Building a Multilingual Vocabulary Database: A Comprehensive Study of 24 Global and Local Languages

Muhammad Farukh Arslan[1] and Mehtab Ahmed[2]

https://doi.org/10.62345/jads.2025.14.2.103

Abstract

The ability to communicate in more than one language is becoming increasingly important in a globalised society, and it is closely related to vocabulary learning, which is central to communicative competence —a multifaceted notion defined by Hymes (1967), of which 'grammatical competence' includes vocabulary as one of its components. In light of the significance of vocabulary for language teaching and learning, this research aimed to design an extensive vocabulary learning application for 20 international and six local Pakistani languages. At the first step, a corpus of the Urdu language was compiled, and a wordlist was prepared based on that corpus using corpus analysis software. Subsequently, a total of the most common 1,000 content words were selected and organised semantically. The top 18 most widely spoken languages in the world were obtained, and six indigenous languages from Pakistan were also included. Using AI, the vocabulary elements were translated into all these languages, and the translated texts were proofread using Google and Google Images. Translated words with accompanying pictures were also added. This database will facilitate the cross-examination of areas in phonology, morphology, translation, and culture.. Blended learning outlines an international education approach that alternates between mobility and online learning for teachers, students, and tourists in the previously chosen languages.

Keywords: Vocabulary Database, Communicative Competence, Global Languages.

Introduction

Language plays a crucial role in human communication and cultural exchange. In today's interconnected world, multilingualism has become an essential skill. Effective language acquisition depends significantly on vocabulary learning, which is a key component of communicative competence, as defined by Hymes (1967). In this framework, grammatical competence, one of the four major components of communicative competence, includes vocabulary as a fundamental element. Recognising the importance of vocabulary for language learning and teaching, this study aims to develop a multilingual vocabulary database encompassing twenty international and six local Pakistani languages. The goal of this initiative is to enhance comparative linguistic studies, support the documentation of languages, and promote a more comprehensive general understanding of linguistic diversity.

Despite the growing demand for multilingual education and cross-cultural communication, existing vocabulary learning tools often focus on dominant global languages while neglecting

[1]*Lecturer, Department of English, National University of Modern Languages, Faisalabad Campus.*
Email: Farukh.arslan@numl.edu.pk
[2]*MPhil, Department of Applied Linguistics, Government College University, Faisalabad, Pakistan.*
Email: mehtabjutt79004@gmail.com

Compliance with ethical standards: There are no conflicts of interest (financial or non-financial). This study did not receive any funding.

indigenous and underrepresented languages. Furthermore, existing multilingual lexical databases frequently struggle to capture cultural nuances, phonological variations, and morphological structures in diverse linguistic contexts (Giunchiglia et al., 2023). The absence of a comprehensive and systematically categorised vocabulary database that integrates both widely spoken and underrepresented languages presents a significant research gap. The crux of the matter is that improving the language experience, supporting the preservation of languages, and enabling one language to be analysed in cross-reference with another is crucial.

Prior studies (Mukherjee et al., 2023; Bella et al., 2022) have highlighted significant concerns regarding multilingual lexical databases, which encompass the co-cultural representation of words across contexts and accurate translation between languages. While the Universal Knowledge Core (UKC) has increasingly been utilised for many language demarcation products, there is still a lack of vocabulary database structuring, semantic organisation, and cultural consideration. Earlier, researchers have demonstrated some sociocultural biases of language models and indicated that Liang et al.'s (2024) count of existing computational representations of linguistic diversity is inadequate. The study aims to bridge the gap by complementing already underrepresented languages while employing AI-aware techniques for accurate translations and vocabulary classification.

The primary aim of this research is to develop a vocabulary-learning application for multilingual environments that systematically organises vocabulary from 20 world languages and six local Pakistani languages. This will solve problems of comparative linguistics, language documentation, and inclusiveness. Indeed, AI-based translation and visual images will pave the way for new avenues in vocabulary learning, unlike the methodologies currently adopted in foreign-language applications.

1. To collect and subject the Urdu vocabulary corpus to analysis for multilingual translation.
2. An ordered standard wordlist of 1000 content words will be generated, whereby frequency is established by analysis of the corpus.
3. Translation and classification into 24 languages of selected vocabulary items, which shall include international languages and local languages of Pakistan.
4. Verify the translations using AI tools, Google Search, and Google Images.
5. Set up an interactive application for vocabulary learning, utilising images to enhance understanding of the experience.
6. Engage in linguistically interesting research comparisons on phonology, morphology, translation equivalence, and cultural semantics.

Recent research has given significant attention to several key issues regarding multilingual lexical representation, including bias in computational models, semantic mismatches, and underrepresentation of smaller languages (Giunchiglia et al., 2023). It is in this spirit that the Universal Knowledge Core (UKC) should be introduced: to enhance the representation of language diversity across more than 2,000 languages and facilitate cross-linguistic analysis (Bella et al., 2022). Such work and studies on sociocultural biases in NLP models would also be helpful, shedding light on the inequalities and inequities in which language technology manifests (Mukherjee et al., 2023). Describing polysemy networks across 61 languages, Liang et al. (2024) outlined that quite a fine array of cognition underlie certain of the semantic relations and that these require culturally aware lexical databases. These studies offer insights into the challenges that the representation of multilingual vocabulary could pose.

This study employs a multi-phase research methodology to develop and validate a multilingual vocabulary database:

1. *Corpus Compilation:* For identifying high-frequency content words, a complete, compact corpus of Urdu vocabulary was compiled.
2. *Wordlist Preparation and Selection:* One thousand most common content words were selected and semantically related.
3. *Language Selection and AI Integration:* This includes twenty global languages spoken worldwide and six indigenous languages of Pakistan, translated using AI-empowered translations validated through Google Search and Google Images.
4. *Visual Representation and Database Design:* The corresponding images were supplemented with the meanings of each word to facilitate understanding, especially for young learners and language enthusiasts.
5. *Development of Learning Application:* A user-friendly interface has been developed to facilitate free-flow interaction across languages and vocabulary categories, promoting a more interactive approach to language learning.

This research aims to bridge the gap in language learning through high-end AI-mediated translation tools from a visual learning perspective. Already, it addresses existing limitations in multilingual vocabulary databases for learning, linguistics research, and even cultural preservation. Moreover, the findings will serve as a stepping stone for future inquiries into multilingualism, cross-linguistic influence, and computational linguistics, as they attract an all-encompassing and inclusive approach to language learning and recording.

Language learning has gained more importance than ever before, as the world becomes increasingly interconnected compared to the past. This capability is built on the realisation of vocabulary, which is one of the aspects of Linguistic competence described by Hymes (1967). In this framework, 'grammatical competence' is one of the four main components, and 'vocabulary' is a portion of this component. To explore the importance of vocabulary in language teaching and learning, this research aimed to develop a comprehensive vocabulary learning application for eighteen international languages and six local languages of Pakistan.

This study focused on the development of a multilingual dictionary. The motivation behind the study was to gain a fundamental understanding of the vocabulary of different languages from around the world. As multilingual speakers in a global society, knowledge of other languages from around the world is a prerequisite for survival and success (González, 2019).

The motivation behind this project was curiosity about the vocabulary items of different languages from around the world. Applications like Doulingo are already available, which help in learning languages from around the world. However, the application developed in this project operates differently. It also included the local languages of stan and, worked on a lexicthe al level to provide readready-to-usewledge regarding the other languages of the world.

Literature Review

Theoretical Framework

According to new studies, there are both current problems and opportunities in the development of multilingual lexical databases. Most existing databases have thus far failed to adequately address various culturally specific words and their meanings, particularly those of non-dominant languages (Giunchiglia et al., 2023). The Universal Knowledge Core is designed to address this gap; it focuses on the issue of language diversity, encompassing over 2,000 languages, and thereby facilitates cross-lingual analysis (Bella et al., 2022). Other researchers have examined the sociocultural biases of language models in various languages, revealing a range of manifestations of these biases across different linguistic and cultural contexts (Mukherjee et al., 2023). Another

study has analysed polysemy networks in simple vocabularies of 61 languages, identifying metacognitive structures common to basic human experiences, such as body parts, motion, and physical features, that play a central semantic role (Liang et al., 2024). Collectively, all these studies highlight the importance of broad, developing, and culture-sensitive approaches in multilingual lexical research.

Language diversity and its computational representations are pressing topics in natural language processing (NLP), multilingual lexical databases, and the sociocultural biases inherent in linguistic technologies. Meanings between languages are processed using lexical databases; however, in most cases, the primary emphasis in such systems is on English or other dominant languages, which induces a filtration of meaning as well as sociocultural injustice (Giunchiglia et al., 2023). At this point, increasing reliance on AI-fueled language models will also require a more accurate depiction of lexical diversity and its specificisation to multilingual contexts (Bella et al., 2022). Research by Mukherjee et al. (2023) states that NLP models also reflect the existing biases of society, which can drastically affect issues such as fairness, inclusivity, and linguistic equity in language technologies. Much more than this would be required to truly address the problem, namely, a well-developed, integrated theoretical framework from the linguistic, cognitive, and computational perspectives.

Earlier studies have provided the paradigm with some valuable insights into multilingual lexical representation and the sociocultural dimension in computational linguistics. Whereas Giunchiglia et al. (2023) discuss the constitutive elements of meaning formation across languages in this specific area of computational lexical databases, Liang et al. (2024) explore how universal human cognitive experiences shape semantic networks across languages. Mukherjee et al. (2023) further demonstrate how the language processing done by computational models serves to reinforce social prejudices. Additionally, Batsuren et al. (2021) examined derivational and inflectional morphology, highlighting crucial shortcomings in NLP regarding the processing of linguistic diversity. An analysis of kinship systems was conducted by Khalilia et al. (2023), highlighting how cultural differences influence the organisation of lexical structure. The conjunction of these studies makes evident the complex interplay between computations, linguistic constructs, cognition, and social structures.

On the one hand, various limitations persist with existing computational models of the lexicon, despite remarkable advancements. Most lexical databases are English-centric, thereby erasing meanings in the context of interlingual debates (Giunchiglia et al. 2023). Polysemy networks provide vague cues to basic human cognition but have not yet been fully incorporated into computational models (Liang et al., 2024). Furthermore, NLP models tend to uphold societal biases and thereby sustain stereotypes across languages (Mukherjee et al., 2023). One central challenge facing modern NLP models is how to account for cross-linguistic morphological variation, which acts as an obstacle to acceptability when less-represented languages are at issue (Batsuren et al., 2021). Therefore, it becomes plausible to argue for the development of a holistic theoretical framework that incorporates multilingual lexical representation, cognate-linguistic meaning construction, and sociocultural awareness as components within the computational linguistic ambit.

This research proposes an integrated theoretical framework that encompasses multilingual lexical representation, cognitive linguistics, critical discourse analysis (CDA), morphological typology, and cultural linguistics. According to recent studies (Giunchiglia et al., 2023; Bella et al., 2022), meaning is encoded differently across languages and therefore requires a computational approach centred on concepts rather than words. Hence, NLP models should be designed to consider

interlingual meaning shifts to promote better linguistic accuracy. From a cognitive-linguistic perspective, Liang et al. further argue that language and mind construct and shape human cognition and conceptual structures, which in turn, affect the transmission of meaning across languages. The integration of polysemy networks and universal cognitive patterns into lexical databases can protect linguistic meaning from becoming heterogeneous across languages.

Mukherjee et al. (2023) apply CDA to illustrate how language is inextricably tied to the relations of power, ideologies, and social strata, which are usually reflected in the functioning of the computational models as well. According to them, NLP systems are required to incorporate bias detection and correction mechanisms to minimise such biases. A conceptual contribution to morphological typology by Batsuren et al. (2021) states that computational models must account for cross-linguistic variations in word formation to explain morphologically different constructions. Thus, multilingual lexical databases would address both derivational and inflectional morphology in a more holistic linguistic representation. The same applies to Khalilia et al. (2023), who explore cultural linguistics in the context of lexical evolution, arguing that language is not merely a linguistic system but also a cultural construction. This warrants the need for context-sensitive NLP models that account for cultural nuances, such as kinship terminologies and differences in lexicon usage across languages.

This framework establishes connections between computational linguistics, cognitive science, and social theory to enhance accuracy and equity in specific terms, as well as cognitive-linguistic structures and sociocultural biases. It enables a more accurate multilingual representation of meaning in computational models and reduces linguistic and cultural biases in applications related to NLP. The framework strengthens morphological processing across languages in AI translation tools. Thus, this proposed framework makes available a rich theoretical foundation on which to develop computational linguistic models that are more inclusive and linguistically equitable.

Empirical Framework

Multilingualism, as a social phenomenon rather than a survival tool, is becoming an emerging feature in present-day society. According to Cenoz (2013), robust individual and public interest in multilingualism has spurred research in this area. The current study distinguishes between two forms of multilingualism: participatory, in which speakers engage freely in code-switching between their languages, and contextual, in which speakers use a particular language depending on the context. This differentiation sparks ongoing debates on the acquisition, processing, and social application of multiple languages, resulting in diverse research designs. The article discusses viewing things from both monolingual and multilingual perspectives on these issues, as opposed to the aforementioned rigorous view, according to which research has recently shifted into a different kind of flow that concerns language use. Singh (2010) argues that real-life multilingual challenges become linguistically real problems for researchers, rather than abstract theoretical issues. Similarly, Tabouret-Keller (1985) mentioned the shift and noted that multilingualism has eventually been accepted by scholars in their work.

Communicative competence has reinforced the argument that, during this epoch of hyper-connectivity, learning multiple languages and knowing them is more vital than ever. Arslan and Shakir (2019) examined the inclusion of social media abbreviations into communicative competence in their paper. Their conclusions recommend incorporating modern linguistic trends into language learning and testing. According to Whyte (2019), communicative competence refers to knowing the rules of a language and how to apply them in context. This line aligns with Hymes (1972), which posits that communicative competence is a complex term, as it varies according to

second language research and education, thereby affecting different modes of assessment. The study highlights the importance of LSP as a component of applied linguistics, where communicative competence is deemed relevant to both the acquisition of a language and specialised linguistic contexts.

In language documentation and dictionary compilation, lexicography constitutes the fundamental base. Bergenholtz and Gouws (2012) analyse different parameters under definitions, discussing the components that distinguish between practical and theoretical lexicography. The contribution highlights the critical role that lexicography plays in the construction of multilingual dictionaries, multiword concept databases, and electronic dictionaries. The latest developments in lexicography also span the latest digital dimension, especially in lesser-represented languages. Arslan et al. (2023) made a significant contribution to the development of the underdeveloped Punjabi language in the field of computational linguistics through the creation of digital resources. A morphological analyser, a POS tagger, and a WordNet for Punjabi were introduced in the study, making it a landmark achievement towards conservation and computational access to the language. Further building on this, Arslan et al. (2024) analysed the morphological structure of Punjabi adjectives, emphasising the case markers and their relevance in understanding morpho-syntactic constructs. Their contribution to these studies is relevant, as it paves the way for linguistic inquiry in indigenous languages with the empirical foundations required for further research into the linguistic density of Pakistan.

This paper has also extensively reviewed the existing empirical studies on lingua franca in multilingual settings. Elder and Davies (2006) explain that a lingua franca refers to a second-language medium of communication for speakers with different native languages. For example, this definition also applies to varying dialects of the same language, as shown by Ur (2013). Building awareness of such lingua francas enhances general understanding of greater global linguistic diversity and provides a basis for formulating language teaching approaches. Furthermore, a theoretical and practical approach contributes to pedagogies of language teaching. Ur (2013) asserts that task-based language teaching (TBLT) is highly effective in utilising practicality to enhance linguistic competence, rather than relying solely on theoretical approaches. While behaviourist methods rely on imitative processes and drills, traditional techniques, such as the Grammar Translation Method (GTM), emphasise grammatical structures and vocabulary items. The audio-lingual method further refines this approach with well-structured practice. On the other hand, communicative language teaching (CLT) has gained popularity, focusing on fluency and efficiency in communication, rather than inaccuracy regarding the rules of grammar (Richards & Rodgers, 2010). All the tools produced in this research project adhere to the principles of communicative competence, enabling language learners to engage in real-life interactions.

As multilingual dictionaries draw on valuable linguistic sources, they provide a comprehensive panorama of language documentation and translation. Prinsloo (2016) critically evaluates multilingual dictionaries in South Africa, assessing their merits and demerits. In the same vein, Ji, Wang, and Carlson (2016) conducted an exploratory study on dictionary sources and web access, examining the issues of information retrieval and data quality. Their work proposed an approach to mapping and merging dictionary entries, making lexical databases more accurate and user-friendly. This research is based on both Princeton WordNet and BabelNet, two widely used linguistic databases. Aker, Paramita, Pinnis, and Gaizauskas (2014) contributed to the construction of a bilingual dictionary for European languages, thereby providing structured linguistic resources. Unlike his previous works, the present study considered both national and local languages of

Pakistan, classified vocabulary items into semantic fields, and utilised images to enhance understanding.

The empirical research covered in this framework sheds light on the trajectory of changes in multilingualism, communicative competence, lexicography, and computational linguistics. The application of digital resources for local languages, particularly Punjabi, highlights the need for linguistic inclusivity in computational models. The principles of communicative and task-based approaches in language learning require reinforcement through research in language teaching methodologies, with a focus on multilingual dictionaries and lexical database development as yet another essential area for ensuring accurate representation and access to diverse languages. In summary, all these empirical studies provide a solid foundation on which future linguistic research will be built, and multilingual education will thrive alongside enhanced computational language processing.

Based on the theoretical and empirical framework, this study aims to address the research questions as outlined below.

1. How can a multilingual vocabulary database, organised by semantic categories, improve comparative linguistic analysis across global and regional languages?
2. How does the integration of underrepresented languages (e.g., Balti, Balochi, Sindhi) enhance language documentation and preservation efforts?

Methodology

This section outlines the key stages followed in the research project. First, it describes the process of corpus compilation. Second, it details the preparation and selection of the wordlist. Third, it explains the integration of languages with AI and the choice of languages. Lastly, a visual interpretation of the vocabulary database in an application is discussed.

Corpus Compilation

The initial segment of the project involved assembling an extensive corpus of the Urdu language to provide the basis for the vocabulary database. A systematic procedure was followed to incorporate as many written varieties of Urdu as possible. The corpus compiled was analysed using specific corpus analysis tools, with a distinct emphasis on frequency and semantic relevance in determining the vocabulary list.

Wordlist Preparation and Selection

A corpus analysis produced a list of the 1000 most commonly used content words. From those, the selected were the words most frequently used in everyday conversation. The words were later placed in different categories derived from semantic relations, allowing learners to be aware of and understand the connections existing between different vocabulary items.

Language Selection and AI Integration

In the next step, twenty of the most widely spoken languages worldwide were chosen, along with six original languages from the local area. These languages would add their place in the world context, both in international communication and within local cross-cultural consumption. The possibility of employing AI-based translation tools to translate the selected vocabulary items into all 24 languages is illustrated. Subsequently, further verification of the translated material was conducted through Google Search and Google Images to identify and correct mistakes and minimise glaring translation inconsistencies.

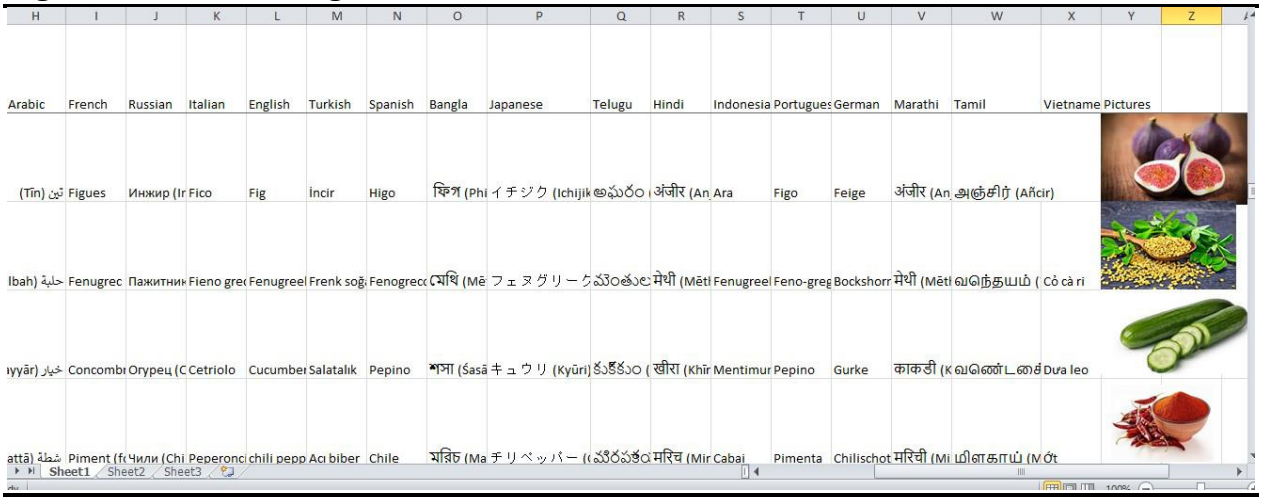
Visual Representation and Database Design

Involved in providing a new learning experience, some visual materials were embedded into the vocabulary database. For any translated word in the database, an appropriate image is made available to facilitate easy understanding or comprehension, especially for children. Additionally, a user-friendly interface was designed to enable seamless navigation between languages and vocabulary categories, thereby enhancing the experience for learners and making it even more engaging and intuitive.

Results

The word list derived from this work can assist in the acquisition of language in all 24 languages. It is also used for vocabulary-building training and serves as a handy search tool for collectively comparing and contrasting various aspects of languages, including phonetics, morphemes, translation, and calendars. By presenting examples along with the translation, the database caters to a wide range of learning modes, making it highly suitable for learners.

Figure 1: Database designed



The picture shown above is an example of the database designed in this study. Words translated into 24 languages were also accompanied by pictures that gave a better understanding of the words. The database altogether included almost 24,000 words from local and global languages of the world.

Discussion

The research introduces the multilingual vocabulary database as a new tool for linguistic analysis and culture preservation. Semantic categorization bridges linguistic diversity and prioritizes underrepresented language documentation, thus responding to the critical research questions of comparative linguistic analysis and language preservation. The discussion amplifies these themes with additional insights showing how the database promotes interdisciplinary applicability in strengthening linguistic research while stressing the urgency of protecting the endangered languages.

Semantic Categorization as a Tool for Comparative Linguistics

The current study findings are further in concordance with earlier studies wherein they placed emphasis on the fact that semantic categorization is indeed essential for comparative linguistics. The resulting structured multilingual word list from this research shows that the organization of words into thematic categories, such as agriculture, education, or health, ultimately enhances linguistic comparison across culture and geography. This follows Giunchiglia et al. (2023), who contend the databases should be organized by concepts, not words because otherwise, interlingual meaning will be destroyed.

The agricultural lexicon in our database is a reflection of culture and environment on the choice of words that has been seen in other studies on linguistic relativity. The word plough (حل in Arabic, aratro in Italian, محراث in Balti) is one such example where geography is said to influence what people say and how they say it as Parv, the one in Khalilia, et al. (2023) pointed out. Arabic expressions manifest agricultural difficulties arid lands have in facing, their European equivalences resonating to the innovations in agriculture among Europeans and Balti languages going back to reference practices of farming among the Himalayas. This trend supports Liang et al. (2024), stressed the basic human experience in producing food shapes semantic networks.

Linguistic Diversity in Educational Terminology

The results of our study concerning educational terminology corroborate the findings of Mukherjee et al. (2023) on the socio-cultural aspects of language. The concept of a teacher being lexically different in several languages, for instance, in Arabic أستاذ, Japanese 教师, or Spanish maestro, adheres to their argument that language exposes underlying power contexts and ideological structures. Arabic أستاذ communicates a sense of hierarchical respect in formal and honorific terms, while Japanese 教师 points towards structures of authority around educational contexts. Spanish maestro, interwoven with notions of mentorship and nurturing, demonstrates how linguistic choices carry cultural attitudes toward education. In this regard, our investigation deepens the scope of these studies by introducing AI-assisted translation and visual analytical tools for comparative cross-linguistics. This thrust by technology is parallel to the contribution of Bella et al. (2022), who employed computational methods in multilingual lexical databases to promote interlingual meaning transference.

Health Terminology and Lexical Evolution

Some forms of consultancy used in medicine for treatment are proven from the research that it confirms an earlier developmental study on linguistic adaptation in medicine. Words for medicine as in Urdu دوائی, French médicament, Japanese 薬 reveal dissimilar cultural effects as pointed out by Batsuren et al. (2021) in their discussion of morphological diversity. Bonds of traditional therapeutic practice define Urdu and Japanese terms, while French align with pharmaceutical development. The pattern reflects Liang et al. (2024), whoadvance that cognition structure shared is biased to how languages absorb scientific and technological developments.

The multilingual database developed here substantiates earlier studies by demonstrating that semantic categorization enables comparison between languages while highlighting socio-cultural determinants and providing a basis for language evolution. Further, our study combines artificial intelligence-amp translation with their respective visual representation to pt it into heretofore theoretical dimensions but also practically puts it at the disposal of linguistic studies and education.

Inclusion of Underrepresented Languages in Language Preservation

This study is consistent with earlier works emphasizing the significance attached to documenting underrepresented languages in fostering the preservation of linguistic diversity. The multilingual database containing languages such as Balti, Balochi, and Sindhi serves as an expression of the awareness on the part of authors toward endangered languages in doing their bit for linguistic preservation, as noted by Giunchiglia et al. (2023). They further claim that computational lexical databases often neglect non-dominant languages, creating gaps in linguistic representation. The incorporation of these languages into our database will help in bridging the gap and ensuring the preservation of local linguistic expressions.

Culturally relevant terms, such as المحراث (mihraath)-plough in Balti and نمك (namak)-salt, included in this study echo the viewpoint of Khalilia et al. (2023), who affirm that kinship terminologies and other culturally significant words ought to be preserved in order to protect linguistic and cultural identities. The documentation of Balti terms related to Himalayan agricultural heritage equally reflects the importance voiced by Liang et al. (2024), wherein they maintain that language reflects the cognitive commonalities essentially derived from environmental and social conditions.

Semantic Adaptations and Cultural Variations

It has also been shown in our study that the socio-culture biases and adaptations shape linguistic structures, which is what Mukherjee et al. (2023) have shown. The different meanings carried by breakfast as represented in Urdu ناشتہ, Arabic فطور (fuṭūr), and Japanese 朝食 are reflective of cultural routines and dietary traditions. Just as Mukherjee et al. mention the ideological effects in the use of language, our results show that mealtime lexicon embodies cultural attitudes toward food and communal practices. Whereas Urdu ناشتہ emphasizes the bonding between family members during meals, Japanese 朝食 underscores a more regimented and balanced approach to nutrition, thus reiterating previous studies on culture-bound linguistic phenomena.

By contrast, with respect to word happiness (Urdu سکھ, Hindi खुशी, French bonheur), the findings of the study have parallels with cognitive-linguistic research. According to Liang et al. (2024), linguistic choices encode psychological and social values, lightly corroborated by our finding. While Urdu سکھ emphasizes collective well-being in observance of South Asian communal traditions, French bonheur stresses the happiness of the individual, which serves to illuminate Western philosophical views. These observations further substantiate the arguments about language as an embodiment of cultural concerns and social values.

Implications for Linguistic Diversity and Preservation

Adding underrepresented languages to our database serves to validate past work and extend findings through a formal approach to documenting computational linguistics. This was the manifesto espoused by Bella et al. (2022), which we put into practice through AI-assisted translation and visualization in our work. In ensuring due representation for underrepresented languages, our database seeks to promote both linguistic equity as well as sustaining efforts toward preserving endangered languages, thereby reinforcing Batsuren et al.'s (2021) argument on morphological diversity in NLP models.

That said, this study supports and expands on previous works by affirming that linguistic preservation embodies the maintenance of cultural heritage. The modeling of semantic shifts, cultural importance, and linguistic rules in endangered languages will substantiate computational models of language designed to embrace linguistic diversity for the future.

Cross-Disciplinary Applications and Technological Integration

The outcomes of the described research yield the same results found in some previous research dealing mostly with the role of multilingual lexical databases in computational linguistics and education. The semantic categorization of vocabularies from 24 languages under study further backs the argument advanced by Noor, Arslan, and Noor (2024) that biases are common in NLP models due to their over-reliance on certain major world languages such as English, Mandarin, and Spanish. The database also goes further to minimize such biases through the addition of endangered and less represented languages, making tools for language processing by AI more general.

Besides this, our multilingual template of words, such as petrol (پٹرول in Urdu, بنزين in Arabic, ガソリン in Japanese), aligns with banknote (نقدية ورقة in Arabic, банкнота in Russian) and previous work by Batsuren et al. (2021), which call for morphological diversity in language processing models. Their research indicates that languages are hardly ever considered in language technologies and that languagetechnologies are not usually effective working on morphologically rich languages, which our database has been built to address by structured translations across the linguistic families involved.

Applications in Education and Linguistic Diversity

This study also reiterates the argument of Mukherjee et al. (2023) that comparative linguistic studies advance language learning by presenting the variety in linguistic structures with their cultural impact. Our database classifies words according to the shared human experiences such as meals (breakfast: Urdu ناشتہ, Arabic فطور, Japanese 朝食) and emotions (happiness: Urdu سکھ, Hindi खुशी, French bonheur), and serves an educational purpose that opens the eyes of students to the world. Giunchiglia et al. (2023) have endorsed this kind of structure in linguistic comparison, as they would argue, to promote cross-cultural understanding in language education, emphasizing the importance of semantic organization.

In addition, the functionality of the database for curriculum development corresponds with what Bella et al. (2022) observe: multilingual resources provide fertile ground for cognitive flexibility in learners. It increases an appreciation for linguistic diversity because it demonstrates how different cultures recognize and engage in common experiences.

Interdisciplinary Applications in Cultural Studies and Historical Research

Our research not only involves computational linguistics and education but also corroborates previous studies on the cross-disciplinary focus of language, anthropology, and studies of environment. According to Liang et al. (2024), linguistic structures reflect societal values and historical contexts, and we have verified this claim through an analysis of the etymology of words such as ash (راکھ in Urdu, 灰 in Japanese, cenere in Italian). Words signify more than the physical remnants; they are endowed with symbols within religious rituals and whispered narratives on the environment or history.

These findings are together with the works of Khalilia et al. (2023), who analyze language as the bank of cultural knowledge. Our data base organizes the meanings of vocabularies derived mainly from nature, economy, and daily life, such that it allows interdisciplinary studies linking linguistic with anthropology, ecology, and theology. In this way, it builds on previous research by making a portioned interface to investigate the coding of a language concerning historical and environmental realities.

The Role of the Database in Safeguarding Linguistic Heritage

The study results further underscore the importance of recording these endangered languages in the context of global linguistic preservation initiatives or UNESCO programs, to help secure cultural heritage. In support of this is the inclusion of languages like Balti and Sindhi which are usually quite under-documented in our database, thus extending Austin and Sallabank's (2014) perspective that linguistic diversity is not only about documentation-it encapsulates cultural-historical knowledge regarded in that lexicon. Examples from the database would include words like محراث (ḥmihraath for plough) and آسمانی (āsmānī for sky blue) which build up Harrison (2007)'s argument that indigenous vocabularies have strong epistemic thrusts potentially pertinent to studies such as ethnobotany and traditional medicine.

The other connection offered in the study parallels Evans' (2010) work, where he elaborates on the important role that computer tools play in revitalizing the language of endangered status. The multilingual vocabulary presented in an easily accessible manner renders the database an effective modern tool for preservation that inhibits the loss of valuable linguistic knowledge, thereby ensuring that it remains accessible for future generations. This digital view has been further affirmed by new studies, including Liu et al. (2022), who argue that AI-based documentation will enhance the sustainability of linguistic heritage.

Adaptation of Languages to Globalization and Modernity

These results add another contribution to already existing conversations on how languages adapt to address and represent modern realities. For example, scholars such as Mufwene (2008) have argued that globalization does not necessarily involve the extinction of languages; instead, it enhances adaptation to linguistic identities. Examples of language adaptation include the words coat (کوٹ in Urdu, manteau in French, пальто in Russian) and pants, which show absorption of foreign terms but retain culture. This representation is reflected in the expressions such as شلوار in Urdu, pantalones in Spanish, and ズボン in Japanese.

Our study also reaffirms the findings of Fishman (1991), who argues that language maintenance should consider both traditional and modern language needs. The database thus connects past knowledge to modern language practice by incorporating terms that signify heritage (for example, محراث for plough) to those signifying life today (for example, coat, pants). This is consistent with the view presented by Pauwels (2016), who reveals that multilingual dictionaries and lexical databases are balancing places under which traditional preservations interact with realities of a globalizing world.

Extending the Scope: Future Directions

Despite being a good start, there remains an immense room for possible growth in the database. The expansion towards semantic categories embracing technology, environmental issues, and even emotions would definitely augment the database. For instance, incorporating computing or internet terms into digital technology keywords could help provide insight into how the local languages are molding their expressions to keep pace with global progress. Adding audio and visual representations combined with contextualized use would maximize the capacity, accessibility, interactivity, and user-friendliness of such digital databases.

The documents will now be enriched through their collaboration with indigenous communities to build an even stronger database that would actually document linguistic usage and thus cultural worth. Indigenous knowledge would record fine pronunciation, idioms, or context meanings making everything richer into this discourse on language resource.

This multilingual vocabulary database would be at the site of the intersection between linguistic research and cultural preservation. The categories systematic arrangement of words into semantic categories allows under category oriented comparisons between languages that might eventually lead to identifying patterns of linguistic adaptation and cultural expression (Cenoz, 2013). The minority languages such as Balti and Sindhi included here indicate the seriousness with which the task of safeguarding endangered linguistic heritage has been taken; this secures their voices within the discourses of the global paradigm. Thus, this database plays a role in connecting cultural with linguistic diversity and understanding it; this makes it all the more valuable to learners and teachers alike in the common pursuit of preserving yet distinct linguistic identity of humankind. Soon, with expected expansion and collaboration with many other disciplines, the tool has an opportunity to become a strong factor of worldwide linkage and celebration of linguistic richness.

The plural vocabulary database is also a life bridge between global and indigenous regional language sets: this point is one more input for these endangered language researchers and preservationists to consider. Its systematic arrangement helps cross-cultural understanding and pops on wider ideals, such as the promotion of linguistic diversity and safeguarding of endangered languages. Such expansion toward real fields, like on indigenous peoples, is even more promising than this database into an essential reliquary for all-the researchers, the teachers, and the technologists alike.

Practical Applications of the Multilingual Vocabulary Database

The database has various practical applications in different fields. In the context of blended mobility involving international education between physical and online courses, the database would serve as a useful language tool. Language teachers could exploit the lexical resources in class, while students could benefit from it as an additional tool for foreign language learning. Likewise, tourists would use its multilingual features to ease their navigation amidst new language surroundings.

One major contribution of this research includes enhancement of the database to cover Pakistani local languages, thereby buttressing the global significance of the idea of promoting indigenous languages into the education system. This inclusion would not only facilitate language learning, but it would also help preserve these lesser-known languages, which in turn would ensure their very existence in a rapidly globalizing world.

Linguistic Analysis and Comparative Studies

The really striking use of this database might be for linguistic analysis, especially in regard to phonological and morphological comparisons between languages. One can see how particular phones that exist in one language but may not be in another show linguistic contrast. The change in just one phone can cause a drastic difference, thus pointing to very fundamental differences in the structuring of two languages.

The database also helps with tracing lexical borrowing among languages. Similar vocabulary items in a number of languages provide insight into language contact and borrowing cases. For example, the word yogurt is found in both Turkish and English, denoting that some culinary terms cross the boundary between languages. Equally, the French word for suit, costume, signals how languages influence one another through exchange.

Cultural Insights and Translation Challenges

The database shows the lexical gaps across languages, which indeed aids in doing a cross-cultural analysis-for instance: there is a lexical item in one language, while there is none in the other-the differences portray the culture that gives rise to such sorts of translation problems. Some topics may arise from one such culture and have no counterparts in another culture; such types of problems tend to bring about barriers for the communication of one language to another.

Again, pictorial references were included to facilitate the understanding, although this section of research had also offered some challenges and some opportunities. It is difficult to illustrate vague terms like question, answer, sympathy, and hatred because their meanings are dependent on the context and not on forms or objects as bases. Nonetheless, the picture adds yet another dimension to the understanding, making the database a more fruitful exercise for both learner and researcher.

Conclusion

This study demonstrates that a multilingual vocabulary database, categorized according to semantic fields, functions as a formidable instrument for comparative linguistic analysis and preservation of marginalized languages. The database organizes its words according to such thematic fields as agriculture, education, health, and emotion, thus making possible specific cross-linguistic comparisons. This evidences patterns of linguistic accommodation and cultural manifestations and demonstrates how common human experiences-such as teaching (teacher) or agricultural activities (plough) - are encoded differently in languages. Findings enrich not only linguistic research but also provide aids in comprehending the intricate web linking language, culture, and societal values.

One important input from this study is the inclusion of languages such as Balti, Balochi, and Sindhi, which are again uncommonly included, justifying the contribution of this database in language documentation and preservation. Often, referred to as truly minority languages, these often speak of the remaining members of communities as being rich in extra-cultural commodities with varied traditions and worldviews. For instance, terms likeمحراث (mihraath, plough) andآسمانی (āsmānī, sky blue) are more than the conservation of vocabulary; they are the saving of cultural properties, oral history, and indigenous knowledge systems that may otherwise be lost. This is in line with the global effort to preserve languages in that these languages will not only be relevant at this time of rapid globalization but will also be accessible.

Semantic structuring makes a point for the angle to the very opening question research has raised. It is a tool of comparative linguistics that offers a more systematic way of exploring the similarities and differences in languages. It provides thus a systematic pattern applicable to more than one subject such as linguistics, pedagogy, and natural language processing technologies so that it can serve other interdisciplinary applications as well. Regarding the second research question, the inclusion of endangered languages in the database symbolizes a tangible commitment to linguistic equality and cultural preservation. It thus encourages linguistic diversity and greatly contributes to the preservation of culture, especially for future generations, by documenting and analyzing concepts from these kinds of marginalized languages.

Significantly, the database is also valuable beyond academia, useful in practice. It attracts teachers and students, even tourists, by offering an interactive platform for language acquisition and cross-cultural communications. Phonological and morphological studies, lexical borrowings, and problems of cultural translation add to its chances of a research tool. The inclusion of pictorial references, though difficult as words such as sympathy or hatred have highly abstract definitions, enhances the learning experience.

Essentially multilingual vocabulary, this database binds the countries of linguistic diversity and cultural understanding, thus forming part of the global effort to conserve what is still left of humankind's language heritage. Depending upon increased audio-visual expansion, the ever-growing technology and partnership with indigenous communities, the database may yet be developed into an even stronger tool to forge global links and celebrate linguistic richness.

References

- Aker, A., Paramita, M. L., Pinnis, M., & Gaizauskas, R. (2014, May). *Bilingual dictionaries for all EU languages. In LREC 2014 Proceedings (pp. 2839-2845).* European Language Resources Association.
- Arslan, M. F., Mahmood, M. A., Akram, A., & Tahir, M. S. (2023). Developing Digital Resources of Shahmukhi Punjabi. *Russian Law Journal, 11*(2), 998-1006.
- Arslan, M. F., Kanwal, A., Mehmood, M. A., & Haroon, H. (2023). Morphemic Analysis of Case Markers in Shahmukhi Punjabi Nouns. *Journal of Namibian Studies: History Politics Culture,* 33, 6501-6517.
- Arslan, M. F., Mehmood, M. A., & Kanwal, A. (2024). Punjabi Adjectives: A Morphological Perspective. *Pakistan Languages and Humanities Review, 8*(1), 543-552.
- Arslan, M. F., Mahmood, M. A., Shoaib, M., Idrees, S., & Tariq, Z. (2023). Morphological Description of Nouns in Shahmukhi Punjabi; A Corpus Based Study. *Journal of Positive School Psychology,* 1259-1269.
- Arslan, M. F., & Shakir, A. (2019, December). Inclusion of social media abbreviations incommunicative language testing. *Linguistic Forum - A Journal of Linguistics, 1*(2), pp. 33-41).
- Batsuren, K., Bella, G., & Giunchiglia, F. (2021). *MorphyNet: A large multilingual database of derivational and inflectional morphology. Proceedings of the Seventeenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 39–48. DOI: https://doi.org/10.18653/v1/2021.sigmorphon-1.5
- Bella, G., Batsuren, K., & Giunchiglia, F. (2021). *A database and visualization of the similarity of contemporary lexicons. Workshop on Time-Delay Systems.* DOI: https://doi.org/10.1007/978-3-030-83527-9_8
- Bella, G., Byambadorj, E., Chandrashekar, Y., Batsuren, K., Cheema, D. A., & Giunchiglia, F. (2022). *Language diversity: Visible to humans, exploitable by machines. Annual Meeting of the Association for Computational Linguistics.* DOI: https://doi.org/10.48550/arXiv.2203.04723
- Bergenholtz, H., & Gouws, R. H. (2012). What is lexicography?. *Lexikos,* 22, 31-42.
- Cenoz, J. (2013). Defining multilingualism. *Annual Review of Applied Linguistics,* 33, 3-18.
- Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics,* 24, 282-304.
- Giunchiglia, F., Bella, G., Nair, N. C., Chi, Y., & Xu, H. (2023). Representing interlingual meaning in lexical databases. *Artificial Intelligence Review*, 56, 11053–11069. DOI: https://doi.org/10.1007/s10462-023-10427-1
- González, S. P. (2019). Growing in another culture.
- Ji, K., Wang, S., & Carlson, L. (2016, October). Multilingual Dictionary Linking and Aggregation: Quality from Consistency. In *LD4IE@ ISWC* (pp. 51-62).

- Khalilia, H., Bella, G., Freihat, A. A., Darma, S., & Giunchiglia, F. (2023). Lexical diversity in kinship across languages and dialects. *Frontiers in Psychology*. DOI: https://doi.org/10.3389/fpsyg.2023.1229697
- Liang, Y., Xu, K., & Ran, Q. (2024). Shared structure of fundamental human experience revealed by polysemy network of basic vocabularies across languages. *Scientific Reports*, 14, 5877. DOI: https://doi.org/10.1038/s41598-024-56571-8
- Mukherjee, A., Raj, C., Zhu, Z., & Anastasopoulos, A. (2023). *Global voices, local biases: Socio-cultural prejudices across languages. Conference on Empirical Methods in Natural Language Processing.* DOI: https://doi.org/10.48550/arXiv.2310.17586
- Nair, N. C., Velayuthan, R. S., Chandrashekar, Y., Bella, G., & Giunchiglia, F. (2022). *IndoUKC: A concept-centered Indian multilingual lexical resource. International Conference on Language Resources and Evaluation.*
- Noor, N., Arslan, M. F., & Noor, H. (2024). Morphological Representation of Gender in Arabic, English, German, And Punjabi: A Comparative Linguistic Analysis. *Journal of Applied Linguistics and TESOL, 7*(4), 628-645.
- Prinsloo, D. J. (2016). Critical analysis of multilingual dictionaries. *Lexikos, 24*(1), 220-240.
- Richards, J. C., & Rodgers, T. S. (2014). Approaches and methods in language teaching. Cambridge University Press.
- Singh, R. (2010). Multilingualism, sociolinguistics and theories of linguistic form: Some unfinished reflections. *Language Sciences, 32*(6), 624-637.
- Whyte, S. (2019). Revisiting communicative competence in the teaching and assessment of language for specific purposes. *Language Education & Assessment, 2*(1), 1-19.
- Ur, P. (2013). Language-teaching method revisited. *ELT journal, 67*(4), 468-474.