3 Specification of the tagset

The primary purpose of this chapter is to fulfil the aim stated in the Introduction, of defining a POS tagset for use in the tagging of Urdu, in compliance with the EAGLES guidelines and with the design principles stated in the previous chapter. This is done in sections 3.1 to 3.14. However, there is also a secondary purpose, which is to establish the claim that it is possible to usefully extend the EAGLES guidelines to Urdu (see Introduction and 2.2.1.3). This claim is evaluated by the very process of attempting to define an EAGLES-compliant Urdu tagset; remarks on the degree to which this claim can be upheld are given at the end in section 3.15.

The tagset as given below was devised by considering the EAGLES guidelines step-by-step and assessing the applicability of each of the categories to Urdu. At the same time, an intermediate tagset was constructed as per the EAGLES guidelines. The specification is organised according to the EAGLES major parts of speech (noun, verb, etc.) as outlined by Leech and Wilson (1999). The tagset as a whole contains 387 tags.

All details of Urdu morphosyntax given in the definition of the tagset are drawn from Schmidt (1999)[1], unless otherwise specified. I refer to works on Hindi as well as works on Urdu for details of morphology and syntax, but have always given priority to authors dealing exclusively with Urdu to ensure that Hindi-only features did not "creep in" to the tagset. For more details, see section 2.3.

[1] As explained in 2.3, Schmidt's grammar is being used as a model of the Urdu language for the purposes of tagset definition.

3.1 Nouns

The EAGLES guidelines give four recommended attributes for nouns: *type*, *gender*, *number* and *case*. There are also two optional attribute, *countability* and *definiteness*. *Type* refers to whether a noun is *common* (denotes one or more members of a class of things[2]) or *proper* (is the name of one or more particular things). This attribute is an example of one which is marginal to morphosyntax, but should be included since the distinction between common and proper might well prove useful to some future linguistic investigation of the text. It has been included in the tagset for now, but with the reservation that it might have to be collapsed in any subtagset for automatic tagging. This is because there may well not be any way for the tagger to make this distinction. Unlike the Roman, Greek and Cyrillic alphabets, the Urdu alphabet has no uppercase letters. In the European languages for which the EAGLES guidelines were designed, which use one of the former alphabets, uppercase letters are often used to identify proper nouns. It is clear that no such simple rule could be employed in Urdu. Furthermore there are no articles in Urdu (Bhatia and Koul 2000: 318), the absence and presence of an article being typical of proper and common nouns respectively in English and similar languages.

The attributes *gender*, *number*, and *case* are familiar linguistic features. Urdu marks all of them by means of suffixes on nouns. Moreover, the suffixes for these features are fused; in other words, Urdu has noun declensions. The creation of appropriate tags and EAGLES intermediate tags using these attributes is discussed

[2] "Thing" here is to be taken in the broadest possible sense – i.e. an entity real or hypothetical, concrete or abstract. This is a purposefully vague definition, since the issue of how a "noun" is to be defined is by no means theoretically uncontroversial.

below.

3.1.1 Gender

Urdu has two genders, masculine and feminine. Some nouns are marked for gender, whereas others are not[3]. This means that there is in effect a four-way distinction among nouns: masculine marked, masculine unmarked, feminine marked and feminine unmarked. For example:

rūpayah	"money"	(marked masculine)
ghar	"house"	(unmarked masculine)
baccī	"female child"	(marked feminine)
kitāb	"book"	(unmarked feminine)

(examples from Schmidt 1999: 1-2.)

Note that since some unmarked nouns coincidentally display the suffixes typical of marked nouns, the diagnostic feature of a marked noun is that its plural inflection follows that of the marked nouns (e.g. masculine *–ā* changing to *–ē*, feminine *–ī* to *–iyā~*, and so on).

This four-way split could be encoded into a tagset in two ways: by creating

[3] Some writers (e.g. Bailey et al. 1956: 1) have captured this fact by saying that Urdu nouns fall into "two declensions". Kachru (1990) goes further, identifying eight separate noun "paradigms". However, this approach is avoided here because many of the suffixes are identical between the so-called declensions. In the EAGLES guidelines, "inflection type is omitted as an attribute, since it is purely morphological". But it would seem better to include this information, although not strictly morphosyntactic, since it could well prove to be of value to the end user.

two new values for the *gender* attribute (the EAGLES guidelines have only *masculine*, *feminine*, *neuter*, and *common*) or by creating a new *markedness* attribute with two values, 1 = *marked for gender* and 2 = *not marked for gender*. The latter approach has been followed since it will almost certainly be easier for software processing the intermediate tagset to ignore an entire attribute than to work out what to do about values it does not recognise in existing attributes. This is especially the case if the extra attribute is added at the end of the tag, as I have done.

3.1.2 Number

Urdu has two numbers, singular and plural. This is well agreed on (Schmidt 1999: 1; Bhatia and Koul 2000: 314; Barz 1977: 36; Bailey et al. 1956: 1, 5). The EAGLES guidelines on noun number allow for exactly this possibility, and thus have been implemented unproblematically.

3.1.3 Case

In the model of the language given by Schmidt, Urdu has three cases, nominative, oblique and vocative. McGregor (1972: 1-2) uses a different classification, treating the vocative as a special form of the oblique case. However, since the special form would still need to be tagged separately, it makes sense to treat it as a vocative case, a phenomenon for which the EAGLES guidelines already allow for.

As Schmidt (1999: 7) points out, some grammarians[4] have treated Urdu

[4] For example, Kellogg (1875) and Butt (1995) are both of this view.

postpositions as being either suffixes or clitics indicating cases, in which case Urdu would possess many more than three cases. However, this is a minority view amongst writers of general grammars: Schmidt (1999), Barz (1977), Bhatia and Koul (2000), McGregor (1972), Bailey et al. (1956) all do not treat postpositions as marking cases. There is an etymological basis for this view. Kellogg (1875: 128-133) reports that the postpositions do not derive from Sanskrit case markers, but rather from independent words (e.g. *kō*, "to", from Sanskrit *kākshe*, "armpit, side"; *mē~*, "in", from Sanskrit *madhye*, "middle", both locative nouns; *tak*, "until", from the Sanskrit past participle *tarita*, "passed to", plus a dative affix *ku*.). Furthermore, the suffix/clitic approach would require case to be determined across multi-token units, which would breach the design principle of including no multiword tags. It would also have implications for the principle of theoretical neutrality, since it would be necessary to take some standpoint on the subject of whether or not Urdu has ergative case marking, a theoretically controversial point (see 1.1.5.4). Thus I use the nominative-oblique-vocative distinction as exemplified below:

laRkā, laRkē	"boy(s)"	(nominative singular/plural)
laRkē, laRkō~		(oblique singular/plural)
laRkē, laRkō		(vocative singular/plural)

(example from Schmidt 1999: 10-12)

There is something of an issue with the names of the cases. *Vocative* is straightforward enough, and is one of the values given for the *case* attribute in the EAGLES guidelines. *Nominative*, however, is usually given meaning by its contrast with *accusative* – a case that does not exist in Urdu. The nominative may in Urdu be

used for either, neither or both of the subject and the direct object. Thus it is not certain whether the *nominative* in Urdu really corresponds with the *nominative* that is value 1 in the EAGLES guidelines[5]. Certainly it does not correspond with the *nominative* as it exists in, for example, German or Latin. However, I have used value 1 in the intermediate tagset for the Urdu case, on the basis that no Urdu case resembles the nominative in the European languages for which the EAGLES guidelines were devised any more closely than the Urdu nominative.

There is no value in the EAGLES guidelines for *oblique*. Nor is there one for *postpositional*, *locative* or *instrumental* (alternative names used by Bailey et al. 1956 for this case[6]). Rather than invent an extra value (undesirable for reasons given with regard to *markedness* above), I have used the value for *dative* to represent *oblique*, on the grounds that in some European languages (e.g. German) prepositions frequently govern the dative, and in Urdu postpositions govern the oblique.

3.1.4 EAGLES attributes for nouns not used in this tagset

The optional attribute *countability* has not been used in this tagset. The count/mass noun distinction in Urdu is fairly similar to that of English. An example of a count noun is *kamrah*, "room"; an example of a mass noun is *pānī*, "water" (Schmidt 1999: 6-7). As in English, in the correct context, normally non-count nouns can be count without any additional morphological marking. For example, *dāl*,

[5] Barz (1977) and McGregor (1972) actually call the nominative case the *direct case*.

[6] In fact, Bailey et al. (1956: 8) suggest that there are in fact four cases: *nominative*, *vocative*, *oblique/postpositional*, and *locative/instrumental*, with the latter two having exactly the same form. However, it is more parsimonious, as the later authors have done, to consider this an example of one case with more than one use (hardly an unknown phenomenon in the annals of linguistic description).

"pulse" is normally non-count, but when it means "a type of pulse" it is count. Therefore this is not a distinction that can be made with reference to morphosyntax, since it is dependent on semantic/pragmatic features – without reference to the sense of the sentence it is not always possible to say whether any given noun is count. It is thus excluded from the tagset, in accordance with design principle that semantic and pragmatic information shall not be included. The optional attribute *definiteness* has also not been used, since definiteness is not marked morphologically on nouns in Urdu (but see section 3.5 on the Arabic definite marker).

3.1.5 The problem of ambiguous suffixes

A potential problem with making the distinctions listed above is that in Urdu, many noun suffixes indicate more than one of the attribute-value combinations that are possible. The inflectional noun suffixes (based on Schmidt 1999) are listed below[7]: Note that it is possible for some of the sequences of letters/sounds given below to occur word-finally without being a suffix (if the noun is unmarked for gender).

Table 3.1

Suffix	Indicates…
(Zero)	Unmarked masculine nominative singular Unmarked feminine nominative singular Unmarked masculine nominative plural

[7] Excepting those that are confined to Perso-Arabic loanwords. There are also derivational suffixes that *determine* the gender of the noun they create, but these do not change for case or number.

	Unmarked masculine oblique singular Unmarked feminine oblique singular Unmarked masculine vocative singular Unmarked feminine vocative singular
–ā(~)	Marked masculine nominative singular (form with ~ described as "rare" by Schmidt 1999)
–ah	Marked masculine nominative singular
–ayah[8]	Marked masculine nominative singular
–ī	Marked feminine nominative singular Marked feminine oblique singular Marked feminine vocative singular
–iyā	Marked feminine nominative singular Marked feminine oblique singular Marked feminine vocative singular
–ē	Marked masculine nominative plural Marked masculine oblique singular Marked masculine vocative singular
–aē	Marked masculine nominative plural Marked masculine oblique singular Marked masculine vocative singular
–iyā~	Marked feminine nominative plural
–ē~	Marked masculine nominative plural ("rare")

[8] The suffixes ending in *choTī he* (ہ) are transcribed "–a" and "–aya" by Schmidt (1999), because the *choTī he* at the end is not pronounced.

	Marked masculine oblique singular (“rare”) Marked masculine vocative singular (“rare”) Unmarked feminine nominative plural
–ō~	Marked masculine oblique plural Unmarked masculine oblique plural Unmarked feminine oblique plural
–iyō~	Marked feminine oblique plural
–ō	Marked masculine vocative plural Marked feminine vocative plural Unmarked masculine vocative plural Unmarked feminine vocative plural

The oblique singular is identical to the nominative singular except for marked masculine nouns, where it is identical with the nominative plural. The vocative singular is identical to the oblique singular. Combined with other multiple-use suffixes, this means that the affix-meaning relationship is simultaneously many-to-one and one-to-many[9]. Thus it might seem wise to have one tag for each affix rather than one tag for each morphosyntactic category. However, this would create some unhappy bedfellows (e.g. oblique singular classed with nominative plural) and breach some of the design principles of the tagset, namely that of tagging for function (i.e. number, gender and case) rather than tagging for form (i.e. the surface form of the suffixes).

[9] This summary does not consider those words which happen to end in one of the “suffix” forms as part of their base form, before any inflection; e.g. *devā* “medicine” is unmarked feminine, not marked masculine (Schmidt 1999: 3). Such words confuse the situation yet further.

3.1.6 The tags for nouns

On the basis of the attributes as discussed above, the intermediate tags for Urdu nouns will be formed as follows (those attributes which are not applicable to Urdu[10], for whatever part of speech, always have the value 0 in the intermediate tagset for Urdu):

Table 3.2

Value	*i) type*	*ii) gender*	*iii) number*	*iv) case*	*vii) markedness*
1	Common	Masculine	Singular	Nominative	Marked
2	Proper	Feminine	Plural		Unmarked
3				Oblique	
4					
5				Vocative	

Logically, there are 2 x 2 x 2 x 3 x 2 = 48 tags that can be produced by these attribute-value pairs. These are given below.

Table 3.3

Description	**Tag (Roman)**	**Tag (Perso-Arabic)[11]**	**Intermediate Tag**
Common marked masculine singular nominative noun	NNMM1N	خ۱امسس	N1111001

[10] In the case of nouns, the non-applicable attributes are *countability* and *definiteness*, as explained above.

[11] See section 2.2.9.2 and Appendix 3.

Common marked masculine singular oblique noun	NNMM1O	س‌س‌ام‌۱‌ص	N1113001
Common marked masculine singular vocative noun	NNMM1V	س‌س‌ام‌۱‌ف	N1115001
Common marked masculine plural nominative noun	NNMM2N	س‌س‌ام‌۲‌خ	N1121001
Common marked masculine plural oblique noun	NNMM2O	س‌س‌ام‌۲‌ص	N1123001
Common marked masculine plural vocative noun	NNMM2V	س‌س‌ام‌۲‌ف	N1125001
Common marked feminine singular nominative noun	NNMF1N	س‌س‌اع‌۱‌خ	N1211001
Common marked feminine singular oblique noun	NNMF1O	س‌س‌اع‌۱‌ص	N1213001
Common marked feminine singular vocative noun	NNMF1V	س‌س‌اع‌۱‌ف	N1215001
Common marked feminine plural nominative noun	NNMF2N	س‌س‌اع‌۲‌خ	N1221001
Common marked feminine plural oblique noun	NNMF2O	س‌س‌اع‌۲‌ص	N1223001
Common marked feminine plural vocative noun	NNMF2V	س‌س‌اع‌۲‌ف	N1225001
Common unmarked masculine singular nominative noun	NNUM1N	س‌س‌ن‌م‌۱‌خ	N1111002
Common unmarked masculine singular oblique noun	NNUM1O	س‌س‌ن‌م‌۱‌ص	N1113002
Common unmarked masculine singular vocative noun	NNUM1V	س‌س‌ن‌م‌۱‌ف	N1115002
Common unmarked masculine plural nominative noun	NNUM2N	س‌س‌ن‌م‌۲‌خ	N1121002
Common unmarked masculine plural oblique noun	NNUM2O	س‌س‌ن‌م‌۲‌ص	N1123002

Common unmarked masculine plural vocative noun	NNUM2V	ف٢منسس	N1125002
Common unmarked feminine singular nominative noun	NNUF1N	خ١عنسس	N1211002
Common unmarked feminine singular oblique noun	NNUF1O	ص١عنسس	N1213002
Common unmarked feminine singular vocative noun	NNUF1V	ف١عنسس	N1215002
Common unmarked feminine plural nominative noun	NNUF2N	خ٢عنسس	N1221002
Common unmarked feminine plural oblique noun	NNUF2O	ص٢عنسس	N1223002
Common unmarked feminine plural vocative noun	NNUF2V	ف٢عنسس	N1225002
Proper marked masculine singular nominative noun	NPMM1N	خ١مانس	N2111001
Proper marked masculine singular oblique noun	NPMM1O	ص١مانس	N2113001
Proper marked masculine singular vocative noun	NPMM1V	ف١مانس	N2115001
Proper marked masculine plural nominative noun	NPMM2N	خ٢مانس	N2121001
Proper marked masculine plural oblique noun	NPMM2O	ص٢مانس	N2123001
Proper marked masculine plural vocative noun	NPMM2V	ف٢مانس	N2125001
Proper marked feminine singular nominative noun	NPMF1N	خ١عانس	N2211001
Proper marked feminine singular oblique noun	NPMF1O	ص١عانس	N2213001
Proper marked feminine singular vocative noun	NPMF1V	ف١عانس	N2215001

Proper marked feminine plural nominative noun	NPMF2N	خ٢عانس	N2221001
Proper marked feminine plural oblique noun	NPMF2O	ص٢عانس	N2223001
Proper marked feminine plural vocative noun	NPMF2V	ف٢عانس	N2225001
Proper unmarked masculine singular nominative noun	NPUM1N	خ١مننس	N2111002
Proper unmarked masculine singular oblique noun	NPUM1O	ص١مننس	N2113002
Proper unmarked masculine singular vocative noun	NPUM1V	ف١مننس	N2115002
Proper unmarked masculine plural nominative noun	NPUM2N	خ٢مننس	N2121002
Proper unmarked masculine plural oblique noun	NPUM2O	ص٢مننس	N2123002
Proper unmarked masculine plural vocative noun	NPUM2V	ف٢مننس	N2125002
Proper unmarked feminine singular nominative noun	NPUF1N	خ١عننس	N2211002
Proper unmarked feminine singular oblique noun	NPUF1O	ص١عننس	N2213002
Proper unmarked feminine singular vocative noun	NPUF1V	ف١عننس	N2215002
Proper unmarked feminine plural nominative noun	NPUF2N	خ٢عننس	N2221002
Proper unmarked feminine plural oblique noun	NPUF2O	ص٢عننس	N2223002
Proper unmarked feminine plural vocative noun	NPUF2V	ف٢عننس	N2225002

3.2 Verbs

There are a considerable number of factors to be taken into account in a description and categorisation of the Urdu verbal system. There are a number of inflected forms, and with the use of one or more auxiliary elements, 15 compound tenses are built up. Furthermore, any part of the compound verb-phrase may be marked for number, person or gender agreement[12]. There are two conceivable approaches to the markup of such a compound verb-phrase. Firstly, each word could be tagged separately, regardless of its context. So for example the form that Schmidt (1999) refers to as the "perfective participle" would be tagged the same regardless of what compound tense it was being used in. Secondly, compound verbs could be treated as multi-word units, each such unit receiving a single tag.

The latter approach was not followed, for three reasons. In the first place, it goes against the principle that every word should have its own tag, using no multiword tags. Secondly, it goes against the suggestion made by the EAGLES guidelines that "In general, compound tenses are not dealt with at the morphosyntactic level, since they involve the combination of more than one verb in a larger construction" (Leech and Wilson 1999: 63). Thirdly, it would result in the tagset being much more complicated than need be. For example, each of the 15 compound tenses would need to be distinguished. By contrast the other approach would require a relatively smaller number of distinctions to be made, between the elements of which the compound tenses are built. The over-complicated tagset design that multi-word tagging of compound verbs would necessitate would also have the drawback of going far beyond the EAGLES guidelines on verbal tags. By treating each word of the

[12] Note that any single Urdu verb form is marked for either gender or person, never both.

compound verb as separate, it is possible to stick fairly closely to the guidelines.

The EAGLES guidelines for verb tags suggest a number of attributes that are not relevant to Urdu. Urdu lacks separable verbs and its passives[13] are phrasal rather than being morphologically marked; nor is reflexivity marked. The attributes *voice*, *separability* and *reflexivity* are thus superfluous. The attribute *auxiliary*, which encodes what auxiliary the verb takes in compound tenses, is also irrelevant, since all verbs in Urdu take the same auxiliaries. The attribute *Aux-function*, designed to distinguish English modal and non-modal auxiliaries, is not relevant as such in Urdu. While there are different types of auxiliary element, the distinctions between them are not of this clear, two-way oppositional nature. Therefore it would seem that there should be a better way to typify them than by attempting to shoehorn them into the categories of an attribute designed to encode something very different.

Of the remaining nine suggested attributes, the agreement attributes *number*, *gender*, and *person* are clearly relevant to the Urdu verbal system. Some writers consider that Urdu displays what has been described as split ergativity (as described in section 1.1.5.4). That is, the verb agrees sometimes with the subject, and sometimes with the direct object. It may also under some circumstances agree with neither (Schmidt 1999: 125). As explained in 1.1.5.4, however, some writers (e.g. Butt 1995) disagree with this analysis. However, for the purposes of defining verbal tags the matter of ergativity is more or less irrelevant. The agreement suffixes which occur on verbs – and, therefore, the morphosyntactic categories displayed by verbs – are exactly the same regardless of which argument of the verb is being agreed with. A single morphosyntactic phenomena receives a single tag; so for example when I give

[13] Except for one marginal case (see discussion of *cāhiē* in section 3.2.2.3 below).

a verb a tag VVYF1N[14] (see 3.2.1.3), it is not specified whether the feminine agreement is with a subject or object. Thus, the principle of theoretical neutrality is upheld: this analysis is as compatible with a theory in which Urdu displays split ergativity as with a theory in which it does not.

Aspect is relevant to all verbs and *tense* is relevant to the auxiliaries. *Status* (i.e. whether a verb is main or auxiliary) is relevant throughout. However, the way in which it has been used is a little different to that given in the EAGLES recommendations. The EAGLES guidelines suggest a main/auxiliary distinction which is context dependent. This can be seen by Leech and Wilson's example tagset for English (1999: 72-74), in which it is made clear that the verb *be* can be either a main verb or an auxiliary verb. However, the distinction I have used is between lexical verbs and non-lexical auxiliary verbs. This is not context-dependent; English *be* would be considered an auxiliary regardless of context. The motivation for this is the decidedly irregular morphology of Urdu auxiliary verbs, most particularly *hōnā*, "be" (see also 3.2.2.4). This goes far beyond the inflectional oddities found in English non-lexical verbs: *hōnā* possesses two tenses that no other verb has, and it possesses them regardless of whether it is a main verb or not. To mark up *hōnā* as a main verb, there would have to be a tag, for example, for a present-tense main verb. But to include such a tag would be to vastly misrepresent the majority of Urdu verbs, which have no inflected present tense. There are similar problems with such non-lexical verbal forms as *cāhiē* and *gā*. Thus it makes sense to use the *status* attribute to distinguish (mostly regular) lexical verbs and (irregular) auxiliary verbs, so that the unique marking on the latter can be tagged exclusively on the latter. The optional

[14] An instance of a verb that would receive this tag is the word *mānī* in the example given later in this section.

third value of the *status* attribute, *semi-auxiliary*, has been used as described below.

The last two attributes, *finiteness* and *mood*, are problematic. Firstly, inherent in the EAGLES guidelines is the problem that the *mood* attribute contains values relevant to both finite and non-finite forms, so that the *finiteness* attribute becomes redundant. Secondly, the finite/non-finite distinction may be hard to draw in Urdu. The forms described below as participles would traditionally be considered non-finite in European languages. However, in Urdu they have certain features which make them seem more like finite forms. For example, they can occur as the only verb in a main clause, and can agree with a subject or object – not a property prototypically associated with non-finite forms. These properties are illustrated by the following example from Schmidt (1999: 126)[15]:

unhō~	nē	an	paRh	kī	bāt	nah
3-PLRL-OBL	ERG	un	educated	of-FEM	speech	not
mānī						
accept-PERF.PART-FEM-SING						
They did not accept what the uneducated person said.						

The verb form *mānī* is a participle, but it is the only verb form in the sentence, and it is marked for agreement (with the object, since this clause is of the ergative type). It, like the postposition *kī*, agrees with the feminine singular noun *bāt*.

A third problem with the mood distinctions made in the EAGLES guidelines is that they are not necessarily those made by Urdu. For example, Urdu has forms which

[15] Schmidt does not give word-by-word glosses, only whole-sentence translations. I have added the glosses using Schmidt (1999) and Haq (2001) as guides. See also Appendix 2.

may be described as subjunctive and imperative moods, but it would seem to lack an indicative (except for the auxiliary *hōnā*). Because of these difficulties, the concepts of *finiteness* and *mood* will not be used to structure the tagset itself, although they are of course inevitable as attributes in the intermediate tagset[16]. This means that in some cases, the intermediate tagset values used to characterise some Urdu verb forms are somewhat arbitrary, since I have had to simply pick the values that seem closest to describing Urdu. For example, considering the "irrealis tense" (the term used by Schmidt 1999 for the finite use of the imperfective participle) to be a past tense subjunctive is not warranted by the Urdu verbal system. It was picked as the "least bad" way to characterise it simply because the Urdu irrealis has a usage similar to that of the past subjunctive in languages included in EAGLES such as German and (vestigially) English (e.g. *ich wäre*, *I were*). For example, Schmidt (1999) translates a sentence from the poet Ghalib as follows:

agar	aur	jītē	rahtē
if	and	alive-MASC-PLRL	stay-IMPERF.PART-MASC-PLRL
yahī		intizār	hōtā
this-very		waiting	be-IMPERF.PART-MASC-SING

If I were to live longer it would only be to wait like this

The presence in the translation of the past tense subjunctive ("I were") in the first – but not the second – of two clauses containing the finite imperfective participle demonstrates the partial parallelism between an Urdu irrealis and an English past subjunctive.

[16] Several of the values for *mood* (e.g. *gerund*, *supine*) are not used.

The intermediate tags as given contain fourteen attributes. This includes the attributes that are not used. There is also one attribute that the EAGLES guidelines did not contain, *case*, which is needed for reasons explained below. As with *markedness* for nouns, I have placed an additional attribute on the end of the word rather than attempt to modify the guidelines for intermediate tagset design in any internal fashion. The following table summarises the intermediate tagset used.

Table 3.4

Value	*i) person*	*ii) gender*	*iii) number*	*iv) finiteness*	
1	First	Masculine	Singular	Finite	*(Table*
2	Second	Feminine	Plural	Non-finite	*continues…)*
3	Third				

Value	*v) mood*	*vi) tense*	*viii) status*	*ix) aspect*	*xiv) case*
1	Indicative	Present	Main	Perfective	Nominative
2	Subjunctive		Auxiliary	Imperfective	
3	Imperative	Future	Semi-auxiliary		Oblique
4		Past			
5	Infinitive				Vocative
6	Participle				

Application of this intermediate tagset as described below gives a grand total of 113 tags for verbs. I will now consider each of the different types of verbs in turn.

3.2.1 Lexical verbs

The EAGLES guidelines do not consider lexical and auxiliary verbs to be separate major parts of speech, although this is a view that some have held (e.g. the ICE tagset – Greenbaum and Yibin 1996). However, in Urdu this distinction is very significant, since auxiliary forms pattern differently to the forms of lexical verbs. Therefore, this tagset will employ a high-level (but not top-level) distinction between lexical verbal elements (whose tags will commence with VV) and non-lexical or auxiliary verbal elements (whose tags will commence with V and one other letter – either one indicating what word it is, for auxiliary verbs whose inflectional behaviour is anomalous, or X for a general auxiliary). Thus both the EAGLES guidelines and the demands of Urdu morphology are complied with.

There exist in Urdu two widely applicable derivational suffixes which attach to the root of a lexical verb and increase its valence, making it transitive or causative in sense. This has been highlighted as a significant feature of the language (e.g. by Kachru 1990: 63)and is described in some detail by Schmidt (1999: 87, 157-175). It might be possible to distinguish such derived verbs from non-derived verbs in the tagset, but I do not, because of the design principle that no derivational information should be included. Furthermore, such a distinction would be difficult to automate, and also probably difficult for humans to annotate.

Lexical verbs occur in a number of inflected forms. The names of these forms are perhaps not very useful, since each of them has a variety of uses hard to capture by one of the traditional grammatical category names. However, rather than resort to letters or numbers which would be unlinkable to any previous writing on the Urdu verb, I use the same names for the forms as Schmidt (1999), as I have been doing thus

far in this thesis.

3.2.1.1 The root

The root consists, as its name suggests, of the root of the verb unadorned by affixation. It is not marked for person, number or gender and cannot occur as the sole verb of a main clause; it is, therefore, non-finite (untensed and also neither imperfective nor perfective in aspect). The exception to this is when it is used as an imperative form (discussed below). However, it does not fit neatly into any of the non-finite values for *mood* (the choices being *infinitive*, *participle*, *gerund* and *supine*). Therefore, in the intermediate tagset it is given a 0 for *mood*. Since this only has one form, there is only one tag. It should be noted that in the intermediate tags for this and all the following forms of lexical verb, all the tags give the *status* attribute the value *main*, since by definition a lexical verb is not an auxiliary (see the discussion of the *status* attribute in 3.2 above).

Table 3.5

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Root form lexical verb	VV0	ف‌ف‌.	V00020001000000

3.2.1.2 The infinitive

The infinitive of the verb is regularly formed. Mostly it is used as a verbal noun or as part of a complex verb phrase. It is also used as a neutral request form, in which case it is the main verb of its clause; however, I do not think that this usage is

sufficient to justify separate tagging; this is better treated example of a secondary usage of the same word, rather than a separate word (which giving it a separate tag would imply). The "default" ending of the infinitive is *–nā*, which is a masculine singular ending. When used as a noun it may occur in the oblique case; when it occurs in a verb phrase it may display gender and number agreement (in a similar way to an adjective). However these conditions cannot both occur[17]; therefore there is no feminine oblique or plural oblique, which reduces the number of tags necessary.

There is a problem creating the intermediate tagset: inasmuch as there is no attribute for "case" in the EAGLES guidelines for verbs (presumably non-finite verb forms in European languages do not display case inflection). An attribute, *case*, has therefore been added to the end of the intermediate tags. Otherwise this set of intermediate tags is fairly unproblematic.

The "N" in the mnemonic tags is derived from the *–nā* suffix that indicates the infinitive.

Table 3.6

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Infinitive lexical verb, masculine singular nominative	VVNM1N	فـفـنـم۱خ	V01125001000001
Infinitive lexical verb, masculine singular oblique	VVNM1O	فـفـنـم۱ص	V01125001000003
Infinitive lexical verb, masculine plural nominative	VVNM2	فـفـنـم۲	V01225001000001
Infinitive lexical verb, feminine singular nominative	VVNF1	فـفـنـع۱	V02125001000001

[17] That is, the language as described by Schmidt does not allow for this possibility (see 2.3).

Infinitive lexical verb, feminine plural nominative	VVNF2	ف‌ف‌ن‌ع۲	V02225001000001

3.2.1.3 The participles

Urdu has two participles, the imperfective and the perfective. However, unlike participles in many European languages, they can be used as the sole verb of a main clause. This creates the tenses referred to as the irrealis and the simple past respectively. However, the presence or absence of an auxiliary makes no difference to the form of the participle. It would therefore be misleading to use two tags for a single form of the verb. These tags are thus used for both finite and non-finite, and the notions of *irrealis* and *simple past* are not referred to in the precise definitions of the tags. The dual finite and non-finite nature of the tags is indicated in the intermediate tagset using the OR operator, | . There is a value in the EAGLES tagset for past tense, but there is not one for irrealis. The closest approximation to an irrealis in the EAGLES guidelines is *subjunctive past* (see the discussion of this point in 3.2 above). This is not a perfect solution, but without adding extra values to the intermediate tagset it is the best that can be managed. Thus, the imperfective is *finite subjunctive past* with *zero* aspect or *non-finite participle imperfective* with *zero* tense. The perfective is *finite indicative*[18] *past* with *zero* aspect or *non-finite participle perfective* with *zero* tense.

The participles are not marked for person, but are marked for gender and

[18] It is hard to justify this use of "indicative", since Urdu lexical verbs do not possess any indicative form as such. Therefore the notion of the indicative is not used in the definitions of the tags themselves, but only in the intermediate tagset (where something is needed to distinguish the finite use of the perfective participle from the finite use of the imperfective participle).

number. Their inflection is the same as that of adjectives, except that in some circumstances a distinction is made between feminine singular and plural which is not made by adjectives. Participles can also function as adjectives (see discussion of adjectives in 3.3 below), in which case this extra feminine singular/feminine plural distinction is not made (though this does not affect the tagging). That is to say, an adjective which agrees with a feminine plural noun or pronoun will always receive an F2 tag, regardless of whether it has the plural ending *–ī~* or the more general feminine ending *–ī*.

When participles are used as adjectives, it would in theory be possible to tag them as if they were adjectives. However, this has not been done, since even when being used attributively, participles appear in structures that normal adjectives do not. For example, they frequently occur in participial phrases with the perfective participle of the auxiliary verb *hōnā* (see below). When used adjectivally rather than verbally, participles may be marked for case as well as number and gender. This feature is also included in the tagset. Of course, the feature *case* only applies to the non-finite usage of the participle; this is reflected in the intermediate tagset by the use of (0 | 1) for the nominative or finite form. As with adjectives (see below), the "oblique" case is (3 | 5) in the intermediate tagset.

The characters Y and T have been used for the perfective and imperfective participles respectively, since these are the consonants that indicate the suffixes for these forms[19].

[19] In fact, the perfective participle is frequently marked by the vowel suffixes alone; only when the root ends in a vowel does the *y* appear.

Table 3.7

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Masculine singular (nominative) imperfective participle lexical verb	VVTM1N	خ۱متففف	V011(1|2)(2|6)(4|0)01 (0|2)0000(0|1)
Masculine singular oblique imperfective participle lexical verb	VVTM1O	ص۱متففف	V0112600120000(3|5)
Masculine plural (nominative) imperfective participle lexical verb	VVTM2N	خ۲متففف	V012(1|2)(2|6)(4|0)01 (0|2)0000(0|1)
Masculine plural oblique imperfective participle lexical verb	VVTM2O	ص۲متففف	V0122600120000(3|5)
Feminine singular (nominative) imperfective participle lexical verb	VVTF1N	خ۱عتففف	V021(1|2)(2|6)(4|0)01 (0|2)0000(0|1)
Feminine singular oblique imperfective participle lexical verb	VVTF1O	ص۱عتففف	V0212600120000(3|5)
Feminine plural (nominative) imperfective participle lexical verb	VVTF2N	خ۲عتففف	V022(1|2)(2|6)(4|0)01 (0|2)0000(0|1)
Feminine plural oblique imperfective participle lexical verb	VVTF2O	ص۲عتففف	V0222600120000(3|5)
Masculine singular (nominative) perfective	VVYM1N	خ۱مىففف	V011(1|2)(1|6)(4|0)01 (0|1)0000(0|1)

participle lexical verb			
Masculine singular oblique perfective participle lexical verb	VVYM1O	ف‌ف‌ی‌م‌ا‌ص	V0112600110000(3|5)
Masculine plural (nominative) perfective participle lexical verb	VVYM2N	ف‌ف‌ی‌م‌ا‌خ	V012(1|2)(1|6)(4|0)01 (0|1)0000(0|1)
Masculine plural oblique perfective participle lexical verb	VVYM2O	ف‌ف‌ی‌م‌ا‌ص	V0122600110000(3|5)
Feminine singular (nominative) perfective participle lexical verb	VVYF1N	ف‌ف‌ی‌ع‌ا‌خ	V021(1|2)(1|6)(4|0)01 (0|1)0000(0|1)
Feminine singular oblique perfective participle lexical verb	VVYF1O	ف‌ف‌ی‌ع‌ا‌ص	V0212600110000(3|5)
Feminine plural (nominative) perfective participle lexical verb	VVYF2N	ف‌ف‌ی‌ع‌ا‌خ	V022(1|2)(1|6)(4|0)01 (0|1)0000(0|1)
Feminine plural oblique perfective participle lexical verb	VVYF2O	ف‌ف‌ی‌ع‌ا‌ص	V0222600110000(3|5)

3.2.1.4 The subjunctive

The subjunctive is the only form that is marked for person in Urdu lexical verbs. It is not, however, marked for gender. Therefore the intermediate tagset forms give *gender* as *zero*, *mood* as *subjunctive* and *tense* as *present*.

Urdu has the three normal persons given in the EAGLES guidelines, each in singular and plural forms. Schmidt (1999: 97) suggests that Urdu verbs also have an additional polite or honorific form, which although second person in meaning (it agrees with a pronoun *āp* that refers to one or more interlocutors) is identical to the third person plural form of the verb. In this case I have deviated from the model described by Schmidt, for reasons discussed in my treatment of the *āp* pronoun in section 3.4.1.2. There will be no tags for honorific verbal forms, and verb forms which agree with *āp* will be tagged as third person forms. The exception to this is the imperative, discussed in the next section.

In the mnemonic tags, the part which varies for person is derived from the first letter of the Urdu pronouns *mai~*, "I", *tū*, "you", and *vah*, "he/she/it".

Table 3.8

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
First person singular subjunctive lexical verb	VVSM1	ف‌ف‌ش‌م۱	V10112101000000
First person plural subjunctive lexical verb	VVSM2	ف‌ف‌ش‌م۲	V10212101000000
Second person singular subjunctive lexical verb	VVST1	ف‌ف‌ش‌ت۱	V20112101000000
Second person plural subjunctive lexical verb	VVST2	ف‌ف‌ش‌ت۲	V20212101000000
Third person singular subjunctive lexical verb	VVSV1	ف‌ف‌ش‌و۱	V30112101000000
Third person plural subjunctive lexical verb	VVSV2	ف‌ف‌ش‌و۲	V30212101000000

3.2.1.5 The imperative

There are three simple imperative forms: second person singular (which is identical to the "root" form), second person plural (which is identical to the second person plural subjunctive form) and second person honorific. Each of these receives a separate tag. The existence of a second person honorific form does not undermine the general principle, stated above, that the *āp* pronoun takes a third person verb form since, in the imperative, there is no third person, and the subject is not expressed anyway. For the purposes of the intermediate tagset the *tense* is considered to be *present*, and the number of the honorific form is considered to be (1 | 2), since both singular and plural "subjects" are possible. This also serves to distinguish the VVIA tag in the intermediate tagset. The mnemonic "A" is the same as that used for the *āp* pronoun, and thus refers to politeness.

Table 3.9

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Second person singular imperative lexical verb	VVIT1	ف‌ف‌رت۱	V20113101000000
Second person plural imperative lexical verb	VVIT2	ف‌ف‌رت۲	V20213101000000
Second person honorific imperative lexical verb	VVIA	ف‌ف‌را	V20(1|2)13101000000

3.2.2 Auxiliary verbs

It should be noted that, whereas I have in this category treated all auxiliary

elements as verbs, in the terms of the EAGLES guidelines for intermediate tagsets some could easily be characterised as *unique* or *unassigned* words (see below). The EAGLES guidelines treat the English infinitive marker *to* in this manner, for example. However, treating them as verbs in the intermediate is firstly in keeping with the structure of the Urdu tagset, and secondly allows verbal attributes such as gender and number to be used (the EAGLES *unique* intermediate tags include no such attributes).

3.2.2.1 gā

The form *gā* indicates future tense when it follows a verb in the subjunctive form. It may also follow the polite imperative as a marker of additional politeness (Bhatia and Koul 2000: 332). It is considered by Schmidt (1999) to be a suffix, although one that is written as a separate word; Bhatia and Koul (2000) go so far as to write the inflected verb and the *gā* as a single word. However, given that the orthography must lead *gā* to be treated by the tagging system as a separate token (see 2.2.6.1), and given that the form of the future is otherwise identical to the subjunctive, it makes sense to tag *gā* separately from the lexical verb. Since *gā* is marked for gender and number and the subjunctive is marked for person and number, the future would, if treated as a simple rather than a compound tense, be marked for all three of these features – which is not true of any other simple tense in Urdu. Furthermore, as Schmidt (1999: 94) explains, *gā* derives from a contraction of the perfective participle of the verb *jānā*, "go". Therefore, *gā* is tagged independently.

In the intermediate tagset it is considered to be *finite*, *indicative*, *future*, and with *zero* aspect.

Table 3.10

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Masculine singular future auxiliary *gā*	VGM1	فگم۱	V01111302000000
Masculine plural future auxiliary *gē*	VGM2	فگم۲	V01211302000000
Feminine singular future auxiliary *gī*	VGF1	فگع۱	V02111302000000
Feminine plural future auxiliary *gī*	VGF2	فگع۲	V02211302000000

3.2.2.2 rahā

This auxiliary element is used in the formation of tenses in the durative aspect. It is itself the perfective participle of the lexical verb *rahnā*, "remain", but as Schmidt (1999: 111) reports, this form "has been delexicalised". It is marked for gender and number. It may seem that treating *rahā* as auxiliary and *rahnā* as lexical goes against the principle laid down in 3.2 that the distinction between lexical and auxiliary should be inherent to the verb and not dependent on context, and conflicts, for example, with the treatment of *hōnā* (see 3.2.2.4 below). However, this is not the case. The verb *hōnā* may be main but it is never lexical; *rahnā* is lexical when it is main, and cannot act as an auxiliary at all except for the one, very particular, delexicalised form *rahā*.

There is a problem in the intermediate tagset, in that the EAGLES guidelines contain no value for durative aspect. Therefore, the *aspect* attribute is given the value *zero*, since the aspect is neither perfective nor imperfective. This is not a very good solution but it is preferable to adding a value, and there is no satisfactory way to mark durative in the intermediate tagset by adding an attribute. This solution also ensures that each form of auxiliary *rahā* has a unique value in the intermediate tagset, since

every other participial element is either imperfective or perfective. Otherwise in the intermediate tagset, *rahā* is considered to be a *non-finite participle* with *zero* tense.

When used lexically, *rahā* receives the tag VVYM1N, *rahī* receives VVYF1N or VVYF2N, and so on.

Table 3.11

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Masculine singular durative auxiliary *rahā*	VRM1	فرم١	V01126002000000
Masculine plural durative auxiliary *rahē*	VRM2	فرم٢	V01226002000000
Feminine singular durative auxiliary *rahī*	VRF1	فرع١	V02126002000000
Feminine plural durative auxiliary *rahī*	VRF2	فرع٢	V02226002000000

3.2.2.3 cāhiē

The word *cāhiē* is used in combination with the infinitive of a lexical verb to express advisability. It is also used (as described by Bhatia and Koul 2000: 60) as a polite form of the verb *cāhnā*, "want". It is derived from an old morphologically marked passive form (Schmidt 1999: 137) of *cāhnā*[20]; however, *cāhnā* is a lexical verb and other than this use of *cāhiē*, it does not deviate from the pattern of other

[20] Bailey et al. (1956: 41) report that forms with the *–iē* suffix which appears in *cāhiē*, and also in the honorific imperative, can generally be used as an impersonal passive. However the more recent and comprehensive grammar of Schmidt (1999) does not report any such usage. The auxiliary use of *cāhiē* is distinguished from the imperative in that auxiliary *cāhiē* may be marked for number.

lexical verbs. Therefore the best approach would seem to be to give *cāhiē* its own tags (it requires two tags because it agrees with the number of the object of the preceding infinitive in certain circumstances[21]). This is the approach taken in many English tagsets for modal auxiliary verbs, which are, like *cāhiē*, anomalous forms. The intermediate tags given to *cāhiē* and its plural form *cāhiē~* list them as being without person or gender, without finiteness (since it can be used with or without a following tense-bearing auxiliary), indicative, present tense and without aspect. In the descriptions, these words are defined as "cāhiē-type", rather than attempt to find an English word to accurately summarise the range of meanings associated with desirability and/or advisability that these words can convey.

Table 3.12

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Singular cāhiē-type auxiliary	VC1	فچ۱	V00101102000000
Plural cāhiē-type auxiliary	VC2	فچ۲	V00201102000000

3.2.2.4 hōnā

The verb *hōnā*, "be", is the auxiliary with the greatest range of application: the Urdu compound tenses are formed with it, and it has other uses, such as the copula. It can also be the sole verb of a main clause, but as explained above (section 3.2) it will be tagged the same whether it is a main verb or an auxiliary. The following examples from Schmidt (1999: 94, 120, 126) demonstrate the range of *hōnā*:

[21] *cāhiē* agrees with its object if that object is not followed by a postposition, and if *cāhiē* is not followed by a past tense auxiliary that itself agrees for number.

āj	mai~	daftar	mē~	nahī~	hū~
today	1-SING-NOM	office	in	not	be-PRES-1-SING

*Today I am not in the office (*hōnā *as copula with postpositional phrase)*

kal	mausam	acchā	thā
yesterday	weather	good-MASC-SING-NOM	be-PAST-MASC-SING

*Yesterday the weather was fine (*hōnā *as copula with adjective)*

ham	farś	par	sōtē	hai~
1-PLRL-NOM	floor	on	sleep-IMPERF.PART-MASC-PLRL	be-PRES-1-PLRL

*we sleep on the floor (*hōnā *as auxiliary marking the habitual present with imperfective participle)*

bāriś	hūī	hai
rain	be-PERF.PART-FEM-SING	be-PRES-3-SING

*It has rained (*hōnā *as auxiliary marking immediate past with perfective participle of* hōnā *as main verb; more literal translation would be "There has been rain")*

Some of the parts of *hōnā* are equivalent to the parts of lexical verbs; this being so, their tags are the same for those of lexical verbs, except that they commence in VH– instead of VV–. In the intermediate tagset, this difference is expressed by the verbs being marked as auxiliary instead of main. Unfortunately, Schmidt (1999) does not give a full listing of all the forms of *hōnā*, and I was forced to use other methods as outlined in 2.3. The first recourse was to refer to other works – in this case Bailey

et al. (1956). However, there were still gaps in the listing of forms of *hōnā*. When initially composing the tagset, I was forced by the underspecification in the literature to infer the existence and shape of some forms of the infinitive and imperative. In the case of an irregular verb like *hōnā*, implying its forms on the basis of regular verbal inflections involves making unwarranted assumptions. Therefore, these forms were treated as highly provisional in nature until the stage of manual tagging was undertaken (as described in the next chapter). At this point, it was possible to find examples in tagged texts for most of the forms. The polite imperative was a very notable exception to this. It did not occur in any of the manually tagged texts, and of two native speaker informants consulted on the issue, one concluded that the form *hōiyē* was not possible. However, the other informant suggested that it *was* possible. This being the case, the VHIA tag stands – since there can be no harm in maintaining the parallelism with other verbs even if this form is rare to vanishing point.

Table 3.13

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Root form *hō*	VH0	ف ه ۰	V00020002000000
Infinitive *hōnā*, masculine singular nominative	VHNM1N	ف ه ن م ۱ خ	V01125002000001
Infinitive *hōnē*, masculine singular oblique	VHNM1O	ف ه ن م ۱ ص	V01125002000003
Infinitive *hōnē*, masculine plural nominative	VHNM2	ف ه ن م ۲	V01225002000001
Infinitive *hōnī*, feminine singular nominative	VHNF1	ف ه ن ع ۱	V02125002000001
Infinitive *hōnī*, feminine plural nominative	VHNF2	ف ه ن ع ۲	V02225002000001

Masculine singular (nominative) imperfective participle *hōtā*	VHTM1N	خ١متەف	V011(1|2)(2|6)(4|0)02 (0|2)0000(0|1)
Masculine singular oblique imperfective participle *hōtē*	VHTM1O	ص١متەف	V0112600220000(3|5)
Masculine plural (nominative) imperfective participle *hōtē*	VHTM2N	خ٢متەف	V012(1|2)(2|6)(4|0)02 (0|2)0000(0|1)
Masculine plural oblique imperfective participle *hōtē*	VHTM2O	ص٢متەف	V0122600220000(3|5)
Feminine singular (nominative) imperfective participle *hōtī*	VHTF1N	خ١عتەف	V021(1|2)(2|6)(4|0)02 (0|2)0000(0|1)
Feminine singular oblique imperfective participle *hōtī*	VHTF1O	ص١عتەف	V0212600220000(3|5)
Feminine plural (nominative) imperfective participle *hōtī / hōtī~*	VHTF2N	خ٢عتەف	V022(1|2)(2|6)(4|0)02 (0|2)0000(0|1)
Feminine plural oblique imperfective participle *hōtī*	VHTF2O	ص٢عتەف	V0222600220000(3|5)
Masculine singular (nominative) perfective participle *hūā*	VHYM1N	خ١مىەف	V011(1|2)(1|6)(4|0)02 (0|1)0000(0|1)
Masculine singular oblique perfective participle *hūē*	VHYM1O	ص١مىەف	V0112600210000(3|5)
Masculine plural (nominative) perfective participle *hūē*	VHYM2N	خ٢مىەف	V012(1|2)(1|6)(4|0)02 (0|1)0000(0|1)
Masculine plural oblique perfective participle *hūē*	VHYM2O	ص٢مىەف	V0122600210000(3|5)

Feminine singular (nominative) perfective participle *hūī*	VHYF1N	فەیع١خ	V021(1|2)(1|6)(4|0)02(0|1)0000(0|1)
Feminine singular oblique perfective participle *hūī*	VHYF1O	فەیع١ص	V0212600210000(3|5)
Feminine plural (nominative) perfective participle *hūī / hūī~*	VHYF2N	فەیع٢خ	V022(1|2)(1|6)(4|0)02(0|1)0000(0|1)
Feminine plural oblique perfective participle *hūī*	VHYF2O	فەیع٢ص	V0222600210000(3|5)
First person singular subjunctive *hū~*	VHSM1	فەش١م	V10112102000000
First person plural subjunctive *hō~*	VHSM2	فەش٢م	V10212102000000
Second person singular subjunctive *hō*	VHST1	فەش١ت	V20112102000000
Second person plural subjunctive *hō*	VHST2	فەش٢ت	V20212102000000
Third person singular subjunctive *hō*	VHSV1	فەش١و	V30112102000000
Third person plural subjunctive *hō~*	VHSV2	فەش٢و	V30212102000000
Second person singular imperative *hō*	VHIT1	فەر١ت	V20113102000000
Second person plural imperative *hō*	VHIT2	فەر٢ت	V20213102000000
Second person honorific imperative *hōiyē*	VHIA	فەرا	V20(1|2)13102000000

The past participle of *hōnā*, as with that of other verbs, can be used alone as a simple past tense. The participial tags above would be used in this case. However, there is also an irregular inflected simple past tense – which, as might be expected,

differs slightly in its meaning (Bailey et al. 1956: 109; Barz 1977: 48-49 considers this to be an instance of two separate verbs with the same infinitive[22]). There is, in addition, an irregular inflected simple present tense (the only one in the whole language). These inflected forms are the basis of the compound tense system and both require separate tags, as follows. Like the regular inflected subjunctive mood, the present indicative of *hōnā* is marked for person and number but not gender.

The intermediate tags for the present tense are the same for those of the subjunctive except that the *mood* is *indicative*. In the mnemonic tags I use H to indicate the present tense, since this tense is entirely characteristic of *hōnā*.

Table 3.14

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
First person singular indicative present *hū~*	VHHM1	فهم١	V10111102000000
First person plural indicative present *hai~*	VHHM2	فهم٢	V10211102000000
Second person singular indicative present *hai*	VHHT1	فهت١	V20111102000000
Second person plural indicative present *hō*	VHHT2	فهت٢	V20211102000000
Third person singular indicative present *hai*	VHHV1	فهو١	V30111102000000
Third person plural indicative present *hai~*	VHHV2	فهو٢	V30211102000000

[22] Kellogg (1875: 232) provides etymological evidence that supports Barz's view; however, Kellogg is of the opinion that the inflected present and past tenses are most conveniently treated as parts of *hōnā*, however inconsistent with the etymology this may be. This has been my course of action.

The irregular past tense is marked for gender and number in the same way as a perfective participle, but it is a finite form. The intermediate tags are the same as those for the present tense, except that 1) *gender* is not *zero*, 2) *person* is *zero*, and 3) *tense* is *past* rather than *present*.

Table 3.15

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Masculine singular indicative past *thā*	VHPM1	ف ہ ض م ۱	V01111402000000
Masculine plural indicative past *thē*	VHPM2	ف ہ ض م ۲	V01211402000000
Feminine singular indicative past *thī*	VHPF1	ف ہ ض ع ۱	V02111402000000
Feminine plural indicative past *thī~*	VHPF2	ف ہ ض ع ۲	V02211402000000

3.2.2.5 Modal and vector verbs

Urdu possesses a number of verbs which frequently carry most of the inflection but little of the semantic content within a compound verb phrase. These are the so-called "vector verbs" (Schmidt 1999: 143-156; both verbs and compound structures are discussed in great depth by Butt 1995; see also 1.1.5.4). The class of vector verbs is closed – Schmidt discusses nine which cover most compound verbs – and therefore has a fair claim to be considered as a class of auxiliary verbs. The modal verbs (Schmidt 1999: 115-117) are also a small closed class and can also be considered auxiliary. However, these verbs do not possess any of the inflectional anomalies of the auxiliary verbs considered so far, and, furthermore, most or all can also function as the main verb of a clause, in which usage they carry all the verb

phrase's semantic content. I have thus named the class of modal and vector verbs *general auxiliary* in the tagset definitions, to distinguish them from the *special auxiliaries* discussed above. In the terms of the EAGLES attribute-value pairs, I consider them to be *semi-auxiliary*, on the grounds that "semi-auxiliary" seems a fairly reasonable description of what they are. When not used as vector verbs, these verbs are tagged as lexical verbs. Thus the distinction between lexical and general auxiliary verbs is context-dependent, unlike the lexical-special auxiliary distinction described above (see 3.2). Aside from being *semi-auxiliary*, they have the same set of features marked on them as lexical verbs.

General auxiliaries are defined as those which follow a lexical verb *in its root form*. This is true of both modal and vector verbs. Schmidt (1999) gives details of a number of two-verb constructions involving "vector verbs" with other forms of the lexical verb (e.g. the perfective participle), but these are considered to be idiomatic verb phrases and thus a feature of syntax-semantics rather than morphosyntax. Thus the vector verbs in this context would not be considered to be general auxiliaries. This boundary between general auxiliaries and other verbs that just happen to be the inflected member of a two-verb construction is somewhat arbitrary, and, in truth, something of a fiction: the distinction is a graded one[23]. However, for tagging purposes the division must be made sharp – either something is tagged one way, or it is tagged the other. The rule that a general auxiliary must follow a root-form lexical verb, however artificial a division, means that there is an unambiguous decision on whether or not a verb is a general auxiliary or a lexical verb in any particular context.

An exception must be made to this otherwise clear rule, for the verb *jānā*,

[23] Such graded distinctions are a known factor in linguistic analysis: see for example Leech's (1997b: 32) discussion of a similar fuzzy boundary in English nouns.

which forms the passive when it is preceded by the perfective participle of a lexical verb, and is considered to be functioning as a general auxiliary when it does so (Schmidt 1999: 130). Schmidt (1999: 155) lists some other exceptional cases where a vector verb is preceded by a non-root form main verb; I will not count these as general auxiliaries, since to do so would complicate the categorisation system greatly for any manual tagger or automated tagging system.

One addition must also be made. The verb *karnā* is neither a vector verb nor a modal verb but is considered to be a general auxiliary when it appears in the construction referred to by Schmidt (1999: 108) as a "conjunctive participle" and by Butt (1995) as a "participial adverb". This consists of the root of a lexical verb followed by the root of *karnā*, *kar* and is highlighted by Kachru (1990: 70) as a particularly significant structure: for this reason *kar* is in this context to be tagged as a general auxiliary.

The most prominent members of the class of general auxiliaries (other than *karnā*) are tabulated below. Note that the list is not exhaustive. For example, Schmidt gives examples of the verb *guzarnā* and *calnā* being used as vector verbs: presumably these are less common than the nine she lists as being important.

Table 3.16

Modal verbs	saknā	pānā	cuknā
Vector verbs	jānā	paRnā	nikalnā
	uThnā	baiThnā	dēnā
	lēnā	Dālnā	rakhnā

The tags for general auxiliaries are as follows. Note that some will be rare, as

there are certain compound tenses and forms that compound verbs are used in infrequently (Schmidt 1999: 152-154). However all may potentially occur.

Table 3.17

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Root form general auxiliary verb	VX0	ف‌غ۰	V00020003000000
Infinitive general auxiliary verb, masculine singular nominative	VXNM1N	ف‌غ‌ن‌م۱‌خ	V01125003000001
Infinitive general auxiliary verb, masculine singular oblique	VXNM1O	ف‌غ‌ن‌م۱‌ص	V01125003000003
Infinitive general auxiliary verb, masculine plural nominative	VXNM2	ف‌غ‌ن‌م۲	V01225003000001
Infinitive general auxiliary verb, feminine singular nominative	VXNF1	ف‌غ‌ن‌ع۱	V02125003000001
Infinitive general auxiliary verb, feminine plural nominative	VXNF2	ف‌غ‌ن‌ع۲	V02225003000001
Masculine singular (nominative) imperfective participle general auxiliary verb	VXTM1N	ف‌غ‌ت‌م۱‌خ	V011(1|2)(2|6)(4|0)03(0|2)0000(0|1)
Masculine singular oblique imperfective participle general auxiliary verb	VXTM1O	ف‌غ‌ت‌م۱‌ص	V0112600320000(3|5)
Masculine plural (nominative) imperfective participle general auxiliary verb	VXTM2N	ف‌غ‌ت‌م۲‌خ	V012(1|2)(2|6)(4|0)03(0|2)0000(0|1)

Masculine plural oblique imperfective participle general auxiliary verb	VXTM2O	فغتم٢ص	V0122600320000(3|5)
Feminine singular (nominative) imperfective participle general auxiliary verb	VXTF1N	فغتع١خ	V021(1|2)(2|6)(4|0)03 (0|2)0000(0|1)
Feminine singular oblique imperfective participle general auxiliary verb	VXTF1O	فغتع١ص	V0212600320000(3|5)
Feminine plural (nominative) imperfective participle general auxiliary verb	VXTF2N	فغتع٢خ	V022(1|2)(2|6)(4|0)03 (0|2)0000(0|1)
Feminine plural oblique imperfective participle general auxiliary verb	VXTF2O	فغتع٢ص	V0222600320000(3|5)
Masculine singular (nominative) perfective participle general auxiliary verb	VXYM1N	فغيم١خ	V011(1|2)(1|6)(4|0)03 (0|1)0000(0|1)
Masculine singular oblique perfective participle general auxiliary verb	VXYM1O	فغيم١ص	V0112600310000(3|5)
Masculine plural (nominative) perfective participle general auxiliary verb	VXYM2N	فغيم٢خ	V012(1|2)(1|6)(4|0)03 (0|1)0000(0|1)
Masculine plural oblique perfective participle general auxiliary verb	VXYM2O	فغيم٢ص	V0122600310000(3|5)
Feminine singular (nominative) perfective participle general auxiliary verb	VXYF1N	فغيع١خ	V021(1|2)(1|6)(4|0)03 (0|1)0000(0|1)

Feminine singular oblique perfective participle general auxiliary verb	VXYF1O	فغیع١ص	V0212600310000(3|5)
Feminine plural (nominative) perfective participle general auxiliary verb	VXYF2N	فغیع٢خ	V022(1|2)(1|6)(4|0)03(0|1)0000(0|1)
Feminine plural oblique perfective participle general auxiliary verb	VXYF2O	فغیع٢ص	V0222600310000(3|5)
First person singular subjunctive general auxiliary verb	VXSM1	فغ ش م١	V10112103000000
First person plural subjunctive general auxiliary verb	VXSM2	فغ ش م٢	V10212103000000
Second person singular subjunctive general auxiliary verb	VXST1	فغ ش ت١	V20112103000000
Second person plural subjunctive general auxiliary verb	VXST2	فغ ش ت٢	V20212103000000
Third person singular subjunctive general auxiliary verb	VXSV1	فغ ش و١	V30112103000000
Third person plural subjunctive general auxiliary verb	VXSV2	فغ ش و٢	V30212103000000
Second person singular imperative general auxiliary verb	VXIT1	فغرت١	V20113103000000
Second person plural imperative general auxiliary verb	VXIT2	فغرت٢	V20213103000000

Second person honorific imperative general auxiliary verb	VXIA	فرغا	V20(1|2)13103000000

3.3 Adjectives

The EAGLES guidelines for adjectives recommend the attributes *degree, gender, number* and *case*, and offer as optional extensions the attributes *inflection-type*, *use*, and *NP-function*.

Adjectives in Urdu are not regularly marked for degree (that is, whether they are positive, comparative, or superlative). As Schmidt (1999: 46-49) describes, this is mostly done in a phrase[24]. There are Persian-derived suffixes that indicate comparative and superlative, but these only apply to Perso-Arabic roots and are often written and/or pronounced as separate words to their root. Furthermore, the meaning they give is frequently intensive rather than comparative/superlative (see Schmidt 1999: 256). Because of all this, I have decided not to treat these as comparatives, but rather as an aspect of derivational morphology, which should therefore be excluded from the tagset in accordance with the design features discussed in the previous chapter. Therefore, the attribute of *degree* always has the value 1 in the intermediate tagset.

Like nouns, adjectives may be marked for gender or unmarked for gender. However, unlike unmarked nouns, unmarked adjectives receive no inflection *at all* and always have the same form. It should also be noted that adjectives lack inherent gender. For this reason it was decided to give all unmarked adjectives the same

[24] Bailey et al. (1956: 19) concur with this view.

gender/number/case markup[25].

As far as marked adjectives are concerned, there is again the problem of tag-to-meaning many-to-one and one-to-many mapping – but with adjectives it is, if anything, even greater a problem than it was with nouns. There is no oblique-vocative distinction at all (Schmidt 1999: 36 goes so far as to say that "An adjective modifying a vocative noun is in the oblique case"), and the entire spectrum of gender/number/case combinations are covered by three suffixes, listed below.

Table 3.18

Suffix	Indicates…
–ā(~)	Masculine nominative singular
–ī(~)	Feminine (all cases, both numbers)
–ē(~)	Masculine nominative plural Masculine oblique/vocative, both numbers

However, in line with the principle of tagging for function rather than for form, there will be eight tags for all the functional gender/number/case combinations, rather than three tags to tag each of the forms above. This is justifiable on the following grounds: for masculine nominative adjectives there is a clear singular/plural distinction (e.g. *dāyā~ – dāē~*, "right"), and for masculine singular adjectives there is a clear nominative/oblique distinction (e.g. *baRā – baRē*, "big"). The masculine/feminine distinction is clear throughout the paradigm (e.g. *acchā – acchī*,

[25] This is the opposite decision to that taken for unmarked nouns. The difference is that the gender of unmarked nouns becomes apparent when verbs and marked adjectives agree with them, whereas nothing will ever indicate any trace of gender in an unmarked adjective.

"good"). Therefore these distinctions should be applied to all adjectives, including those (e.g. feminine adjectives) where the said distinctions are not clearly marked. However, no adjective at all distinguishes the vocative case, so marking it would not be justifiable[26].

Thus the tagset does not distinguish vocative adjectives from oblique adjectives (or participle forms of verbs: see above). In the intermediate tagset, this is represented using the OR and bracket operators, as described in the EAGLES guidelines (Leech and Wilson 1999: 71), as (3 | 5). Otherwise the tagset has been constructed in much the same way as that for nouns.

There is no extra *markedness* attribute, as was needed for nouns, because an unmarked adjective can be annotated by placing a zero in the gender/number/case fields (in the EAGLES notation, 0 means "this attribute is not applicable"). Markedness could theoretically have been coded in the attribute *inflection type*, but this attribute has been primarily designed for "adjectival inflection in the Germanic languages German, Dutch and Danish" (Leech and Wilson 1999: 67). It captures such variation as that found in German, where the case/number/gender suffix given to an adjective varies depending on whether the adjective is preceded by a definite article, an indefinite article, or neither. Since Urdu lacks this type of variation, and also lacks articles, this attribute is not used.

Of the two remaining attributes, *NP-function* would seem to be irrelevant because it annotates the position of the adjective relative to its head noun. In the first case, this is syntactic information, which this tagset is excluding. Furthermore, in Urdu, the adjective precedes its noun (Schmidt 1999: 188).

---

[26] In the absence of a formal difference, it is likely to be impossible to identify vocative adjectives reliably. It should be noted that even identifying vocative nouns proved problematic (see Chapter 6).

This leaves *use*, which refers to whether an adjective may be used in attributive or predicative positions only. The default value for this is naturally *both*. In the absence of a specification in the EAGLES guidelines, I represent this with 0. There are a number of common Perso-Arabic adjectives in Urdu that can only be used in predicative position (Schmidt 1999: 37), for which this attribute can take the value 2. This is the rationale for including this attribute, which is however a prime candidate to be underspecified in a practical subtagset. It is anticipated that it will be difficult for a POS tagger to detect predicate-only adjectives. Since the predicate-only adjectives are Perso-Arabic, it ought to follow that they are all unmarked adjectives. However, this is a point on which Schmidt (1999) is silent. For this reason, tags have been included for predicate-only adjectives that are marked for gender/number/case. These may need to be removed if it turns out from the data that they do indeed describe non-existent categories, as I suspect[27].

The table that follows sums up the attribute-value sets used for adjectives.

Table 3.19

Value	*ii) gender*	*iii) number*	*iv) case*	*vi) use*
0	Not marked	Not marked	Not marked	Both
1	Masculine	Singular	Nominative	
2	Feminine	Plural		Predicative
3 | 5			Oblique/ vocative	

If gender is 0, then number and case are too; if gender is 1 or 2, number and case cannot be 0. This reflects the fact that gender number and case marking are fused

[27] See also 4.2.1.1 and 4.2.2.3.

in one suffix, which is either present or absent. This gives us (1 x 1 x 1 x 2) + (2 x 2 x 2 x 2) = 18 tags.

Table 3.20

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Marked masculine singular nominative adjective	JJM1N	ص‌ص‌م۱‌خ	AJ0111000
Marked masculine singular oblique / vocative adjective	JJM1O	ص‌ص‌م۱‌ص	AJ011(3|5)000
Marked masculine plural nominative adjective	JJM2N	ص‌ص‌م۲‌خ	AJ0121000
Marked masculine plural oblique / vocative adjective	JJM2O	ص‌ص‌م۲‌ص	AJ012(3|5)000
Marked feminine singular nominative adjective	JJF1N	ص‌ص‌ع۱‌خ	AJ0211000
Marked feminine singular oblique / vocative adjective	JJF1O	ص‌ص‌ع۱‌ص	AJ021(3|5)000
Marked feminine plural nominative adjective	JJF2N	ص‌ص‌ع۲‌خ	AJ0221000
Marked feminine plural oblique / vocative adjective	JJF2O	ص‌ص‌ع۲‌ص	AJ022(3|5)000
Marked predicate-only masculine singular nominative adjective	JPM1N	ص‌خ‌م۱‌خ	AJ0111020
Marked predicate-only masculine singular oblique / vocative adjective	JPM1O	ص‌خ‌م۱‌ص	AJ011(3|5)020
Marked predicate-only masculine plural nominative adjective	JPM2N	ص‌خ‌م۲‌خ	AJ0121020
Marked predicate-only masculine plural oblique / vocative adjective	JPM2O	ص‌خ‌م۲‌ص	AJ012(3|5)020

Marked predicate-only feminine singular nominative adjective	JPF1N	صخع۱خ	AJ0211020
Marked predicate-only feminine singular oblique / vocative adjective	JPF1O	صخع۱ص	AJ021(3|5)020
Marked predicate-only feminine plural nominative adjective	JPF2N	صخع۲خ	AJ02210201
Marked predicate-only feminine plural oblique / vocative adjective	JPF2O	صخع۲ص	AJ022(3|5)020
Unmarked adjective	JJU	صنص	AJ0000000
Unmarked predicate-only adjective	JPU	صخن	AJ0000020

3.4 Pronouns and determiners

The EAGLES guidelines treat pronouns and determiners together as a single category, although one of the recommended attributes, *category*, distinguishes between them. Since in Urdu the distinction is not clear (particularly in the area of third person pronouns), I also treat this category as being single at the most fundamental level. The difference between what is considered a determiner and what is considered a pronoun is not made in the EAGLES guidelines, which say "different analyses for different languages entail separating [these parts of speech] out in different ways" (Leech and Wilson 1999: 63). For Urdu, I have mostly followed Schmidt – who does not have a separate "determiner" category – in the divisions I make. However, I have classed together all third person pronouns/demonstratives, interrogative and relative pronouns/determiners, because these form sets of words

displaying morphological symmetry (see 3.4.2).

Schmidt counts pronouns such as *yah*, *vah*, as both personal pronouns and determiners. However, for the purposes of the tagset, the division should be sharp; therefore I have limited the "personal pronouns" category to the first and second persons. The justification for this is given in section 3.4.1.1. I have also diverged from Schmidt in classing together a number of her minor categories of pronoun under the covering title "other" for the purposes of this tagset definition.

This gives the following groups of pronoun/determiner-like words

- first and second person personal pronouns
- third person pronouns/demonstratives, interrogative and relative pronouns and determiners
- reflexive pronouns
- other pronouns and determiners

There is one pronoun, *āp* (a kind of honorific personal pronoun) which does not fit unproblematically into any of these categories. Discussion is devoted to this pronoun in section 3.4.1.2 below.

The EAGLES guidelines suggest eleven attributes for pronouns and determiners. The obviously relevant ones are *category*, *person*, *gender*, *number*, *possessive*, and *case*. *Pronoun-type*, *special pronoun-type*, *wh-type*, and *determiner-type* are also relevant, since they can be used to distinguish the smaller groups of words proposed above; finally *politeness* is relevant as well (since pronouns have the same system of politeness as verbs). All the attributes are relevant, but not all values

are used[28], and of course the structuring of the intermediate tagset does not fit completely with the structuring of the categories in Urdu. For example, the attributes *special pronoun-type* and *wh-type* create subsets of *int./rel.* and *pers./refl.*, which are values of the *pronoun-type* and *determiner-type* attributes. The attributes and values used are as below. There are a grand total of 106 tags defined in this section.

Table 3.21

Value	*i) person*	*ii) gender*	*iii) number*	*iv) possessive*	*v) case*	*vi) category*
1	First	Masculine	Singular	Singular	Nominative	Pronoun
2	Second	Feminine	Plural	Plural		Determiner
3					Oblique	Both

Value	*vii) pron.-type*	*viii) det.-type*	*ix) special pron.-type*	*x) wh-type*	*xi) politeness*
1	Demonstrative	Demonstrative	Personal	Interrogative	Polite
2	Indefinite	Indefinite	Reflexive	Relative	Familiar
3	Possessive	Possessive	Reciprocal	Exclamatory	
4	Int./Rel.	Int./Rel.			
5	Pers./Refl.				

The groups of pronoun-like words are now considered in turn (greater

[28] The *case* attribute possesses a value *oblique*, but I have not used this for oblique-case pronouns for two reasons: to maintain consistency with the noun tags, and because the term "oblique case" is used in the EAGLES guidelines to refer to a case that is used for the direct object of a verb or a preposition – which is not the function of the Urdu oblique. I continue to use the value for *dative* for the Urdu oblique case. Schmidt (1999: 15) fails to include the vocative in the list of cases taken by pronouns, so the value for *vocative* is not used for pronouns.

explanation of the attributes and values above is given as and when it is needed).

3.4.1 First and second person personal pronouns

The issue of what exactly constitutes a personal pronoun is not an easy one in the context of the grammar of Urdu as presented by Schmidt (1999). Therefore, in this section, before discussing the tags of the personal pronouns I elaborate on how I drew the boundary of this category, justifying the minor claim that the pronouns *vah* and *yah* (and their various inflected forms) are *not* personal pronouns, as stated by Schmidt (1999)[29]. I first consider these third person pronouns (3.4.1.1), and subsequently the problematic honorific pronoun *āp* (3.4.1.2). In 3.4.1.3 I deal with the tagging of *mai~* and *tū*, the remaining words in the category of personal pronouns.

3.4.1.1 The non-existence of third person personal pronouns

Urdu has no third person personal pronouns. The demonstrative pronouns/determiners are used in their place. This is claimed contrary to Schmidt, who states (1999: 15) that "The demonstrative pronouns *ye* and *vo* are identical in form to the personal pronouns *ye* and *vo* (meaning 'he', 'she', 'it')". However the differences in behaviour between these pronouns and the first and second person pronouns that I list below, also drawn from Schmidt, make it clear that the statement that began this section is justified.

[29] Incidentally, I have in this the support of Kellogg (1875: 168-181), who also deals with *āp* and *yah / vah* separately to the personal pronouns.

- There are absolutely no differences in case / number inflection between the third person pronouns and the demonstratives (Schmidt 1999: 16)
- In a perfective transitive sentence (the type that some, such as Dixon 1994, would class as "ergative"), a third person pronoun subject appears in the oblique case (like a noun); but a first or second person subject pronoun is in the nominative case at all times (Schmidt 1999: 22)
- The third person pronouns take special plural oblique forms before the postposition *nē* (Schmidt 1999: 22), whereas the first and second do not
- There are no possessive adjectives corresponding to the third person pronouns, whereas there are such adjectives corresponding to the first and second person pronouns (Schmidt 1999: 24)

On these grounds, I exclude the third person pronouns from consideration as personal pronouns, and deal with them as demonstratives/determiners, etc. (see section 3.4.2).

3.4.1.2 The problematic honorific pronoun **āp**

The case of *āp*, the second person honorific pronoun, is by no means as clear as that of the third person pronouns. While the fact of its identical appearance with the reflexive pronoun (also *āp*: see 3.4.3[30]) suggests that, like the third person pronouns, it may be best classified elsewhere, there are two very good reasons for regarding *āp* as a personal pronoun like *mai~* and *tū*.

[30] Kellogg (1875: 180-181) gives the common etymology of (what he sees as) these two pronouns in a single Sanskrit word.

The first is semantic. Semantically and pragmatically, *āp* has a very similar meaning to *tū* and its plural form *tum* – they both mean "you"[31]. The second reason is syntactic. From the examples of *āp* given by Schmidt (1999), it would appear that *āp* has a very similar distribution to *mai~* and *tū*. It is used, for example, as the subject of a sentence; the reflexive pronoun *āp*, by contrast, can never be the subject of a sentence for obvious reasons.

There are, on the other hand, a number of reasons to regard *āp* as unlike *mai~* and *tū* and either identical or at least more akin to the cognate reflexive pronoun (also *āp*. All are morphological. Firstly, *āp* (both the honorific and reflexive pronoun) does not have separate nominative and oblique cases, whereas *mai~* and *tū* do. Secondly, as noted above, *mai~* and *tū* have associated possessive adjectives. *āp* also has such a possessive adjective, *apnā*, but this is only used reflexively (see 3.4.3). When the usage is honorific, possession is expressed phrasally with the postposition *kā*, "of". Thirdly, while *mai~* and *tū* agree with verbal forms distinct from those used with nouns or third person pronouns, *āp* does not, always taking identical verbal inflections to the third person. This is what we would expect if it were simply a special usage of a reflexive pronoun.

So then, is *āp* a second person personal pronoun or is it a special usage of the reflexive pronoun? Either position is tenable. The syntax and semantics of the case supports the former approach while the morphology backs up the latter approach. The EAGLES guidelines cannot help in choosing between them, since this problem is an idiosyncrasy of Urdu: we would therefore not expect it to be covered by a standard

[31] This is a generalisation: Schmidt (1999: 18) describes how *āp* may also be used as a third person honorific pronoun. But this usage seems from her description to be more marginal (this has been confirmed by consulting a native speaker informant).

drawn up for a set of languages which do not include Urdu. Ultimately, this is a case where an arbitrary decision must be taken: the decision I took was not to treat *āp* as a personal pronoun along with *mai~* and *tū*. However, although arbitrary, this decision is consistent: *āp* will *always* be treated separately in this way[32].

In fact the non-reflexive *āp* will be given the tag PA, so that in terms of the hierarchy of the tagset, it is categorised neither with the personal nor the reflexive pronouns, but in a separate subdivision of the pronoun category. This is, to an extent, another arbitrary decision: PPA could have been an equally reasonable tag, emphasising the similarity of syntactic function with *mai~* and *tū*, or PRA, emphasising the similarity of its case inflections to those of the reflexive pronouns, which likewise show no difference between the nominative and oblique cases. However, to impose either of these interpretations might prove theoretically controversial, in breach of a stated design principle[33].

Note however that in terms of the intermediate tagset, *āp* is still treated as a personal pronoun, because the things that it will map onto in other languages will be personal pronouns. Its *number* is (1 | 2), on the grounds that it may refer to one person or to more than one. Note that the intermediate tagset for pronouns contains a value, *politeness*; *āp* has been listed as *polite*, whereas the intermediate tags for *tū* as given in the next section contain the value for *familiar*.

---

[32] It might be thought that the creation of a tag for an honorific form of the imperative, along with singular and plural non-honorific forms, is inconsistent with this. However, this is not the case: the polite / honorific imperative does not necessarily co-occur with *āp*.

[33] An additional consideration is that Bhatia (1987: 82), Schmidt (1999) and Kellogg (1875) all treat the honorific use of *āp* separately from its reflexive use.

Table 3.22

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Honorific pronoun (*āp*)	PA	اپ	PD20(1|2)00150101

3.4.1.3 The tagging of first and second person personal pronouns

Thus, the subcategory of first and second person personal pronouns contains only the pronouns *mai~* and *tū*, and inflectionally related forms such as their plurals and possessive forms. All tags in this subcategory begin PP– (or PG– for possessives).

Personal pronouns are not marked for gender: as with verbs, that which is marked for person is not marked for gender. (The "M" in the tags below signifies "first person", not "masculine".) They *are* marked for number and case.

As noted in the preceding section, the intermediate tagset for pronouns contains an attribute of *politeness*. All pronouns in this section are given as *familiar*, to distinguish their intermediate tags from that for *āp*. In practice, the singular/plural distinction is often also used to indicate formality in the second person pronouns (Bhatia and Koul 2000: 35-36); *tum* may apply to one or more than one person. However, the EAGLES guidelines suggest[34] that such a pragmatic usage of the number distinction may still be encoded as a number distinction. This is what I have done, tagging *tum* as plural, on the basis that for purposes of inflection it is the number of the pronoun, not the number of its referent, that counts.

[34] This suggestion is made with regard to French, whose second-person pronouns have similar uses.

Table 3.23

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
First person singular nominative personal pronoun (*mai~*)	PPM1N	ض‌ض‌م۱‌خ	PD10101150100
First person singular oblique personal pronoun (*mujh*)	PPM1O	ض‌ض‌م۱‌ص	PD10103150100
First person plural nominative personal pronoun (*ham*)	PPM2N	ض‌ض‌م۲‌خ	PD10201150100
First person plural oblique personal pronoun (*ham*)	PPM2O	ض‌ض‌م۲‌ص	PD10203150100
Second person singular nominative personal pronoun (*tū*)	PPT1N	ض‌ض‌ت۱‌خ	PD20101150102
Second person singular oblique personal pronoun (*tujh*)	PPT1O	ض‌ض‌ت۱‌ص	PD20103150102
Second person plural nominative personal pronoun (*tum*)	PPT2N	ض‌ض‌ت۲‌خ	PD20201150102
Second person plural oblique personal pronoun (*tum*)	PPT2O	ض‌ض‌ت۲‌ص	PD20203150102

There are possessive adjectives corresponding to the personal pronouns above. While the intermediate tagset must treat these as pronouns, within the Urdu tagset they could have been treated as adjectives (as has been done with some other determiner-like pronouns; see below). However, this has not been done, since the possessive adjectives have person. This is not true for any adjectival form, and thus the possessive adjectives are better classed as personal pronouns.

As they are adjectival, they may be marked for gender, number and case. The

case and *gender* attributes indicate the features that are in agreement with the head noun rather than inherent features of the pronoun. The *number* attribute is also for agreement; the inherent number of the possessive adjective itself is shown by the attribute *possessive*.

Table 3.24

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
First person singular masculine singular nominative possessive adjective (*mērā*)	PGM1M1N	ض‌ق‌م۱م۱خ	PD11111203000
First person singular masculine singular oblique possessive adjective (*mērē*)	PGM1M1O	ض‌ق‌م۱م۱ص	PD11113203000
First person singular masculine plural nominative possessive adjective (*mērē*)	PGM1M2N	ض‌ق‌م۱م۲خ	PD11211203000
First person singular masculine plural oblique possessive adjective (*mērē*)	PGM1M2O	ض‌ق‌م۱م۲ص	PD11213203000
First person singular feminine singular nominative possessive adjective (*mērī*)	PGM1F1N	ض‌ق‌م۱ع۱خ	PD12111203000
First person singular feminine singular oblique possessive adjective (*mērī*)	PGM1F1O	ض‌ق‌م۱ع۱ص	PD12113203000
First person singular feminine plural nominative possessive	PGM1F2N	ض‌ق‌م۱ع۲خ	PD12211203000

adjective (*mērī*)			
First person singular feminine plural oblique possessive adjective (*mērī*)	PGM1F2O	ض‌ق‌م١ع٢ص	PD12213203000
First person plural masculine singular nominative possessive adjective (*hamārā*)	PGM2M1N	ض‌ق‌م٢م١خ	PD11121203000
First person singular masculine singular oblique possessive adjective (*hamārē*)	PGM2M1O	ض‌ق‌م٢م١ص	PD11123203000
First person singular masculine plural nominative possessive adjective (*hamārē*)	PGM2M2N	ض‌ق‌م٢م٢خ	PD11221203000
First person singular masculine plural oblique possessive adjective (*hamārē*)	PGM2M2O	ض‌ق‌م٢م٢ص	PD11223203000
First person singular feminine singular nominative possessive adjective (*hamārī*)	PGM2F1N	ض‌ق‌م٢ع١خ	PD12121203000
First person singular feminine singular oblique possessive adjective (*hamārī*)	PGM2F1O	ض‌ق‌م٢ع١ص	PD12123203000
First person singular feminine plural nominative possessive adjective (*hamārī*)	PGM2F2N	ض‌ق‌م٢ع٢خ	PD12221203000
First person singular feminine plural oblique possessive adjective (*hamārī*)	PGM2F2O	ض‌ق‌م٢ع٢ص	PD12223203000
Second person singular masculine singular	PGT1M1N	ض‌ق‌ت١م١خ	PD21111203000

nominative possessive adjective (*tērā*)			
Second person singular masculine singular oblique possessive adjective (*tērē*)	PGT1M1O	ض‌ق‌ت‌١‌م‌١‌ص	PD21113203000
Second person singular masculine plural nominative possessive adjective (*tērē*)	PGT1M2N	ض‌ق‌ت‌١‌م‌٢‌خ	PD21211203000
Second person singular masculine plural oblique possessive adjective (*tērē*)	PGT1M2O	ض‌ق‌ت‌١‌م‌٢‌ص	PD21213203000
Second person singular feminine singular nominative possessive adjective (*tērī*)	PGT1F1N	ض‌ق‌ت‌١‌ع‌١‌خ	PD22111203000
Second person singular feminine singular oblique possessive adjective (*tērī*)	PGT1F1O	ض‌ق‌ت‌١‌ع‌١‌ص	PD22113203000
Second person singular feminine plural nominative possessive adjective (*tērī*)	PGT1F2N	ض‌ق‌ت‌١‌ع‌٢‌خ	PD22211203000
Second person singular feminine plural oblique possessive adjective (*tērī*)	PGT1F2O	ض‌ق‌ت‌١‌ع‌٢‌ص	PD22213203000
Second person plural masculine singular nominative possessive adjective (*tumhārā*)	PGT2M1N	ض‌ق‌ت‌٢ م‌١‌خ	PD21121203000
Second person singular masculine singular oblique possessive	PGT2M1O	ض‌ق‌ت‌٢‌م‌١‌ص	PD21123203000

adjective (*tumhārē*)			
Second person singular masculine plural nominative possessive adjective (*tumhārē*)	PGT2M2N	ض ق ت ٢ م ٢ خ	PD21221203000
Second person singular masculine plural oblique possessive adjective (*tumhārē*)	PGT2M2O	ض ق ت ٢ م ٢ ص	PD21223203000
Second person singular feminine singular nominative possessive adjective (*tumhārī*)	PGT2F1N	ض ق ت ٢ ع ١ خ	PD22121203000
Second person singular feminine singular oblique possessive adjective (*tumhārī*)	PGT2F1O	ض ق ت ٢ ع ١ ص	PD22123203000
Second person singular feminine plural nominative possessive adjective (*tumhārī*)	PGT2F2N	ض ق ت ٢ ع ٢ خ	PD22221203000
Second person singular feminine plural oblique possessive adjective (*tumhārī*)	PGT2F2O	ض ق ت ٢ ع ٢ ص	PD22223203000

3.4.2 Third person pronouns/demonstratives, interrogative and relative pronouns and determiners

This class of pronouns consists of all those pronouns that fall into the parallel classes of what Schmidt (1999: 39) calls “symmetrical y-v-k-j word sets”. These classes contain a variety of pronouns and adjectives that are of similar form, the first letter indicating what set they belong to, thus:

- y or a vowel indicates the set of proximal demonstratives (*this*, *now*, etc.)
- v or t[35] indicates the set of distal demonstratives (*that*, *then*, etc.)
- k indicates the set of interrogatives (*who*, *what*, *how*, etc.)
- j indicates the set of relative words (*who*, *where*, *whither*, etc.)

Thus, in Urdu there is 1) a significant distinction between proximal and distal words, for which there is no distinction in the EAGLES guidelines; 2) a significant distinction between interrogatives and relatives, which is only made by the EAGLES guidelines at the secondary optional level (the recommended features include only *int./rel.*, presumably on the basis that these have similar forms in many European languages – the so-called wh-words). This means that the intermediate tags for these pronouns are not as elegant as they might be, and the tags for the y-set and the v-set are the same[36]. However, I will make this distinction in the Urdu tags, which begin with P followed by the letter of the relevant y-v-k-j set.

The proximal and distal demonstratives have not been distinguished for any other language that I am aware of. For example, no English tagset I know of distinguishes *here/hither* from *there/thither*. However, most distinguish *where/whither* from the non-interrogative/relative words. In Urdu, the "near~far" phonological pattern is much more consistent – there are no odd pairs such as English

[35] The words that begin in *t* are actually members of a former set of correlative words, as Schmidt explains. They will be tagged as members of the "far" set, because they function as such (and the words in *v* are used as correlatives).

[36] There does not seem to be any easy way to avoid this by adding an attribute: how could an attribute distinguishing between *proximal*, *distal*, and *neither* be distinguished linguistically from the already existing attributes dealing with type? As stated above, I do not wish to add values to pre-existing attributes.

this~that – and is formally of an equal degree to the "demonstrative~interrogative" distinction. Furthermore, there is a difference of usage between the proximal and distal sets – the latter are used in correlative clauses where the former are not[37]. For this reason I tag the four-way distinction, since it would be odd to arbitrarily merge two of what are on a language-internal basis clearly different categories.

The pronouns in the y-v-k-j sets are used as demonstrative pronouns and third person personal pronouns (so *yah* and *vah*[38] mean both "this" and "that" and "he/she/it"). They can also act as determiners within a noun phrase. I have not tagged these uses differently, because this would fall under the heading of syntactic information, which this tagset does not include. See also section 3.4.1.1.

I do not, as Schmidt (1999: 38-41) does, characterise the determiner-usage as adjectival, since these pronouns do not display gender agreement, as adjectives (including other members of the y-v-k-j sets) do. They are however marked for case and number[39]. They also have the peculiarity that their plurals have a third case-like form, which appear solely before the postposition *nē* (which indicates the subject of an ergative-type clause). This is tagged separately (and, like the proximal/distal distinction, not distinguished in the intermediate tagset, since it is difficult to see how this could be achieved).

There are two interrogative pronouns, both beginning in *k*; one means "what" and one means "who". They both receive the same tags, since tagging an animacy distinction would be odd when this is done nowhere else in the tagset.

---

[37] There is one minor exception to this (Schmidt 1999: 206).

[38] These two words are almost always transcribed as *yē* and *vō*, which is how they are pronounced. However, the spellings with *h* are closer to the Perso-Arabic (Bhatia and Koul 2000: 36).

[39] However, in the nominative case the singular and plural forms are identical.

In the intermediate tagset, following what is done for such pronouns in the example English tagset given in the EAGLES guidelines I give *person* as zero, and for the k-set words the *wh-type* is –2[40], since *kyā* may also be exclamatory. The *category* attribute is *both*, because these words are both pronouns and determiners.

Table 3.25

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Singular nominative proximal demonstrative pronoun (*yah*)	PY1N	خ۱یض	PD00101311000
Singular oblique proximal demonstrative pronoun (*is*)	PY1O	ص۱یض	PD00103311000
Plural nominative proximal demonstrative pronoun (*yah*)	PY2N	خ۲یض	PD00201311000
Plural oblique proximal demonstrative pronoun (*in*)	PY2O	ص۲یض	PD00203311000
Plural oblique proximal demonstrative pronoun before *nē* (*inhō~*)	PY2E	ے۲یض	PD00203311000
Singular nominative distal demonstrative pronoun (*vah*)	PV1N	خ۱وض	PD00101311000
Singular oblique distal demonstrative pronoun (*us*)	PV1O	ص۱وض	PD00103311000

[40] Leech and Wilson (1999: 71) explain this notation of exclusion thus: "the negative operator [is] signalled by the minus (-), so that –4 means 'all values of this attribute except the fourth'."

Plural nominative distal demonstrative pronoun (*vah*)	PV2N	ض۰۲خ	PD00201311000
Plural oblique distal demonstrative pronoun (*un*)	PV2O	ض۰۲ص	PD00203311000
Plural oblique distal demonstrative pronoun before *nē* (*unhō~*)	PV2E	ض۰۲ے	PD00203311000
Singular nominative interrogative pronoun (*kyā, kaun*)	PK1N	ض ک ۱ خ	PD001013440-20
Singular oblique interrogative pronoun (*kis*)	PK1O	ض ک ۱ ص	PD001033440-20
Plural nominative interrogative pronoun (*kyā, kaun*)	PK2N	ض ک ۲ خ	PD002013440-20
Plural oblique interrogative pronoun (*kin*)	PK2O	ض ک ۲ ص	PD002033440-20
Plural oblique interrogative pronoun before *nē* (*kinhō~*)	PK2E	ض ک ۲ ے	PD002033440-20
Singular nominative relative pronoun (*jō*)	PJ1N	ض ج ۱ خ	PD00101344020
Singular oblique relative pronoun (*jis*)	PJ1O	ض ج ۱ ص	PD00103344020
Plural nominative relative pronoun (*jō*)	PJ2N	ض ج ۲ خ	PD00201344020
Plural oblique relative pronoun (*jin*)	PJ2O	ض ج ۲ ص	PD00203344020
Plural oblique relative pronoun before *nē* (*jinhō~*)	PJ2E	ض ج ۲ ے	PD00203344020

There are also in the y-v-k-j sets a number of words that are more like

determiners than pronouns, i.e. they take adjectival inflection and cannot stand alone as pronouns. However they behave in some respects more like adjectives, e.g. they can be predicative rather than attributive. In terms of the EAGLES guidelines they are best characterised within the pronoun/determiner category. They correspond to English words like "such", "this/that much/many" and so on. In terms of the Urdu tagset, I have classified them as JD – determiner-like adjectives[41].

Table 3.26

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Masculine singular nominative proximal demonstrative adjective (*itnā, aisā*)	JDYM1N	خ۱میتص	PD01101201000
Masculine singular oblique proximal demonstrative adjective (*itnē, aisē*)	JDYM1O	ص۱میتص	PD01103201000
Masculine plural nominative proximal demonstrative adjective (*itnē, aisē*)	JDYM2N	خ۲میتص	PD01201201000
Masculine plural oblique proximal demonstrative adjective (*itnē, aisē*)	JDYM2O	ص۲میتص	PD01203201000
Feminine singular nominative proximal demonstrative	JDYF1N	خ۱عیتص	PD02101201000

[41] This somewhat arbitrary decision is taken on the basis that most JD– words take adjectival inflection. It would not necessarily make less sense linguistically to classify them as PD– (determiner-like pronouns).

adjective (*itnī, aisī*)			
Feminine singular oblique proximal demonstrative adjective (*itnī, aisī*)	JDYF1O	ص١عيتص	PD02103201000
Feminine plural nominative proximal demonstrative adjective (*itnī, aisī*)	JDYF2N	خ٢عيتص	PD02201201000
Feminine plural oblique proximal demonstrative adjective (*itnī, aisī*)	JDYF2O	ص٢عيتص	PD02203201000
Masculine singular nominative distal demonstrative adjective (*utnā, vaisā*[42])	JDVM1N	خ١وتص	PD01101201000
Masculine singular oblique distal demonstrative adjective (*utnē, vaisē*)	JDVM1O	ص١وتص	PD01103201000
Masculine plural nominative distal demonstrative adjective (*utnē, vaisē*)	JDVM2N	خ٢وتص	PD01201201000
Masculine plural oblique distal demonstrative adjective (*utnē, vaisē*)	JDVM2O	ص٢وتص	PD01203201000
Feminine singular nominative distal demonstrative adjective (*utnī, vaisī*)	JDVF1N	خ١وتص	PD02101201000
Feminine singular oblique distal demonstrative adjective (*utnī, vaisī*)	JDVF1O	ص١وتص	PD02103201000
Feminine plural nominative distal demonstrative	JDVF2N	خ٢وتص	PD02201201000

[42] The word *taisā* (from the old correlative set) also appears in some idioms. It is tagged JDV– as well.

adjective (*utnī, vaisī*)			
Feminine plural oblique distal demonstrative adjective (*utnī, vaisī*)	JDVF2O	ص٢عوتص	PD02203201000
Masculine singular nominative interrogative adjective (*kitnā, kaisā*)	JDKM1N	خ١مكتص	PD011012040-20
Masculine singular oblique interrogative adjective (*kitnē, kaisē*)	JDKM1O	ص١مكتص	PD011032040-20
Masculine plural nominative interrogative adjective (*kitnē, kaisē*)	JDKM2N	خ٢مكتص	PD012012040-20
Masculine plural oblique interrogative adjective (*kitnē, kaisē*)	JDKM2O	ص٢مكتص	PD012032040-20
Feminine singular nominative interrogative adjective (*kitnī, kaisī*)	JDKF1N	خ١عكتص	PD021012040-20
Feminine singular oblique interrogative adjective (*kitnī, kaisī*)	JDKF1O	ص١عكتص	PD021032040-20
Feminine plural nominative interrogative adjective (*kitnī, kaisī*)	JDKF2N	خ٢عكتص	PD022012040-20
Feminine plural oblique interrogative adjective (*kitnī, kaisī*)	JDKF2O	ص٢عكتص	PD022032040-20
Masculine singular nominative relative adjective (*jitnā,*	JDJM1N	خ١مجتص	PD01101204020

jaisā)			
Masculine singular oblique relative adjective (*jitnē, jaisē*)	JDJM1O	ص‌ت‌ج‌م۱‌ص	PD01103204020
Masculine plural nominative relative adjective (*jitnē, jaisē*)	JDJM2N	ص‌ت‌ج‌م۲‌خ	PD01201204020
Masculine plural oblique relative adjective (*jitnē, jaisē*)	JDJM2O	ص‌ت‌ج‌م۲‌ص	PD01203204020
Feminine singular nominative relative adjective (*jitnī, jaisī*)	JDJF1N	ص‌ت‌ج‌ع۱‌خ	PD02101204020
Feminine singular oblique relative adjective (*jitnī, jaisī*)	JDJF1O	ص‌ت‌ج‌ع۱‌ص	PD02103204020
Feminine plural nominative relative adjective (*jitnī, jaisī*)	JDJF2N	ص‌ت‌ج‌ع۲‌خ	PD02201204020
Feminine plural oblique relative adjective (*jitnī, jaisī*)	JDJF2O	ص‌ت‌ج‌ع۲‌ص	PD02203204020

3.4.3 Reflexive pronouns

Unlike many European languages, Urdu reflexive pronouns are not personal. That is, they have the same form regardless of the person of the pronoun they are reflexing back to. There are two reflexive pronouns, both tagged the same, a reciprocal pronoun (which only appears within a postpositional phrase) and a reflexive possessive adjective. The reflexive possessive adjective is classed with the other possessive adjectives in the hierarchy given in 3.14. See also the discussion of the honorific usage of *āp* in section 3.4.1.2 above.

Table 3.27

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Reflexive pronoun (*āp, xud*)	PRF	ض‌ر‌ج	PD00000150200
Reciprocal pronoun (*āpas*)	PRC	ض‌ر‌پ	PD00000150300
Masculine singular nominative reflexive possessive adjective (*apnā*)	PGRM1N	ض‌ق‌ر‌م۱‌خ	PD01101203000
Masculine singular oblique reflexive possessive adjective (*apnē*)	PGRM1O	ض‌ق‌ر‌م۱‌ص	PD01103203000
Masculine plural nominative reflexive possessive adjective (*apnē*)	PGRM2N	ض‌ق‌ر‌م۲‌خ	PD01201203000
Masculine plural oblique reflexive possessive adjective (*apnē*)	PGRM2O	ض‌ق‌ر‌م۲‌ص	PD01203203000
Feminine singular nominative reflexive possessive adjective (*apnī*)	PGRF1N	ض‌ق‌ر‌ع۱‌خ	PD02101203000
Feminine singular oblique reflexive possessive adjective (*apnī*)	PGRF1O	ض‌ق‌ر‌ع۱‌ص	PD02103203000
Feminine plural nominative reflexive possessive adjective (*apnī*)	PGRF2N	ض‌ق‌ر‌ع۲‌خ	PD02201203000
Feminine plural oblique reflexive possessive adjective (*apnī*)	PGRF2O	ض‌ق‌ر‌ع۲‌ص	PD02203203000

3.4.4 Other pronouns and determiners

In this miscellaneous group of pronouns are included two indefinite pronouns, *kōī* and *kuch*, which may function as pronouns or determiners (just as *yah* and *vah* do). Also included in the PN* category is *sab*, "all", which has an inflected oblique plural (like numerals – see section 3.9) which is tagged as PNO.

Table 3.28

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Nominative indefinite pronoun (*kōī, kuch, sab*)	PNN	ضغخ	PD00001322000
Oblique indefinite pronoun (*kīsī*, *kuch, sabhō~*)	PNO	ضغص	PD00003322000

There is also a tag for indefinite determiners. Two words in this class are *zyādah* "more" and *kāfī* "enough". Following Schmidt (1999) these are classed broadly as adjectives for two reasons: to keep them in line with the possessive adjectives, which are determiners; and because they can also function as adverbs (see section 3.6 below), which is characteristic of adjectives. These are not marked for gender, number or case.

Table 3.29

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Indefinite determiner	JD	صت	PD00000202000

3.5 Articles

Urdu lacks articles. However, some phrases borrowed from Arabic contain the clitic Arabic definite article, which receives the single tag AL (the spelling of the Arabic article). I have not included a C in this tag, as I have done for other clitics (see section 3.12), because this would make the tag less transparent. The use of the AT intermediate tag could be queried here, because the use of the Arabic definite article in Urdu does not parallel that of, for example, *the* in English or *le/la/les* in French. For example, the Arabic definite article is only found with Arabic loanwords[43], whereas of course *the* can appear with the vast majority of nouns in English. However, on balance it seems that this disadvantage is outweighed by the advantage of indicating that the Arabic definite article in Urdu does do pretty much what other languages' articles do. Khoja et al.'s (2001) Arabic tagset does not have a separate tag for the article, but considers definiteness a feature of nouns: this would not be an appropriate approach for Urdu because non-Arabic nouns cannot be made definite by use of the Arabic definite form.

Table 3.30

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Arabic definite article	AL	ال	AT1000

[43] See Chapter 5 for a discussion of guidelines on the tagging of Arabic loanwords.

3.6 Adverbs

As with verbs, there are lexical and non-lexical adverbs, which will be considered in turn.

In the EAGLES guideline, the recommended attribute for adverbs is *degree*[44], which is not relevant morphologically to Urdu (as discussed with reference to adjectives: see 3.3 above). However, the remaining three features are relevant, and have been included. These are *adverb-type*, which distinguishes general and degree adverbs, and *polarity* and *wh-type*, which distinguish interrogative and relative pronouns. The following summarises the features used in the intermediate tagset. There are a total of 13 adverb tags.

Table 3.31

Value	*ii) adverb-type*	*iii) polarity*	*iv) wh-type*
1	General	wh-type	Interrogative
2	Degree	Non-wh-type	Relative
3			Exclamatory

3.6.1 Lexical adverbs

In Urdu these are of two sorts: adverbs which are derived from adjectives by inflecting them to their masculine oblique form or adding a Persian or Arabic loaned

[44] This use of "degree" (i.e. inflected superlative or comparative) should be clearly distinguished from the use of "degree adverb" below (i.e. words with meanings such as "very", "more").

derivational suffix[45] (RRJ), and adverbs which are not (RR). While this unfortunately violates the principle of not including derivational information, this distinction has been included in the tagset for two reasons.

Firstly, it helps avoid ambiguity, since an adverb derived from an adjective has the same form as that adjective in its masculine singular oblique form (see Schmidt 1999: 57). If adjectival adverbs were marked RR, this would lead to a wide ambiguity between RR and JJM1O, which would make non-adjectival adverbs ambiguous as well! Using a separate tag, there is only an RRJ~JJM1O ambiguity, which significantly reduces the scope of the ambiguity. Although this is a pragmatic consideration which should probably be included at the subtagset level, it involves creating a distinction rather than collapsing one, and must thus exist in the top level tagset.

However, there is another motivation for the RRJ tag, which is that it is necessary to maintain theoretical neutrality. It is possible that some analyst might wish to treat the RRJ adverbs as if they were actually adjectives – that is, identify them with JJ– categories instead of RR. Indeed Bailey et al. (1956: 18) come close to saying this. The principle of theoretical neutrality must here override the principle of excluding derivational information.

The EAGLES intermediate tags for RR and RRJ are the same.

[45] Words which may function as adjectives or as adverbs without any morphological modification (e.g. *pās*) would be marked with either an adjective tag or RR, depending on context; the RRJ tag is not intended for such words.

Table 3.32

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
General adverb	RR	ل ل	AV0120
General adverb derived from adjective	RRJ	ل ل ص	AV0120

3.6.2 Non-lexical adverbs

Urdu possesses some degree adverbs[46], which indicate "very" or "more" or a similar notion when they occur before an adjective. These include *bahut*, "very", *zyādah,* "more" and *kāfī*, "quite"[47].

Table 3.33

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Degree adverb	RD	ل ت	AV0220

There are also a number of modal adverbs. Those listed by Schmidt (1999: 62) are *śāyad*, "maybe", *zarūr*, "certainly", *bhī*, "also, too", *phir*, "again", and *sirf*, "only". This category is given a subcategory for negative adverbs following the example of Schmidt. I do this, even though the EAGLES guidelines suggest that negative "particles" should be tagged in the "Unique" class, because Urdu has three negative adverbs – which means that they are *not* unique, unlike (say) the English

[46] The adjective *baRā* may be used pragmatically as a degree adverb, but as it still agrees with the head noun of the noun phrase, it is grammatically an adjective and should be tagged as such. The y-v-k-j adjectives *itnā* and *kitnā* are also used thus.

[47] These last two are also indefinite determiners (see above).

"not". However, for these words I give a Unique intermediate tag as well.

Note that the intermediate tag is the same as that for general adverbs. This is because the EAGLES tagset has no means with which to distinguish modal adverbs. I did not add an additional attribute because, as was the case with the pronoun tags, such an attribute would be from a linguistic point of view identical in purpose to the already existing *type* attribute.

Table 3.34

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Modal adverb	RM	لط	AV0120
Negative modal adverb (*nahī~*, *nah*, *mat*)	RMN	لطن	AV0120 , U20000

There are also a number of adverbs in the y-v-k-j sets of words, with meanings of time, place and manner such as "now", "then", "thus", "thither", etc. This includes all the interrogative and relative adverbs. Adverbs derived from adjectives of the y-v-k-j sets (see above) have separate tags. Again, many of the intermediate tags are identical.

Table 3.35

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Proximal demonstrative adverb (*ab*, *yahā~*, *idhar*, *yū~*)	RY	لی	AV0120
Proximal demonstrative adverb derived from adjective (*aisē)*	RYJ	لیص	AV0120

Distal demonstrative adverb (*tab*, *vahā~*, *udhar*, *tyū~*)	RV	لو	AV0120
Distal demonstrative adverb derived from adjective (*vaisē)*	RVJ	لوص	AV0120
Interrogative adverb (*kab*, *kahā~*, *kidhar*, *kyō~*)	RK	لک	AV011-2
Interrogative adverb derived from adjective (*kaisē)*	RKJ	لکص	AV011-2
Relative adverb (*jab*, *jahā~*, *jidhar*, *jū~*)	RJ	لج	AV0112
Relative adverb derived from adjective (*jaisē)*	RJJ	لجص	AV0112

3.7 Adpositions

It should be noted at the outset that I treat as adpositions those elements of Urdu that some writers (e.g. Kellogg 1875, Butt 1995) describe as case suffixes or clitics. This is firstly because Schmidt (1999), the model of the language being used, does so. Secondly, however, treating *nē* (among other markers) as adpositions allows theoretical neutrality to be maintained on the question of whether Urdu displays ergativity[48].

The EAGLES guidelines give only one attribute for adpositions, *Type*, which has a range of recommended and optional values: *preposition*, *fused preposition-*

[48] See also the discussion of the ergativity controversy in 1.1.5.4 and the discussion of noun cases and the etymology of postpositions in 3.1.3.

article, *postposition*, and *circumposition*. The second and fourth of these do not apply to Urdu, which lacks articles[49] and circumpositions. The vast majority of Urdu adpositions are postpositions, but there are some prepositions borrowed from Persian and Arabic (Schmidt 1999: 68, 250, 267), so this attribute is relevant.

There are two other issues. The first is that of *izāfat* (Bhatia and Koul 2000: 339; Schmidt 1999: 246-247). The izāfat is a Persian enclitic (pronounced as a shorter form of *–ē–*) which in some circumstances can be considered a preposition: it links two nouns in a possessive relationship, although the phrase thus produced may often have a different meaning to a phrase produced with the native Urdu postposition *kā*. However, the izāfat may also join a noun to an adjective, in which case it is not so clearly accurate to describe it as a preposition parallel to the prepositions in European languages for which the EAGLES guidelines were compiled. A better way to treat izāfat is in the context of the *Unique* category of miscellaneous one-member word-classes, discussed below.

The second issue is that in Urdu, the postposition *kā* can be marked for number/gender/case agreement (Schmidt 1999: 68-69). It does not agree with the noun it governs, but with the head noun of the noun phrase that contains its postposition phrase. This is not a phenomenon allowed for by the EAGLES guidelines as they now stand. *kā* takes the same inflectional endings as marked adjectives (having the forms *kā*, *kē*, and *kī*). Therefore, it is necessary for the same number/gender/case categories to be distinguished by the tagset for postpositions as for adjectives[50]. This means that the intermediate tagset contains three more attributes

[49] That is, Urdu lacks articles other than the Arabic definite article in borrowed words and constructions. See section 3.5.

[50] This *only* applies to the agreement categories, not to the *use* attribute, for example.

than are suggested in the EAGLES guidelines.

Table 3.36

Value	*i) type*	*ii) gender*	*iii) number*	*iv) case*
0		Not marked	Not marked	Not marked
1	Preposition	Masculine	Singular	Nominative
2		Feminine	Plural	
3	Postposition			
3 | 5				Oblique/ vocative

Since prepositions do not inflect for gender/number/case, there are 2 + (2 x 2 x 2) = 10 tags. Although Schmidt does not specify whether there are any postpositions other than *kā* with gender/case marking, my native speaker informants report that there are not, so the marked tags are restricted to *kā*, *kē* and *kī*.

Table 3.37

Description	**Tag (Roman)**	**Tag (Perso-Arabic)**	**Intermediate Tag**
Preposition	IB	جگ	AP1000
Unmarked postposition	II	جج	AP3000
Marked masculine singular nominative postposition	IIM1N	ججم۱خ	AP3111
Marked masculine singular oblique / vocative postposition	IIM1O	ججم۱ص	AP311(3|5)
Marked masculine plural nominative postposition	IIM2N	ججم۲خ	AP3121

Marked masculine plural oblique / vocative postposition	IIM2O	ججم۲ص	AP312(3|5)
Marked feminine singular nominative postposition	IIF1N	ججع۱خ	AP3211
Marked feminine singular oblique / vocative postposition	IIF1O	ججع۱ص	AP321(3|5)
Marked feminine plural nominative postposition	IIF2N	ججع۲خ	AP3221
Marked feminine plural oblique / vocative postposition	IIF2O	ججع۲ص	AP322(3|5)

3.8 Conjunctions

The EAGLES guidelines suggest that conjunctions be classified firstly for whether they are coordinating or subordinating, and then secondly as one of four co-ordinating types or one of three subordinating types. I have disregarded the attribute for subordinate-type, since it was developed for German and does not seem relevant to Urdu subordinating conjunction as described by Schmidt (1999: 223-227). Urdu correlative conjunctions (such as *bhī...bhī*, *yā…yā*) do not have initial and non-initial forms, so those features are also not needed. This gives three types of conjunctions: simple coordinating, correlative coordinating, and subordinate. Note that phrases involving the relative j-set of pronouns, adjectives and adverbs are often translated by conjunctions, but are not to be tagged as such. The following are the values used in the intermediate tags:

Table 3.38

Value	*i) type*	*ii) coord-type*
1	Coordinating	Simple
2	Subordinating	Correlative

The three tags are as follows:

Table 3.39

Description	**Tag (Roman)**	**Tag (Perso-Arabic)**	**Intermediate Tag**
Coordinating conjunction	CC	ت ت	C110
Correlative coordinating conjunction	CCC	ت ت ت	C120
Subordinating conjunction	CS	ت ش	C200

The EAGLES guidelines (Leech and Wilson 1999: 68) specify that a conjunction is correlative when it is at the start of the first of a pair of correlated clauses. The conjunction at the start of the second half of the pair is then a simple coordinating conjunction (CC)[51]. This practice will be followed to ensure compliance with the EAGLES guidelines.

3.9 Numerals

The EAGLES guidelines give numerals as a separate major part-of-speech, but

[51] In fact the EAGLES guidelines on this point are significantly more complicated. However, the remainder of the recommendations are concerned with handling phenomena that do not occur in Urdu.

say that “In some languages (e.g. Portuguese) this category is not normally considered to be a separate part of speech, because it can be subsumed under others… We recognise that in some tagsets Numeral may therefore occur as subcategory within other parts of speech” (Leech and Wilson 1999: 65). This approach seems sensible for Urdu, where numerals display very much the behaviour of adjectives. However, for purposes of the intermediate tagset, the numeral class *has* been used, since it contains the very useful attribute *type*. In fact, all the EAGLES attributes have been used (though of course, not all of their values). For *case*, the oblique / vocative value (3 | 5) is used, as with adjectives. There are a total of 19 tags in this category.

Table 3.40

Value	*i) type*	*ii) gender*	*iii) number*	*iv) case*	*v) function*
1	Cardinal	Masculine	Singular	Nominative	Pronoun
2	Ordinal	Feminine	Plural		Determiner
3				Oblique	Adjective
5				Vocative	

Cardinal numbers function as grammatically unmarked determiner-like adjectives (Schmidt 1999: 228). However, they can appear in the oblique plural – with the same suffix as an unmarked noun – to express totality (Schmidt 1999: 10-11). There is therefore an additional tag for this (indicated only by O, since there is no oblique singular to make a contrast). In the intermediate tagset I have given their *function* as *determiner*, in line with the determiners that are in the pronoun category above. Numerals are to be tagged as below, even if written as figures rather than words (and whatever set of figures are used: Urdu uses both the Western European

and the Arabic-Indic digits).

Table 3.41

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Cardinal number	JDNU	ص ت ع ن	NU10002
Oblique cardinal number	JDNUO	ص ت ع ن ص	NU10002
Masculine singular nominative ordinal number	JDNM1N	ص ت ع م ۱ خ	NU21112
Masculine singular oblique / vocative ordinal number	JDNM1O	ص ت ع م ۱ ص	NU211(3|5)2
Masculine plural nominative ordinal number	JDNM2N	ص ت ع م ۲ خ	NU21212
Masculine plural oblique / vocative ordinal number	JDNM2O	ص ت ع م ۲ ص	NU212(3|5)2
Feminine singular nominative ordinal number	JDNF1N	ص ت ع ۱ خ	NU22112
Feminine singular oblique / vocative ordinal number	JDNF1O	ص ت ع ۱ ص	NU221(3|5)2
Feminine plural nominative ordinal number	JDNF2N	ص ت ع ۲ خ	NU22212
Feminine plural oblique / vocative ordinal number	JDNF2O	ص ت ع ۲ ص	NU222(3|5)2

Urdu has a fairly wide range of words for fractions (there are for example words for “plus one quarter” (*savā*), “less one quarter” (*paun, paunā*), “one half” (*ādh, ādhā*), “one and a half” (*DēRh*), “plus one half” (*sāRhē*)), which can modify cardinal numerals as well as nouns. They are therefore tagged separately (although the intermediate tags are not all distinct). Most are unmarked, but two are marked. Two others can also function as nouns, in which case they should receive standard noun

tagging.

Table 3.42

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Unmarked fraction	JDFU	ص ت ب	NU10002
Masculine singular nominative fraction	JDFM1N	ص ت ب م ١ خ	NU11112
Masculine singular oblique / vocative fraction	JDFM1O	ص ت ب م ١ ص	NU111(3|5)2
Masculine plural nominative fraction	JDFM2N	ص ت ب م ٢ خ	NU11212
Masculine plural oblique / vocative fraction	JDFM2O	ص ت ب م ٢ ص	NU112(3|5)2
Feminine singular nominative fraction	JDFF1N	ص ت ب ع ١ خ	NU12112
Feminine singular oblique / vocative fraction	JDFF1O	ص ت ب ع ١ ص	NU121(3|5)2
Feminine plural nominative fraction	JDFF2N	ص ت ب ع ٢ خ	NU12212
Feminine plural oblique / vocative fraction	JDFF2O	ص ت ب ع ٢ ص	NU122(3|5)2

3.10 Interjections

The EAGLES guidelines do not recommend any additional attributes for the class of interjections. Nor have I introduced any of my own. There is thus one tag. The mnemonic tag represent the spelling of *ō* (Schmidt 1999: 217), which has been selected as a representative interjection.

Table 3.43

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Interjection	AU	او	I

3.11 Punctuation

The EAGLES guidelines allow three options for the markup of word-external punctuation: firstly, to use a single tag for all punctuation marks (the obligatory-attribute-only approach); secondly, to give each punctuation mark its own separate tag; and thirdly, to group punctuation marks into a smaller number of tags according to how they may position in a sentence. The first approach I rejected on the grounds that it needlessly excluded potentially useful information. The third approach, likewise, tags different punctuation marks in the same way. Since punctuation marks can be tagged utterly unambiguously – a comma is always a comma – this is needless. The decision was therefore taken to give each punctuation mark a unique tag. This tag is, in fact, the same as the punctuation mark itself (a practice also adhered to in, for example, the C7 tagset: see 2.1.2.1). However, since the tagset is designed to operate in Unicode texts, more forms of punctuation can be distinguished (for example, opening and closing quotation marks). Some of these distinctions may be finer than is necessary (e.g. that between square and normal brackets is useless if one simply wishes to search for brackets in general) but it would be trivial to design search software that could treat the two tags as alike, or to map to a subtagset that collapsed these to a single "bracket" category. There are 13 tags in this section. The EAGLES guidelines underspecify the value of the one attribute, stating values only for the full stop, comma, and question mark, so I have inferred it (using letters when the available

digits ran out).

Table 3.44

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Full stop (U+06D4)	.	۔	PU1
Comma (U+060C)	,	،	PU2
Question mark (U+061F)	?	؟	PU3
Exclamation mark (U+0021)	!	!	PU4
Colon (U+003A)	:	:	PU5
Semi-colon (U+061B)	;	؛	PU6
Neutral quotation mark (U+0022)	"	"	PU7
Open quotation mark (U+201C)	“	”	PU8
Close quotation mark (U+201D)	”	“	PU9
Open parenthesis (U+0028)	(	)	PUA
Close parenthesis (U+0029)	)	(	PUB
Open square bracket (U+005B)	[	]	PUC
Close square bracket (U+005D)	]	[	PUD

For all punctuation marks, the Unicode of the Perso-Arabic tag is the same as that of the punctuation mark being tagged[52]. The Roman tags for full stop, comma,

[52] Single and double quotation marks, however, should both be tagged using the same three tags.

question mark, and semi-colon consist of a different Unicode character to the punctuation mark being tagged, but otherwise likewise use the same Unicode.

With regard to paired punctuation – the quotation marks and brackets – there is a point to be made as regards directionality. The Unicode Standard specifies (Unicode 1996: 6-4) that in bi-directional text[53] the same character – i.e. the same Unicode value – should represent the opening member of the pair whatever its appearance, and the same with the closing member of the pair. That is, the code U+0028 (OPENING PARENTHESIS) ought always to be the first of the pair, and be rendered as " (" in left-to-right text, such as English, and as ") " in right-to-left text, such as Urdu. Other paired punctuation marks should function similarly[54]. Therefore for each of these marks, the Roman and Perso-Arabic tags are mirror images of one another, though they are encoded by the same numeric value.

This could potentially create confusion when an analyst tags text by hand, inasmuch as the (Roman) tag will have the opposite appearance to the (Perso-Arabic) symbol in the actual text[55]. However, this will not be problematic when tagging is automated, "right" and "left" meaning nothing to a computerised tagger.

There remain some problematic points, for example, the ellipsis (…), angle bracket speech marks, and braces. These have not been given tags for now, on the basis that no Urdu text I have yet seen contains these symbols. However, nor does any

---

[53] The corpus texts will be bi-directional because their SGML markup will be left-to-right.

[54] The software used to create many of the Urdu texts for the EMILLE project is unfortunately inconsistent in its implementation of this part of the standard. It *does* reverse the glyph for the quotation marks in Urdu text. It *does not* reverse the glyphs for the brackets. Thus, the directionality of paired punctuation marks cannot be relied upon to be consistent.

[55] This problem could, of course, be avoided by using the Perso-Arabic version of the tagset, and thus Perso-Arabic directionality for the tags as well as the text. See Appendix 3.

work on Urdu rule out their use, so extra punctuation tags may prove necessary.

3.12 Unique/unassigned (including particles, clitics and tags)

The Unique category in the EAGLES guidelines is meant to contain words that are members of a one-word category; for example, the infinitive marker *to* or the existential *there* in English. I will first outline the general nature of the tags defined in this part of the tagset (3.12.1), before going into some depth on the problem that motivated the creation of one particular unique category, that of nongrammatical lexical element: the *zimmah dār* problem (3.12.2).

3.12.1 Tags for the unique categories

The EAGLES guidelines contain for this category no recommended attributes, and only one attribute: *unique-type*, whose values denote unique classes relevant to European languages. Obviously, the same unique categories will not all be relevant to Urdu. In fact, only the U2 tag (*unique-negative particle*) seems relevant – this has already been dealt with in the discussion of adverbs. For this reason, I have created another attribute for Urdu unique types, and added it to the end. Each value simply denotes one of the classes I describe. I have also included below clitic forms of words listed above. Clitics are to be separated from their host word during tagging and given their own tags. For the intermediate tagset, clitics that have corresponding independent forms have two tags: the tag of the word that they are "short for", and the unique tag. Which is used in any given mapping to the intermediate tagset will depend on the purpose of the mapping. The intermediate tags, together with some examples

of the categories[56], follow:

Table 3.45

Value	*ii) Urdu unique type*	***Value***	*ii) Urdu unique type*	***Value***	*ii) Urdu unique type*
1	Question marker (*kyā*)	2	Izāfat	3	Sentence tag-word (e.g. *sāhī*)
4	Clitic postposition (*(h)ē(~)*)	5	Pre-multiplicative clitic numeral	6	Contrastive emphatic particle (*tō*)
7	Exclusive emphatic particle (*hī*)	8	Clitic exclusive emphatic particle (*(h)ī(~)*)	9	Inclusive emphatic particle (*bhī*)[57]
A	Multiplicative marker (*gunā*)	B	Adjectival particle (*sā*)	C	Adjectival / occupational particle (*vālā*)
D	Persian compound-forming conjunction (*ō*)	E	Nongrammatical lexical element		

Because of the presence in this category of words showing adjectival agreement, I have had to add attributes for the adjectival agreement categories of *gender*, *number* and *case* to the intermediate tags, as was done with the postpositions: see 3.7 above. This takes the full number of attributes to five. From these attributes, a total of 34 tags are defined.

It would have been possible to include some of the non-lexical verbal elements (e.g. the present tense of *hōnā*) in the Unique category. However I did not do this, because giving them verbal tags allowed them to be given the agreement categories of verbs. As with *gā*, some of the words listed in this category (e.g. *vālā*) have been described as suffixes by Schmidt (1999) but are written as separate words.

[56] The names of the particles are taken from Schmidt (1999).

[57] Note that Schmidt does not specify the distinction between *bhī* as an inclusive emphatic particle, and *bhī* as a modal adverb "too". This will be investigated in the following chapter.

The Urdu tags are as follows (with the exception of that for the nongrammatical lexical element, which is postponed to the next section). As with all categories of very small membership, I list the actual words. Some of the categories are marked for number/gender/case; these are all given the same intermediate tag. Tags either consist of a two-character mnemonic or, for the particles, of X (for "unclassified") followed by a one-character subclass. The adjective-forming elements are considered as non-lexical adjectival elements (JX–). Clitics are given the same tag as their full-length forms, but with an appended C.

Table 3.46

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Question marker *kyā*	QQ	ککک	U01000
Izāfat[58]	ZZ	ےےے	U02000
Sentence tag-word[59]	TT	ٹٹ	U03000
Clitic postposition[60] *ē*, *ē~*, *hē~*	IIC	ۓۓۓ	U04000 (AP3000)
Pre-multiplicative clitic cardinal number *du-*, *ti-*, *cau-*	JDNUC	چنعتص	U05000 (NU10002)
Contrastive emphatic particle *tō*	XT	غت	U06000

[58] See discussion under *Adpositions*. The izāfat is not always written (or pronounced: Schmidt 1999:247), but where it is written it is to be treated as other clitics.

[59] This category is rather more open than the other "unique" categories, and may in certain circumstances be ambiguous with adverbs.

[60] A form of *kō* added to a pronoun.

Exclusive emphatic particle *hī*	XH	غه	U07000
Clitic exclusive emphatic particle *ī*, *ī~*, *hī~*	XHC	غهچ	U08000 (U07000)
Inclusive emphatic particle *bhī*	XB	غب	U09000
Masculine singular nominative multiplicative marker *gunā*	JXGM1N	صغگم۱خ	U0A111
Masculine singular oblique / vocative multiplicative marker *gunē*	JXGM1O	صغگم۱ص	U0A11(3|5)
Masculine plural nominative multiplicative marker *gunē*	JXGM2N	صغگم۲خ	U0A121
Masculine plural oblique / vocative multiplicative marker *gunē*	JXGM2O	صغگم۲ص	U0A12(3|5)
Feminine singular nominative multiplicative marker *gunī*	JXGF1N	صغگع۱خ	U0A211
Feminine singular oblique / vocative multiplicative marker *gunī*	JXGF1O	صغگع۱ص	U0A21(3|5)
Feminine plural nominative multiplicative marker *gunī*	JXGF2N	صغگع۲خ	U0A221
Feminine plural oblique / vocative multiplicative marker *gunī*	JXGF2O	صغگع۲ص	U0A22(3|5)
Masculine singular nominative adjectival particle *sā*	JXSM1N	صغسم۱خ	U0B111
Masculine singular oblique / vocative adjectival particle *sē*	JXSM1O	صغسم۱ص	U0B11(3|5)

Masculine plural nominative adjectival particle *sē*	JXSM2N	ص غ س م ٢ خ	U0B121
Masculine plural oblique / vocative adjectival particle *sē*	JXSM2O	ص غ س م ٢ ص	U0B12(3|5)
Feminine singular nominative adjectival particle *sī*	JXSF1N	ص غ س ع ١ خ	U0B211
Feminine singular oblique / vocative adjectival particle *sī*	JXSF1O	ص غ س ع ١ ص	U0B21(3|5)
Feminine plural nominative adjectival particle *sī*	JXSF2N	ص غ س ع ٢ خ	U0B221
Feminine plural oblique / vocative adjectival particle *sī*	JXSF2O	ص غ س ع ٢ ص	U0B22(3|5)
Masculine singular nominative adjectival / occupational particle *vālā*[61]	JXVM1N	ص غ و م ١ خ	U0C111
Masculine singular oblique / vocative adjectival / occupational particle *vālē*	JXVM1O	ص غ و م ١ ص	U0C11(3|5)
Masculine plural nominative adjectival / occupational particle *vālē*	JXVM2N	ص غ و م ٢ خ	U0C121
Masculine plural oblique / vocative adjectival / occupational particle *vālē*	JXVM2O	ص غ و م ٢ ص	U0C12(3|5)

[61] This element is the source of the English word / suffix "wallah" (Kachru 1990: 70), which may help the reader to gain some grasp on its meaning.

Feminine singular nominative adjectival / occupational particle *vālī*	JXVF1N	صغوع١خ	U0C211
Feminine singular oblique / vocative adjectival / occupational particle *vālī*	JXVF1O	صغوع١ص	U0C21(3|5)
Feminine plural nominative adjectival / occupational particle *vālī*	JXVF2N	صغوع٢خ	U0C221
Feminine plural oblique / vocative adjectival / occupational particle *vālī*	JXVF2O	صغوع٢ص	U0C22(3|5)
Persian compound-forming conjunction (*ō*)	OO	وو	U0D000

3.12.2 The *zimmah dār* problem[62]

Words that contain an orthographic space which does not actually represent a word break – principally Persian loans such as *zimmah dār*, "responsible", *xūb tarīn*, "best", and *ham zāt*, "of the same caste"[63] – cause a problem for tokenisation as described in 2.2.6.1. This was solved by the decision to treat every orthographic space as a word break, so that *zimmah dār*, etc., are treated as two tokens. However, this leads to another problem, greater if anything, concerned with tagging. How are the two elements to be tagged?

[62] This problem is referred to as such because it was first encountered during an attempt to manually tag a sentence from Schmidt (1999) containing the word *zimmah dār* using an early trial version of the tagset.

[63] All examples from Schmidt (1999: 248-256).

As it happens, *zimmah*, *xūb* and *zāt* are independent words ("duty", "good" and "caste" respectively) and could be given the appropriate tags, nominal and adjectival. The problem then becomes, what to do with *dār*, *tārīn* and *ham*? The former two could be given some tag to indicate that they were adjective forming clitics or affixes, and the prefix *ham* could be marked up as an adverb (according to Haq 2001 the part of speech of *ham* when it occurs independently). However, this has two drawbacks. Firstly, it breaks with the design principle that no derivational information will be included in the tagset by analysing the component morphemes of complex words – for *zimmah dār* etc. are words, not phrases. The word *zimmah dārī*, "responsibility", is clear evidence of this – it has been created by a morphological process (suffixation of *–ī*) and morphological processes apply to words, not to syntactic phrases[64]. Also, the single word *zimmah dār* has been given two tags in this approach – a contravention of the "one word, one tag" principle[65].

Secondly, it introduces inconsistency into the tagging. The derivational information would be present for some words formed with the relevant Persian derivational morphemes, but not for all, because not all words formed with them contain the superfluous orthographic token break. Examples of single-token derived words include *samajh**dār***, "sensible", *kam**tarīn***, "least", and ***ham**dardī*, "sympathy". If *zimmah* and *dār* are to be tagged separately, then for consistency *samajh* would also have to be tagged separately – opening up whole vistas of morphological analysis that are utterly irrelevant to part-of-speech tagging. Indeed, going down this road subverts the entire enterprise: we would find ourselves engaged in derivational analysis instead

[64] This property of morphological processes is discussed by Katamba (1993: 217 and elsewhere).

[65] A more minor difficulty is the possibility that someday, some word followed by a free-standing *dār* or *tarīn* might prove not to exist as a word on its own, and thus be untaggable.

of morphosyntactic analysis.

To take the opposite approach to tagging *zimmah dār*, we might mark a single tag for the whole word (JJU in this case) – however this also breaks the "one word, one tag" principle as there is now an untagged token and multiword tag. The best solution to the problem (although far from ideal) would seem to be to use some kind of special tag on the first part of the two-token word to indicate that this is a case of the *zimmah dār* problem, and put the tag we would like to give to the whole thing on the second token[66].

This tag will be LL, the "nongrammatical lexical element" listed in the previous section, and it will be applied thus[67]:

zimmah_LL dār_JJU

samajhdār_JJU

xūb_LL tarīn_JJU

kamtarīn_JJU

ham_LL zāt_JJU

hamdardī_NNUF1N

The first element is described as a nongrammatical lexical element because while it does not contribute to the morphosyntax of the two-token word, it does contribute to its meaning. Therefore it is entirely lexical in nature. It is to be hoped

[66] Since *dār*, *tarīn* and other affixes involved in the *zimmah dār* problem are derivational suffixes, it is they that determine the part of speech; thus it makes sense for them to carry the actual tag.

[67] I use an underscore format to link the words and their tags for clarity in the examples given here; in practice an XML/SGML markup would be used.

that the usage of the LL tag can be restricted to one context: alongside a relatively small number of affixes such as *dār*.

The formal definition of the LL tag follows.

Table 3.47

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Nongrammatical lexical element	LL	ظظ	U0E000

3.13 Residual

The remaining categories (called "residual" in the EAGLES guidelines) cover, quite simply, everything else. This comprises various semi-linguistic and non-Urdu elements. There are 8 such tags. Although the EAGLES guidelines allows for these elements having number and gender, I have not included this: if such an element is inflected as a verb, noun or adjective, then it may be considered sufficiently a part of that category to be tagged as such. This particularly applies to acronyms and abbreviations. Thus, the second and third EAGLES attributes, *number* and *gender*, are zero in the intermediate tags below. Every value from the first EAGLES attribute, *type*, has been used; with the exception of FX and FS, each tag bears the name of the value in the intermediate tagset it is mapped onto.

The tag for "foreign words" is meant to cover words from other languages written in the Urdu alphabet. It is *not* meant to cover the large number of Persian, Arabic and English loanwords that exist in Urdu, although it remains to be seen how sharp this distinction can be made in actual tagging. The tag for "non-Perso-Arabic string" is for foreign words in other alphabets, or for other non-Perso-Arabic

incursions into the text. FU is a catch-all "Unclassified" category, although it is to be hoped that the vast majority of tokens will be catered for by at least one of the other tags outlined in this chapter.

Table 3.48

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Foreign word	FF	دد	R100
Non-Perso-Arabic string	FX	دل	R100
Formula (e.g. mathematical)	FO	در	R200
Letter of the alphabet	FZ	دت	R300
Other symbol	FS	دح	R300
Acronym[68]	FA	دا	R400
Abbreviation	FB	دم	R500
Other unclassifiable non-Urdu element	FU	دن	R600

3.14 The tagset defined as a hierarchy

One of the design principles was that the tagset should be fully decomposable and hierarchical. That it is decomposable is demonstrated by the fact that it is possible to set up mappings from Roman to Perso-Arabic as described in Appendix 3. The hierarchy is shown diagrammatically below. It is structured around the Urdu tagset as

[68] When manual tagging was undertaken, the FA tag was never used. It therefore remains to be seen whether the category of "acronym" is applicable to Urdu.

described above, with the aim of keeping similar tags (and thus similar categories) as close together as possible. Sometimes arbitrary decisions were necessary: for example, the decision to put the common/proper distinction in nouns higher than the marked/unmarked distinction is a fairly arbitrary one.

I have used abbreviations to represent ranges of inflectional elements in the hierarchy as follows:

Table 3.49

[NOUN]	M1N, M1O, M1V, M2N, M2O, M2V, F1N, F1O, F1V, F2N, F2O, F2V
[ADJ]	M1N, M1O, M2N, M2O, F1N, F1O, F2N, F2O
[VERB]	M1, M2, T1, T2, V1, V2

Other ranges of person/case/gender/number inflectional categories are given in [square brackets] in the hierarchy. I have also used the symbol | to represent the end of a tag, in contrasts such as II| versus IIM1N. The format of the hierarchy diagram is taken from Leech (1997b: 28). Punctuation is not viewed as forming part of the hierarchy.

Table 3.50: The tagset as a hierarchy

Word	→	N	→	N	→	M	→ [NOUN]	– marked common noun
					→	U	→ [NOUN]	– unmarked common noun
			→	P	→	M	→ [NOUN]	– marked proper noun
					→	U	→ [NOUN]	– unmarked proper noun
	→	V	→	V	→	0		– root form lexical verb

→ N → [M1N, M1O, M2, F1, F2]

– infinitive lexical verb

→ T → [ADJ] – imperfective participle lexical verb

→ Y → [ADJ] – perfective participle lexical verb

→ S → [VERB] – subjunctive lexical verb

→ I → [T1, T2, A] – imperative lexical verb

→ G → [M1, M2, F1, F2] – future auxiliary

→ R → [M1, M2, F1, F2] – durative auxiliary

→ C → [1, 2] – cāhiē-type auxiliary

→ H → 0 – root form of *hōnā*

→ N → [M1N, M1O, M2, F1, F2]

– infinitive of *hōnā*

→ T → [ADJ] – imperfective participle of *hōnā*

→ Y → [ADJ] – perfective participle of *hōnā*

→ S → [VERB] – subjunctive of *hōnā*

→ I → [T1, T2, A] – imperative of *hōnā*

→ [VERB] – present tense of *hōnā*

→ P → [M1, M2, F1, F2]– past tense of *hōnā*

→ X → 0 – root form general auxiliary

→ N → [M1N, M1O, M2, F1, F2]

– infinitive general auxiliary

→ T → [ADJ] – imperfective participle general auxiliary

→ Y → [ADJ] – perfective participle general auxiliary

→ S → [VERB] – subjunctive general auxiliary

→ I → [T1, T2, A] – imperative general auxiliary

→ J → J → [ADJ] – marked adjective

→ U – unmarked adjective

→ P → [ADJ] – marked predicate-only adjective

→ U – unmarked predicate-only adjective

→ D → | – indefinite determiner

→ N → U → | – cardinal number

→ O – oblique cardinal number

→ C – pre-multiplicative clitic cardinal number

→ [ADJ] – ordinal number

→ F → U – unmarked fraction

→ [ADJ] – marked fraction

→ Y → [ADJ] – near-demonstrative adjective

→ V → [ADJ] – far-demonstrative adjective

→ K → [ADJ] – interrogative adjective

→ J → [ADJ] – relative adjective

→ X → G → [ADJ] – multiplicative marker

→ S → [ADJ] – adjectival particle

→ V → [ADJ] – adjectival / occupational particle

→ P → P → [M1N, M1O, M2N, M2O, T1N, T1O, T2N, T2O]

– personal pronoun

→ A – honorific pronoun

→ G → M → 1 → [ADJ] – first person singular possessive adjective

→ 2 → [ADJ] – first person plural possessive adjective

→ T → 1 → [ADJ] – second person singular possessive adjective

→ 2 → [ADJ] – second person plural possessive adjective

→ R → [ADJ] – reflexive possessive adjective

→ R → F – reflexive pronoun

→ C – reciprocal pronoun

→ N → [N, O] – indefinite pronoun

→ Y → [1N, 1O, 2N, 2O, 2E] – proximal demonstrative pronoun

→ V → [1N, 1O, 2N, 2O, 2E] – distal demonstrative pronoun

→ K → [1N, 1O, 2N, 2O, 2E] – interrogative pronoun

→ J → [1N, 1O, 2N, 2O, 2E] – relative pronoun

→ R → R → | – general adverb

→ J – general adverb derived from adjective

		→	D				– degree adverb
		→	M	→	|		– modal adverb
				→	N		– negative modal adverb
		→	Y	→	|		– proximal demonstrative adverb
				→	J		– proximal dem. adverb derived from adj.
		→	V	→	|		– distal demonstrative adverb
				→	J		– distal demonstrative adverb derived from adj.
		→	K	→	|		– interrogative adverb
				→	J		– interrogative adverb derived from adjective
		→	J	→	|		– relative adverb
				→	J		– relative adverb derived from adjective
→	I	→	B				– preposition
		→	I	→	|		– unmarked postposition
				→	C		– clitic postposition
				→	[ADJ]		– marked postposition
→	C	→	C	→	|		– coordinating conjunction
				→	C		– correlative coordinating conjunction
		→	S				– subordinating conjunction
→	X	→	T				– contrastive emphatic particle
		→	H	→	|		– exclusive emphatic particle
				→	C		– clitic exclusive emphatic particle
		→	B				– inclusive emphatic particle
→	AU						– interjection
→	AL						– Arabic definite article
→	QQ						– question marker
→	ZZ						– izāfat
→	TT						– sentence tag-word
→	LL						– nongrammatical lexical element
→	F	→	F				– foreign word
		→	X				– non-Perso-Arabic string

→	O	– formula
→	Z	– letter of the alphabet
→	S	– other symbol
→	A	– acronym
→	B	– abbreviation
→	U	– other unclassifiable non-Urdu element

3.15 The extensibility of the EAGLES guidelines

At the outset of this chapter, I stated a claim that the EAGLES guidelines are extensible to Urdu. The very fact that it has proven possible to group the words of Urdu using the EAGLES major word classes suggests that they largely are so extensible. In particular, they are capable of dealing with Urdu's gender, case and number systems with very minor modifications only (such as the use of (3 | 5) as an intermediate value for "oblique case"). There were a few more problems with regard to the verbal system, particularly in matching up tense, mood and finiteness features between Urdu and EAGLES, and in dealing with the phenomenon of case marked on verbs. The greatest difficulty arose with regard to minor, idiosyncratic features of Urdu – such as the y-v-k-j word sets, or the pronoun *āp*, or the *zimmah dār* problem, or the clitic izāfat, which are quite simply not covered by EAGLES. However, these problems too were circumvented with the aid of a few (sometimes arbitrary) decisions and added attributes – such as *case* on verbal participles, or the added classificatory attribute for the *unique* class. The match between Urdu and the EAGLES categories remained generally very good. There was no major group of Urdu words for which there was no equivalent in EAGLES. Contrast Arabic, Chinese and Korean as discussed in the previous chapter (2.1.5.1, 2.1.5.2, 2.1.5.3). There is nothing in

EAGLES to correspond with “modifiers” (Chae and Choi 2000) or “particles” (Khoja et al. 2001) as major word classes which are higher hierarchically than some of the EAGLES word classes.

In light of these results with Urdu, it would seem likely that the EAGLES guidelines might easily be extensible to other Indo-Aryan languages, and possibly Iranian languages such as Persian. Furthermore, it would probably not take a great deal of work to create an extended version of the EAGLES guidelines to cover all Indo-European languages. To include intermediate tagging options for the idiosyncratic features of such languages would not be without precedent: the EAGLES guidelines as they stand include options to cover the idiosyncratic features of Western European languages (e.g. the fused preposition-article *au* in French, or strong and weak adjective inflections in German and Dutch). On the other hand, the extensibility of EAGLES to non-Indo-European languages may well prove very difficult or impossible. However, to confirm or contradict such hypotheses on extensibility lies beyond the scope of this thesis.

3.16 Concluding remarks

In this chapter of the thesis, I have achieved my aim of defining a POS tagset for use in the tagging of Urdu, and successfully validated my claim about the extensibility of the EAGLES guidelines. The tagset, one of the major prerequisites of an automated part-of-speech tagset, is now in place: however, it has not yet been tested, or validated out in the “real world” outside the model of Urdu given by Schmidt (1999). The essential next step is to see how well it stands up when exposed to actual language data. This is done by means of a phase of manual tagging, which is

the topic of the following chapter.