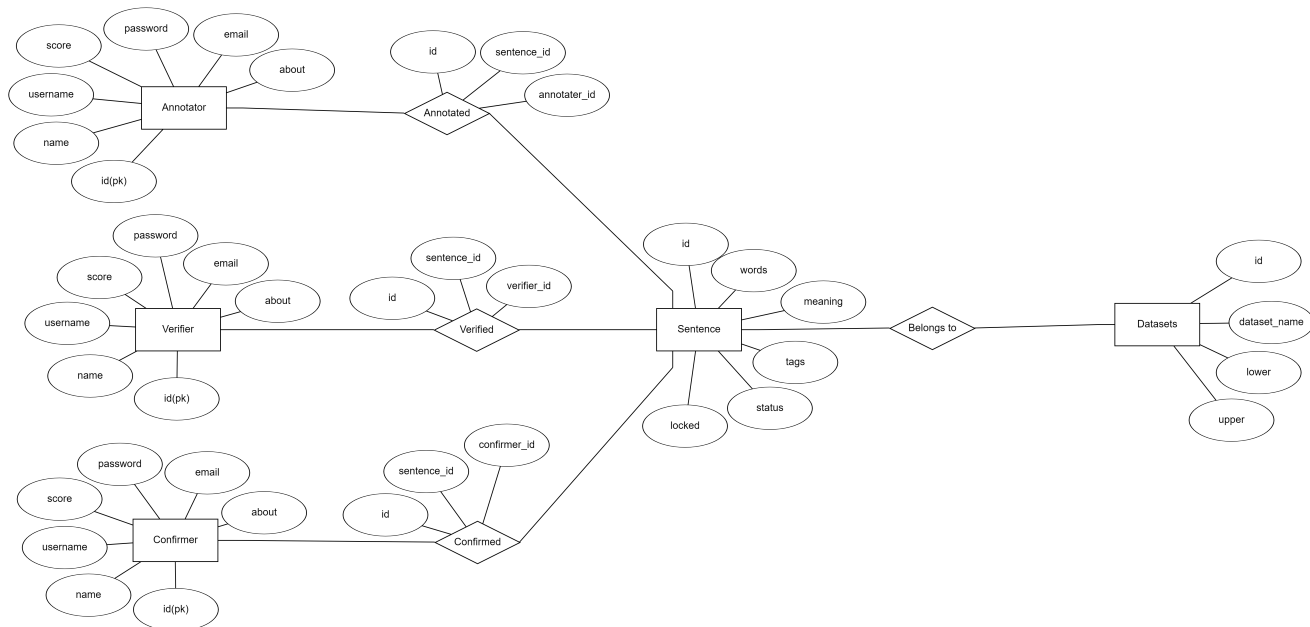- Annotator
- Verifier
- Confirmer
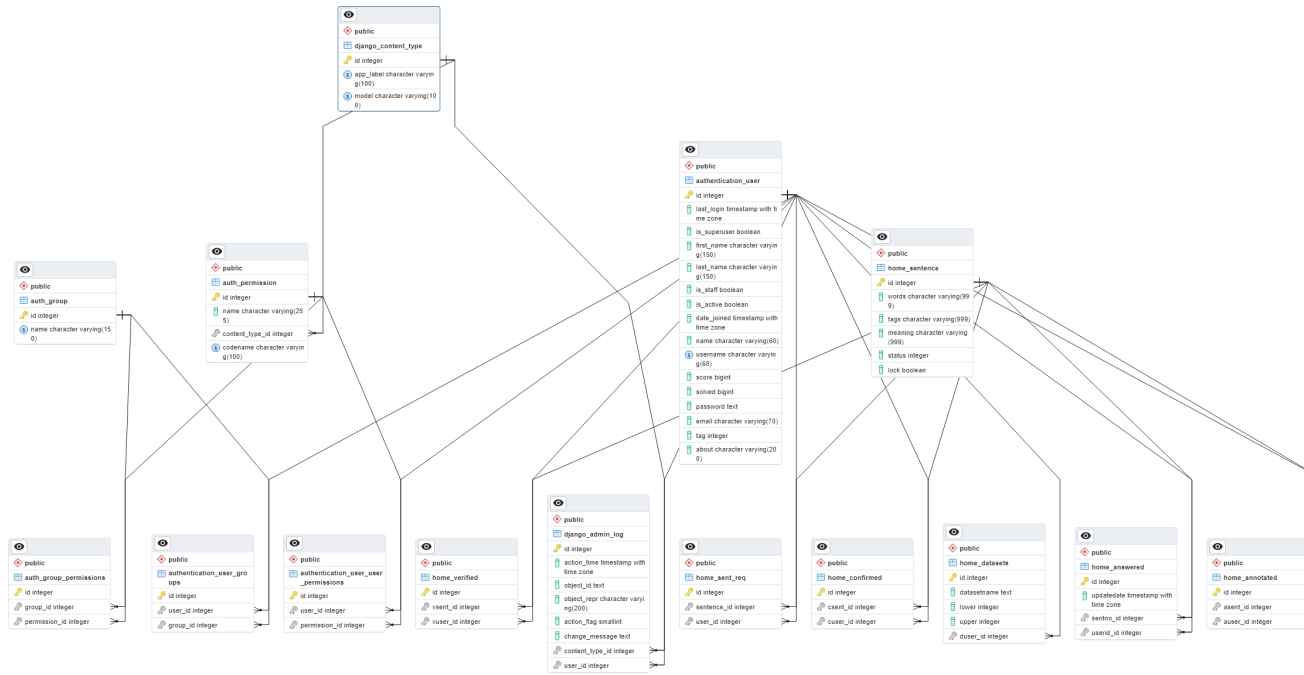- Sentence
- Datasets

Relationships:

- Answered

Initially, the database was in 1st Normal Form. After refining the schema by normalization, it was converted to 3rd Normal Form.

ER-diagram:

7.0.1 ER diagram

- This is a Django model named Sentence which represents a sentence to be transcribed from Hinglish to Hindi. It has the following fields:

  - words: A character field that stores the words of the sentence.

  - tags: A character field that stores the tags for each word in the sentence. These tags are used to identify which words need to be transliterated from Hinglish to Hindi.

  - meaning: A character field that stores the meaning of the sentence in Hindi, after it has been transcribed.

  - status: An integer field that stores the status of the sentence. This can be used to track the progress of the transcription process.

  - lock: A boolean field that indicates whether the sentence is currently being worked on by an annotator, verifier, or confirmer.

- This is a Django model for a table called Answered. The table has three columns:

  - sentno is a foreign key to the Sentence table and specifies the sentence number that has been answered.

  - userid is a foreign key to the User table and specifies the user who answered the sentence.

  - updatedate is a DateTimeField that stores the date and time when the answer was updated.

- This is a Django model for a dataset. It has the following fields:

  - datasetname: A text field representing the name of the dataset.

  - duser: A foreign key to the User model, representing the user who owns the dataset.

  - lower: An integer field representing the lower bound of the dataset (i.e., the ID of the first sentence in the dataset).

  - upper: An integer field representing the upper bound of the dataset (i.e., the ID of the last sentence in the dataset).

8 Conclusion

In conclusion, the development of a website that utilizes the language transcription system offers several benefits for users and organizations working with mixed-language content. The transcription system enables accurate and efficient transcription of mixed-language sentences, which saves time and effort for users who need to work with such content. By providing an easy-to-use platform for transcribing mixed-language sentences, the website enhances the user experience and makes it easier for users to work with content in multiple languages. Moreover, the use of a machine learning model and user input to transcribe sentences allows for scalability and the ability to handle large volumes of content in multiple languages. Finally, leveraging an existing machine learning model and relying on user input make the development and maintenance of the website cost-effective. Overall, the language transcription system offers a powerful solution for working with mixed-language content and can be a valuable addition to any website.