Appendix 3 : Creating the Perso-Arabic tagset

As described in section 2.2.9.2, a tagset using Perso-Arabic characters has been created. The Perso-Arabic tags are given alongside the Roman tags in the descriptions of the tagset in chapter three, and elsewhere the Roman tags alone are used. In this Appendix I outline the principles used to create the Perso-Arabic tags.

Rather than create tags from scratch, I devised a system whereby each character in the Roman tags would map to a given Perso-Arabic character. Creating these mappings as laid out below was by no means unproblematic. To simply map each Roman character to its nearest phonetic equivalent in the Perso-Arabic alphabet would not be useful. Without reference to English grammatical terminology – or the Latin and Greek roots on which much of it is based – the significance of characters mapped in this way would not be clear. Therefore it was necessary to choose characters that "mean something" in the context of Urdu grammatical terminology.

There were two additional difficulties. Firstly, many Urdu grammatical terms begin with the words *harf*, "letter/particle" or *ism*, "name/noun" so that tag initials had to be taken from words modifying these terms. Secondly the grammatical terms I needed and the grammatical terms I was able to find using Haq (2001) did not always coincide. Because of these flaws, this Perso-Arabic version must be considered provisional, pending the approval of some future user whose native language is, if not Urdu, then at least a language written in the Perso-Arabic alphabet.

As a result of this difficulty in choosing appropriate letters, there is a many-to-many mapping between Roman and Perso-Arabic characters, as the table below outlines. Where phonetic / orthographic justification is cited, it means that both the Roman and the Perso-Arabic characters represent some element of an Urdu sound or

letter associated with the decomposable element in question.

Despite the fact that the Perso-Arabic script is cursive, the tags contain only the independent forms of the letters. This is used to parallel the use of capital letters in the Roman tagset. It will also help the reader to distinguish tags visually from the words of the text. In Unicode files, it is to be specified that every character of a Perso-Arabic tag except the last will be followed by the zero-width non-joiner character (U+200C), to ensure that the characters are not joined together when displayed or printed.

Roman	Perso-Arabic	Justification
A (honorific)	ا	Phonetic/orthographic justification
A (in AL)	ا	Phonetic/orthographic justification
A (in AU)	ا	Phonetic/orthographic justification
A (in FA)	ا	I was unable to find an Urdu word for “acronym” and have thus used the transliteration of Roman A.
B (in IB)	گ	Second letter of *āgē*, “before”
B (in FB)	م	First letter of *muxaffaf*, “abbreviation”
B (in XB)	ب	Phonetic/orthographic justification
C (cāhiē)	چ	Phonetic/orthographic justification
C(onjunction)	ت	First letter of second word of *harfe tardīd*, “(disjunctive) conjunction”
C(litic)	چ	First letter of *chōTā*, “small, short”

C (reciprocal)	پ	Second letter of *āpas*, the reciprocal pronoun.
D(eterminer)	ت	First letter of second word of *harfe ta'rīf*, "definite article"[1]
D(egree)	ت	First letter of second word of *isme tafzīl*, "comparative/superlative adjective"
E	ے	Phonetic/orthographic justification
F(eminine)	ع	First letter of *'aurat*, "woman"
F(oreign)	د	First letter of *dūsrā*, "other, alternate"
F(raction)	ب	First letter of *batē*, "divided". (The actual word for "fraction" could not be used since its initial letter is already in use in this context.)
F (reflexive)	ج	Second letter of *rāja'*, "returning", part of the Urdu term for "reflexive"
G (for gā)	گ	Phonetic/orthographic justification
G (for gunā)	گ	Phonetic/orthographic justification
G (possessive)	ق	First letter of *qābiz*, "possessor"
H (for hōnā)	ہ	Phonetic/orthographic justification
H (in XH)	ہ	Phonetic/orthographic justification
I (postposition)	ج	First letter of second word of *harfe jār*, "preposition"
I (imperative)	ر	Third letter of *amro nahī*, "imperative mood"

[1] This is not a precise equivalent, but I could not find a better Urdu grammatical term.

J (relative)	ج	Phonetic/orthographic justification
J (adjective)	ص	First letter of second word of *isme sifat*, "adjective"
K (interrogative)	ک	Phonetic/orthographic justification
L (in AL)	ل	Phonetic/orthographic justification
L (lexical)	ظ	Last letter of *lafz*, "word"
M (first person)	م	Phonetic/orthographic justification
M(arked)	ا	The letter in which all marked adjectives and many marked nouns end, when in their citation form (it is the masculine nominative inflection)
M(asc)	م	First letter of *muzakkar*, "male"
M(odal)	ط	First letter of *tarīqah*, "mode" (probably not in the linguistic sense, but I was unable to find a better word)
N(indefinite)	غ	First letter of *ɣairma'ayyan*, meaning "indefinite"
N(infinitive)	ن	Phonetic/orthographic justification
N(egative)	ن	Phonetic/orthographic justification (based on the Urdu word for "not", which is *nah*).
N(ominative)	خ	First letter of *xātī* (?[2]), "nominative"

[2] Due to unclear printing in Haq (2001), I was unable to make out the word for "nominative" exactly, but I am reasonably certain that *x* is the first letter.

N(oun)	س	Second letter of *ism*, “noun”
N(umber)	ع	First letter of *’adad*, “numeral”
O(blique)	ص	Second letter of *nasb*, “oblique/accusative case”
O (in FO)	ر	First letter of *riyāzī*, “mathematics”
O (in OO)		Phonetic/orthographic justification
P(ast)	ض	Third letter of *māzī*, “past”
P(redicate)	خ	First letter of second word of *sifat xabarī*, “predicate adjective”
P(ronoun)	ض	First letter of second word of *isme zamīr*, “pronoun”
P(roper)	ن	First letter of *nām*, “name”
Q(uestion)	ک	Phonetic/orthographic justification (this is the first letter of all the interrogative words, as well as the question marker)
R (for rahā)	ر	Phonetic/orthographic justification
R (adverb)	ل	Letter occurring twice in *muta’aliq fa’l*, “adverb” (parallel to R in English, the second consonant in the lexeme “verb”)
R(eflexive or reciprocal)	ر	First letter of *rāja’*, “returning”, part of the Urdu term for reflexive
S (for sā)	س	Phonetic/orthographic justification

S(ubjunctive)	ش	Third letter of word meaning "subjunctive"[3]
S(ubordinating)	ش	First letter of second word of *harfe śarT*, "conditional conjunction"
S (in FS)	ح	First letter of *harf*, meaning "letter"
T (2nd person)	ت	Phonetic/orthographic justification
T (imperfective)	ت	Phonetic/orthographic justification
T (in TT)	ٹ	Second letter of *ciT*, meaning "tag, label" (from English "chit")
T (in XT)	ت	Phonetic/orthographic justification
U(nmarked)	ن	Second letter of *nah*, meaning "not" as in "not marked"
U (in AU)	و	Phonetic/orthographic justification
U (in FU)	ن	First letter of word *nāma'lūm*, meaning "unknown"
V (for vālā)	و	Phonetic/orthographic justification
V (3rd person)	و	Phonetic/orthographic justification
V (far-dem.)	و	Phonetic/orthographic justification
V(erb)	ف	First letter of *fa'l*, meaning "verb"
V(ocative)	د	Second letter of second word of *harfe nidā*, "vocative"

[3] Once again, due to unclear printing in Haq (2001), I could not make out the whole word.

X (non-lexical)	غ	Second letter of second word of *ke baGair*, meaning "without", as in "without lexical content". The Urdu grammatical term which is closest to "non-lexical" is probably *harf* but this term is too general.
X (in FX)	ل	First letter of *lafz*, "word" (the closest thing I could find to "string" in the computational sense)
Y (near-dem.)	ی	Phonetic/orthographic justification
Y (perfective)	ی	Phonetic/orthographic justification
Z (in ZZ)	ئ	Phonetic/orthographic justification (fullest way to write the izāfat clitic)
Z (in FZ)	ت	First letter of second word of *hurūfe tahajjī*, "letters of the alphabet"
0 (root form)	۰	The corresponding Urdu numeral is used.
1 (singular)	۱	The corresponding Urdu numeral is used.
2 (plural)	۲	The corresponding Urdu numeral is used.

Punctuation tags are discussed separately in section 3.11.