# Appendix 1 : Description of the transcription scheme

The transcriptions schemes used in the works referred to throughout this study differ greatly. For this reason I outline in this appendix the method of transcription I use, which is so far as I am aware unique to me. It represents a compromise between different representations of Urdu[1] (and Hindi) in the Roman alphabet, combined with some features relevant to my particular concern with the written form as represented in electronic texts.

Most transcription schemes represent the *pronunciation* of Urdu, rather than how it is spelt, because Urdu is written in the Perso-Arabic alphabet, which does not customarily mark short vowels. Furthermore, since Urdu has more long vowels and diphthongs than Arabic, there is a great deal of ambiguity in how long vowels are represented in the Perso-Arabic script. Rather than reproduce this ambiguity in a straight transliteration into the Roman alphabet, I follow previous studies and include short vowels in the transcriptions. However, in other matters I have tried to stick rather more closely to the Perso-Arabic spelling. This is simply a matter of practicality. Pedagogical works and descriptive grammars obviously require as precise a representation of what is *said* as possible. For the purposes of part-of-speech tagging, however, it is better to represent the spelling since that is all the information the tagger (human or software) has access to.

I now discuss each part of the system used to transcribe Urdu in this study. For

[1] "Urdu" and "Hindi" are written as such throughout; the names of the language(s) are given without the long vowels of the original Hindi-Urdu words. Barz (1977) writes "Hindī" and "Urdū", but I have not followed this practice. Long vowels are not thus indicated in English, and the names when they appear in my text must be regarded as English words (albeit loanwords from Hindi-Urdu).

a short discussion of the Perso-Arabic script, see section CROSSREF; I do not discuss the details of the Arabic and Perso-Arabic scripts; see Bhatia and Koul (2000).

## A1.1 Roman symbols common to the majority of transcription schemes

The following symbols are widely agreed on (only the independent forms of the Perso-Arabic characters are shown; the Arial Unicode MS font is used throughout):

| Perso-Arabic | Roman | Perso-Arabic | Roman |
|---|---|---|---|
| پ | p | ب | b |
| ر | r | د | d |
| چ | c | ج | j |
| ک | k | گ | g |
| ق | q | ف | f |
| ل | l | م | m |

## A1.2 Single Roman symbols mapped from multiple Perso-Arabic characters

In Urdu, there are several pairs or groups of Perso-Arabic symbols that represent the same sounds. In Persian and/or Arabic, they represent sounds that Urdu does not distinguish (see also section 1.1.5.2.3). Some writers (e.g. Platt 1884,

Schmidt 1999, Bhatia and Koul 2000) preserve this distinction in their transcriptions by means of subscript dots – but not all do this consistently. In Schmidt's case, she does it only when discussing Persian and Arabic elements in Urdu; in Bhatia and Koul's, only when actually discussing the writing system. Others (e.g. Bailey et al. 1956, and of course every writer who has based their transcription on the Devanāgarī alphabet used to write Hindi) do not even attempt to maintain the distinction. Although a strict transliteration *would* do so, I do not. Only one of the alternate forms is the usual form, and the others are found mainly in Persian and Arabic loanwords; thus the distinction is not critical[2].

| Perso-Arabic | Roman | Perso-Arabic | Roman |
|---|---|---|---|
| ت ط | t | س ث ص | s |
| ذ ظ ض | z | ہ ح | h |

## A1.3 Symbols for retroflex consonants

Retroflex consonants are an important part of Urdu phonology. There are three, corresponding to the dental consonants *t*, *d* and *r*; in both Perso-Arabic and Roman symbols, they are represented by some variation on the symbol for their dental counterpart. However, there is no consensus on what this variation should be in Roman transcription. Schmidt (1999) and Platts (1884) indicate it with a single subscript dot. Bailey et al. (1956) use the IPA symbols ʈ, ɖ and ɽ. Bhatia and Koul (2000) use the uppercase Roman letters T, D, and R. This last approach has been

[2] Of course, in any context where the distinction is critical, it has been pointed out.

followed, to maximally differentiate the visual appearances of the retroflex and dental consonant symbols. The subscript dots are too easily overlooked (and too easily confused with those used to indicate different spellings of the same sound: see above) and the tailed IPA symbols too hard to read.

| Perso-Arabic | Roman | Perso-Arabic | Roman |
|---|---|---|---|
| ٹ | T | ڈ | D |
| ڑ | R | | |

## A1.4 Aspiration

The symbol called *dō caśmī hē* indicates aspiration of a stop consonant. It is most frequently represented by $h$ [3], but Bhatia and Koul (2000) represent it with a small superscript $^{h}$ (so that it is distinct from *h* above). In this case I have gone with the majority: superscript symbols are small and difficult to read. Note that the shapes taken by *dō caśmī hē*, particularly in its initial form, vary considerably from typescript to typescript.

| Perso-Arabic | Roman |
|---|---|
| ھ | h |

[3] In the Devanagari script used for Hindi there are separate symbols for the aspirated and unaspirated versions of consonants, but Barz (1977), whose work makes substantial use of Devanagari, uses the same transcription for aspirated stops as is used here.

## A1.5 Symbols for fricative consonants

The voiceless velar fricative is generally transcribed as *x*, although Platts (1884) transcribes it as *kh*. Despite the underscore, this is too easily confused with *kh* (aspirated *k*). The voiced velar fricative is transcribed by Platts (1884) as a lower case G with a line through its tail – again, too easily confused, with *g*. Bhatia and Koul (2000) represent this sound with an uppercase G, but I follow Schmidt (1999) and Bailey et al. (1956) in transcribing it as a lowercase gamma.

The two palatal fricatives are also represented in several ways. Platts (1884) uses *sh* or *ś* and *zh*, Bailey et al. (1956) use ʃ and ʒ, and Bhatia and Koul (2000) use *sh* and *ž*[4]. To introduce some consistency, whilst avoiding the reader-unfriendly IPA symbols, I use the acute accent to indicate both palatal fricatives.

| Perso-Arabic | Roman | Perso-Arabic | Roman |
|---|---|---|---|
| خ | x | غ | γ |
| ش | ś | ژ | ź |

## A1.6 Representation of vowels and diphthongs

There is only one major difference between most sets of transcriptions: whether or not the macron is used to indicate long vowels. Most writers (e.g. Platts 1884, Schmidt 1999, Bhatia and Koul 2000) do make use of a superscript macron (or

[4] Schmidt (1999) uses *ś* for the voiceless fricative, but I have not been able to determine what she uses for the voiced fricative.

acute accent) for long vowels; others, such as Bailey et al. (1956) use separate symbols for the long and short vowels[5]. This latter approach, while it avoids superscript marks, ultimately creates difficulties, as it means that *y* and *w* may be used for vowels as well as consonants[6] – but *not* for the vowels that the corresponding symbols in Perso-Arabic represent. It also involves use of the inverted "e" sign (IPA schwa / ə /) to represent the "short *a*" – granted that this is the more accurate representation of the vowel's quality, it is still awkward to use and read in running text.

I use the macron for all five long vowels; this is perhaps unwise, as there is no phonemic short *e* or short *o* against which to distinguish the long vowels. (Accordingly, Bhatia and Koul use an unmodified *e* and *o* to represent the long vowels.) However, I use the macron to make clear that these vowels are of the same phonological length as *ā*, *ū* and *ī*. The short vowels are straightforwardly represented by *i*, *a*, and *u*. I do not, as Schmidt does, use short *e* and *o* when one of the phonemic short vowels are pronounced thus; instead I use the vowelling that would be used in Perso-Arabic script, in the (rare) event of the vowel diacritics not being omitted. The symbol for a doubled consonant is treated in much the same way as a short vowel diacritic in terms of being omitted; I always transcribe the double consonants as I always transcribe the short vowels.

The diphthongs are written as such, following Schmidt (1999). Bhatia and Koul (2000) consider them to be pure vowels[7] and write one of them as such, but for

[5] Like English, Hindi-Urdu possesses pairs of corresponding vowels that are distinguished in both length and quality. See also section 1.1.5.1 on Urdu phonology.

[6] For example, in Bailey et al. the vowel / ɪ / is transcribed as *y* , and the vowel / i: / as *i*.

[7] Phonological support for this viewpoint is given by Kachru (1990: 55).

purposes this is an unnecessary deviation from the Perso-Arabic representation.

| Perso-Arabic | Roman | Perso-Arabic | Roman |
|---|---|---|---|
| ( ◌َ ) | a | ا ( آ ) | ā |
| ( ◌ِ ) | i | ی | ī |
| ( ◌ُ ) | u | و (وُ) | ū |
| ے | ē | و | ō |
| ے( ے َ ) | ai | و ( وَ) | au |
| ( ◌ّ ) | (double cons.) | | |

The long vowel symbols are also used to represent consonants (glides); in fact in terms of the Perso-Arabic script, their consonantal use is primary and the indication of long vowels secondary. All the transcription schemes I know of use separate characters to transcribe these two uses.

| Perso-Arabic | Roman | Perso-Arabic | Roman |
|---|---|---|---|
| ی | y | و | v |

## A1.7 Representation of izāfat

The single-vowel clitic known as izāfat is realised in Urdu with several spellings, or with none (Schmidt 1999: 247). Where it is actually present in the written form being transcribed, it will be represented with an *e* at the end of the word

to which it is attached. (The symbol *ē* is not used, because izāfat is often dropped from pronunciation and it would be odd to transcribe as a long vowel a syllable that is frequently so short it is elided altogether!) Traditionally, izāfat is written as *e* hyphenated to both the preceding and following words (e.g. "ism-e-sharīf" – Bhatia and Koul 2000). I will not follow this practice, for reasons given below. For example, I would transcribe the previous example as *isme śarīf* if the izāfat is written, as *ism śarīf* if it is not.

## A1.8 Representation of nasal consonants and vowels

The consonant symbol called *mīm* is, as mentioned above, transcribed unproblematically as *m*. However, the symbols for the remaining nasal sounds are less straightforward. Bhatia and Koul (2000) transcribe the symbol called *nūn* in four different ways, to show the four different places of articulation it can have preceding different stops – perhaps influenced by the Devanāgarī script, which has separate symbols for each of these. Bailey et al. (1956) do not do this; I have not either, since for POS tagging purposes there is little to be gained by making distinctions that are utterly absent in the writing system. The *nūn* symbol is also used to indicate a nasal vowel (although only in medial and initial position; in final position, a character like a *nūn* without a dot is used for this purpose). This is transcribed in two major ways: firstly, by marking the nasal vowel with a superscript tilde (Bhatia and Koul 2000); secondly, with an independent character following the nasal vowel (in Platts 1884, an *n* with a superscript dot; in Schmidt 1999, an *m* with a superscript dot). Neither of these is ideal: the tilde is too easily confused with the macron (and when both must be put over the same vowel, or if a diphthong is nasal, the text becomes very untidy) and

the superscript dots are too easily overlooked, allowing the nasality sign to be mistaken for an independent nasal consonant. I take a compromise approach, and indicate nasality with a tilde following the nasal vowel, thus: *pā~c*, *ammā~*, *hai~*. This looks a little odd, but it is unmistakeable.

| Perso-Arabic | Roman | Perso-Arabic | Roman |
|---|---|---|---|
| ں | n | ں ( ں) | ~ |

## A1.9 Representation of words ending in *chōTī hē*

Many words end in *chōTī hē*, which is transcribed *h*. However in final position most writers transcribe it as *–a* (which is how it is pronounced). I transcribe this as *-ah*, following the spelling.

## A1.10 Other "silent" consonants

The consonant called *ain* is found in loanwords. It represents the Arabic voiced pharyngeal fricative; this is not found in Urdu, and so its pronunciation is unpredictable. It is pronounced as a glottal stop, as zero, as *ā* or as *a*, depending on a range of factors including its environment (Bhatia and Koul 2000: 228). Most writers, including Bhatia and Koul, do not use a consistent transcription for *ain* and transcribe according to pronunciation as often as possible. However, because of its presence in the writing system, I transliterate it as an apostrophe wherever it appears (this is the symbol generally used for it, when it is not left out).

The superscript *hamza* symbol is used to indicate a vowel cluster. Since this is

adequately indicated in the Roman transcription by the presence of two adjacent vowels (which are probably unwritten in the original Perso-Arabic) I do not transliterate *hamza*.

| Perso-Arabic | Roman | Perso-Arabic | Roman |
|---|---|---|---|
| ء | ( zero ) | ع | ’ |

Generally, whenever a consonant is not pronounced, due to whatever irregularity, I have tried to transcribe according to the spelling and not the pronunciation. This is against the general trend of earlier transcription systems, but is justified by the nature of part-of-speech tagging: spelling is very important, and the written word is all there is; thus the Roman written word should mirror the Perso-Arabic as closely as may be.

## A1.11 The Arabic article *al*

In Arabic, the *l* of this word assimilates to a following dental or alveolar consonant (although its spelling does not change). Many phrases have been borrowed into Urdu that contain such an assimilated *al*. These words are normally transcribed according to pronunciation (e.g. “as-salām”, Bhatia and Koul 2000; “ ‘abd-ur-rahmān”, Schmidt 1999) but I have transcribed according to spelling, thus: *alsalām*, *’abd alrahmān*.

## A1.12 Word breaks and hyphenation

Most writers use the hyphen in their transcriptions, for joining together the Arabic article *al* and the following word, and/or for linking together the Persian izāfat and the words before and after it (see above). I have not done this, as there is nothing to justify it in the Perso-Arabic script, or indeed the practice of transcribing izāfat as a separate word from the one that precedes it. Where the Perso-Arabic contains a word break, I have transcribed the words with a space between them; where the Perso-Arabic is a single word, I have transcribed the words as one. For further discussion of the problems of Urdu word breaks, see sections 2.2.6 and 3.12.2.