

Original Research

Phono-lexical similarity between Bahasa Indonesia and Urdu: A corpus-based contrastive analysis study

Hafiza Sana Mansoor^{1*}, Harun Joko Prayitno¹, Laili Etika Rahmawati¹

¹Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

This study investigates phono-lexical similarity between Urdu and Bahasa Indonesia through a corpus-based contrastive analysis. The aim of the research is to categorise phonologically similar lexical pairs using Contrastive Analysis theory and to identify cross-linguistic lexical convergence between Urdu and Bahasa Indonesia. A manually compiled corpus of 326 phonologically similar word pairs was drawn from everyday communication, dictionaries, academic texts, and news sources to ensure broad lexical coverage. The data were organised according to semantic, phonological, syntactic, and etymological criteria and categorised accordingly. The findings indicate that most phonologically similar word pairs are cognates (69.1%), followed by partial cognates (18.4%) and a smaller group of false friends (12.5%). Cognates may support positive lexical transfer, particularly in religious, academic, and formal settings. Partial cognates reflect semantic narrowing or extension and may require additional processing. False friends may produce phono-lexical ambiguity and increase the risk of negative transfer. Most lexemes derive from Arabic (72.4%), followed by Persian (11.7%), Sanskrit (6.7%), Indo-Aryan (2.5%), Portuguese (1.8%), and other languages (4.9%). The study identifies two functions of phono-lexical similarity as a facilitative and a constraining factor in cross-linguistic comprehension. The findings contribute to contrastive linguistics and BIPA instruction and have implications for learners and speakers of Urdu, Bahasa Indonesia, and related languages such as Arabic, Malay, and Hindi. The research is limited to lexical-level contrastive analysis between two languages. Future studies may apply experimental methods to investigate intelligibility or extend the analysis to other linguistic levels or replicate the method with additional languages.

Keywords: *phono-lexical similarity, contrastive analysis, cognates, partial cognates, false friends*

Article history:

Submitted January 17, 2026

Revised March 2, 2026

Accepted March 23, 2026

CRediT Author Statement:

Hafiza Sana Mansoor: Conceptualisation, Methodology, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualisation.

Harun Joko Prayitno: Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing – Review & Editing, Supervision, Project Administration.

Laili Etika Rahmawati: Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing – Review & Editing.

Conflict of interest:

The authors declared no conflict of interest.

Data availability statement:

The data supporting this study's findings are openly available in Zenodo at <https://doi.org/10.5281/zenodo.18739704>.

Funding:

No funding was reported for this research.

Doi:

10.22363/2521-442X-2026-10-1-38-53

For citation:

Mansoor, H. S., Prayitno, H. J., & Rahmawati, L. E. (2026). Phono-lexical similarity between Bahasa Indonesia and Urdu: A corpus-based contrastive analysis study. *Training, Language and Culture*, 10(1), 38–53.



This is an open access article distributed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/) (CC BY-NC 4.0), which allows its unrestricted use for non-commercial purposes, subject to attribution. The material can be shared/adapted for non-commercial purposes if you give appropriate credit, provide a link to the license, and indicate if changes were made.

1. INTRODUCTION

Languages evolve through prolonged interaction among linguistic communities across the world, a process in which vocabulary exchange, the emergence of new

structural forms, and phonological adaptations gradually alter linguistic systems under conditions of language contact (Gumperz, 1964; Matras, 2020; Sato et al., 2025; Thomason, 2019). Such contact-driven processes may

* Corresponding author

© Hafiza Sana Mansoor, Harun Joko Prayitno, Laili Etika Rahmawati 2026

Licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)

generate unexpected lexical and phonological similarities even between languages that are not genetically related (Sulistyono, 2025). Urdu belongs to the Indo-Iranian branch, whereas Bahasa Indonesia originates from the Austronesian language family; despite these distant genealogical origins, the two languages nevertheless present a notable case of contact-induced convergence. With the diffusion of Islam and the expansion of transregional trade across South and Southeast Asia, the communities using these languages maintained sustained commercial, religious, and cultural relations over several centuries (Haque, 2020), and one of the most visible consequences of these interactions is the extensive borrowing of Arabic vocabulary into both languages (Khrisat, 2014; Manfredi, 2017; Versteegh, 2001;). Although Urdu and Bahasa Indonesia are not genetically related, historical contact has produced a shared inventory of Arabic loanwords with similar phonological forms, and these loanwords constitute a significant lexical stratum in both languages, as documented in earlier research (Almurashi, 2024; Islam, 2012; Zainuri, 2019).

The presence of Arabic loanwords with similar phonological forms gives rise to phono-lexical similarity, i.e. resemblance between lexical items at the level of sound and orthographic form. Previous studies indicate that such similarity may facilitate lexical recognition, comprehension, and second language processing (Díaz-Galaz & Torres, 2019; Ramírez et al., 2013). At the same time, phonologically similar lexemes may diverge semantically, while semantically related items may differ in their phonological realisation. This situation may create processing difficulty and has been described as phono-lexical ambiguity (Bordag et al., 2022; Detey et al., 2025). Shared Arabic borrowings in Urdu and Bahasa Indonesia may therefore assist recognition, comprehension, and vocabulary learning, but they may also hinder interpretation when misleading semantic associations arise, particularly in the case of partial cognates and false friends.

Bahasa Indonesia is one of the official languages recognised by the UNESCO General Assembly and is spoken by a large and expanding population of users. Current estimates indicate approximately 80 million native speakers and around 180 million second-language speakers, yielding a total of roughly 260 million users worldwide. Indonesian is also taught internationally in formal educational programmes at 52 institutions outside Indonesia, and more than 150,000 foreign students study the language in different countries, indicating substantial international academic interest. Within Indonesia, Bahasa Indonesia functions as a national lingua franca and coexists with hundreds of regional languages across the archipelago (Lauder & Purwo, 2024). The language is also used in diaspora communities and cross-cultural communication in countries such as Australia, the Netherlands, Saudi Arabia, and Singapore.

Urdu likewise occupies an important position as a widely used international language. It is spoken by more than 70 million native speakers and nearly 100 million second-language users (Abbas et al., 2022). In addition to its extensive use in Pakistan and India, Urdu is maintained in diaspora communities across Europe, the Middle East, and North America, where established speech communities continue to sustain the language.

Bahasa Indonesia and Urdu exist in distinct multilingual environments and maintain wide geographical distribution. The international presence of Bahasa Indonesia has expanded with the development of Bahasa Indonesia for Foreign Learners (BIPA) programmes and sustained cultural diplomacy initiatives, which have increased the number of learners and institutional programmes outside Indonesia (Nugraheni et al., 2021; Susanto et al., 2024). Urdu likewise reflects extensive historical contact with Arabic, Persian, and regional languages and functions as a major lingua franca in South Asia with a well-documented lexical tradition (Billah, 2018). At the same time, growing cultural and diplomatic relations between Pakistan and Indonesia have increased the practical need for cross-language comprehension, particularly in relation to the mobility of tourists, students, migrants, and traders. Hence, the phono-lexical and sociolinguistic characteristics of the two languages make their comparison analytically relevant.

Language interaction among non-native speakers has long been examined in studies of lexical borrowing and the integration of foreign vocabulary into the phonological systems of recipient languages (Best, 1994; Matras, 2020; Thomason, 2019). Loanwords increase perceived similarity between otherwise unrelated languages because borrowed items frequently retain recognisable phonological and morphological features. Research in historical and etymological linguistics shows that Arabic vocabulary in Indonesian undergoes phonological simplification and adaptation to Austronesian phonotactic constraints, whereas Urdu incorporates Arabic lexical items through Persian mediation and preserves relatively complex consonant clusters (Ambreen & To, 2025; Khan et al., 2024). Despite these different pathways of adaptation, the two languages contain hundreds of Arabic loanwords with highly similar phonological forms.

This study constructs a corpus of 326 phono-lexically similar lexical pairs extracted from news texts, everyday conversation, academic discourse, and library sources to examine phonological similarity between Urdu and Bahasa Indonesia. The lexical pairs are classified into three categories – cognates, partial cognates, and false friends – allowing systematic comparison of phonological resemblance and semantic divergence between corresponding items. The analysis focuses on phono-lexical similarity and ambiguity that may facilitate lexical recognition but may also introduce potential sources of misunderstanding in

cross-language communication. The findings have implications for translation practice, second language learning, and language pedagogy, particularly in relation to Bahasa Indonesia for Foreign Learners (BIPA) and Urdu language programmes. The study also contributes to research in second language acquisition, comparative linguistics, and cross-cultural communication through systematic documentation of contact-related lexical correspondences between two historically unrelated languages.

2. THEORETICAL BACKGROUND

Research in sociolinguistics and comparative linguistics has addressed the role of positive lexical transfer in cross-language comprehension (Gooskens & van Heuven, 2021; Karam, 1979). The extent of such transfer depends on several linguistic factors, including phonological transparency, lexical overlap, and syntactic similarity (Gooskens, 2024). Studies of cross-linguistic comprehension have examined both closely related and typologically distinct languages, including Romance languages (Carvalho & Child, 2018) and Scandinavian languages (Platzack, 1987). Urdu and Bahasa Indonesia differ substantially in typology, including their writing systems (Perso-Arabic script in Urdu and Latin script in Indonesian), canonical word order (SOV versus SVO), and morphological structure (Raza, 2017). Nevertheless, extensive borrowing from Persian and Arabic has produced lexical correspondences that may facilitate recognition between the two languages. In contrastive linguistics, corpus-based methods provide systematic evidence for identifying lexical correspondences and classifying cross-linguistic relationships (Granger & Lefer, 2020; Hasselgård, 2020).

Previous research has examined loanword phonology in Malay and Urdu separately, the Persian–Arabic influence on the Urdu lexicon (Billah, 2018; Islam, 2012), and the transmission of Arabic vocabulary into Malay (Alfa et al., 2015; Al-Malki et al., 2014). Phonological and lexical resemblance between items in different languages is generally described as phono-lexical similarity. Psycholinguistic studies maintain that such similarity influences lexical access, word recognition, and meaning inference during language processing (Feldman & Milin, 2018; Orden et al., 1990; Weber & Cutler, 2006). Words in different languages that share similarity in both form and meaning are referred to as cognates, and they are known to facilitate comprehension and accelerate lexical recognition (Carroll, 1992; Sunderman & Schwartz, 2008). At the same time, semantic divergence between phonologically similar items may produce partial cognates and false friends, which may lead to misinterpretation, delayed lexical access, or phono-lexical ambiguity.

Previous studies have examined lexical similarity across languages within contrastive analysis through the classification of cognates, partial cognates, and false

friends. Partial cognates and false friends occur systematically even between closely related languages, as demonstrated in research on English–Spanish lexical relations (Carvalho & Child, 2018; Divasson & León, 2002). These studies relied on dictionary-based etymological analysis and frequency counts to classify lexical correspondences. Experimental research further indicates that false cognates may lead to misinterpretation and phono-lexical ambiguity in second-language communication (Marecka et al., 2021; Mella, 2024). Within second language acquisition research, lexical similarity alone does not guarantee comprehension, as even advanced learners may encounter difficulty distinguishing cognates from false friends (Otwinowska & Szweczyk, 2019).

Shared writing systems or superficially similar word forms may conceal substantial phonological divergence, which may complicate cross-language transfer, as illustrated in studies of phonemic variation in Persian–Urdu–Pashto script adaptation and in research on Urdu–Turkish lexical relations (Farashah, 2010; Khan, 2021). Research on Arabic loanwords in Bahasa Indonesia likewise indicates that phonological similarity may facilitate lexical recognition when systematic phonological adaptation preserves semantic equivalence (Ghazali, 2022; Pangestika et al., 2023). At the same time, several studies report that superficial similarity may mislead second-language learners, although phono-lexical similarity can still support intelligibility even between unrelated languages.

Previous studies conducted in Europe and North America have examined cognate facilitation in language pairs such as Dutch–German (Hamann & Sennema, 2005; Müller, 2005), Spanish–Portuguese (Elvin et al., 2014), and the Scandinavian languages (Plunkett & Strömquist, 2022). Research in Asian language settings has also addressed lexical similarity in pairs such as Malay–Javanese, Hindi–Urdu, and Mandarin–Cantonese, although the available studies remain relatively limited. Much of this research examines the role of phono-lexical similarity in lexical recognition, language learning, and pedagogical applications. In contrast, studies conducted in Asian linguistic environments frequently address religious and cultural borrowings or typological comparison in Urdu and Bahasa Indonesia separately (Khalid et al., 2025; Khan, 2014). Systematic comparison of phono-lexical items between Urdu and Bahasa Indonesia, particularly through corpus-based classification and in relation to cross-linguistic comprehension, remains largely absent from the literature. The present study therefore investigates phono-lexical similarity between Urdu and Bahasa Indonesia and evaluates its role in positive lexical transfer among non-native speakers of the two languages.

The present study is grounded in contrastive analysis (CA), first formulated by Lado (1957). CA involves systematic comparison of similarities and differences between

two or more languages (Fisiak, 1981; Lado, 1957). Ellis (1994) and Chesterman (1998) describe CA as a useful analytical perspective for examining cross-linguistic features and influence. The study adopts the lexical–semantic contrastive model proposed by James (1980), which derives from the principles of contrastive analysis. This model specifies three stages of lexical comparison: examination of formal similarity, semantic analysis, and classification of cross-linguistic correspondences. Awla and Azeez (2021) likewise identify three relevant lexical categories – cognates, partial cognates, and false friends – which are incorporated into contrastive analysis. Earlier studies note that cognates tend to facilitate positive transfer, partial cognates require contextual clarification, and false friends may produce negative transfer in second-language learning (James, 1980; Ringbom, 2007). Distinguishing these categories therefore remains important for understanding both facilitative and misleading effects in lexical recognition (Ellis, 1994; Ringbom, 2007).

The lexical-semantic contrastive perspective described by James (1980), developed within the tradition of contrastive analysis, provides an analytical basis for the examination of phono-lexical similarity in genetically unrelated or contact-influenced languages. Within corpus linguistics, contrastive analysis also enables systematic comparison of lexical items across bilingual and multilingual datasets (Granger & Lefer, 2020; Taylor & del Fante, 2020). A phono-lexical contrastive perspective directs attention to lexical items with similar phonological forms and to the semantic relations between them, which may facilitate comprehension, partially assist interpretation, or lead to misunderstanding in cross-language communication (Ellis, 1994; Ringbom, 2007). The lexical-semantic contrastive tradition therefore provides a basis for classifying correspondences between lexical form and meaning and for distinguishing cognates, partial cognates, and false friends in cross-linguistic comparison.

The present study concentrates on phono-lexical similarity in lexical form and does not attempt to measure mutual intelligibility or psycholinguistic processing outcomes. In this study, lexical similarity refers to shared vocabulary items between Bahasa Indonesia and Urdu, whereas phono-lexical similarity refers to resemblance between lexical items at the level of phonological form together with lexical meaning across the two languages.

Accordingly, the study addresses the following research questions:

1. How can phono-lexically similar lexical pairs between Urdu and Bahasa Indonesia be classified into cognates, partial cognates, and false friends?
2. How does this classification indicate potential areas of positive and negative lexical transfer?
3. What are the principal etymological sources of shared lexical items between Urdu and Bahasa Indonesia?

3. METHODOLOGY

The study employs a corpus-based phono-lexical contrastive analysis to examine similarities and differences between phono-lexically similar lexical items in Urdu and Bahasa Indonesia. Contrastive analysis has long been used to compare linguistic systems across languages and to identify areas of similarity and potential learner difficulty (Fisiak, 1981; James, 1980). In corpus linguistics, such analysis relies on systematic examination of attested language data and avoids dependence on intuitive judgments (McEnery & Hardie, 2012). The present research concentrates on phono-lexically similar lexical items in Urdu and Bahasa Indonesia and examines the relation between phonological form and lexical meaning in corresponding word pairs. The analysis therefore considers two levels of comparison – phonological form and lexical meaning – to classify lexical correspondences and identify cognates, partial cognates, and false friends. The two-level analytical procedure follows principles established in contrastive linguistic research (Haspelmath, 2009; Ringbom, 2007). Phonological similarity was selected as the principal criterion for lexical comparison because phonological form constitutes one of the primary cues involved in lexical recognition and cross-linguistic transfer in second language processing. Figure 1 illustrates the classification scheme used to categorise lexical correspondences.

A corpus of 326 lexical pairs was compiled on the basis of phonological similarity between Urdu and Bahasa Indonesia. To ensure contextual representativeness and broad lexical coverage, the pairs were extracted from multiple sources, including everyday conversation, dictionaries, academic texts, and news materials. A purposive sampling procedure was applied to identify lexical items displaying surface-level phonological resemblance across the two languages. The analysis concentrates on the identification and classification of phono-lexically similar items to document cross-linguistic correspondence in phonological form and lexical meaning, focusing primarily on phonological similarity while orthographic resemblance is treated as secondary.

Data collection relied on phonological similarity as the primary selection criterion, and lexical items were extracted from dictionaries, academic texts, conversational usage, and media discourse. Urdu lexical items were Romanised for phonetic comparison and analytical consistency. This procedure is consistent with established practices in contrastive and psycholinguistic research (Weber & Cutler, 2006). Lexical items displaying full or partial phonological similarity at the surface level were included in the corpus, ensuring that the dataset corresponded to the analytical objectives of the study. Phonological similarity was determined on the basis of comparable syllable structure and segmental correspondence between lexical forms in the two languages.

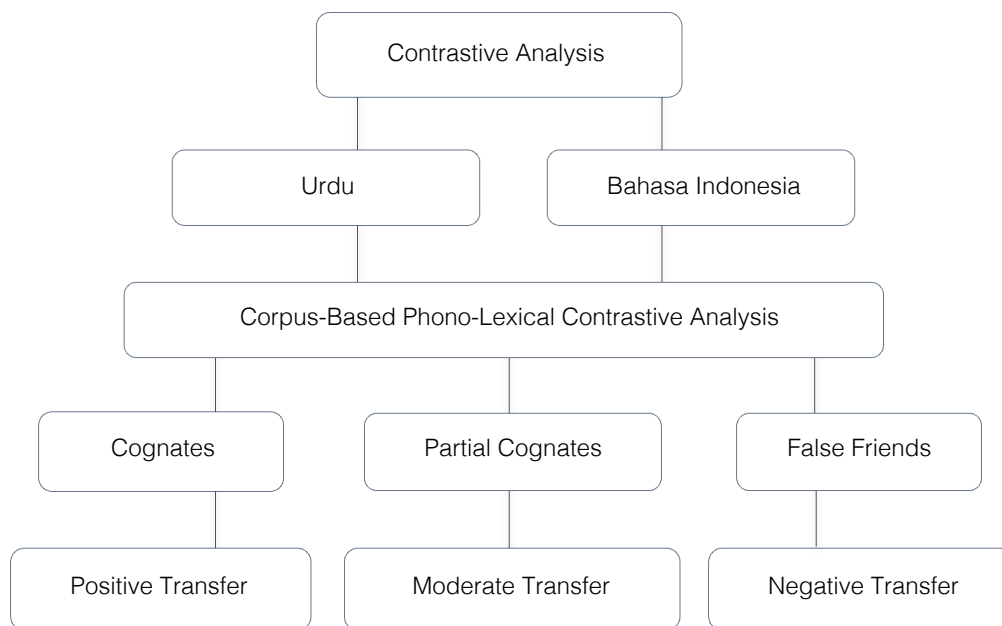


Figure 1. Classification of phono-lexical correspondences between Urdu and Bahasa Indonesia

After compiling the corpus of 326 phono-lexically similar lexical pairs, the dataset was organised to support accurate and replicable analysis. The lexical pairs were classified according to an explicit semantic decision protocol designed to ensure transparency in the categorisation procedure. Core denotational meanings of the lexical items were compared using standard reference dictionaries in each language. Bahasa Indonesia entries were verified through *Kamus Besar Bahasa Indonesia* (KBBI), whereas Urdu items were checked against the *Rekhta Dictionary*. Classification required at least one shared semantic component between the lexical items, as some words displayed multiple meanings in one language while corresponding items in the other showed fewer semantic interpretations. In cases of polysemy, classification was based on the primary dictionary sense shared by the lexical items.

Each lexical pair was recorded in a structured data matrix and treated as a single analytical unit. The dataset was first organised in spreadsheet software and subsequently processed using statistical and qualitative analysis tools. The corpus was organised into clearly defined columns representing the principal linguistic variables: Urdu lexical item, Bahasa Indonesia lexical item, semantic meaning, etymological origin, part of speech, and International Phonetic Alphabet (IPA) transcription. Standardised coding conventions were applied throughout the dataset to maintain internal consistency. Authoritative bilingual and monolingual dictionaries served as reference sources for establishing lexical meanings. Etymological references for Arabic, Persian, and indigenous sources were consulted to determine the historical origin of each

lexical item. A unified part-of-speech tag set applicable to both languages was adopted to classify syntactic categories. To permit direct comparison of phonological forms, all lexical items were transcribed in IPA.

Several procedures were implemented to strengthen the validity and reliability of the dataset. All lexical pairs included in the corpus were required to display full or partial phonological similarity, which served as the principal criterion for corpus inclusion. IPA transcription was applied to document phonological correspondence between lexical items and to make pronunciation transparent for readers who are not familiar with either language.

Coding guidelines and standardised data-entry procedures were maintained throughout corpus construction to preserve internal consistency. Syntactic classification, etymological attribution, and lexical meanings were verified through multiple reference sources. In addition, the dataset was independently reviewed by two native speakers of each language, both university professors specialising in Bahasa Indonesia and Urdu. Their evaluation confirmed the accuracy of the lexical classifications and the adequacy of the corpus structure. Disagreements were resolved through joint discussion until consensus was reached.

This verification procedure reduced the possibility of individual coding bias. Replicability of the dataset was further supported through the specification of operational definitions for each recorded variable.

The criteria used to classify lexical correspondences as cognates, partial cognates, or false friends are presented in Table 1.

Table 1
Classification criteria for cognates, partial cognates, and false friends

CRITERIA	COGNATES (C)	PARTIAL COGNATES (PC)	FALSE FRIENDS (FF)
Phonological similarity	✓	✓	✓
Semantic relationship	Overlap in core meaning	Partial overlap; semantic narrowing, widening, or extension	No semantic overlap
Functional equivalence	Equivalent usage in both languages	Partially equivalent; context-dependent	Non-equivalent usage
Etymology	Often shared, not obligatory	Often shared, meanings diverged	Generally different
Lexical transfer	High; positive transfer	Moderate; may assist comprehension but risks misinterpretation	Low; negative transfer

Each lexical pair was assigned to one of three operational categories — cognates, partial cognates, or false friends — on the basis of combined phonological and lexical-semantic comparison.

Lexical pairs were classified as cognates when phonological similarity coincided with substantial overlap in core denotational meaning and comparable functional usage across the two languages. Shared etymological origin, frequently from Arabic or Persian, was treated as supporting evidence but was not required for classification (Crystal, 2011; Ringbom, 2007).

Lexical pairs displaying phonological similarity together with partial semantic overlap were classified as partial cognates. In such cases, at least one shared meaning was identified, while additional meanings differed through semantic narrowing, semantic widening, or specialised contextual usage in one of the languages. These items therefore represent intermediate lexical correspondence.

Lexical pairs showing phonological similarity but no semantic overlap in denotational or connotational meaning were classified as false friends. In these cases, similar phonological forms correspond to unrelated lexical meanings. Independent semantic development or divergent etymological histories supported this classification (James, 1980; Ringbom, 2007).

A representative sample of the dataset is presented in Table 2, while the complete corpus of 326 lexical pairs is publicly available at <https://doi.org/10.5281/zenodo.18739704>.

4. RESULTS

The classification of the 326 lexical pairs into cognates, partial cognates, and false friends reveals distinct distributions across the dataset (Table 2). A contrastive lexical-semantic analysis was employed to examine cog-

nates, partial cognates, and false friends between Urdu and Bahasa Indonesia, using a manually compiled corpus of 326 phonologically similar lexical pairs. The study focuses on form—meaning relationships, including semantic equivalence, partial overlap, and semantic divergence. Descriptive quantitative analysis, based on frequencies and percentages, was applied to analyse systematic lexical categorisation. Etymological analysis was conducted to account for the sources of cross-linguistic similarity and divergence.

4.1. Contrastive analysis categories

Lexical pairs exhibiting phonological similarity and partial semantic overlap, yet differing in additional meanings, semantic extensions, contextual restrictions, or pragmatic usage, were identified as partial cognates. These pairs share at least one meaning; however, one language displays either semantic widening, semantic narrowing, or specialised usage absent in the other. Partial cognates were therefore treated as representing an intermediate degree of similarity.

Phonologically similar lexical pairs showing little or no semantic overlap in either denotational or connotational meaning were categorised as false friends. In such cases, phonological similarity may mislead speakers or learners into assuming equivalence, despite the meanings being unrelated or contradictory. This classification is supported by independent semantic development or etymological divergence.

Clear patterns of similarity and divergence between Urdu and Bahasa Indonesia are presented in Table 3. Of the 326 phonologically similar lexical pairs, 225 items (69.1%) were classified as cognates (C), indicating a high degree of semantic equivalence alongside phonological similarity. These items exhibit stable meaning correspond-

Table 2
Sample representation of contrastive data

URDU SCRIPT	URDU (ROMAN)	BAHASA INDONESIA	SEMANTICS (URDU / BI)	IPA (URDU / - / BI)	SYNTAX (URDU / BI)	ETYMOLOGY (URDU / BI)	CA
ادب	Adab	adab	politeness, etiquette	/ʔa.ɖab/	N	AR	C
بحث	Baḥs	bahas	discussion	/ba:ʃ/ - /ba.has/	N / V	AR	PC
دکھ	Dukh	duka	sorrow	/ɖʊkʰ/ - /du.ka/	N	Sanskrit	C
دکان	dukān	dukan	shop	/du'ka:n/ - /'du.kan/	N	Persian	C
گیلا	gīlā	gila	wet / crazy	/ʔgi:la:/ - /ʔgi.la/	Adj	Sanskrit / Austronesian	FF
حاصل	Ḥaṣil	hasil	result	/ha:ʃɪ/ - /ha.sil/	N, Adj	AR	C
باتمی	Hāthī	hati	elephant / heart	/ʔha:ʔhɪ:/ - /ʔha.ti/	N	Indo-Aryan/ Austronesian	FF
اجازت	Ijāzat	ijazah	permission /	/i.ɖʒa:zat/ - /i.ɖʒa.zah/	N	AR	PC
کنجی	Kunji	kunci	key	/kun.ɖʒi:/ - /'kun.tʃi/	N	Persian	PC
لذت	Lazzat	Lezat	pleasure / delicious	/laz.zat/ - /lə.zat/	N / Adj	AR	PC
ماہر	māhir	mahir	expert	/ma:ɦɪr/ - /'ma.hir/	Adj	AR	C
میدان	Maidan	medan	field	/mɛ:da:n/ - /me.dan/	N	Persian	C
فیتہ	fiṭā	Pita	ribbon, tape	/fi:ʔa:/ - /'pi.ta/	N	Portuguese	C
سبب	Sabab	Sebab	cause, reason	/sa.bab/ - /sa.bab/	N	AR	C
تالی	Tālī	Tali	clapping / rope	/ta:li/ - /ta.li/	N	AR / Austronesian	FF

ence across both languages and generally retain comparable syntactic functions. The percentages reported refer exclusively to the curated dataset of phonologically similar words.

Partial cognates (PC) account for 60 lexical pairs (18.4%), which share a common semantic core but differ in meaning scope, usage, or contextual constraints. In many cases, these pairs share both phonological form and etymological origin; however, the observed partial overlap reflects independent processes of semantic extension or narrowing in the two languages. In contrast, 41 lexical pairs (12.5%) were categorised as false friends (FF). These items display notable phonological similarity but little or no semantic correspondence between the two languages. Most

phonologically similar lexical items are cognates, while a substantial proportion exhibits partial or complete semantic mismatch (Table 3).

4.2. Cognates

Cognates are characterised by semantic overlap, which enables straightforward categorisation. By contrast, partial cognates exhibit semantic narrowing or extension. True cognates constitute the largest category in the dataset from a semantic perspective. Examples include ادب—*adab/adab* (politeness), علم—*ilm/ilmu* (knowledge), ایمان—*iman/iman* (faith), کتاب—*kitab/kitab* (book), زکات—*zakat/zakat* (charity), مسجد—*masjid/masjid* (mosque), وقت—*waqt/waktu* (time), and انسان—*insan/insan* (human).

Table 3
Distribution of CA categories

CATEGORY	FREQUENCY	PERCENTAGE (%)
C	225	69.1%
PC	60	18.4%
FF	41	12.5%
Total	326	100%

Most cognates retain their original word class from a grammatical perspective. They function as nouns and adjectives and preserve the same syntactic roles in both languages. For instance, *ظاهر*—*zahir* and *واجب*—*wajib* function as

adjectives in both languages, while *حَق*—*hak* functions as a noun in both Urdu and Bahasa Indonesia. These borrowed cognates are used in Bahasa Indonesia in their original form with minimal recategorisation (Figure 2).

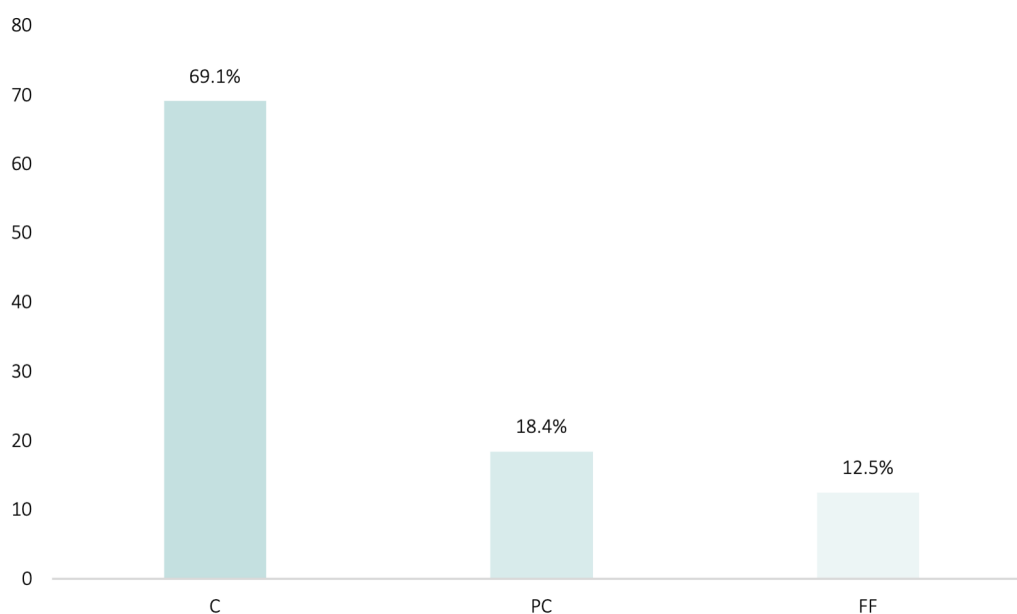


Figure 2. Distribution of cognate types across categories

Cognates exhibit phonological adaptation in Bahasa Indonesia, as certain lexical items undergo simplification of consonant clusters and loss of pharyngeal and emphatic consonants. For instance, *قبر*—*qabr* becomes *kubur*, *حَق*—*haqq* becomes *hak*, and *غائب*—*ghâ ib* becomes *gaib*. Etymologically, most of these cognates originate from Classical Arabic. These lexical items were transmitted either indirectly into Urdu via Persian or directly into Bahasa Indonesia through Islamic scholarship.

4.3. Partial Cognates

Partial cognates differ in semantic scope, collocational patterns, or pragmatic usage, despite overlapping core meanings across Urdu and Bahasa Indonesia. They often

undergo semantic shifts between the two languages. For example, *عادت*—*adat* refers to an individual habit in Urdu, whereas in Bahasa Indonesia it denotes a social custom or practice. Similarly, *علامت*—*alamat* means ‘a sign’ in Urdu but ‘an address’ in Bahasa Indonesia. The word *مقام*—*maqam/makam* refers to a place or status in Urdu, while in Bahasa Indonesia it denotes a grave. Likewise, *حال*—*hal* signifies a state or condition in Urdu but refers to a matter or problem in Bahasa Indonesia.

Some partial cognates exhibit category restriction or functional narrowing in Bahasa Indonesia. In Urdu, *اجازت*—*ijazat/ijazah* refers to permission, whereas in Bahasa Indonesia it denotes a certificate or diploma. Moreover, *عمل*—*amal* refers to any action in Urdu but is often restricted to

religious or moral actions in Bahasa Indonesia. Phonologically, partial cognates maintain close similarity, as in علم—*ilm* becoming *ilmu* in Bahasa Indonesia. Through the addition of final vowels or adaptation to a consonant–vowel syllable structure, Bahasa Indonesia accommodates these forms. Most partial cognates share Arabic origins. Figure 2 shows the distribution of cognates, partial cognates, and false friends in source languages. Items of Arabic origin dominate cognate categories, whereas false friends remain comparatively limited across all language groups.

4.4. False Friends

False friends are lexical items that diverge significantly in meaning despite strong formal similarity across languages. They exhibit clear semantic mismatches between Urdu and Bahasa Indonesia. For example, اگـر—*agar/agar* means *if* in Urdu but *so that* in Bahasa Indonesia, and بـچـہ—*bacha/baca* denotes a child in Urdu but functions as the verb *to read* in Bahasa Indonesia. Moreover, صـورـت—*surat/surat* refers to appearance in Urdu but means ‘a letter’ in Bahasa Indonesia. مـکـان—*makan/makan* means ‘house’ in Urdu but ‘to eat’ in Bahasa Indonesia, while جـوتـا—*juta/juta* denotes a shoe in Urdu but million in Bahasa Indonesia. Likewise, ہـاتـھی—*hathi/hati* means ‘elephant’ in Urdu but ‘heart’ in Bahasa Indonesia.

False friends often belong to different grammatical categories in the two languages and may result in syntactic conflict. For instance, بـچـہ—*bacha* is a noun in Urdu, whereas *baca* functions as a verb in Bahasa Indonesia; نـیک—*nek/naik* is an adjective in Urdu but a verb in Bahasa Indonesia; and مـنـڈی—*mandi/mandi* is a noun in Urdu but functions as a verb in Bahasa Indonesia.

False friends typically retain near-identical syllable structures and similar stress patterns. These lexical items generally originate from different linguistic sources despite their surface similarity. For example, بـدـن—*badan* derives from Persian in Urdu but from Austronesian sources in Bahasa Indonesia. Persian is the source of *dana* in Urdu, whereas it is derived from Sanskrit in Bahasa Indonesia. Similarly, بـازو—*baju* comes from Persian in Urdu but from Austronesian sources in Bahasa Indonesia.

4.5. Etymological distribution

Etymological analysis indicates that Arabic plays a dominant role, accounting for over three-quarters (72.4%) of the total lexical items. Persian represents 11.7% of the lexical sources. Indigenous Indo-Aryan lexemes account for 2.5%, while Sanskrit-derived items constitute 6.7%. A minor proportion (1.8%) originates from Portuguese (Table 4).

Table 4
Source language distribution of lexical items

ETYMOLOGY	NUMBER OF WORDS	PERCENTAGE (%)
Arabic	236	72.4%
Persian	38	11.7%
Sanskrit	22	6.7%
Portuguese	6	1.8%
Indo-Aryan	8	2.5%
Other languages	16	4.9%
Total	326	100%

Some Arabic-derived religious lexical pairs show high semantic stability in both languages, such as وفـات—*wafat/wafat* and تـوبـہ—*taubah/taubat*. The core meaning remains unchanged, although Bahasa Indonesia simplifies syllable structure and neutralises vowel length. Greater semantic divergence is observed in Arabic-derived abstract and administrative vocabulary. For example, the usage of اـخـتـیار—*ikhtiyār/ikhtiar* shifts toward effort-based or institutional

meanings in Bahasa Indonesia, reflecting partial semantic overlap. Similarly, حـکـم—*hukum/hukum* shows semantic narrowing, with ‘law’ in Bahasa Indonesia corresponding to ‘ruling/command’ in Urdu. The Persian-derived pair بـازار—*bāzār/pasar* demonstrates systematic phonological adaptation. While preserving the semantic equivalence ‘market’, it exhibits syllable restructuring in Bahasa Indonesia and consonant substitution, such as /b/ → /p/ and /z/ → /s/.

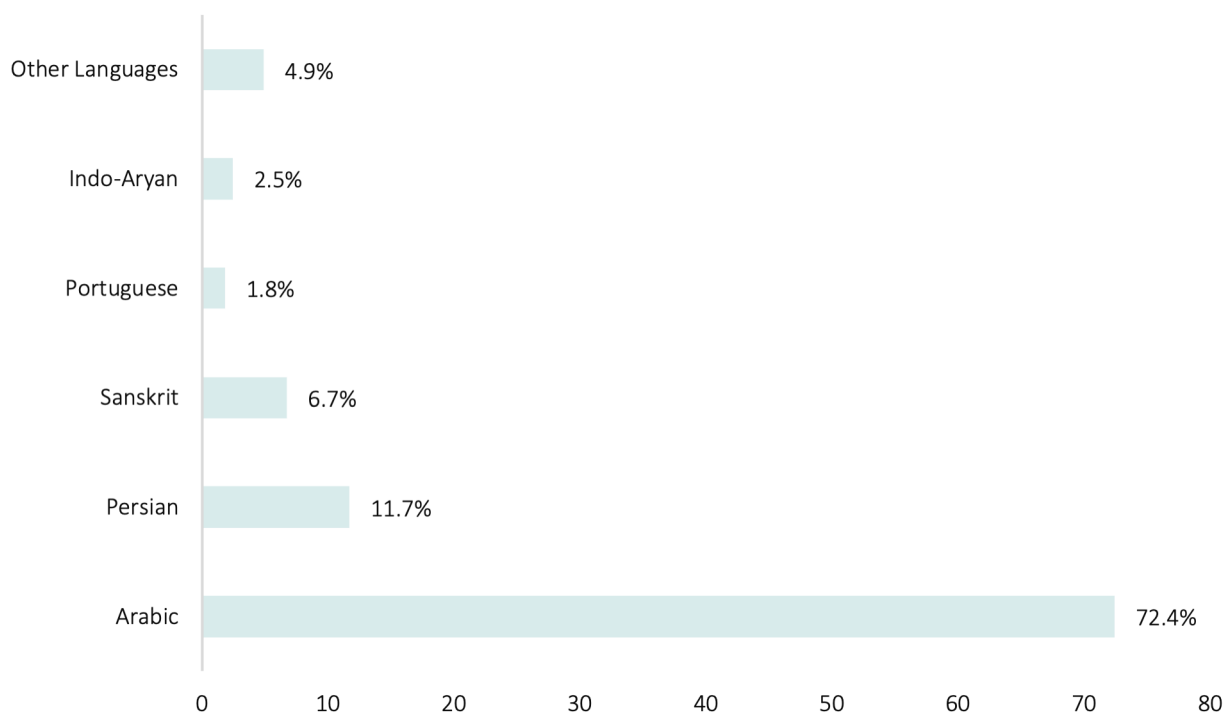


Figure 3. Distribution of lexical items by source language

Table 5 demonstrates the clear dominance of Arabic-origin lexical items in both languages. Of the 236 Arabic lexical items, 173 cognates (73.3%) and 41 partial cognates (17.4%) indicate a high degree of semantic transparency in Arabic-derived vocabulary. Conversely, the relatively small proportion of 22 false friends (9.3%) reflects a high level of semantic stability across both languages. A similar,

though less pronounced, pattern is observed in Persian-origin words, with 25 cognates (65.8%), 9 partial cognates (23.7%), and 4 false friends (10.5%), indicating comparatively strong semantic correspondence. Among the 22 Sanskrit-derived lexical pairs, cognates constitute the largest proportion (59.1%), followed by partial cognates (22.7%) and false friends (18.2%).

Table 5

Distribution of cognates, partial cognates, and false friends by source language

SOURCE LANGUAGE	COGNATES	PARTIAL COGNATES	FALSE FRIENDS	TOTAL
Arabic	173	41	22	236 (72.4%)
Persian	25	9	4	38 (11.7%)
Sanskrit	13	5	4	22 (6.7%)
Portuguese	5	0	1	6 (1.8%)
Indo-Aryan	6	1	1	8 (2.5%)
Other languages	3	4	9	16 (4.9%)
Total	225 (69.1%)	60 (18.4%)	41 (12.5%)	326

In contrast, lexical items from other languages show substantial semantic divergence, with partial cognates and false friends accounting for over two-thirds of this category. Persian-origin words occupy an intermediate position, contributing to all three categories, with a clear predominance of cognates. A similar pattern is observed in Indo-Aryan lexical items, where 75.0% are cognates and 12.5% each are partial cognates and false friends. Portuguese-derived lexical items also show a high proportion of cognates (5 items; 83.3%); however, the

limited sample size restricts broader generalisation. Most notably, the ‘Other Languages’ category exhibits the highest degree of semantic mismatch, with 56.3% false friends, 25.0% partial cognates, and only 18.7% cognates. Overall, Arabic-origin lexical items display the highest degree of semantic similarity, whereas items of mixed etymology show greater semantic divergence.

Figure 4 presents the distribution of cognates, partial cognates, and false friends across source languages (Figure 4).

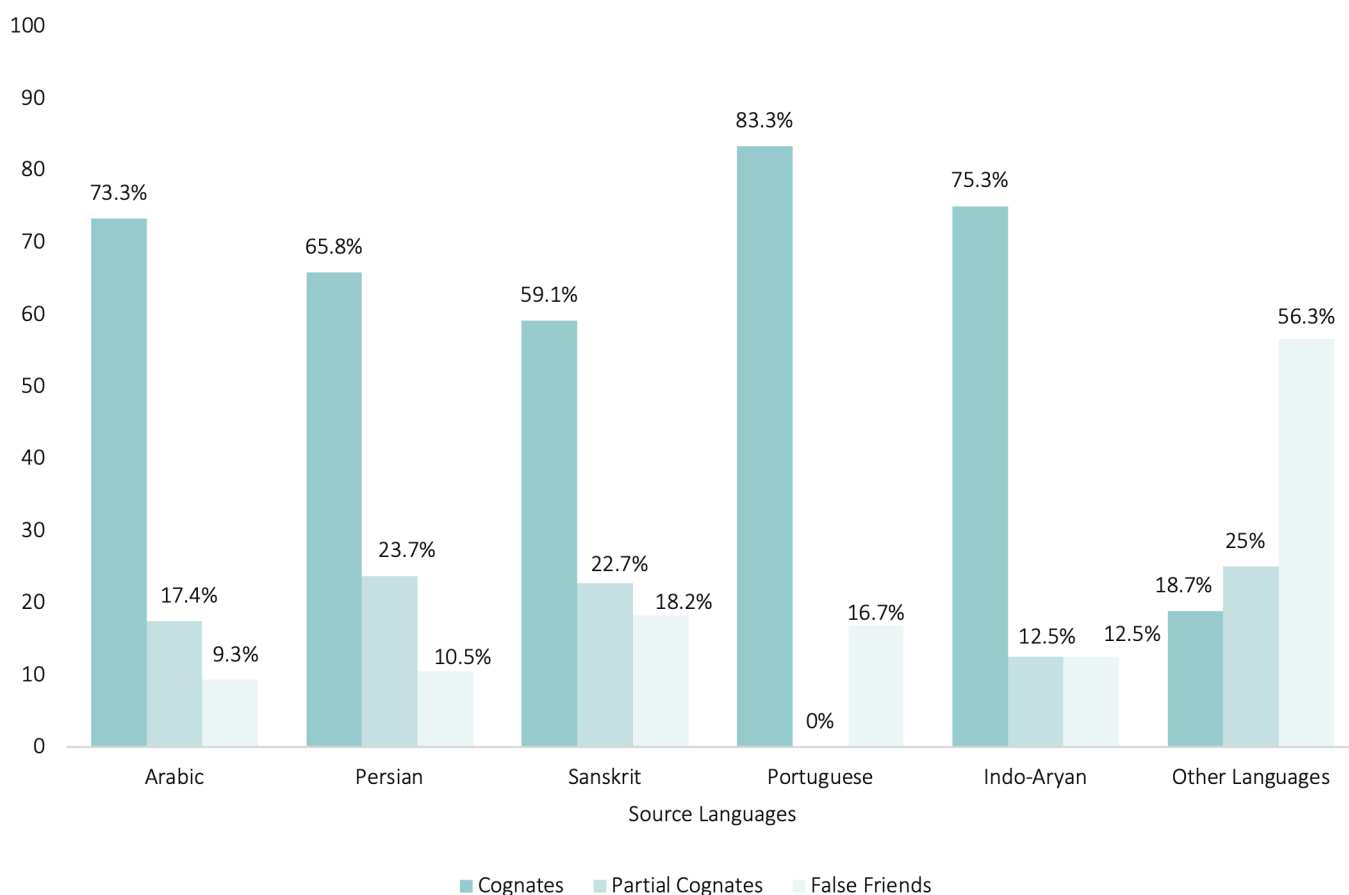


Figure 4. Distribution of cognates, partial cognates, and false friends by source language

The analysis of 326 lexical items indicates that cognates constitute the largest group, followed by partial cognates and false friends. The distribution by source language further shows that Arabic-derived items account for the majority of the dataset, with most of these classified as cognates. In contrast, items derived from Persian, Sanskrit, Indo-Aryan languages, and Portuguese represent a smaller proportion of the dataset and display greater variation in semantic correspondence. Furthermore, items classified under ‘Other Languages’ exhibit a comparatively higher proportion of false friends.

5. DISCUSSION

The aim of this study was to categorise and describe lexical similarities in phonologically similar lexical pairs at the surface level only, as it represents the first systematic contrastive lexical study between Bahasa Indonesia and Urdu.

The findings reveal patterns of lexical similarity between Urdu and Bahasa Indonesia.

The dataset may therefore serve as a foundation for future experimental, historical, and multi-level linguistic research.

5.1. Categorisation of similar sounding pairs into cognates, partial cognates, and false friends

The results indicate that phonologically similar lexical pairs can be reliably categorised into cognates, partial cognates, and false friends. This classification is based solely on semantic equivalence. The high proportion of cognates identified in this study suggests that shared cultural and religious transmission channels significantly increase lexical overlap (Haque, 2020; Khalid et al., 2025; Khan, 2014). Historical, religious, and sociolinguistic contact through intermediary languages, particularly Arabic and Persian, account for these similarities. Although numerically fewer, false friends present the greatest risk of misclassification, as surface similarity masks semantic divergence. This tripartite categorisation aligns with cross-linguistic models of lexical comparison (Jarvis & Pavlenko, 2008; Odlin, 1989) and supports the contrastive analysis hypothesis that high degrees of formal and semantic similarity exert greater influence than differences in historically related lexical systems (Lado, 1957; Ringbom, 2007).

Similarity in form does not necessarily entail similarity in meaning, as reflected in the 12.5% of false friends in the dataset. Lexical pairs such as تالی—*Tālī/tālī* (*clapping/rope*), گیلہ—*gīlā/gīlā* (*wet/crazy*), and ہاتھی—*Hāthī/hatī* (*elephant/heart*) exhibit phonological similarity but semantic divergence and represent clear instances of false friends. Surface similarity is therefore informative but limited and may be misleading in cross-linguistic communication. In addition to fully equivalent or entirely distinct borrowed items, meanings may shift gradually through semantic extension or narrowing. As lexical borrowing is a gradual process, careful contrastive categorisation is required to identify potential cross-linguistic ambiguities.

The categories were defined on the basis of established linguistic theory to reduce ambiguity (James, 1980; Ringbom, 2007). Initial observations suggest that lexemes with minor phonological variation tend to retain more stable meanings, whereas those with greater phonological variation may undergo semantic shift. However, no statistical analysis of pronunciation was conducted. Future research may examine this relationship using statistical modelling. The dataset includes lexemes with identical pronunciation, which may bias the results toward convergence. Hence, the frequency of such similarities across the full vocabularies of Bahasa Indonesia and Urdu cannot be determined. Future studies may employ computational methods or larger, frequency-based datasets to more accurately estimate the prevalence of these similarities.

5.2. Similar-sounding lexical pairs as a source of positive and negative lexical transfer

The results indicate that phonologically similar lexical pairs, particularly cognates, play a role in facilitating positive lexical transfer between Urdu and Bahasa Indonesia.

Cognates support positive transfer by enabling learners to access meanings with minimal cognitive load. Psycholinguistic research has shown that cognate recognition accelerates vocabulary acquisition and reading comprehension in second language learning, in line with the present findings (Babel & Fricke, 2023; Feldman & Milin, 2018; Orden et al., 1990; Rommetveit, 2014; Weber & Cutler, 2006). Positive transfer is also supported by partial cognates, although in a more limited and conditional manner. Their shared semantic core allows for basic comprehension; however, semantic shifts may lead to pragmatic ambiguity if learners rely solely on form-based interpretation. Although less frequent, false friends significantly hinder comprehension, particularly in spoken discourse. Cognates and partial cognates collectively constitute nearly 88% of the dataset, suggesting that lexical similarity between Urdu and Bahasa Indonesia generally facilitates comprehension rather than causing ambiguity. This finding is consistent with the contrastive analysis hypothesis, according to which similarity promotes learning, whereas dissimilarity increases ambiguity (Lado, 1957).

The classification of phonologically similar lexemes does not guarantee full comprehension, as this depends on multiple linguistic and cognitive factors (Otwinowska & Szewczyk, 2019). Nevertheless, the dataset may be of value for BIPA teaching, as it enables learners to identify semantic convergence and divergence and to infer meanings. No classroom experiment was conducted in this study. Cognates may support beginner learners by increasing confidence and reducing anxiety through positive transfer, whereas partial cognates and false friends may raise awareness and help prevent ambiguity resulting from negative transfer.

The corpus developed in this study may serve as a practical lexical resource for cognate identification activities in BIPA classrooms, particularly in view of the growing Pakistani diaspora in Indonesia and the international presence of BIPA programmes. Teachers may design structured comparison tasks based on the categorisation of cognates, partial cognates, and false friends, given that speakers often rely on phonological similarity in verbal communication (Lobley et al., 2005; Ringbom, 2007). For example, classroom activities may include pronunciation comparison drills, matching tasks, and semantic mapping exercises. Although the corpus does not claim to directly improve BIPA teaching for Urdu-speaking learners, it provides a structured basis for incorporating cognates into classroom practice. The study thus establishes a foundation that may be developed further through experimental and psycholinguistic research. Psycholinguistic evidence indicates that phonological similarity influences both accurate comprehension and misperception, affecting access to lexical meaning. Cognates facilitate positive transfer by reducing processing difficulty due to cross-linguistic similarity.

Phonological similarity may support comprehension but may also contribute to misunderstanding. Partial cognates enable partial transfer, consistent with research on cross-linguistic lexical ambiguity (James, 1980; Janke & Kolokonte, 2015). False friends present a high risk of negative transfer and comprehension breakdown, as similarity in form activates incorrect lexical associations (Jafarova, 2025; Weber & Cutler, 2006). While partial overlap may initially facilitate comprehension, it often increases the risk of misinterpretation in academic, legal, and cross-linguistic contexts. False friends therefore pose a significant challenge for language learners. When learners rely on surface form rather than contextual meaning, deceptive similarity frequently results in negative transfer. Such mismatches require immediate interpretative adjustment. Partial cognates likewise require explicit semantic instruction to prevent misuse, whereas false friends necessitate contrastive instruction to avoid the fossilisation of errors. The findings support the contrastive analysis hypothesis: high similarity facilitates learning, partial similarity leads to interference, and formal similarity without semantic correspondence increases error-proneness in second language acquisition.

5.3. Etymological sources

The predominance of Arabic-origin words (72.4%) reflects the historical spread of Islam and the transmission of religious, legal, and scholarly vocabulary into both Urdu and Bahasa Indonesia (Almurashi, 2024; Billah, 2018; Islam, 2012). The results show that Arabic-derived items contribute substantially to the cognate category (173 out of 236 items), indicating a high degree of semantic stability across the two languages. Persian-derived items also show a strong tendency toward cognates and partial cognates, reflecting their role as a mediating language within South Asian Islamic and literary traditions. By contrast, items derived from Sanskrit and Austronesian sources account for the highest proportion of false friends, indicating semantic divergence across different etymological origins.

Previous contact-linguistic studies have shown that shared source languages associated with stable semantic domains, such as religion, governance, and education, significantly increase cross-linguistic comprehension (Gooksens et al., 2015; Matras, 2020). This suggests that lexical convergence is more pronounced in formal and culturally transmitted domains than in everyday communication. The lexical items examined are predominantly located in religious, abstract, and institutional domains and exhibit near-identical meanings. The sustained influence of Arabic is evident in both Urdu and Bahasa Indonesia, largely due to the role of Islam, religious education, administration, and classical scholarly traditions. Persian has exerted greater influence on Urdu, owing to historical processes of Persianisation during Muslim rule in South Asia and its

indirect transmission into the Malay–Indonesian lexicon. Other languages, including Austronesian, Greek, Dravidian, and other Indo-European sources, account for 4.9% of the dataset, reflecting layers of pre-Islamic heritage as well as borrowings associated with trade and colonisation.

The results suggest that Arabic serves as a primary source of positive lexical transfer between Urdu and Bahasa Indonesia, while Persian and other languages play complementary roles. Lexical items tend to preserve their original meanings when embedded within shared Islamic discourse traditions. However, shared etymology does not necessarily entail identical meaning or form in present usage, despite the predominance of Arabic- and Persian-derived cognates. Phonologically similar lexical pairs, borrowed into Bahasa Indonesia and Urdu from different sources or undergoing structural adaptation, may develop divergent outcomes. Overall, Arabic-derived vocabulary plays a central role in facilitating positive transfer, whereas lexical items from heterogeneous sources are more prone to partial or false equivalence. Furthermore, while identical lexemes may facilitate communication in religious and formal contexts, it remains unclear whether this effect extends to everyday language use.

6. CONCLUSION

This study presents a systematic contrastive account of phono-lexical similarity between Urdu and Bahasa Indonesia. Phonologically similar lexical pairs are classified into cognates, partial cognates, and false friends on the basis of semantic equivalence. This classification clarifies similarity and divergence between the two languages. The findings indicate areas of potential positive and negative lexical transfer. Semantic stability in cognates supports comprehension. Partial overlap and semantic divergence introduce risks of ambiguity and misinterpretation.

The study identifies the contribution of etymological roots, particularly the influence of Arabic, together with Persian and other languages, in the formation of lexical similarity, and offers a more precise account of lexical similarity in contrastive linguistics. The results confirm that phonological similarity does not necessarily entail semantic equivalence. The study contributes to contrastive linguistics and contact linguistics and presents empirical evidence of the influence of cultural and historical factors on lexical convergence and divergence in unrelated languages.

The findings have implications for second language acquisition research and the BIPA curriculum. The analysis is limited to a corpus-based examination at the lexical surface level and does not include experimental validation of processing or comprehension effects. The study provides a basis for further research on cross-linguistic transfer, phonological similarity, and semantic change in languages with contact histories.

References

- Abbas, S. N., Akram, H., & Ranra, B., (2022). In quest of language and national identity: A case of Urdu language in Pakistan. *International Journal of Business and Management Sciences*, 3(2), 48–66.
- Al-Malki, E. A., Majid, N. A., & Omar, N. A. M. (2014). Generic reference in English, Arabic and Malay: A cross linguistic typology and comparison. *English Language Teaching*, 7(11), 15–27. <https://dx.doi.org/10.5539/elt.v7n11p15>
- Alfa, M. S., Dollah, H., & Abdullah, N. (2015). Analysis of the impact of Arabic-Malay bilingual dictionaries in Malaysia. *UMRAN – Journal of Islamic and Civilizational Studies*, 2(3), 37–45. <https://dx.doi.org/10.1113/umran2015.2n3.29>
- Almurashi, W. (2024). Exploring the lexical influence of Arabic on Bahasa Indonesia: Phonetically transcribed. *AWEJ for Translation & Literary Studies*, 8(4), 20–30. <https://doi.org/10.24093/awejtls/vol8no4.3>
- Ambreen, S., & To, C. K. S. (2025). Review of the phonological system of contemporary Urdu spoken in Pakistan. *International Journal of Speech-Language Pathology*, 27(1), 101–112. <https://dx.doi.org/10.1080/17549507.2024.2324905>
- Awla, H. A., & Azeez, R. A. (2021). False cognates and friends between English and Kurdish. *International Journal of Social Sciences & Educational Studies*, 8(3), 170–182. <https://doi.org/10.23918/ijsses.v8i3p170>
- Babel, M., & Fricke, M. (2023). Sound structure and the psycholinguistics. In D. Kavitskaya & A. C. L. Yu (Eds.), *The life cycle of language: Past, present, and future* (pp. 339–353). Oxford Academic. <https://doi.org/10.1093/oso/9780192845818.003.0021>
- Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). MIT Press.
- Billah, A. M. M. A. (2018). Linguistic influence of Persian on South Asian languages: Special reference to Urdu. *SACS Special Issue*, 2018, 98–110.
- Bordag, D., Gor, K., & Opitz, A. (2022). Ontogenesis model of the L2 lexical representation. *Bilingualism: Language and Cognition*, 25(2), 185–201. <https://doi.org/10.1017/S1366728921000250>
- Carroll, S. E. (1992). On cognates. *Second Language Research*, 8(2), 93–119. <https://dx.doi.org/10.1177/026765839200800201>
- Carvalho, A., & Child, M. (2018). Expanding the multilingual repertoire: Teaching cognate languages to heritage Spanish speakers. In K. Potowski (Ed.), *The Routledge handbook of Spanish as a heritage language* (pp. 420–432). Routledge. <https://dx.doi.org/10.4324/9781315735139>
- Chesterman, A. (1998). *Contrastive functional analysis*. John Benjamins Publishing Company. <https://doi.org/10.1075/pbns.47?locatt=mode:legacy>
- Crystal, D. (2011). *A dictionary of linguistics and phonetics* (6th ed.). Wiley-Blackwell.
- Detey, S., De Fino, V., & Fontan, L. (2025). Phonetic or morpho-lexical issues in L2 speech? A pilot study of ambiguity assessment for pedagogical purposes among Japanese learners of French. *Language Testing in Asia*, 15(1), Article 65. <https://dx.doi.org/10.1186/s40468-025-00394-5>
- Díaz-Galaz, S., & Torres, A. (2019). Comprehension in interpreting and translation: Testing the phonological interference hypothesis. *Perspectives*, 27(4), 622–638. <https://doi.org/10.1080/0907676X.2019.1569699>
- Divasson, L., & León, I. (2002). Medical English and Spanish cognates: Identification and classification. *ASp*, 35, 73–87. <https://doi.org/10.4000/asp.1607>
- Ellis, R. (1994). *The study of second language acquisition*. Oxford University Press.
- Elvin, J., Escudero, P., & Vasiliev, P. (2014). Spanish is better than English for discriminating Portuguese vowels: Acoustic similarity versus vowel inventory size. *Frontiers in Psychology*, 5, Article 1188. <https://dx.doi.org/10.3389/fpsyg.2014.01188>
- Farashah, M. (2010). Persian, Urdu, and Pashto: A comparative orthographic analysis. *Writing Systems Research*, 2(1), 9–23. <https://doi.org/10.1093/wsr/wsq005>
- Feldman, L. B., & Milin, P. (2018). Psycholinguistic studies of word morphology and their implications for models of the mental lexicon and lexical processing. In R. Berthiaume, D. Daigle, & A. Desrochers (Eds.), *Morphological processing and literacy development* (pp. 16–47). Routledge. <https://doi.org/10.4324/9781315229140>
- Fisiak, J. (1981). *Contrastive linguistics and the language teacher*. Pergamon.
- Ghazali, S. A. B. M. (2022). Contrastive analysis of Arabic and Malay for adjective phrases in short stories. *European Journal of Language and Literature Studies*, 8(1), 82–98.
- Gooskens, C. (2024). *Mutual intelligibility between closely related languages* (Vol. 30). Walter de Gruyter. <https://doi.org/10.1515/9783111134697>
- Gooskens, C., & van Heuven, V. J. (2021). Mutual intelligibility. In M. Zampieri & P. Nakov (Eds.), *Similar languages, varieties, and dialects: A computational perspective* (pp. 51–95). Cambridge University Press. <https://doi.org/10.1017/9781108565080.006>
- Gooskens, C., van Bezooijen, R., & van Heuven, V. (2015). Mutual intelligibility of Dutch-German cognates by children: The devil is in the detail. *Linguistics*, 53(2), 255–283. <https://doi.org/10.1515/ling-2015-0002>
- Granger, S., & Lefer, M. A. (2020). Introduction: A two-pronged approach to corpus-based crosslinguistic studies. *Languages in Contrast*, 20(2), 167–183. <https://dx.doi.org/10.1075/lic.00014.int>
- Gumperz, J. J. (1964). Linguistic and social interaction in two communities. *American Anthropologist*, 66(6), 137–153. https://dx.doi.org/10.1525/aa.1964.66.suppl_3.02a00100
- Hamann, S., & Sennema, A. (2005). Acoustic differences between German and Dutch labiodentals. *ZAS Papers in Linguistics*, 42, 33–41. <https://doi.org/10.21248/zaspil.42.2005.272>

- Haque, S. (2020). Language use and Islamic practices in multilingual Europe. *Signs and Society*, 8(3), 401–425. <https://doi.org/10.1086/710157>
- Haspelmath, M. (2009). Lexical borrowing: Concepts and issues. In M. Haspelmath & U. Tadmor (Eds.), *Loanwords in the world's languages: A comparative handbook* (pp. 35–54). De Gruyter Mouton. <https://dx.doi.org/10.1515/9783110218442>
- Hasselgård, H. (2020). Corpus-based contrastive studies: Beginnings, developments and directions. *Languages in Contrast*, 20(2), 184–208. <https://dx.doi.org/10.1075/lic.00015.has>
- Islam, R. A. (2012). *The morphology of loanwords in Urdu: The Persian, Arabic and English strands* (Doctoral dissertation, Newcastle University). Newcastle University Thesis Archive. <https://hdl.handle.net/10443/1407>
- Jafarova, K. A. (2025). False friends in translation: A lexical source of interference in English–Azerbaijani contexts. *Journal of Linguistics, Culture and Communication*, 3(1), 210–225. <https://dx.doi.org/10.61320/jol-cc.v3i1.210-225>
- James, C. (1980). *Contrastive analysis*. Longman.
- Janke, V., & Kolokonte, M. (2015). False cognates: The effect of mismatch in morphological complexity on a backward lexical translation task. *Second Language Research*, 31(2), 137–156. <https://dx.doi.org/10.1177/0267658314545836>
- Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge. <https://dx.doi.org/10.4324/9780203935927>
- Karam, F. X. (1979). Processes of increasing mutual intelligibility between language varieties. *International Journal of the Sociology of Language*, 22, 115–138. <https://dx.doi.org/10.1515/ijsl.1979.22.115>
- Khalid, A., Anwar, P., & Khan, A. A. (2025). A sociolinguistics study of Urdu-Hindi lexical borrowings in English from perspectives of language contact. *Qlantic Journal of Social Sciences and Humanities*, 6(1), 447–461. <https://doi.org/10.55737/qjssh.vi-i.25399>
- Khan, B., Asad, M., & Zahid, M. (2024). Consonantal variation of Hindi-Urdu loanwords in standard English: A phonological analysis. *Language, Technology, and Social Media*, 2(2), 145–159. <https://dx.doi.org/10.70211/ltsm.v2i2.77>
- Khan, I. A. (2014). Lexical borrowings from Arabic and semantic change in Urdu: A cross linguistic analysis. *Hope Journal of Research*, 2(3), 47–66.
- Khan, T. A. (2021). Phonemic variations in similar words of Urdu and Turkish. *Journal of Language and Linguistic Studies*, 17(1), 517–533.
- Khrisat, A. (2014). Language's borrowings: The role of the borrowed and Arabized words in enriching Arabic language. *American Journal of Humanities and Social Sciences*, 2(2), 133–142.
- Lado, R. (1957). *Linguistics across cultures: Applied Linguistics for language teachers*. University of Michigan Press.
- Lauder, A., & Purwo, B. (2024). English in multicultural Indonesia. In A. J. Moody (Ed.), *The Oxford handbook of South-east Asian Englishes* (pp. 231–242). Oxford Academic.
- Lobley, K. J., Baddeley, A. D., & Gathercole, S. E. (2005). Phonological similarity effects in verbal complex span. *The Quarterly Journal of Experimental Psychology*, 58(8), 1462–1478. <https://dx.doi.org/10.1080/02724980443000700>
- Manfredi, S. (2017). Arabic as a contact language. In E. Benmamoun & R. Bassiouney (Eds.), *The Routledge handbook of Arabic linguistics* (pp. 407–420). Routledge. <https://doi.org/10.4324/9781315147062>
- Marecka, M., Szewczyk, J., Otwinowska, A., Durlik, J., Forys-Nogala, M., Kutylowska, K., & Wodniecka, Z. (2021). False friends or real friends? False cognates show advantage in word form learning. *Cognition*, 206, Article 104477.
- Matras, Y. (2020). Theorizing language contact: From synchrony to diachrony. In R. D. Janda, B. D. Joseph, & B. S. Vance (Eds.), *The handbook of historical linguistics* (pp. 375–392). Wiley. <https://dx.doi.org/10.1002/9781118732168.ch18>
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Mella, S. M. (2024). An approach to the lexical ambiguity caused by false cognates in Spanish L2: A corpus-based exploratory study. *Studia Linguistica*, 78(1), 186–205. <https://doi.org/10.1111/stul.12225>
- Müller, K. (2005, September 4–8). Revealing phonological similarities between German and Dutch. In *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)* (pp. 1609–1612). ISCA. <https://dx.doi.org/10.21437/Interspeech.2005-469>
- Nugraheni, A. S., Lestari, S., Husain, A. P., & Susanto, E. (2021). Development of online SABASIA (Suka Bahasa Indonesia) teaching materials for Indonesian language learning for international students as a pioneer for FIBAA international accredited study programs at State Islamic Universities. *Jurnal Pendidikan Islam*, 10(1), 133–158. <https://dx.doi.org/10.14421/jpi.2021.101.133-158>
- Odlin, T. (1989). *Language transfer* (Vol. 27). Cambridge University Press.
- Orden, V. G. C., Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review*, 97(4), 488–522. <https://psycnet.apa.org/doi/10.1037/0033-295X.97.4.488>
- Otwinowska, A., & Szewczyk, J. M. (2019). The more similar the better? Factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism*, 22(8), 974–991. <https://doi.org/10.1080/13670050.2017.1325834>
- Pangestika, E., Musthofa, T., & Nasiruddin, N. (2023). Differences in Arabic-Indonesian vocabulary absorption in religious terms: Phonological studies. *Al-Irfan: Journal of Arabic Literature and Islamic Studies*, 6(1), 190–207. <https://doi.org/10.58223/alirfan.v6i1.6797>
- Platzack, C. (1987). The Scandinavian languages and the null-subject parameter. *Natural Language & Linguistic Theory*, 5(3), 377–401.

- Plunkett, K., & Strömqvist, S. (2022). The acquisition of Scandinavian languages. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition* (pp. 457–556). Psychology Press. <https://dx.doi.org/10.4324/9781315808208>
- Ramírez, G., Chen, X., & Pasquarella, A. (2013). Cross-linguistic transfer of morphological awareness in Spanish-speaking English language learners: The facilitating effect of cognate knowledge. *Topics in Language disorders*, 33(1), 73–92. <https://dx.doi.org/10.1097/TLD.0b013e318280f55a>
- Raza, A. (2017). A review on Urdu language parsing. *International Journal of Advanced Computer Science and Applications*, 8(4), 93–97. <https://dx.doi.org/10.14569/IJACSA.2017.080413>
- Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning*. Multilingual Matters.
- Rommetveit, R. (2014). Words, meaning, and messages: Theory and experiments in psycholinguistics. Academic Press.
- Sato, M., Thorne, S. L., Michel, M., Alexopoulou, T., & Hellermann, J. (2025). Language, people, classrooms, world: Blending disparate theories for united language education practices. *The Modern Language Journal*, 109(S1), 15–38. <https://doi.org/10.1111/modl.12976>
- Sulistiyono, Y. (2025). Dialectal divergence in Alorese: Evidence from lexical, phonological, and morphological variation across Alor and Pantar. *Linguistik Indonesia*, 43(2), 243–261. <https://dx.doi.org/10.26499/li.v43i2.840>
- Sunderman, G., & Schwartz, A. I. (2008). Using cognates to investigate cross-language competition in second language processing. *TESOL Quarterly*, 42(3), 527–536. <https://doi.org/10.1002/j.1545-7249.2008.tb00145.x>
- Susanto, G., Pickus, D., Espree-Conaway, D., Suparmi, Rusiandi, A., & Noviyya, H. (2024). Indonesian language policy and perspectives on its implementation in promoting Bahasa Indonesia as an international language. *Cogent Arts & Humanities*, 11(1), Article 2364511. <https://doi.org/10.1080/23311983.2024.2364511>
- Taylor, C., & del Fante, D. (2020). Comparing across languages in corpus and discourse analysis: Some issues and approaches. *Meta*, 65(1), 29–50. <https://dx.doi.org/10.7202/1073635ar>
- Thomason, S. G. (2019). Historical linguistics since 1968: On some of the causes of linguistic change. In H. C. Boas & M. Pierce (Eds.), *New directions for historical linguistics* (pp. 110–131). Brill. https://dx.doi.org/10.1163/9789004414075_006
- Versteegh, K. (2001). Linguistic contacts between Arabic and other languages. *Arabica*, 48(4), 470–508. <https://doi.org/10.1163/157005801323163825>
- Weber, A., & Cutler, A. (2006). First-language phonotactics in second-language listening. *Journal of the Acoustical Society of America*, 119(1), 597–607. <https://doi.org/10.1121/1.2141003>
- Zainuri, M. (2019). Perkembangan Bahasa Arab di Indonesia. *Tarling: Journal of Language Education*, 2(2), 231–248. <https://doi.org/10.24090/tarling.v2i2.2926>

ABOUT THE AUTHORS

Hafiza S. Mansoor

Doctoral Researcher,

Department of English Education, Faculty of Teacher Training and Education

Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

Postal address: Jl. Ahmad Yani, Pabelan, Kartasura, Kab. Sukoharjo, Jawa Tengah, 57169, Indonesia

Email: Q300249009@student.ums.ac.id

ORCID ID: <https://orcid.org/0009-0006-4418-6160>

Harun J. Prayitno

PhD in Indonesian Language and Literature Education, Professor, Rector

Department of Indonesian Language and Literature, Faculty of Teacher Training and Education

Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

Postal address: Jl. Ahmad Yani, Pabelan, Kartasura, Kab. Sukoharjo, Jawa Tengah, 57169, Indonesia

Email: hjp220@ums.ac.idORCID ID: <https://orcid.org/0000-0002-4997-5891>

Laili E. Rahmawati

PhD in Indonesian Language and Literature Education, Associate Lecturer

Department of Indonesian Language and Literature, Faculty of Teacher Training and Education

Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

Postal address: Jl. Ahmad Yani, Pabelan, Kartasura, Kab. Sukoharjo, Jawa Tengah, 57169, Indonesia

Email: ler211@ums.ac.idORCID ID: <https://orcid.org/0000-0003-2453-7809>