

TOMASZ GACEK

(Jagiellonian University, Cracow, Poland)

ORCID: 0000-0002-8111-2797 

Reproducing Sections of a Research in the Field of Indo-Iranian Linguistics Using ChatGPT AI Model – an Experiment

Abstract

In recent years, practical applications of AI models have become globally accessible and increasingly commonplace. These models have begun to be integrated into scientific research, marking the initial stages of their use in this domain. Among such tools, ChatGPT, recognized as an “AI language model”, is particularly intriguing for examining how it handles linguistic problems. This paper seeks to replicate selected parts of a research previously conducted using traditional methods and evaluate the results produced by the model. The research focused on analyzing the vocabulary of Persian and Arabic origin in Hindi songs from Bollywood movies. The goal of this paper is to determine how effectively an AI model can perform specific tasks originally carried out by a human researcher in the original study.

Keywords: AI, ChatGPT, Hindi, Persian, Arabic, vocabulary, Bollywood songs



Introduction

1. AI models in linguistic studies

In recent years, the widespread availability of practical applications of AI models has become the new normal. These models are already employed in various aspects of everyday life, both personal and professional, though the practical value of their outcomes can vary. Among the numerous AI models available today, this paper focuses on the first widely adopted model, ChatGPT, precisely its free variant ChatGPT-3.5.

ChatGPT is described as an “AI language model developed by OpenAI, based on the GPT (Generative Pre-trained Transformer) architecture.” (ChatGPT answer to: “please describe what chat gpt is”, ChatGPT conversation, 2024-02-16). Therefore, it is particularly appealing to evaluate its performance on practical problems within the field of linguistic studies.

There is, of course, an ongoing debate regarding the acceptability of using AI tools in research and scientific writing. The concerns raised can be categorised into at least three different areas: 1. ethical issues revolving around the question of plagiarism; 2. academic integrity (including problems such as fictitious references, and 3. various forms of bias.¹ The third can be understood more broadly as a question of the reliability of results.

In the author’s view, some ethical concerns can be addressed by establishing proper citation standards for references for data generated by ChatGPT and similar tools. In fact, work on this issue is already underway.² Thus, the key question is not whether using ChatGPT and other AI models is ethical, but rather whether their use is meaningful and can contribute new value to scientific research. If AI tools like ChatGPT can enhance research outcomes, it may be more appropriate to ask whether it is ethical *not* to use them. As Teel, Wang, and Lund suggest, “Rather than viewing ChatGPT and similar technologies as threats, academics should embrace the challenges they present and use them as opportunities to broaden and deepen their understanding of ethical and responsible boundaries”.³

2. Hindi and Persian languages

Hindi is one of the New Indo-Aryan languages, spoken in a large part of northern India (the states of Uttar Pradesh, Bihar, Madhya Pradesh, Rajasthan, Haryana, Himachal Pradesh, and the Union Territory of Delhi).⁴ Its literary variant, Standard Modern Hindi,

¹ Zoë Abbie Teel et al. ‘ChatGPT conundrums: Probing plagiarism and parroting problems in higher education practices’, *College & Research Libraries News* 84/6 (2023), pp. 205–206.

² Timothy McAadoo, ‘APA Style. How to cite ChatGPT’, *California State University Channel Islands website*, Viewed 17 October 2024, <<https://www.csuci.edu/wmc/documents/apa-citingchatgpt.pdf>>.

³ Teel et al., ‘ChatGPT conundrums’, p. 207.

⁴ Colin Masica, *The Indo-Aryan Languages*, Cambridge 2001, pp. 8–9.

continues the traditions of various earlier forms, particularly Braj and Awadhi.⁵ Numerous variants of Hindi exist, primarily regional ones.⁶

Hindi underwent a period of intense lexical borrowing from Persian, which infused it with words of both Iranian and Arabic origin, with Persian serving as the intermediary for the latter. As a result, the “Vocabulary of Persian and Arabic origin constitutes a large, extremely important and unalienable part of Hindustani lexical resources.”⁷

There is a subtle and elusive, not clearly defined boundary between Hindi and Urdu. There is not even a simple answer as to whether the two ethnolects should be treated as distinct languages or as varieties of one language.⁸ Investigating this problem lies far beyond the scope of the present study. However, let us note just one thing: Hindi is traditionally written in the Devanāgarī script, while Urdu uses a modified version of the Arabic alphabet.

New Persian belongs to the southwestern branch of the Iranian languages⁹. Iranian languages are the western group of the Indo-Iranian side of the Indo-European language family.¹⁰ (Windfuhr, 2009, 1). This means Persian and Hindi are relatively closely related.

The Muslim conquest marked the beginning of the epoch of intense borrowing from Arabic into Persian. It is estimated that by the 10th century, around 1/3 of New Persian vocabulary was of Arabic origin, reaching nearly half of the lexicon two centuries later.¹¹ Persian is written using a modified variant of the Arabic alphabet (distinct from the one used for Urdu).

3. The present study

In this paper, we aim to take a small step toward answering the question of whether ChatGPT and similar tools can be useful in linguistic research, by reproducing parts of a previously conducted study¹². We will attempt to achieve this goal by asking questions to ChatGPT and evaluating whether, and to what extent, the results are consistent with the original research. We will also try to improve the answers by asking follow-up questions, providing sample data to the model, and so on.

⁵ Ibidem.

⁶ Ibidem, p. 9ff.

⁷ Agnieszka Kuczkiewicz-Fraś, *Perso-Arabic Loanwords in Hindustani. Part ii, Linguistic Study*, Kraków 2012, p. 1.

⁸ Christine Everaert, *Tracing the Boundaries between Hindi and Urdu. Lost and Added in Translation between 20th Century Short Stories*, Leiden 2010, pp. 1ff.

⁹ Gernot Windfuhr, ‘Introduction to the Iranian Languages’, in: Gernot Windfuhr (ed.), *The Iranian Languages*, London-New York 2009, p. 2.

¹⁰ Ibidem, p. 1.

¹¹ Ali Ašraf Sādeqi, ‘Arabic Language. i. Arabic elements in Persian’, *Encyclopedia Iranica online*, Viewed 17 October 2024, <<https://iranicaonline.org/articles/arabic-i>>.

¹² Tomasz Gacek, ‘Some Remarks on the Perso-Arabic Nominal Lexica in Bollywood Movie Song Lyrics’, *Acta Asiatica Varsoviensia* 36 (2023), pp. 39-67, <https://doi.org/10.60018/AcAsVa.njyu2866>.

The original research was conducted according to the following algorithm:

- a. Isolating nominal lexemes from the analyzed texts (see below);
- b. Reconstructing lemmata for the nominal forms found;
- c. Tagging the forms based on their origin as either native (tadbhava, tatsama, borrowed from other Indo-Aryan languages) or foreign (loanwords from non-Indo-Aryan languages, foreign words).
- d. Conducting statistical analysis.

The original research was conducted on a corpus of lyrics from 93 Bollywood movie songs spanning the period between 1940 and 2020, the details of which are described in the relevant publication.¹³ In this corpus, 1074 nominal forms have been identified (1909 including repetitions), which can be reduced to 956 distinct lemmata. Due to restrictions on the length of user input to ChatGPT (so called quota), only samples of this material (presented in specific questions to the model, below and in the appendix) were used in the current study.

Both the original writing systems (especially Devanāgarī) and transcription were used when communicating with ChatGPT. To avoid confusion, both the data input to the model and its replies will be presented exactly as they were provided or generated during the experiment, even though it is common practice, especially in Indological studies, to substitute original writing systems with transcription. In this particular case, the romanization of actual queries and model replies would, in the author's opinion, distort the representation of the conducted research, especially considering that existing transcription systems do not capture some minor details of Devanāgarī orthography as used in Hindi.

1. Identifying nominal lexemes

We will begin with an attempt to indicate nominal lexemes in the first two of the songs analyzed in the original research: *Divānā hū̃, divānā hū̃*¹⁴ from the 1940s movie *Zindagī* and *Cāne Zor Garam Bābū*¹⁵ from the film *Bandhan* from the same year. Let us start with the first text¹⁶.

¹³ Ibidem.

¹⁴ The text was retrieved from: Viewed 17 October 2024, <https://www.hindilyrics4u.com/song/deewana_hoon_deewana_hoon_raahat_se_main_begana_hoon.htm> and verified against the recording.

¹⁵ The text was retrieved from: Viewed 17 October 2024, <https://bharatdiscovery.org/india/%E0%A4%9A%E0%A4%A8%E0%A5%87_%E0%A4%9C%E0%A5%8B%E0%A4%B0_%E0%A4%97%E0%A4%B0%E0%A4%AE_%E0%A4%AC%E0%A4%BE%E0%A4%AC%E0%A5%82> and verified against the recording.

¹⁶ For the sake of clarity, all the interactions with the model will be reproduced in tables, with “U” marking the user's input and “C” used for the model's replies.

Table 1. ChatGPT conversation, 2024-02-16

U	Please, indicate nouns in the following Hindi text: हे दीवाना हूँ दीवाना हूँ राहत से मैं बेगाना हूँ राहत से मैं बेगाना हूँ दीवाना हूँ दीवाना हूँ राहत से मैं बेगाना हूँ राहत से मैं बेगाना हूँ दिल को कैसे बहलाऊँ मैं हाँ ये ज़ख्म किसे दिखलाऊँ मैं दिल को कैसे बहलाऊँ मैं हाँ ये ज़ख्म किसे दिखलाऊँ मैं आंसू सुन सुन कर हँसते हैं क्या दर्द भरा अफ़साना हूँ हैं क्या दर्द भरा अफ़साना हूँ आयी न बहार ख़िज़ाँ ही सही आयी न बहार ख़िज़ाँ ही सही उजड़े दिल में अरमान ही सही उजड़े दिल में अरमान ही सही हैरान हूँ की आखिर क्या हूँ मैं हैरान हूँ की आखिर क्या हूँ मैं आबाद हूँ या विराना हूँ घायल हूँ तुम से दूर हूँ मैं घायल हूँ तुम से दूर हूँ मैं पर फूँक चुका मजबूर हूँ मैं पर फूँक चुका मजबूर हूँ मैं ए समां बलाएं ले आ कर ए समां बलाएं ले आ कर आखिर मैं भी परवाना हूँ आखिर मैं भी परवाना हूँ आखिर मैं भी परवाना हूँ
C	Nouns in the provided Hindi text: दीवाना (deewana) – madman राहत (raahat) – relief दिल (dil) – heart ज़ख्म (zakhm) – wound अफ़साना (afsana) – story ख़िज़ाँ (khizaan) – autumn अरमान (arman) – desire हैरान (hairaan) – surprised विराना (viraana) – desolation परवाना (parwaana) – moth

We will discuss the poor transcription used by the model in its reply later; for now, we will focus on identifying nouns in the text. The forms राहत *rāhat*, दिल *dil*, ज़ख़म *zaxm*, अफ़साना *afsānā*, ख़िज़ाँ *xizā*, अरमान *armān* and परवाना *parvānā* raise no doubts. However, we can be indeed surprised to see हैरान *hairān* “surprised” classified as a noun, as it is obviously an adjective.¹⁷ Moreover, it is used in the text as an adjectival predicate, with no indication of possible nominalization. विराना *virānā* is disputable. It could be an adjective, meaning “estranged” or “alienated,”¹⁸ or a noun meaning “a desolate place”.¹⁹ However, as it occurs here paired with आबाद *ābād*, which is exclusively an adjective,²⁰ it seems logical to classify it as an adjective as well, and exclude it from the list.

दीवाना *divānā* “crazy, mad, insane,” is unanimously classified as an adjective in dictionaries,²¹ even though it is imaginable to assume that in some situations it could acquire the nominal meaning “a mad man”. However, no indication of such a process is observed in the text.

It would certainly be more natural to classify the form बेगाना *begānā* as a noun here. While it can function both as an adjective “(...) strange, foreign, alien (...)” and as a noun “(...) stranger, foreigner, outsider (...)”,²² it seems to fit the context better in the latter role.

We also do not find in the list some other words we would likely expect the model to identify: आँसू (better: आँसू) *āsū* “tear”, दर्द *dard* “pain”, बहार *bahār* “spring” and फूँक *phūmk* “(...) blowing, (...) breath (...)” as they are all nouns.²³ The form आखिर *āxir* could also be interpreted as a noun,²⁴ especially as it is followed by a postposition in the text.

Let us ask the model if it is confident that all the forms it had singled out are indeed nouns.

Table 2. ChatGPT conversation, 2024-02-16

U	Are you sure all the words in the list you provided are nouns?
C	You're correct; I apologize for the oversight. Not all the words in the list provided are nouns. Let me correct that: Nouns in the provided Hindi text: (...) <i>And here follows exactly the same list, as before (sic!)</i> ²⁵

¹⁷ Ronald Stuart McGregor, *Oxford Hindi-English Dictionary*, Oxford-Delhi 1993, p. 1080; Hardev Bahri, *Learners' Hindi-English dictionary*, Delhi 1989, p. 706.

¹⁸ See Śyāmsundar Dās, *Hindī śabdsāgar*, 11 vols. 2nd edition, Benares 1965–1975, p. 4522.

¹⁹ McGregor, *Oxford Hindi-English Dictionary*, p. 933.

²⁰ Ibidem, p. 89; Mahendra Caturvedi, *A practical Hindi-English dictionary*, Delhi 1970, p. 68; Śyāmsundar Dās, *Hindī śabdsāgar*, p. 454.

²¹ McGregor, *Oxford Hindi-English Dictionary*, p. 500; Caturvedi, *A Practical Hindi-English Dictionary*, p. 319; Bahri, *Learners' Hindi-English dictionary*, p. 311; Dās, *Hindī śabdsāgar*, p. 2302.

²² McGregor, *Oxford Hindi-English Dictionary*, p. 745.

²³ See respectively McGregor, *Oxford Hindi-English Dictionary*, pp. 77, 482, 716 and 686.

²⁴ Ibidem, p. 79.

²⁵ The italic font in the “C” section of tables will be used to summarize, paraphrase, or comment on ChatGPT output.

Only when we explicitly object to this (“But you have given exactly the same list now!”), ChatGPT will finally drops at least *hairān* from the set. (ChatGPT conversation, 2024-02-16).

Now, what about the overlooked forms?

Table 3. ChatGPT conversation, 2024-02-16

U	Are you sure आँसू, दर्द, बहार, आँखि and फूँक are not nouns in this text?
C	You’re correct; I appreciate your patience. Here’s the corrected list with additional nouns: Nouns in the provided Hindi text: दीवाना (deewana) – madman राहत (raahat) – relief दिल (dil) – heart ज़ख्म (zakhm) – wound अफ़साना (afsana) – story ख़िज़ान (khizaan) – autumn अरमान (arman) – desire हैरान (hairaan) – surprised विराना (viraana) – desolation परवाना (parwaana) – moth आँसू (aansoo) – tears दर्द (dard) – pain बहार (bahaar) – spring आँखि (aankh) – eye फूँक (phoonk) – blow Thank you for bringing it to my attention.

We can see some improvement, but unfortunately, हैरान *hairān* is back on the list. This is an important point, as it shows that we are unable to perform a real supplementary training of the model. All the improvements we achieve do not affect the model permanently. If we want the model to behave according to our rules, we will always have to refer to them in our prompts.

Let us now see how the model copes with the other text.

Table 4. ChatGPT conversation, 2024-02-17

U	Please, find nouns in the following Hindi text: चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम मेरे चने हैं चटपटे भैया और बड़े लासानी और कैसे चाव से खाते देखो और रमजानी और चुनू मुनू की जबान भी हो गयी पानी पानी और कहें कबीर सुनो भई साथी सुनो गुरू की बानी चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम पढ़ें मदरसे क्राज़ी बन तो चंद दिनों का ठाट और पढ़ लिख कर सब चल दोगे तुम अपनी अपनी बाट फिर कोई तुममें अफसर होगा कोई गवरनर लाट तब मैं आऊंगा दफ़्तर घूमने लिये चने की चाट चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम मेरा चना बना है आला इसमें डाला गरम मसाला चखते जाना जी तुम लाला कहता हूं मैं दिल्ली वाला इसका स्वाद है बड़ा निराला चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम आई चने की बहार खाते जाना जी सरकार मेरे चने जायकेदार अगर तुमको ना होय एतबार मैं कहता हूं ललकार चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम देख लो मेरा ये दरबार जहां पर खड़े सिलसिलेवार रियासत भर के सरदार एक से एक सभी हुसियार ये देखो मेरे सूबेदार ये देखो मेरे तहसीलदार ये हैं मेरे थानेदार और ये बड़े सिपहसालार वानर सेना के सरदार चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम चने जोर गरम बाबू मैं लाया मजेदार चने जोर गरम
---	---

Table 4. cont.

C	Here are the nouns in the Hindi text you provided: चने (gram) जोर (strength) बाबू (term of endearment) मजेदार (delicious) भैया (brother) लासानी (spicy) चाव (taste) रमजानी (enjoyment) चुनू (a name) मुनू (a name) जबान (tongue) मदरसे (school) क्राज़ी (judge) दिनों (days) ठाट (position) गवरनर (governor) दफ़्तर (office) स्वाद (taste) निराला (unique) आला (superior) मसाला (spice mix) जी (term of respect) दिल्ली (Delhi) बहार (spring) सरकार (government) जायकेदार (excellent) एतबार (trust) ललकार (call) दरबार (royal court) सिलसिलेवार (chain-wearer) रियासत (kingdom) सरदार (leader) हुसियार (clever) तहसीलदार (revenue officer) थानेदार (police officer) सिपहसालार (commander) वानर (monkey) सेना (army)
---	--

Several nouns in the text have once again not been detected by the model: पानी *pānī* “water (...)”,²⁶ साधो *sādhō*: voc. pl. of *sādh* “good, virtuous (...)”; an ascetic”,²⁷ गुरू *gurū* var. of गुरु *guru* “heavy, (...) venerable (...)”; a spiritual teacher (...)”,²⁸ वानी *bānī* “sound, voice, (...) teachings (...)”,²⁹ बाट *bāṭ* “way, road, path (...)”,³⁰ अफ़सर *afsar* “an officer, official (...)”,³¹ लाट *lāṭ* “a lord (...)”,³² चाट *cāṭ* “licking, tasting; taste (...)”,³³ लाला *lālā* “(...) a term of respectful address (...)”,³⁴ स्वाद *svād* “a taste, flavour (...)”,³⁵ and सूबेदार *sūbedār* “a governor (...)”.³⁶

On the other hand, the words लासानी *lā-sānī* “without a second; unequalled”,³⁷ हुसियार *husiyār* “wise, clever”,³⁸ and आला *ālā* “(...) most high, exalted (...)”³⁹ are adjectives and should, in fact, be excluded based on the meaning provided by the model itself. निराला *nirālā* “separate, lonely (...) a lonely place (...)”⁴⁰ may be both a noun and an adjective, but only the adjectival sense fits the context here.

Whether the honorific expression जी may be classified as a noun is another problem. Let us set it aside for now.

Let us try to find out why the model classified *lā-sānī*, *nirālā* and *ālā* as nouns.

Table 5. ChatGPT conversation, 2024-02-17

U	Explain, why did you classify the words लासानी, निराला and आला as nouns
C	I must apologize for the mistake. The words “लासानी,” “निराला,” and “आला” are not nouns; they are adjectives. Adjectives in Hindi describe or modify nouns by providing additional information about their qualities, characteristics, or attributes. Here’s the correct classification: लासानी: This word is an adjective meaning “spicy.” निराला: This word is an adjective meaning “unique” or “unusual.” आला: This word is an adjective meaning “superior” or “excellent.” Thank you for bringing this to my attention, and I appreciate the opportunity to correct the mistake. If you have any further questions or need clarification, please feel free to ask.

²⁶ McGregor, *Oxford Hindi-English Dictionary*, p. 623.

²⁷ Ibidem, p. 1004.

²⁸ Ibidem, p. 271.

²⁹ Ibidem, p. 724.

³⁰ Ibidem, p. 721.

³¹ Ibidem, p. 45.

³² Ibidem, p. 891.

³³ Ibidem, p. 310.

³⁴ Ibidem, p. 893.

³⁵ Ibidem, p. 1051.

³⁶ Caturvedi, *A Practical Hindi-English Dictionary*, p. 829.

³⁷ McGregor, *Oxford Hindi-English Dictionary*, p. 894.

³⁸ See Dās, *Hindī śabdsāgar*, p. 5538.

³⁹ McGregor, *Oxford Hindi-English Dictionary*, p. 94.

⁴⁰ Ibidem, p. 567.

Some experience with ChatGPT may suggest that the model has a tendency – a specific fault in its algorithm – to always agree with the user’s objections. This should be investigated.

Table 6. ChatGPT conversation, 2024-02-17

U	are you sure एतबार (trust), ललकार (call) and दरबार (royal court) are nouns?
C	You’re correct; I made a mistake in classifying these words. Let me correct that: एतबार (etbaar): This word can function as both a noun and a verb, depending on the context. As a noun, it can mean “trust” or “confidence.” So, it is indeed a noun. ललकार (lalkaar): This word is a noun and means “shout” or “call.” So, it is classified correctly as a noun. दरबार (darbaar): This word primarily functions as a noun and means “royal court” or “king’s court.” So, it is also correctly classified as a noun. Thank you for pointing out the oversight. If you have any further questions or need clarification, feel free to ask!

This is interesting. The model initially politely admits a mistake but then essentially stands by its opinion, which is quite fortunate in this particular case. Therefore, our suspicion was unfounded.

Let us try a different approach now. Perhaps we should not use terms like “nouns” or “adjectives”? So the next query was: “User: Please, find in the following Hindi text all the words, which are names of persons, objects, abstract ideas, natural phenomena. Don’t include the words that are mainly used with the attributive function. Use only exactly the words in the form, in which they are used in the text. (...)” However, the results were very similar to those achieved above. Most nouns were correctly identified, but some were omitted. Additionally, several adjectives were included. Furthermore, contrary to the explicit guidelines, the model returned forms that are indeed nouns but do not appear in the text! They are just word-formationally related to the attested forms, such as *hairānī* < *hairān*. (ChatGPT conversation, 2024-02-17)

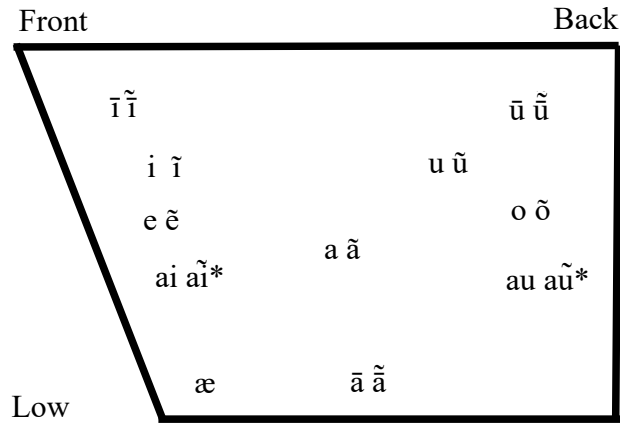
The author of the present paper must admit that he has no idea how to improve the model’s results in this area. Let us assume that the model is far from perfect in classifying parts of speech (at least in Hindi) and move on to the next task.

2. Providing transcription for selected Hindi forms

Transcription is often preferred in linguistic research, even when the language in question is traditionally written with a precise and efficient writing system. At least two factors encourage to use it: first, probably every naturally developed writing system has at least some ambiguities, regardless of how well designed it is, especially if it has a long history.⁴¹ Second, transcription allows researchers not being specialists in this particular language to access and understand the information provided in a work.

Before we can start persuading ChatGPT to use a particular standard of transcription, it would be useful to present a phoneme inventory of modern Hindi. The vowel system consists of 21 sounds⁴²:

Table 7. Hindi Vowels



* /ai/ and /au/ are pronounced monophthongally in most dialects (as ε and ɔ), but we will retain the traditional transcription, compatible with that of McGregor’s dictionary.

While there are minor differences in the classification of specific sounds, in general, the consonant system of Hindi may be presented as follows:

⁴¹ In traditional Hindi orthography we may put forward the problem of not using virāma and ligatures to mark the lack of a vowel, especially in words of foreign origin, e.g. अफसर *afsar* (not *afasara*) instead of अफसर̣. Another problem is the inconsistent marking of nasal elements.

⁴² The table has been compiled on the basis of Manjari Ohala, *Aspects of Hindi Phonology*, Delhi-Varanasi-Patna 1983, p. 1; modified using: Manjari Ohala, ‘Hindi’, in: *Handbook of the International Phonetic Association. A Guide to Use of the International Phonetic Alphabet*, Cambridge 1999, p. 102. However, for the sake of consistency, the transcription used in McGregor, *Oxford Hindi-English Dictionary* is used, instead of the symbols used by Ohala (i, ɪ, e, ε, æ – for front vowels; ə, a – central vowels; u, ʊ, o, ɔ – back vowels; nasalization marked in the same way).

We are following Ohala's classification here.⁴³ However, as with the vowel system, we will use the transcription characters introduced by McGregor in his dictionary.⁴⁴ Additionally, we include other potential consonants of Persian origin apart from *f* (all in parentheses). The reason for that is that the pronunciation of Perso-Arabic loanwords in Bollywood songs is sometimes more more faithful, approaching Urdu standards.⁴⁵

Let us now analyze in detail the method of transcription ChatGPT uses. Then we will attempt to persuade the model to use a coherent and reliable system instead. Since we were unable to get the AI to correctly identify nouns in the analyzed texts, we will confront it with a manually compiled list.

Table 9. ChatGPT conversation, 2024-02-17

U	Provide transcription for the following Hindi forms: राहत, बेगाना, दिल, ज़ख्म, आँसू, दर्द, अफ़साना, बहार, खिज़ाँ, अरमान, आखिर, फूँक, समां, परवाना, चने, बाबू, भैया, चाव, ज़बान, पानी, भई, साधो, गुरू, बानी, मदरसे, क़ाज़ी, दिनों, ठाट, बाट, अफ़सर, गवरनर, लाट, दफ़्तर, चाट, मसाला, लाला, स्वाद, बहार, सरकार, एतबार, ललकार, दरबार, रियासत, सरदार, सूबेदार, तहसीलदार, थानेदार, सिपहसालार, वानर, सेना ⁴⁶
C	Here are the transcriptions for the provided Hindi forms: राहत – raahat, बेगाना – begaana, दिल – dil, ज़ख्म – zakhm, आँसू – aansoo, दर्द – dard, अफ़साना – afsaana, बहार – bahaar, खिज़ाँ – khizaaan, अरमान – armaan, आखिर – aakhir, फूँक – phoonk, समां – samaan, परवाना – parwaana, चने – chane, बाबू – baabu, भैया – bhaiyaa, चाव – chaav, ज़बान – zubaan, पानी – paani, भई – bhai, साधो – saadho, गुरू – guru, बानी – baani, मदरसे – madrase, क़ाज़ी – qaazi, दिनों – dino, ठाट – thaath, बाट – baat, अफ़सर – afsar, गवरनर – governor, लाट – laaT, दफ़्तर – daftar, चाट – chaat, मसाला – masala, लाला – laala, स्वाद – swaad, बहार – bahaar, सरकार – sarkaar, एतबार – etbaar, ललकार – lalkaar, दरबार – darbaar, रियासत – riyaaasat, सरदार – sardaar, सूबेदार – subedaar, तहसीलदार – tahsildaar, थानेदार – thanedaar, सिपहसालार – sipahsaalaar, वानर – vaanar, सेना – sena These transcriptions reflect the pronunciation of the words in Hindi. If you have any further questions or need clarification, feel free to ask!

⁴³ The table is based on Manjari Ohala, 'Hindi', *Handbook of the International Phonetic Association. A Guide to Use of the International Phonetic Alphabet*, Cambridge 1999, p. 100; However, as in the case of the vowel system, we are going to use the transcription characters as introduced by McGregor in his dictionary (McGregor, *Oxford Hindi-English Dictionary*).

⁴⁴ McGregor, *Oxford Hindi-English Dictionary*.

⁴⁵ Let us recall the anecdote how Lata Mangeshkar needed Urdu lessons to perfect her pronunciation so that she could pursue her career as a Bollywood singer (Dilip Kumar, *The Substance and the Shadow. An Autobiography*, New Delhi 2014, p. 369).

⁴⁶ Both the query and the answer originally contained lists with each item in a separate row. They were re-formatted to make the text more compact.

As we can see, the transcription provided is far from what we would expect in a scientific paper. Many phonologically relevant differences are neglected (e.g. $t : \bar{t}$, $d : \bar{d}$, $u : \bar{u}$). Additionally, some differences that may not be universal but are important for Bollywood songs – where there is often a greater effort to pronounce Persian or Arabic sounds more clearly than in everyday speech (e.g., $kh : x$, see above) – are also neglected. The transcription provided is inconsistent as oo is rendered either as u or as oo , and length is marked for $a : \bar{a}$, but not for $u : \bar{u}$, among other issues. Let us see if it is possible to persuade Chat GPT to use a more precise system. First, let us examine whether asking GPT to transcribe the whole text in its original, integral form yields better results:

Table 10. ChatGPT conversation, 2024-08-15

U	Provide transcription in Latin alphabet for the following Hindi text: (...) <i>the text as given in Table 1</i>
C	He deewana hoon deewana hoon Rahat se main begaana hoon Rahat se main begaana hoon Deewana hoon deewana hoon Rahat se main begaana hoon Rahat se main begaana hoon Dil ko kaise behlaoon main Haan ye zakhm kise dikhlaaoon main Dil ko kaise behlaoon main Haan ye zakhm kise dikhlaaoon main Aansoo sun sun kar hanste hain Kya dard bhara afsaana hoon Hain kya dard bhara afsaana hoon Aayi na bahaar khizaan hi sahi Aayi na bahaar khizaan hi sahi Ujre dil mein armaan hi sahi Ujre dil mein armaan hi sahi Hairaan hoon ki aakhir kya hoon main Hairaan hoon ki aakhir kya hoon main Aabaad hoon ya viraana hoon Ghayal hoon tum se door hoon main Ghayal hoon tum se door hoon main Par phoonk chuka majboor hoon main Par phoonk chuka majboor hoon main Ae samaan balaayein le aa kar Ae samaan balaayein le aa kar Aakhir main bhi parwaana hoon Aakhir main bhi parwaana hoon Aakhir main bhi parwaana hoon

As we can see, the transcription is generally more consistent but otherwise no better than in Table 9. Let us explicitly instruct the model to change the transcription in the desired way.

Table 11. ChatGPT conversation, 2024-02-17

U	Provide accurate phonological transcription for the forms above, i.e. one that would take into consideration all the existing phonological differences
C	Sure, here's the accurate phonological transcription for the forms provided: राहत – /ra:ɦət/, बेगाना – /be:ga:na:/, दिल – /dil/, ज़ख़्म – /zəxm/, आँसू – /ā:su/, दर्द – /dərd/, अफ़साना – /əfsa:na:/, बहार – /bəɦa:r/, खिज़ाँ – /xi'za:n/, अरमान – /ərma:n/, आखिर – /a:xir/, फूँक – /pʰu:ŋk/, समां – /səma:n/, परवाना – /pərva:na:/, चने – /tʃəne/, बाबू – /ba:bu/, भैया – /bʰəja:/, चाव – /tʃa:v/, ज़बान – /zəba:n/, पानी – /pa:ni/, भई – /bʰəi/, साधो – /sa:ɖʰo/, गुरू – /guru/, बानी – /ba:ni/, मदरसे – /məɖrase/, क़ाज़ी – /qa:zi/, दिनों – /dino:/, ठाट – /tʰa:t/, बाट – /ba:t/, अफ़सर – /əfsər/, गवरनर – /gəvərnər/, लाट – /la:t/, दफ़्तर – /dəftər/, चाट – /tʃa:t/, मसाला – /masa:la:/, लाला – /la:la:/, स्वाद – /sva:ɖ/, बहार – /bəɦa:r/, सरकार – /sərka:r/, एतबार – /etba:r/, ललकार – /ləlka:r/, दरबार – /dərba:r/, रियासत – /rija:sat/, सरदार – /sərda:r/, सूबेदार – /su:bəda:r/, तहसीलदार – /təɦsi:lɖa:r/, थानेदार – /tʰa:ne:ɖa:r/, सिपहसालार – /sipəɦsa:la:r/, वानर – /va:nər/, सेना – /se:na/

The result is quite astonishing. Even though the system applied is very different from the one we have chosen, it is actually quite good. It is based on the IPA standard and it is much more coherent than the model's initial attempts. Vowel length is generally marked consistently, retroflex consonants are rendered differently from dental ones, the difference between /x/ and /kʰ/ is shown, vowel nasalization is indicated rather by diacritics (e.g. *ā:kʰō:*, *ā:su*) than by adding the character *n* after a vowel. This looks as if the model indeed understood what is meant by a request for more accurate transcription.

However, this new transcription is not flawless. We can identify two types of problems: a) those resulting from imperfect application of the system (“errors against the system”); b) problems inherent to the system itself (“errors of the system”). Regarding the category a), examples include the incorrect transcription of the initial retroflex consonant in *tʰa:t* (instead of *tʰa:ɖ*), rendering गुरू as if there were two identical vowels (*guru* instead of *gurū*).

As far as the “errors of the system” are concerned, it is particularly noteworthy that in some cases the model went a bit too far, leaning more towards phonetic transcription rather than phonological one. For example, the difference between *h* and *ɦ* in the forms *təɦsi:lɖa:r* and *bəɦa:r* is not phonologically relevant. Similarly, the difference in aspiration marking in voiced and voiceless consonants is redundant, too (*bʰəja:* vs. *pʰu:ŋk*).

In this situation, let us ask the model about the difference between phonetic and phonological transcription.

Table 12. ChatGPT conversation, 2024-02-17

U	What is the difference between the phonological and phonetic transcription?
C	Phonological and phonetic transcriptions are both systems used to represent spoken language, but they serve different purposes and focus on different aspects of speech. Phonological Transcription: Phonological transcription represents the abstract underlying sounds of a language, often referred to as phonemes. It is concerned with the distinctive sounds (phonemes) of a language and the rules governing their distribution and behavior in words. Phonological transcription ignores variations in pronunciation that do not affect the meaning of words. For example, in English, the sounds [p] in “pin” and [ph] in “spin” are considered allophones of the same phoneme /p/. It is typically enclosed in slashes (/ /) to indicate that it represents phonemes rather than actual sounds. Phonological transcription is useful for studying the structure and patterns of sounds in a language. Phonetic Transcription: Phonetic transcription represents the actual sounds produced in speech, including variations in pronunciation such as accent, intonation, and speech disorders. It captures the precise articulatory details of speech sounds, including variations in vowel quality, consonant articulation, and prosody. Phonetic transcription uses symbols from the International Phonetic Alphabet (IPA) to represent specific speech sounds. Each symbol corresponds to a unique articulatory or acoustic feature. Phonetic transcription is often enclosed in square brackets ([]) to indicate that it represents actual speech sounds. Phonetic transcription is useful for detailed analysis of individual speech sounds and for accurately representing spoken language in linguistic research, language teaching, and speech therapy. In summary, phonological transcription focuses on the abstract phonemic structure of language, while phonetic transcription captures the concrete articulatory and acoustic details of speech.

If ChatGPT were taking an Introduction to Linguistics exam, this answer would certainly bring it closer to a passing grade. However, just as with living students, storing the necessary information and using it properly are two different things. Let us see if the model can use it.

Table 13. ChatGPT conversation, 2024-02-17

U	Taking into consideration the difference between phonological and phonetical transcription, can you provide transcription of the forms given above that would be strictly phonological. I want you to omit all the pronunciation features that are not phonologically relevant.
C	ChatGPT: Certainly, here are the forms transcribed with a focus on phonological features: (...) <i>Here comes exactly the same list as presented in table 11.</i>

Quite disappointingly, after agreeing so politely, ChatGPT provided exactly the same transcription as it had produced in response to the previous query. Perhaps it is time to inform the model about our specific concerns. Unfortunately, even explicitly pointing out the deficiencies (“User: u and u: are distinct phonemes in Hindi and the difference should always be marked.”, ChatGPT conversation, 2024-02-17) results in no improvement.

Let us try another approach. As we have seen, one of the persistent problems is issue of differentiation between *u* and *u:*. Maybe we can force it to distinguish between them by providing more words containing both vowels.

Table 14. ChatGPT conversation, 2024-02-18

U	Provide phonological transcription for the following Hindi forms: शुरू, खुशबू, राहत, बेगाना, दिल, ज़ख्म, आँसू, दर्द, अफ़साना, बहार, खिज़ाँ, अरमान, आखिर, फूँक, समाँ, परवाना, चने, बाबू, भैया, चाव, ज़बान, पानी, भई, साधो, गुरू, बानी, मदरसे, क़ाज़ी, दिनों, ठाट, बाट, अफ़सर, गवरनर, लाट, दफ़्तर, चाट, मसाला, लाला, स्वाद, बहार, सरकार, एतबार, ललकार, दरबार, रियासत, सरदार, सूबेदार, तहसीलदार, थानेदार, सिपहसालार, वानर, सेना
C	Here are the phonological transcriptions for the provided Hindi words: शुरू – /ʃu:ru:/, खुशबू – /xʊʃbu:/, राहत – /ra:ɦət/, बेगाना – /be:ga:na:/, दिल – /dɪl/, ज़ख्म – /zəxm/, आँसू – /āsu:/, दर्द – /dərd/, अफ़साना – /əfsa:na:/, बहार – /bəɦa:r/, खिज़ाँ – /xɪza:n/, अरमान – /ərna:n/, आखिर – /a:xɪr/, फूँक – /fu:ŋk/, समाँ – /səma:n/, परवाना – /pərva:na:/, चने – /tʃəne/, बाबू – /ba:bu:/, भैया – /bʰaɪja:/, चाव – /tʃa:v/, ज़बान – /zəba:n/, पानी – /pa:ni:/, भई – /bʰəi:/, साधो – /sa:ðʰo:/, गुरू – /guru:/, बानी – /ba:ni:/, मदरसे – /mədərse/, क़ाज़ी – /qɑ:zi:/, दिनों – /dino:/, ठाट – /tʰa:t/, बाट – /ba:t/, अफ़सर – /əfsər/, गवरनर – /gəvərnər/, लाट – /la:t/, दफ़्तर – /dəftər/, चाट – /tʃa:t/, मसाला – /məsa:la:/, लाला – /la:la:/, स्वाद – /sva:ɖ/, बहार – /bəɦa:r/, सरकार – /sərkɑ:r/, एतबार – /etba:r/, ललकार – /ləlkɑ:r/, दरबार – /dərbɑ:r/, रियासत – /rija:sət/, सरदार – /sərdɑ:r/, सूबेदार – /su:be:da:r/, तहसीलदार – /təɦsi:lda:r/, थानेदार – /tʰa:ne:da:r/, सिपहसालार – /sipəɦsa:la:r/, वानर – /va:nər/, सेना – /se:na:/

We can see some improvement now. Although the first vowel of शुरू *śuru* was mistakenly classified as a long one (**fu:ru:*), in general, the model performed better now. However, this is just one of the problems. All the other issues remain, especially those arising from the fact that, contrary to our expectations and its own declaration, the model produces a transcription that is more phonetic than phonological (or, to be precise, a peculiar hybrid system). The mechanism we have used for *u* vs. *u:* will not work in this case. Moreover, if we pay close attention, we will notice that some new, previously unseen mistakes have appeared in the last variant. For example the model uses the spirant *f* instead of the aspirated *p^h* (**fu:ŋk* instead of *p^hu:ŋk*). Aspiration in *tha:t* is now marked with the normal letter *h*, whereas in all other forms, it is superscripted. Let us try to eliminate at least some of these problems.

Table 15. ChatGPT conversation, 2024-02-18

U	Repeat the last list, but use always /a/ for /ə/, /u/ for /ʊ/, /i/ for /ɪ/, /d/ for /d̪/, /t/ for /t̪/, /v/ for /v/
C	शुरू – /fu:ru:/, खुशबू – /xʊʃbu:/, राहत – /ra:ɦat/, बेगाना – /be:ga:na:/, दिल – /dɪl/, ज़ख्म – /zəxm/, आँसू – /āsu:/, दर्द – /dərd/, अफ़साना – /əfisa:na:/, बहार – /bəɦa:r/, खिज़ाँ – /xɪza:n/, अरमान – /ərma:n/, आखिर – /ɑ:xɪr/, फूँक – /fu:ŋk/, समां – /səma:n/, परवाना – /pərva:na:/, चने – /tʃəne/, बाबू – /ba:bu:/, भैया – /b^hajja:/, चाव – /tʃa:v/, ज़बान – /zəba:n/, पानी – /pa:ni:/, भई – /b^hai:/, साधो – /sa:dho:/, गुरू – /guru:/, बानी – /ba:ni:/, मदरसे – /mədərse/, क़ाज़ी – /qa:zi:/, दिनों – /dɪno:/, ठाट – /t̪ha:t̪/, बाट – /ba:t̪/, अफ़सर – /əfəsər/, गवरनर – /gəvərnər/, लाट – /la:t̪/, दफ़्तर – /dəftər/, चाट – /tʃa:t̪/, मसाला – /məsa:la:/, लाला – /la:la:/, स्वाद – /sva:d/, बहार – /bəɦa:r/, सरकार – /sərkɑ:r/, एतबार – /etba:r/, ललकार – /ləlka:r/, दरबार – /dərbɑ:r/, रियासत – /rija:sət/, सरदार – /sərdɑ:r/, सूबेदार – /su:be:da:r/, तहसीलदार – /t̪əɦsi:lɖa:r/, थानेदार – /t̪hɑ:ne:da:r/, सिपहसालार – /sɪpəɦsa:lɑ:r/, वानर – /va:nər/, सेना – /se:na:/

As we can see, despite the model expressing a willingness to follow our guidelines, it still repeats the same output as before. Even if we explicitly reiterate the list and include our guidelines, the result remains largely unchanged:

Table 16. ChatGPT conversation, 2024-02-18

U	Provide phonological transcription for the following Hindi forms: शुरू, खुशबू, राहत, बेगाना, दिल, ज़ख्म, आँसू, दर्द, अफ़साना, बहार, खिज़ाँ, अरमान, आखिर, फूँक, समां, परवाना, चने, बाबू, भैया, चाव, ज़बान, पानी, भई, साधो, गुरू, बानी, मदरसे, क़ाज़ी, दिनों, ठाट, बाट, अफ़सर, गवरनर, लाट, दफ़्तर, चाट, मसाला, लाला, स्वाद, बहार, सरकार, एतबार, ललकार, दरबार, रियासत, सरदार, सूबेदार, तहसीलदार, थानेदार, सिपहसालार, वानर, सेना. Use always always /a/ instead of /ə/, /u/ for /ʊ/, /i/ for /ɪ/, /d/ for /d̪/, /t/ for /t̪/, /v/ for /v/
C	ChatGPT: Here are the phonological transcriptions for the provided Hindi words with the modifications: शुरू – /fu:ru:/, खुशबू – /xʊʃbu:/, राहत – /ra:ɦat/, बेगाना – /be:ga:na:/, दिल – /dɪl/, ज़ख्म – /zəxm/, ... (the same forms as before follow).

This method does not seem to work. What else can we do? What is so attractive about AI models, is their ability to learn. While it is not possible for an ordinary user to perform real supplementary training of the model, maybe we can provide a large set of examples and ask ChatGPT to use them as a guide?

We will use the data from the same research. Let us select the next 500 forms that immediately follow the list based on the first two songs we have been working on so far. To simplify the process, we will use a system closer to the IPA transcription, which the model seems to prefer. Of course, we will not include in the list the forms we expect the model to transcribe.

Table 17. ChatGPT conversation, 2024-02-18

U	Transcribe the following words: राहत, बेगाना, दिल, ज़ख्म, आँसू, दर्द, अफ़साना, बहार, खिज़ाँ, अरमान, आखिर, फूँक, समाँ, परवाना, चने, बाबू, भैया, चाव, ज़बान, पानी, भई, साधो, गुरू, बानी, मदरसे, क़ाज़ी, दिनों, ठाट, बाट, अफ़सर, गवरनर, लाट, दफ़्तर, चाट, मसाला, लाला, स्वाद, बहार, सरकार, एतबार, ललकार, दरबार, रियासत, सरदार, सूबेदार, तहसीलदार, थानेदार, सिपहसालार, वानर, सेना using the list below as a transcription guide: भगवान b ^h agva:n, दुखी duk ^{hi} :, दान da:n, भूखे b ^h u:k ^h e, अन्न ann, बुरे bure (...) ⁴⁷
C	Here are the phonological transcriptions for the provided Hindi words: राहत – ra:hat, बेगाना – bega:na:, दिल – dil, ज़ख्म – zakhm, आँसू – ā:su:, दर्द – dard, अफ़साना – afsa:na:, बहार – baha:r, खिज़ाँ – xiza:, अरमान – arma:n, आखिर – a:khir, फूँक – p ^h u:ŋk, समाँ – sama:, परवाना – parwa:na:, चने – tʃane, बाबू – ba:bu:, भैया – b ^h a:ja:, चाव – tʃa:v, ज़बान – zaba:n, पानी – pa:ni:, भई – b ^h ai:, साधो – sa:d ^h o:, गुरू – guru:, बानी – ba:ni:, मदरसे – madrase, क़ाज़ी – qa:zi:, दिनों – dinə, ठाट – t ^h a:t, बाट – ba:t, अफ़सर – afsar, गवरनर – gavarnar, लाट – la:t, दफ़्तर – daftər, चाट – tʃa:t, मसाला – masa:la:, लाला – la:la:, स्वाद – sva:d, बहार – baha:r, सरकार – sarka:r, एतबार – etba:r, ललकार – lalka:r, दरबार – darba:r, रियासत – ri:ja:sat, सरदार – sarda:r, सूबेदार – subeda:r, तहसीलदार – tahsi:l-da:r, थानेदार – t ^h a:ne-da:r, सिपहसालार – sipəhsa:la:r, वानर – va:nar, सेना – sena:.

Compared to all the previous results, this output is nearly perfect. To improve it further, we could try adjusting the forms *parwa:na:*, *sva:d*, and *sipəhsa:la:r* to comply with our transcription system. The best approach would be to add some additional forms to our guidelines list, teaching the model how to handle similar cases.

⁴⁷ For the complete list, see Appendix 1.

Table 18. ChatGPT conversation, 2024-02-18

U	Transcribe the following words: राहत, बेगाना, दिल, ज़ख्म, आँसू, दर्द, अफ़साना, बहार, ख़िज़ाँ, अरमान, आख़िर, फूँक, समाँ, परवाना, चने, बाबू, भैया, चाव, ज़बान, पानी, भई, साधो, गुरू, बानी, मदरसे, क़ाज़ी, दिनों, ठाट, बाट, अफ़सर, गवरनर, लाट, दफ़्तर, चाट, मसाला, लाला, स्वाद, बहार, सरकार, एतबार, ललकार, दरबार, रियासत, सरदार, सूबेदार, तहसीलदार, थानेदार, सिपहसालार, वानर, सेना using the list below as a transcription guide: भगवान bʰagva:n, दुखी dukhi:, दान da:n, भूखे bʰu:khe, अन्न ann, बुरे bure (...), उतरवाना utarva:na:, पिरवाना pirva:na:, स्वात sva:t, स्वादु sva:du, सिपहगिरी sipahgiri:
C	राहत – ra:hat, बेगाना – bega:na:, दिल – dil, ज़ख्म – zakm, आँसू – āsu:, दर्द – dard, अफ़साना – afsa:na:, बहार – baha:r, ख़िज़ाँ – xi:zā:, अरमान – arma:n, आख़िर – a:xir, फूँक – pʰu:ŋk, समाँ – sama:, परवाना – parva:na:, चने – tʃane, बाबू – ba:bu:, भैया – bʰa:jaa, चाव – tʃa:v, ज़बान – zaba:n, पानी – pa:ni:, भई – bʰa:i:, साधो – sa:dʰo:, गुरू – guru:, बानी – ba:ni:, मदरसे – madrase, क़ाज़ी – qa:zi:, दिनों – dino:, ठाट – tʰa:t, बाट – ba:t, अफ़सर – afsar, गवरनर – gavarna:r, लाट – la:t, दफ़्तर – dafta:r, चाट – tʃa:t, मसाला – masa:la:, लाला – la:la:, स्वाद – sva:d, बहार – baha:r, सरकार – sarka:r, एतबार – etba:r, ललकार – lalka:r, दरबार – darba:r, रियासत – ria:sat, सरदार – sarda:r, सूबेदार – subeda:r, तहसीलदार – tahsila:da:r, थानेदार – tʰa:ne:da:r, सिपहसालार – sipahsala:r, वानर – va:nar, सेना – sena:

Better, but it is worth noting that some of the issues tend to reappear (e.g. *ḡ* in *tahsila:da:r*, *ə* in *va:nar*, etc.), and some new ones appear (like *zakm* alongside *a:xir*). But let us try a technique we have not used so far. We are going to repeat the last query several times. At some point, we get an answer like this:

Table 19. ChatGPT conversation, 2024-02-18

U	(...) ⁴⁸
C	राहत: ra:hat, बेगाना: bega:na:, दिल: dil, ज़ख्म: zakhm, आँसू: a:nsu:, दर्द: dard, अफ़साना: afsa:na:, बहार: baha:r, ख़िज़ाँ: xi:zā:, अरमान: arma:n, आख़िर: a:xir, फूँक: phu:ŋk, समाँ: sama:, परवाना: parva:na:, चने: tʃane, बाबू: ba:bu:, भैया: bʰa:i:a, चाव: tʃa:v, ज़बान: zaba:n, पानी: pa:ni:, भई: bʰa:i:, साधो: sa:dʰo:, गुरू: guru:, बानी: ba:ni:, मदरसे: madarase, क़ाज़ी: qa:zi:, दिनों: dino:, ठाट: tʰa:t, बाट: ba:t, अफ़सर: afsar, गवरनर: gavarnar, लाट: la:t, दफ़्तर: daftar, चाट: tʃa:t, मसाला: masa:la:, लाला: la:la:, स्वाद: sva:d, बहार: baha:r, सरकार: sarka:r, एतबार: etba:r, ललकार: lalka:r, दरबार: darba:r, रियासत: riyasat, सरदार: sarda:r, सूबेदार: subeda:r, तहसीलदार: tahsi:lda:r, थानेदार: tha:ne:da:r, सिपहसालार: sipahsala:r, वानर: va:nar, सेना: sena:

⁴⁸ Exactly the same as the previous query.

As we can see, most of the issues we were trying to eliminate have been resolved, however, unfortunately, new ones have appeared (*zakhm* instead of *zaxm*, *tha:ne:da:r* instead of *tʰa:ne:da:r*). Nevertheless, it is worth noting that the results we see now are much better than at the very beginning.

It is probably a good moment to draw attention to an issue that is probably, at the moment, the biggest challenge in using ChatGPT for any serious scientific research: the lack of repeatability. McAdoo notes that “the results of a ChatGPT »chat« are not retrievable by other readers”.⁴⁹ In fact, they are not even retrievable by the same user, especially if a considerable amount of time passes between attempts. The author of the present paper repeated the same or similar queries as described above within the span of a few days in February 2023, which produced consistent or very similar results. However, an attempt to reproduce the same results two months later led to complete failure. However, at least in some cases, a minor change in query syntax, completely insignificant from a human perspective, led to the model cooperating again.

Let us leave aside this issue now and return to our transcription challenge. An interesting question is whether providing a different reference list could prompt ChatGPT to use another transcription standard, e.g. one used in McGregor’s Oxford Hindi-English Dictionary⁵⁰.

Table 20. ChatGPT conversation, 2024-04-19

U	Transcribe the following words: राहत, बेगाना, दिल, ज़ख्म, आँसू, दर्द, अफ़साना, बहार, ख़िज़ाँ, अरमान, आख़िर, फूँक, समाँ, परवाना, चने, बाबू, भैया, चाव, ज़बान, पानी, भई, साधो, गुरू, बानी, मदरसे, क़ाज़ी, दिनों, ठाट, बाट, अफ़सर, गवरनर, लाट, दफ़्तर, चाट, मसाला, लाला, स्वाद, बहार, सरकार, एतबार, ललकार, दरबार, रियासत, सरदार, सूबेदार, तहसीलदार, थानेदार, सिपहसालार, वानर, सेना using the list below as a transcription guide: भगवान /bʰagvān/, दुखी /dukʰi/, दान /dān/, भूखे /bʰūkʰe/ ... ⁵¹
C	राहत – rāhat, बेगाना – begānā, दिल – dil, ज़ख्म – zakhm, आँसू – ānsū, दर्द – dard, अफ़साना – afsānā, बहार – bahār, ख़िज़ाँ – khizān, अरमान – armaan, आख़िर – ākhir, फूँक – phoonk, समाँ – samān, परवाना – parwānā, चने – chane, बाबू – bābū, भैया – bhaiyā, चाव – chāv, ज़बान – zabān, पानी – pānī, भई – bhai, साधो – sādho, गुरू – guru, बानी – bānī, मदरसे – madrase, क़ाज़ी – qāzi, दिनों – dinon, ठाट – thāt, बाट – bāt, अफ़सर – afsar, गवरनर – gavarnar, लाट – lāt, दफ़्तर – daftar, चाट – chāt, मसाला – masālā, लाला – lālā, स्वाद – swād, बहार – bahār, सरकार – sarkār, एतबार – etbār, ललकार – lalakār, दरबार – darbār, रियासत – riyāsāt, सरदार – sardār, सूबेदार – sūbedār, तहसीलदार – tahsīldār, थानेदार – thānēdār, सिपहसालार – sipahsālār, वानर – vānar, सेना – senā

⁴⁹ Timothy McAdoo, ‘APA Style. How to cite ChatGPT’.

⁵⁰ We will introduce minor changes into McGregor’s transcription, the most systemic one being marking aspiration with Unicode superscript ^h sign (U+02B0) instead of standard h, which is left to render only the glottal fricative.

⁵¹ The complete list is available in Appendix 2.

The result is not bad, although the model ignored our idea to use the special superscript ^h to mark aspiration. Interestingly, this inadvertently brings the result closer to the standard used in McGregor’s dictionary. However, there are still some forms where the model clearly defaulted to its standard transcription system, such as: armaan, phoonk, guru, dinon. Let us try to address this issue.

Table 21. ChatGPT conversation, 2024-04-19

U	This is exactly the same transcription as before.
C	Apologies for the oversight. Here are the corrected transcriptions: अरमान – aramān, फूँक – phūnk, गुरू – gurū, दिनों – dinō

Good job, ChatGPT! Let’s show mercy and refrain from commenting on the additional vowel in **aramān*. The rest is perfect.

Extending the guide list or carefully choosing the forms to include could help to eliminate most mistakes. However, I would argue that the current error rate is no higher than in a human-written paper submitted for proofreading. Therefore, we may conclude that task number 2 has been completed.

3. Reconstructing lemmata of the selected textual forms

Let us set aside pronunciation and transcription for now and focus on morphology. An important step in analyzing the assembled corpus was identifying the lemmata, or dictionary forms, of the words occurring in the studied texts. Let us see if ChatGPT can assist us with this.

Table 22. ChatGPT conversation, 2024-02-24

U	Provide dictionary forms for the following nouns: rāhat, begānā, dil, zakhm, ānsū, dard, afsānā, bahār, khizā, armān, ākhir, phūk, samā, parvānā, cane, bābū, bhaiyā, cāv, zabān, pānī, bhaī, sādho, gurū, bānī, madrase, qāzi, dinō, ṭhāt, bāt, afsar, gavranar, lāt, daftar, cāt, masālā, lālā, svād, bahār, sarkār, etbār, lalkār, darbār, riyāsat, sardār, sūbedār, tahsildār, thānedār, sipahsālār, vānar, senā
C	Sure, here are the dictionary forms for the provided nouns: rāhat: राहत (rāhat), begānā: बेगाना (begānā), dil: दिल (dil), zakhm: ज़ख्म (zakhm), ānsū: आंसू (ānsū), dard: दर्द (dard), afsānā: अफ़साना (afsānā), bahār: बहार (bahār), khizā: खिज़ाँ (khizā), armān: अरमान (armān), ākhir: आखिर (ākhir), phūk: फूँक (phūk), samā: समाँ (samā), parvānā: परवाना (parvānā), cane: केन (cane) – It seems like there might be a typo here. If you meant “cane” as in sugar cane, the Hindi term is गन्ना (gannā), bābū: बाबू (bābū), bhaiyā: भैया (bhaiyā), cāv: चाव (cāv), zabān: ज़बान (zabān), pānī: पानी

Table 22. cont.

	(pānī), bhaī: भाई (bhaī), sādho: साधो (sādho), gurū: गुरू (gurū), bānī: बानी (bānī), madrase: मद्रसे (madrase), qāzi: काज़ी (qāzi), dinō: दिनों (dinō), ṭhāt: ठाठ (ṭhāt), bāt: बाट (bāt), afsar: अफ़सर (afsar), gavranar: गवरानार (gavranar), lāt: लाठ (lāt), daftar: दफ़्तर (daftar), cāt: चाट (cāt), masālā: मसाला (masālā), lālā: लाला (lālā), svād: स्वाद (svād), bahār: बहार (bahār), sarkār: सरकार (sarkār), etbār: इतबार (etbār), lalkār: ललकार (lalkār), darbār: दरबार (darbār), riyāsat: रियासत (riyāsat), sardār: सरदार (sardār), sūbedār: सूबेदार (sūbedār), tahsīldār: तहसीलदार (tahsīldār), thānedār: थानेदार (thānedār), sipahsālār: सिपाहसालार (sipahsālār), vānar: वानर (vānar), senā: सेना (senā)
--	--

First of all, let us note that the model has trouble interpreting *c* as the sign for the palatal affricate. However, what is for us far more important right now is that it seems not to understand the term “dictionary form”. For every single word submitted, it simply repeated it. While this approach works accidentally for many words, for some (*sādho*, *madrase*, *cāne*, *dinō*) we expected a different response.

Since the results are not satisfactory, let us try another approach. What if we specify that we expect the direct case singular form, instead of referring to it as a “dictionary form”? Additionally, let us include some more complex examples to see clearly how the model handles the problem.

Table 23. ChatGPT conversation, 2024-02-24

U	Provide direct case singular forms for the following nouns: rāhat, begānā, dil, zakhm, ānsū, dard, afsānā, bahār, khizā, armān, ākhir, phūk, samā, parvānā, cane, bābū, bhaiyā, cāv, zabān, pānī, bhaī, sādho, gurū, bānī, madrase, qāzi, dinō, ṭhāt, bāt, afsar, gavranar, lāt, daftar, cāt, masālā, lālā, svād, bahār, sarkār, etbār, lalkār, darbār, riyāsat, sardār, sūbedār, tahsīldār, thānedār, sipahsālār, vānar, senā, vinatiyā, diśāō, ghumṛe, sapne, ḍalō, bātō, ākhō.
C	(ChatGPT produces exactly the same list as presented in Table 22)

Unfortunately, the result is the same. For all the words in the list (including the newly added ones), the model simply repeated their form as rect. sg. Let us see now what happens if we use the Devanāgarī script instead of transcription? Initially, there was no improvement. However, after repeating the same question several times (out of frustration, let us admit that honestly), we can finally obtain something like this:

Table 24. ChatGPT conversation, 2024-02-24

U	provide Direct Case Singular form for the following Hindi nouns: राहत, बेगाना, दिल, ज़ख्म, आँसू, दर्द, अफ़साना, बहार, खिज़ाँ, अरमान, आखिर, फूँक, समां, परवाना, चने, बाबू, भैया, चाव, ज़बान, पानी, भई, साधो, गुरू, बानी, मदरसे, क़ाज़ी, दिनों, ठाट, बाट, अफ़सर, गवरनर, लाट, दफ़्तर, चाट, मसाला, लाला, स्वाद, बहार, सरकार, एतबार, ललकार, दरबार, रियासत, सरदार, सूबेदार, तहसीलदार, थानेदार, सिपहसालार, वानर, सेना, विनतियाँ, दिशाओं, घुमड़े, सपने, डलों, बातों, आँखों.
C	Here are the Direct Case Singular forms for the provided Hindi nouns: राहत: राहत (rāhat), बेगाना: बेगाना (begānā), दिल: दिल (dil), ज़ख्म: ज़ख्म (zakhm), आँसू: आँसू (ānsū), दर्द: दर्द (dard), अफ़साना: अफ़साना (afsānā), बहार: बहार (bahār), खिज़ाँ: खिज़ाँ (khizā), अरमान: अरमान (armān), आखिर: आखिर (ākhir), फूँक: फूँक (phūṅk), समां: समां (samā), परवाना: परवाना (parvānā), चने: चना (chane), बाबू: बाबू (bābū), भैया: भैया (bhaiyā), चाव: चाव (chāv), ज़बान: ज़बान (zabān), पानी: पानी (pānī), भई: भई (bhaī), साधो: साधू (sādhū), गुरू: गुरू (gurū), बानी: बानी (bānī), मदरसे: मदरसा (madrase), क़ाज़ी: क़ाज़ी (qāzi), दिनों: दिन (din), ठाट: ठाठ (thāt), बाट: बाट (bāt), अफ़सर: अफ़सर (afsar), गवरनर: गवरनर (gavranar), लाट: लाठ (lāt), दफ़्तर: दफ़्तर (daftar), चाट: चाट (chāt), मसाला: मसाला (masālā), लाला: लाला (lālā), स्वाद: स्वाद (svād), बहार: बहार (bahār), सरकार: सरकार (sarkār), एतबार: एतबार (etbār), ललकार: ललकार (lalkār), दरबार: दरबार (darbār), रियासत: रियासत (riyāsat), सरदार: सरदार (sardār), सूबेदार: सूबेदार (sūbedār), तहसीलदार: तहसीलदार (tahsīldār), थानेदार: थानेदार (thānedār), सिपहसालार: सिपहसालार (sipahsālār), वानर: वानर (vānar), सेना: सेना (senā), विनतियाँ: विनति (vinatiyā), दिशाओं: दिशा (dīśā), घुमड़े: घुमड़ा (ghumṛe), सपने: सपना (sapne), डलों: डला (ḍalā), बातों: बात (bāt), आँखों: आँख (āṅkh).

As we can see, at least in some cases the model provided different forms from the submitted ones. Incorrect rect. sg. forms were given for *thāt* and *lāt*. It is unclear why the non-aspirated consonant changes to an aspirated one there. The dictionary forms provided for *sādhū* and *ghumṛe* are also incorrect, but they at least follow real inflection patterns of Hindi. However, the remaining forms are inflected correctly. This is a step forward.

If we complain to the model (“you have provided wrong direct case singular forms for साधो, ठाट, लाट, घुमड़े”), it will immediately correct the forms for *thāt* and *lāt*. However, the issue of *sadhū* and *ghumṛe* is more persistent.

Table 25. ChatGPT conversation, 2024-02-24

U	The forms you propose for साधो and घुमड़े belong actually to different words.
C	You are correct. I apologize for the confusion. Here are the corrected forms: साधो: साध (sādh), घुमड़े: घुमड़ा (ghumṛā).

As far as *sādh* is concerned, *sādh* is finally identified as the correct lemma. However, instead of the form *ghumṛā*, consistently provided by the model, we would rather expect *ghumar* in this context. Nevertheless, the mistake rate is not very high and, just as with the transcription, we may conclude we have achieved a satisfactory level of accuracy.

4. Classifying the lemmata on the basis of their origin

The next stage of the original work was tagging the lemmata based on their origin: native (tadbhava/tatsama) and borrowed (from what language). This is the first attempt:

Table 26. ChatGPT conversation, 2024-02-26

U	Please, classify the following Hindi words on the basis of their origin. Tag every word as either Indian (IN), Persian/Arabic (FA/AR), English (EN), Mixed Indian and Persian/Arabic (IN-FA/AR) or Other (O): begānā, dil, zaḡḡm, āmsū, dard, afsānā, bahār, ḡḡizām, armān, āḡḡhir, phūmk, parvānā, canā, bābū, bhaiyā, cāv, zabān, pānī, bhaī, sādha, guru, bānī, madrasā, qāzī, din, ṡṡāt, bāt, afsar, gavarnar, lāt, daftar, cāt, masālā, lālā, svād, bahār, sarkār, etbār, lalkār, darvār, riyāsāt, sardār, sūbedār, tehsīldār, ṡṡānedār, sipahsālār, vānar, senā
C	Here’s the classification of the provided Hindi words based on their origin: rāhat – FA/AR, begānā – FA/AR, dil – IN, zaḡḡm – FA/AR, āmsū – FA/AR, dard – FA/AR, afsānā – FA/AR, bahār – FA/AR, ḡḡizām – FA/AR, armān – FA/AR, āḡḡhir – FA/AR, phūmk – FA/AR, parvānā – FA/AR, canā – FA/AR, bābū – FA/AR, bhaiyā – FA/AR, cāv – EN, zabān – FA/AR, pānī – FA/AR, bhaī – FA/AR, sādha – FA/AR, guru – FA/AR, bānī – FA/AR, madrasā – FA/AR, qāzī – FA/AR, din – FA/AR, ṡṡāt – FA/AR, bāt – FA/AR, afsar – FA/AR, gavarnar – FA/AR, lāt – FA/AR, daftar – FA/AR, cāt – FA/AR, masālā – FA/AR, lālā – FA/AR, svād – FA/AR, bahār – FA/AR, sarkār – FA/AR, etbār – FA/AR, lalkār – FA/AR, darvār – FA/AR, riyāsāt – FA/AR, sardār – FA/AR, sūbedār – FA/AR, tehsīldār – FA/AR, ṡṡānedār – FA/AR, sipahsālār – FA/AR, vānar – FA/AR, senā – FA/AR Key: FA/AR: Persian/Arabic, IN: Indian, EN: English, O: Other (for the unknown word)

The result is far from satisfactory. Almost all the forms were classified as Perso-Arabic, with the only exception being *dil*, which was classified as Indian, despite there being no doubt about its Persian origin. Let us try using different wording in the input.

Table 27. ChatGPT conversation, 2024-02-26

U	Indicate words of Persian and Arabic origin among the following Hindi words: Please, classify the following Hindi words on the basis of their origin. Tag every word as either Indian (IN), Persian/Arabic (FA/AR), English (EN), Mixed Indian and Persian/Arabic (IN-FA/AR) or Other (O): begānā, dil, zaḥm, āmsū, dard, afsānā, bahār, khizām, armān, ākhir, phūmk, parvānā, canā, bābū, bhaiyā, cāv, zabān, pānī, bhaī, sādha, guru, bānī, madrasā, qāzī, dīn, thāt, bāt, afsar, gavarnar, lāt, daftar, cāt, masālā, lālā, svād, bahār, sarkār, etbār, lalkār, darvār, riyāsāt, sardār, sūbedār, tehsildār, thānedār, sipahsālār, vānar, senā.
C	Here are the Hindi words classified based on their origin: Persian/Arabic (FA/AR): begānā, zaḥm, āmsū, dard, afsānā, bahār, khizām, armān, ākhir, phūmk, parvānā, zabān, pānī, sādha, qāzī, dīn, thāt, bāt, afsar, gavarnar, lāt, daftar, cāt, masālā, lālā, bahār, sarkār, etbār, lalkār, darvār, riyāsāt, sardār, sūbedār, tehsildār, thānedār, sipahsālār, vānar. Indian (IN): dil, bābū, bhaiyā, cāv, pānī, bhaī, guru, bānī, madrasā, svād, senā. Mixed Indian and Persian/Arabic (IN-FA/AR): canā

The outcome is different but not actually better. A random classification might yield more reliable results. What approach could we use now? We have previously noted that using Devanāgarī forms instead of transcription can influence the results. Let's try this method again:

Table 28. ChatGPT conversation, 2024-02-26

U	Indicate words of Persian and/or Arabic origin in the following list: बेगाना, दिल, ज़ख्म, आँसू, दर्द, अफ़साना, बहार, खिज़ाँ, अरमान, आख़िर, फूँक, परवाना, चना, बाबू, भैया, चाव, ज़बान, पानी, भई, साथ, गुरू, बानी, मदरसा, क़ाज़ी, दिन, ठाट, बाट, अफ़सर, गवर्नर, लाट, दफ़्तर, चाट, मसाला, लाला, स्वाद, बहार, सरकार, एतबार, ललकार, दरबार, रियासत, सरदार, सूबेदार, तहसीलदार, थानेदार, सिपहसालार, वानर, सेना
C	In the given list, the words of Persian and/or Arabic origin are: बेगाना (begānā), ज़ख्म (zaḥm), आँसू (āmsū), दर्द (dard), अफ़साना (afsānā), बहार (bahār), खिज़ाँ (khizām), अरमान (armān), आख़िर (ākhir), फूँक (phūmk), परवाना (parvānā), ज़बान (zabān), अफ़सर (afsar), गवर्नर (gavarnar), दफ़्तर (daftar), मसाला (masālā), लाला (lālā), सरकार (sarkār), एतबार (etbār), ललकार (lalkār), दरबार (darbār), रियासत (riyāsāt), सरदार (sardār), सूबेदार (sūbedār), तहसीलदार (tehsildār), थानेदार (thānedār), सिपहसालार (sipahsālār). These words have their origins in Persian and/or Arabic.

This looks much better, indeed. Only two Indian forms have been erroneously classified as Perso-Arabic, ie. *āmsū* and *lalkār*. The word *thānedār* is of mixed Indian/Persian origin, so this is not a complete mistake. The immediate source of *gavarnar*

is surely English. On the other hand, *afsar*, while probably ultimately of English origin⁵² might have been at least influenced by its Persian variant (originating from the same source), as the Hindi pronunciation is much closer to the Persian form than to the original English one.⁵³ The only Perso-Arabic form the model failed to identify correctly is *dil*, once again.

To further improve the output, we will now check if requiring the model to provide precise etymologies instead of merely classifying the words has any impact.

Table 29. ChatGPT conversation, 2024-02-29

U	Please, provide etymology for the following Hindi words: बेगाना, दिल, ज़ख्म, आँसू, दर्द, अफ़साना, बहार, ख़िज़ाँ, अरमान, आख़िर, फूँक, परवाना, चना, बाबू, भैया, चाव, ज़बान, पानी, भई, साध, गुरू, बानी, मदरसा, क़ाज़ी, दिन, ठाट, बाट, अफ़सर, गवरनर, लाट, दफ़्तर, चाट, मसाला, लाला, स्वाद, बहार, सरकार, एतबार, ललकार, दरबार, रियासत, सरदार, सूबेदार, तहसीलदार, थानेदार, सिपहसालार, वानर, सेना
C	Here are the etymologies for the provided Hindi words: बेगाना (begānā) – From Persian بیگانه (begānē), ultimately from Middle Persian. It means “stranger” or “unrelated.” दिल (dil) – From Persian دل (dil), ultimately from Middle Persian. It means “heart.” ज़ख्म (zakhm) – From Persian زخم (zakhm), ultimately from Middle Persian. It means “wound.” आँसू (ānsū) – From Sanskrit अश्रु (aśru). It means “tear” (as in crying). दर्द (dard) – From Persian درد (dard). It means “pain” or “ache.” अफ़साना (afsānā) – From Arabic أفسانه (afsāna). It means “story” or “tale.” बहार (bahār) – From Persian بهار (bahār). It means “spring” (the season). ख़िज़ाँ (khizān) – From Persian خزان (khizān). It means “autumn” or “fall” (the season). अरमान (armaan) – From Persian آرمان (ārmān). It means “desire” or “wish.” आख़िर (ākhir) – From Arabic آخر (ākhir). It means “end” or “finally.” फूँक (phūnk) – From Sanskrit पूक (pūka). It means “blow” or “puff” (as in blowing air). परवाना (parvānā) – From Persian پروانه (parvāna). It means “moth.” चना (chanā) – From Sanskrit चणक (chaṇaka). It means “gram” or “chickpea.” बाबू (bābū) – From Persian بابا (bābā), a term of endearment. It means “father” or “sir.” भैया (bhayyā) – From Sanskrit भ्रातृ (bhrāṭṛ). It means “brother.” चाव (chāv) – From Sanskrit चाव (chāva). It means “taste” or “flavor.” ज़बान (zabān) – From Persian زبان (zabān). It means “language” or “tongue.” पानी (pānī) – From Sanskrit पानीय (pānīya). It means “water.” भई (bhai) – From Sanskrit भ्रातृ (bhrāṭṛ). It means “brother.” साध (sādh) – From Sanskrit साधु (sādhū). It means “saint” or “virtuous person.” गुरू (gurū) – From Sanskrit गुरु (guru). It means “teacher” or “spiritual guide.”

⁵² McGregor, *Oxford Hindi-English Dictionary*, p. 45.

⁵³ See *afsar*, Hassan Amid, *Farhang-i Amid*, Tehrān 1373 SH, p. 174.

Table 29. cont.

C	बानी (bānī) – From Sanskrit वाणी (vāṇī). It means “speech” or “voice.” मदरसा (madrasā) – From Arabic مدرسة (madrasa). It means “school” or “educational institution.” क्वाज़ी (qāzī) – From Arabic قاضي (qāḍī). It means “judge.” दिन (din) – From Sanskrit दिन (dina). It means “day.” ठाट (thāt) – Uncertain, possibly derived from Sanskrit. बाट (bāt) – Uncertain, possibly derived from Sanskrit. अफ़सर (afsar) – From Persian افسر (afsar). It means “officer.” गवरनर (gavarnar) – From English “governor.” लाट (lāt) – Uncertain, possibly derived from Sanskrit. दफ़तर (daftar) – From Persian دفتر (daftar). It means “office” or “register.” चाट (chāt) – Uncertain, possibly derived from Sanskrit. मसाला (masālā) – From Arabic مسح (mash). It means “spice” or “seasoning.” लाला (lālā) – From Persian لاله (lālā). It means “tulip” or “darling.” स्वाद (svād) – From Sanskrit स्वादु (svādu). It means “taste” or “flavor.” बहार (bahār) – From Persian بهار (bahār). It means “spring” (the season). सरकार (sarkār) – From Persian سرکار (sarkār). It means “government” or “authority.” एतबार (etbār) – From Persian اعتبار (etbār). It means “trust” or “belief.” ललकार (lalkār) – From Sanskrit लल्कार (lalkār). It means “cry” or “shout.” दरबार (darbār) – From Persian دربار (darbār). It means “royal court” or “palace.” रियासत (riyāsāt) – From Persian ریاست (riyāsāt). It means “state” or “kingdom.” सरदार (sardār) – From Persian سردار (sardār). It means “leader” or “chief.” सूबेदार (sūbedār) – From Persian سوبیدار (sūbedār). It means “subordinate officer[”]
---	---

Let us assess the provided etymologies now.

Table 30. Etymology assessment

Original form + ChatGPT transcription	Actual transcription	ChatGPT etymology	Actual etymology
बेगाना (begānā)	begānā	Persian بیگانه (begānē)	correct ⁵⁴
दिल (dil)	dil	Persian دل (dil)	correct ⁵⁵
ज़ख़्म (zakhm)	zaxm	Persian زخم (zakhm)	correct ⁵⁶

⁵⁴ Francis Joseph Steingass, *A Comprehensive Persian-English Dictionary, Including the Arabic Words and Phrases to be Met with in Persian Literature*, London 1892, p. 223.

⁵⁵ Ibidem, p. 530.

⁵⁶ Ibidem, p. 611.

Table 30. cont.

Original form + ChatGPT transcription	Actual transcription	ChatGPT etymology	Actual etymology
आँसू (ānsū)	āṣū	Sanskrit अश्रु (aśru)	correct ⁵⁷
दर्द (dard)	dard	Persian درد (dard)	correct ⁵⁸
अफ़साना (afsānā)	afsānā	Arabic أفسانة (afsāna)	Persian < Middle Persian afsānag ⁵⁹
बहार (bahār)	bahār	Persian بهار (bahār)	correct ⁶⁰
खिज़ाँ (khizān)	xizā	Persian خزان (khizān)	correct, originally probably from Bactrian ⁶¹
अरमान (armaan)	armān	Persian آرمان (ārmān)	correct ⁶²
आखिर (ākhir)	āxir	Arabic آخر (ākhir)	correct ⁶³
फूँक (phūnk)	phūnk	Sanskrit पूँक (pūnka)	Indian, indeed, but the immediate source is rather Pk. phumkā ⁶⁴
परवाना (parvānā)	parvānā	Persian پروانه (parvāna)	correct ⁶⁵
चना (chanā)	canā	Sanskrit चणक (chaṇaka)	The Sanskrit protoform is caṇa ⁶⁶
बाबू (bābū)	bābū	Persian بابا (bābā)	correct ⁶⁷

⁵⁷ See Ralph Lilley Turner, *A Comparative Dictionary of Indo-Aryan Languages*, London 1966, p. 40.

⁵⁸ Steingass, *A Comprehensive Persian-English Dictionary*, p. 510.

⁵⁹ It occurs in the title of the lost Middle Persian text Hazār Afsānag, see Werner Sundermann, “Belles Lettres i. Sasanian Iran”, *Encyclopædia Iranica*, online edition, 2006, Viewed 03 January 2024, <<https://www.iranicaonline.org/articles/belles-lettres-sasanian-iran>>.

⁶⁰ Steingass, *A Comprehensive Persian-English Dictionary*, p. 209.

⁶¹ Mahmud Ja’fari Dehaqi, Amir Omad-od-Din Sadri, ‘Wām-wāzehā-ye balxi dar pārsi-ye nou’, *Zabān-šenāxt* 2/2 (1390 SH), p. 7.

⁶² Steingass, *A Comprehensive Persian-English Dictionary*, p. 39.

⁶³ Ibidem, p. 25.

⁶⁴ Turner, *A Comparative Dictionary of Indo-Aryan Languages*, p. 512.

⁶⁵ Steingass, *A Comprehensive Persian-English Dictionary*, p. 245.

⁶⁶ Turner, *A Comparative Dictionary of Indo-Aryan Languages*, p. 248.

⁶⁷ Steingass, *A Comprehensive Persian-English Dictionary*, p. 135.

Table 30. cont.

Original form + ChatGPT transcription	Actual transcription	ChatGPT etymology	Actual etymology
भैया (bhayyā)	bhaiyyā	Sanskrit भ्रातृ (bhrātr̥)	correct ⁶⁸
चाव (chāv)	cāv	Sanskrit चाव (chāva)	Indian, but the original Old Indian source is rather like cāh ⁶⁹
ज़बान (zabān)	zabān	Persian زبان (zabān)	correct ⁷⁰
पानी (pānī)	pānī	Sanskrit पानीय (pānīya)	correct ⁷¹
भई (bhaī)	bhaī	Sanskrit भ्रातृ (bhrātr̥)	correct ⁷²
साध (sādh)	sādh	Sanskrit साधु (sādhū)	possibly correct, Turner gives two options, the other being śrāddha ⁷³
गुरू (gurū)	gurū	Sanskrit गुरु (guru)	correct; gurū as variant of guru ⁷⁴ shares the same etymology ⁷⁵
बानी (bānī)	bānī	Sanskrit वाणी (vāṇī)	correct ⁷⁶
मदरसा (madrasā)	madrasā	Arabic مدرسة (madrasa)	correct ⁷⁷
काज़ी (qāzī)	qāzī	Arabic قاضي (qāḍī)	correct ⁷⁸

⁶⁸ Turner, *A Comparative Dictionary of Indo-Aryan Languages*, p. 552.⁶⁹ Ibidem, p. 259.⁷⁰ Steingass, *A Comprehensive Persian-English Dictionary*, p. 608.⁷¹ Turner, *A Comparative Dictionary of Indo-Aryan Languages*, p. 456.⁷² Ibidem, p. 552.⁷³ Ibidem, p. 736.⁷⁴ Caturvedi, *A Practical Hindi-English Dictionary*, p. 180.⁷⁵ See McGregor, *Oxford Hindi-English Dictionary*, p. 271 and Turner, *A Comparative Dictionary of Indo-Aryan Languages*, p. 226.⁷⁶ Dās, *Hindī śabdsāgar*, p. 3456; Turner, *A Comparative Dictionary of Indo-Aryan Languages*, p. 670.⁷⁷ Steingass, *A Comprehensive Persian-English Dictionary*, p. 1201.⁷⁸ Ibidem, p. 948.

Table 30. cont.

Original form + ChatGPT transcription	Actual transcription	ChatGPT etymology	Actual etymology
दिन (din)	din	Sanskrit दिन (dina)	correct ⁷⁹
ठाट (ṭhāt)	ṭhāt	Uncertain, possibly derived from Sanskrit.	Sanskrit thaṭṭha ⁸⁰
बाट (bāt)	bāt	Uncertain, possibly derived from Sanskrit.	Sanskrit vārtman ⁸¹
अफ़सर (afsar)	afsar	Persian افسر (afsar)	correct ⁸²
गवर्नर (gavarnar)	gavarnar	english “governor”	seems obviously correct, as the word follows pronunciation of the English governor.
लाट (lāt)	lāt	Uncertain, possibly derived from Sanskrit.	From English lord ⁸³ possibly via other North Indian language
दफ़्तर (daftar)	daftar	Persian دفتر (daftar)	correct, as the immediate source must have been Persian indeed. However, ultimately the word may be traced to Ancient Greek διφθέρα “hide”, whence it came into Aramaic and then Arabic. ⁸⁴
चाट (chāt)	cāt	Uncertain, possibly derived from Sanskrit.	S. caṭṭ ⁸⁵

⁷⁹ Turner, *A Comparative Dictionary of Indo-Aryan Languages*, p. 362.

⁸⁰ Ibidem, p. 347.

⁸¹ Ibidem, p. 663.

⁸² Amid, *Farhang-i Amid*, p. 174.

⁸³ McGregor, *Oxford Hindi-English Dictionary*, p. 891.

⁸⁴ Bernard Lewis, ‘Daftar’, in: P. Bearman (ed.), *Encyclopaedia of Islam New Edition Online (EI-2 English)*, Brill, https://doi.org/10.1163/1573-3912_islam_COM_0145.

⁸⁵ Turner, *A Comparative Dictionary of Indo-Aryan Languages*, p. 248.

Table 30. cont.

Original form + ChatGPT transcription	Actual transcription	ChatGPT etymology	Actual etymology
मसाला (masālā)	masālā	Arabic مسح (mash)	Arabic, but different word: maṣāliḥ ⁸⁶
लाला (lālā)	lālā	Persian لاله (lālā)	Probably correct, though the Persian form is incorrect (should be لالا not لاله) ⁸⁷
स्वाद (svād)	svād	Sanskrit स्वादु (svādu)	Borrowed from Sanskrit svāda ⁸⁸
सरकार (sarkār)	sarkār	Persian سرکار (sarkār)	correct ⁸⁹
एतबार (etbār)	etbār	Persian اعتبار (etbār)	Arabic via Persian ⁹⁰
ललकार (lalkār)	lalkār	Sanskrit लल्कार (lalkār)	Different form – Sanskrit lallakka ⁹¹
दरबार (darbār)	darbār	Persian دربار (darbār)	correct ⁹²
रियासत (riyāsāt)	riyāsāt	Persian ریاست (riyāsāt)	Arabic via Persian ⁹³
सरदार (sardār)	sardār	Persian سردار (sardār)	correct ⁹⁴
सूबेदार (sūbedār)	sūbedār	Persian سویدار (sūbedār)	actually صوبدار, Arabic šūba + Persian suffix -dār ⁹⁵

⁸⁶ McGregor, *Oxford Hindi-English Dictionary*, p. 797.⁸⁷ See ibidem, p. 893 and Steingass, *A Comprehensive Persian-English Dictionary*, p. 1112.⁸⁸ Turner, *A Comparative Dictionary of Indo-Aryan Languages*, p. 805.⁸⁹ Amid, *Farhang-i Amid*, p. 733.⁹⁰ Steingass, *A Comprehensive Persian-English Dictionary*, p. 72.⁹¹ Turner, *A Comparative Dictionary of Indo-Aryan Languages*, p. 637.⁹² Steingass, *A Comprehensive Persian-English Dictionary*, p. 508.⁹³ Ibidem, p. 600.⁹⁴ Ibidem, p. 673.⁹⁵ Ibidem, p. 795.

Of the 42 forms in question, the etymology provided by ChatGPT for 27 of them can be deemed correct, including, finally, the form *dil*. Nine more forms are correctly classified as Indian, even though the provided protoform is not correct or, in one case, the etymology is objectively unclear. One form (*sūbedār*) is classified as Persian, which is not entirely wrong, as it contains both Arabic and Persian elements. In other words, the source language was indicated correctly in 37 out of 42 cases, even though protoforms were sometimes incorrect.

Unfortunately, further analysis, like that conducted in the original study, is impossible due to the current restrictions on ChatGPT’s input and output (2049 characters). This limitation is far below what is needed for any meaningful statistical analysis.

5. Other models

Of course, the free ChatGPT 3.5 is not the only general AI model available. When this article was nearly finished, on May 14, 2024, OpenAI made its more advanced model, GPT-4, available for free with a limited quota. Other tools also exist, some of which are dedicated to supporting academic research. Most of these are tools specialized for particular applications; for example, *Scholarcy* is designed for extracting critical information from scientific texts, while *Consensus* gathers information from various published materials. General models are available, too (e.g. Scite.ai). Comparing the performance of all these models with GPT-3.5 in detail would certainly be interesting, but it would also turn this paper into a book. To provide at least a limited comparison, we will test the performance of two additional models – GPT-4 and Scite.ai – on one of the tasks: identifying nominal lexems in a text (see point 1 above).

5.1. ChatGPT-4

Table 31. ChatGPT-4 conversation, 2024-08-14

U	Please, indicate nouns in the following Hindi text: (...) – see Table 1
C	Here are the nouns in the given Hindi text: दीवाना (Deewana) राहत (Rahat) बेगाना (Begana) दिल (Dil) ज़ख्म (Zakhm) आंसू (Aansu) दर्द (Dard) अफ़साना (Afsana) बहार (Bahaar)

Table 31. cont.

	खिज़ाँ (Khizaan) दिल (Dil) अरमान (Armaan) समाँ (Sama) बलाऐँ (Balaayein) परवाना (Parwana)
--	--

As we can see, GPT-4 performs much better out of the box. It identified all the nouns except for फूँक phūnk “blowing, breath”. On the other hand, there are no false indications.

5.2. Scite.ai

Table 32. Scite.ai conversation, 2024-08-14

U	Please, indicate nouns in the following Hindi text: (...) – see Table 1
Scite.ai	The text provided contains a mix of nouns, verbs, adjectives, and pronouns. To identify the nouns in the text accurately, we need to consider the words that represent people, places, things, or ideas. Here is a breakdown of the nouns in the given Hindi text: 1. हे (Hey) 2. दीवाना (Crazy/Passionate) 3. राहत (Relief) 4. मैं (I) 5. बेगाना (Stranger) 6. दिल (Heart) 7. कैसे (How) 8. बहलाऊँ (Console) 9. ज़ख्म (Wound) 10. किसे (Whom) 11. आंसू (Tears) 12. हँसते (Laughing) 13. दर्द (Pain) 14. भरा (Filled) 15. अफ़साना (Story) 16. आयी (Arrived) 17. बहार (Spring) 18. ख़िज़ाँ (Autumn)

Table 32. cont.

U	Please, indicate nouns in the following Hindi text: (...) – see Table 1
	19. अरमान (Desire) 20. हैरान (Surprised) 21. आखिर (Finally) 22. आबाद (Populated) 23. विराना (Desolate) 24. घायल (Wounded) 25. तुम (You) 26. दूर (Far) 27. फूँक (Blown) 28. मजबूर (Helpless) 29. ए समाँ (O Time) 30. बलाएँ (Invitations) 31. परवाना (Moth) These are the nouns present in the provided Hindi text. Each of these words represents a person, place, thing, or idea, making them nouns in the context of the text.

As we can see, the outcome is not significantly better than the results achieved by GPT-3.5 and is much worse than GPT-4. Scite.ai identified correctly many nouns but also included some pronouns, verb forms, etc. in the list.

6. Other languages

Another intriguing idea would be to assess ChatGPT’s performance on a similar task using texts in languages other than Hindi. As with testing other models, conducting a complete experiment in one or more different languages is far beyond the scope of this paper. However, even a limited test can be valuable. Therefore, we will test the ability of GPT-3.5 to transcribe similar texts (task two in the original study) in two other languages using traditionally non-Latin orthographies.

6.1. Transcription of a Japanese song⁹⁶

Japanese uses a challenging mix of logograms of Chinese origin (kanji), two syllabaries (hiragana and katakana) and occasionally the Latin alphabet. While hiragana and katakana are unambiguous⁹⁷, the kanji characters usually have more than one reading (typically between 2 and 4, though some may have more than 10, e.g. 行: *kou, gyō, an, i.ku, yu.ku, -yu.ki, -yuki, -i.ki, -iki, okona.u, yasū, nami, name, tsura, michi, taka*⁹⁸). The choice of a particular reading depends on a lot of factors, including the specific word in which the character is used (and its origin), whether it is a personal name, etc.

We will encourage the model to provide a Latin alphabet transcription for the song 幸福はあの空から *Shiawase-wa ano sora-kara* “Happiness [comes] from this sky” by Haruo Oka (actually Tatsuo Sasaki).⁹⁹

Table 33. ChatGPT conversation, 2024-08-15

U	Provide transcription in Latin alphabet for the following Japanese song: 黄昏の並木路 ひとり見てましよう 淋しさに 恋しさに ひとり窓辺で見てましよう やがてつく 街の灯が 暗い心を 照らして 照らして 来る来る 来る来る来る 来るよ 幸福は あの街から かならず やって来る 来るよ あの頃の思い出を そっと呼びましよう 過ぎし日を 夢の日を そっと小声で呼びましよう やがて出る 月さえも 愛し面影 浮かべて 浮かべて 来る来る 来る来る来る 来るよ 幸福は あの空から かならず やって来る 来るよ
---	--

⁹⁶ I would like to express my gratitude to prof. Tomasz Majczak from the Department of Japanese Studies, Jagiellonian University, Cracow, Poland for his precious remarks on the two unusual readings of two Japanese forms (see below).

⁹⁷ With only minor inconsistencies resulting from historical factors, such as writing the subject particle *wa* as は <ha>.

⁹⁸ Jack Halpern (ed.), *The Kodansha Kanji Learner's Dictionary*, Tokyo-New York–London 1999, p. 69.

⁹⁹ The text of the song has been established based on the webpages <https://www.uta-net.com/song/108419/> and <https://utaten.com/lyric/yc19091806/>, confronting them with available recordings.

Table 33. cont.

	思い出を抱きしめて じっと待ちましょう 苦しみも かなしみも じっと耐えて待ちましょう やがて来る 幸福に 涙なんかは おさらば おさらば 来る来る 来る来る来る 来るよ 幸福は いつの日にか かならず やって来る 来るよ
C	Here is the transcription of the Japanese song into the Latin alphabet: Tasogare no namiki ji Hitori mite mashou Samishisa ni koishisa ni Hitori madobe de mite mashou Yagate tsuku machi no hi ga Kurai kokoro o terashite terashite Kuru kuru kuru kuru yo Koufuku wa ano machi kara Kanarazu yatte kuru kuru yo Ano koro no omoide o Sotto yobimashou Sugishi hi o yume no hi o Sotto kogoede yobimashou Yagate deru tsuki sae mo Itoshi omokage ukabete ukabete Kuru kuru kuru kuru yo Koufuku wa ano sora kara Kanarazu yatte kuru kuru yo Omoide o dakishimete Jitto machimashou Kurushimi mo kanashimi mo Jitto taete machimashou Yagate kuru koufuku ni Namida nanka wa osaraba osaraba Kuru kuru kuru kuru yo Koufuku wa itsu no hi ni ka Kanarazu yatte kuru kuru yo

The result is quite impressive. In fact, only two words were transcribed inaccurately: 幸福 which is clearly pronounced *shiawase* in the recordings, not *koufuku* (both meaning “happiness”¹⁰⁰) and 耐えて which is sung *koraete*, while dictionaries do not support this reading, unanimously reading it *taete* (*te*-form of the verb *taeru*). Again, both *koraeru* (normally written 堪える) and 耐える *taeru* have similar meanings revolving around “to bear, to stand, to endure, etc.”¹⁰¹ What we must emphasize here is that both cases where the model failed involve very atypical readings, especially *koraete*¹⁰². Such non-standard readings (especially when additionally clarified by furigana, i.e. phonetic reading guide above or beside the characters) is in fact part of the tradition of the Japanese script and may sometime be used to achieve specific results.¹⁰³

Another thing worth noting is the consistency of the transcription. In general, this is a big contrast to Hindi, especially if we consider the fact that the Devanāgarī script phonetic in principle.

6.2. Transcription of a Persian song

Modern Persian uses the Arabic alphabet with several additional letters. It is a semi-consonantal script, meaning that only some vowels are written in standard situations. This poses a significant challenge for the model, especially since one of the unwritten elements, the *ezāfe* marker *-e*, is an extremely important morphological feature.¹⁰⁴

We will use as an example the song آمدم *Āmadam* “I came” by Delkash (actually Esmat Bāqerpur Bāboli).¹⁰⁵

¹⁰⁰ Robert M.V. Collick, *Kenkyusha's New Collegiate Japanese-English Dictionary*, Tōkyō 1983, pp. 418, 497.

¹⁰¹ Ibidem, pp. 450 & 743.

¹⁰² In the case of 幸福 read as *shiawase*, there are some precedents, see e.g. the movie title 幸福のスイッチ *Shiawase no switchi* <https://www.imdb.com/title/tt0872018/>.

¹⁰³ See e.g., Tomasz Majtczak, ‘Translating the Untranslatable: A Yiddish Text in Japanese Rendering’, *Studia Linguistica Universitatis Iagellonicae Cracoviensis* 136 (2019), p. 305ff.

¹⁰⁴ To put it short, *ezāfe* in Modern Persian is a construction consisting of the head (often a noun) and a modifier (another noun, an adjective, etc.). The marker of the construction is *-e* after consonants (unwritten) and *-ye* after vowels (mostly written). Identifying places where *ezāfe* is proves crucial to understanding of any Persian text – see Ann K.S. Lambton, *Persian Grammar*, Cambridge 1960, p. 9, 20.

¹⁰⁵ The text was established based on the <https://www.musixmatch.com/lyrics/Delkash/Amadam> website checked against available recordings.

Table 34. ChatGPT conversation, 2024-08-15

U	provide transcription for the following Persian text: آمدم، آمدم، که سر آید انتظار ت آمدی، آمدی، که بمیرم در کنار تو می دانستم، در این خلوت، به امیدم بنشستی، شده ای بی تاب آری دارم، شوق رویت، چه بخوابم، چه نباشد خواب به عهد خود پابندم من، ز عشق تو خرسندم من به دامن ت سر بگذارم، جور زمان گر بگذرد بیم رقیب گر برخیزد، جور زمان گر بگذارد با تو سرخوشم جانم ز وصال ت، خرسند است بختم ز وفایت، بیدار است بر لب به نگاهت، لبخند است قلبم ز محبت، سر شار است گل با رویت به چشمم، خار است جز این دل شکسته، گو چه دارم من، تا بیفشانم، زیر پای تو ای آرزوی جانم، من نمی خواهم، جز وفای تو، جز وفای تو
C	Here is the transcription of the Persian text you provided: Āmadam, āmadam, ke sar āyad entezār-e to Āmadi, āmadi, ke bemīram dar kenār-e to Mī dānestam, dar īn khalvat, be omīdam benshaste, shodeī bī tāb Ārī dāram, shauq-e rūyat, che bekhābam, che nabāshad khwāb Be 'ahd-e khod pābandam man, ze 'eshq-e to khorsandam man Be dāmanat sar begzāram, jūr-e zamān gar begzārad Bīm-e raqīb gar barkhīzad, jūr-e zamān gar begzārad Bā to sarkhosham Jānam ze vesālat, khorsand ast Bakhtam ze vofāyat, bīdār ast Bar lab be negāhat, labkhand ast Qalb-am ze mohabbat, sar shār ast Gol bā rūyat be cheshmam, khār ast Joz īn del shekaste, gu che dāram man, tā bīfshānam, zīr pāy-e to Ey ārezū-ye jān-am, man namī khāham, joz vofā-ye to, joz vofā-ye to

The result is not bad, although it does have several imperfections, including some systemic ones. First of all, the model distinguishes between long and short vowels, while the quantity opposition in Modern New Persian is disputable at best.¹⁰⁶ Traditionally, the symbols *a* and *ā* (or *â*) are often used in transcription, but not as markers of length difference. The former represents a front vowel, while the latter is a back one, also a bit lower. Even if we agree to mark the historical vowel length in transcription, it should be

¹⁰⁶ Andrzej Pisowicz, *Origins of the New and Middle Persian Phonological Systems*, Kraków 1985.

done consistently. For example, we should either write *shodeī* and *āmadi* or better *shodei* and *āmadi*, but not *shodeī* and *āmadi*, as the final sound in both cases is the same (historically long) vowel. Similarly either *rūyat* and *gū* or better *ruyat* and *gu*, but not *rūyat* beside *gu*.

Apart from that, enclitic pronouns are written both attached to the proceeding word (e.g. *rūyat*) and with a hyphen (e.g. *jān-am*). The form **benshaste* should be corrected to *beneshasti*. The phrase *ze 'eshq* should be written *z-eshq*, in accordance with the pronunciation audible in recordings. **Khwāb* should be corrected to *khāb* (The former is an archaism, as it reflects the labialisation which was there in the Early New Persian, but not in the middle of the 20th century). The reading **vofāyat* is incorrect. It should be amended to *vafāyat*, while **bīfshānam* should be changed to *bi-afshānam* and **namī khāham* to *nemī-khāham* (or better: *nemi-khāham*). However, what is quite remarkable is that the model correctly identified most of the *ezāfe* phrases in the text, failing to notice only one in the penultimate verse (it should be *zir-e pāy-e to* instead of **zir pāy-e to*).

In general, despite several problems, the provided transcription is understandable, and all the words and most grammatical forms (except *benshaste*) are correctly identified. On the other hand, inconsistencies and mistakes in the transcription of the Persian text make us evaluate its performance at this task as closer to the results achieved in the case of Hindi than in the case of the Japanese text.

Conclusion

The experiment described in the present paper has shown that it is practically possible to obtain from ChatGPT results with a relatively low percentage of errors in performing the following tasks:

- i. identifying nouns in a Hindi text;
- ii. providing satisfactory phonological transcription for a list of words, complying with a chosen system;
- iii. reconstructing lemmata of selected forms;
- iv. classifying these lemmata based on their origin.

However, practically none of these results was achieved out of the box. All required a lot of additional effort. Moreover, the quality of the results often depended on secondary or even unrelated factors, such as specific wording, syntax of the query, and the use of original script versus transcription. In other words, we cannot rely on the model to perform these tasks reliably; we need to know a priori what the answer should be.

Another significant issue is the questionable repeatability of the results. While submitting identical queries within a few days generally led to similar (and sometimes even identical) results, a longer interval resulted in significantly different behaviour of the model. This is a well-known problem.¹⁰⁷

Among the factors that influence ChatGPT's answers, we should particularly consider those that we are unable to control, such as the random component or updates to the training set by the model administrators. Randomness is an inherent element of models like ChatGPT, as the text they generate is based on probability patterns derived from the training data, i.e., the texts used to train the model.¹⁰⁸ While some models may allow adjusting the level of randomness (the "temperature" parameter) it cannot be eliminated altogether (Ibid.)

Another issue that reduces the usability of ChatGPT for research is the limited quota for user queries.

Apart from that, there are clear differences in the model's 'competence' in various languages. Most probably, this is related to the proportion of materials in particular languages included in the GPT's training sets. The Japanese text transcription was notably the best, despite the challenge of identifying the readings of every single kanji character. Persian suffered from some problems and the transcription was inconsistent. The initially provided transcription of the Hindi text was fairly consistent but not satisfactory from a scientific perspective, as it did not capture several phenomena important from the point of view of the Hindi phonological system.

Still, let us bear in mind that general models are still a very recent development. Even ChatGPT cannot predict their evolution in the coming decades. The progress since the Eliza program of 1966 is astounding. While Eliza simulated participation in a dialogue and did not aim at actually solving any problems, ChatGPT provides answers and solutions even if only some of them are usable. It is worth noting that even a very limited comparison of ChatGPT-3.5 and ChatGPT-4 performance shows significant progress between these two versions.

We will end our considerations with one last question to ChatGPT. While poetry is not particularly significant for the results of our study, the sole fact that the model was able to create the text presented below is really impressive:

¹⁰⁷ Suzanne Fergus, Michelle Botha, Mehrnoosh Ostovar, 'Evaluating Academic Answers Generated Using ChatGPT', *Journal of Chemical Education* 100 (2023), p. 1674.

¹⁰⁸ Fabio Motoki et al., 'More human than human: measuring ChatGPT political bias', *Public Choice* 198 (2024), pp. 3–23, <https://doi.org/10.1007/s11127-023-01097-2>.

Table 35. ChatGPT conversation, 2024-04-15

U	Can you write a short poem in Persian about Persian and Arabic loanwords in Hindi?
C	در آواز هندی آمیخته‌ی هم دوست پارسی و عربی، دوستی بی‌پایان هر کلمه یک داستان، هر نغمه دلیل پیوندهای معنا، پیوندهای جان گل زبان پارسی، بوی عربی در آمیخته در هندی، همه‌ی زبان‌ها رقصان هر لحظه یک داستان، هر کلمه یک آواز در این همبستگی، دوستی می‌سازد پایان

Probably a translation could be of use at least for some of the readers...

Table 36. ChatGPT conversation, 2024-04-15

U	translate this poem to English and Hindi
C	Here is the translation of the poem into English and Hindi: English: In the melody of Hindi, intertwined with love, Persian and Arabic, an endless friendship. Each word a story, each note a reason, Bonds of meaning, bonds of soul. The rose of Persian language, infused with the scent of Arabic, In Hindi, all languages dance. Every moment a story, every word a melody, In this unity, friendship finds its end. Hindi: हिंदी की सुरमई में, प्रेम से उलझे हुए, पारसी और अरबी, एक अंतहीन दोस्ती। हर शब्द एक कहानी, हर नोट एक कारण, अर्थ के बंधन, आत्मा के बंधन। पारसी भाषा के गुलाब, अरबी की सुगंध के साथ लबालब, हिंदी में, सभी भाषाएँ नृत्य करती हैं। हर पल एक कहानी, हर शब्द एक सरगम, इस एकता में, दोस्ती अपना अंत ढूंढती है।

The epoch of Chatgpt-nāme is coming...

References

- Ali Ašraf Sādeqi, 'Arabic Language. i. Arabic elements in Persian', *Encyclopedia Iranica online*, Viewed 17 October 2024, <<https://iranicaonline.org/articles/arabic-i>>.
- Amid Hassan, *Farhang-i Amid*, Tehrān, Sephar, 1373 SH.
- Bahri Hardev, *Learners' Hindi-English dictionary*, Rajpal & Sons, Delhi 1989.
- Caturvedi Mahendra, *A Practical Hindi-English Dictionary*, National Publishing House, Delhi 1970.
- Collick Robert M. V., *Kenkyusha's New Collegiate Japanese-English Dictionary*, Kenkyusha, Tōkyō 1983.
- Dās Śyāmsundar, *Hindī śabdsāgar*, 11 vols. 2nd edition, Nāgarīpracārīnī Sabhā, Benares 1965–1975.
- Everaert Christine, *Tracing the Boundaries between Hindi and Urdu. Lost and Added in Translation between 20th Century Short Stories*, Brill, Leiden 2010.
- Fergus Suzanne, et al. 'Evaluating Academic Answers Generated Using ChatGPT', *Journal of Chemical Education* 100 (2023), pp. 1672–1675.
- Gacek Tomasz, 'Some Remarks on the Perso-Arabic Nominal Lexica in Bollywood Movie Song Lyrics', *Acta Asiatica Varsoviensia* 36 (2023), pp. 39–67, <https://doi.org/10.60018/AcAsVa.njyu2866>
- Halpern Jack (ed.), *The Kodansha Kanji Learner's Dictionary*, Kodansha, Tokyo–New York–London 1999.
- Ja'fari Dehaqi, Mahmud & Omad-od-Din Sadri, Amir, 'Wām-wāzehā-ye balxi dar pārsi-ye nou', *Zabān-senāxt* 2/2 (1390 SH).
- Kumar Dilip, *The Substance and the Shadow. An Autobiography*, Hay House, New Delhi 2014.
- Kuczkiwicz-Fraś Agnieszka, *Perso-Arabic Loanwords in Hindustani. Part ii, Linguistic Study*, Księgarnia Akademicka, Kraków 2012.
- Lambton Ann K. S., *Persian Grammar*, Cambridge University Press, Cambridge 1960.
- Lewis Bernard, 'Daftar', in: P. Bearman (ed.), *Encyclopaedia of Islam New Edition Online (EI-2 English)*. Brill. https://doi.org/10.1163/1573-3912_islam_COM_0145.
- Majtczak Tomasz, 'Translating the Untranslatable: A Yiddish Text in Japanese Rendering', *Studia Linguistica Universitatis Iagellonicae Cracoviensis* 136 (2019), pp. 297–307.
- Masica Colin, *The Indo-Aryan Languages*, Cambridge University Press, Cambridge 2001.
- McAdoo Timothy, 'APA Style. How to cite ChatGPT', *California State University Channel Islands website*, Viewed 17 October 2024, <<https://www.csuci.edu/wmc/documents/apa-citingchatgpt.pdf>>.
- McGregor Ronald Stuart, *Oxford Hindi-English Dictionary*, Oxford University Press, Oxford-Delhi 1993.
- Motoki Fabio, Pinho Neto, Valdemar, Rodrigues, Victor, 'More human than human: measuring ChatGPT political bias', *Public Choice* 198 (2024), pp. 3–23, <https://doi.org/10.1007/s11127-023-01097-2>.
- Ohala Manjari, *Aspects of Hindi Phonology*, Motilal Banarsidass, Delhi-Varanasi-Patna 1983.
- Ohala Manjari, 'Hindi', w: *Hanbook of the International Phonetic Association. A Guide to Use of the International Phonetic Alphabet*, Cambridge University Press, Cambridge 1999.
- Pisowicz Andrzej, *Origins of the New and Middle Persian Phonological Systems*, Uniwersytet Jagielloński, Kraków 1985.
- Steingass Francis Joseph, *A Comprehensive Persian-English Dictionary, Including the Arabic Words and Phrases to be Met with in Persian Literature*, Routledge & K. Paul, London 1892.
- Sundermann Werner, 'Belles Lettres i. Sasanian Iran', *Encyclopædia Iranica*, online edition, 2006, Viewed 03 January 2024, <<https://www.iranicaonline.org/articles/belles-lettres-sasanian-iran>>.
- Teel Zoë Abbie, et al. 'ChatGPT conundrums: Probing plagiarism and parroting problems in higher education practices', *College & Research Libraries News* 84/6 (2023), pp. 205–208.
- Turner Ralph Lilley, *A Comparative Dictionary of Indo-Aryan Languages*, Oxford University Press, London 1966.
- Windfuhr Gernot, 'Introduction to the Iranian Languages', in: G. Windfuhr (ed.), *The Iranian Languages*, Routledge, London–New York, 2009, pp. 1–3.

Appendix 1

Supplementary transcription guide (IPA-style variant)

- | | | |
|---|------------------------------------|--------------------------------------|
| 1. भगवान – b ^h agva:n | 40. लगन – lagan | 79. फूलों – p ^h u:lō |
| 2. दुखी – duk ^h i: | 41. सवेरे – savere | 80. खुशबू – xuśbu: |
| 3. दान – da:n | 42. रूप – ru:p | 81. कलियों – kaliyō |
| 4. भूखे – b ^h u:k ^h e | 43. चाँद – tʃā:d | 82. मोहब्बत – mohabbat |
| 5. अन्न – ann | 44. मन्न – mann | 83. तारों – ta:rō |
| 6. बुरे – bure | 45. बेकली – bekali: | 84. चेहरा – tʃehra: |
| 7. बदला – badla: | 46. तरहदार – tarahda:r | 85. रात – ra:t |
| 8. भला – b ^h ala: | 47. दिलरुबा – dilruba: | 86. दिन – din |
| 9. दास – da:s | 48. दिलदार – dilda:r | 87. दिलों – dilō |
| 10. वरदान – varda:n | 49. चितवन – tʃitvan | 88. ज़माना – zama:na: |
| 11. लाख – la:k ^h | 50. यार – ya:r | 89. राजा – ra:dʒa: |
| 12. बंसी – baṁsi: | 51. धोखा – d ^h oka: | 90. कोयल – koyal |
| 13. नाच – na:tʃ | 52. अय्यर – ayyar | 91. खेतों – k ^h etō |
| 14. मन – man | 53. आँखों – ā:k ^h ō | 92. जवानी – dʒava:ni: |
| 15. स्वर – svar | 54. जादू – dʒa:du: | 93. रानी – ra:ni: |
| 16. बिष – biṣ | 55. बातों – ba:tō | 94. धीर – d ^h i:r |
| 17. प्रेम – prem | 56. शोला – śola: | 95. रिया – riya: |
| 18. गीत – gi:t | 57. चरका – tʃarka: | 96. आरी – a:ri: |
| 19. तट – taṭ | 58. मोर – mor | 97. चुनरिया – tʃunriya: |
| 20. हाथ – ha:t ^h | 59. जीवन – dʒi:van | 98. नज़रिया – nazariya: |
| 21. जोड़ – dʒoṛ | 60. छाया – tʃ ^h a:ya: | 99. अम्बुआ – ambua: |
| 22. विनतियों – vinatiyā: | 61. दिशाओं – diśa:ō | 100. ओट – oṭ |
| 23. बात – ba:t | 62. रिमझिम – rimdʒ ^h im | 101. सैयाँ – saiya: |
| 24. रुत – rut | 63. राज – ra:dʒ | 102. आँखियाँ – ā:k ^h iyā: |
| 25. गाल – ga:l | 64. बन – ban | 103. काजल – ka:dʒal |
| 26. रंग – raṅg | 65. डालों – dā:lō | 104. धरा – d ^h ara: |
| 27. आँख – ā:k ^h | 66. शोर – śor | 105. इशारा – iśa:ra: |
| 28. रस – ras | 67. तन – tan | 106. जोबन – dʒoban |
| 29. डाल – dā:l | 68. आनंद – a:naṁd | 107. रेट – reṭ |
| 30. लचक – latʃak | 69. घुमड़े – g ^h umaṛe | 108. माता – ma:ta: |
| 31. पात – pa:t | 70. भावों – b ^h a:vō | 109. चरणों – tʃaraṇō |
| 32. लहक – lahak | 71. बादल – ba:dal | 110. उमरिया – umriya: |
| 33. फूल – p ^h u:l | 72. मोती – moti: | 111. नगरिया – nagariya: |
| 34. महक – mahak | 73. नींद – ni:ṁd | 112. गगन – gagan |
| 35. बलमा – balma: | 74. सपने – sapne | 113. सोना – sona: |
| 36. कली – kali: | 75. चितचोर – tʃittʃor | 114. अटरिया – aṭariya: |
| 37. चंचल – tʃaṁtʃal | 76. सितारो – sita:ro | 115. रहमत – rahmat |
| 38. रूप – rup | 77. मान – ma:n | 116. बदरिया – badriya: |
| 39. दामिन – da:mani | 78. रातों – ra:tō | 117. पाँव – pā:v |

118. वादों – va:dō
 119. डगरिया – ḍagariya:
 120. धाम – dh'a:m
 121. गरीबों – ḡari:bō
 122. द्वार – dva:r
 123. नैनों – nainō
 124. नैना – naina:
 125. लाज – la:dʒ
 126. मारी – ma:ri:
 127. प्रीत – pri:t
 128. घूँघट – ḡhũ:ḡhaṭ
 129. बलमवा – balamva:
 130. हार – ha:r
 131. तूफ़ान – tu:fa:n
 132. . प्राण – pra:ṇ
 133. नस – nas
 134. खुमार – xuma:r
 135. नीर – ni:r
 136. हथीले – haṭhi:le
 137. नैया – naiya:
 138. भँवर – bhāvar
 139. पार – pa:r
 140. छोरे – tʃhore
 141. गली – gali:
 142. छोरी – tʃhori:
 143. रोज़ – roz
 144. मुलाकात – mula:qa:t
 145. प्यार – pya:r
 146. जुदाई – dʒuda:i:
 147. चौदनी – tʃā:dni:
 148. जिया – dʒiya:
 149. लड़ाई – laṛa:i:
 150. दुहाई – duha:i:
 151. आँखियों – ā:khiyō
 152. चेहरे – tʃehre
 153. गुस्सा – ḡussa:
 154. सफ़ाई – safa:i:
 155. घड़ी – ḡhaṛi:
 156. बातें – ba:tē
 157. क़सम – qasam
 158. बचपन – batʃpan
 159. याद – ya:d
 160. रुम-झुम – rum-dʒhum
 161. सावन – sa:van
 162. बिजली – bidʒli:
 163. बालम – ba:lam
 164. पवन – pavan
 165. हिचकोले – hitʃkole
 166. कागा – ka:ga:
 167. पी – pi:
 168. संदेस – sandes
 169. छम-छम – tʃham-tʃham
 170. गुलशन – ḡulšan
 171. हँसी – hāsi:
 172. खुशी – xuši:
 173. ग़म – ḡam
 174. साज़ – sa:z
 175. आवाज़ – a:va:z
 176. पट – paṭ
 177. पिया – piya:
 178. घट – ḡhaṭ
 179. बचन – batʃan
 180. धन – dhan
 181. गरब – garab
 182. मोल – mol
 183. महल – mahal
 184. मंदिर – maṁdir
 185. दिया – diya:
 186. आसन – a:san
 187. घर – ḡhar
 188. प्रीतम – pri:tam
 189. कल – kal
 190. प्रभू – prabhu:
 191. बार – ba:r
 192. बतियाँ – batiyā:
 193. जग – dʒag
 194. मैना – maina:
 195. कानन – ka:nan
 196. सपनों – sapnō
 197. हाथों – ha:tʃhō
 198. कंगन – kaṁgan
 199. पैरों – pairō
 200. पायल – pa:yal
 201. गले – gale
 202. सिंगार – siṁga:r
 203. सखि – sakhi
 204. कर – kar
 205. जोगिन – dʒogin
 206. भगत – bhagat
 207. कमर – kamar
 208. बेड़ा – beṛa:
 209. संग्राम – saṁgra:m
 210. हिम्मत – himmat
 211. काम – ka:m
 212. जीत – dʒi:t
 213. दाम – da:m
 214. जाते – dʒa:te
 215. तोप – top
 216. तीर – ti:r
 217. तलवार – talva:r
 218. वादा – va:da:
 219. रतियाँ – ratiyā:
 220. बिरहा – biraha:
 221. खुदकुशी – xudkuši:
 222. इलज़ाम – ilza:m
 223. गला – gala:
 224. दस्त – dast
 225. छुरी – tʃhuri:
 226. छींटे – tʃhi:mte
 227. शोले – šole
 228. जुल्म – zulm
 229. करवटें – karavaṭē
 230. जूता – dʒu:ta:
 231. पतलून – patlu:n
 232. सर – sar
 233. टोपी – ṭopi:
 234. सड़क – saṛak
 235. सीना – si:na:
 236. मंज़िल – maṁzil
 237. दरिया – dariya:
 238. किनारे – kina:re
 239. राह – ra:h
 240. वतन – vatan
 241. कहानी – kaha:ni:
 242. मौत – maut
 243. निशानी – niša:ni:
 244. राजे – ra:dʒe
 245. राजकुंवर –
 ra:dʒkuṁva:r

246. शहज़ादे – śahza:de
 247. सिंघासन – siṅg^ha:saṇ
 248. इरादे – ira:de
 249. सूरत – su:rat
 250. दौरे – daure
 251. मानवा – ma:nva:
 252. साँवरिया – sā:variya:
 253. गगरिया – gagariya:
 254. कारण – ka:raṇ
 255. छलिया – tʃ^haliya:
 256. मुरलिया – muraliya:
 257. इनक – dʒ^hanak
 258. पायलिया – pa:yaliya:
 259. रूनाक – ru:na:q
 260. झुनाक – dʒ^huna:k
 261. झंकार – dʒ^haṃka:r
 262. धरती – d^harti:
 263. संसार – saṃsa:r
 264. राग – ra:g
 265. सज़ा – saza:
 266. दुनिया – duniya:
 267. रो-रो – ro-ro
 268. हालत – ha:lat
 269. दम – dam
 270. आग – a:g
 271. सलाम – sala:m
 272. पयाम – paya:m
 273. उलफ़त – ulfat
 274. ज़ाम – dʒa:m
 275. हवाओं – hava:õ
 276. छाँव – tʃ^hã:v
 277. गाँव – gã:v
 278. ज़माँन – zami:n
 279. यक़ीन – yaqi:n
 280. उड़ान – uṛa:n
 281. पलकों – palakõ
 282. जान – dʒa:n
 283. तबीयत – tabi:yat
 284. नज़रें – nazarẽ
 285. हाल – ha:l
 286. मलाल – mala:l
 287. हसरत – hasrat
 288. प्यारे – pya:re
 289. क़िस्मत – qismat
 290. दिल्लगी – dillagi:
 291. ख़बर – xabar
 292. ज़ालिम – za:lim
 293. आहट – a:haṭ
 294. बिजलियाँ – bidʒliyã:
 295. अदाओं – ada:õ
 296. उजाला – udʒa:la:
 297. हसीनों – hasi:nõ
 298. गुलाम – ġula:m
 299. मुखड़े – muk^hṛe
 300. होंठों – hoṃṭ^hõ
 301. बंदर – baṃdar
 302. शहर – śahr
 303. बनमानुष –
 banama:nuṣa
 304. चाचा – tʃa:tʃa:
 305. भतिजा – b^hatidʒa:
 306. बाप – ba:p
 307. भाई – b^ha:i:
 308. कान – ka:n
 309. दादा – da:da:
 310. संतान – saṃta:n
 311. रिश्ता – riśta:
 312. नाक – na:k
 313. दूरी – du:ri:
 314. दाता – da:ta:
 315. दया – daya:
 316. देस – des
 317. बादलों – ba:dalõ
 318. सितारा – sita:ra:
 319. शरण – śaraṇ
 320. भक्त – b^hakt
 321. माला – ma:la:
 322. कमरिया – kamariya:
 323. बालमा – ba:lma:
 324. नैन – nain
 325. मैया – maiya:
 326. आँचल – â:tʃal
 327. पनघट – pang^hṭ
 328. कलाईयाँ – kala:iyã:
 329. कंकड़ी – kaṅkaṛi:
 330. साड़ी – sa:ṛi:
 331. जियरा – dʒiyrã:
 332. घूँघटा – ġ^hũ:ġ^haṭa:
 333. सखी – sak^hi:
 334. मेघा – meġ^ha:
 335. जियरवा – dʒiyrã:va:
 336. मीन – mi:n
 337. पल – pal
 338. छिन – tʃ^hin
 339. नैनन – nainan
 340. सजरिया – sadʒariya:
 341. अगन – agan
 342. नगरी – nagari:
 343. फुहार – p^huha:r
 344. पिचकारियों –
 pitʃka:riyõ
 345. होली – holi:
 346. मस्ती – masti:
 347. आशाओं – a:śa:õ
 348. दामन – da:man
 349. सजना – sadʒana:
 350. साजन – sa:dʒan
 351. अंग – aṃg
 352. पिचकारी – pitʃka:ri:
 353. पकड़ – pakaṭ
 354. अनाड़ी – ana:ṛi:
 355. धड़कन – d^haṭkan
 356. रँग-रूप – raṃg-ru:p
 357. चमन – tʃaman
 358. हुस्न – husn
 359. इश्क़ – iśq
 360. मुक़ाम – muqa:m
 361. शाम – śa:m
 362. शबाब – śaba:b
 363. शेर – śer
 364. अह्ल – ahl
 365. इल्म – ilm
 366. नगर – nagar
 367. मंज़िलों – maṃzilõ
 368. गोद – god
 369. रह-गुज़र – rah-guʒar
 370. नज़र – nazar
 371. कलि – kali
 372. नाज़नी – na:zani:m

373. रवायते – rava:yaṭē
 374. अदब – adab
 375. शाहकार – śa:hka:r
 376. अमीर – ami:r
 377. गरीब – ġari:b
 378. जां-निसार – dʒa:n-nisa:r
 379. शाख – śa:x
 380. बुलबुलों – bulbulō
 381. चह-चहें – tʃah-tʃahē
 382. जिंदगी – zindagi:
 383. क़दम – qadam
 384. नज़ारा – naza:ra:
 385. दोस्त – dost
 386. मोहब्बतों – mohabbatō
 387. उम्र – umr
 388. आन – a:n
 389. शान – śa:n
 390. दोस्तों – dostō
 391. नाम – na:m
 392. घराना – ġhara:na:
 393. जहान – dʒaha:n
 394. कविराज – kavi:ra:dʒ
 395. ताज – ta:dʒ
 396. राजघराना – ra:dʒġhara:na:
 397. खज़ाना – xaza:na:
 398. धूल – dhu:l
 399. सफ़र – safar
 400. मेला – mela:
 401. मान – ma:n
 402. अभिमान – abhima:n
 403. मेहमान – mehma:n
 404. डफ़ली – dʌfali:
 405. तराना – tara:na:
 406. शीशा – śi:śa:
 407. हवाएँ – hava:ē
 408. मौजों – maudʒō
 409. अदाएँ – ada:ē
 410. जुल्फ़ – zulf
 411. फ़िज़ाए – fiza:e
 412. निगाहे – niga:he
 413. नज़ारे – naza:re
 414. कुरबान – qurba:n
 415. नज़ारों – naza:rō
 416. लहरें – lahrē
 417. इशारों – iśa:rō
 418. अदा – ada:
 419. धारों – dha:rō
 420. बहारों – baha:rō
 421. मछलियाँ – matʃhliyā:
 422. किनारों – kina:rō
 423. होश – hoś
 424. खयाल – xya:l
 425. ख़्वाब – xva:b
 426. शाखों – śakhō
 427. मुद्दत – muddat
 428. सदियों – sadiyō
 429. शनासाई – śina:sa:i:
 430. झरनों – dʒharnō
 431. जज़्बात – dʒazba:t
 432. क़दमों – qadamō
 433. तले – tale
 434. बर्फ़ – barf
 435. फ़र्श – farś
 436. गरमी – ġarmi:
 437. संगमरमर – samġmarmar
 438. तरह – tarah
 439. बदन – badan
 440. नरमी – narmi:
 441. पाटों – pa:tō
 442. खयालात – xya:la:t
 443. करवट – karavaṭ
 444. हया – haya:
 445. कोहरे – kohre
 446. वादी – va:di:
 447. ज़रुरत – zaru:rat
 448. धुंध – dhu:ndh
 449. चादर – tʃa:dar
 450. परदेसियों – pardesiyō
 451. फ़साना – fasa:na:
 452. पंछी – paṁtʃhi:
 453. ठिकाना – tʃhika:na:
 454. बाग़ों – ba:ġō
 455. हरजाई – hardʒa:i:
 456. पतझड़ – patdʒhʌṛ
 457. महबूबा – mahbu:ba:
 458. क़स्में – qasmē
 459. रस्में – rasmē
 460. शिकवे – śikve
 461. वादे – va:de
 462. खयालों – xya:lō
 463. खुशियाँ – xuśiyā:
 464. ख़्वाब – xva:b
 465. महलों – mahalō
 466. राहगुजारों – ra:hġuza:rō
 467. फ़िज़ा – fiza:
 468. भेष – bheṣ
 469. पते – patte
 470. चनारों – tʃana:rō
 471. राहें – ra:hē
 472. झोंका – dʒhōmka:
 473. हवा – hava:
 474. गुमान – ġuma:n
 475. झील – dʒhi:l
 476. मंज़र – maṁzar
 477. किरणों – kiranō
 478. बरसातें – barsa:tē
 479. पहरों – paharō
 480. दर्पण – darpaṇ
 481. रातें – ra:tē
 482. हमराही – hamra:hi:
 483. बाँह – bā:h
 484. गर्दिश – ġardiś
 485. दूरियाँ – du:riyā:
 486. मजबूरी – madʒbu:ri:
 487. बहाना – baha:na:
 488. इंतज़ार – intaza:r
 489. हिसाब – hisa:b
 490. आँखियों – ākhiyō
 491. जवाब – dʒava:b
 492. बाँहों – bā:hō
 493. जंजीरों – zaṁdʒi:rō
 494. कलाई – kala:i:
 495. मौक़े – mauqe
 496. वफ़ा – vafa:
 497. क़दर – qadr
 498. अंगारे – aṅga:re
 499. शहद – śahd
 500. शर्म – śarm

Appendix 2

Supplementary transcription guide (McGregor Dictionary based variant)

- | | | |
|--|-----------------------------------|------------------------------------|
| 1. भगवान – b ^h agvān | 40. लगन – lagan | 79. फूलों – p ^h ūlō |
| 2. दुखी – duk ^h ī | 41. सवेरे – savere | 80. खुशबू – xušbū |
| 3. दान – dān | 42. रूप – rūp | 81. कलियों – kaliyō |
| 4. भूखे – b ^h ūk ^h e | 43. चाँद – cāḍ | 82. मोहब्बत – mohabbat |
| 5. अन्न – ann | 44. मन्न – mann | 83. तारों – tārō |
| 6. बुरे – bure | 45. बेकली – bekalī | 84. चेहरा – cehrā |
| 7. बदला – badlā | 46. तरहदार – tarahdār | 85. रात – rāt |
| 8. भला – b ^h alā | 47. दिलरुबा – dilrubā | 86. दिन – din |
| 9. दास – dās | 48. दिलदार – dildār | 87. दिलों – dilō |
| 10. वरदान – vardān | 49. चितवन – citvan | 88. ज़माना – zamānā |
| 11. लाख – lāk ^h | 50. यार – yār | 89. राजा – rājā |
| 12. बंसी – baṁsī | 51. धोखा – d ^h okā | 90. कोयल – koyal |
| 13. नाच – nāc | 52. अय्यर – ayyār | 91. खेतों – k ^h etō |
| 14. मन – man | 53. आँखों – āk ^h ō | 92. जवानी – javānī |
| 15. स्वर – svar | 54. जादू – jādū | 93. रानी – rānī |
| 16. बिष – biṣ | 55. बातों – bātō | 94. धीर – d ^h īr |
| 17. प्रेम – prem | 56. शोला – śolā | 95. रिया – riyā |
| 18. गीत – gīt | 57. चरका – carkā | 96. आरी – ārī |
| 19. तट – taṭ | 58. मोर – mor | 97. चुनरिया – cunriyā |
| 20. हाथ – hāth | 59. जीवन – jīvan | 98. नज़रिया – nazariyā |
| 21. जोड़ – joṛ | 60. छाया – c ^h āyā | 99. अम्बुआ – ambuā |
| 22. विनतियाँ – vinatiyā | 61. दिशाओं – diśāō | 100. ओट – oṭ |
| 23. बात – bāt | 62. रिमझिम – rimj ^h im | 101. सैयाँ – saiyyā |
| 24. रुत – rut | 63. राज – rāj | 102. आँखियाँ – āk ^h iyā |
| 25. गाल – gāl | 64. बन – ban | 103. काजल – kājal |
| 26. रंग – raṁg | 65. डालों – ḍālō | 104. धरा – d ^h arā |
| 27. आँख – āk ^h | 66. शोर – śor | 105. इशारा – iśārā |
| 28. रस – ras | 67. तन – tan | 106. जोबन – joban |
| 29. डाल – ḍāl | 68. आनंद – ānaṁd | 107. रेट – reṭ |
| 30. लचक – lacak | 69. घुमड़े – g ^h umare | 108. माता – mātā |
| 31. पात – pāt | 70. भावों – b ^h āvō | 109. चरणों – caranō |
| 32. लहक – lahak | 71. बादल – bādal | 110. उमरिया – umriyā |
| 33. फूल – p ^h ūl | 72. मोती – motī | 111. नगरिया – nagariyā |
| 34. महक – mahak | 73. नींद – nīṁd | 112. गगन – gagan |
| 35. बलमा – balmā | 74. सपने – sapne | 113. सोना – sonā |
| 36. कली – kalī | 75. चितचोर – citcor | 114. अटरिया – aṭariyā |
| 37. चंचल – caṁcal | 76. सितारो – sitāro | 115. रहमत – rahmat |
| 38. रुप – rup | 77. मान – mān | 116. बदरिया – badriyā |
| 39. दामिन – dāmani | 78. रातों – rātō | 117. मंज़िल – manzil |

118. पाँव – pāv
 119. वादों – vādō
 120. डगरिया – ḍagariyā
 121. धाम – dhām
 122. गरीबों – ḡarībō
 123. द्वार – dvār
 124. नैनों – nainō
 125. नैना – nainā
 126. लाज – lāj
 127. मारी – mārī
 128. प्रीत – prīt
 129. घूँघट – ḡhūḡḥaṭ
 130. बलमवा – balamvā
 131. हार – hār
 132. तूफ़ान – tūfān
 133. प्राण – prāṇ
 134. नस – nas
 135. खुमार – xumār
 136. नीर – nīr
 137. हथीले – haṭhīle
 138. नैया – naiyā
 139. भँवर – bhāvar
 140. पार – pār
 141. छोरे – cḥore
 142. गली – galī
 143. छोरी – cḥorī
 144. रोज़ – roz
 145. मुलाक़ात – mulāqāt
 146. प्यार – pyār
 147. जुदाई – judāi
 148. चौदनी – cādnī
 149. जिया – jiyā
 150. लड़ाई – laṛāi
 151. दुहाई – duhāi
 152. आँखियों – āṅḥiyō
 153. चेहरे – cehre
 154. गुस्सा – ḡussā
 155. सफ़ाई – safāi
 156. घड़ी – ḡḥarī
 157. बातें – bātē
 158. क़सम – qasam
 159. बचपन – bacpan
 160. याद – yād
 161. रुम-झुम – rum-jḥum
 162. सावन – sāvan
 163. बिजली – bijlī
 164. बालम – bālam
 165. पवन – pavan
 166. हिचकोले – hickole
 167. कागा – kāgā
 168. पी – pī
 169. संदेस – sandes
 170. छम-छम – cḥam-cḥam
 171. गुलशन – ḡulśan
 172. हँसी – hāsī
 173. खुशी – xuśī
 174. ग़म – ḡam
 175. साज़ – sāz
 176. आवाज़ – āvāz
 177. पट – paṭ
 178. पिया – piyā
 179. घट – ḡḥaṭ
 180. बचन – bacan
 181. धन – dhān
 182. गरब – garab
 183. मोल – mol
 184. रंग – raṅg
 185. महल – mahal
 186. मंदिर – maṁdir
 187. दिया – diyā
 188. आसन – āsan
 189. घर – ḡḥar
 190. प्रीतम – prītam
 191. कल – kal
 192. प्रभू – prabhū
 193. बार – bār
 194. बतियाँ – batiyā
 195. जग – jag
 196. मैना – mainā
 197. कानन – kānan
 198. सपनों – sapnō
 199. हाथों – hāṭhō
 200. कंगन – kaṁgan
 201. पैरों – pairō
 202. पायल – pāyal
 203. गले – gale
 204. सिंगार – siṁgār
 205. सखि – sakḥi
 206. कर – kar
 207. जोगिन – jogin
 208. भगत – bhagat
 209. कमर – kamar
 210. बेड़ा – beṛā
 211. संग्राम – saṁgrām
 212. हिम्मत – himmat
 213. काम – kām
 214. जीत – jīt
 215. दाम – dām
 216. जाते – jāte
 217. तोप – top
 218. तीर – tīr
 219. तलवार – talvār
 220. वादा – vādā
 221. रतियाँ – ratiyā
 222. बिरहा – birahā
 223. खुदकुशी – xudkuśī
 224. इलज़ाम – ilzām
 225. गला – galā
 226. दस्त – dast
 227. छुरी – cḥurī
 228. छींटे – cḥīṁṭe
 229. शोले – śole
 230. जुल्म – zulm
 231. करवटें – karavaṭē
 232. जूता – jūtā
 233. पतलून – patlūn
 234. सर – sar
 235. टोपी – ṭopī
 236. सड़क – saṛak
 237. सीना – sīnā
 238. मंज़िल – maṁzil
 239. दरिया – dariyā
 240. किनारे – kināre
 241. राह – rāh
 242. वतन – vatan
 243. कहानी – kahānī
 244. मौत – maut
 245. निशानी – niśānī

246. राजे – rāje
 247. राजकुंवर – rājkuṁvār
 248. शहजादे – śahzāde
 249. सिंघासन – siṅghāsana
 250. इरादे – irāde
 251. सूरत – sūrat
 252. दौरे – daure
 253. मानवा – mānvā
 254. साँवरिया – sāvariyā
 255. गगरिया – gagariyā
 256. कारण – kāraṇ
 257. छलिया – cʰaliyā
 258. मुरलिया – muraliyā
 259. झनक – jʰanak
 260. पायलिया – pāyaliyā
 261. रूनाक – rūnāq
 262. झुनाक – jʰunāk
 263. झंकार – jʰamkār
 264. धरती – dʰartī
 265. संसार – saṁsār
 266. राग – rāg
 267. सज़ा – sazā
 268. दुनिया – duniyā
 269. रो-रो – ro-ro
 270. हालत – hālat
 271. दम – dam
 272. आग – āg
 273. सलाम – salām
 274. पयाम – payām
 275. उलफ़त – ulfat
 276. जाम – jāṁ
 277. हवाओं – havāō
 278. छाँव – cʰāṁv
 279. गाँव – gāṁv
 280. ज़माँन – zamān
 281. यक़ीन – yaqīn
 282. उड़ान – uḍān
 283. पलकों – palakō
 284. जान – jān
 285. तबीयत – tabīyat
 286. नज़रें – nazarē
 287. हाल – hāl
 288. मलाल – malāl
 289. हसरत – hasrat
 290. प्यारे – pyāre
 291. क़िस्मत – qismat
 292. दिल्लगी – dillagī
 293. ख़बर – xabar
 294. ज़ालिम – zālim
 295. आहट – āhaṭ
 296. बिजलियाँ – bijliyā
 297. अदाओं – adāō
 298. उजाला – ujālā
 299. हसीनों – hasīnō
 300. गुलाम – gūlām
 301. मुखड़े – mukʰṛe
 302. होंठों – hoṁṭʰō
 303. बंदर – baṁdar
 304. शहर – śahr
 305. बनमानुष –
 banamānuṣa
 306. चाचा – cācā
 307. भतिजा – bʰatijā
 308. बाप – bāp
 309. भाई – bʰāī
 310. कान – kān
 311. दादा – dādā
 312. संतान – saṁtān
 313. रिश्ता – riśtā
 314. नाक – nāk
 315. दूरी – dūri
 316. दाता – dātā
 317. दया – dayā
 318. देस – des
 319. बादलों – bādalō
 320. सितारा – sitārā
 321. शरण – śaraṇ
 322. भगवन – bʰagvān
 323. भक्त – bʰakt
 324. माला – mālā
 325. कमरिया – kamariyā
 326. बालमा – bālmā
 327. नैन – nain
 328. मैया – maiyā
 329. आँचल – ācal
 330. पनघट – pangʰaṭ
 331. कलाईयाँ – kalāiyā
 332. कंकड़ी – kaṅkaṛī
 333. साड़ी – sāṛī
 334. जियरा – jiyṛā
 335. घूँघटा – gʰū: gʰaṭā
 336. सखी – sakʰī
 337. मेघा – megʰā
 338. जियरवा – jiyṛavā
 339. मीन – mīn
 340. पल – pal
 341. छिन – cʰin
 342. नैनन – nainan
 343. सजरिया – sajariyā
 344. अगन – agan
 345. नगरी – nagarī
 346. फुहार – fuʰhār
 347. पिचकारियों – pickāriyō
 348. होली – holī
 349. मस्ती – mastī
 350. आशाओं – āśāō
 351. दामन – dāman
 352. सजना – sajanā
 353. साजन – sājan
 354. अंग – aṁg
 355. पिचकारी – pickārī
 356. पकड़ – pakar
 357. अनाड़ी – anārī
 358. धड़कन – dʰaṛkan
 359. रँग-रूप – raṁg-rūp
 360. चमन – caman
 361. हुस्न – husn
 362. इश्क – iśq
 363. मुक़ाम – muqām
 364. शाम – śām
 365. शबाब – śabāb
 366. शेर – śer
 367. अह्ल – ahl
 368. इल्म – ilm
 369. नगर – nagar
 370. मंज़िलों – maṁzilō
 371. गोद – god
 372. रह-गुज़र – rah-guzar
 373. नज़र – nazar

374. कलि – kali
 375. नाज़नी – nāzanīm
 376. रवायते – ravāyatē
 377. अदब – adab
 378. शाहकार – śāhkār
 379. अमीर – amīr
 380. गरीब – ḡarīb
 381. जां-निसार – jān-nisār
 382. शाख – śāx
 383. बुलबुलों – bulbulō
 384. चह-चहें – cah-cahē
 385. ज़िंदगी – zindagī
 386. क़दम – qadam
 387. नज़ारा – nazārā
 388. दोस्त – dost
 389. मोहब्बतों – mohabbatō
 390. उम्र – umr
 391. आन – ān
 392. शान – śān
 393. दोस्तों – dostō
 394. नाम – nām
 395. घराना – ḡharānā
 396. जहान – jahān
 397. कविराज – kavīrāj
 398. ताज – tāj
 399. राज्यराना – rājḡharānā
 400. खज़ाना – xazānā
 401. धूल – d^hūl
 402. सफ़र – safar
 403. मेला – melā
 404. मान – mām
 405. अभिमान – ab^himām
 406. मेहमान – mehmām
 407. डफ़ली – ḡafalī
 408. तराना – tarānā
 409. शीशा – śīśā
 410. हवाएं – havāē
 411. मौजों – maujō
 412. अदाएं – adāē
 413. जुल्फ़ – zulf
 414. फ़िज़ाए – fizāe
 415. निगाहे – nigāhe
 416. नज़ारे – nazāre
 417. कुरबान – qurbān
 418. नज़ारों – nazārō
 419. लहरें – lahrē
 420. इशारों – isārō
 421. अदा – adā
 422. धारों – d^hārō
 423. बहारों – bahārō
 424. मछलियाँ – mac^hliyā
 425. किनारों – kinārō
 426. होश – hoś
 427. खयाल – xyāl
 428. ख़्वाब – xvāb
 429. शाखों – śak^hō
 430. मुद्दत – muddat
 431. सदियों – sadiyō
 432. शनासाई – śināsāi
 433. झरनों – j^harnō
 434. जज़्बात – jazbāt
 435. क़दमों – qadamō
 436. तले – tale
 437. बर्फ़ – barf
 438. फ़र्श – farś
 439. गरमी – ḡarmī
 440. संगमरमर – samḡmarmar
 441. तरह – tarah
 442. बदन – badan
 443. नरमी – narmī
 444. पाटों – pātō
 445. खयालात – xyālāt
 446. करवट – karavaṭ
 447. हया – hayā
 448. कोहरे – kohre
 449. वादी – vādī
 450. ज़रूरत – zarūrat
 451. धुंध – d^hund^h
 452. चादर – cādar
 453. परदेसियों – pardesiyō
 454. फ़साना – fasānā
 455. पंछी – paṃc^h
 456. ठिकाना – ṭ^hikānā
 457. बाग़ों – bāḡō
 458. हरजाई – harjāi
 459. पतझड़ – patj^har
 460. महबूबा – mahbūbā
 461. क़समें – qasmē
 462. रस्में – rasmē
 463. शिकवे – śikve
 464. वादे – vāde
 465. खयालों – xyālō
 466. खुशियाँ – xuśiyā
 467. ख़्वाब – xvāb
 468. महलों – mahalō
 469. राहगुजारों – rāhguzārō
 470. फ़िज़ा – fizā
 471. भेष – b^heṣ
 472. पत्ते – patte
 473. चनारों – canārō
 474. राहें – rāhē
 475. झोंका – j^homkā
 476. हवा – havā
 477. गुमान – ḡumām
 478. झील – j^hīl
 479. मंज़र – maṃzar
 480. किरणों – kiraṇō
 481. बरसाते – barsātē
 482. पहरों – paharō
 483. दर्पण – darpaṇ
 484. रातें – rātē
 485. हमराही – hamrāhī
 486. बाँह – bāh
 487. गर्दिश – gardiś
 488. दूरियाँ – dūriyā
 489. मजबूरी – majbūrī
 490. बहाना – bahānā
 491. इंतज़ार – intazār
 492. हिसाब – hisāb
 493. आँखियों – āk^hiyō
 494. जवाब – javāb
 495. बाँहों – bāhō
 496. जंजीरों – zaṃjīrō
 497. कलाई – kalāi
 498. मौक़े – mauqe
 499. वफ़ा – vafā
 500. क़दर – qadr