4 Manual tagging of Urdu texts

In this chapter, I discuss various aspects of the process of manual tagging which was undertaken using the tagset described in chapter 3.

My initial aim in this chapter is to justify this exercise in manual tagging within the context of this project (section 4.1). Then, in 4.2.1, I move on to explain and justify some changes that were made to the tagset outlined in the previous chapter. A subsidiary aim in this section is to evaluate the model of Urdu presented by Schmidt (1999), and its utility in practical applications, by examining how much a tagset based on that model must be modified to be useful. I will also describe the removal of certain category distinctions in the subtagsets[1] used for manual and automatic tagging, and substantiate my claim that removing these categories was a necessary step (section 4.2.2). In 4.3 I will outline categorisation difficulties that remain for the manual tagger using these slightly reduced versions of the tagset.

Also in this chapter, I aim to demonstrate the necessity of tagging guidelines for both manual tagging and accurate implementation of automated tagging. This is included in my discussion of how the said guidelines were created (see section 4.4). Finally, I will briefly discuss the nature of the data that was actually tagged, as a preliminary to the uses to which this data will be put in chapter 6 (section 4.5).

4.1 Why undertake manual tagging?

It is not obvious *a priori* why a study such as this one, whose goal is

[1] For the definition of this term, see 2.2.5.

automated part-of-speech tagging, should include a phase of manual tagging. Two reasons may be given to justify this manual tagging.

Firstly, trying out a tagset manually is an essential prerequisite to implementing an automatic tagger. In the attempt to apply the categories to tokens in natural language data, it can be established whether or not those categories actually reflect valid distinctions in the language. This is particularly the case for tagsets, such as the one discussed in the previous chapter, which have been formulated on the basis of a published grammar of the language rather than with direct reference to data. While one would not anticipate major word-class categories such as "noun" and "verb", clearly attested in every one of the Urdu grammars discussed Chapter 2 (see section 2.3), to prove invalid, one might well question categories such as "preposition" and "Arabic definite article", whose extent and productiveness in Urdu, beyond loanwords, has not been quantified. Exposing these categories to language data in the process of manual tagging can shed light on how clearly established they are in the structure of the language[2].

Utilising the tagset in manual tagging can also help to identify those phenomena which are difficult to categorise, not because of a flaw in the system of categories, but because the phenomenon is genuinely ambiguous in the Urdu language. For example, the boundary between the categories of JJU ("unmarked adjective") and JD ("indefinite determiner") is a fuzzy one. Words in these two categories have a similar syntactic distribution (i.e. prior to nouns) and morphological marking (none), so the division between the categories basically depends on semantic criteria. There is therefore genuine room for disagreement as to whether words such as *cand*, "few", and *har*, "every", belong in one category or the other. In this case, the

[2] See, for example, my discussion of the XB category in section 4.2.1.2 below.

process of manual tagging allowed such words to be encountered and discussed, and a decision taken. This decision, while certainly not definitive, and possibly even arbitrary, could then be applied consistently[3]. In the case of *cand* and *har*, they were judged to be examples of indefinite determiners. In this way, the process of manual tagging allows the boundary of a fuzzy category to be "mapped", as it were, and problematic examples pushed into one category or the other. In other cases, a firm decision – even an arbitrary one – was not possible. An example of this would be the category FF (foreign word): it proved impossible to draw a clear line as to what exactly constitutes a foreign word in Urdu[4]. However, even in such cases, it was useful to know – as a result of manual tagging – which areas exactly these were.

The information acquired by identifying and dealing with the phenomena within the language that are difficult to classify or intrinsically ambiguous is expressed in the tagging guidelines. These guidelines are an indispensable adjunct to the tagset if it is to be used for either manual or automatic tagging. However, without testing the classification system by means of manual tagging, it would be very difficult to create such a set of guidelines at all[5].

The second reason to undertake manual tagging is that tagged text is vitally necessary for many computational part-of-speech tagging methods, as will be outlined in the following chapter. It is needed in some cases as training data. But even if no training data is required, some tagged data at least is needed as a benchmark to

[3] For a discussion of the necessity of arbitrary but consistent decisions, see 2.2.7.

[4] See the discussion of *'alaikum* and *TikaT* in 4.3 below.

[5] In theory, the discovery of areas of problematic classification, and the creation of tagging guidelines, could be done in the process of developing an automated tagger. However, it does not seem conceivable that this could be an easier way to produce the guidelines than via the process of manual tagging.

evaluate the performance of an automated tagger.

The exercise in defining category boundaries and stating them as tagging guidelines is discussed in sections 4.3 and 4.4 below, and the uses the tagged text will be put to in 4.5. Having argued for the necessity of this process, the following section will describe modifications made to the tagset prior to or as a result of the manual tagging process.

4.2 Modifying the tagset

As was explained in chapter 2 (see 2.3), I do not speak Urdu. This meant that the tagset in its initial form was devised without any input from a native speaker (except at second hand in the form of descriptions written by native speakers). This being the case, it is unsurprising that some aspects of the tagset did not initially reflect the realities of Urdu. Therefore, the tagset was modified slightly in the light of native speaker input. This input took two forms. Firstly, through discussion with a native speaker informant, I was able to resolve some points of difficulty and eliminate some blemishes. Secondly, feedback from the process of manual tagging[6] allowed this process to be completed. The changes made are detailed in 4.2.1.

A separate but simultaneous process was the specification of appropriate subtagsets for use in manual and automatic tagging. This was also accomplished through the use of these two forms of native speaker input, as will be described in section 4.2.2.

[6] The tagging was also, for obvious reasons, undertaken by a native speaker informant. See also 4.5.

4.2.1 Changes to tagset on the basis of manual tagging[7]

As stated at the outset of this chapter, an aim pursued here is to evaluate the utility of Schmidt's (1999) model of the Urdu language for a practical application such as part-of-speech tagging. Therefore, I will now discuss the modifications which needed to be made, and their ramifications for this evaluation.

Although the initial tagset was reduced in the creation of subtagsets, as discussed below, the changes which I outline in this section do not apply to any of the categories which were reduced. Therefore, the discussion in this section can apply to any of the three variants of the tagset (U0, U1 or U2 – see below). The changes to the tagset include both additions and deletions; the additions are tabulated at the end of this section[8].

4.2.1.1 Deletion of the tags for marked predicate-only adjectives

The tags JPM1N, JPM2O, JPF1N, etc., for morphologically marked adjectives which only occur as predicates were from their inception a point of uncertainty (see

[7] Although presented here as a series of discreet phases, in reality there was a certain degree of overlap between the initial phase of tagset creation and the stage discussed here, of modification in the light of native speaker input. For that reason, some of the results of conferring with native speakers were actually implemented at that initial phase; where this is so, I have commented on it in my discussion in the previous chapter and do not discuss it again here (see for example section 3.7 for an example of this with regard to marked postpositions).

[8] I do not comment on the EAGLES intermediate tags assigned to these new tags, as these are in every case formed in line with analogous tags in the previous chapter (and involving, in most cases, the "Unique" word class).

3.3). Discussion with one of my informants lent support to what I had gathered from Schmidt (1999), that these categories never occurred, since all predicate-only adjectives are unmarked in Urdu (being originally loanwords). Therefore, these categories were deleted.

4.2.1.2 Deletion of the tag for the inclusive emphatic particle

In contrast to this, a category which had initially seemed firmly based was the tag XB for the inclusive emphatic particle *bhī*, "also, even, too" (Schmidt 1999: 215-217). However, in the categorisation system as originally described *bhī* could also be considered a modal adverb (RM) or a correlative conjunction (CCC). The informant who was performing manual tagging was unable to make a clear distinction between the uses of *bhī* in each of these three categories. After some discussion of Schmidt's descriptions and examples of the categories, it was decided that *bhī* should always be treated as being RM. This necessitated deleting the category of XB, since *bhī* was the only word that could be XB.

4.2.1.3 Addition of tags for forms of fused adverbs plus **hī**

The particle *hī* can occur as a clitic of varying form after certain pronouns and adverbs (Schmidt 1999: 212-214). The tag for this is XHC (see 3.12.1). However, there are four words, within the parallel Y-V-K-J set, which, when followed by the clitic form of *hī*, merge with it. An example is *yahā~*, "here", which merges with *hī* to form *yahī~*, "right here". During manual tagging it became clear that the strategy employed for other words with the *hī* clitic – to separate the word from the clitic,

tagging the former as normal and the latter as XHC – could not be employed. There is, in the case of *yahī~*, etc., simply no discernible boundary between the two. That no provision was originally made for this was due to an oversight during the design of the tagset.

This oversight was rectified by adding special tags for the fused forms. These are RYXHC, RVXHC, RKXHC, and RJXHC. This is a less than perfect solution. It violates the design principle that clitics should be tagged separately from the word that they are attached to (see 2.2.6.2.1). However, this solution does maintain the hierarchical nature of the tagset – RYXHC falls under RY in the hierarchy. There is also a precedent for such a step in the EAGLES guidelines, where a single tag is suggested for the phonologically fused preposition plus article combinations that occur in French (Leech and Wilson 1999: 67).

4.2.1.4 Addition of a tag for "verbal postposition"

A structure which Schmidt (1999: 108-109) refers to as the "conjunctive participle" (which was discussed briefly in section 3.2.2.5) has an alternate form where the root of the verb is followed by *kē* rather than *kar*; for example, *dē kar* or *dē kē*, "having given". I did not originally create a tag for *kē* used in this way, considering it to be a special use of the marked postposition *kē* (IIM1O), as Schmidt did not specify that this *kē* was a different word. However, the informant undertaking manual tagging did not accept this, saying that *kē* in this context was certainly not IIM1O. Since it seemed a little odd to use VX0 (the tag given to *kar* in this context) for a token that was on the face of it a postposition, I introduced a tag IV for "verbal postposition" to cover this use of *kē*. I had in mind the analogy of English *to*, which is

frequently given a different tag when used as part of a "to-infinitive" than when it occurs as a preposition.

Later, research in the grammar of Kellogg (1875: 341) revealed that etymologically speaking, *kē* (like *kar*) is here descended from Sanskrit *kṛitya*, the indeclinable past active participle of *kṛi*, "to do". Although the postposition *kā / kē / kī* is also derived from a participle of *kṛi* (Kellogg 1875: 129), the lines of descent appear to be separate, as Kellogg reports them. This implies that the classification of *kē* as any kind of postposition – verbal or otherwise – in this context is dubious (as, indeed, may be the classification of *kar* as a root-form verb). However, I have not altered the tagset as a result of this further information, as to do so would introduce etymological information into the tagset, in breach of a design principle (see 2.2.2).

4.2.1.5 Addition of a further punctuation tag

As a matter of practical necessity, an extra tag was defined for "other punctuation". This is because, while the punctuation tags I described in the last chapter cover the punctuation marks commonly used in Urdu text, it transpired during manual tagging that other symbols might occur – for example, the forward slash sign – in a manner that indicated they were intended as punctuation rather than as any kind of formula. The principle underlying the punctuation tags already defined, that punctuation is tagged as itself, could not be extended indefinitely, because some of the punctuation marks are used as control characters in the file layout used for manual tagging[9]. Therefore, another tag was introduced, using a symbol which is not a control

[9] This format was the Unitag file layout (see 6.2.1.3), and the forward slash is an example of a control character in this function. It is used to separate a tag form its probability. Thus, it could not appear as

character.

4.2.1.6 Tabulated definition of the new tags

Table 4.1

Description	Tag (Roman)	Tag (Perso-Arabic)	Intermediate Tag
Fused proximal demonstrative adverb and exclusive emphatic particle: *yahā~* + *hī* = *yahī~*	RYXHC	ﭼەغیل	AV0120
Fused distal demonstrative adverb and exclusive emphatic particle: *vahā~* + *hī* = *vahī~*	RVXHC	ﭼەغول	AV0120
Fused interrogative adverb and exclusive emphatic particle: *kahā~* + *hī* = *yahī~*	RKXHC	ﭼەغکل	AV011-2
Fused relative adverb and exclusive emphatic particle: *jahā~* + *hī* = *jahī~*[10]	RJXHC	ﭼەغجل	AV0112
Verbal postposition	IV	ف ج	U0D000
Other punctuation	~	~	PUE

the tag for a forward slash in the actual text, since it would cause a program designed to read this format to misread the entire line on which it appeared.

[10] *jahī~* has been described as obsolete, so this tag might only be used in old-fashioned texts.

4.2.1.7 Evaluating Schmidt's model in practical applications

The relatively minor nature of the adjustments above indicates that the model of Urdu presented by Schmidt (1999) is, in the larger part, practically applicable to such tasks as part-of-speech tagging. However, Schmidt's three-way classification of the uses of *bhī* (as modal adverb, inclusive emphatic particle, and correlative conjunction) is a clear instance of a grammatical distinction which was not applicable. It is only through the process of manual tagging that it was possible to bring such a point to light.

4.2.2 Collapsing the tagset

The Urdu tagset system permits categories to be collapsed together to create "subtagsets" for use in manual or automatic tagging (see section 2.2.5). It was clear while the tagset was being devised that it included categories which might prove difficult to implement in practice. These might therefore have to be removed in the subtagsets used for tagging. There are three major candidate categories to be removed in this way, discussed below.

4.2.2.1 Proper nouns versus common nouns

In English and many other European languages, proper nouns receive capital letters and are rarely preceded by articles. Neither of these distinctions applies in Urdu. The Perso-Arabic alphabet lacks the uppercase/lowercase distinction and there

are no articles, except for the marginal Arabic definite article, which occurs with some proper nouns anyway – for instance, the names of many cities in the Arabic-speaking world.

During manual tagging, two more distinctions between proper and common nouns were observed. Proper nouns are more likely to vary in gender – for instance, surnames, which can be masculine or feminine depending on the individual who holds them. By contrast only a few common nouns vary in gender; my informant suggests that this may be due to dialect differences. Secondly, proper nouns are overwhelmingly (but *not* always) unmarked, whereas many more common nouns are marked. However, it is difficult to see how either of these tendencies could be used to determine whether a given instance of a noun in a given context is common or proper, since instances that go against both these tendencies can be, and have been, observed.

Therefore, common and proper nouns are only distinct in terms of semantic considerations (based on what the sentence means) or lexical considerations (based on exhaustive lists of common and proper nouns). These considerations are readily accessible to a human analyst, in the form of their understanding of the text and their mental lexicon. But for a computer program they are clearly impracticable. Therefore it could well prove advantageous to remove this distinction.

4.2.2.2 Oblique case versus vocative case nouns

The oblique and vocative cases are identical in form for all word classes except plural nouns. The vocative case might be anticipated to be marginal in its occurrence (in the sense that a vocative case noun, by definition, does not play a major part in the grammar of the sentence in which it occurs). It might therefore prove

beneficial to merge the categories of oblique and vocative nouns. This would reduce the potential for confusion and ambiguity in those cases (the vast majority) where the correct analysis is clearly oblique.

4.2.2.3 Predicate-only adjectives versus general adjectives

The evidence required to add an adjective to the predicate-only category is negative evidence (i.e. evidence of the form, "Sentence X, where the adjective in question is used attributively, would be impossible"). This is a type of evidence that a computer can never have. Therefore this distinction could only be made by listing all predicate-only adjectives in a lexicon. This may not be an achievable goal. Even in manual tagging, the distinction will rely on the tagger's intuition to provide the negative evidence, rather than any evidence in the text being tagged. The status of an adjective as *predicate-only* or *general* is irrelevant to the tagging of surrounding text – it can have no effect on the syntax that surrounds it, since general adjectives can occur in predicate position as well. So collapsing this distinction would be a useful simplification of the tagset.

4.2.2.4 Removing distinctions in the subtagsets

Although there are arguments in favour of all these collapses, only the removal of the distinction between predicate-only and general adjectives is implemented in the subtagset used for manual tagging. The common/proper distinction in nouns has been maintained in a very large number of previous tagging projects. Furthermore, it can easily be made by human analysts. Therefore it was

deemed pre-emptive, and potentially unnecessary, to exclude it at this stage. The oblique-vocative distinction, likewise, is easy enough for human analysts to perceive[11]. Excluding this distinction would be inconsistent when there is some degree of morphological difference marking them out, given that all other distinctions involving morphological marking have been maintained in the subtagset.

By contrast, there is no particular motivation for retaining the predicate-only/general adjective distinction. It is a lexical consideration rather than a morphosyntactic one and thus marginal to the tagset in the first place. It is also likely to be equally difficult to capture this distinction in manual and automated tagging.

Therefore, in the subtagset used for manual tagging (which I refer to as the **U1** tagset in contrast the original **U0** tagset), the categories whose tags begin in JJ– and JP– are merged, and the JJ– tags used for the merged categories. Another minor collapse was in the categories for mirrored quotation marks. The texts being manually tagged were drawn from the EMILLE Corpus (see section 1.3), which contains only neutral quotation marks.

However, at the later stage of automatic tagging (see chapter 6), it became necessary to remove the common noun versus proper noun distinction as well. This was because, in the system which will be described in that chapter, there was simply no reliable way to make the distinction between the two types of noun, for reasons outlined above.

Also, in this subtagset (referred to as the **U2** tagset), the ZZ category for the

[11] Note that a minor alteration was made to the tagset in this respect: in the formal tagset description in Chapter 3, the oblique forms of adjectives and other words that inflect on the same pattern are described as "oblique / vocative", to indicate the possibility of their agreeing with an oblique or vocative noun. However in the tagset descriptions used for manual tagging, the term used is simply "oblique", to reduce confusion.

enclitic izāfat was removed. It was not used in any of the texts in the dataset used as a training corpus[12]. At least one of its forms in the written language (where it appears at all) is indistinguishable from some non-clitic inflectional endings – both being represented by the letter ـه. Therefore I could not devise any means to tokenise the izāfat without separating many inflectional endings from their bases. Since there was thus no easy way to manually tag it, and from the experience of the manual tagging the izāfat actually appearing in the writing seemed to be vanishingly rare at best, this tag has not been implemented in the automatic tagger, and therefore does not form part of the U2 tagset.

Text tagged using the U1 tagset was transferred automatically to the U2 tagset for use in the tagger design reported in chapter 6.

4.3 Categorisation difficulties in manual tagging

Aside from the problems (described above) that led to the abandonment or merging of some categories at different stages in the process, a number of categorisations proved somewhat problematic in the stage of manual tagging. This seems to have created tendencies for certain tags to be used where they were not necessarily appropriate. These, and other difficulties with the manual tagging, are detailed in this section.

It should be noted that since tagging was undertaken by one informant only

[12] It did occur a small number of times in the written text which was manually tagged, suggesting it may be more a feature of the written than spoken language. However, without more evidence it is impossible to be certain on this. The ZZ tag was removed from this text when it was transferred to the U2 tagset.

(see 4.5 below), the problems discussed here might well, in practice, be particular to that informant's tagging practice rather than the tagging scheme *per se*.

The informant was frequently inconsistent in the application of the FF tag (which was, as a general rule, also overused). Some words would be tagged as FF on their first appearance, but then as NNUM1N, JJU, etc., on their next appearance. This was standardised, as far as possible, when the texts were edited. However, there was inconsistency between words as well. For example the word *'alaikum* (from the Arabic phrase *alsalām 'alaikum*, "peace be on you") was tagged as FF, even though *alsalām* did not receive the FF tag, and the phrase is exceedingly common in Urdu. By contrast, the English word "ticket" (*TikaT* in Urdu) was given noun tags. Ultimately such decisions had to be left to the intuition of the informant, simply because of the lack of any practical informant. However, the inconsistency remains a difficulty.

Other tags that tended to be overused were FU and LL. The former was intended only to be used in the case of completely unanalysable elements, and thus to be quite rare; the latter was intended to occur only before derivational suffixes written as separate tokens. However, FU in particular seems to have been used as a response to bits of text that puzzled the informant, and LL in some phrases that were not of the type intended. Having no means to amend this, it had to be accepted. The FB tag developed a use which had not been envisioned, for tagging English initials spelt phonetically in Perso-Arabic, for example بی بی سی *bī bī sī*, "BBC". The CC tag was vastly overused for words which were really CS.

There was also inconsistency in the gender and markedness assigned to some nouns (see also the next section for a similar difficulty). This was standardised as far as possible, but not completely.

A persistent typographic error[13] in some of the texts resulted in many verb forms containing a superfluous *chōTī yē* character. In some cases, this removed the distinction between the subjunctive form of the verb end in *–ē* and the polite imperative form ending in *–iē* or *–iyē*.

4.4 The tagging guidelines

In this section, I will discuss the creation of a tagging manual, for use by informants undertaking manual tagging. It is my aim here to demonstrate that such a manual is a necessity not merely for manual tagging but also for automated tagging.

The tagging manual for the U1 Urdu tagset consists of two documents[14], the tagset definition and the tagging guidelines. The initial version of the former was based on the discussion of the tagset in the previous chapter, but with academic argument excluded and simplified terminology in some cases. The Perso-Arabic forms of the tags were removed (so as not to confuse matters). Examples of words falling into each category, written in Perso-Arabic script, were added. The initial version of the tagging guidelines was simply a selection of points of advice regarding aspects of tagging that I suspected on introspective grounds would prove problematic and require clarification.

The initial tagging manual was written prior to any manual tagging, and was used to train those undertaking it. As manual tagging got under way, many more

[13] This was due to a flaw in the set-up of the Urdu keyboards in the *Global Writer* word-processing software used to construct the EMILLE Corpus.

[14] A third document that was included for convenience was the tagset listed in chart form for easy reference.

problematic points were identified, as listed above[15].

In practice, no hard-and-fast distinction was maintained between these two parts of the tagging manual. Much information that was technically a "guideline" was inserted into the tagset definition document. It often makes more sense (from the reader's point of view) for a hint or a comment on difficulties concerning a particular category to be given at the time when the tags for that category are introduced. However, the "guidelines" section of the manual grew much more than the other, becoming the repository for much "miscellaneous information".

Additions made to the tagging manual as a result of feedback from the informant doing the tagging tended to fall into two groups (other than the simple addition of illustrative examples). The first were additions to clarify points already covered in the manual that were, in practice, understood. Thus, the process of manual tagging allowed the manual to be made "fool-proof", to a degree, which it clearly was not in the first place. An example of this is the distinction between marked and unmarked nouns.

Originally, the informant doing the tagging interpreted a marked noun as being one which had any sort of ending to indicate its gender. Thus, for instance, a large number of Sanskrit and Arabic-derived feminine nouns (especially proper nouns) ending in *–ā* were tagged as marked. However, they did not display the characteristic inflections of a feminine marked noun (inflecting singular *–ī* or *–iyā* to *–iyā~/–iyō~* in the nominative/oblique plural). Indeed, *–ā* is a masculine ending in marked nouns. The correct tagging was thus as an unmarked noun. It was necessary to write

[15] Not all the difficulties discussed in the previous section were dealt with in the tagging guidelines, since some only became evident at a later stage, when the tagged data was being utilised in the creation of the automatic tagger.

additional guidelines to clarify the notion of markedness as based on the type of plural inflection to prevent this error being maintained throughout the manually tagged text.

The other type of information that was added was to cover situations not originally anticipated. An example of this is the various greetings used in Urdu (and Hindi). Nothing was said about this in the initial tagging manual[16], but as a result of manual tagging the following guideline was added:

> Tagging greetings can be problematic. The Muslim greeting *alsalām 'alaikum* should be tagged as AL–NNUM1N–FF, with the *al* separated off from *salām*. The Hindu greetings *namaskār* and *namastē* should be tagged as AU when they are used as greetings, and as masculine unmarked nouns when they are used within the grammar of the sentence (e.g. as the subject or object of a verb). The English loan-word *hello* should be tagged as AU when used as a greeting, and as a masculine unmarked noun when used within the grammar of the sentence.

As part of this second type of information, arbitrary decisions that were made during tagging were catalogued so that they could be consistently adhered to subsequently. An example of this is the treatment of loanwords from Arabic or English which have retained the plural forms they had in the original language[17]. These do not follow the case/number inflection patterns of either marked or unmarked nouns. Indeed, they appear not to show case inflection at all, even in the plural. The decision was taken that such nouns should be tagged as if they were unmarked. This

[16] This blind spot had led to the very strange situation where the informant tagged *namastē* and *namaskār* as FF, even though they appear in Urdu dictionaries such as Haq (2001), and are very commonly heard in Hindi-Urdu discourse.

[17] For example, *śakayat* (sing), *śakayāt* (plural), "complaint(s)", from Arabic. There is also a regularly-formed unmarked plural *śakayatē~*.

can be justified on the basis that the singular of such words has the "no ending" expected of an unmarked noun. But since they do not show the ending for unmarked plural, this is essentially an arbitrary decision with regard to what may be classed as "irregular" nouns.

A brief list of the topics added to the tagging manual after its initial version follows: handling typing errors; handling ambiguous words and suffixes; spotting case in unmarked singular nouns; subject/object verb agreement; agreement rules for *cāhiē*; the distinction between auxiliary *rahā* and lexical verb *rahnā*; participles as adjectives; number versus politeness in personal pronouns; case of *mai~* and *tū* before *nē*; tagging the *in* and *un*, *is* and *us* ambiguities; the distinction between *kyā* as PK1N and QQ; compound postpositions, like the common *kē liē*; the use of LL; the use of OO; tagging reduplicated words; tagging clitics; tagging greetings; and finally, a list of individually problematic words and phrases, and the decisions made with regard to them.

The full tagging guidelines are given in Appendix 4.

Any of these annotations and clarifications could be used to demonstrate the claim I make in this section, that a good tagging manual is a necessity not just for manual tagging but for automated tagging too. I will discuss two.

Let us first consider the guideline on *rahā*. As explained in 3.2.2.2, this delexicalised form (VRM1) is homonymous with the perfect participle of *rahnā* (VVYM1N), from which it is derived. The tagging manual clearly defines the context in which *rahā* is to be considered VRM1, in terms of what category must precede it and follow it; otherwise it is to be considered VVYM1N. Without this guideline, mistakes would inevitably be made in manual tagging (indeed, they were, which is what prompted the guideline in the first place). However, the guideline must also

inform automatic tagging, as a means of ensuring that the computer and the informant apply the tagset in the same way. Without that decision, there could be no consistency between texts – and since the automatic tagger's performance is measured by how well it conforms with the manual analysis (see chapter 6), this would put limitations on the power of the tagger.

Similarly, it is vitally necessary that a reference list be kept of the sometimes arbitrary decisions made with regard to particular words and phrases. If there were not, there could be no guarantee that the same decision would be taken again the next time the phenomenon was encountered. This would have the same implications for consistent manual tagging, and for the power of an automatic tagger, that I have noted above.

I am therefore justified in my conclusion that the tagging manual, including the extended guidelines, are vitally necessary, and in the not inconsiderable effort and resources that were expended to devise them.

4.5 Creating the manually tagged dataset

The process of creating the manually tagged dataset (referred to throughout this chapter and used as a basis for automatic tagging in chapter 6) was subject to a number of serious practical limitations. In this section, I will explain these limitations, and how they impacted on the makeup of the manually tagged dataset.

As a non-native-speaker of Urdu, I could not undertake manual tagging myself. It was therefore necessary to find informants to perform this task on my behalf. Since this work had to be paid for from a limited pool of resources, this imposed a limit on how much could be done. However, finding capable people is not

easy. It is by no means as easy as might be anticipated to find native speakers of Urdu, with enough knowledge of linguistics to understand the tagset, who are also *willing* to undertake such an onerous task[18]. These two factors together meant that there were major practical limits on what could be done in terms of manual tagging.

For example, it would have been extremely desirable to perform an inter-annotator consistency test using the U1 tagset. This is an excellent way to check that the categories, their definitions, and the accompanying guidelines are valid, since if two annotators cannot reach at least broad agreement using them, something is clearly wrong. However, it simply did not prove possible to get the two informants necessary to work independently on the same text. Thus no such test was conducted, at some cost to the thoroughness with which the validity of the tagset was assessed.

My intention was to analyse 90 thousand words of data – 45,000 from a spoken corpus, and 45,000 from written data – as a training and test dataset. However, in the event it only proved possible for 49,000 words to be annotated – 42,000 spoken and 7,000 written. The imbalance was due to the late date[19] at which written data became available. All this was tagged by a single informant. The written data consisted of a single file, an excerpt from a history book published in 1972. The spoken data consisted of four transcripts from the BBC Asian Network's Urdu and Hindi-Urdu programming[20], transmitted between July and September 2000.

The written text and the shortest of the four spoken texts (4,000 words) were set aside as test datasets. The remaining three spoken texts composed my training

[18] This problem was compounded by a number of false starts: people who said they would work as informants and manual taggers who did not do so and dropped out after wasting precious project time.

[19] Less than four months before the end of the project, two years after spoken data had been available.

[20] See http://www.bbc.co.uk/asiannetwork .

dataset (referred to as such in chapter 6)[21]. These texts, once tagged, were combined into a single disk file and were thoroughly edited by myself to correct annotator errors[22]. I used a lexicon derived automatically from the text, to identify words which had received potentially problematic readings. These were then assessed and removed where possible. Unfortunately, as this was an extremely painstaking process, time did not permit a thorough examination of some of the problems mentioned in 4.3 above, in particular the excessive use of the LL tag.

Last of all, the training set and test texts' U1 tagging was mapped automatically to the U2 tagset.

It would be preferable to have two sets of training and test data, written and spoken, as per my original intentions, as we cannot assume *a priori* that a tagger modelled on spoken data will cope with written data, or vice versa. However, given that this was not possible, the next best option was to have a single test text from the other medium. This made it possible to assess whether or not the spoken data-based tagger was equally effective with written data. This is undertaken in Chapter 6.

4.6 Concluding remarks

In this chapter I aimed to describe and justify the manual tagging as a phase of the project. This has been done. I have also described the manually tagged dataset,

[21] Setting some texts aside as test data, to which the tagger and lexicon have not been exposed, allows the performance of the tagger when used on unknown texts to be simulated.

[22] These errors were extensive and in some cases systematic; some of the tagging instructions had simply been ignored. I continued to identify errors all the way through the process of developing the tagger.

and outlined a number of categorisation difficulties which emerged during this phase.

I also followed up on my aim of evaluating the applicability to tagset creation of Schmidt's (1999) grammar as a model of Urdu in section 4.2.1.7, where I concluded that it was indeed valid to use Schmidt's description in this way. However, the process of manual tagging did highlight one or two minor points (such as the use of *bhī*) where it was necessary, for the purpose of tagging, to deviate slightly from Schmidt's description. The discussion of manual tagging also allowed me to substantiate my claims regarding the necessity of using subtagsets with some categories removed (4.2.2) and the necessity of tagging guidelines (4.4).

I will now leave the topic of Urdu and the analysis of its morphosyntactic categories, and move in the next chapter to another topic, in a review of part-of-speech tagging methodologies. Chapter 6 will draw together these two threads – the analysis of Urdu presented in chapters 3 and 4, and the background to automated part-of-speech tagging in general in chapter 5 – in a discussion of the tagger experiment that I conducted.