Corpus Based Urdu Lexicon Development

Madiha Ijaz
Centre for Research in Urdu Language Processing
National University of Computer and Emerging Sciences
madiha.ijaz@nu.edu.pk

Sarmad Hussain
Centre for Research in Urdu Language Processing
National University of Computer and Emerging Sciences
sarmad.hussain@nu.edu.pk

Abstract

The paper discusses various phases in Urdu lexicon development from corpus. First the issues related with Urdu orthography such as optional vocalic content, Unicode variations, name recognition, spelling variation etc. have been described, then corpus acquisition, corpus cleaning, tokenization etc has been discussed and finally Urdu lexicon development i.e. POS tags, features, lemmas, phonemic transcription and the format of the lexicon has been discussed .

1. Introduction

The project focuses on the creation of an Urdu lexicon needed for speech-to-speech translation components i.e. flexible vocabulary speech recognition, high quality text-to-speech synthesis and speech centered translation following the guidelines of LC-STAR II (http://www.lc-star.org/).

A broad range of common domains and domains for proper names was chosen to be collected from electronically available resources and print media as well. A corpus of 19.3 million was collected and then a large lexicon was created based on that corpus listing detailed grammatical, morphological, and phonetic information suited for flexible vocabulary speech recognition and high quality speech synthesis.

This paper deals with issues regarding Urdu orthography, corpus development (e.g. corpus acquisition, pre-processing, tokenization, cleaning e.g. typos, name recognition etc) and then finally lexicon development for common words.

2. Urdu Orthography

Urdu is written in Arabic script in Nastaleeq style using an extended Arabic character set. The character set includes basic and secondary letters, aerab (or diacritical marks), punctuation marks and special symbols [1]. Urdu support in Unicode is given in Arabic Script block. Further details regarding Urdu letters, diacritics, numbers, special symbols and Unicode variation are described ahead.

Urdu text comprises of the alphabets as show in Figure 1. [9].

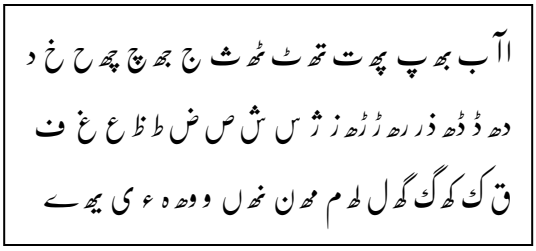

Figure 1: Urdu alphabet

Diacritics described in Table 1 exist in Urdu text [10, 11].

Diacritic	Symbol (Unicode)	Example	IPA
Zabar (Fatah)	(E064) ◌َ	لَب	ləb
Fatah Majhool	(E064) ◌َ	زَہر	zɛhɛr
Zair (Kasra)	(0650) ◌ِ	دِل	d̪ɪl
Kasra Majhool	(0650) ◌ِ	اِہتمام	eh.t̪e.mɑm
Paish (Zamma)	(F064) ◌ُ	گُل	gʊl
Zamma Majhool	(F064) ◌ُ	عُہدہ	oh.d̪ɑ
Sakoon (Jazm)	(0652) ◌ْ	سبْز	səbz
Tashdeed (Shad)	(0651) ◌ّ	ڈبّا	ɖəb.bɑ
Tanween	(B064) ◌ً	فوراً	fɔ.rən
Khara Zabar	(0670) ◌ٰ	عیسٰی	i.sɑ
Elaamat-e-Ghunna	(0658) ◌ٗ	جنگ	ʤəŋ

Table 1: Diacritics in Urdu

Digits from 0 to 9 are represented in Urdu are shown in Figure 2.

Figure 2: Urdu digits

Special symbols that may occur in Urdu text are shown in Figure 3. Their details can be found in Arabic script block in Unicode (http://www.unicode.org/charts/).

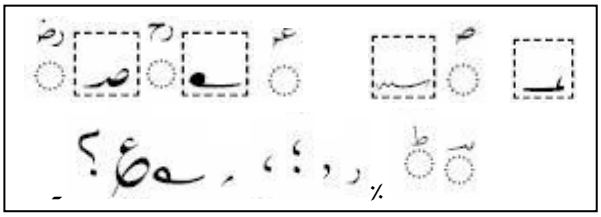

Figure 3: Urdu special symbols

The following sections discuss some issues that arise due to Unicode and Urdu orthography.

2.1. Unicode Variations

The Unicode standard provides almost complete support for Urdu. However, there are a few discrepancies, for example in Unicode, the character Hamza (ء) is declared a non-joiner (i.e. it does not connect with the letter following it). However, in Urdu language words e.g., قائل / kɑ.ɪl / require a Hamza to be joined with the characters following it. For such words Unicode provides a separate character ئ (joining Hamza) instead of ء. Similarly, the character Bari Yay (ے) is also considered a non-joiner in Unicode (with the following character), but the word بے کار /be.kɑr/ (adjective: "useless"). is also commonly written in Urdu as بیکار /be.kɑr/. To write the latter, we need to put ی instead of ے so that the Yay joins with Kaaf ک. These issues still need to be resolved with the Unicode standard for complete Urdu support.

Some characters like ک، ہ، ی etc. have more than one Unicode value in different keyboards. Such characters are replaced by one standard character (depending on their position within the word) in order to normalize them before any processing is done on them. Appendix A provides the currently handled characters for normalization.

2.2. Optional vocalic content

Urdu is normally written only with letters, diacritics being optional. However, the letters represent just the consonantal content of the string and in some cases (under-specified) vocalic content. The vocalic content may be optionally or completely specified by using diacritics with the letters [1]. Every word has a correct set of diacritics, however, it can be written with or with out any diacritics at all, therefore, completely or partially omitting the diacritics of a word is permitted.

In certain cases, two different words (with different pronunciations) may have exactly the same form if the diacritics are removed, but even in that case writing words without diacritics is permitted. One such example is given below:

تَیر /ṭær/ (swim)

تیر /ṭir/ (arrow)

However, there are exceptions to this general behavior, certain words in Urdu require minimal diacritics without which they are considered incomplete and cannot be correctly read or pronounced. Some of these words are shown in Table 2.

Actual word	English translation	With diacritics (correct)	Without diacritics (incorrect)
/ɑ.lɑ/	High quality	اعلیٰ	اعلی
		/ɑ.lɑ/	/ɑ.li/
/ṭəq.ri.bən/	almost	تقریباً	تقریبا
		/ṭəq.ri.bən/	/ṭəq.ri.bɑ/

Table 2: Some Urdu words that require diacritics

2.3. Proper name identification and spelling variation

In Urdu, there is no concept of capitalization. Proper names cannot be identified through script analysis and there is no 'Urdu specific' algorithm for named entity tagging.

Spelling variations are quite common in Urdu. The main reason for these variations is that there are many homophone characters (different letters representing the same phoneme) in Urdu. Also people tend to confuse different homophones for each other, so, as a result, incorrect spelling of words having homophones becomes quite common. For example, "ز" and "ذ" are homophone characters and are very frequently confused with each other. The word "پذیر" /pə.zir/ is commonly written in news papers, books and some dictionaries with letter "ز" instead of "ذ", which is correct.

Urdu collation sequence is fully standardized. In Urdu, three levels of sorting are required for letters,

diacritics and special symbols. The complete table of collation element of Urdu is given in [8].

3. Urdu Corpus development

A large amount of text is needed in order to build the corpus which is used for lexicon extraction. Electronically available resources are the most suitable for collection of text but unfortunately it is not easy to collect Urdu text as first of all there is no publicly available large amount of Urdu text and secondly most of the websites containing Urdu text display it in graphics i.e. gif format which makes it unfit to be used in any text based application [5, 6].

3.1. Corpus acquisition

The data was gathered from a broad range of domains mentioned in Table 3 keeping in view the end user perspective.

Domains C1. Sports/Games	Sub domains C1.1.Sports (special events)
C2. News	C2.1. Local and international affairs C2.2. Editorials and opinions
C3. Finance	C3.1. Business, domestic and foreign market
C4. Culture/Entertainment	C4.1. Music, theatre, exhibitions, review articles on literature C4.2. Travel / tourism
C5. Consumer Information	C5.1. Health C5.2. Popular science C5.3. Consumer technology
C6. Personal communications	C6.1. Emails, online discussions, editorials, e-zines

Table 3: Corpus domains

It was ensured while collecting text from the above mentioned domains that [14]

1. Each domain was represented by at least 1 million tokens.
2. The cut-off date for all corpora used was 1990 as it has been shown that corpora structure and time of appearance of corpora has a large impact on the extracted word lists.
3. Data from chat rooms was not included

Text was collected from two news websites i.e. Jang (www.jang.com.pk) and BBC (http://www.bbc.co.uk/urdu/) and it was made sure that the data collected was not older than 2002.

Apart from the news websites text was also collected from books and magazines related to required domains and the data collected from these sources was not older than 1990.

3.2. Pre-processing

Data that was gathered had different character encoding schemes and before doing any further processing it was to be converted to a standard character encoding scheme i.e. UTF-16.

Data gathered from news websites was in HTML format so it was converted to UTF-16. Similarly data gathered from magazines was in inpage format and hence it was also converted to UTF-16.

3.3. Tokenization

For the development of Urdu lexicon, words are derived from the corpus by assuming white spaces (tab, space character, carriage return and linefeed) and punctuation marks (hyphen, semicolon, backslash, caret, vertical line, Arabic ornamental left parenthesis and right parenthesis, comma, apostrophe, exclamation mark, Arabic semicolon, colon, quotation mark, Arabic starting and ending quotes, Arabic question mark), special symbols (dollar, percent, ampersand, asterisk, plus), digits (0-9 and ۰-۹) and English alphabets (A-Z and a-z) as word boundaries. Thus words like "خوش مزاج" /xʊʃ.mɪ.zɑʤ/ (adjective: "pleasant"), erroneously get split into two separate words "خوش" /xʊʃ/ (adjective: "happy") and "مزاج" /mɪ.zɑʤ/ (noun: "temperament"). Also words like "ذمہ داری" /zɪm.mɑ.ɖɑ.ri/ (noun: "responsibility") erroneously get split into "ذمہ" /zɪm.mɑ/ (noun: "responsibility") and "داری" /ɖɑ.ri/ (non-word suffix) [13]. In order to cater to words like "ذمہ داری" the tokenizer was modified and a list of prefixes and suffixes was used to determine that whether the token under consideration is an affix or not and if it was an affix then depending on whether it is prefix or suffix, the tokenizer picked the next and previous word respectively e.g. "داری" is a suffix so in this case it picked the previous word etc.

Description of procedure of word list extraction is as follow

- The Html and Inpage files were converted to Unicode text files (UTF-16).
- The text in those files was tokenized on characters like white space, punctuation marks, special symbols etc.
- Some characters like ک، ہ، ی etc. have more than one Unicode values in different keyboards. Such characters were replaced by

one standard character (depending on their position within the word) in order to normalize them before any processing was done on them.

- Diacritics were removed from the word list e.g. تَیر /ʈær/ (swim) and تِیر /ʈir/ (arrow) were both mapped to تیر.
- Word frequencies were updated.
- The tokenization based on space does not completely identify the words from the corpus correctly. The output needs to be reviewed in order to remove non-words which may occur due to erroneous output of tokenizer or due to typing errors. Proper names, typos etc were removed from the word list manually and the words that were written without space were separated (space insertion problem) e.g. the token طاہرکوکھلادیا comprises of four words, طاہر /ʈa.h ɪ r/ (proper name and an adjective), کو /ko/ (case marker), کھلا / k^h ɪ .la/ (verb) and دیا / ɖ ɪ .ja/ (verb). Word frequencies were updated after space insertion.

When non-words were analyzed, it was revealed that most of them were affixes apart from proper names and typos. Hence a list of valid Urdu affixes was developed and tokenizer was modified to pick next or previous word if it encountered a prefix or suffix respectively and frequencies were adjusted accordingly e.g. "ذمہ داری" /zɪm.mɑ.ɖɑ.ri/ (noun: "responsibility") is a word with affix "داری" if its frequency was 10 then 10 was subtracted from the frequency of "ذمہ" and from the frequency of "داری " as well.

4. Urdu Lexicon Development

Urdu lexicon development involved decisions regarding part-of-speech tags and their respective features, lemmas, transcription and lexicon format.

4.1. POS tags

Since the lexicon is to be used for speech-to-speech translation components, a high-level POS tag set covering main categories is adequate.

POS tags decided for Urdu lexicon development are as follow

1. Noun.
2. Verb.
3. Adjective.
4. Adverb.
5. Numerals.
6. Post positions.
7. Conjunctions.
8. Pronouns
9. Auxiliaries.
10. Case marker.
11. Harf.

All the recognizable POS tags of the word were identified, regardless of the context in which the word is used in the corpus. The details of the POS tags are given in Appendix B.

Two of the above listed POS are particular to Urdu. Their details are given below:

4.1.1 Harf: Harf is a word which is not meaningful unless used with other words to give meaning [10]. This category includes words like اے /æ/, اوہو /o ho/, واہ /vɑ/, پر /pər/ etc.

4.1.2 Case markers: Case markers are a special word class in Urdu. In some languages case marking is a morphological process, but in Urdu case markers are written with a space. Therefore they are considered as a separate word and are assigned a separate POS tag. There are mainly three case markers: ergative, نے /ne/, dative/accusative, کو /ko/ and genitive, کا /ka/. Sometimes سے /se/ is also included in this category as being an instrumentative case marker. Some grammarians [10] consider case markers as a subset of Haroof[1], but due to their distinct role of case marking (agent/patient role etc), it is better to separate them from other Haroof.

Urdu lexicon does not include respect feature. It also does not include separate POS tag for the light verb and aspectual auxiliary because both light verbs and aspectual auxiliaries have the same surface forms as a verb in the language. Once the wordlists are prepared from the corpus the context of the word is lost. In order to identify a word as a light verb or aspectual auxiliary it is essential to know whether it occurred in the corpus in combination with some other word or as an independent verb.

4.2. Lemmas

Lemma is a canonical form of a word. Morphological forms considered as lemma according to well-known guidelines of Urdu are the following:

1. Common noun: singular, nominative with no respect

1 Plural of Harf

2. Verb: masculine, singular, nominative, infinitive
3. Adjective: masculine, singular, nominative if the word ends at vowel otherwise the word itself.
4. Adverb: usually same as the word itself
5. Numeral: in case of cardinal it is same as the word but in case of ordinal it is masculine, singular, nominative.
6. Post-position (ad positions): same as the word itself
7. Conjunction: same as the word itself
8. Pronoun: masculine, singular, nominative if the word ends at vowel otherwise the word itself.
9. Auxiliary verb: masculine, singular, nominative, infinitive
10. Case marker: same as the word itself except the genitive case markers, in that case it is masculine, singular, nominative.
11. Harf: same as the word itself

4.3. Phonemic Transcription

Urdu phonemic inventory consists of 44 consonants, 8 long oral vowels, 7 long nasal vowels, 3 short vowels and numerous diphthongs (set of Urdu diphthongs is still under analysis). This phonemic inventory is presented in Table 4 [3]. The italicized phonemes are those whose existence is still controversial. These are mainly aspirated versions of nasal stops, lateral and retroflex. Different studies conducted at CRULP to explore the existence of these aspirated phonemes show that they were once part of the Urdu phonemic inventory and are now either already extinct or slowly dying out [4].

(a)

p	b	p^h	b^h	m	m^h	
t̪	d̪	$t̪^h$	$d̪^h$	n	n^h	
ʈ	ɖ	$ʈ^h$	$ɖ^h$			
k	g	k^h	g^h	ŋ	$ŋ^h$	
tʃ	dʒ	$tʃ^h$	$dʒ^h$	q	ʔ	
f	v	s	z			
ʃ	ʒ	x	ɣ	h		
r	r^h	ɽ	$ɽ^h$	j	l	l^h

(b)

i	e	ɛ	æ
u	o	ɔ	ɑ
ɪ	ʊ	ə	
ĩ	ẽ	æ̃	
ũ	õ	ɔ̃	ɑ̃

Table 4: Urdu Phonemic Inventory [1]

Letter to sound rules are mostly regular. Consonantal letters have a one-to-one mapping with consonantal sounds. There are no letter-clusters to single phoneme mappings as in case of English (ph for example). The mapping of diacritics and some basic letters to vocalic sounds is generally determined by their context. For a detailed discussion on 'grapheme to phoneme mapping rules', refer to "Letter to Sound Rules for Urdu Text to Speech System" [1].

Based on the results of Wells [12] and Hussain [1], the table given in Appendix C was derived and used for Urdu lexicon development. It shows the mapping of Urdu letters to IPA [1] and the mapping of that respective IPA to SAMPA [12].

4.4. Lexicon Format

The XML format given in LC-STAR D2.1 document [7] was followed. A sample containing some of the lexicon entries is given below

```
<ENTRYGROUP orthography="مردوں">
        <ENTRY>
          <NOM class="common" case="oblique"
          number="plural" gender="masculine"/>
           <LEMMA>مرد</LEMMA>
            <PHONETIC" m @ r - d_d o~</PHONETIC>
        </ENTRY>
        <ENTRY>
           <NOM class="common" case="oblique"
          number="plural" gender="invariant"/>
            <LEMMA>مردہ</LEMMA>
            <PHONETIC" m U r - d_d o~</PHONETIC>
           </ENTRY>
</ENTRYGROUP>
<ENTRYGROUP orthography="ایکٹیو">
        <ENTRY>
            <ADJ case="invariant" number="invariant"
            gender="invariant"/>
            <LEMMA>ایکٹیو</LEMMA>
            <PHONETIC>{ k - " t' i v</PHONETIC>
        </ENTRY>
</ENTRYGROUP>
<ENTRYGROUP orthography="اٹھارہ">
        <ENTRY>
          <NUM case="invariant" number="invariant"
          type="cardinal" gender="invariant"/>
       <LEMMA>اٹھارہ</LEMMA>
       <PHONETIC>@ t' - " t'_h A - r A</PHONETIC>
        </ENTRY>
</ENTRYGROUP>
<ENTRYGROUP orthography="سجائے">
        <ENTRY>
        <VER number="plural" person="invariant"
            gender="masculine"/>
          <LEMMA>سجانا</LEMMA>
          <PHONETIC>s @ - " d_Z A -e</PHONETIC>
```

</ENTRY>
</ENTRYGROUP>

5. Results

Domain wise corpus size distribution is given in Table 5.

Domains	Raw Corpora	
	Size	Distinct words
C1. Sports/Games	1,666,304	23,118
C2. News	8,957,259	67,365
C3. Finance	1,162,019	17,024
C4. Culture/Entertainment	3,845,117	59,214
C5. Consumer Information	1,980,723	34,151
C6. Personal communications	1,685,424	30,469
Total	19,296,846	104,341

Table 5: Domain wise corpus size distribution (raw)

The big difference in total number of words and unique words shows that closed class words are used extensively in written Urdu text [2] which is also apparent from the 100 high frequency words given in Appendix D.

Domain wise corpus size distribution after cleaning is given in Table 6.

Domains	Clean Corpora	
	Size	Distinct Words
C1. Sports/Games	1,529,066	15,354
C2. News	8,425,990	36,009
C3. Finance	1,123,787	13,349
C4. Culture/Entertainment	3,667,688	34,221
C5. Consumer Information	1,929,732	24,722
C6. Personal communications	1,632,353	23,409
Total	18,308,616	50,365

Table 6: Domain wise corpus size distribution (clean)

Hence a total of 50,365 words were obtained from a corpus of 19.3 million that had 104,341 orthographically unique words.

The difference in the number of words in the raw corpus and the clean corpus is due to the fact that proper names and typos were removed from the wordlist and in case of spelling variations only standard orthography of a word was kept and rest were discarded.

6. Conclusion

Various stages for lexicon development from corpus have been discussed in this paper. Also the approaches followed in those stages have been described and among which two of these approaches need further exploration and consideration.

First, there are issues of space insertion and deletion in Urdu text so the word boundary detection in Urdu requires a more sophisticated algorithm than simple tokenization on spaces.

Secondly, once the wordlists are prepared from the corpus the context of the word is lost and context is most of the time very important in determining the POS and actual pronunciation of a word e.g. سونا /so.na/ could be the verb "to sleep", noun "gold" and adjective "deserted", so while determining POS and pronunciation context should be considered. Hence some other approach should be developed and instead of losing context, it should be kept and in this way the lexicon developed will contain the recent and frequent usage of a word not the old or deprecated one.

References

[1] S. Hussain, "Letter to Sound Rules for Urdu Text to Speech System", in Proceedings of Workshop on *Computational Approaches to Arabic Script-based Languages*, COLING 2004, Geneva, Switzerland (2004).

[2] Humayoun, "Urdu morphology, orthography and lexicon extraction", master's thesis, Chalmers University of Technology, Sweden, 2006.

[3] H. Kabir, A. M. Saleem, "Speech Assessment Methods Phonetic Alphabet (SAMPA): Analysis of Urdu". CRULP Annual Student Report published in *Akhbar-e-Urdu*, April-May 2002, National Language Authority, Islamabad, Pakistan.

[4] A. M. Saleem, H. Kabir, M.K. Riaz, M.M. Rafique, N. Khalid, and S.R. Shahid. "Urdu Consonantal and Vocalic Sounds". CRULP Annual Student Report published in *Akhbar-e-Urdu*, April-May 2002, National Language Authority, Islamabad, Pakistan.

[5] Riaz and Becker, "A study in Urdu corpus construction". The 3rd Workshop on *Asian Language Resources and International Standardization*, COLING 2002.

[6] Baker, JP, Hardie, A, McEnery, AM and Jayaram, BD (2003) "Corpus data for South Asian language processing." In: Proceedings of the EACL Workshop on South Asian Languages, Budapest

[7] Giulio Maltese, Chiara Montecchio (IBM), LC-STAR Deliverables D2.1, Post final version May 2004 (http://www.lc-star.org/).

[8] S. Hussain et. al. "Urdu Encoding and Collation Sequence for Localization", 2004.

[9] National Language Authority (Cabinet Division), Government of Pakistan (www.nla.gov.pk)

[10] Maulvi Abdul Haq, قوائد اردو *(Qawaid-i-Urdu),* Lahore Academy, Pakistan.

[11] *Urdu Lughat*, Taraqqi, Urdu Board Karachi, Pakistan.

[12] Wells, J.C., "Computer-coding the IPA: a proposed extension of SAMPA", University College, London, 1995 (www.phon.ucl.ac.uk/home/sampa/x-sampa.htm)

[13] Hardie, "The computational analysis of morphosyntactic categories in Urdu", PhD. thesis, Lancaster, 2003.

[14] Ziegenhein, U., et al. (2003): "Specification of corpora and word lists in 12 languages", LC-Star deliverable D1.1 (http://www.lc-star.org/).

Appendix A

Character with variation	Standard characters		
	Start position	Intermediary position	End position
ه (U0647)	ھ (U06BE)	ھ (U06BE)	ه (U06C1)
ه (U06D5)	ه (U06C1)	ه (U06C1)	ه (U06C1)
ة (U0629)	ة (U06C3)	ة (U06C3)	ة (U06C3)
ى (U0649)	ى (U06CC)	ى (U06CC)	ى (U06CC)
ي (U064A)	ى (U06CC)	ى (U06CC)	ى (U06CC)
؟ (U061F)	? (U003F)	? (U003F)	? (U003F)
٭ (U066D)	* (U002A)	* (U002A)	* (U002A)
آ (U0627+U0653)	آ (U0622)	آ (U0622)	آ (U0622)
ئ (U06D3)	ئـ (U06D2+U0626)	ئـ (U06D2+U0626)	ئـ (U06D2+U0626)
ۓ (U06D2+U0654)	ئـ (U06D2+U0626)	ئـ (U06D2+U0626)	ئـ (U06D2+U0626)
ؤ (U0648+U0654)	ؤ (U0624)	ؤ (U0624)	ؤ (U0624)
أ (U0627+U0654)	أ (U0623)	أ (U0623)	أ (U0623)
ۂ (U06C1+U0654)	ۂ (U06C2)	ۂ (U06C2)	ۂ (U06C2)

Appendix B

POS	Features	Possible values
Verb (VER)	Number	singular, plural, invariant
	Gender	Masculine, feminine, invariant
	Person	1, 2, 3, not_1, invariant
	Case	nominative, oblique, vocative, invariant
Adjective (ADJ)	Gender	Masculine, feminine, invariant
	Number	singular, plural, invariant
	Case	nominative, oblique, vocative, invariant
Common Noun (NOM)	Class	common
	Gender	masculine, feminine, invariant
	Number	singular, plural, invariant
	Case	nominative, oblique, vocative, invariant
Adverb (ADV)	None	
Numeral (NUM)	Gender	masculine, feminine, invariant
	Number	singular, plural, invariant
	Case	nominative, oblique, vocative, invariant
	Type	Cardinal, ordinal, ratio
Conjunction (CON)	Type	Subordinating, coordinating
Auxiliary (AUX)	Number	singular, plural, invariant
	Gender	Masculine, feminine, invariant
Post-positions (ADP)	None	
Case Markers (CM)	Case	Ergative, instrumental, dative/accusative, genitive
	Gender	masculine, feminine, invariant
	Number	singular, plural, invariant
Harf (HAR)	*None*	
Pronoun (PRO)	Gender	masculine, feminine, invariant
	Number	singular, plural, invariant
	Case	nominative, oblique, dative, accusative, genitive, invariant
	Type	personal, demonstrative, indefinite, interrogative, relative

Appendix C

Urdu	IPA	SAMPA
Consonants		
پ	p	p
ب	b	b
پھ	p^h	p_h
بھ	b^h	b_h
م	m	m
مھ	m^h	m_h
ت, ط	ʈ	t_d
تھ	$ʈ^h$	t_d_h
د	ɖ	d_d
دھ	$ɖ^h$	d_d_h
ن	n	n
نھ	n^h	n_h
نگ	ŋ	N
ٹ	ʈ	t'
ڈ	ɖ	d'
ٹھ	$ʈ^h$	t'_h
ڈھ	$ɖ^h$	d'_h
ک	k	k
گ	g	g
کھ	k^h	k_h
گھ	g^h	g_h
ق	q	q
ع	ʔ	?
ف	f	f
و	v	v
س, ص, ث	s	s
ذ, ض, ظ, ز	z	z
ش	ʃ	S
ژ	ʒ	Z
غ	ɣ	7
خ	x	x
ہ ، ح	h	h
چ	ʧ	t_S
چھ	$ʧ^h$	t_S_h
ج	ʤ	d_Z
جھ	$ʤ^h$	d_Z_h
ر	r	r

رھ	r^h	r_h
ڑ	ɽ	r'
ڑھ	$ɽ^h$	r'_h
ی	j	j
ل	l	l
لھ	l^h	l_h
وھ	v^h	v_h
یہ	j^h	j_h
Vowels		
ی	i	i
ے	e	e
ـَے	æ	{
ـُو	u	u
و	o	o
ـَو	ɔ	O
ا ، آ	ɑ	A
ـِ	ɪ	I
	ɛ	E
ـُ	ʊ	U
ء ، ـَ	ə	@
ـِیں	ĩ	i~
یں	ẽ	e~
ـَیں	æ̃	{~
ـُوں	ũ	u~
وں	õ	o~
اں	ɑ̃	A~
ـَوں	ɔ̃	O~
Special symbols		
Syllable boundary	.	-
Stress marker	'	"
Word boundary	#	#

Appendix D

100 high frequency words are given below

Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
کے	743949	تھا	61762	گے	36725	ہوا	27535
میں	582882	کہا	61090	کرتے	34818	دی	27265
کی	575545	گیا	57360	ہونے	34212	ہوتا	26341
ہے	466908	جو	57041	گئے	34143	کام	26218
اور	413788	و	56038	اپنی	33693	حاصل	25881
سے	368155	گا	49586	والے	33589	رہی	25044
کا	306103	ہی	48594	وقت	33424	سب	24486
کو	281922	جائے	47470	بہت	33032	کرنا	24242
نے	254385	نہ	46989	کسی	32246	جاتا	24185
اس	244017	جب	46894	جا	32228	ہر	24102
کہ	237419	اپنے	46829	تھی	31627	سی	23927
ہیں	217948	آپ	46299	یا	31361	طور	22284
پر	207921	جس	45733	گئی	31279	خان	21801
کر	173407	دیا	45394	ہم	30962	پہلے	21641
ہو	145669	تھے	43418	ہوں	30623	کچھ	21568
بھی	140157	ہوئے	42427	کیلئے	30164	صرف	21425
ایک	129851	تک	42422	زیادہ	29541	انہیں	20853
یہ	128103	بعد	40860	ملک	29528	سال	20564
نہیں	115341	انہوں	39903	بات	29415	وجہ	20301
ان	110463	رہے	39784	رہا	29272	کم	20261
کیا	108414	ساتھ	38807	طرح	28965	جنگ	20095
تو	88137	لیکن	38482	اگر	28729	اسے	19942
وہ	83613	گی	37533	اب	28345	ہوتی	19935
لئے	66765	کریں	37351	حکومت	28332	لیے	19820
کرنے	64589	کوئی	36801	دو	28236	ٹیم	19654