Hindustani or Hindi vs. Urdu: A Computational Approach for the Exploration of Similarities Under Phonetic Aspects

Muhammad Suffian Nizami
Department of Computer Science
FAST National University
Chiniot-Faisalabad, Pakistan

Tafseer Ahmed
Center for Language Computing
Mohammad Ali Jinnah University
Karachi, Pakistan

Muhammad Yaseen Khan
Center for Language Computing
Mohammad Ali Jinnah University
Karachi, Pakistan

***Abstract*—The semantic coexistence is the reason to adopt the language spoken by other people. In such human habitats, different languages share words typically known as loan words which appears not only as of the principal medium of enriching language vocabulary but also for creating influence upon each other for building stronger relationships and forming multilingualism. In this context, the spoken words are usually common but their writing scripts vary or the language may have become a digraphia. In this paper, we presented the similarities and relatedness between Hindi and Urdu (that are mutually intelligible and major languages of Indian sub-continent). In general, the method modifies edit-distance; and works in the fashion that instead of using alphabets from the words it uses articulatory features from the International Phonetic Alphabets (IPA) to get the phonetic edit distance. This paper also shows the results for the languages consonant under the method which quantifies the evidence that the Urdu and Hindi languages are 67.8% similar on average despite the script differences.**

***Keywords*—Lexical Similarity; Urdu; Hindi; Edit Distance; Pho-netics; Natural Language Processing; Computational Linguistics**

I. Introduction

In the Indian sub-continent hundreds of different languages are spoken throughout the area it spans, most of which belong to the Dravidian and Aryan families. It is the accepted fact by linguists that the Aryan family of languages evolved from Sanskrit [1]. Historically, during the medieval period of India, Sanskrit was the language of rulers and of the people from the upper-class, this period also shows the witness for Prakrits and the other languages derived from the Sanskrit [2]. Followed by time, we see the rule of Persian language in Indian courts; and at the near end of Mughal era, Urdu eventually became the official language of the court [2], [3]. Many of the researchers argue that the languages Urdu and Hindi are same because they share the same grammar and a large number of words in their common vocabulary; while in the same context, many other researchers express their findings in refutal. The debate engages the development and origin of the Urdu. A common understanding behind the development of Urdu language shows that it is a creole language which came into being through the mixing of local Indian people and the foreign invaders from the different background and ethnicities [4]; and hence, often referred with the 'camp language'. In contrast, veteran Urdu lexicographer Parekh [5] rejected the theories describing Urdu as a creole language; and maintained that it is the 'Khariboli' of the central zone of India with the exception that if its vocabulary draws words from the Persian and Arabic it would become Urdu, and similarly, if it uses words from Sanskrit it would be Hindi. The Urdu and Hindi are the mutually intelligible languages; however, are the victim of language-split which resulted in the usage of modified Perso-Arabic script called 'Nastaliq' and Devanagari script, respectively, for writing. Amongst many other characteristics of these scripts the two which appear salient are: Nastaliq is a cursive scripted and supposed to be written in right-to-left direction; whereas, Hindi is block-letter and follows left-to-right direction. Figure 1 depicts the languages of the world and their respective sizes (in terms of size of leaves spread), where specifically Urdu and Hindi appear on the top-left, in the Indo-European→Indo-Iranian→Indo-Aryan branch. In the same context, the two languages in the world of today can be combined under the common term 'Hindustani' and also recognized as separate Persianised and Sanskritised registers of the Hindustani language.

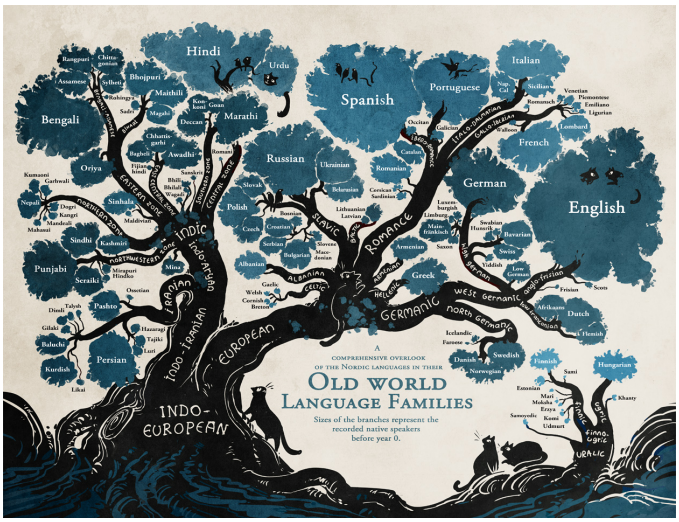

Figure 1: Major languages of the world presented in the format of family tree. Courtesy Minna Sundberg.

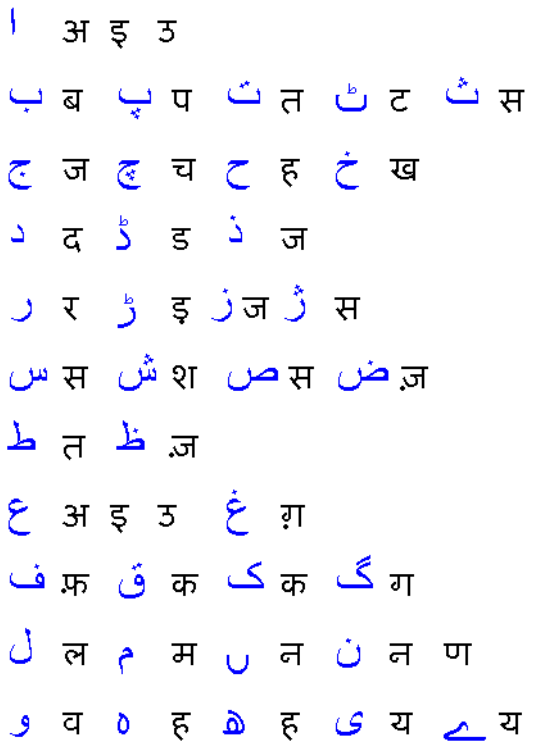

Figure 2: Alphabets of Urdu and Hindi languages in modified Perso-Arabic script (shown in blue colour) and Devanagari script (show in white colour) respectively.

Before we proceed further, another behaviour is noteworthy for which we see that since the division of India happened in 1947, a special focus has been made on official grounds for inducting Persian and Arabic words in Urdu and Sanskrit in Hindi by the right governments and print and electronic media associates of the two countries (i.e., Pakistan and India). Thus, where these languages share a vast vocabulary and morphological structure, their speakers attempt to distinguish them through the word borrowing or with *loan words* from the source languages as mentioned earlier. Hence, it is observable that it may be very difficult for the youth of contemporary time to comprehend the news bulletin announced officially in the pure Urdu and Hindi. However, movies and other channels of entertainment can be accounted for as the principal medium of vocabulary enhancement.

With more than 329.1 million native speakers all around the world, [2] and being a victim of digraphia; the main challenges and point of research investigation—w.r.t the similarities between Hindi and Urdu languages—taken into the consideration for this paper are discussed in the subsequent paragraphs.

The difference in writing script leads the matter not only towards the inabilities of reading but also to the pronunciation. We see that the pronunciation of certain Perso-Arabic alphabets are improper w.r.t the core Hindi speakers such that they are not able to differentiate 'ज', 'ज़', 'ज़', 'ज़', and 'ज़'; as they tend to pronounce 'ज' for all of them. It is seen very often to associate a diacritic symbol, namely *bindi*, in the transformation of 'ज'→'ज़' for differentiating 'ض' and 'ظ' from the rest of aforementioned Urdu alphabets. Similarly, for Urdu alphabets 'گ' and 'غ' they use a single Hindi alphabet'ग' and add bindi in it ('ग़') to substitute it for 'غ'; in a very similar fashion, it uses 'क' and 'क़' for 'ک' and 'ق' respectively. In a contrasting manner, for the two Urdu alphabets such as 'ا' and 'ع' Hindi corresponds with the three alphabets 'अ', इ', and उ'; similarly, for 'ح' and 'ھ', and 'ہ' Hindi has only one alphabet 'ह'; and lastly, for 'ث', 'س', and 'ص' Hindi has got only one alphabet i.e. 'स'; so with the Urdu alphabets 'ے' and 'ی' it has got a single alphabet i.e. 'य'. Collectively all of the alphabet marking are mapped in the figure 2. Thus, with these many-to-many mappings among the alphabets of two scripts, we can easily anticipate the production of very severe semantic mistakes. For example, take the Urdu word 'ذلیل' [ðaˈliːl] (*humiliated*) for which the Hindi may have the chance to pronounce as 'जलील' [dʒaˈliːl] (*exalted*, *magnificent*). Similarly, for the multi-words, the Urdu language has to give an additional space hence a single word would consist of multiple tokens; for example, 'ان پڑھ' [ənˌpəɽʱ] (*illiterate*) which is 'ان' + 'پڑھ', however; Hindi language has no compulsion of giving a white-space in between tokens, so for the given Urdu example 'ان پڑھ' , it will render 'अनपढ़'(pronounced as per same IPA and meant into the same thing).

Thus, in order to find the similarity between the two languages, we are required to transform every cognate as per a similar scheme. For such scheme, Romanized transliteration is the popular way but it undergoes with the same issue i.e. many-to-many alphabet mapping; for example 'ث', 'س', and 'ص' will have only substitute in the Latin script i.e. 'S' *et cetera* [6]. The alternate approach, as used in this paper, is taking the IPA into account for the transformation. In addition to it, this paper presents a modified version of conventional edit distance, namely 'Phonetic Edit Distance' (PED), where the articulatory features of the IPA are employed. This will also help us to see the relatedness of the same word, spoken/pronounced by the people of core Urdu and Hindi backgrounds, at a slight/negligible distance; instead of getting a hard distance through standard edit distance metric (yielded on romanized transliteration). Likewise, Nizami *et al.* [7], we considered to find the similarities w.r.t the Parts-of-Speech (PoS); such that it would be more interesting to find the right cognate of *book* as a noun in the list of nouns of the other language, rather the make a generalized look-up on all possible words.

The rest of the paper organization is as The literature review about the lexical similarity of languages and earlier techniques is in Section II, the methodology is described in Section III, the detailed results are shared in Section IV, followed by the conclusion and future works in Section V, and bibliographical references in the end.

II. Literature Review

In this section, the literature review, the existing techniques, and approaches for lexical similarity concerning script and sound are described.

The two most frequent method for resolving the problem of this kind are string matching algorithms and employment of the Soundex algorithm. The edit distance algorithm has different variants for different types of tasks like string alignment and

spells correction in language processing [8]. The problem is that it takes the characters or letters as distinct units, in such cases if the characters are completely similar then no operation needed. It depends on the user that it may use different weights for the operations. There is another algorithm known as Soundex which works on sound-based matching instead of letter or spell matching [9].

Jinugu [10] presented which is the variant of the Tarhio-Ukkonen algorithm [11] for maximizing the matching of a string by finding the longest patterns in the string and ignores the mismatches of characters. This algorithm works in multiple shifts on the variant lengths of strings for matching purpose, the shift distance and number of characters involved in matching also matters for its performance.

The work [12] shows usage of the Soundex algorithm for retrieving noun words from the database consisting of vowels and consonants for the Hindi language. Likewise, the Soundex algorithm provides classes for letters as their agreement classes which are six in number, where the vowels are eliminated and only consonants are changed into their relevant phonetic class [13]. Other similar work is by [14] and [15] which is phonetic matching using rule-based algorithms and utilizing encoding scheme for homophone words matching scripted in different languages.

The Soundex algorithm considers many letters due to the articulation of sound. The IPA chart is present at the website[1]. The IPA's are ordered according to the manner and place of articulation. Few letters have same articulation i.e. *plosive*, *bilabial* [16]. Similarly few letters are *voiced* and *unvoiced* consonant [17], *aspirated* and *pharyngeal* etc. The set of features can represent IPA symbols. In Germanic languages, according to [18] and [19], there are some voiced-stops that can change into voiceless stops and vice versa.

There are different studies found on the lexical similarity like [20] worked on the dialectal differences among the pair of texts using cosine similarity, Hamming distance, and Levenshtein distance and [21] worked on cognates identification among different languages based on inter-related vocabulary. It shows that the lexical similarity can be computed by using the phonetic level features of words rather than orthographic features.

Another work is done for the similarity of words on a limited PoS by using synsets in WordNet, to extend this lexical similarity on phrase level and sentence level computed with the help of word-level similarity [22]. Similarly, the lexical similarity computed for the source code by using string level matching [23]. Some other researchers find multiple dimensions for lexical similarity like knowledge-based, string-based, and corpus-based [24]. An experiment conducted by [25] on cross-language similarity for the cognates of Dutch and English, on similar grounds [26] lexical similarity was computed by using phonetics based cognates with high frequencies.

As the researchers in [25] and [26] showed that in cross-language cognates and loan words similarity matters phonologically. Similarly, the historic background and origin of these languages are analyzed like [27] did for Urdu and French words. Another work in support of phonological level similarity of languages was done on English word structures as a network of language where links were made between words phonologically [28].

III. Methodology

In this section, the main components are described as: The detailed discussion about the languages chosen for the experiment, the source of the dataset, and the specific PoS word lists for lexical similarity is done in section III-A. Discussion about the standard edit distance and the proposed phonetic modifications based on articulatory features is described in section III-B. The proposed method modified phonetic edit distance is explained in section III-C. The detailed discussion for the computation of lexical similarity for the chosen languages is given in section III-D.

The basic task of lexical similarity calculated on similar words ratio or count in between two languages. But, there are few reservations like:

1) How it will be inferred that two words are similar?
2) To which extent comparison should be done and which criteria should follow to choose the words?

The answer to the raised questions is that there should be a dynamic method to decide whether two words are cognates or not, the edit distance should be employed as a measure of similarity, the comparison should be on the equal words or some acceptable count of words and lastly, the criteria is also regarding the origin and source of words picked for the lexical comparison. To explain the first part of the answer we need to propose a modified edit distance method for computing lexical similarity which is explained in the coming part as phonetic edit distance. The next parts of the answers are related to word lists and their specific feature or aspect for selecting comparison. For this, we have chosen different parts-of-speech (PoS). This decision is made due to the importance of PoS as these are the rich and main content of any language, also some previous similar work of lexical similarity was done by using PoS tag set [7]. For the complete pictorial view of the proposed phonetic edit distance method, the system diagram is given in figure 3.

A. Dataset

We used the Universal Dependency[2] (UD) corpora for extracting PoS word lists. UD has some standard data about all languages in a standard format. Another reason to choose the only PoS for similarity is that in the textual corpora each language can be divided into PoS tag set. In this experiment, we have chosen Urdu and Hindi with majorly two scripts Devanagari and Perso-Arabic. Also, the conversion system for these scripts to IPA is developed. The part-of-speech (PoS) used for the similarity purpose are verbs, proper nouns, nouns, particles, auxiliary, pronoun, coordinating and subordinating conjunctions, and adposition. The length of each PoS tag word list is shown in Figure 4 comparing the size of both Hindi and Urdu languages.

[1] https://www.internationalphoneticassociation.org/content/full-ipa-chart

[2] https://universaldependencies.org/

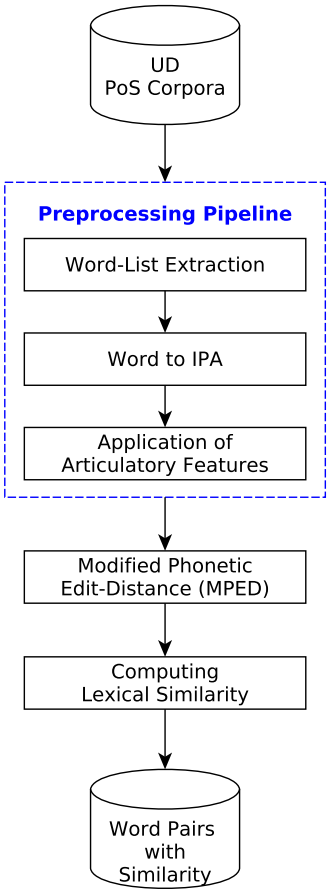

Figure 3: Phonetic Edit Distance based Lexical Similarity System Diagram.

B. Understanding the Phonetic Edit Distance

The standard edit distance [29] takes two words or strings and returns the distance between them. In this process, the internal mechanism of the edit distance method is based on the insertion, deletion, and update operations which compute the cost for two strings as a distance. Each operation is given a unit cost which aggregates during the comparison of two strings. This is simple string matching which doesn't provide any information about the sound-related features like phonetic articulatory features. In our proposed method the edit distance is modified based on these articulatory features and called here as the future of edit distance as phonetic edit distance.

The proposed Phonetic Edit Distance (PED) works the same as standard edit distance but its internal mechanism is based on phonetic features, it is explained in section III-C. It takes IPA encoded two words and then returns the phonetic distance between them. If the sound of both words is the same then the phonetic distance will be zero. But, if the words are not similar then the insertion and deletion operation computes cost as well as the phonetic cost of the words also aggregates to total in case of mismatch of sound.

If we take the standard edit distance, the distance of two IPA strings /bæd/ and /pæd/ is Φ(/bæd/, /pæd/) = 1, these IPA strings represent English words *bad* and *pad* respectively;

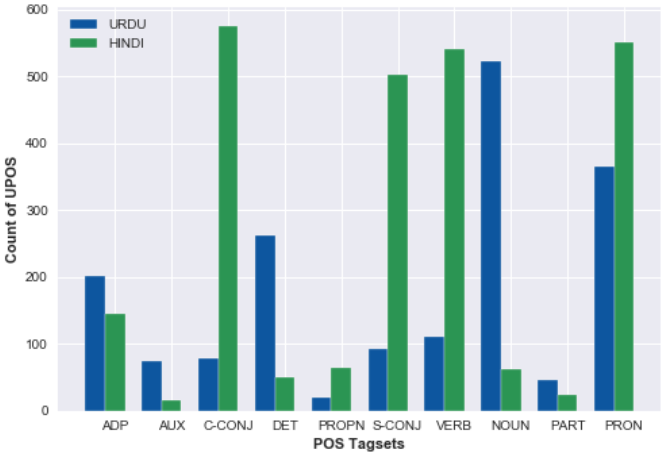

Figure 4: word-list size PoS-wise for Hindi and Urdu.

and Φ is the edit distance. There is only one replace operation of 'b' with 'p' to make the similar string but, with our proposed PED method, by using the same pair of IPA encoded strings is 0.2, in this case, the cost of replacement operation of 'b' with 'p' is 0.2 as the sound of both letters make less difference and take place near in the articulatory feature-based IPA chart, and thus phonetic similarity of the 'bad' and 'pad' is much lesser than the standard ED.

The lexical similarity in true sense is by calculating for both types of words. But vowels in different languages contain add-on features like in Urdu vowels contain short vowels as their composing part [30]. In our paper, we are skipping the vowels as future work due to their complex nature of features for phonetics. Thus, for *consonants*, we have proposed the following features (and their respective values) *voiced* (binary), *airflow* (discrete), *place* (continuous), *aspirated* (binary), *pharyngeal* (binary) and *manner* (discrete).

The features are picked from their positions, the *place* is the articulation of place and inside the human mouth, these places are present. we have assigned the value as per their feature positions. Like lips (bilabial) position have 0.05, teeth position have 0.15, and the throat (glottal) position has 0.95. The other two important features are *type* and *label*, the *label* is the IPA of sound and *type* ensures that sound is a vowel or a consonant.

C. Algorithm for Phonetic Edit Distance

As the sound based articulatory features presented in section III-B and represented with their corresponding values. In this work, we didn't compare the vowels with consonants but only consonants with consonants.

For the comparison of consonants, as shown in algorithm: 1, we have given $^2/_3$ value to the *place* and *manner* features; and $^1/_3$ value is assigned to the remaining features. The *voiced* feature has $^1/_5$ value; and the other features have the remaining weight/value shown as $1 - {}^2/_3 - {}^1/_5$. At present, in the proposed system, the remaining features are *airflow* and *aspirated*. Although, we can increase the features without decreasing the weight of major features.

Further, *manner* and *place* features are more substantial. So, we have considered the distance only when *manner* and *place* represented as (δ_{m+p}) is equal or lesser than the threshold which is $^1/_2$. If the joint distance is above the threshold level then we don't add the distance of other features and return only this distance. Also, rule-based distance for feature *manner* is done by using dictionary look-up when the key of $\langle a_{\text{manner}}, b_{\text{manner}}\rangle$ is given. Finally, a_D,b_D, shows the Manhattan Distance for the remaining all features as presented in-line 10.

Algorithm 1 Phonetic Difference for Consonant (PDC)

1: **procedure** PDC(a, b) ▷ a and b are strings, where $a \neq b$.
2: $\delta_{\text{manner}} \leftarrow \Theta[\langle a_{\text{manner}}, b_{\text{manner}}\rangle]$ ▷ Calculating Manhattan Distance between a and b using open and back features.
3: $\delta_{\text{placed}} \leftarrow \Theta(\langle a_{\text{placed}}, b_{\text{placed}}\rangle)$
4: **if** $\delta_{\text{manner+placed}}$ >threshold **then**
5: **return** $\delta_{\text{manner+placed}}$
6: **else**
7: $\delta_{\text{manner+placed}} \leftarrow \delta_{\text{manner+placed}} \cdot \frac{2}{3}$
8: $\delta_{\text{voiced}} \leftarrow \Theta(\langle a_{\text{voiced}}, b_{\text{voiced}}\rangle) \cdot \frac{1}{5}$
9: $\delta_{\text{remain\_features}} \leftarrow \Theta(\langle a_D, b_D\rangle) \cdot \beta$
10: **return** $\delta_{\text{manner+placed}} + \delta_{\text{voiced}} + \delta_{\text{remain\_features}}$

Algorithm 2 Modified Phonetic Edit Distance (MPED)

1: **procedure** MPED($\mathbf{X}, \mathbf{Y}$) ▷ X and Y are the IPA encoded strings
2: $x \leftarrow$ length of $\mathbf{X}$.
3: $y \leftarrow$ be the length of $\mathbf{Y}$.
4: **if** $\min(x, y) = 0$ **then**
5: **return** $\max(x, y)$
6: **if** $\mathbf{X}[x-1] = \mathbf{Y}[y-1]$ **then**
7: cost= 0
8: **else**
9: ins_cost $\leftarrow$ ED($\mathbf{X}[0 : x-1], \mathbf{Y}$) + 1
10: del_cost $\leftarrow$ ED($\mathbf{X}, \mathbf{Y}[0 : y-1]$) + 1
11: rep_cost $\leftarrow$ ED($\mathbf{X}[0 : x-1], \mathbf{Y}[0 : y-1]$) + PDC($\mathbf{X}[x-1], \mathbf{Y}[y-1]$)
12: **return** $\min$(ins_cost, del_cost, rep_cost)

In Algorithm 1, PDC is the phonetic difference of consonants, Θ is computing unit for two features, MPED is the modified phonetic edit distance and ED is the edit distance. The pseudo-code of overall lexical similarity for experimental languages is described in Algorithm 3, in which on finding similarity the result is in the range of (0,1) if the sound is the same then 0 otherwise 1. This way all PoS words from the Urdu language compared with Hindi language using this modified phonetic edit distance to find the lexical similarity and the ratio of loan words or cognates between the languages.

A brief example. If we take the example of Urdu word 'صاحب' and its standard IPA [sɑː.ɦɪb][3] (*sir* or *mister*) we may have many similar words/cognates in Hindi which are pronounced as gv[s̪äː.ɦɪb] , [sɑː.ɦəb] , [s̪äː.ɦəb] , [sɑːɦb] , [s̪äːɦb] , [sɑːb] , [s̪äːb][4]. To sense, the difference between

[3]https://en.wiktionary.org/wiki/صاحب

[4]https://en.wiktionary.org/wiki/साहिब#Hindi

Table I: Mapping of articulatory features for the Urdu word 'صاحب' [sɑː.ɦɪb].

	Meta Features	IPA letters				
	Label →	s	ɑː	ɦ	ɪ	b
	Type →	c	v	c	v	c
Phonetic Articulatory Features	Method	0	-	0	-	0
	Place	.45	-	.95	-	0.15
	Manner	fr	-	nsfr	-	pl
	Voice	0	-	0	-	0
	Aspirated	0	-	0	-	0
	Open	NA	-	NA	-	NA
	Back	NA	-	NA	-	NA
	Rounded	NA	-	NA	-	NA

Table II: Mapping of articulatory features for the Hindi word 'साहिब '[s̪äː.ɦəb] .

	Meta Features	IPA letters				
	Label →	s̪	äː	ɦ	ə	b
	Type →	c	v	c	v	c
Phonetic Articulatory Features	Method	0	-	0	-	0
	Place	.35	-	.95	-	0.15
	Manner	fr	-	nsfr	-	pl
	Voice	0	-	0	-	0
	Aspirated	0	-	0	-	0
	Open	NA	-	NA	-	NA
	Back	NA	-	NA	-	NA
	Rounded	NA	-	NA	-	NA

Table III: Comparison of standard edit distance and proposed method.

Source & Target Words	Standard Edit Distance	Proposed Phonetic Edit Distance
[sɑː.ɦɪb] vs. [s̪äː.ɦɪb]	2	.1
[sɑː.ɦɪb] vs. [sɑː.ɦəb]	1	0
[sɑː.ɦɪb] vs. [s̪äː.ɦəb]	3	.1
[sɑː.ɦɪb] vs. [sɑːɦb]	1	0
[sɑː.ɦɪb] vs. [s̪äːɦb]	4	.1
[sɑː.ɦɪb] vs. [sɑːb]	3	.1
[sɑː.ɦɪb] vs. [s̪äːb]	5	1.05

the IPA letter /sɑ/ and /s̪/ we can substitute/suppose Urdu alphabets ص and س respectively; and for the analogy of the romanized variant, both of them would be producing sound for *s* but former one has low whistle sound in comparison to the later one, where whistle sound is bit higher due to dental place. Thus, with the romanized equivalents of these (Hindi) words we can get a higher distance through the standard edit distances (see table III); whereas, with the proposed model the PED will give results very closely. Further for reference, the tables I and II show the articulatory features of Urdu word 'صاحب' [sɑː.ɦɪb] and Hindi 'सहब ' [s̪ɑː.ɦəb], where in the table the type *c* indicates the alphabet is consonant otherwise it vowel (for which the PED is not working); similarly, NA shows the feature is not applicable on the very alphabet.

D. Computing Lexical Similarity

The flow of finding lexical similarity is described in this section as; The word lists created from UD corpora, then these word lists converted into respective IPA codes, after this IPA codes enriched with phonetic articulatory features and in last the lexical similarity computed based on these phonetic features between the languages using the proposed method and the pseudo-code is presented in Algorithm 3.

Universal Dependencies website[5] provide the corpora for all languages in CoNLL-U format with tagged PoS. The tagged

[5]https://universaldependencies.org/u/pos/index.html

dependency structure includes word lemma and the Universal Parts-of-Speech (UPoS). We used lemma in computation rather than words, for which there exists an inflectional nature.

After extraction of word-lists, we converted word-lists into IPA strings. There are many online platforms and dictionaries which converts words into respective IPA strings. Keeping in mind that the chosen languages hold short-vowels, izafat-letters, and diacritics causing issues of conversion for such platforms [2]. Underlying this, we have created our mapper of words-to-IPA (in fact script-to-IPA) for both languages.

Further, the articulatory features from the IPA phonetic chart used for a one-to-one mapping of words. The IPA chart is the standard chart for phonetic level weight-age of letters in any word. Based on these features, each word is compared with other words. These articulatory features are described in section III-B and III-C.

Algorithm 3 Computing Lexical Similarity between two Languages.

1: **procedure** LEXSIM($\mathbf{Lang_1}, \mathbf{Lang_2}$) ▷ $\mathbf{Lang_1}$ and $\mathbf{Lang_2}$ are the list of words.
2: $\Phi_{\text{Tot}} \leftarrow 0$
3: **for** every word a in $\mathbf{Lang_1}$ **do**
4: **for** every word b in $\mathbf{Lang_2}$ **do**
5: $\text{x} \leftarrow \text{PED}(a, b) \cdot \frac{1}{\max(\|a\|, \|b\|)}$
6: [Tot] ← least value as key.
7: $_{\text{Tot}} \leftarrow _{\text{Tot}} + \text{x}$
8: $\beta \leftarrow \frac{_{\text{Tot}}}{\|\mathbf{Lang_1}\|}$
9: **return** β

Finally, the lexical similarity (LS) for Urdu and Hindi is computed on all PoS. Let's take the word-lists of $\mathbf{Lang_1}$ and $\mathbf{Lang_2}$; $\mathbf{Lang_1} \neq \mathbf{Lang_2}$ including the same PoS lists. In Algorithm 3 for the comparison of languages $\mathbf{Lang_1}$ and $\mathbf{Lang_2}$, supposing words a of $\mathbf{Lang_1}$ are compared with all words b of $\mathbf{Lang_2}$ in the step-3 and step-4. Also, we have normalized the PED result value with a maximum length of the word for the comparative words in step-5. Otherwise, the smaller words will be getting less value of phonetic distance.

Here in algorithm 3, every word 'a' compared with 'b' and the minimum value recorded in edit distance, this aggregates the overall edit distance; and β is the average distance per letter for both lists. if the value of β is equal to zero '0' then both lists (languages) are similar as identical but if the value is more closer to '1' then words are different in sounds and vice versa.

IV. Results

We have computed the lexical similarity for Urdu and Hindi with the proposed method on articulatory features PoS-wise. The results are shown in Figure 5. It is identified that these languages have cognates and genetic affinity. Urdu and Hindi are quite similar in spoken level but Hindi is written in Devanagari script which is entirely different from Urdu script. The similarity index shows that the Hindi and Urdu are top similar in auxiliaries, determiners, articles, coordinating conjunctions, and pronouns PoS.

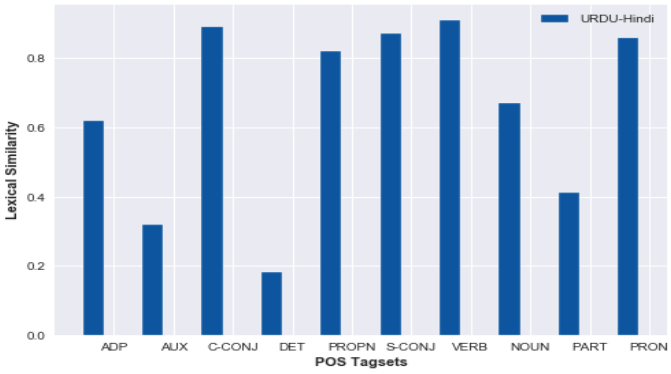

Figure 5: Lexical Similarity PoS-Wise Hindi-Urdu.

It is analyzed from our experiment that Urdu and Hindi on average 67.8% similar languages despite different scripts. This average similarity is computed from the similarity of each PoS employed in this study. In all PoS-wise comparisons, most of the results are comprehensible the only determiner is less similar as shown in Figure 5. Similarly, few PoS have shown high similarity that is due to the unbalanced size of words in those PoS like coordinating conjunction, subordinating conjunction, and verbs. The point which is important to discuss here is the low similarity in adpositions and determiners; this is due to the erroneous tagging of Hindi and Urdu PoS tags in UD dataset. Thus, if specifically the tagging for adposition and determiner is done properly, then, the results would support the fact with more emphasize that the Urdu and Hindi are the same languages though they are the victim of language split.

V. Conclusion

In this paper, we have introduced a modified algorithm for the lexical similarity of Urdu and Hindi languages based on articulatory features. This algorithm has identified the intelligibility, cognates, and borrowed words despite the spelling, script, and phonetic difference. In the conducted experiment with the proposed algorithm, the majority of similarity pairs of PoS are in agreement as per their genetic affinity. The proposed algorithm has given better and understandable results which are far better than the simple string matching with standard edit distance on such a phonetic level parameter. The proposed method is also found effective under the situation where a speaker does not qualify or unable to pronounce a certain alphabet of other languages (for example Arabs cannot pronounce the sound 'p' and 'ch'); so for these situations, they have to look into the similar or near-by sounds for substitution. In such scenarios, PED will give minute results edit distance in comparison to standard edit distance.

The $\approx$ 67.8% similarity is fair enough to stay positive on the question that whether Urdu and Hindi languages are mutually intelligible or not? Since the similarity under the phonetic aspect is high, therefore, we maintain that, within the context of the speech, it is very rightful to term both of these languages as 'Hindustani'; however, the difference of script may produce a very trivial excuse to differentiate either one of them as 'Hindi' or 'Urdu.'

Though there are some limitations in the used UD corpora (variation in the size of languages, format errors, and basic processing) and the issues itself in the languages like silent letters, diacritics, and short-long vowels. It could be improved by using digital lexicographic resources and dictionaries rather than letters to the IPA scheme. In the future, a comprehensive work could be done for the lexical similarity of the whole family of Indo-Aryan languages by extending and enriching the proposed algorithm with vowels along with consonants.

References

[1] A. K. Dutt, C. C. Khan, and C. Sangwan, "Spatial pattern of languages in india: A culture-historical analysis," *GeoJournal*, vol. 10, no. 1, pp. 51–74, 1985.

[2] M. Y. Khan, M. A. Rao, S. Wasi, T. A. Minai, and S. M. K.-u.-R. Raazi, "Edit distance-based search approach for retrieving element-wiseprosody/rhymes in hindi-urdu poetry," *Indian Journal of Science and Technology*, vol. 13, no. 39, pp. 4189–4201, 2020.

[3] M. Y. Khan and M. S. Nizami, "Urdu sentiment corpus (v1. 0): Linguistic exploration and visualization of labeled dataset for urdu sentiment analysis," in *2020 International Conference on Information Science and Communication Technology (ICISCT)*. IEEE, 2020, pp. 1–15.

[4] M. Y. Khan, S. M. Emaduddin, and K. N. Junejo, "Harnessing english sentiment lexicons for polarity detection in urdu tweets: A baseline approach," in *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. IEEE, 2017, pp. 242–249.

[5] R. Parekh, "Urdu's origin: it's not a 'camp language," *Dawn*, Dec 2011. [Online]. Available: https://www.dawn.com/news/681263/urdus-origin-its-not-a-camp-language

[6] M. Y. Khan and T. Ahmed, "Pseudo transfer learning by exploiting monolingual corpus: An experiment on roman urdu transliteration," in *International Conference on Intelligent Technologies and Applications*. Springer, 2019, pp. 422–431.

[7] M. S. Nizami, M. Y. Khan, and T. Ahmed, "Towards a generic approach for pos-tagwise lexical similarity of languages," in *International Conference on Intelligent Technologies and Applications*. Springer, 2019, pp. 493–501.

[8] M. Smith, K. T. Cunningham, and K. L. Haley, "Automating error frequency analysis via the phonemic edit distance ratio," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 6, pp. 1719–1723, 2019.

[9] A. J. Lait and B. Randell, "An assessment of name matching algorithms," *Technical Report Series-University of Newcastle Upon Tyne Computing Science*, 1996.

[10] J. U. Rekha, "Approximate multiple string matching algorithm," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 11, 2020.

[11] S. Mojsilovic and A. Ukkonen, "Relative distance comparisons with confidence judgements," in *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 2019, pp. 459–467.

[12] V. Gautam, A. Pipal, and M. Arora, "Soundex algorithm revisited for indian language," in *International Conference on Innovative Computing and Communications*. Springer, 2019, pp. 47–55.

[13] L. Xu and J. Chamberlain, "Cipher: A prototype game-with-a-purpose for detecting errors in text," in *Workshop on Games and Natural Language Processing*, 2020, pp. 17–25.

[14] I. Mustafin, M.-C. Frunza, and J. Lee, "Multilingual entity matching," in *International Conference on Advanced Information Networking and Applications*. Springer, 2019, pp. 810–820.

[15] M. Llompart Garcia, "Bridging the gap between phonetic abilities and the lexicon in second language learning," Ph.D. dissertation, lmu, 2019.

[16] R. Janssen, S. R. Moisik, and D. Dediu, "The effects of larynx height on vowel production are mitigated by the active control of articulators," *Journal of Phonetics*, vol. 74, pp. 1–17, 2019.

[17] K. R. Kluender, C. E. Stilp, and F. L. Lucas, "Long-standing problems in speech perception dissolve within an information-theoretic perspective," *Attention, Perception, & Psychophysics*, vol. 81, no. 4, pp. 861–883, 2019.

[18] J. Howard, ""pig" or "fig?": Grimm's law, phonemic difference, and linguistic agency in alice's adventures in wonderland," *The Explicator*, vol. 78, no. 1, pp. 41–43, 2020.

[19] A. V. Botsman and O. V. Dmytruk, "Germanic preterite-present verbs and their morphological and semantic peculiarities," *Current issues of Ukrainian linguistics: theory and practice*, no. 39, pp. 74–88, 2019.

[20] J. Nerbonne, W. Heeringa, J. Prokic, and M. Wieling, "Dialectology for computational linguists," 2019.

[21] D. Kanojia, M. Kulkarni, P. Bhattacharyya, and G. Haffari, "Cognate identification to improve phylogenetic trees for indian languages," in *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 2019, pp. 297–300.

[22] T. Vakare, K. Verma, and V. Jain, "Sentence semantic similarity using dependency parsing," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2019, pp. 1–4.

[23] H. Kaur and R. Maini, "Assessing lexical similarity between short sentences of source code based on granularity," *International Journal of Information Technology*, vol. 11, no. 3, pp. 599–614, 2019.

[24] W. H. Gomaa, "A multi-layer system for semantic relatedness evaluation," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 23, 2019.

[25] H. Carrasco-Ortiz, M. Amengual, and S. T. Gries, "Cross-language effects of phonological and orthographic similarity in cognate word recognition: the role of language dominance," *Linguistic Approaches to Bilingualism*, 2019.

[26] E. Lefever, S. Labat, and P. Singh, "Identifying cognates in english-dutch and french-dutch by means of orthographic information and cross-lingual word embeddings," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 4096–4101.

[27] A. A. Khan, "Lexical affinities between urdu and french," *Eurasian Journal of Humanities Vol*, vol. 1, no. 1, 2015.

[28] C. S. Siew and M. S. Vitevitch, "The phonographic language network: Using network science to investigate the phonological and orthographic similarity structure of language." *Journal of Experimental Psychology: General*, vol. 148, no. 3, p. 475, 2019.

[29] G. Cormode and S. Muthukrishnan, "The string edit distance matching problem with moves," *ACM Transactions on Algorithms (TALG)*, vol. 3, no. 1, pp. 1–19, 2007.

[30] M. Kamran Malik, T. Ahmed, S. Sulger, T. Bögel, A. Gulzar, G. Raza, S. Hussain, and M. Butt, "Transliterating urdu for a broad-coverage urdu/hindi lfg grammar," in *LREC 2010, Seventh International Conference on Language Resources and Evaluation*, 2010, pp. 2921–2927.