



Department of Education, University of Oxford

Lafzon ka Guldasta: An Exploratory Study of Urdu Vocabulary Development in
Grades 1-3

Rida Zafar

St. Anne's College

Trinity Term, 2020

Word Count: 19,940 words

Dissertation submitted in part-fulfilment of the requirements for the degree of MSc in Education (Child
Development and Education)

Acknowledgements

Firstly, I would like to extend my gratitude to my supervisor, Sonali Nag. Without her support and guidance over the last year, I would not have been able to complete this dissertation. A continuously evolving global pandemic is a strange backdrop to conduct research in. Sonali steered me in the right direction, caught my anxious thoughts before I spiralled in despair every time we had to reorient ourselves and move from Plan A to Plan B to Plan C, challenged me to engage more critically and to think more deeply, respected my ideas, and helped me feel immense agency and ownership of my research and the dissertation I produced.

This dissertation would not have been possible without the kindness, love, and unconditional (and often unjustified) belief of my family. Unexpectedly moving back home in March when COVID-19 hit meant that my parents witnessed the entire progression of my work. They took care of me when I felt too stressed to eat or sleep, and when I was irritable and not the most present daughter. They created space for me to be frustrated, but also celebrated every small breakthrough and tiny victory with me. Although my brother lived in a different country, his patience with me and words of encouragement meant the world. For this, and for an entire lifetime of knowing my family trust me and value my education, I will always be grateful.

I would also like to thank everyone whose participation in varying degrees allowed me to conduct my research: all the participants who labored through Zoom calls with unreliable WiFi connections, patiently answered my questions, and very graciously agreed to participate in my project as they learned to navigate their own lives as teachers and parents in an age of remote learning and teaching, and my independent coders who helped me tag and prepare my word lists.

Everyone on the CDE pathway - tutors, TAs, and my compassionate and downright brilliant cohort - also hold a special place in my heart for everything they have taught me. Having the opportunity to study this course was wonderful in and of itself, but learning from my peers and their diverse experiences made the journey even more extraordinary.

Finally, I'd like to thank my friends whose support and unwavering belief in me was all I needed from the time I couldn't think of an idea I wanted to work on to the day before I submitted when I was working without electricity or WiFi and wasn't sure I would be able to submit. So to my childhood friends, Alizeh, Lalarukh, Nayab, and RAP, and my friends from university, Ghaniba, Farzan, Mubasher, and Simran, thank you for choosing to be my friend despite it all, and for the support and love you held for me this last year.

Abstract

Urdu is a member of the Indo-European language family. When children understand words and how schools promote children's vocabulary development has important educational implications both in terms of vocabulary growth, and the inter-connectedness of vocabulary and literacy domains like reading comprehension. This study aimed to explore Urdu vocabulary development by assessing the age of acquisition ratings of words derived from a child-directed print corpus of first to third-grade textbooks. A related aim was investigating the associations between the rated age of acquisition and other psycholinguistic variables, such as frequency, number of letters and syllables, and orthographic transparency. A secondary aim was to analyze activities included in the same first to third-grade Urdu textbooks for their promotion of morphological and orthographic awareness.

A word corpus of 1,968 words was created by digitizing connected text (i.e. poems, stories, etc.) found in nine children's textbooks. The corpus was annotated for word classes and orthographic transparency. After inter-rater reliability was assessed for these annotations, the frequencies of six word classes (nouns, verbs, adjectives, adverbs, conjunctions, and postpositions) in the word corpus were calculated. A word list of 600 words representing the same proportions of word classes as the word corpus was extracted for an online AoA study. Ten teachers responded to three questionnaires on Qualtrics containing 200 words each. The first and third-grade textbooks were analyzed using Vagh & Nag's (2014) template for South Asian languages after appropriate amendments were made to reflect Urdu's linguistic properties.

Results reveal that, on average, teachers think children acquire words in the word list between the ages of three and five years. Consensus analysis revealed strong inter-rater reliability for all word classes except conjunctions. According to teachers, conjunctions and nouns are acquired later than other word classes, high frequency content words were acquired earlier than low frequency content words, and vowelized nouns are acquired earlier than semi-vowelized or un-vowelized nouns. The textbook analysis revealed main effects of component skills in the morphological and orthographic domains with a higher proportion of activities coded for vocabulary and letter knowledge than other component skills in the domain. Qualitative analysis suggests a restricted range of activities in both grade levels although the types of activities differ by publisher. Instructional implications for vocabulary development and other component skills are discussed and the need for future experimental research is emphasized.

Table of Contents

Acknowledgments	ii.
Abstract	iii.
List of Tables and Figures	vi.
Chapter 1: Introduction	1
1.1. Morphology and Vocabulary Development	1
1.2. Urdu Morphology	2
1.3. Morphology and Orthography	3
1.4. Morphology and Word Classes	4
1.5. Rationale for the Study	5
1.6. Research Questions	6
Chapter 2: Literature Review	7
2.1. Chapter Overview	7
2.2. Age of Acquisition (AoA)	7
2.3. Corpus Linguistics	10
2.4. Textbook Analysis	14
2.5. Theoretical Framework	18
Chapter 3: Methodology	19
3.1. Chapter Overview	19
3.2. Aims	19
3.3. Research Questions	19
3.4. Research Design for the AoA Study	19
3.5. Developing a Child-Directed Print Corpus	20
3.6. Recruitment	24
3.7. Participants	25
3.8. Tools and Measures	25
3.9. Procedure	27
3.10. Ethical Considerations	29
Chapter 4: Results	31
4.1. AoA Study	31
4.2. Textbook Analysis	36
Chapter 5: Discussion	46
5.1. Chapter Overview	46
5.2. AoA Study	46
5.3. Textbook Analysis	48
5.4. Implications for Education	51
5.5. Limitations	54
5.6. Directions for Future Research	56
5.7. Conclusions	57
References	58

Appendix A: Urdu Phonology and Orthography	68
Appendix B: Adaptations to the Library of Congress' Urdu Romanization Table	72
Appendix C: Kolmogorov-Smirnov Tests for Normality and Descriptive Statistics for Frequency of Word Classes	73
Appendix D: Text for Facebook Recruitment Post	74
Appendix E: Project Information Email for Teachers	75
Appendix F: Project Information Email for Parents	77
Appendix G: Screenshots of a Sample of the AoA Questionnaire	79
Appendix H: Adapted Vagh & Nag (2014) Textbook Analysis Template	81
Appendix I: Sample Script for Zoom AoA Sessions	103
Appendix J: CUREC Approval Notification	104
Appendix K: Z-Scores of Variables in the AoA Study	105
Appendix L: Results from the Shapiro-Wilk Test of Normality of Domains in the Textbook Analysis	106
Appendix M: Levene's test of Homogeneity of Variance for Domains in the Textbook Analysis	107
Appendix N: Complete Word List from the AoA Study	108
Appendix O: Sample Spelling and Reading Word List	131

List of Tables and Figures

Table 1.1 Five Nominal Markers in Urdu with Transcription and Transliteration	2
Table 3.1 Inter-Rater Agreement for Word Class Annotation	22
Table 3.2 Frequencies by Word Classes in Word and Sentence Corpus	23
Table 3.3 Transparency Category Distribution of AoA Word List	24
Table 3.4 Numbers and Percentages of Chapters and Activities in Each Textbook	29
Table 4.1 Means, Standard Deviations, and ICC Indices of Complete Word List and Word Classes	32
Table 4.2 Three Sources of Discrepancy in the AoA Ratings for Identified Words with an Example and Explanation	33
Table 4.3 Spearman’s Rank-Order Correlations for AoA Ratings and Psycholinguistic Variables	35
Table 4.4 Percentages and Number of Activities in the Language Domain by Grade and Publisher	37
Table 4.5 Percentages and Number of Activities in the Literacy Domain by Grade and Publisher	38
Figure 4.1 Estimated Marginal Means Difference of Proportion of Activities between Skills in the Morphological and Orthographic Domains by Grade	41
Table 4.6 Means, Standard Deviations, Main and Interaction Effects of Activity Proportions by Component Skills and Grade in the Morphological Domain	41
Table 4.7 Means, Standard Deviations, Main and Interaction Effects of Activity Proportions by Component Skills and Grade in the Orthographic Domain	42
Table 4.8 Means, Standard Deviations, and Effects of Domain by Publisher	43

Chapter 1: Introduction

1.1 Morphology and Vocabulary Development

Morphological awareness, or the knowledge of the structure of different parts of a word such as prefixes, stems, and suffixes, has been linked to vocabulary development. Inflectional morphology (e.g. the use of suffixes to mark tense) is believed to be acquired earlier than derivational morphology which involves structural changes to the root morpheme (Tyler & Nagy, 1989). Knowledge of an unfamiliar word's structure provides contextual cues that help determine its meaning and can therefore build vocabulary (Nagy & Anderson, 1984). For instance, knowledge of the way certain suffixes work in changing the tense or part of speech can help children infer the meaning of a similarly structured new word.

Morphological awareness of derived words has also been associated with vocabulary growth (Freyd & Baron, 1982) and reading comprehension (Deacon & Kirby, 2004). Anglin (1993) found that fifth-grade English speaking children analyzed the derivational structure of complex words to decode them. Knowledge of multi-morphemic words also grew between first and fifth grades. Similarly, in a study with second-grade children, Vaknin-Nusbaum and Saeigh-Haddad (2020) found a predictive effect of awareness of Arabic derivational morphology on reading comprehension. Tibi and Kirby (2017) also found a predictive effect of Arabic morphological awareness on reading comprehension in third-grade students, and call for the development of morphological interventions for vocabulary and reading comprehension.

Morphological awareness' contribution to word-learning, vocabulary growth, and reading comprehension has led to the development of specific instructional practices that target morphology, particularly in English. Prince (2009) found that teaching morphology as a cognitive strategy to analyze unfamiliar word produced higher scores on a reading comprehension post-test for English L2 learners in secondary school. Results from a meta-analysis with studies from the United States, Denmark, New Zealand, and Canada found significant effects of morphological instruction on literacy domains including vocabulary and reading comprehension (Goodwin & Ahn, 2010).

The next section provides an overview of Urdu morphology before returning to a more specific commentary on the relationship between morphology and orthography, and morphology and word classes.

1.2. Urdu Morphology

Sharing many of its morphological features with other Indo-European languages such as Hindi, Urdu has a predominantly concatenative inflectional morphology. Since it utilizes both linear (i.e. inflectional) and non-linear (i.e. derivational) morphology, it also shares its morphological properties with Arabic, which has a non-concatenative morphology (Daud et al., 2017). In Urdu, morphological changes to root words occur both through the use of affixes (that do not change the root) and infixes (that alter the root word and do not require consonants in the root to be sequentially strung together) (Fatima et al., 2018). Inflectional morphology is most apparent in Urdu in its use of suffixes. Table 1.1 provides examples of Urdu's five nominal markers along with their IPA transcriptions, and translation to demonstrate the variety of suffixes used in Urdu.

Table 1.1.

Suffix Type	Urdu <i>Nastaliq</i> with IPA Transcription	Translation
ی (/i:/)	سردی ، گرمی /sa:r̥d̪i:/, /ga:r̥mi:/	سردی ، گرمی <i>summer , winter</i>
گاہ (/ga:h/)	تجربہ گاہ ، عید گاہ /t̪a:d̪ʒa:r̥ba:hga:h/, /ʔi:d̪ga:h/	تجربہ گاہ ، عید گاہ <i>Eid-site , laboratory</i>
ہٹ (/ha:t/)	گھبراہٹ ، کڑواہٹ /g̱ha:bra:ha:t/, /ka:r̥va:ha:t/	گھبراہٹ ، کڑواہٹ <i>bitterness, restlessness</i>
یت (/i:t̪/)	انسانیت ، حیوانیت /insa:ni:t̪/, /he:va:ni:t̪/	انسانیت ، حیوانیت <i>inhumanity, humanity</i>
ای (/a:i:/)	برائی ، لمبائی /bra:i:/, /la:m̪ba:i:/	برائی ، لمبائی <i>badness, tallness</i>

Five Nominal Markers in Urdu with Transcription and Transliteration

Derivational morphology also plays a central role in Urdu linguistics because gender, word class, and grammatical number inhere the changes made to root words. However, it is important to note that such changes can be either inflectional or derivational and the exact alteration depend on individual words. There is some agreement in the literature on Urdu linguistics that words with infixes tend to follow specific patterns in terms of where certain vowels and consonants (particularly /t̪/ ت and /m/ م) are inserted in the order of graphemes in the root word (Abdul Jabbar & Khan, 2016; Fatima et al., 2018).

Some of Urdu's morphological properties can pose challenge to learners. One such example is that of ambiguous affixes. Some words contain units that would otherwise be classified as inflectional morphemes (prefixes or suffixes) or separate words, but are actually part of the stem itself. Separating or removing such units would render the remaining stem meaningless (Fatima et al., 2018). Another consideration that must be made in Urdu is the dual presence of altered and unaltered plural forms (Fatima et al., 2018). Altered plural forms (*jama mukassar*) occur when the consonantal content of stems has to be non-sequentially altered (through the insertion or deletion of vowels or other consonants) to transform singular nouns to their plural forms. Unaltered transformations (*jama salim*) occur when adding affixes changes the grammatical number of the noun. Urdu also contains broken plurals whereby a single stem can be transformed to its plural form in several ways (which in turn can contain a combination of altered and unaltered forms). With broken plurals, correct usage depends on context. Derivational morphology, ambiguous affixes, and multiple plural forms accentuate Urdu's morphological complexity, perhaps drawing it closer to Arabic. Appendix A provides an overview of Urdu phonology and orthography, particularly considering the influence of other languages like Arabic.

1.3. Morphology and Orthography

The role of morphological awareness in vocabulary acquisition in *abjad* (or consonantal) languages like Arabic is more relevant to the present study as Arabic shares many morphological and orthographic properties with Urdu. Studies on the acquisition of Arabic morphology have focused on the link between morphology and orthography (e.g., Taha & Saeigh-Haddad, 2017). From priming studies, Boudelaa and Marlsen-Wilson (2005; 2013) deduce that root morphemes may be stored in the mental lexicon. Since root morphemes are made up of consonantal units, the link between morphology and orthography becomes salient.

Unlike studies in English in which derivational morphology awareness occurs in secondary school, research from Arabic suggests an earlier emergence of derivational morphology awareness in primary school (Shalhoub-Awwad & Leikin, 2016). Taha and Saeigh-Haddad (2017) comment on the derivational structure of Semitic languages whereby the consonantal unit containing the root along with the word pattern (prefixes, suffixes, and affixes) provide semantic cues to the listener and reader. The early emergence of derivational morphological awareness has been linked to Arabic's orthographic properties to develop interventions for word decoding that

include teaching the morphological structure of words (Saiegh-Haddad, 2018). These morpho-orthographic interventions are based on the complementation of morphological and orthographic properties of abjad languages that can help readers both decode and understand a word's meaning.

Even for languages like English with distinct derivational properties, the role of vocabulary in word decoding is documented (Scarborough, 2001). In a study with fourth-grade students, Oullette (2006) found that oral vocabulary breadth (i.e. how many words are known) predicted word decoding whereas vocabulary depth (i.e. how well word meanings are known) predicted reading comprehension. In morphologically rich languages like Arabic (and Urdu) in which the meaning of derived words is related to their orthographic representation (Saiegh-Haddad, 2018), vocabulary and word decoding may be even more interconnected. This relationship between morphology and orthography can also explain why many Arabic morphological interventions target literacy activities like spelling as early as second grade (Taha & Saiegh-Haddad, 2017).

1.4. Morphology and Word Classes

While it is argued that adult word-classes are irrelevant to children acquiring language, high frequency word classes, like nouns, are acquired earlier than low frequency word classes (Tomaseollo, 2003). The specific morphological properties of a language can influence the vocabulary acquisition between and within different word classes. For instance, Waxman and Lidz (2006) suggest an almost universal dependence of adjectives on nouns, which complements Gentner's (1982) finding of a noun advantage in language acquisition. Waxman and Lidz' (2006) finding of a dependence of adjectives on nouns may be particularly evident in languages (like Urdu) that mark for grammatical number and gender as adjectives must agree with the nouns they modify. Saiegh-Haddad et al. (2012) found differences in the acquisition and developmental trajectory of plural forms based on differences in Arabic noun pluralization (through linear affixes or nonlinear infixes). Derivational morphology can also change word classes. For example, in Urdu, the consonantal root morpheme کتب (/k/, /t/, /b/) are found in both the noun کتاب (/kʰa:b/, book) and the semantically related verb کتب بینی (/kʰb bi:ni:/, reading). Therefore, particularly for derivationally rich languages in which morphology and orthography work together to support accurate reading and inferences of word-meanings, differences between and within the acquisition of word classes may have important implications for vocabulary development.

1.5. Rationale for the Study

Cross-linguistic empirical research from languages such as English and Arabic highlights the contributions of morphological awareness in vocabulary development. As the examples cited above portray, explicit morphological interventions have been developed to enhance word-learning. Vocabulary development, including when children learn certain words is relevant to education because aspects of vocabulary (such as breadth and depth) are related to other literacy skills like word decoding and reading comprehension. Therefore, when words are learned, and strategies for learning new words that are taught in schools carry import beyond the domain of vocabulary.

At present, Urdu is an under-studied language, as research on morphology and word classes is restricted to computational linguistics with adult-directed speech (e.g., Hussain, 2004; Hardie 2003). However, the link between written input and vocabulary in languages rich in derivational morphology such as Arabic (Saeigh-Haddad, 2018) indicate the need to extend such research to orthographically and morphologically similar languages like Urdu.

Child language acquisition and vocabulary development have not been emphasized in Urdu research. Considering this lack of emphasis, the initial plan for this dissertation was an intervention designed for third and fourth-grade classrooms to investigate effects of explicit morphology instruction on vocabulary and reading comprehension. School closures caused a shift in plans to a more exploratory project. Subsequently, the present study aimed to explore the age at which children are believed to acquire the understanding of spoken words from different word classes, and how this vocabulary acquisition relates to features of Urdu morphology and orthography.

Vocabulary acquisition is explored through the development of a small print corpus using descriptive corpus linguistics method. Children's first to third-grade textbooks were selected as reliable sources of child-directed written input. Since the aim of corpus development was extracting a morphologically representative word list, parts-of-speech tagging and frequency calculations were planned to accurately represent the proportions of word classes children encounter in their textbooks. A secondary aim of the study was producing tools that can be used in other Urdu research projects. Therefore, consensus analysis (assessed through inter-rater agreement) and a post-hoc analysis of discrepant words was also planned. Since a school-based intervention study was not viable, the study also aimed to explore learning and literacy practices

available to children through an analysis of the activities in their Urdu textbooks with an identical focus on the role of Urdu morphology and orthography for vocabulary development.

1.6. Research Questions

Considering the importance of morphological and orthographic awareness, two research questions were formulated to assess the rated age at which children acquire certain words, and the psycholinguistic associations of these ratings with frequency, letter and syllable count, and orthographic transparency. The role of textbooks in vocabulary development was explored through a third research question. The research questions are as follows:

1. What are the age of acquisition ratings for different word classes in a word-list derived from child-directed written input?
2. What psycholinguistic associations do these ratings have for each word class?
3. How do activities in early grade Urdu textbooks contribute to the development of children's morphological and orthographic awareness?

Chapter 2: Literature Review

2.1. Chapter Overview

First, this chapter outlines the cross-linguistic literature on Age of Acquisition (AoA). It then provides a summary of corpus linguistics methods, particularly in the study of child language acquisition. An overview of methods used to analyze textbooks follows. The chapter concludes with the theoretical framework that informs the present study on Urdu vocabulary development.

2.2. Age of Acquisition (AoA)

The AoA of a word is defined as the age at which that word is first learned (Morrison et al., 1997). This section begins with a summary of AoA findings in English. The following section details the study of AoA in other linguistic contexts.

2.2.1. Evidence from English

The AoA effect – or the tendency of earlier acquired words to be better retained than late acquired word– has been documented in English (Morrison & Ellis, 1995) and has been used to study the relationship between AoA and other psycholinguistic variables. One such variable that has been linked to object (or noun) naming is word frequency (Oldfield & Wingfield, 1965). However, Morrison et al. (1992) conclude that AoA is the main determiner of object naming speed. The role of frequency in the observed AoA effect still remains contentious. For instance, Zevin and Seidenberg (2002) propose the cumulative frequency hypothesis to explain the AoA effect using computational modelling. Cumulative frequency is defined as the rate of occurrence of a word in a language. These researchers also hypothesized that an AoA effect is observed for words with less predictable orthographic mappings. In addition to naming speed and frequency, AoA is also associated with other psycholinguistic variables, including semantic variables such as concreteness or the imageability of a word (e.g. Monaghan & Ellis, 2002), and lexical variables such as word length (e.g., New et al., 2006).

Not only are several explanations posited to explain the AoA effect, but it is also studied using different methods. The two main methods are objective and subjective measures, which are deemed comparable (Gillhooly & Gillhooly, 1980). Objective measures test children's vocabularies by asking child participants of various ages to name words represented by pictures. Scoring is based on the average number of children in an age band that correctly name the word (Morrison et al., 1997). In subjective measures, adult participants are asked to guess the age at

which they think a given word is learned on a Likert-type scale (Gilhooly & Logie, 1980). Several studies comparing the two methods have found strong associations between them. For example, Gilhooly and Gilhooly (1980) used multiple regression analysis to assess the effect of subjective AoA on a word list rated on word learning age by children. They found that subjective AoA predicted the age at which children had learned words on the list. Similarly, Walley and Metsala (1992) found strong significant correlations between adults' subjective ratings of words and children's ratings of the same words for early, current, and late acquired words. Based on these converging findings from different measures of AoA, Morrison et al. (1997) conclude that objective measures are preferred, but subjective ratings can be used reliably and validly.

The evidence from English is both theoretical and methodological in nature. Theoretically, the AoA effect is not disputed, but there is contention about the validity of the construct with some researchers suggesting a masked effect of frequency, or an effect limited to words with arbitrary orthographic mapping (e.g., Zevin & Seidenberg, 2002). Such theoretical debates become interesting when the linguistic context changes to a less alphabetic language than English (such as Urdu). The methodological evidence allows for flexibility in conducting AoA studies when objective measurement may not be possible.

2.2.2. Cross-linguistic Evidence

2.2.2.1. Establishing AoA Norms in Other Languages. To use AoA as an independent variable in language and literacy studies for various under-studied languages, researchers first need to develop AoA norms for when children can reliably be expected to acquire certain words. Norming procedures require word-level data for several related psycholinguistic variables such as frequency (written and spoken depending on the research questions of interest) and letter, syllable, and phoneme counts as well as morphological features such as word classes (Alonso et al., 2015). Such data is more readily available in languages like Spanish, and large-scale subjective norming studies like the one reported by Alonso et al. (2015) are more frequent. Previous norming studies in Spanish had relied on object names and used both a smaller number of materials and participants (Cuetos et al., 1999).

However, such data is not readily accessible in all languages. In a study on AoA effects in Persian, Bakhtiar and Weekes (2015) first collected subjective AoA ratings from 40 adult Persian speakers on a word list extracted from a Persian word corpus. Similar measures had to be collected by Göz et al. (2017) on their study on Turkish. Researchers are limited to the type of

corpora available. Göz et al.'s (2017)'s word list was derived from a written corpus and Bakhtiar and Weekes (2015) used a corpus developed by the Iranian government. Neither corpus may accurately represent the language as a whole or the type of language children are exposed to. Similarly, these studies also necessitate norming for other variables of interest (such as imageability ratings or orthographic properties). The AoA norms developed may also not be representative of the entire language. Bakhtiar and Weekes (2015) only extracted monosyllabic words whereas Göz et al. (2015) only used nouns and adjectives as opposed to other word classes in Turkish. Most norming studies provide AoA norms collected solely through subjective measures, even if there may be divergence in objective and subjective measures in these languages. For example, Göz et al. (2017) found an age effect with twelve year-olds giving lower mean AoA ratings than three adult groups.

Despite these challenges, norming studies in languages with limited access to word corpora and psycholinguistic databases report strong reliability on both inter-rater and internal consistency measures. The methods developed by such researchers (as those mentioned above) can be applied to languages like Urdu in which both psycholinguistic norms and word corpora may not be accessible.

2.2.2.2. Evidence for AoA Effects in Other Languages. AoA studies in many languages replicate the negative correlation between AoA and word frequency that are reported in English (e.g., for Portuguese, see Cameirão & Vicente, 2010; for French, see Bonin et al., 2004; for American English, Czech, Gaelic, Lebanese Arabic, Malay, Persian and Western Armenian, see Łuniewska et al., 2019). Such findings further the universality of early word-learning and vocabulary development's association with word frequency (Tomasello, 2003).

However, some language specific conclusions can also be drawn from such cross-linguistic work, especially with regards to the orthographic and morphological properties of languages. For example, Chen et al. (2007) furthered the arbitrary mapping hypothesis proposed by Zevin and Seidenberg (2002) in a study on Chinese by manipulating the predictability of the phonology to orthography mapping of words. Chen et al. (2007) found a larger AoA effect for words with lower orthographic predictability. This finding, complements findings from Persian as Bakhtiar and Weekes (2015) found larger AoA effects in naming speeds for words with opaque rather than transparent spellings. Both sets of researchers extend support for the arbitrary

mapping hypothesis, but their research also points to the unique contributions of a word's orthography to when it is learned and how it is stored and processed later.

Not only have more recent norming studies included AoA ratings for word classes besides nouns (such as Alonso et al., 2015; Cameirão & Vicente, 2010), but AoA studies in other languages have also looked at the relationships between word classes and AoA. For example, Khwaileh et al. (2018) found significantly higher mean AoA ratings for abstract nouns and verbs than concrete (and more imageable) nouns and verbs, pointing to the correlation of imageability and AoA. Additionally, unlike English (Bird et al., 2001), these researchers found higher AoA ratings for nouns compared to verbs and adjectives. While this finding points to the association between AoA and imageability, it may also be indicative of morphological or orthographic differences between language as Khwaileh et al. (2018) did not find rated AoA differences in concrete nouns and verbs in an earlier study. The development of AoA norms that represent multiple word classes allows for greater representation of the language and its morphological features (as some word classes like adverbs and adjectives are often morphologically marked). Differences in the associations between word classes and AoA ratings that diverge from those reported in English imply the need for further exploration of the role of word classes in derivationally rich languages like Arabic and Urdu.

To conclude, the age at which a word is acquired can influence how it is stored and processed as studies on naming and reading speed suggest. Research from other languages complements evidence from English by providing support for the universality of negative associations with frequency, and support for the arbitrary mapping hypothesis. Simultaneously, cross-linguistic research also calls for further exploration of features of the AoA effect in English, such as lower mean ratings for nouns, because the AoA of different word classes may vary depending on the morphological and orthographic features of other languages.

The next section outlines some methods from corpus linguistics that can be used to extract words for an AoA study.

2.3. Corpus Linguistics

Corpus linguistics can be described as a branch of linguistics in which researchers apply tools to exploit texts derived from both speech and print corpora to answer research questions (McEnery & Hardie, 2011). This section first outlines how corpora have been used to investigate research questions in the field of child language acquisition, before detailing extant Urdu corpora, and

finally discussing ethical and methodological considerations in corpus methods and generalization of findings.

2.3.1. Studying Children's Language Development

Corpus methods have been used to study different aspects of child language acquisition since the 1950s (Templin, 1957) and have become increasingly common and ambitious with technological innovations and digitization techniques. Corpus methodology's popularity in the field of child language acquisition can be partially attributed to children's limited metalinguistic awareness (McEnergy & Hardie, 2011), making corpora of children's language an important source of otherwise inaccessible data. Additionally, corpora can be used to study theories of language development that include frequency at the level of individual words and word co-occurrence patterns as predictors of children's language learning. Under such theories like the encounter- and usage-based account of Tomasello (2003), corpora can be exploited for linguistic frequency and concordance analyses. These analyses enabled by corpora are also useful from an educational standpoint as both child-produced and child-directed input (such as children's schoolwork and textbooks) are indicative of the quality of language learning in the school context.

A variety of research questions about child language acquisition are now being asked because of advancements in the development of speech and print corpora. For instance, CHILDES is a speech corpus that contains datasets in several languages with child-produced and child-directed speech (MacWhinney, 2000). Ashkenazi et al. (2016) used dyadic parent-child interactions in Hebrew from CHILDES to identify distributional cues that help infants detect morphological patterns for verbs. While CHILDES is a speech corpus, the Lancaster Corpus of Children's Writing Project (LCCWP) (Smith & McEnergy, 1998) is a collection of nine to twelve-year-old children's visual and textual project work in English. Ormerod and Ivanič (2002) used the LCCWP to study material resources children use in meaning-making through content analyses and interviews with the children whose projects were included in the LCCWP. However, as the next subsection illustrates, much of the research on child language acquisition studied through corpus methods is only possible in languages with extensive, accessible, and user-friendly corpora.

2.3.2. Available Urdu Corpora

Usable print corpora are limited in under-studied languages like Urdu. Developing corpora requires both manual and technological resources in addition to linguistic expertise in terms of the textual markup and annotation of the corpus (McEnergy & Hardie, 2011). Moreover, while a wide variety of annotators (for e.g. annotations for word classes) have been developed for languages like English (see for e.g., Leech et al. 1994; Piao et al., 2005), linguistic annotation has not been automated and must be undertaken manually for several languages (McEnergy & Hardie, 2011). This manual annotation reduces the practical feasibility of corpus development and compromises the reliability and validity of using corpora. As the following sections show, existing Urdu corpora face several practical challenges.

2.3.2.1. Focus on Adult-Directed Written Input. Currently, only a few adult corpora exist in Urdu, but these may not be relevant for the study of child language acquisition. To the best of our knowledge, some word-level corpora (e.g., see Ijaz & Hussain, 2007; Urooj et al., 2012) and a single sentence level corpus (Humayoun et al., 2016) can be identified in Urdu. All of these are examples of print corpora that have been developed mostly from news websites (e.g., Ijaz & Hussain, 2007) or Urdu magazines (Urooj et al., 2012). While the source materials for these corpora are convenient to sample, they might not be reflective of the range or frequency of Urdu vocabulary in other print materials such as books, and might be even less representative of child-directed material. A large corpus size, especially one derived from the internet, does not mean the language is represented well. Drawing conclusions from an unrepresentative corpus affects the validity of findings (Leech, 2007).

2.3.2.2. Unreliable Annotation Software. Some Urdu corpus developers have also constructed and tested annotation software, particularly for word classes, but such taggers are not without their own shortcomings. For example, Jawaid et al. (2014) consolidated three different parts-of-speech taggers for Urdu and developed a standalone tagger to analyze their corpora of 94.5 million tokens extracted from Urdu webpages. They found that while all taggers had a general accuracy rate of around 87%, the accuracy for unknown words was only 55%. Their stand-alone tagger is open-access, but it has not been validated by other researchers, requires extensive machine-learning knowledge, and can produce multiple tags for the same word. These factors coupled with the low accuracy rate and a general space insertion problem in Urdu corpora (Humayoun et al., 2016) make this annotation technique unsuitable for a small corpus. Additionally, since it has only been trained on adult print corpora from online publications, its

generalizability to semantically motivated child-directed print corpora cannot be ascertained. Therefore, at the present level of Urdu corpus linguistics, developing a child-directed print corpus might require manual annotation, especially since there is no consensus on accepted and reliable taggers. In addition to suitability, reliability, and validity, researchers must also consider the ethics of corpus methodology, as the next section briefly discusses.

2.3.3. Ethical and Practical Considerations in Corpus Linguistics

Researchers working with corpora have to take several ethical principles into consideration that are particular to corpus methodology. For instance, researchers constructing and using corpora have to make layered decisions about data protection as speech and written samples may involve discussions of third parties (who are not participants themselves). Removing names might not be sufficient as participants may mention other identifiable information (McEnery & Hardie, 2011). Data protection issues are compounded when extended to corpora that include speech or written material from children as children, especially infants, cannot provide informed consent, particularly if their speech or written samples are made accessible to other researchers (through open-access corpora like CHILDES).

The corpus' representativeness as a sample of the entire language (i.e. the population) introduces bias. As mentioned in the previous subsection, generalizing findings drawn from a corpus from a limited range of sources may be unrepresentative of the language as a whole (Leech, 2007; McEnery & Hardie, 2011). This representativeness includes the numbers and types of words as well as word and pattern frequencies. Speech corpora may not accurately represent the speech directed to and produced by children as the act of wearing recording devices may be especially unfamiliar to them.

Ethical and practical considerations make it almost impossible to ensure a corpus is fully representative, but acknowledging the limits of representation in the research may be helpful (McEnery & Hardie, 2011). Researchers can also be cautious in the claims they make about their findings. Therefore, attention must be paid to ethical, practical, and methodological concerns when constructing, using, and analyzing a corpus, in order to appropriately incorporate corpus methods in linguistics research.

Like corpus methods, textbooks can be a useful and convenient source of child-directed written input. The next section of the literature review provides an overview of textbook analysis.

2.4. Textbook Analysis

This section outlines the methods of textbook analysis that have been used for language textbooks in different linguistic contexts. A definition of textbooks and textbook analysis, however, is warranted before a review of the extant research in the field. Valverde et al. (2002) describes textbooks as those tools that are most closely related to instruction and learning because of their ability to structure the course and operationalize curricula. Tomlinson (2012) lists several advantages of using course books, including (but not limited to) standardizing curriculum, providing resources to teachers to structure their lessons, and enabling learners to track their progress. Textbooks' intermediary role between curriculum aims and teaching and learning in the classroom, coupled with the advantages presented above, necessitate pre-use (i.e. before selecting a textbook for a course) as well as in-use (i.e. during the course of teaching from a textbook) analysis. Cunningsworth (1984) provides a framework for textbook evaluation in that teaching materials should be related to the aims of the course, consider the relationship between the language and learner with the learning process, and enable learners to use language effectively and communicatively. While this framework can be generalized to all types of language textbooks, tools developed for textbook analyses have only been applied in a narrow range of linguistic contexts, may not be relevant for primary school Urdu textbooks, and may not be empirically grounded.

2.4.1. Limited Linguistic and Instructional Contexts

Most analyses of language textbooks are restricted to textbooks used to teach a foreign language, especially English (see Tomlinson, 2012; Ansary & Babaii, 2002 for a review). A process of checklist evaluation has been developed by several researchers (e.g., Miekley, 2005) to analyze foreign language textbooks. Such checklists can be used by curriculum developers and teachers when selecting a textbook for their course or evaluating the appropriateness of a textbook already in use. These checklists for foreign-language instruction have been adapted to different cultural contexts with an intentional incorporation of the learner's social context in the tool. For example, in their application of Miekley's (2005) checklist to the Iranian and Turkish L2 English-learning context, Mobarakeh and Arani (2012) situate the textbook's use in the learner's life by including a subsection on the relevance of the content to the learner.

Methods such as comparing item frequencies or using teacher questionnaires are also used. For example, Arslan (2016) and Bahrami (2014) were both interested in the relative

representation of Turkish and Western cultures in their respective studies, and gathered teachers' perceptions through questionnaires. Furthermore, there is an emphasis on focusing on the four skills deemed especially important in second language learning (reading, writing, listening, and speaking) (Tomlinson, 2012). Fatima et al. (2015) analyzed English textbooks used in secondary schools in Pakistan and included these four skills (as well as vocabulary and grammar) in their checklist for an analysis of textbook activities based on item frequencies. This variability in tools (i.e. checklists, item frequency comparisons, and teacher interviews) however has largely been limited to foreign language textbooks, specifically to English.

When methods of textbook analysis have been extended to other languages, the extension has been reserved for second language teaching in higher education, thereby limiting the range of instructional contexts of analyzed textbooks. Rababah and Al-Rababah (2013), Maskor et al. (2016), and Machmudah et al. (2019) all analyze Arabic textbooks used to teach Arabic as an additional language in the context of Oman, Malaysia, and Indonesia respectively. However, like the research on English textbooks reviewed above (except for Fatima et al., 2015), much of the research in Arabic is conducted with either high school students or adults at the university level and the checklist is designed for teachers constructing course syllabi. This target age band means criteria relevant for young learners (such as the age appropriateness of texts and activities) or first language learners (such as specific morphological properties of the language) are not included in the checklists.

2.4.2. Relevance for Early Grade Urdu Textbooks

Methods developed for analyzing foreign language textbooks in English for older students may not be applicable to early grade Urdu textbooks for first-language learners. Such applications may be limited in the component skills they target. Moreover, these tools cannot easily be adapted to a different linguistic and instructional context. Since Urdu adaptation has not taken place yet, Rahman and Sinha's (2010) work on fifth-grade Hindi textbooks in India is perhaps the most relevant to the present study due to morphological similarities between the two languages. For their study with students and teachers of Hindi, the researchers adapted and translated questionnaires developed to analyze foreign language English textbooks taught in Turkey. The questionnaires consisted of 47 items that were rated on a five-point Likert scale from "totally lacking" to "excellent." Each questionnaire was divided into five close-ended sections on physical appearance of the textbook, aims and goals, subject matter, vocabulary

structure, and exercises and activities. The last section consisted of different open-ended questions presented to teachers and students. Each item collected students' and teachers' opinions on vague statements. For example, in the vocabulary section, they were asked to rate items such as "reasonable vocabulary load," the operational meaning of which was left to the interpretation of students and teachers.

Rahman and Sinha's (2010) adaptation of a checklist designed for evaluating textbooks that teach a foreign language to evaluating an early-grade first-language textbook raises serious concerns about the validity of adapting tools across cultural and instructional contexts. Additionally, these checklists are not adapted to include the linguistic contexts of the learners. Instead, general items about language domains such as vocabulary are included. Even after their textbook analyses, the distribution and appropriateness to Hindi of each activity or text in the textbook remains unassessed. Features particular to Hindi, such as its orthographic features that influence reading and writing are also excluded from the checklist. This study, therefore, highlights the narrow scope of textbook analysis to date.

Research using textbooks in orthographically similar languages such as Arabic is not relevant to Urdu either as the focus here is exclusively on linguistic properties. The role of the learner, the learning process, and course aims (Cunningsworth, 1984) are not analyzed. For instance, Belkhouche et al. (2010) compared the linguistic properties of Arabic textbooks published in the UAE to textbooks from Libya. While their research is useful in highlighting the type of child-directed print language available (for example, an over-emphasis on action verbs), such work in corpus linguistics does not analyze features of the textbooks themselves. On the contrary, Rahman & Sinha's (2010) research completely lacks a linguistic analysis of the textbooks. Even in-depth item frequency analyses like those proposed by Fatima et al. (2015) do not encompass specific linguistic properties of the language. This methodological divergence means all three types of studies fail to individually follow Cunningsworth's (1984) framework in that these analyses do not consider the relationship between the language, learner, and learning process, or the aims of the course. Interestingly, the checklist method may include effective communication and course aims (Cunningsworth, 1984) in their items, but are too broad in their scope to analyze the learning process or relationships between learners and languages. Perhaps it is unsurprising then that adaptations of such tools also lack this important feature of Cunningsworth's (1984) framework.

2.4.3. Empirical and Contextual Grounding

The focus on either exclusive corpus linguistics methods or exclusive qualitative methods of data collection and simple descriptive statistics when quantitative data is collected problematizes the empirical standing of textbook analysis by limiting the conclusions researchers can draw from their analyses (Tomlinson, 2012). When researchers like Rahman and Sinha (2010) then use such global checklists whose validity is not well established in other linguistic contexts where the phonology, orthography, and morphology are different, their conclusions about the appropriateness and effectiveness of a particular textbook might equally suffer from a lack of empiricism. These problems are relatively unique to first-language textbooks simply because of differences in psycholinguistic and sociolinguistic contexts. Extending global checklists might be more useful for subjects like mathematics (O’Keeffe, 2013) or science (Eltinge & Roberts, 1993) where curriculum aims are shared between different countries, or when looking at foreign language textbooks of the same language in different cultural contexts (Guilloteaux, 2013; Laabidi & Nfissi 2016; Tang, 2019) as the psycholinguistic context of the target language remains static. Regardless of their greater generalizability, these evaluation criteria still suffer from a lack of empiricism as authors tend to report percentages without performing any tests of inferential statistics.

To the best of our knowledge, no textbook analysis studies have been published for Urdu textbooks, particularly in the early school years when native Urdu speaking children are first formally exposed to Urdu print. At this level, schools and governments may have different curriculum objectives than they do for English (where English is the general medium of instruction), thereby highlighting the relevance of the social context. Nag et al.’s (2014) review outlines the importance of using assessments that use linguistically and contextually relevant psycholinguistic measures. This finding can be applied to textbook analysis as well. The context of Urdu instruction and distinct linguistic features (such as its un-vowelized orthography and derivational morphology) must therefore be incorporated in any checklist or instrument developed to analyze Urdu textbooks, instead of relying on broad checklists developed for teaching an additional language.

2.5. Theoretical Framework

This chapter has highlighted gaps in the resources available for researchers interested in the development of children’s Urdu vocabulary, and related instructional implications. The

present research aims to use established methods of corpus linguistics (such as parts-of-speech tagging, and frequency counts) reviewed by McEnery and Hardie (2011) to develop a word list from child-directed written input. Children's textbooks will be digitized to develop a sentence corpus that lends itself to linguistic annotation. A word corpus will be derived from the annotated sentence corpus, and a word list will be extracted from this corpus based on frequency counts. This word list is to be used for a subjective AoA questionnaire with adult raters, a method that has been validated cross-linguistically.

Evidence from AoA studies in languages with rich derivational morphology also suggests that the age at which different word classes are acquired may differ across languages (Khwaileh, 2018). Therefore, the theoretical framework of the AoA study includes an orthographic dimension whereby Zevin and Seidenberg's (2002) arbitrary mapping hypothesis is assessed through variance in the vowelization of Urdu orthography. The theoretical framework also contains a morphological dimension whereby the word list includes both lemmas and inflected forms, data is collected for six different word classes, introductory psycholinguistic norms are developed for each word class, and the noun advantage (Gentner, 1982) is explored.

Additionally, when children learn certain words also has educational and instructional implications. Therefore, textbook analyses of activities found in children's primary school Urdu textbooks are planned in conjunction with the AoA study. The learning process aspect of Cunningsworth's (1984) framework is central to this analysis of activities in first and third-grade textbooks' that promote children's morphological and orthographic awareness. It is hoped that the focus on morphological and orthographic activities will highlight how children specifically learn Urdu through their textbooks as opposed to more general language learning. Together, the AoA study and textbook analysis can begin to fill the current gap in the literature on Urdu vocabulary development between the first and third grades. The research objective is to study when children understand spoken words in different word classes, and textbooks' role in promoting children's vocabulary development by combining three methods— corpus methods of parts-of-speech annotation and frequency calculations, subjective AoA ratings collected from knowledgeable participants with high exposure to children, and an items frequency analysis of textbooks' activities.

Chapter 3: Methodology

3.1. Chapter Overview

This chapter delineates the methodology used to investigate children's Urdu vocabulary development through an AoA study and textbook analysis. After outlining the aims, research questions, and research design, the chapter provides a detailed account of the techniques used to develop a child-directed sentence corpus from early grade textbooks. The following subsections focus on the sequential process by which a semantically motivated word list is extracted from the sentence corpus. Then, the chapter reports the materials (i.e. the AoA questionnaire and the textbook analysis template), recruitment, and procedure. A discussion of the ethics of online data collection concludes the chapter.

3.2. Aims

The primary aim of this study was to explore children's Urdu vocabulary development between grades one and three, through an online AoA questionnaire containing a word list that was representative of child-directed print. A secondary aim of the study was to analyze the language and literacy learning promoted through prescribed early grade course material from a range of publishing backgrounds, with a special focus on textbook activities that stimulate the development of children's morphological and orthographic knowledge.

3.3. Research Questions

The following research questions were examined in this exploratory study:

1. What are the age of acquisition ratings for different word classes in a word-list derived from child-directed written input?
2. What psycholinguistic associations do these ratings have for each word class?
3. How do activities in early grade Urdu textbooks contribute to the development of children's morphological and orthographic awareness?

3.4. Research Design for the AoA Study

Punch and Oancea (2014) suggest that four components, namely the conceptual framework, strategy, the instruments used for data collection and analysis, and the participants of a study (i.e. who is studied), collectively contribute to the research design of a study. Considering the exploratory nature of project, the independent variables of interest are the morphological properties of words (i.e. the word classes) in the word sorting activity while the dependent variable is participants' ratings of the age at which they think Urdu-speaking

children understand spoken words. Psycholinguistic variables, such as frequency, letter and syllable count, and orthographic transparency category are include in the analysis. The word list was annotated for word classes and other psycholinguistic variables. This annotation allowed comparisons to be made across individual items as well as across word classes.

The AoA online questionnaire is the main tool used to explore this conceptual relationship. Early years' teachers and parents of young children are identified as the population to be studied due to their knowledge of children's language abilities. Finally, a within-subject strategy is selected to analyze the data to not only control for the effect of confound variables such as individual differences in Urdu proficiency or frequency of exposure with children, but also to compare the psycholinguistic properties of individual words. Such a within-subjects design allows for more robust conclusions to be drawn about children's Urdu vocabulary development, which, given the dearth of research investigating children's language acquisition in Urdu, was of special import when designing the study.

3.5. Developing a Child-Directed Print Corpus

This section provides a step-by-step account of the various stages of corpus development.

3.5.1. Textbook Selection

The lack of a pre-existing child-directed corpus necessitated the development of one that could be used to extract a theoretically informed word list to be used in the AoA word sorting activity. Three publishers of Urdu textbooks were selected and their first to third-grade textbooks (i.e. a total of nine textbooks) were identified as the source material for the word corpus. Two sets of books were published for use in private schools in Pakistan. The first set, is published by Oxford University Press, whereas the second book is published by Kifayat Publishers. The third set, published by the National Council of Educational Research and Training, is used in Indian schools that offer Urdu as an additional language.

3.5.2. Transliteration Technique

All connected text from the textbooks (including poems, passages, stories, and essays) were transliterated and digitized using the American Library of Congress' Urdu Romanization table (<https://www.loc.gov/catdir/cps/romanization/urdu.pdf>). As the textbook analysis template focuses solely on the learning activities in each textbook, these activities were not included in the print corpus. This transliteration scheme accounts for Urdu's deep, predominantly un-vowelized orthography, homophonic graphemes, and loan-words from other languages like Arabic and

English. These nuanced considerations allowed for homonyms and orthographically similar words to be distinct in their transliteration. The scheme's reliance on special symbols (such as 'ṭ' for the Urdu grapheme 'ط') were not compatible with statistical software packages like R and SPSS that were needed to generate the word corpus and meaningfully extract words for the AoA study. Therefore, the original scheme was adapted to replace these special symbols (for example, 'ṭ' was replaced by 'ttt'; see Appendix B for a full list of changes). Due to timing constraints, every 4th sentence (or verse in a poem) were transliterated. The final transliterated sentence-level corpus consists of 1,008 sentences and verses.

3.5.3. Word Class Annotation and Inter-Rater Reliability in the Sentence Corpus

To answer the first two research questions that focus on the role of word classes in children's vocabulary learning, the sentence corpus was annotated according to six word classes: nouns, verbs, adjectives, adverbs, conjunctions, and postpositions. The selection of these word classes was based on Schmidt's (1999) framework for parts-of-speech in Urdu. Schmidt's (1999) framework categorizes all words in Urdu into ten word classes. In the absence of a pre-existing print corpus that provides the frequency distribution of different word classes, it was not possible to rely on already established frequency norms. Therefore, the selection of the six word classes in the present study was based on their expected frequencies and their ability to provide a range of both content and function words (conjunctions and postpositions are classified as function words).

In order to reliably annotate the sentences according to their word classes, the researcher and two different coders tagged the sentences according to the six word classes selected from Schmidt's (1999) framework. Widely researched languages like English now have analysis software that automatically tag corpora by their parts-of-speech, such as CLAWS (Garside 1987), making corpus annotation a purely digital enterprise. As Urdu does not have a reliable and valid parts-of-speech tagger, the corpus had to be manually annotated.

To ensure reliability, 504 sentences from the sentence corpus (i.e. 50% of the sentence corpus) were annotated by two separate coders and inter-rater reliability was calculated. The researcher annotated all 1,008 sentences, while two additional coders rated 252 sentences (or 25% of the sentence-corpus) each, which meant 504 sentences could be used in the inter-rater reliability analysis. The two additional raters were first-language Urdu users with undergraduate degrees. Both received a table based on Schmidt's framework and disagreements were discussed

and resolved after every 50 sentences had been annotated. Inter-rater agreement of categorical variables, such as the different word classes of interest in the present study, can be measured through Cohen’s kappa coefficient (1960), a robust measure of agreement. Like other correlation coefficients, the kappa coefficient assesses the degree of similarity between different variables (in this case, the researcher’s and additional raters’ word class tagging), whilst accounting for the possibility of chance agreements. Generally, a kappa value of .81-1.0 is considered strong to near perfect (Landis & Koch, 1977). As Table 3.1 illustrates, the kappa coefficient ranged from .84-.93 for the six word classes selected. Therefore, inter-rater agreement for these parts of speech can be considered strong.

Table 3.1

Word Class	Kappa
Nouns	.92
Verbs	.84
Adjectives	.87
Adverbs	.84
Conjunctions	.93
Post-Positions	.91

Inter-Rater Agreement for Word Class Annotation

3.5.4. Frequency Calculations

After the sentence corpus was annotated for word classes, token and type frequencies were calculated for individual words as well as each word class. While token frequency is a measure of the total number of all words, type frequency is a measure of the number of different words (Pan et al., 2005). A word corpus was developed, which was then used to extract the word list for the AoA questionnaire. The transliterated sentence corpus consisted of 10,186 total words (tokens) and 2,168 unique words (types). After annotating this corpus for word classes, 1,968 unique words (types) and 7,036 total words (or tokens) were identified and used for further analysis. The word corpus has fewer types and tokens than the sentence corpus because some word classes (such as interjections and particles) were not included in the word corpus as the present study was interested in the most frequently occurring content and function words.

As Table 3.2 shows, nouns had the highest frequency (44.5% of the word types in the annotated word corpus were classified as nouns), followed by verbs (33.2%), adjectives

(13.8%), adverbs (6.8%), conjunctions (.9%) and postpositions (.7%). These absolute frequencies were calculated to ensure the final word list represented the proportion of word classes in the word corpus.

Table 3.2

Word Class	Tokens	Types	Frequency Range	Word Type Frequency (as % of word corpus)	Word Type Frequency (as % of sentence corpus)
Nouns	2009	875	1 to 41	44.50%	40.30%
Verbs	2280	653	1 to 124	33.20%	30.10%
Adjectives	662	272	1 to 64	13.80%	12.50%
Adverbs	565	134	1 to 45	6.80%	6.20%
Conjunctions	474	19	1 to 173	0.97%	0.88%
Post-Positions	1046	15	1 to 168	0.76%	0.69%
Total	7036	1968		100.00%	90.67%

Frequencies by Word Classes in Word and Sentence Corpus

To further ensure the word list created was representative of the frequency distribution of individual words in the print corpus, the frequency of each word in the word corpus was calculated. All frequency calculations were performed on a corpus linguistics software, AntConc (Anthony, 2019), which has an in-built frequency calculator in its “Word List” tab. Once the frequency of individual words was determined, the “quintiles” function on R was used to calculate the frequency bands of each word class. As Appendix C demonstrates, none of the word classes were normally distributed and all of them had very large frequency ranges, with most of the items occurring only once in the word corpus (e.g., 60% of the nouns occurred just once).

Considering the small word corpus size, and the aim of developing a word list that was representative of child-directed print in terms of children’s frequency of exposure to different words, all items in the top quintile (with the most frequently occurring words) were selected and an equal number of words from the remaining quintiles were included in the AoA word list. The total number of words selected from each word class was based on the absolute frequency distribution presented in Table 3.2. The word class distribution of the final word list (N = 600) was as follows: 44.5% of the words were nouns (N =265), 33.2% verbs (N = 201), 13.8%

adjectives (N = 81), 6.8% adverbs (N = 41), .9% were conjunctions (N = 6), and .7% were post-positions (N = 6).

3.5.5. Orthographic Tagging

Another psycholinguistic variable of interest was orthographic transparency because of Urdu’s deep and predominantly (but not completely) un-vowelized orthography (Ijaz & Hussain, 2007). The 600 words selected through the word class annotation and frequency calculation were then subjected to transparency categorization. Words spelled with just graphemes representing consonants (i.e. they contained neither graphemes representing long vowels or diacritics to mark vowelization) were tagged as “opaque.” Words were tagged as “transparent” if they were spelled with either graphemes representing long vowels or contained diacritics to mark short vowels. Finally, multisyllabic words which contained at least one vowelized and at least one un-vowelized syllable were tagged as “mixed.” The researcher independently tagged all words and inter-rater reliability was assessed by having a native Urdu user check all tags, but there were no disagreements. Table 3.3 illustrates the orthographic distribution of the AoA word list.

Table 3.3

Orthographic Type	Frequency	Cumulative Percent
Opaque	103	17.1
Transparent	245	58
Mixed	252	100

Transparency Category Distribution of AoA Word List

The final word list of 600 word types was therefore annotated for word classes, frequency, and orthographic transparency. As letter and syllable counts are included in most AoA studies, the researcher also noted the letter and syllable count for each of the extracted words.

3.6. Recruitment

Recruitment efforts were initially organized around directly contacting the head teachers of private schools in Pakistan. However, due to the indefinite closure of schools, snowball sampling with the researchers’ personal contacts was used to recruit parents and teachers of children between the ages of three and six (or nursery to first grade in Pakistan’s education system). This group was selected as previous research has established that subjective AoA ratings are a valid estimate of AoA (Gillhooly & Gillhooly, 1980). Additionally, this group’s high exposure to children growing up as native Urdu users means they are better

positioned to rate words based on their experience with children compared to general adult raters. All participants, except one, were recruited directly from the researcher's pre-existing contacts. The remaining participant was recruited through a Facebook group for schools in Karachi. The researcher posted a short description of the study on the group (see Appendix D) and interested teachers and parents commented and shared their email addresses with the researcher. Once an email address had been received, the same procedure of sending an email introducing the study (see Appendix E) with an attached consent form followed.

3.7. Participants

Eleven participants (nine teachers, one parent, and one participant who fit both the teacher and parent criteria) consented to participating in the study and completed the first session of the AoA questionnaire. However, due to technical difficulties with Zoom (i.e. the video-conferencing platform used in the study), the parent participant had to be dropped from the study. All remaining participants were female between the ages of 24 and 60 years ($M = 37.6$, $SD = 13.7$). Seven participants taught at either the first or second levels at a religious institution that offers after-school Quran learning and Islamic Studies classes to children. Students in the first two levels of this institution attend kindergarten or first-grade, and are approximately between the ages of five and six years. One participant teaches first-grade Urdu at a private school and also has a six-years old child. One participant offered private tuitions (in all subjects, including Urdu) to children in the first and second grades. Lastly, one participant who had recently taught at the primary school level is now working as a speech pathologist with children and adolescents.

3.8. Tools and Measures

3.8.1. Qualtrics Surveys

As detailed in section 3.5, 600 words were extracted from the word corpus for use in the AoA questionnaire. The AoA questionnaire was developed on Qualtrics' online survey software and was designed to be both mobile-friendly and desktop-friendly. To prevent fatigue effects, the word-list was divided into three questionnaires, each containing 200 words randomly selected from the complete word list. The questionnaires were identical, except for the first questionnaire which collected some demographic information from the participants (i.e. age, gender, the number and age of their children (if they were parents), and the subjects and grade-levels they taught (if they were teachers). A seven point Likert-type scale with scale points ranging from 2-3

years to 12 years+ (adapted from Morrison et al., 1997) was used for participants to match the word to the age at which they thought children understand a spoken word. All information, including instructions and words were presented first in English and then in Urdu. Qualtrics' default font settings (i.e. typeface Calibri and a font size of 11) were used for the surveys. The Urdu translations of the information and instructions were written by the researcher and a native Urdu user checked the translations and suggested amendments, which were incorporated. After reading a summary of the research and providing consent for participation in the session, participants were presented with five words on a screen which were written in the Urdu script and a colloquially transliterated form right next to it, making a total of 40 questions per questionnaire. A progress bar at the bottom of the screen allowed participants to see how much of the survey had been completed. Participants saw a notification prompting them to answer a skipped question. Appendix F contains a sample of the first questionnaire as it was presented to participants.

3.8.2. Textbook Analysis Template

Vagh & Nag's (2014) textbook analysis template for South Asian languages was adapted for Urdu. While a single researcher can code the data, the template developers encourage including two raters for inter-rater reliability.

The template contains two sections. The first section, "overview," provides a general analytical framework for each textbook as a whole by coding number of chapters and activities in each chapter, and genres, such as poetry and narrative texts. In the adapted version, section two, called "area-wise analysis," includes fifteen language and literacy component skills, such as "Vocabulary," "Reading Comprehension," and "Phonological Awareness." Each activity in the textbook is coded under the various subsections included in each component skill. For instance, "Vocabulary" contains ten different subsections and coders judge which of those subsections is suitable for each activity. The same activity can be coded for multiple component skills. The chapter, page and activity numbers, instructions for each activity, and three examples of the activity are noted.

Changes were made to the template to capture Urdu's distinct linguistic properties. To this end, sections particular to the orthography of Indian languages using the *akshara* writing system were edited to make them more specific to the nastaliq script and Urdu's orthographic rules. Therefore, a new component skill, "*Huroof e Tahajji* Knowledge," which includes separate

subsections on joiner and non- joiner huroofs, huroof positions, and homophonic graphemes (Fatima et al., 2018) replaced the section on “Akshara Knowledge” in the original template. Additionally, a component skill entitled “Morphological Awareness” with six subsections was added to the template. Fatima et al.’s (2018) research on Urdu informed the development of this component skill in the template, particularly the subsection entitled “*wahid/jama* (singular/plural),” which is further subdivided into three sections that capture different plural forms in Urdu (simple, altered, and broken). “Derivational morphology” is another subsection developed especially for Urdu as this section codes activities in which children learn how to use or attach infixes to word stems to create new words.

Some of the existing component skills were also retitled or removed. As both the component skills of “Decode Sentences” and “Reading Comprehension” capture reading comprehension, but at the levels of sentences and passages respectively, these skills were reframed. In the current template, “Reading Comprehension: Sentences” replaces “Decode Sentences,” and “Reading Comprehension: Passages” replaces “Reading Comprehension.” The subsections within the component skills were not changed. The titles were changed to more accurately represent the types of activities that can be coded in each component skill. These changes rendered the component skill of “Decode Passages” redundant, which was consequently removed from the template. Similarly, since our interest lay in activities that teach the orthographic system as opposed to the accuracy of children’s spellings, the component skill, “Dictation” was also removed. Lastly, “Functional Literacy” was removed as it was not related to our interest in textbooks’ promotion of morphological and orthographic awareness. The adapted template can be found in its entirety in Appendix G).

The tools and measures (the AoA questionnaire and the textbook analysis template) target similar variables of interest, particularly the development of children’s morphological and orthographic knowledge. While the Qualtrics questionnaire was developed to collect data directly from teachers and parents of children under the age of seven, the textbook analysis template was adapted to gather data from the textbooks.

3.9. Procedure

3.9.1 Pilot

As the AoA questionnaire was developed specifically for this study, a pilot was conducted to assess the clarity of instructions, and whether individual words as well as the three

questionnaires functioned as intended (Bryman, 2016). The pilot was conducted with one 55-year-old female teacher from the same religious institution several participants were recruited from. All three questionnaires were conducted on the same day with a minimum one-hour break between sessions. Conducting the pilot allowed the researcher to identify one misspelled word in the first questionnaire and three misspelled words in the third questionnaire. The spellings of these items were corrected before collecting data from participants. Additionally, the pilot enabled the researcher to practice data collection, which involved hosting a meeting on Zoom, asking and answering questions, and audio-recording data.

3.9.2. Procedure for the AoA Study

This subsection outlines the method of obtaining consent from participants, and the way data was collected in the AoA questionnaire through Zoom, Qualtrics, and participants' audio recordings.

Participants were contacted via email, whereby they received information regarding the research and a consent form to be completed and submitted by replying to the researcher's email. Two teachers did not consent to be audio-recorded. A follow-up email with instructions on how to use Zoom was sent once correctly completed consent forms were received. Zoom is a free web-conferencing platform (<https://zoom.us/>) that does not require participants to make personal accounts as Zoom meetings can be accessed simply using the meeting link. This follow-up email also asked participants to select dates and times for their three sessions. Participants were sent links for their Zoom meeting and the relevant questionnaire link five minutes before the set time.

Once participants joined the session, the researcher read out the same set of instructions to each participant (see Appendix H). For ease of access, the survey link was also made available in the chat box. Participants were reminded to share their screens and were facilitated by the researcher if they had trouble with Zoom's "share content" feature. Participants then began the survey, which took 34.2 minutes to complete on average ($N = 30$, $SD = 18.8$). If participants changed their selections or took a long time on a word, the researcher asked them to explain their rating for that word. Participants also asked the researcher questions about individual words. Twenty-eight sessions were completed with the researcher as technical constraints prevented one participant from participating in Zoom video sessions for her second and third sessions. She completed these two sessions independently, while the researcher remained

available to answer the participant’s questions through email and phone-call. All sessions were audio-recorded, except the sessions for the two participants who did not consent to be audio-recorded. More than half (N = 13) of the audio-recorded sessions were transcribed for qualitative analysis.

3.9.3. Procedure for the Textbook Analysis

The amended template described in section 3.8 was used to analyze six textbooks (three from the first and three from the third-grade) that were also used to develop the sentence corpus. Table 3.4 presents the number (and percentages) of analyzed chapters and activities from each textbook.

Table 3.4

	Total Chapters	Analyzed Chapters	Total Activities	Analyzed Activities
First Grade				
Publisher A	17	17 (100%)	70	70 (100%)
Publisher B	6	6 (100%)	148	148 (100%)
Publisher C	27	27 (100%)	123	123 (100%)
Third Grade				
Publisher A	24	5 (20.8%)	123	25 (20.3%)
Publisher B	22	5 (22.7%)	214	73 (34.6%)
Publisher C	20	5 (25%)	211	53 (25.1%)

Numbers and Percentages of Chapters and Activities in Each Textbook

The three first-grade textbooks were analyzed in their entirety. Between 21% and 25% of the chapters from all three third-grade textbooks were randomly selected and analyzed. Research on textbook analysis studies recommend sampling 10% of a text (Dukes & Kelly, 1979) for readability studies, but other researchers who have performed linguistic analyses of textbooks, deem a slightly larger chapter selection as more suitable for their analyses (e.g., 20%; Eltinge & Roberts, 1993). Therefore, a random selection of 21% to 25% of chapters from the three textbooks meets the criteria for textbook analysis. The student researcher coded all textbooks.

3.10. Ethical Considerations

Bryman (2016) notes the importance of ethical considerations in research, all of which also apply to online data collection. Accordingly, the present study adhered to British Educational Research Association Ethical Guidelines for Educational Research (2018) and the

British Psychological Society's Code of Ethics and Conduct (2018). These guidelines primarily focus on in-person data collection or data collected from online communities, but the same guiding principles of respect, competence, responsibility, and integrity (The British Psychological Society, 2018) were upheld in the present study by providing clear information to participants, encouraging them to ask questions during the Zoom meetings, allowing them to choose the times and dates of each session, and committing to sending participants a short brief of the findings at the end of the research period if they wished to learn more about the study. Additionally, the University of Oxford's Central University Research Ethics Committee (CUREC) approved the study (see Appendix I for CUREC approval).

3.10.1 Consent and Withdrawing

Participants' voluntary consent and knowledge of their ability to withdraw at any time is a central ethical guideline. Using snowball sampling and recruiting personal contacts can add extra pressure on potential participants when they are learning about a research project and deciding whether to participate. Similarly, participating in research on one's first language, which also happens to be under-studied can add to the pressures of participating. For this reason, participants were reminded of the voluntary nature of their participation and their ability to withdraw at any point during the study in all information letters provided and at the start of each session. Only one participant withdrew from the study due to technical difficulties. Participants were also given the option to reschedule to a more convenient time. Participants who did not wish to be audio-recorded were not probed or questioned.

3.10.2. Safeguarding and Confidentiality

Online data collection complicates the process of safeguarding participants' personal data and responses, and ensuring confidentiality. While the data collected was not of a sensitive nature, participants' survey responses and transcripts were anonymized by using a unique identification number for each participant. Additionally, participants who did not wish to turn their video cameras on were not asked to do so, but the researcher kept her video camera on at all times. Participants were encouraged to begin sharing their screens only when they had already opened the survey link on a browser so any other information on their device would not be visible to the researcher. As most participants attended the Zoom meetings from their homes, they were encouraged to find a quiet space for the sessions to ensure the researcher could not hear any background noises, which could have caused discomfort to participants. All electronic

data was stored in an encrypted form on a password-protected laptop and was deleted as soon as the data was transferred online to the university's secure Nexus 365 OneDrive.

Chapter 4: Results

This chapter is divided into two broad sections: the analysis strategy and findings from the AoA online questionnaire and the textbook evaluation. Each section will be presented in turn.

4.1. AoA Study

This section aims to answer the following research questions:

1. What are the age of acquisition ratings for different word classes in a word-list derived from child-directed written input?
2. What psycholinguistic associations do these ratings have for each word class?

4.1.1. Analysis Strategy

Results from the AoA study are divided into two sections, each of which independently answer both research questions. The first section is a comparison of the AoA ratings across raters, whereas the second section is an analysis of the psycholinguistic correlates of different word classes. Intra-class correlation indices using the two-way mixed, consistency, average measures model (McGraw & Wong, 1996) were planned for the complete word list, and word-classes. The model was set to two-way mixed because of the purposive sampling approach as raters were not a randomly selected sample from the population and each participant rated every word. The average measures model was selected as ratings across coders was of greater interest than ratings from any single coder. Since rank-order differences (as opposed to absolute agreements) in participants' ratings was of interest, the model was set to consistency. An analysis for discrepant word ratings between participants was planned by comparing standard deviations of individual words with the standard deviation of the complete word list. For the psycholinguistic analysis section, Pearson's correlations were planned for each of the six word classes and psycholinguistic variables.

4.1.2. Consensus Analysis

In addition to an overall calculation of the inter-rater reliability, separate intra-class coefficient (ICC) indices were calculated for the different word classes. Table 4.1 provides the means, standard deviations, and ICC indices for the complete word list and for the six word classes.

Table 4.1

	Mean	SD	ICC
Complete Word List	2.94	.89	.892
Individual Words by Word Class			
Noun	3.03	1.04	.89
Verb	2.86	.66	.84
Adjective	2.93	.97	.93
Adverb	2.83	.74	.87
Conjunction	3.07	.42	.62
Postposition	2.72	.56	.92

Means, Standard Deviations, and ICC Indices of Complete Word List and Word Classes

The complete word list mean as well as the means of different word classes fall between two and three on the Likert-type scale in the questionnaire, which represent age of acquisition in the two to five-years old age range. The inter-rater reliability for the complete word list is of good quality (ICC > 0.75; Portney & Watkins, 2000). Similarly, the ICC indices for all word classes are of good or excellent inter-rater agreement (ICC > .90), except the index for conjunctions. The ICC index for conjunctions suggests moderate inter-rater agreement (.60 < ICC < .75).

The second step in the consensus analysis was identifying discrepant words with low inter-rater agreement. Previous studies use correlations between individual participants' response latencies and mean ratings to remove participants' data that fall below two standard deviations from the main correlation (e.g., Schock et al., 2012). Since participants' response time was affected by technical and practical constraints, such as unfamiliarity with Zoom and connectivity issues, Schock et al.'s (2012) approach was not selected. Instead, the means and standard deviations of individual words were calculated. Words with standard deviations more than twice the size of the standard deviation of the complete word list were flagged as discrepant words. This generated a list of twenty-six words with 18 nouns, seven verbs, and one adverb.

Discrepancies arose from disagreements between raters with some raters sorting each of the words into the earliest age bands while others sorted them into the highest age bands. The

means of the discrepant words falls into a higher age band (i.e. 4-5 years old) than the mean of the complete word list (3-4 years), but this difference in means is non-significant ($t(26) = 1.98, p = .058$). In the word list, eight of the 19 discrepant nouns were presented in the plural inflected form instead of lemmas. Similarly, four of the seven discrepant verbs were inflected for gender marking. These morphological features were not controlled so the vocabulary of child-directed print could be accurately represented. However, such inflectional properties of words may have influenced inter-rater reliability. Transcripts of participants' interviews were assessed to identify three potential sources of discrepancy. These are presented below with an example of a discrepant word that can be explained through each source of discrepancy.

Table 4.2

Potential Source of Discrepancy	Discrepant Word	Explanation
Social context	سمندر (/smndr/, ocean)	One participant commented that children who live in parts of the country without access to the beach would not learn this word early.
Influence of English	گیند (/ge:nd/, ball)	Four participants commented on the dilution of Urdu because of a preference for English, particularly for labeling nouns.
Confusing regarding spoken or written understanding	بھینا (/b ^h jjɑ:/, older brother)	Participants were unsure about the written complexity of certain words. This word has a digraph and diacritic marker, which can be difficult to decode. Although the instructions specified spoken understanding, participants did not always seek clarification.

Three Sources of Discrepancies in AoA Ratings for Identified Words with an Example and Explanation

Table 4.2 provides preliminary explanations for discrepancy based on participants' comments during data collection. These sources of discrepancy may also have influenced participants' ratings for other words in the list. However, the high ICC indices overall suggest strong inter-rater reliability for the word list.

4.1.3. Psycholinguistic Analysis

This section presents the correlations between each word class in the word list with other psycholinguistic variables, namely written frequency (calculated as the number of occurrences in the sentence corpus), the, syllable count, letter count, and transparency category of each word. All variables were checked for normality through the Kolmogorov–Smirnov test, and z-scores for skewness and kurtosis by dividing these statistics by their standard deviations (see Appendix K for results). According to Field (2009), significant results for the Kolmogorov–Smirnov test and skewness and kurtosis z-score values above 1.96 should be checked in conjunction with each other to assess non-normality. The Kolmogorov–Smirnov test and z-score calculations for all variables suggested significant deviations from normality in their distributions. Log transformations did not normalize the variables. Consequently, instead of bivariate Pearson’s correlations the non-parametric Spearman’s correlations were run. Table 4.3 presents these correlations.

Table 4.3

Nouns				
	Frequency	Syllable Count	Letter Count	Transparency
AoA Rating	-.52**	.19**	.20**	-.13*
Frequency		-.25**	-.24**	.06
Syllable			.64**	-.22**
Letter				.16*
Verbs				
	Frequency	Syllable Count	Letter Count	Transparency
AoA Rating	-.19**	.31**	.30**	.01
Frequency		-.28**	-.41**	.23**
Syllable			.65**	-.08
Letter				.02
Adjectives				
	Frequency	Syllable Count	Letter Count	Transparency
AoA Rating	-.50**	.24*	.39**	.00
Frequency		-.35**	-.37**	.12
Syllable			.66**	.08
Letter				.20
Adverbs				
	Frequency	Syllable Count	Letter Count	Transparency
AoA Rating	-.26	.32*	.36*	.19
Frequency		-.24	-.40*	-.29
Syllable			.79**	.19
Letter				.52**
Conjunctions				
	Frequency	Syllable Count	Letter Count	Transparency
AoA Rating	-.64	-.14	-.62	-.29
Frequency		-.66	-.15	.54
Syllable			.71	-.47
Letter				0.07
Postpositions				
	Frequency	Syllable Count ^a	Letter Count	Transparency ^a
AoA Rating	-.17	-	-.66	-
Frequency		-	-.13	-
Syllable			-	-
Letter				-

Spearman's Rank-Order Correlations for AoA Ratings and Psycholinguistic Variables

Note: ^a: correlations could not be computed because these variable were constant.

*: correlation is significant at .05 level (2-tailed). **: correlation is significant at the .01 level (2-tailed).

Results from Spearman's rank-order correlations show that nouns, verbs, and adjectives are significantly negatively associated with written frequency. However, the strength of these correlations is very weak for verbs ($r_s(196) = -.19, p < .001$) and moderate for nouns ($r_s(260) = -$

.52, $p < .05$) and adjectives and ($r_s(76) = -.50, p < .05$) (Cohen, 1988). According to raters, more frequently occurring nouns, verbs, and adjectives are acquired earlier than less frequent words in those word classes. The remaining word classes are not significantly correlated with written frequency.

Additionally, the AoA ratings for all content words (i.e. nouns, verbs, adjectives, and adverbs) are moderately positively associated with both letter and syllable counts, which means the AoA ratings increase as the number of syllables and letters increases. However, this association is not significant for both types of function words. Finally, only nouns were significantly associated with orthographic transparency ($r_s(260) = -.13, p < .05$). The negative direction of this correlation indicates that participants think that more vowelized nouns have an earlier age of acquisition than less vowelized nouns. However, the strength of this correlation is very weak with the rank orders of AoA ratings for nouns and the transparency rating sharing very little variance ($R^2 = .017$).

4.2. Textbook Analysis

This section provides an overview of the language and literacy learning analyzed in the textbooks before more specifically addressing the third research question, i.e., how do activities in early grade Urdu textbooks contribute to the development of children's morphological and orthographic awareness?

4.2.1. Language and Literacy Learning

This subsection presents the breakdown of activities coded under each component skill in the textbook analysis template. Component skills grouped under the domain of language learning ($N = 8$) are assessed before component skills grouped under the domain of literacy ($N = 7$). This breakdown of activities is calculated as the number of activities coded in each component skill as a percentage of all activities in a textbook (i.e. number of skill-specific activities/total activities x 100).

Table 4.4 presents the percentages of activities (with the raw number of activities in parenthesis) in the language learning domain by grade and publisher. Due to double-coding, the sum of activities in the component skills' breakdown does not match the total number of analyzed activities. The table shows how several component skills are not captured by the textbooks' activities. For example, none of the analyzed activities capture "Listening Comprehension," "Conversational Oral Comprehension," and "Syntax." Similarly, both

“Narrative Expression” and “Other Literary Devices” are coded under only two activities each across the range of textbooks analyzed. While the percentage of activities may vary by grade and publisher, the component skills of “Vocabulary,” “Morphological Awareness,” and “Oral Comprehension (repetition and recitation)” are represented in the analyzed activities.

Table 4.4

	First Grade			Third Grade		
	Publisher A (N = 70)	Publisher B (N = 281)	Publisher C (N = 123)	Publisher A (N = 25)	Publisher B (N = 73)	Publisher C (N = 53)
Language Domains						
Vocabulary	30% (21)	3.2% (9)	70.7% (87)	36% (21)	28.8% (21)	49% (26)
Morphological Awareness	0% (0)	4.6% (13)	0% (0)	4% (1)	6.8% (5)	9.4% (5)
Oral Language (Recitation)	42.9% (30)	3.2% (9)	4.1% (5)	8% (2)	4.1% (2)	15.1% (8)
Listening Comprehension	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Oral Language (Conversational)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Narrative Expression	0% (0)	0% (0)	0% (0)	4% (1)	4.1% (1)	0% (0)
Syntax	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Other Literary Devices	0% (0)	0% (0)	0% (0)	4% (1)	0% (0)	1.9% (1)

Percentages and Number of Activities in the Language Domain by Grade and Publisher

Table 4.5. presents the percentages of analyzed activities (with the raw number of activities in the parenthesis) under each of the seven component skills grouped together under a general literacy domain, by grade and publisher.

Table 4.5

	First Grade			Third Grade		
	Publisher A (N = 70)	Publisher B (N = 281)	Publisher C (N = 123)	Publisher A (N = 25)	Publisher B (N = 73)	Publisher C (N = 53)
Literacy Domains						
Huroof Knowledge ^a	0% (0)	26.7% (75)	76.4% (94)	0% (0)	0% (0)	3.8% (2)
Phonological Awareness ^b	0% (0)	22.4% (73)	21.9% (25)	0% (0)	0% (0)	0% (0)
Word List Decoding ^c	18.6% (13)	3.9% (11)	37.4% (56)	4% (1)	1.4% (1)	1.9% (1)
Reading Comprehension: Sentences	65% (15)	1.4% (4)	.8% (1)	4% (1)	6.8% (5)	9.4% (5)
Reading Comprehension: Passages	35.7% (25)	3.9% (11)	0% (0)	52% (13)	9.6% (7)	20.8% (11)
Copy Writing	0% (0)	69.6% (196)	19.5% (24)	4% (1)	4.1% (3)	9.4% (5)
Writing ^d	25.7% (18)	14.2% (50)	66.7% (82)	76% (19)	32.9% (24)	64.2% (34)

Percentages and Number of Activities in the Literacy Domain by Grade and Publisher

Note: ^a: The first-grade textbook for Publisher A starts with passage comprehension and word-level writing activities. Unlike the other publishers that prioritize grapheme teaching in the first grade, students entering first grade in schools using this textbook are likely expected to have already mastered reading and writing activities. ^b: This component skill is rarely captured directly and activities under other skills (such as “Huroof Knowledge”) get double coded for Phonological Awareness ^c: The original adaptation of the textbook analysis template contained activities under word decoding that extended further than reading word lists as can be seen in Appendix H. However, none of the analyzed activities could be coded under these subsections of “Word Decoding.” To accurately represent the component skill being captured, “Word Decoding” was changed to “Word List Decoding.” ^d: “Writing” in the template intends to capture several literacy features, including narrative expression. The activities that were coded for writing primarily consisted of writing individual words or sentences.

The percentages of activities in each literacy domain also reflect differences between grades and publishers. For example, Publisher A does not prioritize “Huroof Knowledge” and “Phonological Awareness” in the first grade. Similarly, the percentage of activities coded under “Reading Comprehension: Passages” increases between all three publishers from first to third-grade. There is a simultaneous decline in the percentage of “Word List Decoding” activities. Lastly, the literacy component skill of “Writing” appears to be emphasized by each publisher in both grades.

As Tables 4.4 and 4.5 depict, the percentages of activities classified under the language and literacy domains vary by grade and publisher. Since the third research question focuses on morphological and orthographic knowledge development specifically, the rest of the analyses exclusively addresses these domains.

4.2.2. Analysis of Morphological and Orthographic Domains

Component skills that captured morphological and orthographic awareness were selected. “Huroof Knowledge” and “Word List Decoding” were selected under the orthographic domain. “Vocabulary” and “Morphological Awareness” were selected under the morphological domain. The unit of observation was the number of activities coded in each component skill in each textbook. To avoid any biases in the distribution arising from a smaller sampling of third-grade textbooks, raw scores were converted to proportion scores by dividing the number of activities coded in each component skill by the total number of activities analyzed in the textbook. To analyze the data, separate two-way ANOVAs were planned for a 2 (grade-level) x 2 (component

skill) analysis for each of the morphological and orthographic domains. Publisher type was also included in the analysis. However, the small sample size of the textbooks ($N = 6$) did not allow for this 2 (grade-level) x 2 (domain type) x 3 (publisher) comparison. Therefore, two separate one-way ANOVAs were planned for each of the domains selected with publisher type as the independent variable.

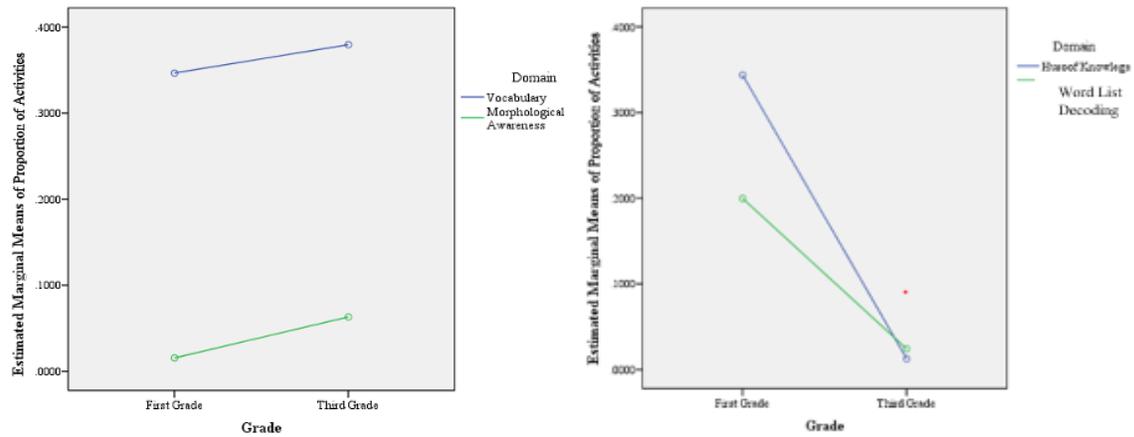
4.2.2.1. Grade-Wise Analysis.

4.2.2.1.1 Assumptions for ANOVAs. According to Field (2009), the two main assumptions for parametric tests are that the dependent variable needs to be normally distributed and there should be homogeneity of variance. Before running the ANOVAs in the present study, the DV, i.e. the number of activities in each skill was tested for normality. Normality was assessed using the Shapiro-Wilk test because of the small sample size ($N = 6$). Results from the Shapiro-Wilk test (see Appendix L) indicated that the two dependent variables, i.e. the morphological and orthographic domains were normally distributed ($ps > 0.01$). Therefore, Field's (2009) stated assumption of normality can be reliably assumed to have been met in the analysis.

The second assumption, i.e. homogeneity of variance was tested using Levene's test separately for each of the two ANOVAs. These tests indicated that neither the variances for "Huroof Knowledge" and "Word List Decoding" nor the variances for "Vocabulary" and "Morphological Awareness" differ significantly from each other (see Appendix M). Since both main assumptions of parametric tests were met, the analysis strategy was pursued as planned.

4.2.2.2.2. Main Analyses. Comparisons were made between first and third-grade to assess the textbooks' contributions at the beginning and end of the primary school period. Before running the analyses, double-coding was removed. The code was retained for the component skill the activity had been directly coded under and removed from the component skill it had been secondarily coded under. Figure 4.1 portrays the estimated marginal means difference of activity proportions between skills and grade levels.

Figure 4.1



Estimated Marginal Means Difference of Proportion of Activities between Skills in the Morphological and Orthographic Domains by Grade

A visual inspection of Figure 4.1. suggests that the proportion of morphological activities increases between first and third-grade, while the proportion of orthographic activities declines in the same period. The proportion of “Vocabulary” activities is greater in the textbooks than the proportion of “Morphological Awareness” activities. Conversely, while the proportion of “Huroof Knowledge” activities in each textbook is initially greater than that of “Word List Decoding” activities, by grade three, this difference is not observed. However, to test whether these differences within categories and between grades were significant, ANOVAs were run. Table 4.6. shows the results from the morphological analysis.

Table 4.6

	Descriptive Statistics				Effects					
	First Grade		Third Grade		Skill		Grade		Skill x Grade	
	M	SD	M	SD	F	p	F	p	F	p
Vocabulary	.35	.34	.37	.12	9.57	.01**	.09	.77	.006	.94
Morphological Awareness	.02	.03	.05	.03						

Means, Standard Deviations, Main and Interaction Effects of Activity Proportions by Component Skill and Grade in the Morphological Domain

*: effect is significant at .05 level **: effect is significant at the .01 level

As shown in Table 4.6, a large main effect ($\eta^2 > .50$; Field, 2009) of component skill was found ($F_{[1,4]} = 9.57$; $p < .05$, $\eta^2 = .55$), but no main effect of grade was found ($F_{[1,4]} = .09$; $p = .77$, $\eta^2 = .01$). Furthermore, the interaction between grade and skill was non-significant ($F_{[1,4]} = .00$; $p = .95$, $\eta^2 = .001$). A higher proportion of activities were coded under vocabulary than the proportion of activities coded under morphology. However, the proportion of activities coded under vocabulary and morphology did not significantly differ between the first and third grades.

Table 4.7. presents the results from the ANOVA for the orthographic domain.

Table 4.7

	Descriptive Statistics				Effects					
	First Grade		Third Grade		Skill		Grade		Skill x Grade	
	M	SD	M	SD	F	p	F	p	F	p
Huroof Knowledge	.34	.39	.01	.02	5.50	.04*	4.17	.07	.47	.51
Word List Decoding	.19	.16	.02	.01						

Means, Standard Deviations, Main, and Interaction Effects of Activity Proportions by Component Skill and Grade in the Orthographic Domain

*: effect is significant at .05 level

As Table 4.7 suggests, there was a main effect of component skill ($F_{[1,4]} = 5.50$; $p < .05$, $\eta^2 = .41$). The ANOVA did not find a main effect of grade ($F_{[1,4]} = 4.17$; $p = .07$, $\eta^2 = .34$) or an interaction effect between grade and skill ($F_{[1,4]} = .47$; $p = .51$, $\eta^2 = .06$). Results from the orthographic analysis mirrored those of the morphological analysis in that only a main effect of component skill was found. The lack of an interaction effect for both domains differs from the visual inspection of Figure 4.1.

4.2.2.2. ANOVAs for Different Publishers.

The third research question of the study aims to assess the contributions made by textbooks to the development of children's morphological and orthographic knowledge, which could differ based on publishers' priorities and principles. Section 4.2.1. commented on differences in publishers' assumptions about children's language competencies upon entry into

primary school. Therefore, separate ANOVAs for the morphological and orthographic domain by publisher were run. Table 4. presents the results from the analyses.

Table 4.8

Domain	Publisher						Effects	
	A		B		C		Domain	
	M	SD	M	SD	M	SD	F	p
Morphology	.17	.18	.11	.12	.32	.33	.937	.427
Orthography	.06	.09	.08	.13	.30	.35	1.462	.282

Means, Standard Deviations, and Effects of Domains by Publishing Scheme

As Table 4.8 presents, the type of publisher did not have a significant effect on the proportion of activities coded either under the morphological domain ($F_{[1,4]} = .937; p = .427, \eta^2 = .172$) or the orthographic domain ($F_{[2,9]} = 1.46; p = .282, \eta^2 = .245$). While comparing the percentages of activities in each component skill across publishers suggested differences in the number of activities in the component skills among publishers (for e.g. no “Huroof Knowledge” activities in Publisher A’s textbooks), these differences do not appear to be statistically significant and the proportion of activities within the larger domains of morphology and orthography do not differ from each other.

4.2.3. Qualitative Textbook Analysis

Knowledge of the types of activities coded under each component skill can deepen our understanding of the processes through which textbooks contribute to children’s morphological and orthographic development. Therefore, this section provides a narrative overview of the textbooks’ analyzed activities, particularly in terms of morphology and orthography.

4.2.3.1. Morphology.

Eight activities aiming to capture the component skill of “Morphological Awareness” were included in the template, but textbooks did not cover this range of activities. For example, none of the activities were coded under the subskill of derivational morphology. Inflectional morphology activities were coded in just one publisher’s textbooks. Similarly, “*Harf* Knowledge” that codes for understanding and use of function words was only captured by two activities across the range of grades and publishers. In both instances, “*Harf* Knowledge” activities were fill-in-the-blanks activities in which students had to select the correct root morpheme (for e.g. chapter 20 of Publisher C’s third-grade textbook contained the activity صحیح

استعمال کیجیے خالی جگہوں میں کا/کی/کے/کو کا (use /ka:/, /ki:/, /ke:/, /ko:/ correctly in the empty spaces). Students have to read sentences and write the correct inflectional form of the function word. Inflections are marked for gender, grammatical number, and tense. Additionally, while the template included the three types of plural forms in Urdu (i.e. jama salim, jama mukassar, and broken plurals), activities in the textbooks exclusively focused on the jama salim (i.e. unaltered plural forms). Lastly, only one textbook included two activities that directly captured gender marking with each activity requiring students to write the names of relations (for e.g. خالہ، ماموں، نانا (maternal aunt, maternal uncle, maternal grandfather) were provided and students had to write the masculine/feminine form of the relation). Interestingly, all but one morphological activities were double-coded for writing.

In terms of “Vocabulary,” ten activity types (ranging from matching pictures and words to using a dictionary) were included in the component skill, but the most frequently occurring activity type differed by publisher. Publisher B preferred activities involving word pairs or words under the same category (such as an activity in chapter six of the third-grade textbook in which students have to find and write four nouns from the preceding lesson). Publisher C’s most frequently occurring “Vocabulary” activity was “reading words with meanings.” After every lesson, difficult words were presented with their meanings. Interestingly, some word-meaning activities at the third-grade level also involved explicit instruction about word classes. For example, the instructions for activity six in chapter ten just asked students to "پڑھیے اور سمجھیے" (read and understand) before providing definitions and examples of common and proper nouns from the expository lesson presented before. Interestingly, Publisher A’s vocabulary activities were less explicit and most of the analyzed activities are double-coded from other component skills. For example, the first-grade textbook contains five explicit and 16 double-coded activities. Double coded activities include six activities from the component skill of writing that ask students to write certain words or names (such as the names of wild animals in chapter four).

The emphasis on writing across all three publishers is perhaps apparent in the way many activities that promote vocabulary development and morphology contain a component of writing words or completing sentences.

4.2.3.2. Orthography.

The component skills considered in this subsection are “Huroof Knowledge” and “Word List Decoding.” Eight activity types were included in the template for “Huroof Knowledge” to

capture different features of Urdu orthography (from grapheme and diacritic recognition to recognition of letter shapes i.e. the changing grapheme form in initial, medial, and final word positions). Ten activities between Publisher B and Publisher C's first-grade textbooks included activities for diacritic recognition and use. For example, in chapter one in Publisher B's textbook, students have to encircle the diacritic that produces the long vowel \bar{a} (/a:/), which is written with a diacritic on the grapheme \bar{a} (/a/). Majority of the activities coded under "Huroof Knowledge" (26 activities in Publisher C and three activities Publisher B) involve single-harf identification. Many of the activities under "Huroof Knowledge" are double-coded with writing as several activities (more than 50 in each textbook) focus on students practicing writing either a single letter or CV pairs. For example, 24 activities in Publisher C's first-grade textbook either ask students to write huroof or to separate simple (and mostly monosyllabic) words into the different huroof. Publisher B's first-grade textbook's double-coded activities mostly consist of copy-writing single huroof or CV pairs (N = 67). Therefore, while "Huroof Knowledge" was only coded in two textbooks, both textbooks focus on teaching students how to write huroof.

"Word List Decoding" was accessed through reading simple word lists (consisting of words with long vowel sounds) or reading lists with mostly vowelized (and therefore transparent words). While the template including several sections for *murakkab lafz* (different combinations of compound words in Urdu), homonyms, words with multiple correct spellings, and non-transparent words, these aspects were rarely captured by activities. Only one activity across the six textbooks (in Publisher A's third-grade textbook) explicitly taught the use of *murakkab lafz* by instructing students to connect two stems by adding the vowel 'و' (/o:/). Interestingly, the instructions for this activity included a definition of *murakkab lafz* and was therefore double-coded for vocabulary. However, this was the only activity (out of 86 total activities across grades and publishers) in "Word List Decoding" that did not involve reading word lists. "Word List Decoding" may therefore be one component skill where publishers have similar expectations of students' competency at entering first grade and learning outcomes as the selected activities appear to be similar.

Chapter 5: Discussion

5.1. Chapter Overview

In this chapter, the main findings from both the AoA study and textbook analysis will be summarized and discussed along with their educational implications. A section on limitations and future directions will follow before the conclusions are detailed.

5.2. AoA study

The AoA study aimed to answer the following research questions:

1. What are the age of acquisition ratings for different word classes in a word-list derived from child-directed written input?
2. What psycholinguistic associations do these ratings have for each word class?

5.2.1. Summary of Findings

On average, participants' ratings on a morphologically representative word list derived from first to third-grade textbooks suggest children acquire words in all word classes between the ages of three and five. Correlations between AoA ratings and frequency are in the predicted direction for all word classes as more frequent words were rated as being acquired early. Content words are moderately associated with letter and syllable counts, which complements findings from earlier norming studies (for e.g., Bird et al., 2001; Cameirão & Vicente, 2010). However, the correlations for function words were not significant. Only nouns are significantly associated with orthographic transparency, but this association is weak.

5.2.2. Discussion of Findings

The following subsections answer research questions one and two in detail by first discussing two word classes with nuanced findings, before exploring morphological sources of disagreement for discrepant words, and finally commenting on Zevin and Seidenberg's arbitrary mapping hypothesis (2002).

5.2.2.1. Conjunctions.

Inter-rater reliability was of high quality for the complete word list and for individual word classes except conjunctions. The late acquisition of conjunctions is especially interesting considering its high written frequency ($M = 69.8$, $SD = 56.4$) as it may be expected that conjunctions are understood early because they are more frequent. Participants' higher mean ratings might reflect potential differences in spoken frequency, particularly in child-directed

speech as adults might not use conjunctions very often when they interact with children. Conjunctions hold phrases and sentences together and may be more frequent in narrative (decontextualized) language, which in turn is associated with children's vocabulary (Rowe, 2012). Therefore, if caregivers and adults are not using conjunctions around children, or using them at a lower frequency, children will, in turn, understand conjunctions later. The ICC model was set to consistency, which means that lower inter-rater reliability for conjunctions reflects more variability in rank-ordering between participants compared to variability in other word classes. In other words, participants did not agree on which conjunctions were acquired earlier than others. This lack of agreement in the rate of acquisition of words within conjunctions is perhaps indicative of the ambiguity of this word class in terms of children's vocabulary development as participants' interactions brought them to different conclusions about individual word learning.

5.2.2.2. Nouns.

The present study did not reveal a noun advantage in order of word acquisition despite nouns having the highest frequency in child-directed print. This finding does not align with cross-linguistic evidence of earlier noun acquisition (Gentner, 1982). In fact, teachers thought nouns were acquired later than all other word classes except conjunctions and had only low to moderate associations with the psycholinguistic variables measured (i.e. frequency, letter and syllable counts, and orthographic transparency). Perhaps the inclusion of both concrete (such as سڑک /sɾk/ which means road) and abstract nouns (such as ہمت /hmt/ which means strength) contributed to relatively late acquisition. According to McDonough (2011), the noun advantage in early word acquisition is attributable to the imageability of concrete nouns. However, in a morphologically similar language, i.e. Arabic, Khwaileh et al. (2018) found no significant differences between AoA ratings for concrete nouns and verbs. While the present study does not directly measure children's actual vocabulary, it begins to question the universality of an earlier noun acquisition (based on both frequency of occurrence as well as imageability) in morphologically similar languages like Arabic and Urdu, but more research is needed before any conclusions can be drawn.

5.2.2.3. Discrepant Words.

An analysis of discrepant words might illuminate potential challenges in children's Urdu vocabulary development assessed through knowledgeable raters' disagreements surrounding

those words. Apart from one adverb, discrepant words were either nouns (69%) or verbs (26.9%), the two word classes that lend themselves the most to inflections in Urdu. Several discrepant words were inflected for either grammatical number or gender. This lack of consensus on when inflected words are understood by children is not fully explored in the present study as the word list was developed to represent the breadth and depth of the vocabulary children are exposed to in child-directed print. However, it may still indicate the role of inflectional morphology in Urdu vocabulary acquisition. For example, the lemma چیز (/tʃiːz/, thing) and two inflected plural forms چیزیں and چیزوں (/tʃiːzɛː/ and /tʃiːzɔː/ both meaning things) were included in the word list. The broken plurals are used in different sentential contexts and correct usage depends on harf knowledge (Fatima et al., 2018). While participants agreed on ratings for the lemma, both inflected forms were discrepant and were rated as being acquired later than the lemma.

5.2.2.4. The Arbitrary Mapping Hypothesis.

Zevin and Seidenberg's (2002) arbitrary mapping hypothesis is directly related to the AoA effect (i.e. the tendency of certain words to be learned earlier and retained better), as they hypothesize that words with predictive orthographic mappings will be retained better than words with less predictive orthographic mappings. The present study did not capture the AoA effect as our interest lay in the first instance of learning of the spoken form, i.e. before children acquire literacy. Still, the psycholinguistic variable of orthographic transparency (which was operationalized in terms of the orthographic representation of vowels) was added in the analysis because the word list was derived from written and not spoken input. Only nouns were significantly associated with orthographic transparency such that the AoA ratings decreased as words become more vowelized and transparent, and even this correlation was weak. Since the earliest learned words by children are not related to the transparency of their written forms, it would then be interesting to see whether the arbitrary mapping hypothesis holds true in potential observations of AoA effect in languages like Urdu.

5.3. Textbook Analysis

Textbooks were analyzed to answer the third research question, i.e., how do activities in early grade Urdu textbooks contribute to the development of children's morphological and orthographic awareness?

5.3.1. Summary of Findings

Analyzed activities in the textbooks did not capture some component skills, such as listening comprehension. Quantitative analysis of skills grouped under the domains of morphology and orthography revealed main effects of component skills. Textbooks included a higher proportion of “Vocabulary” rather than “Morphological Awareness” activities, and a higher proportion of “Huroof Knowledge” activities compared to “Word List Decoding” activities. Interaction effects of grade and component skill, and main effects of publisher type were not found. Qualitative analysis highlighted distinctions within the component skills of interest such that publishers emphasized different types of activities within all component skills of interest besides “Word List Decoding.” These findings are discussed in more detail below through the lens of the language learning process (Cunningsworth, 1984) imolcantly advanced by Urdu textbooks.

5.3.2. Discussion of Main Findings

5.3.2.1. Development of Morphological Awareness.

“Vocabulary” activities more generally promote vocabulary development across languages, while “Morphological Awareness” activities in the template aim to specifically promote Urdu vocabulary development. A significantly higher proportion of “Vocabulary” activities is indicative of early-grade Urdu textbooks’ emphasis on a more generic language learning process where the relationship between the learner and specific language may not be the most important part of word-learning. However, while the textbook analysis template did not include “Vocabulary” activities specific to Urdu morphology, the qualitative analysis points to textbook developers’ incorporation of elements of Urdu morphology in “Vocabulary” activities. For instance, some of the word meanings in Publisher C’s third-grade textbook contained root morphemes as the new word to be learned and an inflectional variant as the meaning. For example, in chapter ten, the root word ‘سرا’ (/sa:ra:/) is presented as the word to be learned and the more commonly used word ‘سرائے’ (/sa:ra:e:/) as its meaning. Both words mean resting place and can be used interchangeably but ‘سرائے’ is inflected for the plural form. These specific features of Urdu morphology are not ignored by the textbook developers even if they are not explicitly targeted in the textbooks.

The activities coded directly under “Morphological Awareness” target very few skills included in the textbook analysis template. Across the textbooks, there is an emphasis on

learning the jama salim plural form as opposed to the more morphologically complex derivational altered forms. Nonetheless, it is important to note that unaltered plurals still encompass the nuance specific to Urdu's inflectional morphology in terms of marking both gender and number (Daud et al., 2017). To illustrate, in an activity in chapter one of Publisher B's third-grade textbook, pupils are instructed to write the plural forms of words. Two such words are باغ (/ba:ɡʰ/; garden) and تتلی (/t̪t̪li:/; butterfly). The plural forms of either word is created by using distinct suffixes (-ات (/a:t̪/) for former and -اں (/ã:/ for the latter). However, in making this selection, gender marking also needs to be considered as باغ is masculine whereas تتلی is feminine rendering only the specific suffixes mentioned for each word suitable. Therefore, the textbooks, perhaps less directly, retain unique features of Urdu morphology even if these features are not expansive.

5.3.2.2. Development of Orthographic Awareness.

The second part of this research question concerns textbooks' contribution to children's orthographic awareness. Although "Huroof Knowledge" was only deemed a relevant component skill in the first-grade by publishers, the proportion of activities coded was significantly higher than those coded for "Word List Decoding." However, activities coded under "Huroof Knowledge" did not address several features specific to Urdu's deep and mostly un-vowelized orthography. The proportion of activities in the orthographic domain is very low for both component skills in the third grade. Therefore, orthography, operationalized at the letter and word level, may only be relevant in the first-grade as other component skills (such as "Reading Comprehension") represent a higher proportion of literacy activities in the third grade.

Both types of component skills are also similar in their limited range of activities. Explicit instruction on Urdu's visually complex orthography is not given as none of the activities target changing grapheme shapes based on word positions. Students are only taught to recognize and write the grapheme in the word-initial position. Less explicitly however, examples of words provided in subsequent word lists (coded under "Word List Decoding") often include the grapheme in multiple positions. For instance, chapter 16 of Publisher C's textbook includes words in which the letter 'ل' (/l/) appears in all three positions. However, preceding activities do not seem to prepare students to have knowledge of changing grapheme positions.

Word lists in Publisher C's first-grade textbook contained words that were often marked with diacritics, but this was not true for either Publisher A or Publisher B. Publisher C included eight

activities for the explicit instruction of diacritics while Publisher B included two and Publisher C did not include any. It can be argued that perhaps Publisher C's textbooks align with an instructional style suggested for Arabic literacy instruction (Saiegh-Haddad, 2018) whereby diacritics are initially taught in their entirety to ease students' phoneme to grapheme mappings (through letters and phonemic diacritics) and then removed in older grades. Here, the language learning process appears to be gradual. However, Publishers A and B might have a different approach where the Urdu script is taught in its un-vowelized form from the earliest levels of instruction, and the relationship between the learner and written form of the language is meant to be consistent across age groups as activities targeting orthography do not assume a stage-wise learning process of the writing system.

5.4. Implications for Education

Results from the AoA study suggest that on average, words in all types of word classes appearing in children's textbooks through connected text (i.e. poems, stories, and essays) may be acquired before children enter primary school according to informed raters. Findings from the textbook analysis show that activities in children's early-grade Urdu textbooks may be less specific to Urdu morphology and orthography and may promote vocabulary development more generically without exclusive focus on features (such as derivational morphology) specific to Urdu. Considering the importance of textbooks (Tomilson, 2012), these findings have important implications for education. The following subsections discuss these implications through the lenses of vocabulary growth, and the link between vocabulary and literacy instruction.

5.4.1. Vocabulary Growth

Vocabulary learning has been described as a continuous process of encountering unfamiliar words in contexts that allow an understanding of the words' meaning (Harmon et al., 2009). There is evidence from research in English that the root word vocabulary increases from around 5,200 root words in the second grade to approximately 8,400 words in the fifth grade (Biemiller & Slonim, 2001), implying the importance of the early school years for vocabulary growth. For this reason, vocabulary learning is deemed an important instructional aim for teachers (Harmon et al., 2009). Considering the importance of vocabulary learning from the points of view of both learning and teaching, the early rated acquisition of spoken words in children's textbooks across word classes may be a cause of concern for children's vocabulary growth.

In the context of Urdu vocabulary development, language textbooks, central to the course curriculum (Tomlinson, 2012), may not be conducive to vocabulary growth beyond the vocabulary presented in connected text. In terms of textbook activities, only one textbook included intentional word-learning activities by providing definitions of difficult words in connected text. While textbook developers' may believe these activities help build children's vocabularies, there is evidence against such strategies (Stahl & Fairbanks, 1986). Activities that have been shown to promote word-learning, such as teaching morphological awareness as a cognitive strategy when students encounter a new word (Goodwin & Ahn, 2012; Prince, 2009) or an emphasis on derivational morphology (Saeigh-Haddad, 2018) in derivationally rich languages (like Arabic and Urdu) are not included in any of the textbooks. In both first and third-grades, there is very little emphasis on activities related to any aspect of Urdu morphology.

In fact, the findings from the AoA study and textbook analysis jointly signify limited instruction for morphological awareness, particularly in the context of function words. Harf knowledge activities that encourage students to understand and use function words (which are mostly mono-morphemic in Urdu) are not represented across publishers or grades. Textbook activities that were coded under harf knowledge only targeted postpositions. However, learning conjunctions is also important vocabulary growth as conjunctions convey semantic relationships between content words (Halliday & Hanson, 1976) and can therefore provide cues that elucidate the meanings of unfamiliar words. In the AoA study, conjunctions were considered to be acquired later than other word classes. The lower inter-rater reliability for conjunctions also signifies uncertainty among participants regarding when conjunctions are learned. Putting results from the AoA study and the textbook analysis together, therefore, points to a combination of differential exposure to function words in the language input children receive before entering school (as participants agreed the least on when children understand words in this word class) and fewer explicit learning opportunities in an instructional context when they start learning Urdu through school.

Together, the two findings (i.e. children's prior average familiarity with words in the different genres included in textbooks, and some ineffective strategies to directly promote vocabulary), are concerning. During the Zoom sessions, many participants commented on the dilution of Urdu with schools and families preferring communicating in English. Additionally, one teacher noted the limited amount of time children have to learn Urdu as the medium of

instruction is English for all subjects besides Urdu across private schools in Pakistan as early as kindergarten. In many schools, children are also penalized by teachers and administrators for speaking in Urdu outside of the designated Urdu period. With limited exposure, incidental vocabulary learning through reading or listening, often considered one of the main sources of vocabulary growth in children (Nagy et al., 1985), may not be as central for first-language Urdu learners as it is for learners of other first languages.

Vocabulary learning is related to several aspects of language and literacy, making the limited opportunity for children to encounter new words and develop effective strategies to decode the meanings of new words when they do encounter them impact other component skills. For instance, Oullette (2006) found that by the fourth-grade, vocabulary measures (such as receptive or oral vocabulary knowledge) predicted performance on word decoding, word recognition, and reading comprehension tasks. If students are already familiar with many of the words they read in school and are not taught strategies tailored to unique features of Urdu morphology that can help them analyze and understand new words, performance on the component skills mentioned by Oullette (2006) may also be affected. However, as the next section illustrates, there may also be some advantages in the vocabulary exposure and instruction currently available.

5.4.2. Vocabulary and Literacy Instruction

The fact that, on average, children hear and read the same words in their textbooks which they already know may not directly benefit their vocabulary development. However, this pre-existing receptive vocabulary knowledge can help them in other literacy domains. Vocabulary knowledge plays a facilitative role in word decoding and reading comprehension (Garlock et al., 2001). Therefore, if subjective ratings are taken as representative estimates of children's vocabulary knowledge, then this knowledge may help them in learning to read and write when they enter primary school.

The most consistently emphasized component skills in the textbooks are "Writing" and "Reading Comprehension" with each publisher including multiple activities targeting both skills in their first-grade and third-grade textbooks. If the aim of Urdu instruction is centered around these component skills, then children's prior vocabulary knowledge may help them acquire these skills. In a study on Dutch elementary school students (between the ages of six and eight years), Verhoeven et al. (2011) found that early vocabulary knowledge predicted word decoding as early

as the second-grade. They also found a reciprocal relationship between children's vocabulary and reading comprehension, such that first-grade vocabulary predicted second-grade reading comprehension which subsequently predicted third-grade vocabulary. More research is needed to study the relationships between these skills in Urdu, but the strong link between vocabulary and literacy skills may promote vocabulary growth even if the written input and activities do not do so directly.

The specific association between Urdu morphology and orthography is also important. Saeigh-Haddad's interventions in Arabic (e.g, 2018) include both morphological and orthographic features because Arabic orthography is deep and often un-vowelized. Perhaps Urdu textbook developers' focus on literacy activities, particularly writing and comprehension (at the sentence and passage level), is based on a similar conceptualization of the language learning process in Urdu, which may rely on both morphology and orthography. Appendix A details some of the challenges faced by learners of Urdu orthography. Perhaps already having learned the receptive vocabulary associated with the words children have to read and write removes the additional burden of simultaneously processing semantic information. The limited range of activities in both the morphological and orthographic domain may also be more reflective of publishers' preference for other literacy activities that may then improve children's vocabulary. The overlap between Arabic and Urdu orthography is also important in this respect. Most Muslim children in Pakistan, a predominantly Muslim country, learn Arabic alphabets and word-decoding in order to read the Quran before primary school. The limited range of activities in the orthographic domain may not be deemed as relevant as other literacy activities because many students would already have mastered letter and word recognition in a very similar script.

5.5. Limitations

5.5.1. Sampling

5.5.1.1. Corpus.

Practical considerations restricted the size of the corpus to 2,168 word types in the sentence corpus and 1,968 words in the word corpus from which the word list was extracted. The corpus had to be digitized and then manually annotated. To ensure that methodological rigor was not compromised, quality was preferred over quantity. The corpus was limited both in the number of textbooks (N =9), and the data taken from each textbook (i.e. every fourth line from connected text). Frequency calculations from corpora are usually based on very large corpus

sizes, but, the calculations for the AoA study were bound by the small corpus size. However, this corpus was developed specifically for the present study because of an absence of an existing child-directed print corpus and was not created with the intention of representing all child-directed print. It is best to be cognizant of the corpus' size and add it to a much larger database of child-directed written materials when one is created for Urdu.

5.5.1.2. Participants.

Indefinite school closure due to a public health crisis severely limited the number and types of participants that could be recruited. Only eleven participants were recruited in the study, primarily through the researcher's personal contacts. Most of the participants taught at the same after-school Quran teaching program. The small sample size and sampling technique collectively make the sample less representative of adult responders or even other teachers. While all participants frequently engaged with children and were therefore more knowledgeable than most adult raters, data from other types of teachers (especially school teachers) and parents is needed to be able to generalize the AoA ratings to teachers and parents who are first-language users of Urdu.

5.5.1.3. Textbooks.

While three different publishers were included in the textbook analysis from different contexts such as international, local, and government level publishers, only the first and third-grade textbooks were selected due to time limitations. This meant progression through primary school could not be documented. Additionally, the template focuses purely on activities. Other features of the textbook (such as physical layout, and socio-cultural implications of texts) were not included. Therefore, the findings reflect the appropriateness of the textbooks only with respect to language and learning activities, and not more generally.

5.5.1.4. Online Data Collection.

Due to health and safety reasons, data had to be collected online. Many participants were unfamiliar with Zoom and struggled with the online format of data collection. While these issues could not be avoided, this form of data collection prevented potential participants from consenting to participate. It also meant some participants were more familiar with Zoom than others. Although inter-rater agreement was good, this variability in familiarity may have contributed to some differences in understanding instructions, and level of comfort in asking

questions, which may have then contributed to ratings that did not accurately represent participants' beliefs about the AoA of some words.

5.5.2. Reliability of Textbook Analysis

Vagh & Nag (2014) encourage at least two coders to analyze textbooks. However, all textbooks were coded by the researcher as a second coder with some understanding of Urdu language and literacy and access to the textbooks was not found. Once such a coder is identified, inter-rater reliability can be calculated to draw more objective conclusions than is currently possible.

5.6. Directions for Future Research

5.6.1. Observational and Experimental Research

As mentioned earlier, initially, a language intervention was planned to assess the effects of explicit morphology instruction on vocabulary and reading comprehension. The present study's joint focus on AoA of words used in children's textbooks and analysis of the textbooks themselves lays the foundation for future observational and experimental work of the type initially planned. Objective AoA measures as well as other vocabulary assessments can be used to explore any differences between knowledgeable raters' beliefs about child language acquisition and actual child language acquisition. Such studies may be particularly useful in exploring acquisition of discrepant words. Discrepancy in ratings of inflected words may then arise from a gap in some raters' knowledge about children's vocabulary depth (i.e. how well words are understood) as opposed to semantic processing difficulty in children. As textbooks are only one element of instructional materials, researchers interested in Urdu literacy can observe how teachers use textbooks, how strictly activities are followed, and other classroom practices that may all contribute to vocabulary growth in the early school years. Finally, the AoA effect has not yet been studied in Urdu. Psycholinguistic norms from the present study can be used to further test the arbitrary mapping hypothesis as well as the AoA effect on reading speed and other variables.

5.6.2. Word Lists.

Methodological rigor in developing the word list and high inter-rater agreement on the AoA ratings point to the robustness of the word list as a tool for research on children's language learning. The use of this word list also extends beyond the present study. Appendices N-O provide materials other researchers can use for Urdu language and literacy research. Appendix N

provides details of all 600 words extracted for the AoA study as well as information on other psycholinguistic variables. Appendix O presents a graded word list of thirty words that can be used as a spelling test in the classroom as well as a literacy measure in other research. This word list has been stratified according to the three transparency categories from no vowelization to complete vowelization. Due to negative associations with written frequency, each ten-words strata of the word list was matched for frequency, but the list represents a range of frequencies (1-33). Words were also matched for syllable and letter counts across strata. Discrepant words were not included and conjunctions were excluded to ensure inter-rater reliability of included words was of high quality. The proportions of word classes identified from the word corpus were replicated. The final word list contains 14 nouns, ten verbs, four adjectives, and two adverbs. This graded list can be used for spelling and reading, and can also be used for other types of psycholinguistic studies. For example, researchers can study the role of diacritics in Urdu by adding them artificially to semi-vowelized or un-vowelized words in the list and then checking performance on measures such as reading speed. Details from the original word list can be specifically curated to answer research questions through a similar process of stratification. It is hoped that the example and complete word list is helpful for researchers and encourages them to pursue more research in Urdu child language acquisition and literacy.

5.7. Conclusions

In part, this study aimed to investigate when children learn words intentionally selected from child-directed print and divided into different word classes, and the relationships between this word-learning and other psycholinguistic variables. The study also aimed to investigate primary school textbooks' contributions to children's vocabulary development through their promotion of morphological and orthographic awareness. Results indicate that adult raters reliably agree that, on average, children acquire words contained in their textbooks before entering primary school. Textbooks contain a limited range of morphological and orthographic activities. Within that limited range, there is a preference for the component skills of vocabulary and letter knowledge respectively. The current research highlights potential challenges of such limited vocabulary exposure and restricted types of activities in the textbook for children's vocabulary growth, but also acknowledges the advantages of these findings for literacy instruction. Future research can use the findings and materials produced in the present research for more observational and experimental work.

References

- Alonso, M., Fernandez, A., & Díez, E. (2015). Subjective age-of-acquisition norms for 7,039 Spanish words. *Behavior Research Methods*, 47(1), 268–274.
<https://doi.org/10.3758/s13428-014-0454-2>
- Anglin, J. M., Miller, G. A., & Wakefield, P. C. (1993). *Vocabulary development: A morphological analysis*. Chicago University Press.
- Ansary, H., & Babaii, E. (2002). Universal characteristics of EFL/ESL textbooks: A step towards systematic textbook evaluation. *The Internet TESL Journal*, 8(2), 1-9.
- Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from <https://www.laurenceanthony.net/software>
- Arslan, S. (2016). An analysis of two Turkish EFL books in terms of cultural aspects. *Procedia - Social and Behavioral Sciences*, 232, 217–225.
<https://doi.org/10.1016/j.sbspro.2016.10.049>
- Ashkenazi, O., Ravid, D., & Gillis, S. (2016). Breaking into the Hebrew verb system: A learning problem. *First Language*, 36(5), 505–524. <https://doi.org/10.1177/0142723716648865>
- Bahrani, N. (2014). Evaluating the representation of cultural elements in an in-use EFL textbook. *Advances in Language and Literary Studies*, 6(3), 128–137.
<https://doi.org/10.7575/aial.v.6n.3p.128>
- Bakhtiar, M., & Weekes, B. (2015). Lexico-semantic effects on word naming in Persian: Does age of acquisition have an effect?. *Memory & Cognition*, 43(2), 298–313.
<https://doi.org/10.3758/s13421-014-0472-4>
- Belkhouche, B., Harmain, H., Al-Taha, H., Al-Najjar, L., & Tibi, S. (2010). Analysis of primary school Arabic language textbooks. In *Proceedings of the Arab Conference on Information Technology*, 10 (pp. 1-5).
- Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, 93(3), 498–520. <https://doi.org/10.1037/0022-0663.93.3.498>
- Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, and Computers*, 33(1), 73–79. <https://doi.org/10.3758/BF03195349>

- Bonin, P., Barry, C., Méot, A., & Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: A never ending story?. *Journal of Memory and Language*, 50(4), 456–476. <https://doi.org/10.1016/j.jml.2004.02.001>
- Boudelaa, S., & Marslen-Wilson, W. D. (2005). Discontinuous morphology in time: Incremental masked priming in Arabic. *Language and Cognitive Processes: Current Issues in Morphological Processing*, Ram Frost, Jonathan Grainger and Kathleen Rastle (Eds), 20(1–2), 207–260. <https://doi.org/10.1080/01690960444000106>
- Boudelaa, S., & Marslen-Wilson, W. D. (2013). Morphological structure in the Arabic mental lexicon: Parallels between standard and dialectal Arabic. *Language and Cognitive Processes*, 28(10), 1453–1473. <https://doi.org/10.1080/01690965.2012.719629>
- Bryman, A. (2016). *Social research methods* (Fifth ed.). Oxford.
- Cameirão, M. L., & Vicente, S. G. (2010). Age-of-acquisition norms for a set of 1,749 Portuguese words. *Behavior Research Methods*, 42(2), 474. <https://doi.org/10.3758/BRM.42.2.474>
- Cardona, G. (2019, November 15). Sanskrit language. Retrieved August 28, 2020, from <http://www.britannica.com/topic/Sanskrit-language>.
- Chen, B. G., Zhou, H. X., Dunlap, S., & Perfetti, C. A. (2007). Age of acquisition effects in reading Chinese: Evidence in favour of the arbitrary mapping hypothesis. *British Journal of Psychology*, 98(3), 499–516. <https://doi.org/10.1348/000712606X165484>
- Code of Ethics and Conduct (2018). Retrieved August 28, 2020, from <https://beta.bps.org.uk/news-and-policy/bps-code-ethics-and-conduct>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum Associates.
- Cuetos, F., Ellis, A. W., & Alvarez, B. (1999). Naming times for the Snodgrass and Vanderwart pictures in Spanish. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, 31(4), 650. <https://doi.org/10.3758/BF03200741>
- Cunningsworth, A. (1984). *Evaluating and selecting EFL teaching material*. London: Heinemann.

- Daud, A., Khan, W., & Che, D. (2017). Urdu language processing: a survey. *Artificial Intelligence Review*, 47(3), 279–311. <https://doi.org/10.1007/s10462-016-9482-x>
- Deacon, S. H., & Kirby, J. R. (2004). Morphological awareness: Just more phonological? The roles of morphological and phonological awareness in reading development. *Applied Psycholinguistics*, 25(2), 223–238. <https://doi.org/10.1017/S0142716404001110>
- Dukes, R. J., & Strauch, K. (1979). Readability of college astronomy and physics texts III. *The Physics Teacher*, 21(7), 447–451. <https://doi.org/10.1119/1.2341355>
- Eltinge, E., & Roberts, C. (1993). Linguistic content analysis: A method to measure science as inquiry in textbooks. *Journal of Research in Science Teaching*, 30(1), 65–83.
- Ethical Guidelines for Educational Research, fourth edition (2018). Retrieved August 28, 2020, from <https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018>
- Fatima, G., Shah, S. K., & Sultan, H. (2015). Textbook analysis and evaluation of 7th & 8th Grade in Pakistani context. *International Journal of English Language Teaching*, 3(4), 79-97. <https://doi.org/10.7575/aiac.all.v.6n.3p.128>
- Fatima, T., Islam, R. U., & Anwar, M. W. (2018, MaY). *Morphological and orthographic challenges in Urdu language processing: A review* [Paper presentation]. The 13th Workshop on Asian Language Resources, Miyazakai.
- Field, A. (2009). *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll)* (3rd ed., ISM (London, England)). Los Angeles [i.e. Thousand Oaks, Calif.]; London: SAGE Publications.
- Freyd, P., & Baron, J. (1982). Individual differences in acquisition of derivational morphology. *Journal of Verbal Learning and Verbal Behavior*, 21(3), 282–295. [https://doi.org/10.1016/S0022-5371\(82\)90615-6](https://doi.org/10.1016/S0022-5371(82)90615-6)
- Garlock, V., Walley, A., & Metsala, J. (2001). Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language*, 45(3), 468–492. <https://doi.org/10.1006/jmla.2000.2784>
- Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech, & G. Sampson (Eds.), *The Computational analysis of English: A corpus-based approach*, (pp. 30-41). London: Longman.

- Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; no. 257*.
- Gilhooly, K. J., & Gilhooly, M. L. (1980). The validity of age-of-acquisition ratings. *British Journal of Psychology, 71*, 105-110.
- Gilhooly, K., & Logie, R. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12*(4), 395–427. <https://doi.org/10.3758/BF03201693>
- Goodwin, A. P., & Ahn, S. (2013). A meta-analysis of morphological interventions in English: Effects on literacy outcomes for school-age children. *Annals of Dyslexia, 60*, 183-208. <https://10.1007/s11881-010-0041-x>
- Göz, I., Tekcan, A., & Erciyes, A. (2017). Subjective age-of-acquisition norms for 600 Turkish words from four age groups. *Behavior Research Methods, 49*(5), 1747–1747. <https://doi.org/10.3758/s13428-016-0817-y>
- Guilloteaux, M. (2013). Language textbook selection: Using materials analysis from the perspective of SLA principles. *The Asia-Pacific Education Researcher, 22*(3), 231–239. <https://doi.org/10.1007/s40299-012-0015-3>
- Gutman, A., & Avanzati, B. (2013). Urdu. Retrieved from <http://www.languagesgulper.com/eng/Urdu.html>
- Hardie, A. (2003). Developing a tagset for automated part-of-speech tagging in Urdu. In *Corpus Linguistics 2003*. https://eprints.lancs.ac.uk/id/eprint/103/1/cl03_urdu.pdf
- Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London: Longmans
- Hardie, A. (2003). Developing a tagset for automated part-of-speech tagging in Urdu. In *Corpus Linguistics 2003*. https://eprints.lancs.ac.uk/id/eprint/103/1/cl03_urdu.pdf
- Harmon, J. M., Wood, K. D., & Kiser, K. (2009). Promoting vocabulary learning with the interactive word wall. *Middle School Journal, 40*(3), 58–63. <https://doi.org/10.1080/00940771.2009.11495588>
- Hussain, S. (2004). To-sound conversion for Urdu text-to-speech system. In *Proceedings of the workshop on computational approaches to Arabic script-based languages* (pp. 74-79). <https://dl.acm.org/doi/pdf/10.5555/1621804.1621823>

- Humayoun, M., Nawab, R. M. A., Uzair, M., Aslam, S., & Farzand, O. (2016, May). Urdu summary corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož (pp. 796-800).
- Ijaz, M., & Hussain, S. (2007). Corpus based Urdu lexicon development. In *Proceedings of Conference on Language Technology (CLT07)*, Pakistan (pp. 73-84).
- Jabbar, A., Iqbal, S., Khan, M., & Hussain, S. (2016). A survey on Urdu and Urdu like language stemmers and stemming techniques. *Artificial Intelligence Review*, 49(3), 339–373. <https://doi.org/10.1007/s10462-016-9527-1>
- Jawaid, B., Kamran, A., & Bojar, O. (2014, May). *A tagged corpus and a tagger for Urdu*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik (pp. 2938-2943).
- Khwaileh, T., Mustafawi, E., Herbert, R., & Howard, D. (2018). Gulf Arabic nouns and verbs: A standardized set of 319 object pictures and 141 action pictures, with predictors of naming latencies. *Behavior Research Methods*, 50(6), 2408–2425. <https://doi.org/10.3758/s13428-018-1019-6>
- Khwaileh, T., Mustafawi, E., Howard, D., & Herbert, R. (2018). Imageability, familiarity, and age of acquisition ratings for Arabic abstract nouns, abstract verbs and adjectives. *Mental Lexicon*, 13(3), 354–387. <https://doi.org/10.1075/ml.18016.khw>
- Laabidi, H., & Nfissi, A. (2016). Fundamental criteria for effective textbook evaluation. *EFL Journal*, 1(2), 141-159. <http://dx.doi.org/10.21462/eflj.v1i2.13>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>
- Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In *Corpus linguistics and the Web* (133-149). Brill Rodopi.
- Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4: the tagging of the British National Corpus. In *COLING The 15th International Conference on Computational Linguistics*, 1 (pp.622-628).
- Luniewska, M., Wodniecka, Z., Miller, C. A., Smolik, F., Butcher, M., Chondrogianni, V., Hreich, E. K., Messarra, C., A. Razak, R., Treffers-Daller, J., Yap, N. T., Abboud, L., Talebi, A., Gureghian, M., Tuller, L., & Haman, E. (2019). Age of acquisition of 299 words in seven languages: American English, Czech, Gaelic, Lebanese Arabic, Malay,

- Persian and Western Armenian. *PLoS ONE*, 14(8), 1-19.
<https://doi.org/10.1371/journal.pone.0220611>
- Machmudah, U., Khuzaimah, K., & Sholihah, M. (2019). The characteristics of Arabic language textbooks of madrasah ibtidaiyah. *Jurnal Al-Bayan: Jurnal Jurusan Pendidikan Bahasa Arab*, 11(1), 179-193.
- MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk (third edition): Volume I: Transcription format and programs, Volume II: The database. *Computational Linguistics*, 26(4), 657–657. <https://doi.org/10.1162/coli.2000.26.4.657>
- McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R. M., & Lannon, R. (2011). An image is worth a thousand words: Why nouns tend to dominate verbs in early word learning. *Developmental science*, 14(2), 181–189. <https://doi.org/10.1111/j.1467-7687.2010.00968.x>
- Maskor, Z. M., Baharudin, H., Lubis, M. A., & Yusuf, N. K. (2016). Teaching and learning Arabic vocabulary: From a teacher's experiences. *Creative Education*, 7(3), 482-490. <http://dx.doi.org/10.4236/ce.2016.73049>
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- McGraw, K. O., & Wong, S. P. (1996). Forming Inferences about Some Intraclass Correlation Coefficients. *Psychological Methods*, 1(1), 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>
- Miekley, J. (2005). ESL textbook evaluation checklist. *The reading Matrix*, 5(2). <https://doi=10.1.1.534.2299&rep=rep1&type=pdf>
- Mirdehghan, M. (2010). Persian, Urdu, and Pashto: A comparative orthographic analysis. *Writing Systems Research*, 2(1), 9–23. <https://doi.org/10.1093/wsr/wsq005>
- Mobarakeh, S. D., & Arani, H. K. (2012). Textbook Evaluation: A Comparative Study between Iranian and Turkish. *The Iranian EFL Journal*, 8(4), 242-260.
- Monaghan, J., & Ellis, A. W. (2002). What exactly interacts with spelling-sound consistency in word naming? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 28, 183-206.

- Morrison, C., Chappell, T., & Ellis, A. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology*, 50(3), 528–559.
- Morrison, C., & Ellis, A. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1), 116. <https://doi.org/10.1037/0278-7393.21.1.116>
- Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition*, 20(6), 705. <https://doi.org/10.3758/BF03202720>
- Nag, S., Chiat, S., Torgerson, C., & Snowling, M. J. (2014). Literacy, foundation learning and assessment in developing countries. *DFID publication*.
- Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, 19(3), 304. <https://doi.org/10.2307/747823>
- Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly*, 20(2), 233–253. <https://doi.org/10.2307/747758>
- New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: new evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, 13(1), 45. <https://doi.org/10.3758/BF03193811>
- O’Keeffe, L. (2012). A framework for textbook analysis. *International Review of Contemporary Learning Research*, 2(1), 1-13. <https://doi.org/10.1007/s10763-013-9463-3>
- Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, 17(4), 273–281. <https://doi.org/10.1080/17470216508416445>
- Ormerod, F., & Ivanič, R. (2002). Materiality in children’s meaning-making practices. *Visual Communication*, 1(1), 65–91. <https://doi.org/10.1177/147035720200100106>
- Oullette, G. P. (2006). What’s meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, 98(3), 554. <https://doi.org/10.1037/0022-0663.98.3.554>
- Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, 76(4), 763–782. <https://doi.org/10.1111/j.1467-8624.2005.00876.x>

- Piao, S., Archer, D., Mudraya, O., Rayson, P., Garside, R., McEnery, T., & Wilson, A. (2005). A large semantic lexicon for corpus annotation. *Corpus Linguistics*.
http://stig.lanacs.ac.uk/people/paul/publications/cl2005_estlex.pdf
- Portney, L. G., & Watkins, M. P. (2000). *Foundations of clinical research: Applications to practice* (2nd ed.). Prentice Hall.
- Prince, R. E. C. (2009). Usable knowledge from Harvard Graduate School of Education- Morphological analysis: New light on a vital reading skill, HGSE Nonie Lesaux. Retrieved from: <http://www.uknow.gse.harvard.edu/teaching/TC102-407.html>
- Punch, K., & Oancea, A. (2014). *Introduction to research methods in education* (2nd edition). London.
- Rababah, H. S., & Al-Rababah, I. H. (2013). A criteria for evaluating textbooks used for teaching Arabic to non-Arabic speakers: Language teaching methodology. *European Journal of Social Sciences*, 38, 453-469.
- Rahman, M. M., & Sinha, R. (2010). Class V Hindi textbooks: Materials analysis in the perspective of language teaching principles. *English for Specific Purposes World*, 30(9), 1-23.
- Rao, C., Vaid, J., Srinivasan, N., & Chen, H.-C. (2011). Orthographic characteristics speed Hindi word naming but slow Urdu naming: evidence from Hindi/Urdu biliterates. *Reading and Writing*, 24(6), 679–695. <https://doi.org/10.1007/s11145-010-9256-9>
- Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development, *Child Development*, 83(5), 1762–1774. <https://doi.org/10.1111/j.1467-8624.2012.01805>.
- Saiegh-Haddad, E. (2018). MAWRID: A model of Arabic word reading in development. *Journal of Learning Disabilities*, 51(5), 454–462. <https://doi.org/10.1177/0022219417720460>
- Saiegh-Haddad, E., Hadieh, A., & Ravid, D. (2012). Acquiring noun plurals in Palestinian Arabic: Morphology, familiarity, and pattern frequency. *Language Learning*, 62(4), 1079–1109. <https://doi.org/10.1111/j.1467-9922.2012.00727.x>
- Saleem, A. M., Kabir, H., Riaz, M. K., Rafique, M. M., Khalid, N., & Shahid, S. R. (2002). Urdu consonantal and vocalic sounds. *CRULP Annual Student Report*.

- Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp.97– 110). New York: Guilford Press.
- Schmidt, R. L. (1999). *Urdu, an essential grammar*. Routledge.
- Schock, J., Cortese, M., Khanna, M., & Toppi, S. (2012). Age of acquisition estimates for 3,000 disyllabic words. *Behavior Research Methods*, 44(4), 971–977.
<https://doi.org/10.3758/s13428-012-0209-x>
- Shalhoub-Awwad, Y., & Leikin, M. (2016). The Lexical status of the root in processing morphologically complex words in Arabic. *Scientific Studies of Reading*, 20(4), 296–310.
<https://doi.org/10.1080/10888438.2016.1180525>
- Smith, N., & McEnery, T. (1998). Issues in transcribing a corpus of children’s handwritten projects. *Digital Scholarship in the Humanities*, 13(4), 217–225.
<https://doi.org/10.1093/llc/13.4.217>
- Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, 56(1), 72–110.
<https://doi.org/10.3102/00346543056001072>
- Taha, H., & Saiegh-Haddad, E. (2017). Morphology and spelling in Arabic: Development and interface. *Journal of Psycholinguistic Research*, 46(1), 27–38.
<https://doi.org/10.1007/s10936-016-9425-3>
- Tang, X. (2019). Learnability of grammatical sequencing: A processability perspective of textbook evaluation in EFL Settings. *Chinese Journal of Applied Linguistics*, 42(2), 236–257. <https://doi.org/10.1515/CJAL-2019-0014>
- Library of Congress. (n.d.). *Urdu Romanization Table*. The Library of Congress.
<https://www.loc.gov/catdir/cpso/romanization/>
- Templin, M. C. (1957). *Certain language skills in children: their development and interrelationships*. University of Minnesota Press.
- Tibi, S., & Kirby, J. R. (2017). Morphological awareness: Construct and predictive validity in Arabic. *Applied Psycholinguistics*, 38(5), 1019–1043.
<https://doi.org/10.1017/S0142716417000029>
- Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Harvard University Press.

- Tomlinson, B. (2012). Materials development for language learning and teaching. *Language Teaching*, 45(2), 143–179. <https://doi.org/10.1017/S0261444811000528>
- Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language*, 28(6), 649–667. [https://doi.org/10.1016/0749-596X\(89\)90002-8](https://doi.org/10.1016/0749-596X(89)90002-8)
- Urooj, S., Hussain, S., Adeeba, F., Jabeen, F., & Parveen, R. (2012). CLE Urdu digest corpus [Paper presentation]. In S. Hussain, A. Khan, M. Butt, & T. Ahmed. *Language & Technology*, Lahore (47-54). Society for Natural Language Processing (SNLP) Pakistan.
- Vagh, S. B., & Nag, S. (2014). A template for the analysis of primary school language textbooks in Indian languages. The promise foundation. http://www.linguaakshara.org/resources/assessment_and_tool_construction
- Valverde, G., Bianchi, L., Wolfe, R., Schmidt, W., Houg, R., Sabates, R., Feinstein, L., & Gilbert, J. (2002). According to the book: using TIMSS to investigate the translation of policy into practice through the world of textbooks. *British Journal of Educational Studies*, 51(4), 437–438.
- Vaknin-Nusbaum, V., & Saiegh-Haddad, E. (2020). The contribution of morphological awareness to reading comprehension in Arabic-speaking second graders. *Reading and Writing*, 1-24. <https://doi.org/10.1007/s11145-020-10048-y>
- Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading: Vocabulary Growth and Reading Skill*, 15(1), 8–25. <https://doi.org/10.1080/10888438.2011.536125>
- Walley, A. C., & Metsala, J. L. (1992). Young children’s age-of-acquisition estimates for spoken words. *Memory & Cognition*, 20(2), 171.
- Wattad, H., & Rabia S. A., (2020). The advantage of morphological awareness among normal and dyslexic native Arabic readers: A literature review. *Reading Psychology*, 41(3), 130-156. <https://doi.org/10.1080/02702711.2020.1768974>
- Waxman, S. R., & Lidz, J. L. (2006). Early word learning. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology* (pp. 299-335). J. Wiley.
- Zevin, J. D., & Seidenberg, M. S. (2002). Age-of-acquisition effects in reading and other tasks. *Journal of Memory & Language*, 47, 1-29. <https://doi.org/10.1037/0278-7393.21.1.116>

Appendix A: Urdu Phonology and Orthography

Urdu, a member of the Indo-European language family, is the national language of Pakistan and has over 170 million users (Cardona, 2019) around the world including India, Nepal, Bangladesh, the U.S., and the U.K., Several languages, including Hindi, Sanskrit, Persian, and Arabic have influenced the development of Urdu. The subsections below elaborate Urdu's convergences and divergences with some of these languages and highlight Urdu's distinct linguistic properties.

Urdu phonology

Urdu's phonetic inventory consists of 44 consonants (sixteen of which are aspirated), and 23 vowels (with a combination of long and short, nasalized and non-nasalized vowel sounds) (Saleem et al., 2002; Hussain, 2007). Most consonants are stops and affricates, but the phonetic inventory also contains fricatives, nasals, and glides (Saleem et al., 2002). Additionally, Urdu phonology includes both aspirated and retroflex stops (Cardona, 2019). Additionally, Urdu phonology includes both aspirated and retroflex stops (Cardona, 2019).

Urdu and Hindi share a common historical and geographic ancestry and as such, are highly phonologically similar (Cardona, 2019). Where Urdu phonology diverges from Hindi phonology, it borrows from other languages that have influenced its development, such as Arabic and Persian. An example of this type of borrowing can be the inclusion of the glottal ʔ (/q/) consonant, which is not found in Hindi. In terms, of pronunciation, Gutman and Avanzati (2013) note that while the stress often falls on the penultimate syllable, words are not solely marked by syllable stress. As the next section suggests, Urdu's predominantly shared phonology with Hindi is complicated by its reliance on Persian and Arabic influences in its orthography. A complete, open-access Urdu phonetic inventory with IPA transcriptions can be found through the Centre for Language Engineering's website (<http://cle.org.pk/>).

Urdu orthography

Written in the Nastaliq style, Urdu derives its writing system from Arabic (Ijaz & Hussain, 2007). Urdu has a cursive writing system (Mirdehghan, 2010) and is written from right to left. The Urdu alphabet consists of 39 basic letters (most of which are consonants) and 13 extra symbols (including diacritics) that collectively build its 52 characters. Additionally, the Urdu writing system contains eleven digraphs to represent aspirates such as ڀ (/bh/) and ڄ (/djh/).

Due to the influence of Arabic and Persian in its developed, Urdu is often listed as an *abjad*, or consonantal writing system with opaque orthography. However, Urdu also contains many transparent words, leading to variability in grapheme-phoneme mapping. These words can contain diacritics to mark short vowel sounds or include graphemes such as ا (/a:/), و (/o:/), ع (/i:/), and ے (/e:/) that produce long vowel sounds. While Urdu contains diacritics (◌َ), (◌ِ), (◌ُ), (◌̣), and (◌̤)), these markers are mostly considered optional as most words are written only with letters that represent the consonantal content of a word (Hussain, 2004). Even when semantically different words are spelled in the same way (unless diacritics are added to distinguish their pronunciation), diacritics remain optional (Ijaz & Hussain, 2007). The exclusion of diacritics may draw Urdu closer to an abjad writing system. Long vowels are always represented through the graphemes mentioned above, drawing Urdu closer to a more transparent alphabetic writing system. Urdu may therefore be considered an impure-abjad.

In addition to limited vowelization, another source of inconsistency in phoneme to grapheme mapping can be attributed to the uniqueness of several graphemes. For example, Urdu is rich in homographs and contains seven sets of letters that represent the same phoneme. These letters are phonologically distinguishable in Arabic, but not in Urdu (see Table A.1.). Like Arabic, however, Urdu also contains a silent alphabet, ء, whose use is infrequent and purpose is to document the presence of a vowel cluster (Humayoun, 2006). Lastly, certain alphabets can serve as representations of both consonants and vowel sounds depending on their place in a written word. For example, the three ‘/h/’ alphabets (ه and ه, ح) can be read as ‘/h/’ when they appear in the initial or medial part of a word, but two of these (ه and ه) can render the long vowel sounds (/a:/ or /e:/) if they appear at the end of a word. The nontransparent representation of vocalic content in the writing system coupled with the presence of several homographs and unique, multipurpose letters adds to Urdu’s orthographic depth.

Homographs	IPA Transcription
ا، آ، ع	/ɑ:/, /ɑ:/, /ɑ:/
س، ث، ص	/s/, /s/, /s/
ح، ہ، ه	/h/, /h/, /h/
ت، ط	/t/, /t/
ق، ک	/q/, /k/
ز، ذ، ض، ظ	/z/, /z/, /z/, /z/
ی، یٰ	/j/, /j/

Table A.1. Homophonic letters in Urdu: Inconsistent phoneme to grapheme mapping

Urdu's often opaque orthography is not the only challenge learners potentially face as Urdu is also visually complex (Rao et al., 2011). There is a lack of standardization rules for writing in Urdu (Fatima et al., 2018). For example, Urdu contains both joiner and non-joiner alphabets. Joiner alphabets (ی، ہ، ہ، م، ل، گ، ق، ک، ف، غ، ع، ظ، ط، ض، ص، ش، س، خ، ح، چ، ج، ٹ، ت، پ، ب) take a different written shape based on their position in a word (i.e. initial, medial, or final), whereas non-joiners (ا، آ، ع، ہ، ہ، ہ، ی، یٰ) are always written with the same shape and do not connect to the letters that precede or follow them. In addition to being aware of the three written shapes of all joiner alphabets, if a word ends with a joiner alphabet and the next word begins with a joiner alphabet, the writer must insert space between the two words. Digraph representation is also complicated by the rules applied to the initial grapheme as the same type of phonemes are either visually represented as a single unit or as two separate graphemes (ہج versus ہه). Space also becomes an issue in the case of connected stems (Fatima et al., 2018). Two stems can be joined together to create a new word. However, some of these new words can be written both by joining them (if the first stem ends with a joiner alphabet) to make a single stem or by keeping the space between them. Both spellings would be accepted as correct (Fatima et al., 2018), but this acceptance is not extended to all connected stems and the rules are arbitrarily applied to specific words (such as اس کا (/us ka:/) can also be written as اسکا (/uska:/).

Similarly, there is a lack of consensus on how to write compound words in terms of whether space should be inserted between different stems or morphological units and if so, where. This problem is accentuated by the fact that two of the four possible ways of creating compound words (murakkab lafz) require adding specific morphemes whose insertion depends on the letters before and after them as they are joiner alphabets, one requires adding the diacritic

marker *zair* (◌ِ), and one requires combining two stems (Fatima et al., 2018). Such inconsistencies in vowelization, phoneme to grapheme representation, lack of standardized rules based on word position (i.e. initial, medial, and final) and grapheme type (joiner or non-joiner) and lack of standardization in the rules of the writing system not only contribute to Urdu's orthographic depth, but also complicates the Urdu teaching and learning process.

Appendix B: Adaptations to the Library of Congress' Urdu Romanization Table

Nastaliq Script	LoC Romanization Table	Adaptation
ث	<u>s</u>	sss
خ	<u>kh</u>	kkhh
ذ	<u>z</u>	zz
ط	<u>t</u>	ttt
ظ	<u>z</u>	zzz
غ	<u>gh</u>	gghh
ں	<u>n</u>	nn

Appendix C: Kolmogorov-Smirnov Tests for Normality and Descriptive Statistics for Frequency of Word Classes

Word Class	Kolmogorov-Smirnov normality test			% of words with frequency of 1
	Statistic	df	Sig.	
Nouns	.33	875	< .001	60.7
Verbs	.39	653	< .001	57.4
Adjectives	.37	272	< .001	61
Adverbs	.32	134	< .001	45.5
Conjunctions	.33	19	< .001	42.1
Postpositions	.29	15	< .001	26.7

Appendix D: Text for Facebook Recruitment Post

I am a graduate student at the University of Oxford pursuing my MSc in Education (Child Development & Education). I am looking for elementary school teachers (especially Urdu teachers) to help me in my dissertation research on Urdu linguistics and language acquisition in children. If you are interested or know anyone who is, please let me know by providing your email address and I can email the details to you 😊 Thank you in advance!

Appendix E: Project Information Email for Teachers



University of Oxford
Department of Education

Principal Investigator: Associate Professor xx
Student Researcher: xx

Tel: +44(0)1865 274024
Fax: +44(0)1865274027



Lafzon ka Guldasta

An Exploratory Study of Urdu Vocabulary Development in Grades 1-3

Ethics Approval Reference: ED-CIA-20-205

Dear Ms. xx,

My name is xx and I am a research student at the Department of Education, University of Oxford. I have a Bachelor's of Arts Degree in Psychology from xx and am currently studying Child Development and Education for my MSc course. A more experienced researcher, xx, who is an Associate Professor of Education and the Developing Child, is the main investigator of this project and she supervises my research.

I am writing to invite you to take part in my research study, with the rest of your colleagues. To participate in the study, you will be invited to attend three different one-on-one Zoom video-call sessions with me. You can choose dates and times for each session as per your convenience. Each session will take around thirty minutes. In these sessions, you will be asked to complete an online questionnaire, which is based on a word sorting activity, by sharing your screen with me as you go through the activity. The complete questionnaire consists of a list of approximately 600 words that have been selected from students' Urdu textbooks. In an individual session, you will be presented with a section of this word list (approximately 200 words) and asked to select the age at which you think children learn each of those words in Urdu.

Once you submit the consent form attached to this letter, you will receive further information about selecting a date and time for the first Zoom session as well as instructions on how to use Zoom. You will also have the opportunity to email the researcher if you have questions about the task or face technical issues.

Additionally, the questionnaire will ask you to respond to questions about your demographic information (such as age, gender, and teaching experience) All of your information will be anonymized and will only be seen by myself and my research supervisor. All of your information will be destroyed at the end of the research period.

Your data will be used in my MSc dissertation and may be published in academic journals. However, your personal information will always be protected. You do not need to take part in this study. If you choose to take part in the study, you can withdraw your participation at any time.

If you would like to ask any questions about the project before or during the study, please contact me (xx.xx@xx-ox.ac.uk). You can also contact the lead researcher, xx (xx.xx@education.ox.ac.uk). I will be happy to talk with you in more detail. You can withdraw your participation from the study at any time. I hope you agree to participate in the study. If you do, please fill in the consent form attached in this document and added separately as a Word doc and return it to me via email.

Thank you!

xx

Graduate Student,

MSc in Education (Child Development & Education)

Appendix F: Project Information Email for Parents



University of Oxford

Department of Education

Principal Investigator: Associate Professor xx

Tel: +44(0)1865 274024

Student Researcher: xx

Fax: +44(0)1865274027



Lafzon ka Guldasta

An Exploratory Study of Urdu Vocabulary Development in Grades 1-3

Ethics Approval Reference: ED-CIA-20-205

Dear Ms. xx,

My name is xx and I am a research student at the Department of Education, University of Oxford. I have a Bachelor's of Arts Degree in Psychology from xx and am currently studying Child Development and Education for my MSc course. A more experienced researcher, xx, who is an Associate Professor of Education and the Developing Child, is the main investigator of this project and she supervises my research.

I am writing to invite you to take part in my research study. To participate in the study, you will be invited to attend three different one-on-one Zoom video-call sessions with me. You can choose dates and times for each session as per your convenience. Each session will take around thirty minutes. In these sessions, you will be asked to complete an online questionnaire, which is based on a word sorting activity, by sharing your screen with me as you go through the activity. The complete questionnaire consists of a list of approximately 600 words that have been selected from your students' Urdu textbooks. In an individual session, you will be presented with a section of this word list (approximately 200 words) and asked to select the age at which you think children learn each of those words in Urdu.

Once you submit the consent form attached to this letter, you will receive further information about selecting a date and time for the first Zoom session as well as instructions on how to use Zoom. You will also have the opportunity to email the researcher if you have questions about the task or face technical issues.

Additionally, the questionnaire will ask you to respond to questions about your demographic information (such as age, gender, and teaching experience) All of your information will be anonymized and will only be seen by myself and my research supervisor. All of your information will be destroyed at the end of the research period.

Your data will be used in my MSc dissertation and may be published in academic journals. However, your personal information will always be protected. You do not need to take part in this study. If you choose to take part in the study, you can withdraw your participation at any time.

If you would like to ask any questions about the project before or during the study, please contact me (xx.xx@xx.ox.ac.uk). You can also contact the lead researcher, xx (xx.xx@education.ox.ac.uk). I will be happy to talk with you in more detail. You can withdraw your participation from the study at any time. I hope you agree to participate in the study. If you do, please fill in the consent form attached in this document and added separately as a Word doc and return it to me via email.

Thank you!

xx

Graduate Student,
MSc in Education (Child Development & Education)

Appendix G: Screenshots of a Sample of the AoA Questionnaire



English



English

Welcome to Part I of "Lafzon ka Guldasta"

We are interested in understanding Urdu vocabulary development in children by looking at when children learn certain words. For this task, you will be presented with a word list containing 250 words and will be asked to select the age at which you think children first learn that word. I will ask you some questions on Zoom during this process. Your responses will be kept private.

This questionnaire study should take you around 30-40 minutes to complete. Your participation in this research is voluntary. You have the right to withdraw at any point during the study. You can contact the student researcher at [REDACTED]

The Principal Investigator of this study can be contacted at [REDACTED]

By clicking the button below, you acknowledge: You are a native Urdu user. Your participation in the study is voluntary. You are aware that you may choose to terminate your participation at any time for any reason.

اردو

"لفظوں کا گلدستہ" کے حصہ اول میں خوش آمدید

ہم اردو زبان میں بچوں کے ذخیرہ الفاظ میں اضافے کی رفتار کے بارے میں تحقیق کر رہے ہیں۔ اس مقصد کے لئے 250 الفاظ کی ایک فہرست مرتب کی گئی ہے۔ اس فہرست میں دیئے گئے الفاظ بچے مختلف عمروں میں سیکھتے ہیں۔ آپ سے درخواست ہے کہ ہر لفظ کو دیکھ کر وہ عمر چن لیں جس میں وہ لفظ بچے نے سیکھا۔ اس کام کے دوران میں بذریعہ "زوم" آپ سے کچھ سوالات کروں گی۔ آپ کی دی گئی معلومات کو محفوظ رکھا جائے گا۔ یہ سوال نامہ آپ کے 30-40 لے گا۔ اس تحقیق میں آپ کی شمولیت رضا کارانہ ہوگی اور آپ کسی وقت بھی اس شرکت سے دستبردار ہونے کا اختیار رکھتے ہیں۔
رابطے کے لئے:

اگر آپ متفق ہیں تو نیچے دیئے گئے بٹن کو کلک کیجیے: اردو آپ کی مادری زبان ہے، اس تحقیق میں آپ کی شمولیت رضاکارانہ ہے، آپ کے پاس اختیار ہے کہ جب چاہیں اپنی شرکت ختم کر دیں۔

I consent, begin the study (میں رضامند ہوں، مطالعہ شروع) (کیجیے)

I do not consent, I do not wish to participate (میں رضامند نہیں ہوں، مجھے حصہ نہیں لینا)

Based on your interactions with children, please select the age at which you think children understand the spoken form of the words presented below. You do not need prior knowledge of linguistics to answer these questions. Select the age range you think is most appropriate while bearing in mind that there are no right or wrong answers.

نیچے دی ہوئی فہرست کو دیکھیے اور ہر لفظ کو عورتوں سے پڑھ کر اس عمر کا انتخاب کریں جب بچے نے یہ لفظ سمجھا۔ اس مقصد کے لیے آپ کا علم لسانی سے واقف ہونا ضروری نہیں۔ صرف وہ عمر منتخب کیجئے جو آپ کے خیال میں سب سے زیادہ درست ہو۔ خیال رہے کوئی جواب درست یا غلط نہیں ہے

	2-3yrs (Pre-nursery)	3-4yrs (Nursery)	4-5yrs (Kindergarten/KG)	5-6yrs (1st Grade)	7-9yrs (2nd-4th Grade)	10-11yrs (5th-6th Grade)	12yrs+
چھت (chhat)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ببلی (bulbul)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
حد (had)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
کروٹ (karwat)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
فصل (fasal)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

<<

>>

0%

100%

Appendix H: Adapted Vagh & Nag (2014) Textbook Analysis Template

Section 1: Overview

State / Country					
Textbook name and publication year					
Grade					
Language					
Prescribed since which academic year (note first year to current year of use)					
Total no. of units/chapters (by number of activities)					
Chapter	No. of activities	Chapter	No. of activities	Chapter	No. of activities
Genres:		No. of units (with page and activity number)			
1. Poetry					
2. Narrative texts (The main purpose of narrative text is to tell a story. Narrative text has beginning, middle and end, characters, plot or conflict, and setting. Usually, narrative texts are written from the authors imagination.)					
3. Plays / Text in dialogue form					

<p>4. Expository texts (The main purpose of expository text is to inform or describe. The texts is based on a topic and factual information is organized in a logical and interesting manner. Essays on Gandhiji, Diwali and Solar Eclipse are examples of Expository texts)</p>	
<p>5. Picture Reading (activities to describe pictures - descriptive)</p>	
<p>6. Picture Story (activities to describe a story sequence given through pictures - sequential)</p>	
<p>Teacher's manual: Included in the text book with instruction in each chapter/ Supplementary text/ instructions included in the textbook/ none available</p>	
<p>Support (supplementary) materials for remediation and enrichment with children.</p>	

Additional details (add note sheets)

Section 2: Area-wise analysis

Skill targeted	Type of activity (describe)	Units/chapters (with page and activity number)	Instruction	Three Examples
1. Listening comprehension Operational Definition: Understanding spoken language	a. listen to poem/story and respond to questions			
	b. listen about picture			
	c. follow simple instructions			
2. Oral language (Repetition and recitation) Operational Definition: Producing language through imitation, i.e. by repetition or recitation	a. repeat huroof-e-tahajji			
	b. repeat words			
	c. repeat word lists with multiple stems (such as compound words)			
	d. recite morphologically related word pairs (such as singular/plural(s), masculine/feminine)			
	e. repeat simple sentences/recite poem/song			

	f. sing rhymes/song/ poem (put poem to tune and sing; dramatize)			
	g. complete sentences			
	h. memorize and recite poems/story			
Oral language (conversational skills) Operational Definition: Communicating with others by producing language	a. answer simple questions (general questions that lead to chapter topic)			
	b. pose questions			
	c. discuss with friend (Note: use superscript ^{vocab} if also coded under vocabulary, picture identification/ naming)			
	d. interview to gather information			

	(ask teacher) (also see narrative expression (d) gather information about)			
<p>3. Vocabulary</p> <p>Operational Definition: Learning or using both familiar and unfamiliar words</p>	<p>a. picture identification/ picture naming/ real life sightings</p> <p>(Note: use superscript ^{write} if also coded under writing (words); ^{discuss} if coded under oral language conversation; ^{narrate} if coded under narrative expression)</p>			
	<p>b. use words in sentences / make sentences</p>			
	<p>c. matching pictures with words</p>			
	<p>d. ask simple questions to elicit word to name picture</p>			

	e. ask simple questions/ give clues to get child to guess the word (Word identification)			
	f. reading words with meanings (Definitions: descriptions, functions, synonyms)			
	g. give/ find words that rhyme/ have the same beginning or ending/ is a word pair			
	h. give/ find words that mean the same as XX (Synonyms/ belong to the same category)			
	i. give/find words that are morphologically related (such as singular/plural(s), masculine/feminine)			

	j. using a dictionary			
<p>4. Narrative expression</p> <p>Operational Definition: Orally present a narrative involving description and/or recounting of information</p>	a. narrate story just narrated (orally heard)/just read/ already known to child			
	b. narrate story based on picture information			
	c. narrate based on pictures/ poem/ words (Note: use superscript ^{vocab} if also coded under vocabulary; ^{oral} if coded under oral language; ^{reading comprehension} if coded under reading comprehension)			
	d. narrate experiences (what you ate this morning, what you saw in the garden)			
	e. gather information about			
<p>5. Huroof-e-Tahaji Knowledge</p> <p>Operational Definition:</p>	<p>a. single harf identification:</p> <p>I. joiner huroof-e-tahajji (i.e. پ, ب, ح, چ, ج, ٹ, ٹ, ت)</p>			

Knowledge of the Urdu script including graphemes and diacritics	<p>ض, ص, ش, س, خ, ك, ف, غ, ع, ظ, ط, ه, ه, ن, م, ل, گ, ق, ی)</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>II. non-joiner huroof-e-tahajji (ڙ, ز, ر, ڙ, ڙ, ڙ, ا, ا, ے, و, ڙ)</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>b. identification of vowels represented by diacritics (such as ِ and َ)</p>			
	<p>c. complete sequence of huroof-e-tahajji (for e.g. ... پ, پ, ت)</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>d. unique huroof-e-tahajji sequences (for e.g. read/write/find/identify)</p>			

	<p>huroof-e-tahaji from the same grapheme group)</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>e. homophones: differentiate between homophonic graphemes or identify/find words that begin with homophonic graphemes (for e.g. ت or ط)</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>f. huroof-e-tahaji position:</p> <p>I. trace joiner graphemes in multiple positions (i.e. word-initial, -medial, & -final)</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>II. join multiple huroof to create words with target joiner graphemes in</p>			

	<p>multiple positions (i.e. word-initial, - medial, & -final) (Note: use superscript ^{write} if also coded under writing)</p>			
<p>7. Phonological awareness</p> <p>Operational Definition: Ability to attend to and discriminate between phonemes at the letter and word level</p>	<p>a. beginning sounds: identify/ list all words beginning with the same sound/same vowel/ same consonant in the beginning (maa, maal, maan, maar)</p>			
	<p>b. same vowel but added to different consonants (/oo/ in sooraj, phool, chooha, bhookh; and /ee/ in qeemat, zeenat, papeeta, meezan)</p>			
	<p>c. ending sounds: Identify/list all words ending with the same sound/same vowel/ same consonant</p>			
	<p>d. sound games: recite/write as many words with the same beginning sound (alone or in groups)</p>			

	e. Rhyming words			
	All may be from text. Other responses perhaps also allowed.			
8. Word Decoding Operational Definition: Ability to read, understand, and use words	a. Read simple word lists (with monosyllabic words with -a, -aa, or -o vowel sounds)			
	b. read word lists that contain mostly vowelized words (through diacritics or presence of long vowels) (typically extracted from chapter text, note if words are in inflected form or only roots)			
	c. Note whether word lists that contain with mixed letter conventions are presented horizontally or vertically			
	d. Reading lists with words that are non-transparent:			

	non-vowelized and non-diacritized)			
	e. Homonyms:			
	I. read meanings of/write/ identify homonyms that are spelled differently (such as أم and علم)			
	II. read meanings of/use homonymous words that are spelled in the same way without diacritics (such as دور)			
	III. add diacritics to similarly spelled homonymous words to semantically separate them			

	<i>f.</i> Murakkab Lafz:			
	I. read/write/join stems of AB compound words (such as پاس آس)			
	II. read/write/join stems of/ add diacritic to A-e-B compound (such as خلق خدمت)			
	III. read/write/join stems of/ add و to A-o-B compound (such as شب و روز)			
	IV. read/write A-al-B compound (such as الاقوامی بین)			
	<i>g.</i> Knowledge of connected stems in target words with multiple correct spellings:			
	Word	Units/ Chapter & Page No.	Joiner/ Non-Joiner Stem	
	i. Konsa			
ii. Usska				

	iii. Kyunke		
	iv. Dilchasp		
	v. Bezaar		
	vi. Dilkash		
	vii. Khoobsurat		
	viii. Maahvaar		
	ix. Khushhaal		
	x. Khusgawaar		
9. Reading Comprehension: Sentences	a. Read with focus on context		
Operational Definition: Ability to read and understand sentences and sentence structures	Fill in the blanks (Note: use superscript ^{write} if also coded under writing)		
	b. Jumbled sentence		
	(Note: use superscript ^{write} if also coded under writing, ^{syntax} if also coded under syntax)		
	c. Joining given word patterns to make sentences		

<p>10. Copy writing</p> <p>Operational Definition: Imitating writing already presented</p>	a. Lines/Circle			
	b. Colouring			
	c. Consonants			
	d. Huroof-e-tahajji that make up long vowel sounds			
	e. Joining letters (CV pairs)			
	f. Simple words, rhyming words			
	g. Words with mixed letter conventions (containing mix of joiner and non-joiner letters i.e. رذذدآا (ےءوژژز)			
	h. Words that are not transparent (non-vowelized and non-diacritized)			
	i. Copy sentences			

	j. Copy numbers			
<p>11. Writing</p> <p>Operational Definition: Following prompts for independent writing</p>	<p>a. Write huroof-e-tahajji</p> <p>(Note: use superscript^{huroof} if also coded under huroof-e-tahaji)</p>			
	<p>b. Name pictures, write familiar words</p> <p>(Note: use superscript^{vocab} if also coded under vocabulary, ^{discuss.} if coded under oral language, conversation)</p>			
	<p>c. Write short sentences</p> <p>(Note: use superscript^{vocab} if also coded under vocabulary, ^{reading comprehension} if coded under reading comprehension)</p>			
	<p>a. Complete sentences/ Using words in a sentence such as ‘fill in the blanks’ /match</p>			

	sentences/ Question& Answers			
	b. Composing new word with changing position of/ matching huroof-e-tahajji or huroof-e-tahajji clusters			
	c. Composing new words: I. connect only joiner letters to create words			
	II. connect only non-joiner letters to create words			
	III. Connect letters that follow mixed conventions			
	d. Composing sentences based on a picture			
	e. Writing sentences in sequence			

	f. Composing narratives			
<p>12. Syntax</p> <p>Operational Definition: Understanding the arrangement of language structures at the word and sentence level</p>	<p>a. Suffixes</p> <p>(Note: use superscript ^{morph} if also coded under morphological awareness)</p>			
	<p>b. Compound words</p> <p>(Note: use superscript ^{morph} if also coded under morphological awareness)</p>			
	<p>c. Prefixes</p> <p>(Note: use superscript ^{morph} if also coded under morphological awareness)</p>			
	<p>d. Jumbled sentences</p> <p>(Note: use superscript ^{write} if coded under write)</p>			
	a. Factual			

<p>13. Reading comprehension: Passages</p> <p>Operational Definition: Understanding information presented in passages</p>				
	b. Inferential			
	c. Synthesizing			
	d. Elaborations			
	e. Analyzing (providing an opinion)			
<p>14. Other Literacy Devices</p> <p>Operational Definition: Understanding and/or using language specific literary devices such as idioms</p>	a. Idioms			
	b. Proverbs			
	c. Riddles			
<p>15. Morphological Awareness</p> <p>Operational Definition: Understanding and/or using language specific morphological features</p>	<p>a. <i>Wahid/Jama</i> (singular/plural)</p> <p>I. Jama Salim remove/add وں, ہیں, ے ان, ات, ین to transform words from singular to plural or vice-versa</p>			

	(Note: use superscript ^{write} if also coded under writing)			
	<p>II. Jama Mukkasar</p> <p>Recite/write the missing singular/plural for altered plurals (such as ilm and uloom)</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>III. Broken Plurals:</p> <p>Read/write multiple plurals for the same word (such as rasam)</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>b. Harf Knowledge:</p> <p>(find/use morphemes or words that join nouns and vowels to form sentences such as عز)</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			

	<p>c. Morpheme Matching: (connect/join/add multiple morphemes together to form new words)</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>d. Inflectional Morphology: Make new words/ find words using the same root morpheme by adding suffixes or prefixes</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>e. Derivational Morphology: Make new words/find words using the same root morpheme by using infixes</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
	<p>f. Gender Marking:</p>			

	<p>(Match/ fill in the blank/ list the missing gender</p> <p>(Note: use superscript ^{write} if also coded under writing)</p>			
--	---	--	--	--

Appendix I: Sample Script for Zoom AoA Sessions

Assalam-Alaikum/ Hello, my name is Rida. Thank you for participating in this study. I hope this is still a good time for us to conduct this session.

I understand that this may be an unfamiliar situation for you. I want to remind you that you can withdraw your participation at any time. If this timing does not work for you, we can reschedule to a later date and time. I want to remind you that your data is private and only my supervisor and I will have access to it.

The reason we are conducting the research on Zoom is so that I am available to answer any questions you might have in general but also with regard to a particular item on the questionnaire. I might also ask you to explain the thought process behind particular responses. For this reason, I will ask you to share your screen with me.

This session should take thirty minutes to complete. Please let me know if you need to take a break.

I have already emailed the link to the online questionnaire to you and it is also available in the chat box on Zoom. Please do not use this chat box for any other purpose. Please open the link and begin sharing your screen once you are ready to start responding to the questionnaire.

Appendix J: CUREC Approval Notification

Dear [REDACTED]

Lafzon ka Guldasta: An Exploratory Study of Urdu Language and Literacy Development through the Examination of Early-Grade Textbooks

The above application has been considered on behalf of the Departmental Research Ethics Committee (DREC) in accordance with the procedures laid down by the University for ethical approval of all research involving human participants.

I am pleased to inform you that, on the basis of the information provided to DREC, the proposed research has been judged as meeting appropriate ethical standards, and accordingly, approval has been granted.

Please continue to follow all current guidance issued by CUREC during the pandemic, notably COVID-19: CUREC guidance on research involving human participants, <https://researchsupport.admin.ox.ac.uk/governance/ethics/coronavirus>

Should there be any subsequent changes to the project which raise ethical issues not covered in the original application you should submit details to research.office@education.ox.ac.uk for consideration.

Good luck with your research study, and keep well and safe,

Yours sincerely,

All good wishes,

Liam Gearon

Chair, DREC

Appendix K: Z-Scores of Variables in the AoA Study

Variable	Skewness	Std. Error	Z-Score	Kurtosis	Std. Error	Z-Score
Age of Acquisition Ratings	1.068	.100	10.68	1.061	.199	5.33
Written Frequency	6.366	.100	63.66	45.647	.199	229.38
Syllable Count	.739	.100	7.39	3.458	.199	17.37
Letter Count	.941	.100	9.41	2.592	.199	13.03
Orthographic Category	-.424	.100	-4.24	-1.032	.199	-51.86

Appendix L: Results from the Shapiro-Wilk Test of Normality of Domains in the Textbook Analysis

Domain	F Statistic	df	Sig.
Morphology	.26	12	.02
Orthography	.34	12	.01

Appendix M: Levene's test of Homogeneity of Variance for Domains in the Textbook Analysis

Domain	F Statistic	df1	df2	Sig.
Morphology	3.86	3	8	.056
Orthography	.465	3	8	.063

Appendix N: Complete Word List from the AoA Study

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
چہت	chat	Roof	Noun	Opaque	1	1	3	1.8	1.03
بلبل	bulbul	Nightingale	Noun	Opaque	1	2	4	2.8	1.40
حد	had	Limit	Noun	Opaque	1	1	2	4	1.41
کروٹ	karvaṭ	Lying on your side	Noun	Opaque	1	2	4	3.6	1.07
فصل	faṣal	Crop	Noun	Opaque	1	2	3	4.3	0.82
کمر	kamar	Back	Noun	Opaque	1	2	3	2.6	1.17
جلسہ	jalsāh	Procession	Noun	Opaque	1	2	4	5	1.25
جڑ	jar	Root	Noun	Opaque	1	1	2	4	1.25
ترس	taras	Pity	Noun	Opaque	1	2	3	4.4	1.07
قلم	qalam	Pen	Noun	Opaque	1	2	3	3	1.25
توجہ	tavajjāh	Attention	Noun	Opaque	1	3	4	4.2	1.40
پرہت	parbat	Mountain	Noun	Opaque	1	2	4	5.1	1.79
بدھ	budh	Wednesday	Noun	Opaque	2	1	3	3	0.82
حل	hal	Solve	Noun	Opaque	2	1	2	4	0.67
بیل	bel	Vine	Noun	Opaque	3	1	3	3.6	0.97
ہمت	himmat	Strength	Noun	Opaque	3	2	3	3.5	1.27
مرتبہ	martabāh	Position	Noun	Opaque	3	3	5	3.9	1.29
مسجد	masjid	Mosque	Noun	Opaque	3	2	4	2.3	1.42
محبت	muhubat	Love	Noun	Opaque	3	3	4	2.9	1.45
کرکٹ	kirkīṭ	Cricket	Noun	Opaque	3	2	4	2.6	1.17
بل	bil	Burrow	Noun	Opaque	3	1	2	3.7	1.25
بندر	bandar	Monkey	Noun	Opaque	3	2	4	2.2	1.03
تحفہ	tuhfāh	Gift	Noun	Opaque	3	2	4	2.7	1.57
پسند	pasand	Like	Noun	Opaque	3	2	4	2.1	1.45
سبق	sabaq	Lesson	Noun	Opaque	3	2	3	2.9	1.45
پتہ	patāh	Leaf	Noun	Opaque	4	2	3	3.3	1.25

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
کمرہ	kamrāh	Room	Noun	Opaque	4	2	4	1.9	1.52
سمندر	samandar	Sea	Noun	Opaque	4	3	5	3.1	1.79
منہ	mūnh	Mouth	Noun	Opaque	4	1	3	2.1	1.45
پھل	phal	Fruit	Noun	Opaque	4	1	3	2.7	1.70
وجہ	vajāh	Reason	Noun	Opaque	4	2	3	3.9	1.37
سڑک	saṛak	Road	Noun	Opaque	5	2	3	2.8	0.79
نظر	nazzzar	Sight	Noun	Opaque	5	2	3	2.4	1.17
اندر	andar	Inside	Noun	Opaque	5	2	4	1.8	1.23
بہن	behen	Sister	Noun	Opaque	5	2	3	1.8	1.23
صبح	ṣubah	Morning	Noun	Opaque	6	2	3	2	1.49
حضرت	haẓrat	Sir	Noun	Opaque	6	2	4	4.1	0.99
وطن	vatttan	Country	Noun	Opaque	6	2	3	4	0.94
سورج	sūraj	Sun	Noun	Opaque	6	2	4	1.9	1.20
جگہ	jagāh	Place	Noun	Opaque	6	2	3	2.4	1.26
ملک	mulk	Country	Noun	Opaque	6	1	3	3.3	0.67
مدد	madad	Help	Noun	Opaque	7	2	3	2.1	1.10
جنگل	jangal	Jungle	Noun	Opaque	7	2	4	2.8	1.32
دل	dil	Heart	Noun	Opaque	7	1	2	2.2	1.20
پھول	phūl	Flower	Noun	Opaque	7	1	4	2.2	1.40
عزت	izzat	Respect	Noun	Opaque	7	2	3	4.2	1.14
شہر	sheher	City	Noun	Opaque	8	2	3	4.1	0.99
موسم	mausam	Weather	Noun	Opaque	8	2	4	3	1.15
درخت	darakkhht	Tree	Noun	Opaque	13	2	4	2.2	1.32
سب	sab	Everyone	Noun	Opaque	15	1	2	2.4	1.35
وقت	vaqt	Time	Noun	Opaque	17	1	3	2.8	1.14
دن	din	Day	Noun	Opaque	30	1	2	2.1	1.20
گھر	ghar	House	Noun	Opaque	41	1	3	1.7	1.25

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
بھینس	bhens	Buffalo	Noun	Transparent	1	1	5	3.3	1.83
بوری	borī	Sack	Noun	Transparent	1	2	4	3.5	1.78
چیلیں	cīlenn	Falcons	Noun	Transparent	1	2	5	2.8	1.75
بھیڑ	bher	Crowd	Noun	Transparent	1	1	4	3.3	1.83
بانس	bāns	Bamboo	Noun	Transparent	1	1	4	4	1.63
بوڑھے	būrhe	Old	Noun	Transparent	1	2	5	3.1	1.60
باورچی خانہ	bāvarcī kkhhānāh	Kitchen	Noun	Transparent	1	2	4	3.9	1.91
گاؤں	gā'onn	Village	Noun	Transparent	1	2	4	3.3	1.64
جال	jāl	Trap	Noun	Transparent	1	1	3	3.6	1.35
نان	nān	Bread	Noun	Transparent	1	1	3	3.5	1.72
جیب	jeb	Pocket	Noun	Transparent	1	1	3	2.6	1.43
سیڑھی	sīrhī	Ladder	Noun	Transparent	1	2	5	2.7	1.77
قینچی	qencī	Scissor	Noun	Transparent	2	2	5	2.9	1.91
گھوڑے	ghore	Horses	Noun	Transparent	2	2	5	2.5	2.01
بار	bār	Instance	Noun	Transparent	3	1	3	2.6	1.17
بیٹی	beṭī	Daughter	Noun	Transparent	3	2	4	2.7	1.57
گیند	gend	Ball	Noun	Transparent	3	1	4	2.7	2.16
دیوار	dīvār	Wall	Noun	Transparent	3	2	5	2.5	1.65
آپا	āpā	Older sister	Noun	Transparent	3	2	3	2.4	1.26
باتوں	bātonn	Talks	Noun	Transparent	3	2	5	2.5	1.65
چیزوں	cīzonn	Things	Noun	Transparent	3	2	5	2.6	1.84
اڑو	ārū	Peach	Noun	Transparent	3	2	3	3.2	1.99
مالی	mālī	Gardner	Noun	Transparent	3	2	4	3.2	1.81
چوہا	cūhā	Mouse	Noun	Transparent	3	2	4	2	1.15
بیج	bīj	Seed	Noun	Transparent	3	1	3	3.1	1.45
دوست	dost	Friend	Noun	Transparent	4	1	4	2.1	1.10
حال	hāl	Condition	Noun	Transparent	4	1	3	3.8	1.14

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
کھانے	khāne	Foods	Noun	Transparent	4	2	5	2.6	1.58
سال	sāl	Year	Noun	Transparent	4	1	3	2.6	1.17
پودا	podā	Plant	Noun	Transparent	4	2	4	2.6	1.35
شوق	shoq	Interest	Noun	Transparent	4	1	3	3.5	1.18
آدمی	ādmī	Man	Noun	Transparent	5	2	4	3.1	1.37
رات	rāt	Night	Noun	Transparent	5	1	3	1.9	1.20
کھیل	khel	Game	Noun	Transparent	5	1	4	2.1	1.37
روٹی	roṭī	Bread	Noun	Transparent	5	2	4	1.8	1.23
روز	roz	Daily	Noun	Transparent	6	1	3	2.8	1.40
شور	shor	Noise	Noun	Transparent	6	1	3	2.4	1.43
ہاتھ	hāth	Hand	Noun	Transparent	6	1	4	1.7	1.25
مان	mān	Agree	Noun	Transparent	6	1	3	3.8	1.32
جان	jān	Life	Noun	Transparent	6	1	3	4	1.49
چیزیں	cīzenn	Things	Noun	Transparent	6	2	5	2.8	1.99
کھانا	khānā	Food	Noun	Transparent	7	2	5	1.6	1.26
چاند	cānd	Moon	Noun	Transparent	7	1	4	1.9	1.29
دانتوں	dāntonn	Teeth	Noun	Transparent	7	2	6	2.1	1.60
بابا	bābā	Father	Noun	Transparent	7	2	4	1.5	1.08
لوگوں	logonn	People	Noun	Transparent	7	2	5	3.1	1.37
باتیں	bātenn	Talks	Noun	Transparent	8	2	5	2.7	1.89
بادشاہ	bādshāh	King	Noun	Transparent	8	2	6	3.2	1.55
بھائی	bhāī	Brother	Noun	Transparent	13	2	4	1.6	1.26
بات	bāt	Talk	Noun	Transparent	17	1	3	2.3	1.57
کام	kām	Work	Noun	Transparent	23	1	3	2.3	1.34
عمارتیں	imāratenn	Buildings	Noun	Mixed	1	4	7	4.1	1.52
کتابوں	kitābonn	Books	Noun	Mixed	1	3	6	3.5	1.58
محله	muhalle	Neighborhoods	Noun	Mixed	1	3	4	4.5	1.27

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
حصّے	hiṣṣe	Parts	Noun	Mixed	1	2	3	3.8	1.55
دنبے	dunbe	Lambs	Noun	Mixed	1	2	4	3.5	2.07
کشتی	kashṭī	Boat	Noun	Mixed	1	2	4	2.5	1.43
ترکیب	tarkīb	Process/Instruction	Noun	Mixed	1	2	5	4.7	1.49
لڑکے	larke	Boys	Noun	Mixed	1	2	4	2.2	1.32
نوالے	nivāle	Morsels	Noun	Mixed	1	3	5	3	1.76
تدریس	tadrīs	Teach	Noun	Mixed	1	2	5	5.9	1.29
تحریک	tehrīk	Movement	Noun	Mixed	1	2	5	5.9	0.99
ہیرا	hīrā	Diamond	Noun	Mixed	1	2	4	4.6	1.17
چراغ	cirāgghh	Lantern	Noun	Mixed	2	2	4	4.2	1.03
انگور	angūr	Grapes	Noun	Mixed	3	2	5	2.1	1.52
کنارے	kināre	Corners	Noun	Mixed	3	3	5	3.7	1.64
نبی	nabī	Prophet	Noun	Mixed	3	2	3	2.7	0.95
اندھیرا	andherā	Darkness	Noun	Mixed	3	3	7	2.5	1.58
لکڑی	lakṛī	Wood	Noun	Mixed	3	2	4	2.7	1.34
مسلمانوں	musalmānonn	Muslims	Noun	Mixed	3	4	8	3.6	1.58
لڑکا	larkā	Boy	Noun	Mixed	3	2	4	2.1	1.29
مالک	mālik	Master	Noun	Mixed	3	2	4	4	1.15
دروازے	darvāze	Doors	Noun	Mixed	3	3	6	2	1.41
بادل	bādal	Clouds	Noun	Mixed	3	2	4	2.5	1.27
جھولا	jhūlā	Swing	Noun	Mixed	3	2	5	2	1.41
جماعت	jamā't	Class	Noun	Mixed	4	1	5	4.2	1.40
نماز	namāz	Prayer	Noun	Mixed	4	2	4	2	1.25
بچو	bacco	Children	Noun	Mixed	4	2	3	2.3	1.42
دانت	dānt	Tooth	Noun	Mixed	4	1	4	1.8	1.32
سلام	salām	Greeting	Noun	Mixed	4	2	4	1.9	1.37
شاعر	shā'er	Poet	Noun	Mixed	4	2	4	5.4	0.84

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
پتے	patte	Leaves	Noun	Mixed	5	2	3	2.3	1.34
لنگور	langūr	Baboon	Noun	Mixed	5	2	5	4	1.89
کاغذ	kāgghazz	Paper	Noun	Mixed	5	2	4	2.5	0.97
بھیا	bhayyā	Older brother	Noun	Mixed	5	2	4	2.1	1.97
چڑیا	ciṛyā	Bird	Noun	Mixed	6	2	4	1.8	1.23
احسان	ehsān	Favor	Noun	Mixed	6	2	5	5	0.94
تربوز	tarbūz	Watermelon	Noun	Mixed	6	2	5	3.1	1.73
بارش	bārish	Rain	Noun	Mixed	6	2	4	2	1.33
گلاب	gulāb	Rose	Noun	Mixed	6	2	4	3.1	1.52
زمین	zamīn	Earth	Noun	Mixed	6	2	4	2.1	1.20
چوہے	cūhe	Mice	Noun	Mixed	7	2	4	2.4	1.07
دنیا	dunyā	World	Noun	Mixed	7	2	4	3	1.25
امان	ammān	Mother	Noun	Mixed	7	2	4	1.5	1.08
بلی	billī	Cat	Noun	Mixed	7	2	3	1.5	1.08
سردی	sardī	Winter	Noun	Mixed	8	2	4	2.1	1.29
انسان	insān	Human	Noun	Mixed	10	2	5	3	0.94
جانور	jānvar	Animal	Noun	Mixed	13	2	5	2.6	1.17
بچوں	bacconn	Children	Noun	Mixed	15	2	4	2.2	1.32
بچے	bacce	Children	Noun	Mixed	18	2	3	2.5	1.35
امی	ammī	Mother	Noun	Mixed	30	2	3	1.5	1.08
رنگ	rang	Color	Noun	Opaque	7	1	3	2.4	1.35
پیغام	paigghhām	Message	Noun	Mixed	2	2	5	4.8	1.14
انگریزوں	angrezonn	English	Noun	Mixed	3	3	7	5.1	1.29
کہانی	kahānī	Story	Noun	Mixed	3	3	5	1.9	1.29
لکڑہارا	lakaṛhārā	Woodcutter	Noun	Mixed	3	4	7	3.8	1.62
گھروں	gharonn	Houses	Noun	Mixed	3	2	5	3.2	1.87
الگنی	algaṇī	Clothesline	Noun	Mixed	3	3	5	6	1.15

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
مسلمان	musalmān	Muslim	Noun	Mixed	3	3	6	3.2	1.62
دوائیں	davā'enn	Medicines	Noun	Mixed	3	3	6	3.1	1.79
جھنڈا	jhandā	Flag	Noun	Mixed	3	2	5	2.4	1.78
جوتوں	jūtonn	Shoes	Noun	Mixed	3	2	5	2.4	1.65
سواری	savārī	Ride	Noun	Mixed	3	3	5	3.9	1.45
نقصان	nuqṣān	Loss	Noun	Mixed	3	2	5	4.2	1.55
خزاں	kkhhizānn	Autumn	Noun	Mixed	4	2	4	4.1	1.45
پری	parī	Fairy	Noun	Mixed	4	2	3	2.5	1.58
گرمی	garmī	Summer	Noun	Mixed	4	2	4	2.3	1.42
بہار	bahār	Spring	Noun	Mixed	4	2	4	3.8	1.03
پودے	pode	Plants	Noun	Mixed	4	2	4	2.4	1.17
جواب	javāb	Answer	Noun	Mixed	4	2	4	3	0.82
پرندے	parinde	Birds	Noun	Mixed	4	3	5	2.4	1.51
میاں	miyānn	Sir	Noun	Mixed	5	2	4	4.8	1.75
باہر	bāhir	Outside	Noun	Mixed	5	2	4	2.1	1.45
جانوروں	jānvaronn	Animals	Noun	Mixed	5	3	7	2.7	1.77
تتلی	titlī	Butterfly	Noun	Mixed	5	2	4	2.2	1.23
دریا	daryā	River	Noun	Mixed	5	2	4	3.4	1.43
ملکوں	mulkonn	Countries	Noun	Mixed	5	2	5	4.3	0.95
بچہ	baccā	Child	Noun	Mixed	5	2	3	2.4	1.43
درختوں	darakkhhtonn	Trees	Noun	Mixed	5	3	6	3	1.33
عورت	orat	Woman	Noun	Mixed	5	2	4	3.4	1.78
بکری	bakrī	Goat	Noun	Mixed	5	2	4	2.3	1.16
دودھ	dūdh	Milk	Noun	Mixed	9	1	4	1.7	1.16
کتاب	kitāb	Book	Noun	Mixed	9	2	4	2.3	1.42
کپڑے	kapṛe	Clothes	Noun	Mixed	9	2	4	2	1.33
ابو	abbū	Father	Noun	Mixed	9	2	3	1.6	1.35

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
پیار	pyār	Love	Noun	Mixed	10	2	4	2.3	1.42
اسکول	iskūl	School	Noun	Mixed	12	2	5	1.8	1.32
ہوا	havvā	Wind	Noun	Mixed	12	2	3	2.5	1.90
تعلیم	tā'līm	Education	Noun	Transparent	1	2	5	4.7	0.82
شاخیں	shākkhhenn	Branches	Noun	Transparent	1	2	5	4.2	1.40
آندھی	āndhī	Dust storm	Noun	Transparent	1	2	5	4.3	1.25
سانپ	sānp	Snake	Noun	Transparent	1	1	4	2.9	1.45
روزے	roze	Fasting	Noun	Transparent	1	2	4	2.9	1.45
چونچ	conc	Beak	Noun	Transparent	2	1	4	3.3	1.42
کیڑے	kīṛe	Insects	Noun	Transparent	2	2	4	2.6	1.26
باری	bārī	Turn	Noun	Transparent	2	2	4	2.7	1.06
نیم	nīm	Neem	Noun	Transparent	3	1	3	4.6	0.70
طوطے	ttottte	Parrots	Noun	Transparent	4	2	4	2.3	1.16
مونگ	mūng	Yellow lentil	Noun	Transparent	4	1	4	5.3	1.06
جوتے	jūte	Shoes	Noun	Transparent	4	2	4	2.1	1.29
ہاتھی	hāthī	Elephant	Noun	Transparent	4	2	5	2.1	0.99
آسمان	āsmān	Sky	Noun	Transparent	4	2	5	2.1	1.29
آبا	abbā	Father	Noun	Transparent	4	2	3	1.6	1.26
دھوپ	dhūp	Sunshine	Noun	Transparent	4	1	4	3.1	1.45
میز	mez	Table	Noun	Transparent	4	1	3	2.7	1.16
پان	pān	Betel	Noun	Transparent	4	1	3	3.8	1.23
اونٹ	ūṅṭ	Camel	Noun	Transparent	4	1	4	2.6	1.26
آپ	āp	You (term of respect)	Noun	Transparent	4	1	2	2.2	1.40
چیز	cīz	Thing	Noun	Transparent	5	1	3	2.1	1.37
بیٹا	beṭā	Son	Noun	Transparent	5	2	4	2.5	1.43
دادا	dādā	Grandfather	Noun	Transparent	5	2	4	1.6	1.07
سائیکل	sā'ikal	Bicycle	Noun	Transparent	5	2	5	1.8	1.23

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
لومڑی	lomrī	Fox	Noun	Transparent	5	2	5	3.5	1.18
آواز	āvāz	Voice	Noun	Transparent	5	2	4	2.4	1.58
بازار	bāzār	Market	Noun	Transparent	5	2	5	2.6	1.26
طوطا	tttotttā	Parrot	Noun	Transparent	6	2	4	2.1	1.10
آم	ām	Mango	Noun	Transparent	8	1	2	1.8	1.03
شام	shām	Evening	Noun	Transparent	8	1	3	2.4	1.35
گاڑی	gārī	Car	Noun	Transparent	9	2	4	1.6	0.97
خوشی	kkhhushī	Happiness	Noun	Transparent	10	2	4	3	1.33
باغ	bāgghh	Garden	Noun	Transparent	10	1	3	2.8	1.48
چاچا	cācā	Paternal uncle	Noun	Transparent	10	2	4	1.7	1.34
سامان	sāmān	Items	Noun	Transparent	10	2	5	3.6	1.35
پانی	pānī	Water	Noun	Transparent	10	2	4	1.5	1.27
دادی	dādī	Paternal grandmother	Noun	Transparent	12	2	4	1.5	1.08
لوگ	log	People	Noun	Transparent	12	1	3	3.3	1.06
نام	nām	Name	Noun	Transparent	12	1	3	1.9	0.99
ایک	ek	One	Noun	Mixed	1	1	3	1.6	0.97
چٹنی	catnī	Chutney	Noun	Mixed	1	2	4	3.1	1.29
دوسرا	dūsra	Other	Noun	Mixed	1	2	5	2.9	1.66
گڑیا	guryā	Doll	Noun	Mixed	1	2	4	2	1.49
دولت	dolat	Wealth	Noun	Mixed	1	2	4	5.3	0.48
بکرا	bakrā	Goat	Noun	Mixed	1	2	4	1.9	1.29
انگوروں	angūronn	Grapes	Noun	Mixed	1	3	7	3.7	1.89
دکاندار	dukāndār	Shop Keeper	Noun	Mixed	1	3	7	3.8	1.55
چھتری	chatrī	Umbrella	Noun	Mixed	1	2	4	3.1	1.60
فضا	fizā	Atmosphere	Noun	Mixed	1	2	3	5.4	0.84
درباریوں	darbāryonn	Courtiers	Noun	Mixed	1	3	8	5.8	1.23
چھٹیوں	chutṭyonn	Holidays	Noun	Mixed	1	2	6	3.5	1.51

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
چوزے	cūze	Chicks	Noun	Mixed	1	6	10	2.8	1.48
جگنو	jugnū	Firefly	Noun	Mixed	1	2	4	3.6	1.17
کرتا	kurta	Shirt	Noun	Mixed	1	2	4	3.4	1.26
عمارت	imārat	Building	Noun	Mixed	1	3	5	4.6	1.43
مہمانوں	mehmānonn	Guests	Noun	Mixed	1	3	7	3.7	1.49
جوابات	javābāt	Answers	Noun	Mixed	1	3	6	4.6	1.07
ہفتے	hafte	Weeks	Noun	Mixed	1	2	4	4.1	1.45
		Someone whose mother has passed away	Noun	Mixed	1	2	4	6.4	0.97
یسیر سائنس	yasīr	away	Noun	Mixed	1	2	4	6.4	0.97
دانوں	sā'ensdānonn	Scientists	Noun	Mixed	1	4	10	5.6	0.97
گنتی	gintī	Count	Noun	Mixed	1	2	4	3.2	1.23
پروں	paronn	Wings	Noun	Mixed	1	2	4	3.3	0.95
تجربہ گاہ	tajarbah gāh	Laboratory	Noun	Mixed	1	4	8	6.2	0.79
پھلیاں	phallyānn	Beans	Noun	Mixed	1	2	6	5.2	0.92
زخمی	zakkhhmī	Injured	Noun	Mixed	1	2	4	4.7	1.06
سزا	sazā	Punishment	Noun	Mixed	1	2	3	3.4	1.58
سالگرہ	sālgirāh	Birthday	Noun	Mixed	2	3	6	3.4	1.78
صفائ	ṣafā'ī	Cleanliness	Noun	Mixed	2	3	4	2.8	1.48
غذا	gghhizzā	Diet	Noun	Mixed	2	2	3	4.9	0.74
استانی	ustānī	Teacher	Noun	Mixed	2	3	6	3.9	1.37
مرغے	murgghhe	Chicken	Noun	Mixed	2	2	4	2.8	1.32
اتوار	itvār	Sunday	Noun	Mixed	2	2	5	3.1	1.20
سڑکوں	saṛkonn	Streets	Noun	Mixed	2	2	5	3.9	1.20
ٹکڑے	tukṛe	Pieces	Noun	Mixed	2	2	4	3.4	0.97
کیں	kīnn	Did	Verb	Opaque	1	1	3	3.9	1.29
جوت	jot	Plough	Verb	Opaque	1	1	3	5.5	1.18

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
ٹہل	tehel	Pace	Verb	Opaque	1	2	3	4.5	0.97
پھدک	phudak	Prance	Verb	Opaque	1	2	4	4.3	1.70
پوچھ	pūc	Ask	Verb	Opaque	1	1	4	2.6	1.17
رک	ruk	Stop	Verb	Opaque	1	1	2	1.9	1.20
شرمندہ	sharmindāh	Ashamed	Verb	Opaque	1	3	6	4.4	1.35
بند	band	Closed	Verb	Opaque	3	1	3	1.8	1.03
بن	ban	Make	Verb	Opaque	3	1	2	2.6	1.35
اڑ	uṛ	Fly	Verb	Opaque	3	1	2	2.1	1.10
لیں	lenn	Take	Verb	Opaque	3	1	3	2.2	1.40
تھک	thak	Tired	Verb	Opaque	3	1	3	2	1.05
لگ	lag	Seem	Verb	Opaque	4	1	2	2.9	1.85
سمجھ	samajh	Understand	Verb	Opaque	4	2	4	2.8	1.40
مل	mil	Join	Verb	Opaque	4	1	2	2.5	1.35
کرو	karo	Do	Verb	Opaque	4	1	3	1.8	1.14
دوڑ	dauṛ	Run	Verb	Opaque	4	1	3	2.5	1.08
رہ	reh	Left	Verb	Opaque	5	1	2	2.9	1.52
کہ	keh	Says	Verb	Opaque	5	1	2	3	1.33
ہوں	honn	Am	Verb	Opaque	5	1	3	2.7	1.34
ختم	kkhhatam	End	Verb	Opaque	6	2	3	2.1	1.45
سن	sun	Listen	Verb	Opaque	6	1	2	1.9	1.20
چل	cal	Walk	Verb	Opaque	6	1	2	2.2	1.32
ہوں	hūnn	Am	Verb	Opaque	24	1	3	2.2	1.32
کھایا	khāyā	Ate	Verb	Transparent	1	2	5	2	1.41
جھانکا	jhānkā	Peek	Verb	Transparent	1	2	6	4	1.89
جیت	jīt	Victory	Verb	Transparent	1	1	3	2.6	1.51
لاؤ	lā'o	Bring	Verb	Transparent	1	2	3	2.5	1.65
کھیلتی	kheltī	Playful	Verb	Transparent	1	2	6	2.7	2.06

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
پھولے	phūle	Swell	Verb	Transparent	1	2	5	3.9	1.37
ہار	hār	Loss	Verb	Transparent	1	1	3	2.3	1.34
بیٹھا	beṭhā	Sat	Verb	Transparent	3	2	5	2.6	1.96
دینا	denā	Give	Verb	Transparent	3	2	4	2.4	1.65
آتے	āte	Coming	Verb	Transparent	3	2	3	2.5	1.65
چاہتا	cāhtā	Want	Verb	Transparent	3	2	5	3.6	1.51
آتی	ātī	Coming	Verb	Transparent	3	2	3	2.8	1.40
ہوتا	hotā	Is	Verb	Transparent	4	2	4	2.9	1.52
بنایا	banāyā	Made	Verb	Transparent	4	1	3	2.6	1.58
کاٹ	kāṭ	Cut	Verb	Transparent	4	2	4	2.2	1.14
آئیں	ā'enn	Come	Verb	Transparent	4	2	5	3.1	1.73
کام	kām	Work	Verb	Transparent	4	1	3	2.5	1.51
گھاس	ghās	Grass	Noun	Transparent	5	1	4	2.6	1.35
پیدا	pedā	Born	Verb	Transparent	5	2	4	3.4	1.35
کھائے	khā'e	Ate	Verb	Transparent	5	2	4	2.6	1.35
بولے	bole	Said	Verb	Transparent	6	2	4	2.7	1.70
آؤ	ā'o	Come	Verb	Transparent	6	2	2	1.9	1.45
معلوم	mā'lūm	Know	Verb	Transparent	6	2	5	3.4	1.65
دی	dī	Gave	Verb	Transparent	9	1	2	2.6	1.51
لو	lo	Take	Verb	Transparent	10	1	2	2.5	1.58
گا	gā	Will	Verb	Transparent	15	1	2	2.5	1.58
آیا	āyā	Came	Verb	Transparent	18	2	3	2.4	1.51
گے	ge	Will	Verb	Transparent	19	1	2	2.7	1.49
گئے	ga'e	Would	Verb	Transparent	22	2	2	2.8	1.40
تھے	the	Were	Verb	Transparent	33	1	3	2.7	1.42
خریدتے	kkhharīdte	Buying	Verb	Mixed	1	3	6	3.7	1.70
مچا	macā	Done	Verb	Mixed	1	2	3	3.6	1.35

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
جگمگا	jagmagā	Shining	Verb	Mixed	1	3	5	4.4	1.35
ہنسنے	hansne	Laughing	Verb	Mixed	1	2	5	2.4	1.65
ہوتیں	hotīnn	Are	Verb	Mixed	1	2	5	3.3	1.70
نکلتی	nikaltī	Leaving	Verb	Mixed	1	3	5	3.1	1.91
خلاف	kkhhilāf	Compete against	Verb	Mixed	1	2	4	5.1	0.88
رکھا	rakhā	Put	Verb	Mixed	2	2	4	2.5	1.51
بتا	batā	Tell	Verb	Mixed	3	2	3	2.4	1.43
بتاؤ	batā'o	Tell	Verb	Mixed	3	3	4	2.4	1.35
بنائے	banā'e	Made	Verb	Mixed	3	3	4	2.6	1.07
کرنی	karnī	Action	Verb	Mixed	3	2	4	2.8	1.48
رکھتے	rakhte	Putting	Verb	Mixed	3	2	5	2.8	1.62
کہنے	kehne	Telling	Verb	Mixed	4	2	4	3.3	1.34
دینے	dī'e	Given	Verb	Mixed	4	2	3	2.6	1.58
لکھنا	likhnā	Writing	Verb	Mixed	4	2	5	2.4	1.17
نکلے	nikle	Leaving	Verb	Mixed	4	2	4	2.3	1.34
سیر	sair	Outing	Verb	Mixed	4	3	5	2.9	1.37
پڑی	parī	Fell	Verb	Mixed	5	2	3	2.9	1.45
لکھا	likhā	Wrote	Verb	Mixed	5	2	4	2.5	1.18
رہتا	rehtā	Lives	Verb	Mixed	5	2	4	2.8	1.32
پڑھنا	paṛhnā	Reading	Verb	Mixed	6	2	5	2.7	1.25
رہنے	rehne	Living	Verb	Mixed	6	2	4	2.9	1.37
ملے	mile	Met	Verb	Mixed	7	2	3	2.8	1.62
پوچھا	pūcā	Ask	Verb	Mixed	10	2	5	2.7	1.49
کرنے	karne	Doing	Verb	Mixed	14	2	4	2.6	1.65
لگا	lagā	About to	Verb	Mixed	15	2	3	2.8	1.62
دیا	diyā	Gave	Verb	Mixed	19	2	3	2.5	1.43
کرنے	karte	Doing	Verb	Mixed	26	2	4	2.5	1.51

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
کہا	kahā	Said	Verb	Mixed	21	2	3	2.5	1.51
رہی	rahī	Remaining	Verb	Mixed	23	2	3	2.7	1.57
گیا	gayā	Left	Verb	Mixed	40	2	3	2.3	1.25
کر	kar	Do	Verb	Opaque	120	1	2	2.4	1.51
سکتی	saktī	Can	Verb	Mixed	2	2	4	2.9	1.20
اترے	utre	Taken off	Verb	Mixed	2	2	4	2.9	1.52
اٹھا	uṭhā	Pick up	Verb	Mixed	2	2	4	2.6	1.43
شامل	shāmil	Include	Verb	Mixed	2	2	4	4.3	0.82
سنیں	sunenn	Listen	Verb	Mixed	2	2	4	2.2	1.23
سننے	sunte	Listening	Verb	Mixed	2	2	4	2.7	1.34
رہنا	rehnā	Live	Verb	Mixed	2	2	4	2.8	1.40
اگائے	ugā'e	Grow	Verb	Mixed	2	3	4	3.7	0.95
رکھے	rakhe	Kept	Verb	Mixed	2	2	4	2.4	1.43
سمجھایا	samjhayā	Explain	Verb	Mixed	2	3	7	3.3	1.49
اڑنے	uṛne	Fly	Verb	Mixed	2	2	4	2.8	1.55
رکھنا	rakhnā	Keep	Verb	Mixed	2	2	5	2.5	1.65
سنی	sunī	Heard	Verb	Mixed	2	2	3	2.4	1.26
اڑتی	uṛtī	Fly	Verb	Mixed	2	2	4	2.5	1.43
پڑھتے	paṛhte	Study	Verb	Mixed	3	2	5	2.8	1.32
سنو	suno	Listen	Verb	Mixed	3	2	3	2.5	1.27
لیا	liyā	Brought	Verb	Mixed	3	2	3	2.4	1.17
کھڑا	kharā	Stood	Verb	Mixed	3	2	4	2.3	1.25
لگتا	lagtā	Feel	Verb	Mixed	3	2	4	2.9	1.52
رہو	raho	Stay	Verb	Mixed	3	2	3	2.9	1.29
پڑھنے	paṛhne	Study	Verb	Mixed	4	2	5	2.7	1.34
گھبرا	ghabrā	Frightened	Verb	Mixed	4	2	5	4.2	1.03
لگیں	lagīnn	Attached	Verb	Mixed	4	2	4	3.4	1.71

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
رہتی	rehtī	Lives	Verb	Mixed	4	2	4	2.8	1.48
چلے	cale	Walked	Verb	Mixed	4	2	3	2.4	1.17
بنائے	banā'ī	Made	Verb	Mixed	5	3	4	2.6	1.35
کہتا	kehtā	Says	Verb	Mixed	6	2	4	2.7	1.64
لگی	lagī	Happen	Verb	Mixed	6	2	3	2.6	1.58
کوشش	koshish	Attempt	Verb	Mixed	7	2	4	3.7	1.06
کرنا	karnā	Doing	Verb	Mixed	7	2	4	2.8	1.14
شروع	shurū'	Start	Verb	Mixed	7	2	4	3	1.49
سکتے	sakte	Could	Verb	Mixed	8	2	4	2.7	1.42
بنا	banā	Made	Verb	Mixed	8	2	3	2.4	1.35
کہتے	kehte	Said	Verb	Mixed	8	2	4	2.9	1.73
کرتا	karte	Does	Verb	Mixed	8	2	4	2.5	1.43
رہتے	rehte	Lived	Verb	Mixed	10	2	4	2.8	1.62
سکتا	saktā	Can	Verb	Mixed	11	2	4	3	1.41
کرتی	kartī	Does	Verb	Mixed	12	2	4	2.7	1.57
کیا	kiyā	Did	Verb	Mixed	13	2	3	2.1	1.29
رہے	rahe	Lived	Verb	Mixed	13	2	3	2.8	1.62
چاہیے	cāhī'e	Want	Verb	Mixed	15	3	5	2.6	1.71
کرتے	karte	Does	Verb	Mixed	25	2	4	2.6	1.58
رہا	rahā	Doing	Verb	Mixed	26	2	3	2.9	1.37
گائیں	gā'enn	Singing	Verb	Transparent	1	2	5	2.9	1.20
سوتے	sote	Sleeping	Verb	Transparent	1	2	4	2.3	1.06
سوچنے	socne	Thinking	Verb	Transparent	1	2	5	3.4	0.97
تولو	tolo	Weigh	Verb	Transparent	1	2	4	4.3	0.82
جھانکتیں	jhānktīnn	Peeking	Verb	Transparent	1	2	8	4.8	1.55
رونے	rone	Crying	Verb	Transparent	2	2	4	2.2	1.23
یاد	yād	Remember	Verb	Transparent	2	1	3	2.7	1.42

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
آنا	ānā	Coming	Verb	Transparent	3	2	3	3.1	1.91
بولو	bolo	Speak	Verb	Transparent	3	2	4	2.2	1.48
سونے	sone	Sleeping	Verb	Transparent	3	2	4	2.3	1.34
کھیلنے	khelne	To play	Verb	Transparent	3	2	6	2.3	1.57
خوش	kkhhush	Happy	Verb	Transparent	3	1	3	2.5	1.51
جاؤ	jā'o	Go	Verb	Transparent	3	2	3	2	1.25
کھانے	khāne	Eating	Verb	Transparent	3	2	5	1.9	1.29
رو	ro	Cry	Verb	Transparent	3	1	2	2.1	1.29
ادا	adā	Paid	Verb	Transparent	4	2	3	4.5	1.18
بیٹھے	bethe	Sat	Verb	Transparent	4	2	5	2.7	1.42
سوچ	soc	Thought	Verb	Transparent	4	1	3	3.4	1.35
دیکھے	daikhe	Saw	Verb	Transparent	4	2	5	2.7	1.42
بھاگ	bhāg	Ran	Verb	Transparent	4	1	4	2.2	1.62
کھاتے	khāte	Ate	Verb	Transparent	4	2	5	2.4	1.43
سو	so	Sleep	Verb	Transparent	4	1	2	1.9	1.10
ہوئی	hū'ī	Happen	Verb	Transparent	4	2	3	3.2	1.69
کھینچے	khencne	Pulled	Verb	Transparent	4	2	6	3.9	1.45
بھاگے	bhāge	Ran	Verb	Transparent	4	2	5	2.7	1.42
لی	lī	Took	Verb	Transparent	5	1	2	2.6	1.58
ہوتے	hote	Be	Verb	Transparent	5	2	4	3.6	1.51
ہوگا	hogā	Have to	Verb	Transparent	5	2	4	3.7	1.49
بولی	bolī	Said	Verb	Transparent	6	2	4	3	1.76
جاتی	jātī	Goes	Verb	Transparent	7	2	4	2.7	1.49
جانا	jānā	Go	Verb	Transparent	7	2	4	2.7	1.49
سوچا	socnā	To think	Verb	Transparent	7	2	4	3.6	1.51
لا	lā	Bring	Verb	Transparent	8	1	2	3	1.33
دیکھ	daikh	Look	Verb	Transparent	8	1	4	2.8	1.48

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
دیتے	dete	Gave	Verb	Transparent	8	2	4	2.7	1.64
تھیں	thīnn	Were	Verb	Transparent	8	1	4	3.5	1.43
آتا	ātā	Came	Verb	Transparent	8	2	3	2.9	1.29
جاتے	jāte	Gone	Verb	Transparent	9	2	4	2.8	1.55
بولا	bolā	Said	Verb	Transparent	10	2	4	2.6	1.43
بیٹھ	beth	Sit	Verb	Transparent	11	1	4	2.1	1.29
جا	jā	Go	Verb	Transparent	11	1	2	2.2	1.14
دیکھو	daikho	Look	Verb	Transparent	12	2	5	2.1	1.52
جاتا	jatā	Goes	Verb	Transparent	12	2	4	2.4	1.17
ہوا	huā	Happen	Verb	Transparent	12	2	3	3	1.49
آئے	ā'e	Came	Verb	Transparent	13	2	2	3	1.63
دے	de	Give	Verb	Transparent	13	1	2	3	1.63
گی	gī	Would	Verb	Transparent	14	1	2	3	1.83
جائے	jā'e	Go	Verb	Transparent	14	2	3	3.3	1.89
لے	le	Take	Verb	Transparent	15	1	2	2.9	1.37
ہوئے	hū'e	Were	Verb	Transparent	16	2	3	3.4	1.65
ہوتی	hotī	Would	Verb	Transparent	17	2	4	3	1.41
دیکھا	daikhā	Saw	Verb	Transparent	17	2	5	2.9	1.60
آئی	āī	Came	Verb	Transparent	17	2	2	2.6	1.43
گئی	gaī	Went	Verb	Transparent	18	2	2	2.6	1.51
تھی	thī	Was	Verb	Transparent	19	1	3	2.7	1.34
دو	(do)	Give	Verb	Transparent	41	1	2	2	1.05
ہو	(ho)	Are	Verb	Transparent	60	1	2	2.7	1.42
بتاؤں	batā'unn	Tell	Verb	Mixed	1	3	5	2.9	1.29
خرید	kkhharīd	Buy	Verb	Mixed	1	2	4	3.1	1.29
کہے	kahe	Said	Verb	Mixed	1	2	3	3.6	1.78
لگانے	lagā'e	Apply	Verb	Mixed	1	3	5	3.5	1.43

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
ٹکراتی	ṭakrātī	Colliding	Verb	Mixed	1	3	6	3.9	1.85
نکالا	nikālā	Remove	Verb	Mixed	1	3	5	3.3	1.34
اتار	utār	Remove	Verb	Mixed	1	2	4	2.8	1.55
پڑھی	parhī	Have read	Verb	Mixed	1	2	4	3	1.49
پھدکنے	phudakne	Prancing	Verb	Mixed	1	3	6	4.5	1.51
سوجھا	sūjhā	Thought	Verb	Mixed	1	2	5	5.1	0.99
پہنچنے	puhancne	Reach	Verb	Mixed	1	3	6	4.4	1.51
ہیں	hainn	Are	Verb	Transparent	60	1	3	3.3	1.42
تھا	thā	Was	Verb	Transparent	61	1	3	2.9	1.45
ہے	hai	Is	Verb	Transparent	124	1	2	2.5	1.43
مکمل	mukammal	Complete	Adjective	Opaque	1	3	4	3.3	1.34
سرد	sard	Cold	Adjective	Opaque	1	1	3	3.4	1.35
کل	kul	Total	Adjective	Opaque	3	1	2	4.6	0.70
دس	das	Ten	Adjective	Opaque	3	1	2	2.3	1.16
کچھ	kuch	Some	Adjective	Opaque	4	1	3	2.2	1.40
گرم	garam	Hot	Adjective	Opaque	4	2	3	1.6	1.07
یہ	yeh	This	Adjective	Opaque	10	1	2	2.2	1.40
بہت	buhāt	Many	Adjective	Opaque	10	2	3	2.1	1.37
ہر	har	Every	Adjective	Opaque	13	1	2	2.9	1.37
قومی	qomī	National	Adjective	Transparent	1	2	4	4.6	1.07
یادگار	yādgār	Memorable	Adjective	Transparent	1	2	6	4.9	1.10
آخری	ākhhārī	Last	Adjective	Transparent	2	2	4	2.6	1.65
کالے	kāle	Black	Adjective	Transparent	3	2	4	2.2	1.14
کوئی	ko'ī	Any	Adjective	Transparent	3	2	3	3	1.41
بھاری	bhārī	Heavy	Adjective	Transparent	4	2	5	2.5	1.43
موٹا	moṭā	Fat	Adjective	Transparent	4	2	4	2	1.15
اچھے	ache	Good	Adjective	Transparent	8	2	4	2	1.49

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
چار	cār	Four	Adjective	Transparent	10	1	3	2.2	1.23
خوش	kkhhush	Happy	Adjective	Transparent	13	1	3	1.7	1.06
گندا	gandā	Dirty	Adjective	Mixed	1	2	4	1.8	1.48
آہستہ	āhistāh	Slowly	Adjective	Mixed	1	3	5	2	1.41
خراب	kkhharāb	Ruin	Adjective	Mixed	2	2	4	2.3	1.42
جنگلی	janglī	Savage	Adjective	Mixed	3	2	5	3.8	1.32
بہلا	bhalā	Benefit	Adjective	Mixed	3	2	4	4.2	1.03
ضروری	zarūrī	Necessary	Adjective	Mixed	4	3	5	3.5	1.27
پوری	pūrī	All	Adjective	Mixed	4	2	4	2.7	1.34
نہی	nanhī	Small	Adjective	Mixed	6	2	4	3.7	1.25
بڑے	barē	Big	Adjective	Mixed	8	2	3	2.1	0.99
بڑا	barā	Big	Adjective	Mixed	13	2	3	1.7	1.06
سب	sab	All	Adjective	Opaque	15	1	2	2.3	1.16
لمبی	lambī	Long	Adjective	Mixed	2	2	4	2.3	1.42
سردی	sardī	Winter	Adjective	Mixed	2	2	4	2	1.41
ذرا	zzarā	Some	Adjective	Mixed	2	2	3	3.3	1.16
برا	burā	Bad	Adjective	Mixed	3	2	3	2.1	1.10
کسی	kisī	Any	Adjective	Mixed	3	2	3	2.9	1.37
بری	burī	Bad	Adjective	Mixed	3	2	3	2	0.94
دوسری	dūsri	Others	Adjective	Mixed	3	2	5	2.9	1.37
مشہور	mashūr	Famous	Adjective	Mixed	3	2	5	4.6	1.07
دوسرا	dūsra	Other	Adjective	Mixed	4	2	5	2.8	1.32
اپنے	apne	Own	Adjective	Mixed	4	2	4	2.7	1.49
سفید	safed	White	Adjective	Mixed	5	2	4	2.7	1.25
ہری	harī	Green	Adjective	Mixed	6	2	3	2.9	1.20
پیارے	pyāre	Loved	Adjective	Mixed	8	3	5	2.2	1.23
بڑی	barī	Big	Adjective	Mixed	8	2	3	2.1	1.29

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
اعلیٰ	ālā	Superior	Adjective	Transparent	1	2	4	4.7	0.95
عام	ām	Ordinary	Adjective	Transparent	1	1	3	4.6	0.84
خالی	kkhhālī	Empty	Adjective	Transparent	1	2	4	2.7	1.16
نادان	nādān	Naive	Adjective	Transparent	1	2	5	4.9	0.99
سیدھا	sīdhā	Straight	Adjective	Transparent	1	2	5	3	1.41
سجا									
سجایا	sajā sajāyā	Decorated	Adjective	Transparent	1	3	8	5	0.94
والا	vālā	Someone specific	Adjective	Transparent	3	2	4	2.6	1.58
دوسرے	dūsre	Other	Adjective	Transparent	4	2	5	3	1.49
صاف	ṣāf	Clean	Adjective	Transparent	4	1	3	2.2	1.23
سات	sāt	Seven	Adjective	Transparent	4	1	3	2.3	1.16
تین	tīn	Three	Adjective	Transparent	4	1	3	1.8	1.03
ساتھ	sāth	Together	Adjective	Transparent	4	1	4	2.6	1.17
چھوٹے	choṭe	Small	Adjective	Transparent	4	2	5	2.3	1.34
خوب	kkhhūb	Well	Adjective	Transparent	4	1	3	3.6	1.43
سارا	sārā	All	Adjective	Transparent	4	2	4	2.4	1.26
سارے	sāre	All	Adjective	Transparent	5	2	4	2.8	1.32
کئی	ka'ī	Many	Adjective	Transparent	5	2	2	3.9	1.37
خاموش	kkhhāmosh	Silent	Adjective	Transparent	5	2	5	3.1	1.29
والے	vāle	Doers	Adjective	Transparent	5	2	4	3.3	1.34
ساری	sār	All	Adjective	Transparent	6	2	4	2.6	1.35
زور	zor	Pressure	Adjective	Transparent	6	1	3	2.5	1.27
اچھی	achī	Good	Adjective	Transparent	6	2	4	2.1	1.29
چھوٹا	choṭā	Small	Adjective	Transparent	6	2	5	2.1	1.29
بوڑھی	būrhī	Old	Adjective	Transparent	6	2	5	3.8	1.23
لال	lāl	Red	Adjective	Transparent	7	1	3	2.5	0.97
نئے	na'e	New	Adjective	Transparent	8	2	2	2.3	1.34
دو	do	Give	Adjective	Transparent	8	1	2	2.1	1.29

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
خوبصورت	kkhhūbṣūrat	Beautiful	Adjective	Transparent	9	2	7	3.8	1.55
اچھا	achā	Okay	Adjective	Transparent	9	2	4	2	1.15
افسوس	afsos	Sorrow	Adjective	Mixed	1	2	5	4.5	1.27
ایک دم	ek dam	At Once	Adjective	Mixed	1	2	5	4.4	1.35
علیحدہ	alehdāh	Separate	Adjective	Mixed	1	3	6	4.9	1.37
جسمانی	jismānī	Bodily	Adjective	Mixed	1	3	6	5.5	0.97
لمبے	lambe	Tall	Adjective	Mixed	1	2	4	3.2	1.40
مٹی	munnī	Little	Adjective	Mixed	1	2	3	2.8	1.14
تازہ	tāzāh	Fresh	Adjective	Mixed	1	2	4	3.9	1.10
ایک	eik	One	Adjective	Transparent	64	1	3	1.8	1.03
کم	kam	Few	Adverb	Opaque	1	1	2	2.3	1.16
یوں	yūnn	Way	Adverb	Opaque	2	1	3	3.5	0.97
کل	kal	Tomorrow	Adverb	Opaque	5	1	2	2.2	1.23
دور	dūr	Far	Adverb	Opaque	6	1	3	2.1	1.20
تک	tak	Till	Adverb	Opaque	8	1	2	3.4	1.43
سب	sab	All	Adverb	Opaque	9	1	2	2.3	1.34
طرف	ttaraf	Direction	Adverb	Opaque	14	2	3	3.4	0.97
طرح	ttarāh	Like	Adverb	Opaque	18	2	3	3.2	1.40
جب	jab	Then	Adverb	Opaque	19	1	2	2.8	1.23
اب	ab	Now	Adverb	Opaque	23	1	2	2.5	1.35
نہ	nāh	No	Adverb	Opaque	24	1	2	1.9	1.45
دوران	daurān	During	Adverb	Transparent	1	2	4	4.9	1.37
پیچھے	pīche	Behind	Adverb	Transparent	5	2	5	2.6	1.84
نیچے	nīce	Down	Adverb	Transparent	6	2	4	2.4	1.71
بعد	bā'd	Later	Adverb	Transparent	14	1	3	2.8	1.40
آج	āj	Today	Adverb	Transparent	20	1	2	2.3	1.25
کتنے	kitne	How many	Adverb	Mixed	1	2	4	2.2	1.40

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
کہاں	kahānn	Where	Adverb	Mixed	5	2	4	2.2	1.32
واپس	vāpis	Return	Adverb	Mixed	6	2	4	2.7	1.25
ابھی	abhī	Now	Adverb	Mixed	8	2	4	2.4	1.43
کبھی	kabhī	Ever	Adverb	Mixed	9	2	4	3	1.33
نہیں	nahīnn	No	Adverb	Mixed	44	2	4	1.6	1.35
کچھ	kuch	Some	Adverb	Opaque	10	1	3	2.9	1.37
بہت	buhāt	Much	Adverb	Opaque	15	2	3	2.2	1.32
آپس	āpas	Between	Adverb	Mixed	2	2	3	3.6	1.35
ہمیشہ	hameshāh	Always	Adverb	Mixed	2	3	5	3.8	1.14
ادھر	idhar	Here	Adverb	Mixed	5	2	4	2.5	1.43
ادھر	udhar	Here	Adverb	Mixed	6	2	4	2.8	1.48
پہلے	pehle	First	Adverb	Mixed	7	2	4	2.5	1.35
یہاں	yahānn	Here	Adverb	Mixed	10	2	4	2.6	1.65
وہاں	vahānn	There	Adverb	Mixed	10	2	4	2.8	1.40
ایک	eik	One	Adverb	Transparent	1	1	3	1.8	1.32
بائیں	bā'enn	Left	Adverb	Transparent	1	2	5	4.3	1.57
زور	zor	Daily	Adverb	Transparent	1	1	3	2.5	1.51
جیسے	jaise	Like	Adverb	Transparent	5	2	4	3.7	1.16
کیسے	kaise	How	Adverb	Transparent	5	2	4	3.4	1.17
سامنے	sāmne	Front	Adverb	Transparent	5	2	5	3	1.56
دیر	dair	Late	Adverb	Transparent	7	1	3	2.9	1.29
ساتھ	sāth	Together	Adverb	Transparent	11	1	4	3	1.25
پاس	pās	Near	Adverb	Transparent	16	1	3	2.4	1.35
موجود	mojūd	Present	Adverb	Mixed	1	2	5	4.8	0.92
کہ	keh	Say	Conjunction	Opaque	47	1	2	3.2	1.40
ہی	hī	Is	Conjunction	Transparent	41	1	2	3.6	1.35
بھی	bhī	Also	Conjunction	Transparent	65	1	3	3.1	1.29

Urdu	LoC Transliteration ^a	Translation	Word Class	Transparency Category	Written Frequency ^b	Syllable Count	Letter Count	AoA ^c	SD
لیکن	lekin	But	Conjunction	Mixed	9	2	4	3.1	1.20
اور	aur	And	Conjunction	Transparent	173	1	3	2.3	1.06
تو	to	Then	Conjunction	Transparent	84	1	2	3.1	1.60
سی	sī	Like	Postposition	Transparent	6	1	2	3.8	1.23
نے	ne	Has	Postposition	Transparent	132	1	2	2.5	1.27
میں	menn	In	Postposition	Transparent	138	1	3	2.2	1.14
کی	kī	Is	Postposition	Transparent	139	1	2	2.7	1.16
سے	se	From	Postposition	Transparent	154	1	2	2.5	1.27
کے	ke	Is	Postposition	Transparent	168	1	2	2.6	1.17

Note. ^a: The American Library of Congress' Urdu Romanization Table was used as the transliteration scheme to represent

vowelization. ^b: This estimate is based on the frequency calculations performed on the sentence corpus derived from first to third-

grade Urdu textbooks (N = 9). ^c: This column contains the mean AoA ratings from the ten participants included in the present study.

Appendix O: Sample Spelling and Reading Word List

Urdu	LoC Transliteration	Translation	Spelling Type	Written Frequency	Syllable Count	Letter Count	Word Class	Mean AoA Rating	SD AoA Rating
تھے	the	Were	Transparent	33	1	3	Verb	2.7	1.4
نئے	na'e	New	Transparent	8	2	2	Adjective	2.3	1.4
چاند	cānd	Moon	Transparent	7	1	4	Noun	1.9	1.3
جھانکتیں	jhānktīnn	Peeking	Transparent	1	2	8	Verb	4.8	1.6
نیچے	nīce	Down	Transparent	6	2	4	Adverb	2.4	1.7
کھانے	khāne	Foods	Transparent	4	2	5	Noun	2.6	1.6
سال	sāl	Year	Transparent	4	1	3	Noun	2.6	1.7
پودا	podā	Plant	Transparent	4	2	4	Noun	2.6	1.3
بولے	bole	Said	Transparent	6	2	4	Verb	2.7	1.7
اؤ	ā'o	Come	Transparent	6	2	2	Verb	1.9	1.4
امی	ammī	Mother	Mixed	30	2	3	Noun	1.5	1.1
بڑے	baṛe	Big	Mixed	8	2	3	Adjective	2.1	1
بلی	billī	Cat	Mixed	7	2	3	Noun	1.5	1.1
خلاف	kkhhilāf	Compete against	Mixed	1	2	4	Verb	5.1	0.9
ننھی	nanhī	Small	Mixed	6	2	4	Adjective	3.7	1.3
جماعت	jamāt	Class	Mixed	4	1	5	Noun	4.2	1.4
دانت	dānt	Tooth	Mixed	4	1	4	Noun	1.8	1.3
سلام	salām	Greeting	Mixed	4	2	4	Noun	1.9	1.4

Urdu	LoC Transliteration	Translation	Spelling Type	Written Frequency	Syllable Count	Letter Count	Word Class	Mean AoA Rating	SD AoA Rating
پڑھنا	paṛhnā	Reading	Mixed	6	2	5	Verb	2.7	1.3
رہنے	rehne	Living	Mixed	6	2	4	Verb	2.9	1.4
دن	din	Day	Opaque	30	1	2	Noun	2.1	1.2
بہت	buhāt	Much	Opaque	10	2	3	Adjective	2.1	1.4
جنگل	jangal	Jungle	Opaque	7	2	4	Noun	2.8	1.3
شرمندہ	sharmindāh	Ashamed	Opaque	1	3	6	Verb	4.4	1.3
دور	dūr	Far	Opaque	6	1	3	Adverb	2.1	1.2
کمرہ	kamrāh	Room	Opaque	4	2	4	Noun	1.9	1.5
منہ	mūnh	Mouth	Opaque	4	1	3	Noun	2.1	1.4
پھل	phal	Fruit	Opaque	4	1	3	Noun	2.7	1.7
ختم	kkhhatam	End	Opaque	6	2	3	Verb	2.1	1.4
چل	cal	Walk	Opaque	6	1	2	Verb	2.2	1.3

Note. ^a: The American Library of Congress' Urdu Romanization Table was used as the transliteration scheme to represent

vowelization. ^b: This estimate is based on the frequency calculations performed on the sentence corpus derived from first to third-

grade Urdu textbooks (N = 9). ^c: This column contains the mean AoA ratings from the ten participants included in the present study.

