Scripting History: A Diachronic Urdu Text and Image Corpus from the 18th to 19th Centuries

Sana Shams[1], Sahar Rauf[1], Asad Mustafa[1], Muhammad Zeeshan Javed[1], Qurat-ul-Ain Akram[2], Sarmad Hussain[1], Miriam Butt[3]

[1]Center for Language Engineering (CLE), Al-Khawarizmi Institute of Computer Science (KICS), University of Engineering and Technology (UET), Lahore, Pakistan
[2]Department of Computer Science, UET, Lahore, Pakistan
[3]Department of Linguistics, University of Konstanz, Konstanz, Germany
{sana.shams, sahar.rauf, asad.mustafa, zeeshan.javed, sarmad.hussain} @kics.edu.pk
ainie.akram@uet.edu.pk, miriam.butt@uni-konstanz.de

Abstract

This paper presents the Diachronic Urdu Text and Image Corpus, a one-million-word resource covering Urdu's development across the 18th and 19th centuries. The corpus is compiled from 328 printed books published between 1800 and 1950, representing a diverse range of genres, authors, and publishers. A 140,000-word sub-corpus has been manually annotated with Urdu part-of-speech tags to facilitate linguistic and computational analysis. The dataset enables systematic investigation of historical changes in Urdu orthography, morphology, and syntax, providing new insights into the language's history and standardization. To preserve the original printed form, each text is paired with its corresponding page image, creating the first multimodal diachronic corpus for Urdu. The paper outlines the corpus compilation pipeline, digitization workflow, text-image alignment, and annotation strategy designed to ensure accuracy, consistency, and authenticity. This multimodal Urdu diachronic corpus establishes a benchmark for research in computational linguistics, digital humanities, and South Asian language technology, supporting corpus-based exploration of Urdu's linguistic history and cultural heritage.

Keywords: Diachronic corpora, Corpus design, Corpus digitization, Part of Speech tagging

1. Introduction

Urdu, an Indo-Aryan language, emerged from the Hindi/Hindustani spoken in North India, particularly in and around the Delhi region (Rehman, 2011). From approximately the 15th to 18th centuries, Urdu underwent a process by which it transitioned into a Persianized register of Hindustani (King, 1995; Rehman, 2011). In the 18th century the Perseo-Arabic script variant of Urdu was standardized though prominent Urdu writers persisted in referring to it as Hindi. In 1850, Urdu was officially recognized as the language of authority in British India (Rai, 1991). Subsequently, from 1850 onward, Urdu witnessed significant alterations in its morphology, syntax, and orthography, culminating in its standardization in 1950 as the official language of Pakistan. This standardization involved the adoption of a uniform publishing character set and grammar. In this paper we present a first linguistically annotated diachronic corpus of Urdu that can be used to trace and analyze the morphological, syntactic, semantic and orthographic changes that have shaped Urdu's diachronic trajectory over the centuries.

In what follows, we provide information on the design, development and digitization of this first diachronic Urdu corpus in text and image formats, covering a chronological span of 150 years—from 1800 to 1950. It encompasses multiple genres, including history, religion, science, literature. The text corpus consists of one million words extracted from 3439 pages of 328 books. We have additionally constructed a parallel image corpus, which includes 3439 images of these pages in jpeg format, aligning seamlessly with the text corpus. Furthermore, 140,000 words from this one million diachronic text corpus have already been annotated with Urdu Part-of-Speech (POS) tags. The result is a first multimodal diachronic corpus for Urdu that includes linguistic annotations.

The paper is organized as follows. Section 2 presents existing studies related to diachronic corpora. Section 3 provides details about the design and development of our corpus. Section 4 describes the corpus digitization process and the challenges encountered during this process. Section 5 details the annotation process at the POS level. Finally, Section 6 showcases the historical changes found for Urdu orthography, and Section 7 concludes the paper and presents future work.

2. Related Work

Extensive research on the design, development and digitization of diachronic text and image corpora has been reported in literature. A good collection of historical corpora exists for high-resourced languages like English, with around thirty to forty English historical corpora currently available or under development (Osório and Lopes Cardoso, 2024), totaling over 630 million words (Kytö, 2011). Among these, the largest is the Corpus of Historical American English (COHA), a 400-million-word resource widely used

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1622–1632
11–16 May 2026.

for studying lexical, syntactic, and semantic changes (Davies, 2012). Developed by Brigham Young University, COHA comprises carefully selected historical texts from newspapers, magazines, and both fiction and nonfiction books published between 1810 and 2009. Other prominent resources include: the British National Corpus (2007), a 100-million-word collection of late 20th century British English (1991–1994); ARCHER (2014), which spans 1650–1999 and includes both British and American English; the Helsinki Corpus of English Texts (1993), a diachronic corpus covering Old English to the early 18th century. Additional corpora include the 100-million-word Time Magazine Corpus (Davies, 2007) and the 52-million-word Old Bailey Corpus (2016). The Penn Corpora of Historical English (Kroch, 2020) contains 4.4 million tokens of British English prose from the earliest Middle English period up to World War I.

Beyond English, historical corpora have also been developed for other languages. A large-scale historical Arabic corpus (Belinkov et al., 2016) has been developed using texts from the Al-Maktaba Al-Shamela website, covering the 7th century to the modern era. The corpus comprises over 6,000 texts (around 1 billion words, with 800 million words from diachronic texts). A semi-automatic process was used for text cleaning and metadata organization, and the corpus was lemmatized to facilitate semantic analysis, addressing the complexities of Arabic morphology. Similarly, a multi-institutional effort led to the development of OpenITI, a large-scale diachronic Arabic corpus comprising approximately 1.5 billion words (Belinkov et al., 2019). For Persian, a literary text corpus has been developed, incorporating poems from the 9th to the 21st century as well as two main collections of myths and stories with 9.1 million tokens (Raji et al., 2024). For Panjabi/Gurmukhi NLP research, Kang et al. (2024) compiled a corpus of 242.4 million words across 30.1 million sentences, covering 261,599 documents from the 14th to the 21st century, with a vocabulary size of 1.18 million unique words. Historical corpora have also been developed for Hindi and Marathi, comprising 1.02 million and 2.11 million unique words, respectively, with word frequency distributions spanning eleven decades (pre-1920, 1920 to 2020). Encompassing a diverse range of textual sources, including fiction, non-fiction, historical writings, autobiographies, and periodicals, these corpora offer valuable insights into the diachrony of both languages (Belhekar and Bhargava, 2023).

Historical image corpora also exist. These have been developed by creating high resolution images of the diachronic text materials e.g., books, records, and newspapers. These images are then annotated at multiple levels to conduct research focused on document layout analysis, automatic recognition of printed or handwritten text, writer identification, etc. The IMPACT (Papadopoulos et al., 2013) dataset is an extensive collection of more than 600K images sourced from various European libraries. The dataset includes 80% of documents from the 19th to the early 20th century, 17% from the 17th or 18th centuries and the remaining images are from the 15th century. The PAGE XML files (Pletschacher and Antonacopoulos, 2010) within this dataset contain valuable information such as layout details, reading order, and text transcription annotations for over 45K samples. Additionally, users can access metadata that includes bibliographic information, digitization details, physical characteristics of the documents themselves, copyright information as well as administrative data and comments. This collection is even more impressive in its diversity as it encompasses documents written in eighteen different languages including Bulgarian, Catalan, Czech, Dutch, English, French, German, Greek, Hebrew, Latin, Norwegian, etc. It is important to note that no benchmark results are presented for this particular dataset. However, this comprehensive dataset undoubtedly serves as a valuable resource for researchers working on a wide range of topics requiring visual data analysis or document understanding. The GW database (Fischer et al., 2012), consists of historical documents from the George Washington Papers dating back to the 18th century. The data consists of 656 text and 4894 binarized and normalized word images, alongside their transcription annotations. These pages are written in English using longhand script by two different writers. The dataset also provides statistics such as 1471 word classes and 82 letters. It is widely used for evaluating word spotting algorithms.

The KERTAS dataset (Adam et al., 2018) comprises handwritten Arabic manuscripts spanning fourteen centuries, collected from the Qatar National Library for age and writer detection. It includes 2000 high-resolution images annotated with the Islamic century of authorship and accompanied by XML metadata detailing source, title, description, and writer information. Saad et al. (2016) introduced BCE-Arabic-v1, containing 1833 pages from 180 books, later expanded by Elanwar et al. (2021) to over 9000 scanned images from 700 Arabic books. Of these, 1500 documents were segmented and labeled by layout region and content type. Barakat et al. (2019) presented a medieval Hebrew manuscript dataset with 30 annotated pages verified by historical experts, while Stökl Ben Ezra et al. (2021) compiled 202 pages from 100 Hebrew manuscripts covering diverse scripts and periods, annotated at region, line, and page levels.

With respect to Urdu, there is a noticeable gap in the current state of the art. A variety of linguistic resources have been developed for contemporary Urdu and have been used for studies on modern

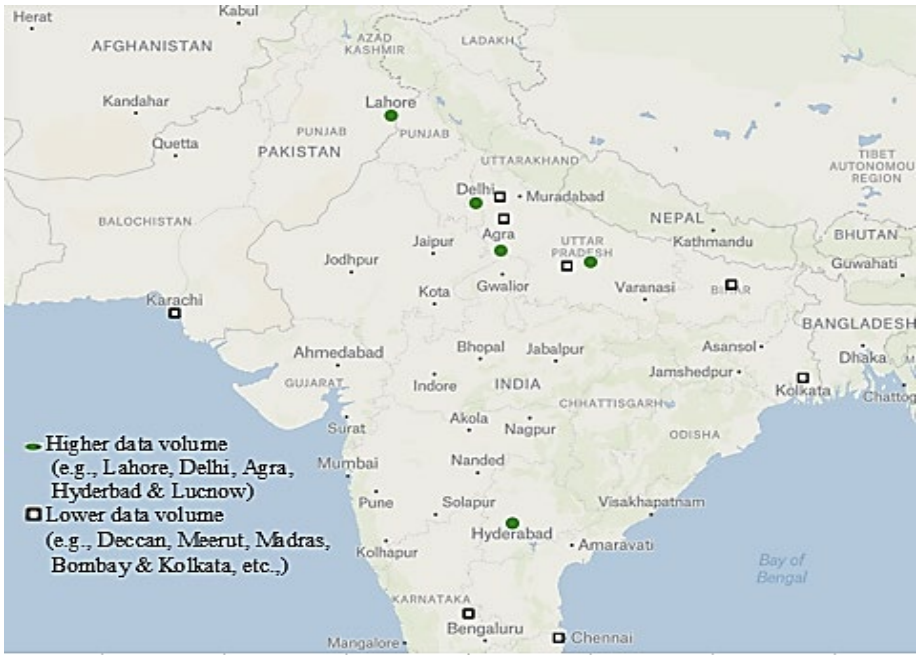

Figure 1: Geographical distribution of data points: A visualization of key cities

Urdu (e.g., Ahmed et al., 2020; Akram and Hussain, 2017; Ehsan and Butt, 2020; Urooj et al., 2021). However, similar resources are not available for earlier stages of Urdu. We thus decided to undertake the effort of developing a first diachronic corpus of Urdu covering 150 years.

3. The Diachronic Urdu Text and Image Corpus (DUTI)

The Diachronic Urdu Text and Image (DUTI) corpus comprises Urdu publications spanning 150 years, from 1800 to 1950, a crucial period for Urdu's emergence as a standardized register distinct from Hindi. Limited access to original materials in libraries and institutions necessitated using online libraries to build the corpus. The final corpus was compiled from 328 different manuscripts including 12 from the World Digital Library,[1] and 49 from the Iqbal Cyber Library.[2] The remaining manuscripts were gathered from Urdu Channel,[3] and the Internet Archive.[4]

To achieve an equitable *chronological representation* within this corpus, the temporal span was systematically segmented into 15 decades (1800-09, 1810-19, etc.). The objective was to ensure an even distribution of texts across these decades, with a focus on publications exhibiting consistent genres and geographic coverage. For this purpose, texts were selected from diverse publishing houses and authors. This corpus ensures a well-distributed coverage of texts from key publishing centers, such as Lucknow, Lahore, Hyderabad, and Delhi, known for their significant contributions to Urdu publishing in the 18th and 19th centuries (Qureshi, 1996). Additionally, texts from other regions in India, including Agra, Meerut, Aligarh, Mumbai, and Pathankot were also incorporated for broader representation. Figure 1 presents a visual representation of the key cities covered.[5]

Achieving comprehensive geographical and genre coverage posed several challenges. Urdu publishing flourished primarily after 1850, resulting in a scarcity of Urdu publications from 1800 to 1850. To address this limitation, more text was included from each available book from the 1800-1850 timeframe. Figure 2 provides an overview of the text distribution within the corpus. The collection also incorporates different handwriting styles and early lithographed texts, enhancing its representativeness for Optical Character Recognition (OCR) applications. Book pages were selected based on high image quality (minimal noise or fading), predominantly Urdu content (excluding dominant Arabic, Persian, or English sections), and minimal non-text elements such as diagrams.

[1]https://www.loc.gov/collections/world-digital-library/about-this-collection/#government-college-university-lahore
[2] http://iqbalcyberlibrary.net/
[3] http://urduchannel.in
[4] https://archive.org/
[5]https://felt.com/map/Locations-88DAsMCEQAO5BAT3B9BPjBA?loc=22.63,77.69,5.01z

Decades	No of Pages	Total Characters	Unique Urdu Characters	Total Ligatures	Unique Ligature
1800-09	172	191,450	41	98,528	4447
1810-19	191	184,066	42	92,284	3751
1820-29	258	223,314	42	111,778	6231
1830-39	276	256,168	42	128,450	6252
1840-49	325	367,384	44	182,685	8361
1850-59	234	231,699	42	118,024	5582
1860-69	241	266,309	42	136,847	5755
1870-79	226	231,990	42	119,337	5701
1880-89	294	280,384	42	141,648	5840
1890-99	187	213,985	42	109,294	5292
1900-09	210	233,569	42	119,778	5768
1910-19	205	225,150	42	116,506	5141
1920-29	202	209,224	42	107,802	4924
1930-39	215	216,982	42	111,507	5018
1940-50	203	214,743	42	110,012	5081

Table 1: Urdu character set and ligature coverage

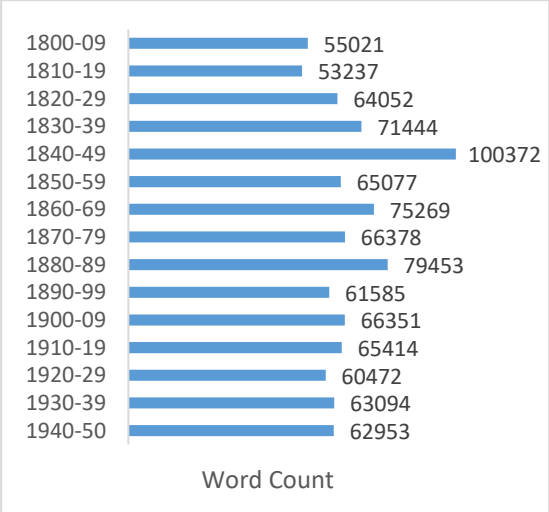

Figure 2: Distribution of texts within the corpus

Regarding genre diversity, the corpus encompasses material from 6 distinct genres. The distribution of the corpus word count percentage by genre is depicted in Figure 3. As illustrated in Figure 3, Science and Religion constitute the majority of diachronic texts, followed by Literature and Language. Content related to History and Philosophy was comparatively limited.

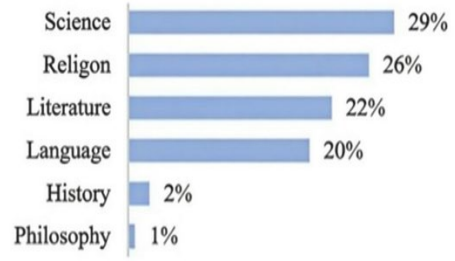

Figure 3: Corpus word count percentage with respect to genre in the DUTI Corpus

The diachronic image corpus consists of 3439 image files along with text transcriptions. Adequate Urdu character and ligature[6] coverage was ensured in the image corpus, as this was necessary for developing a character based Urdu OCR system. Table 1 highlights the coverage of Urdu characters and ligatures in the images and

[6] In cursive scripts like Urdu, characters are joined to form ligatures.

shows that the corpus covers all Urdu characters and a variety of different Urdu ligatures across each decade. The image corpus included 8 different page layouts; i) single column, ii) single column and rectangular enclosed text, iii) stylish page border, iv) comments on page borders, v) text with images, vi) double column text, vii) diagonal and horizontal text alignment, viii) table with text. Sample images pertaining to each layout category are shown in Figure 4 and are quantified in Table 2.

Figure 4: Sample images from each of the eight page layouts

Sr. No	Page layout	Total instances
1	Single column text	2558
2	Single column rectangular enclosed text	346
3	Stylish page border	28
4	Comments on page borders	425
5	Text with images	20
6	Double column text	27
7	Diagonal and horizontal text alignment	33
8	Table with text	2

Table 2: Page layouts in the DUTI Corpus

In addition, metadata for each file has been collected and maintained during corpus compilation. This includes: publication name, author, genre, date, publisher/place of publication, volume (if applicable), compiler/translator (if applicable), number of pages, source of the scanned copy, text quality (original or republished).

4. Corpus Digitization

The text corpus was created through manual transcription of each selected image file in the diachronic dataset. This process resulted in the generation of line-level ground truth for every corresponding image in the corpus.

4.1 Typing Tool

For efficient digitization of the corpus, a custom typing tool was created to streamline the typing process and to ensure consistency in naming for both the image and text corpora. The interface of the typing utility is presented in Figure 5.

Key features of this tool included displaying the corpus image alongside the typing area and saving typed files in UTF-8 format with identical names to the corresponding image file, appended with a typist identifier. Hard copies of the corpus were furnished to the typists for reference in instances where the images were unclear. Typists received brief training on using the tool, practicing on sample corpus pages. The typed content was reviewed by an expert linguist, with revisions provided until 100% accuracy against the reference pages was achieved. Prior to the digitization of the selected corpus, a brief pilot typing analysis was conducted to identify potential transcription challenges. During this process, unidentified Urdu characters were observed in the diachronic text.

Figure 5: Corpus typing tool

4.2 Unidentified Urdu Characters

Detailed examination of the diachronic texts revealed that certain characters written in the historical Urdu texts are absent from the contemporary Urdu character set and therefore also the keyboard we were using. Among these characters, three had corresponding Unicode representations, while four were not included in the Arabic script's Unicode page. To accommodate the former, the Urdu keyboard was enhanced, whereas the latter were represented by typing a specific contemporary Urdu character for each, followed by the special character "$" for identification in the corpus. Figure 6 presents examples of these characters.

Sr. No.	Char.	Unicode	Typed Char.	Example from Diachronic Corpus
1.	ڈ	0690	ڈ	
2.	ي	064A	ي	
3.	ٹ	067F	ٹ	

(a)

Sr. No.	Char.	Unicode	Typed Char.	Example from Diachronic Corpus
4.	ڈ	-	$ڈ	
5.	ے	-	$ے	
6.	ر	-	$ر	
7.	ف	-	$ف	

(b)

Figure 6: Old Urdu characters in diachronic corpus. (a) Characters with an existing valid Unicode were added to the keyboard. (b) Characters that did not have a digital representation, instead a character from the existing Urdu character set appended with $ was used for typing

4.3 Typing Guidelines

To maintain uniformity in the typing process, specific typing guidelines were formulated, as detailed here:

- •Retain original content: Diachronic content must be typed on as-is basis i.e., without altering the spellings or word forms.
- •Addition of a special character at the end of each line: A specific character sequence (پیپ) must be typed at every line break as present in the image file. This additional character sequence was inserted to keep track of line breaks, for exact alignment with the image corpus.
- •Usage of the new keyboard: Typing must be done using the new keyboard.
- •Typing of old characters absent on the keyboard: For characters that are not present in the Urdu keyboard, the specified Urdu character appended with a $ as shown in Figure 6b should be used.

4.4 Typing Review

For quality assurance, the text corpus was double-typed by two Urdu typists. This approach enabled easy detection of inconsistencies by means of file comparisons as shown in Figure 7. Finally, both files were reviewed and discrepancies resolved by an expert linguist.

Figure 7: Highlighted discrepancies between the two same documents typed by two typists

5. POS Annotated Layer

To study diachronic morpho-syntactic properties of Urdu, a 140,000-word sub-corpus was extracted from the one million-word diachronic text corpus for POS annotation. The CLE Urdu POS tagset (Ahmed et al., 2014),[7] containing 35 tags was used for annotation.

5.1 Corpus Extraction

The DUTI corpus consisted of 3439 files. To extract a 140,000-word sub-corpus, 15% of the data was manually selected from every file. Only complete sentences were included. Identifying the sentence boundary was challenging as punctuation marks to indicate phrasal and sentence boundary were used only infrequently in our texts.

[7]https://www.cle.org.pk/software/langproc/POStagset.htm

5.2 Challenges in Diachronic POS Tagging

To bootstrap the tagging procedure, we initially employed the existing CLE POS tagger[8] trained on contemporary Urdu. However, the data annotated in this way required extensive review to rectify tagging errors and to observe diachronic syntactic variations. Specific challenges we identified for POS tagging and stemming are as follows.

Phonological Variation: The following differences were observed between diachronic and contemporary Urdu, with stems and POS tags provided as well:

•Consistent substitution of 'ہ' /h/ with 'ھ' /◌ʰ/ (aspiration) e.g.,

i) Old Urdu: بہاگ bhāg →Modern Urdu: بھاگ bhāg /bʰɑ:g/ (run-VBF)
ii) تہا thā → تھا thā /t̪ʰɑ:/ (was-AUXT)

•Consistent substitution of 'ن' /n/ with 'ں' /◌̃/ (nasalization) e.g.,

i) نہین nahīñ →نہیں nahīñ /nəhĩ:/ (no-NEG)
ii) مین meñ → میں meñ /mæ̃:/ (I-PRP)

Spelling Variation: Systematic differences were observed between diachronic and contemporary spellings. These included the following:

•Omission of diacritic 'ۂ' e.g.,

i) ہوے hue →ہوئے hue /hu:e:/ (happened-AUXA)
ii) پاون pāon →پاؤں pāon /pɑ:õ:/ (foot-NN)

•Addition of 'و' /v/ e.g.

i) سہاونا suhāvnā → سہانا suhānā /sʊhɑ:nɑ:/ (pleasant-NN)
ii) اوڑ uṛ → اڑ uṛ /ʊɽ/ (fly-VBF)

•Addition of 'ہ' /h/ e.g.,

i) یہہ yeh → یہ yeh /je:h/ (this-PDM)
ii) مونہہ munh → منہ munh /mʊ̃h/ (mouth-NN).

For all words in the diachronic texts exhibiting this spelling variation, the POS of its contemporary Urdu word was assigned to the target word.

Orthographic Variation: The following three cases of orthographic variations were observed in the diachronic texts.

•Two words in the diachronic text being written as a single word in contemporary text, e.g.,

i) شاہ shāh /ʃɑ:h/ (wide-JJ) + راہ rāh /rɑ:h/ (passage-NN) → شاہراہ shāhrāh /ʃɑ:hrɑ:h/ (highway-NN)
ii) پس pas /pəs/ (backward-JJ) + پا pā /pɑ:/ (steps-NN) → پسپا paspā /pəspɑ:/ (push back-NN)

In this case, a separate POS label was assigned to each word in the diachronic text.

•A single word in the diachronic text being written as two words in the contemporary language, e.g.,

i) اسوقت iswaqt /ɪsvəqt̪/ (this time-PDM+NN) → اس is /ɪs/ (this-PDM) + وقت waqt /vəqt̪/ (time-NN)
ii) برسکی baraskī /bərəski:/ (of the year-NN+PSP) → برس baras /bərəs/ (year-NN) + کی kī /ki:/ (of-PSP)
iii) رہگیا rehgayā /ræhgæɑ:/ (was left-VBF+AUXA) → رہ reh /ræh/ (left-VBF) گیا gayā /gæɑ:/ (was-AUXA)

•Negation with a prefix ن in certain verbs instead of نہ as done in contemporary Urdu. E.g.,

i) نہوتا nahotā /nəho:t̪ɑ:/ (not happened-NEG+VBF)→ نہ nah /nɑ:/ (not-NEG) + ہوتا hotā /ho:t̪ɑ:/ (happened-VBF)
ii) نمانا namānā /nəmɑ:nɑ:/ (not agree-NEG+VBI) → نہ nah /nɑ:/ (not-NEG) مانا mānā /mɑ:nɑ:/ (agree-VBI)

Due to such orthographic variation new Urdu words were identified that the POS tagger for contemporary Urdu had not been aware of. Consequently, to assign POS tags to these compound words, novel POS-tag combinations were formulated, see Section 5.3.

5.3 Combined POS Tagging

For items formed by the orthographic combination of two to three words (perhaps due to the orthography rendering prosodic phrasing rather than morphosyntactic word boundaries), a combined POS tag was created by incorporating the individual POS tags. Table 3 provides an overview of all the tag combinations and their occurrences in the diachronic corpus. Overall, 17 new POS-tag combinations were formed to annotate 215 new words found 1695 times in the diachronic corpus. It is interesting to note that majority of these orthographically combined words were formed by combining postpositions, pronouns, auxiliaries and negation particles with other words. All of these tend to be prosodically weak items cross linguistically and are prone to cliticization. It is therefore highly likely that the orthography was representing prosodic phrasing whereby prosodically weak elements incorporate into another prosodic phrase. This possibility remains to be investigated in more detail and is one of several avenues of research that can be conducted now that the DUTI corpus is available.

[8] https://tech.cle.org.pk/services/text/pos

Sr. no	POS-tag combinations	Total number of occurrences	Unique words for each tag
1.	PRP-PSP	1152	81
2.	PRR-PSP	129	11
3.	NN-PSP	41	33
4.	Q-PSP	8	1
5.	VBI-PSP	8	2
6.	VBF-SCK	66	14
7.	VBF-AUXT	60	23
8.	VBF-AUXA	17	11
9.	PDM-NN	64	13
10.	PSP-NN	31	4
11.	PRD-NN	21	2
12.	CD-NN	8	2
13.	PRP-NN	2	2
14.	NEG-VBF	58	10
15.	NEG-VBI	12	2
16.	VBI-VALA	12	2
17.	NEG-VBF-AUXT	6	2

Table 3: New POS-tag combinations in 140K corpus

5.4 Corpus Validation and Quality Assurance

Both manual and automatic methods were used to ensure consistency and completeness. Automated methods verified file numbers, word counts within each file, and word distribution across decades. Additionally, POS label consistency and format were meticulously validated to ensure corpus reliability for language analysis and future research.

6. Temporal Analysis of Orthographic Variation

As discussed in section 5.2, many phonological, spelling and orthographic variations in Urdu words have been observed chronologically. In this section we study their temporal frequency and analyze the pattern of change over time.

Phonological variations in words i.e., تہا ,تھا and تہی، تہی as observed in diachronic text have been analyzed in Figure 8. It can be observed that تہا was widely used in the diachronic texts as an alternate spelling for تھا /ʈʰɑ:/ `was.Masc' and began declining around 1910-1919. On the other hand, تہی used alternatively for تھی /ʈʰi:/ `was.Fem' had already been discontinued around 1820. One explanation for the early extinction of تہی could be that تہی /ʈəhi:/ also means empty a distinct word in Urdu, though spelled the same as /ʈʰi:/.

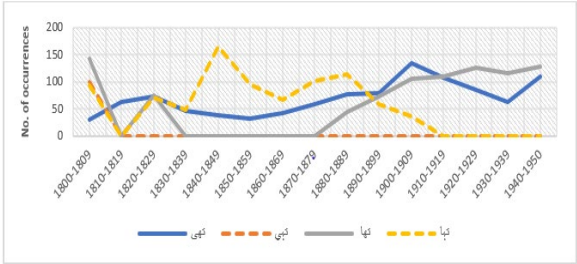

Figure 8: Occurrences of تھا and تھی across 15 decades

In Figure 9, phonological variation of the words نہیں ‘nahīñ’ /nəhĩ:/ ‘no’ as نہین and میں ‘meñ’ /mæ̃:/ ‘I’ as مین due to the alternate use of ‘ن’ /n/ with ‘ں’ (use to represent nasalization) in the diachronic texts are presented. It seems that the alternate use of the letter ‘ن’ with ‘ں’ became standardized after 1929, possibly because of a new distinction between full phonemes and aspiration/nasalization. This seems to have led to the specific use of ‘ھ’ for aspirated sounds (e.g., تھ /ʈʰ/ and ٹھ /ʈʰ/) and ‘ں’ /ə̃/ for nasal sounds (e.g., ماں ‘mãñ’ /mə̃:/ (mother-NN). Additionally, the alternative use of these sounds can sometimes impact the meanings of words, as مان was used alternatively for ماں /mə̃:/ (mother-NN) although the same spelling can also refer to ‘mān’ /mɑ:n/ (proud-NN/agree-VBI).

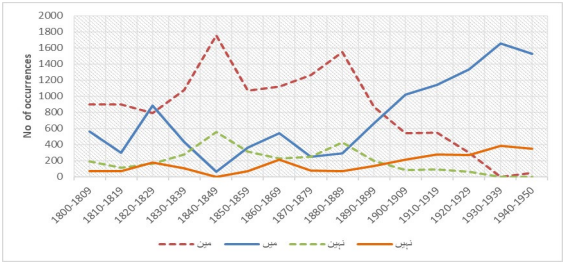

Figure 9: Occurrences of the words نہیں ‘nahīñ’ and میں ‘meñ’ across 15 decades

In Figure 10, occurrences of an alternate spelling of the word یہ ‘yeh’ /je:h/ (this-PDM) as یہہ with an insertion of an extra 'ہ' /h/ is presented. It can be observed that beyond 1900, the alternate form of یہ with additional 'ہ' discontinued, perhaps indicative of standardization effects.

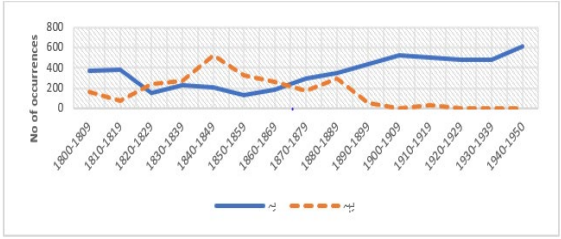

Figure 10: Occurrences of the word یہ ‘yeh’

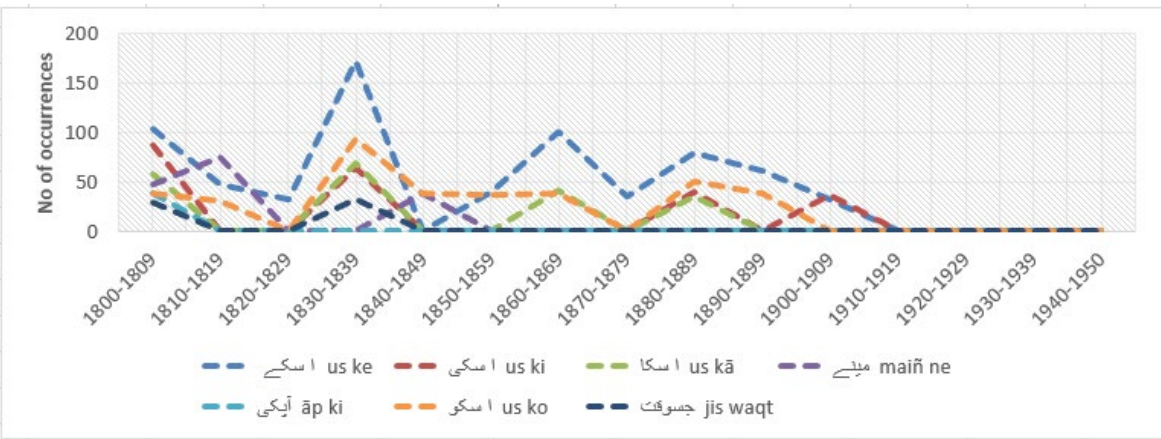

Figure 11: Occurrences of joined words across 15 decades

Orthographic variations of seven items: مینے 'maiñ ne' (PRP-PSP), آپ کی 'āp ki' (PRP-PSP),جسوقت 'jis waqt' (PRD-NN), اسکی 'us ki' (PDM-PSP), اسکے 'us ke' (PDM-PSP), اسکا 'us kā' (PDM-PSP),اسکو 'us ko' (PDM-PSP) were analyzed, as shown in Figure 11. These words exist as a single word in diachronic text, but are commonly written as two words in contemporary text. It can be observed from the figure that these variations were frequently used during 1800-1890, but were less common beyond 1910-1919, although they continue to be rarely used to date, specifically in Urdu website.

Figure 12 traces the historical use of character "ي" choti yeh /j/ and emergence of the alternate spelling of کي as کی /ki:/ by replacing "ي" with "ی" in contemporary Urdu texts. The figure illustrates that the letter 'ي' ceased to appear after 1879, while its alternate 'ی' continued in usage till the present.

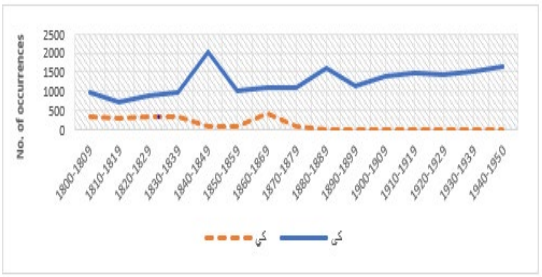

Figure 12: Occurrences of the word کی 'ki' across 15 decades

7. Conclusion

This paper describes the various steps that went into constructing a unique DUTI corpus, encompassing one million words spanning 150 years, along with a sub-corpus of 140,000 POS-tagged data. The corpus development process involved three distinct phases: i) text acquisition and selection, ii) digitization, and iii) POS tagging. Stored in UTF-8 format, each file within the corpus corresponds to one page of historical text. This meticulously constructed resource fills a crucial gap in existing resources for historical and linguistic research to explore diverse research questions, including lexical and grammatical changes, orthographic variations, particularly within the expanding domain of digital humanities.

8. Acknowledgements

This work was funded through a grant, project no. 16141370202 from Deutscher Akademischer Austauschdienst (DAAD), Germany. We deeply acknowledge and thank the services of Mr. Tariq Najmi and Mr. Abdul Hafeez for their dedicated efforts in typing the corpus.

9. Data Availability

Our Dataset is freely available at https://github.com/clekics/diachronic-urdu-corpus/tree/main

10. Bibliographical References

Adam, K., Baig, A., Al-Maadeed, S., Bouridane, A., and El-Menshawy, S. (2018). KERTAS: dataset for automatic dating of ancient Arabic manuscripts. *International Journal on Document Analysis and Recognition (IJDAR)*, 21(4):283-290.

Ahmed, T., Ehsan, T., Ashraf, A., Mutee u Rahman, Hussain, S., and Butt, M. (2020). A multilayered Urdu treebank. In *Proceedings of 7th International Conference on Language and Technology 2020*, pages 1–8, Lahore, Pakistan.

Ahmed, T., Urooj, S., Hussain, S., Mustafa, A., Parveen, R., Adeeba, F., Hautli, A., and Butt, M. (2014). The CLE Urdu POS Tagset. In *Proceedings of the Language Resources and Evaluation Conference (LERC 14)*, pages 2920–2925, Reykjavik, Iceland. European Language Resources Association (ELRA).

Akram, Q., and Hussain, S. (2017). Ligature-based font size independent OCR for Noori Nastalique writing style. In *Proceedings of 1st*

International Workshop on Arabic Script Analysis and Recognition (ASAR 2017), LORIA, Nancy, France.

A representative corpus of historical English registers (ARCHER). (2014). https://www.projects.alc.manchester.ac.uk/archer/

Barakat, B. K., El-Sana, J., and Rabaev, I. (2019). The Pinkas dataset. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, Australia.

Belhekar, V., and Bhargava, R. (2023). Development of word count data corpus for Hindi and Marathi literature. *Applied Corpus Linguistics*, 3(3), 100070.

Belinkov, Y., Magidow, A., Barrón-Cedeño, A., Shmidman, A., and Romanov, M. (2019). Studying the history of the Arabic language: language technology and a large-scale historical corpus. *Language Resources and Evaluation*, 53(4):771-805.

Belinkov, Y., Magidow, A., Romanov, M., Shmidman, A., and Koppel, M. (2016). Shamela: A Large-Scale Historical Arabic Corpus. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, Osaka, Japan.

Burnard, L. (2007). *Reference Guide for the British National Corpus (XML Edition).* Oxford University Computing Services on behalf of the BNC Consortium. https://www.natcorp.ox.ac.uk/docs/URG/

Davies, M. (2007). *Time Magazine Corpus*. Retrieved February 21, 2025 from https://www.english-corpora.org/time/

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora*, 7(2):121-157.

Ehsan, T., and Butt, M. (2020). Dependency parsing for Urdu: Resources, conversions and learning. In *Proceedings of the 12th Language Resources and Evaluation Conference,* pages 5202–5207, Marseille, France.

Elanwar, R., Qin, W., Betke, M., and Wijaya, D. (2021). Extracting text from scanned Arabic books: a large-scale benchmark dataset and a fine-tuned Faster-R-CNN model. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(4):349-362.

Fischer, A., Keller, A., Frinken, V., and Bunke, H. (2012). Lexicon-free handwritten word spotting using character HMMs. *Pattern Recognition Letters*, 33(7):934-942.

Huber, M., Nissel, M., and Puga, K. (2016). *Old Bailey Corpus 2.0*. https://fedora.clarin-d.uni-saarland.de/oldbailey/downloads/OBC_2.0_Manual%202016-07-13.pdf

Kang, A., Le, T., and Chen, Y. (2024). Toshakhana: A Multidimensional Panjabi Corpus in Gurmukhi Script. In *Proceedings of the 2024 ACM Southeast Conference,* Marietta, Georgia, USA.

King, C. R. (1999). *One language, two scripts: The Hindi movement in nineteenth century North India*. Oxford University Press.

Kroch, A. (2020). *Penn Parsed Corpora of Historical English LDC2020T16.* Retrieved February 21, 2025 from https://catalog.ldc.upenn.edu/LDC2020T16

Kytö, M. (2011). Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada*, *11*:417-457.

Osório, T. F., and Lopes Cardoso, H. (2024). Historical Portuguese corpora: a survey. *Language Resources and Evaluation*, 59:1797-1832.

Papadopoulos, C., Pletschacher, S., Clausner, C., and Antonacopoulos, A. (2013). The IMPACT dataset of historical document images. In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing,* Washington, District of Columbia, USA.

Pletschacher, S., and Antonacopoulos, A. (2010). The PAGE (Page Analysis and Ground-Truth Elements) format framework. *20th International Conference on Pattern Recognition,* Istanbul, Turkey.

Qureshi, O. (1996). *Twentieth-century Urdu literature*. In N. Natarajan (Ed.), Handbook of Twentieth-Century Literatures of India. Greenwood Press, Westport, CT.

Rai, A. (1991). *A house divided: The origin and development of Hindi-Urdu*. Oxford University Press.

Raji, S., Alikhani, M., de Melo, G., and Stone, M. (2024). A corpus of Persian literary text. *Language Resources and Evaluation*, 58(2):409-425.

Rehman, T. (2011). *From Hindi to Urdu: A social and political history*. Oxford University Press.

Saad, R. S. M., Elanwar, R. I., Kader, N. S. A., Mashali, S., and Betke, M. (2016). BCE-Arabic-v1 dataset. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, Corfu Island, Greece.

Stökl Ben Ezra, D., Brown-DeVost, B., Jablonski, P., Kiessling, B., Lolli, E., and Lapin, H. (2021). BibIIA - a General Model for Medieval Hebrew Manuscripts and an Open Annotated Dataset. *The 6th International Workshop on Historical*

Document Imaging and Processing, Lausanne, Switzerland.

The Helsinki corpus of English texts. (1993). Helsinki: Department of English, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/

Urooj, S., Mumtaz, B., Hussain, S., and Haq, E. u. (2021). Acoustic and prosodic correlates of emotions in Urdu speech. *Interspeech 2021*, pages 396–400, Brno, Czechia.