Developing a tagset for automated part-of-speech tagging in Urdu

Andrew Hardie
Department of Linguistics and Modern English Language, University of Lancaster
a.hardie@lancaster.ac.uk

1. Abstract

While part-of-speech tagging is an established technology for Western European languages such as English or Spanish, extending the technique to Urdu presents a range of interesting issues. There are some problems associated with the writing system, e.g. the problems of locating token boundaries in the Urdu version of the Arabic script. However, there are also linguistic issues.

Little work has hitherto been done in the area of tagset creation for Urdu. The tagset discussed here was created in accordance with the EAGLES guidelines for morphosyntactic annotation of corpora. Although these guidelines were written to cover the languages of the European Union, they can be applied fairly easily to Urdu, which, coming as it does from another branch of the Indo-European family, is structurally quite similar. They can also be extended to deal with the idiosyncrasies presented by Urdu grammar.

This paper will look at the process of creating one of the necessary resources for the development of a POS tagging system for Urdu, that of a suitable tagset, considering some of the problems encountered along the way.

2. Introduction

As part of the EMILLE project[1], it was decided to develop a POS tagger for one of the languages of South Asia covered by the project. Urdu was chosen as the language in question for a number of reasons. Firstly, it is widely spoken in the UK, both as a first and second language, and native speakers were available to be consulted at Lancaster where this part of the EMILLE project is taking place. Secondly, as the *lingua franca* of a multilingual community (that of South Asian Muslims) and the official language of Pakistan, Urdu has considerable political and cultural importance. Thirdly, there are a number of factors that we anticipated would make tagging Urdu more complicated than tagging any other EMILLE language. For example, the right-to-left directionality of the Indo-Perso-Arabic script in which Urdu is written and the presence of grammatical forms borrowed from Arabic and Persian, which are structurally quite distinct from Indo-Aryan forms, mean that Urdu represents a unique challenge within the EMILLE corpora. It seemed the best course of action to confront these problems by choosing Urdu as the language for which to develop POS tagging.

The first stage of the work was to develop a tagset for use in Urdu texts and corpora, an area which has not been research extensively heretofore[2]. The next stage, now underway, is to test the tagset's usability in manual tagging, and build up a set of tagged texts to serve as training data for the final phase of this part of the project. This will be to automate the tagging and subsequently tag the whole of the EMILLE Urdu corpus. In this paper, the first, completed stage of this process is discussed: the devising of a tagset for Urdu based on the Urdu grammar of Schmidt (1999).

3. Some background on the Urdu language

Urdu is an Indo-European languages of the Indo-Aryan branch of the family. It is spoken in India and Pakistan (where it is the main official language) and also throughout the world

Urdu is more closely related to Hindi than either is to any other language. Indeed, their high level of similarity has led some to consider them dialects of the same language (as reported by Bhatia and Koul 2000: ix-x). Masica (1991: 27) goes to so far as to suggest that by one definition of a dialect, Urdu and Hindi "are different *literary styles* based on the *same* linguistically defined subdialect". Both originate from the dialect of the Delhi region and share their phonology, morphology and syntax in all but the smallest details. However, Urdu has borrowed a great deal of vocabulary (and its writing system) from Persian and Arabic, whilst Hindi has borrowed much vocabulary from Sanskrit.

The most notable features of Urdu grammar are as follows (see Schmidt 1999 for further detail). Its word order is principally SXOV, with some flexibility in the order of these elements; subject pronouns are frequently dropped. It possesses postpositions rather than prepositions. Inflection on

[1] A project to develop language engineering resources for the languages of South Asia undertaken at the Universities of Lancaster and Sheffield: see Baker et al. (2003).

[2] We are only currently aware of one other study into this area, undertaken by the Department of Electronics of the Indian government (personal communication, Dr. I. Hasnain). We did not become aware of this study until a late stage in the research and so it is not discussed further in this paper.

verbs, nouns and adjectives takes the form of fusional affixes, many of which are homophonous with one another. Nouns are inflected for case and number (singular/plural); the suffixes also indicate their gender (masculine/feminine). Gender agreement is marked by suffixes on verbs and adjectives; verbs show agreement either with the subject or with the direct object, although not both at once. Urdu verbs have one simple finite verb form (the subjunctive), two simple forms that may be finite or non-finite (the perfective and imperfective participles), and two further non-finite simple forms (the root and the infinitive). Tense and aspect, however, are mostly expressed through the use of irregular auxiliary elements within the verb phrase; there are also a number of frequently-used semi-auxiliary elements which confer semantic shading.

The history of linguistic investigation into Urdu (and Hindi[3]) is described by Bhatia (1987). The current standard grammar is that of Schmidt (1999), although a great many pedagogical books have also been published (e.g. Bailey et al. 1956). There has also been a certain amount written on the language within the field of theoretical linguistics (e.g. Butt 1995). However, there remains some contention about certain points of its grammar.

There is for instance some disagreement as to whether Urdu possesses three cases, or a much larger number. Urdu nouns generally display two clear cases, oblique (most commonly before postpositions) and nominative (elsewhere). A third case, the vocative, is identical with the oblique except in the plural. However, because of the structure of the Urdu noun phrase, postpositions always occur directly after the head noun of the noun phrase which they govern – in stark contrast to English, for instance, where a variety of elements may come between a preposition and the noun it governs. This phenomenon has led some (e.g. Kellogg 1875) to conclude that Urdu postpositions are actually "case suffixes" and that Urdu thus has a much larger variety of cases, possibly including accusative, dative, genitive, ergative, and so on. In this context it may be noted that the case system for other parts of speech (e.g. adjectives) displays the two-way nominative/oblique split that we would predict if the postpositions were *not* nominal affixes.

Another contentious issue relates to whether the language should be considered to display split-ergativity or not. In past and perfective clauses in Urdu, the subject is marked with the postposition *nē*, which has no other use than to indicate such a subject, whereas the object remains in the nominative case. This can be treated as evidence of split ergativity in the language. However, it has been argued (e.g. by Butt 1995) that *nē*, rather than marking an ergative case, is a semantic marker of agentivity or volitionality.

These differences of opinion on matters of Urdu grammar are not insignificant for the task of designing a set of morphosyntactic categories. Ideally one would wish to compose a markup scheme that does not commit the user to a particular theoretical analysis (as suggested, for example, by Leech 1997), thus making the categories equally acceptable and useful to researchers on either side of the two debates mentioned above.

4. A model for categorisation of the Urdu language

To create the linguistic categories of a tagset, it is necessary to have a model of the language to categorise. An ideal approach would be to derive this model from empirical data – however, this cannot be done prior to the creation of a tagset. A native speaker of a language could use their own intuitions about the language as a model, but as none of the researchers on the EMILLE project are native speakers of Urdu, this is not an option. The only remaining option is to make use of a published description of Urdu grammar as a model of the language.

The decision was taken to rely on the current standard grammar of Urdu by Schmidt (1999) to furnish a model of the language. It would probably have been preferable to rely on a synthesis of a range of published descriptions; however, in practice this was impossible. Most other recent works fall into two categories: pedagogical manuals, and works in theoretical linguistics that look at Urdu (or "Hindi-Urdu"). It might be assumed that the latter group of studies could be used to compile, in conjunction with Schmidt (1999), a synthesised model of the language on which to base the tagset categories. However, this is not so. Works in theoretical linguistics which concentrate on Hindi-Urdu tend to focus on one aspect of the language to the exclusion of the rest[4]. Thus they are of little use in developing a complete model of the language. Most of the studies listed in Masica's comprehensive bibliography (1991: 493-497, 510) are of this kind. Another type of linguistic study is the language survey, some of which have been published covering Urdu (e.g. Kachru 1990). These, while they cover the whole language, are not sufficiently detailed to constitute adequate models. Likewise, pedagogical

[3] Until the early 20th Century, the distinction between Hindi and Urdu was not as clearly defined as it later became.
[4] For example, Butt (1995) looks in detail at complex verb phrases, touching in cursory fashion or not at all on other aspects of Urdu grammar.

works[5] – aimed at language learners and covering mostly no more than what it is anticipated a learner would need to know first – are generally too partial to contribute a large amount to a composite model.

For these reasons, Schmidt (1999) was used as the model of Urdu grammar for the definition of the tagset. This necessitated the assumption that the model of Urdu presented by Schmidt is identical to the actual language, which may well be unwarranted. However, there was no alternative to this assumption.

5. Some background on the EAGLES guidelines

The Urdu tagset described in this paper was created in accordance with the EAGLES guidelines on morphosyntactic annotation (Leech and Wilson 1999). These guidelines were designed to help standardise tagsets for what were then the official languages of the European Union[6].

The EAGLES guidelines outline a set of features for tagsets, some recommended, some optional. Simultaneously, a scheme of encoding all these features into an "intermediate tagset" is given. This is an encoding using numerical values for the assorted EAGLES attributes. The choice of how the features are encoded within a given EAGLES-compliant tagset is left to the user, as long as the categories thus created can also be expressed using the intermediate tagset. The purpose of the intermediate encoding is to allow mapping between any two tagsets created in compliance with the EAGLES guidelines, thus ensuring their compatibility. EAGLES tags are defined as sets of morphosyntactic attribute-value pairs (e.g. *Gender* is an attribute that can have the values *Masculine*, *Feminine* or *Neuter*). In the intermediate tagset, these attribute-value pairs are arranged in a hierarchical structure.

The obligatory feature of an EAGLES-compatible tagset is a set of "major word categories"; the EAGLES guidelines suggest thirteen of these, such as *noun*, *verb*, *adjective* etc. The recommended and optional attributes are then organised by these major word categories, and do not necessarily correspond across word classes. For example, the first recommended attribute is *Type* (*Common*/*Proper*) for nouns but *Person* (*First*/*Second*/*Third*) for verbs and *Degree* (*Positive*/*Comparative*/*Superlative*) for adverbs. The recommended attributes also cover number, gender, case, finiteness, tense, voice, and other important features which one would anticipate being of relevance to a range of languages. The optional part of the recommendations consists of similar attributes of more narrow applicability, and some additional values – mainly specific to one language or a small group of languages[7] – for the recommended attributes.

The EAGLES guidelines provide a flexible framework that in theory encompasses all the things which one would wish to mark up, without restricting the freedom of the tagset designer. It promotes consistency and reusability of linguistic resources for different languages and discourages "reinvention of the wheel". However, as Leech and Wilson point out, "[i]t remains to be seen how far these guidelines can be extended, without substantial revision, to other languages" (1999: 58).

6. Extending the EAGLES scheme to Urdu

While Urdu did not fall under EAGLES' EU remit, it was decided to work with this international standard in order to ensure the maximum utility of the final tagged corpus. Furthermore, from a typological perspective it is not unreasonable to expect that the EAGLES guidelines would prove compatible with Urdu on the grounds that both Urdu and the original EAGLES languages are all from the Indo-European family. Genetically speaking, Urdu is no more distant from EAGLES languages such as German and Italian than they are from one another (although of course this does not take into account areal features which would make Urdu more divergent due to its geographical distance from the original EAGLES languages). Furthermore, impressionistically, the vast majority of the inflectional system of Urdu is very reminiscent of EAGLES languages such as French, German, and so on.

Indeed, it transpired that most of design features of the attribute-value system outlined in the EAGLES guidelines were suitable for application in the design of the Urdu tagset. The major categories in Urdu – nouns, pronouns, verbs, adjectives, adverbs, postpositions and conjunctions – are virtually identical to the equivalent categories as defined in EAGLES. Note that this is not the case for all languages. Others who have compiled tagsets for genetically unrelated languages such as Arabic (Khoja et al. 2001) and Korean (e.g. Chae and Choi 2000) have of necessity employed categorisation strategies alien to the system laid out in the EAGLES guidelines.

It was therefore possible to link the Urdu tagset (see the Appendix) directly to the EAGLES guidelines, using the EAGLES intermediate tagset as described by Leech and Wilson (1999), with only

[5] For instance, Bhatia and Koul (2000), Barz (1977), Bailey et al. (1956)

[6] English, Dutch, German, Danish, French, Spanish, Portuguese, Italian and Greek

[7] For example, the fairly language-specific attribute *aux-function*, which relates to the difference in English between the primary auxiliaries (*do/be/have*) and the modals (*can/may/will* etc.)

some minor modifications. For instance, although the EAGLES guidelines deal very well with the gender, case and number system of Urdu (as described above), since there was no value for "oblique case" in the EAGLES system, the value for "dative case" was used instead, on the grounds that the usage of the Urdu oblique corresponds quite closely to that of the dative in some EU languages, such as German. Most other general attributes in the EAGLES system were similarly amenable to being extended to Urdu. The verbal system proved a little more problematic, in the sense that mapping the mood, tense and finiteness features outlined in the EAGLES attribute-value system onto those found in the Urdu language was less straightforward.

However, the greatest difficulty arose in dealing with the minor, idiosyncratic features of Urdu – whilst the idiosyncratic features of the EU languages are covered by the EAGLES guidelines this is obviously not the case for Urdu. Some of these difficulties are discussed in the following section. However, on the whole, the EAGLES guidelines have proved a robust and useful framework within which to approach Urdu POS tagging.

7. Points of difficulty in devising an EAGLES-compliant tagset for Urdu

7.1. Minor idiosyncratic features of the language

Some idiosyncratic features of Urdu corresponded to nothing built into the EAGLES guidelines. These features include: the appearance of case on some verbal elements[8]; the distinction between 'marked' and 'unmarked' nouns; the Urdu honorific pronoun *āp*, which does not fit easily into any of the EAGLES categories for pronouns; the borrowed Persian enclitic called *izāfat*; However, none of these problems were insurmountable. For instance, the EAGLES guidelines include a "Unique" category for words that effectively form a class on their own; this category was used for some idiosyncratic features of Urdu which resisted classification elsewhere (e.g. *izāfat*, or the use of a marker-word for *yes/no* questions). Others could be handled by means of minor extensions to the EAGLES system, e.g. to handle case on verbs a *case* attribute was added to the end of the EAGLES intermediate tags for verbs[9].

However, some idiosyncrasies could not be resolved this easily and uncontroversially. An example is the definite article. Urdu does not really have a definite article, but the Arabic definite article *al–* does occur in loans from that language. The question arises as to whether it is really appropriate to equate this via the intermediate tagset with the definite articles of the EAGLES languages, given that the distribution of *al–* in Urdu is much more restricted than, say, that of *the* in English or *le/la/les* in French.

Another example of this type relates to the distinction between relative and interrogative pronouns. In the EAGLES guidelines this distinction is made only at the optional level (attribute *Wh-type*), whereas the distinction between demonstrative and interrogative/relative is made by the recommended-level attribute *Pronoun-type*. This maps easily onto sets of words such as *this~that~what* in English and similar EAGLES languages. In Urdu, however, there is a four-way distinction between proximal demonstratives, distal demonstratives, interrogatives and relatives (for example, the pronouns *yah*, *vah*, *kyā*, *jō* respectively). This distinction is maintained throughout a system of pronouns, determiners and adverbs. Thus, in Urdu the distinction between interrogatives and relatives, which is only made by the EAGLES guidelines at the secondary optional level, appears a priori to be as significant as that between demonstrative and relative. To categorise this in the tagset, it was necessary for a high-level distinction in the Urdu tags (PY… PV… PK… PJ…) to map to distinctions made at varying levels in the intermediate tagset – an unfortunately inelegant solution.

7.2. Token division in Urdu

Urdu is written in Indo-Perso-Arabic, a form of Perso-Arabic which adds letters for some characteristically Indo-Aryan sounds (e.g. the retroflex consonants) and uses some original Arabic

[8] The participles and the infinitive can all display case.

[9] A similar approach was taken to solve problem of encoding the marked/unmarked distinction found in Urdu nouns, and the marking of case, gender and number on one Urdu preposition. In all these cases, entire attributes were added to the end of the EAGLES intermediate tags as described by Leech and Wilson (1999), rather than adding more values to the attributes that already existed. This was to make it less problematic for an EAGLES-compliant computer application to ignore the extra elements added to handle Urdu (simply by passing over the extra attributes), while retaining in the intermediate tagset all the information in the full Urdu tagset. A concrete example is the intermediate tag V10212101000000 (first person plural subjunctive lexical verb) which has an added "0" on the end compared to what would be the equivalent tag in Leech and Wilson's description of the intermediate tagset: the extra attribute allows the tagset to encode case on those verb forms that display it.

letters with different phonetic values[10]. One aspect of this writing system and its application to Urdu is that many things described in the literature on Urdu grammar as suffixes are actually written as independent words (for example, the verbal auxiliary element indicating future tense: see Schmidt 1999: 106, Bhatia and Koul 2000: 331-332).

For consistency, the (essentially arbitrary) decision was taken to treat every orthographic space as a word break even if it occurs within a lexical word[11]. However, this meant that the tagset had to contain some means of tagging elements which did not constitute entire words.

For example, the word *zimmah dār* ("responsible") consists of a root plus a derivational affix, with an orthographic space between them. Although it would be attractive to describe this as a phrase for the purposes of tagging, this cannot be supported linguistically: the same suffix appears without a word break in other contexts (e.g. *samajhdār*, "sensible"), and moreover other derivational suffixes can be added (e.g. *zimmah dārī*, "responsibility"). A number of other derivational suffixes behave in the same way, as do some simple lexemes, for example *Tēlī fōn*, "telephone". The phenomenon appears to be common in borrowed vocabulary (*dār* derives from Persian, *Tēlī fōn* from English).

This created a problem[12]: how to tag two tokens which make up a single morphological word? After considering various solutions to this problem, it was decided to incorporate a tag into the categorisation system for a *non-grammatical lexical element*, i.e. a token which has no effect on the syntax of the clause and is dependent for its grammar on the subsequent token. This tag is LL, and it would be used thus[13]:

*samajhdār*_JJU
*zimmah*_LL *dār*_JJU

7.3. Loan words and inflections

As mentioned above, Urdu has borrowed a great deal of vocabulary form Arabic and Persian (and also, more recently, from English). Some of the words thus borrowed are inflected forms (Schmidt 1999: 253, 259-264). However, it is not currently clear a) how extensive these are in Urdu, and b) whether they are used as lexical roots just like any other loan word or whether they are used as actual inflected forms. The latter would pose a problem, as some special tagging might be necessitated for grammatical forms with no equivalent in the "native" Urdu vocabulary.

While this is evidently an important issue, there is not currently enough knowledge of all aspects of the "loan-word" phenomenon in Urdu to incorporate a comprehensive means of handling it into the tagset. Therefore it is our intention at a later date to study this phenomenon in a corpus of Urdu text tagged according to the current tagset, on the basis of which changes to subsequent tagsets may be necessary.

8. Conclusion

Experience in developing an annotation scheme for Urdu has demonstrated that the scope of the EAGLES guidelines may productively be extended beyond the languages for which they were originally devised, to genetically and typologically similar languages such as those of the Indo-Aryan group. Furthermore, although certain modifications to the scheme were required, these were only minor and did not involve any disruption to the large-scale organisation of the EAGLES categories. Although some features of the Indo-Perso-Arabic alphabet create problems for tokenisation and the assignation of tags to some tokens, these difficulties are not insuperable. As a result of this process, a tagset for use with Urdu texts and corpora has been developed (see Appendix).

It should also be noted that, while a tagset is the most obvious prerequisite resource for tagging, there are other significant resources which must be in place *prior* to the development of an

[10] Because this form of the Arabic writing system is used by other Indo-Aryan languages (e.g. some forms of Western Punjabi), I refer to it as Indo-Perso-Arabic; it is popularly referred to simply as the "Urdu alphabet" or "Urdu script" (e.g. by Nakanishi 1980: 36).

[11] Word breaks are also introduced in some places where there is no orthographic space, e.g. where clitics precede/follow another word without a break.

[12] This is only partially analogous to the problem of multi-word idioms in English and similar languages that leads, for example, to phrases such as "given that" being tagged as the two parts of a single subordinating conjunction. In these cases, there is also an analysable internal syntactic structure (in this case, verbal past participle followed by conjunction). In the Urdu case, it would be very difficult to assign any internal structure to *Tēlī fōn*, and the internal structure of *zimmah dār* would of necessity be morphological rather than morphosyntactic – both undesirable analyses.

[13] The underscore character is used here for visual clarity: the system under development actually uses a columnar format for communication between different modules and outputs SGML compliant word tags.

automated tagger. These include a set of guidelines for application of the tags, and a lexicon (or else a reliable means of acquiring such automatically from tagged text). The guidelines are of course needed for the generation of a manually tagged training/test corpus, a *sine qua non* of automated tagger design[14]. Development of these resources is currently underway.

9. References

Bailey, TG, ed. by Firth, JR and Harley, AH 1956 *Teach Yourself Urdu*. New York: David McKay Company.

Baker, JP, Hardie, A, McEnery, A and Jayaram, BD 2003 Corpus data for South Asian language processing. Paper given at the Corpus Linguistics 2003 conference, Lancaster.

Barz, RK 1977 *An Introduction to Hindi and Urdu*. Canberra: Australian National University Press.

Bhatia, TK and Koul, A 2000 *Colloquial Urdu*. London: Routledge.

Brill, E and Pop, M 1999 Unsupervised learning of disambiguation rules for part of speech tagging. In: Armstrong, S, Church, K, Isabelle, P, Manzi, S, Tzoukermann, E and Yarowsky, D (eds.) *Natural language processing using very large corpora*. Dordrecht: Kluwer Academic Publishers.

Butt, M 1995 *The Structure of Complex Predicates in Urdu*. Stanford, California: CSLI Publications.

Chae, Y-S and Choi, K-S 2000 Introduction of KIBS (Korean Information Base System) Project. In: Gavrilidou, M, et al. (eds.) *Second International Conference on Language Resources and Evaluation: Proceedings*. 3 volumes. Athens: European Language Resources Association. Vol. 3, 1731-1735.

Kachru, Y (1990) Hindi-Urdu. In: Comrie, B (ed.) 1990 *The Major Languages of South Asia, the Middle East and Africa*. London: Routledge.

Kellogg, SH 1875 *A grammar of the Hindí language, in which are treated the High Hindí, Braj, and the eastern Hindí of the Rámáyan of Tulsí Dás*. (Reprinted 1965.) London: Routledge and Kegan Paul.

Khoja, S, Garside, R, and Knowles, G 2001 A tagset for the morphosyntactic tagging of Arabic. Paper given at the Corpus Linguistics 2001 conference, Lancaster.

Leech, G 1997 Grammatical tagging. In: Garside, R, Leech, G and McEnery A (eds.) (1997) *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

Leech, G and Wilson, A. 1999 Standards for tagsets. (Edited version of *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora* 1996: available on the internet at http://www.ilc.pi.cnr.it/EAGLES96/annotate/annotate.html .) In: van Halteren, H (ed.) (1999) *Syntactic wordclass tagging*. Dordrecht: Kluwer Academic Publishers.

Masica, CP 1991 *The Indo-Aryan languages*. Cambridge: Cambridge University Press.

Merialdo, B 1994 Tagging English text with a probabilistic model. In: *Computational Linguistics*, 20 (2): 155-171.

Nakanishi, A. 1980 *Writing systems of the world*. Tokyo: Charles E. Tuttle Company.

Platts, JT. 1884 *A Dictionary of Urdu, Classical Hindi, and English*. Oxford: Oxford University Press (reprinted 1965).

Schmidt, RL 1999 *Urdu: an essential grammar*. London: Routledge.

Appendix: the U1 tagset

Tag	Description
AL	Arabic definite article
AU	Interjection
CC	Coordinating conjunction
CCC	Correlative coordinating conjunction
CS	Subordinating conjunction
FF	Foreign word
FX	Non-Perso-Arabic string
FO	Formula (e.g. mathematical)
FZ	Letter of the alphabet
FS	Other symbol

[14] Although techniques have been developed for training a tagger based on untagged data (see for instance Merialdo 1994, Brill and Pop 1999), all such techniques appear to depend on the availability of a reliably accurate lexicon in which words are linked to all their possible tags. In the case of a language that has not previously been tagged automatically, such a lexicon must also be derived from manually tagged data – or else written item by item using a linguist's intuition, a task which may well be as onerous as manual tagging in any case. Thus there is no way to avoid the effort associated with manual tagging of tens of thousands of words of text.

FA	Acronym
FB	Abbreviation
FU	Other unclassifiable non-Urdu element
IB	Preposition
II	Unmarked postposition
IIC	Clitic postposition *ē*, *ē~*, *hē~*
IIM1N	Marked masculine singular nominative postposition *kā*
IIM1O	Marked masculine singular oblique postposition *kē*
IIM2N	Marked masculine plural nominative postposition *kē*
IIM2O	Marked masculine plural oblique postposition *kē*
IIF1N	Marked feminine singular nominative postposition *kī*
IIF1O	Marked feminine singular oblique postposition *kī*
IIF2N	Marked feminine plural nominative postposition *kī*
IIF2O	Marked feminine plural oblique postposition *kī*
JJM1N	Marked masculine singular nominative adjective
JJM1O	Marked masculine singular oblique adjective
JJM2N	Marked masculine plural nominative adjective
JJM2O	Marked masculine plural oblique adjective
JJF1N	Marked feminine singular nominative adjective
JJF1O	Marked feminine singular oblique adjective
JJF2N	Marked feminine plural nominative adjective
JJF2O	Marked feminine plural oblique adjective
JJU	Unmarked adjective
JD	Indefinite determiner
JDNU	Cardinal number
JDNUO	Oblique cardinal number
JDNUC	Pre-multiplicative clitic cardinal number *du–*, *ti-*, *cau–*
JDNM1N	Masculine singular nominative ordinal number
JDNM1O	Masculine singular oblique ordinal number
JDNM2N	Masculine plural nominative ordinal number
JDNM2O	Masculine plural oblique ordinal number
JDNF1N	Feminine singular nominative ordinal number
JDNF1O	Feminine singular oblique ordinal number
JDNF2N	Feminine plural nominative ordinal number
JDNF2O	Feminine plural oblique ordinal number
JDFU	Unmarked fraction
JDFM1N	Masculine singular nominative fraction
JDFM1O	Masculine singular oblique fraction
JDFM2N	Masculine plural nominative fraction
JDFM2O	Masculine plural oblique fraction
JDFF1N	Feminine singular nominative fraction
JDFF1O	Feminine singular oblique fraction
JDFF2N	Feminine plural nominative fraction
JDFF2O	Feminine plural oblique fraction
JDYM1N	Masculine singular nominative proximal demonstrative adjective (*itnā, aisā*)
JDYM1O	Masculine singular oblique proximal demonstrative adjective (*itnē, aisē*)
JDYM2N	Masculine plural nominative proximal demonstrative adjective (*itnē, aisē*)
JDYM2O	Masculine plural oblique proximal demonstrative adjective (*itnē, aisē*)
JDYF1N	Feminine singular nominative proximal demonstrative adjective (*itnī, aisī*)
JDYF1O	Feminine singular oblique proximal demonstrative adjective (*itnī, aisī*)
JDYF2N	Feminine plural nominative proximal demonstrative adjective (*itnī, aisī*)
JDYF2O	Feminine plural oblique proximal demonstrative adjective (*itnī, aisī*)
JDVM1N	Masculine singular nominative distal demonstrative adjective (*utnā, vaisā*)
JDVM1O	Masculine singular oblique distal demonstrative adjective (*utnē, vaisē*)
JDVM2N	Masculine plural nominative distal demonstrative adjective (*utnē, vaisē*)
JDVM2O	Masculine plural oblique distal demonstrative adjective (*utnē, vaisē*)
JDVF1N	Feminine singular nominative distal demonstrative adjective (*utnī, vaisī*)
JDVF1O	Feminine singular oblique distal demonstrative adjective (*utnī, vaisī*)
JDVF2N	Feminine plural nominative distal demonstrative adjective (*utnī, vaisī*)
JDVF2O	Feminine plural oblique distal demonstrative adjective (*utnī, vaisī*)
JDKM1N	Masculine singular nominative interrogative adjective (*kitnā, kaisā*)
JDKM1O	Masculine singular oblique interrogative adjective (*kitnē, kaisē*)
JDKM2N	Masculine plural nominative interrogative adjective (*kitnē, kaisē*)
JDKM2O	Masculine plural oblique interrogative adjective (*kitnē, kaisē*)
JDKF1N	Feminine singular nominative interrogative adjective (*kitnī, kaisī*)
JDKF1O	Feminine singular oblique interrogative adjective (*kitnī, kaisī*)
JDKF2N	Feminine plural nominative interrogative adjective (*kitnī, kaisī*)
JDKF2O	Feminine plural oblique interrogative adjective (*kitnī, kaisī*)
JDJM1N	Masculine singular nominative relative adjective (*jitnā, jaisā*)
JDJM1O	Masculine singular oblique relative adjective (*jitnē, jaisē*)
JDJM2N	Masculine plural nominative relative adjective (*jitnē, jaisē*)

JDJM2O	Masculine plural oblique relative adjective (*jitnē, jaisē*)
JDJF1N	Feminine singular nominative relative adjective (*jitnī, jaisī*)
JDJF1O	Feminine singular oblique relative adjective (*jitnī, jaisī*)
JDJF2N	Feminine plural nominative relative adjective (*jitnī, jaisī*)
JDJF2O	Feminine plural oblique relative adjective (*jitnī, jaisī*)
JXGM1N	Masculine singular nominative multiplicative marker *gunā*
JXGM1O	Masculine singular oblique multiplicative marker *gunē*
JXGM2N	Masculine plural nominative multiplicative marker *gunē*
JXGM2O	Masculine plural oblique multiplicative marker *gunē*
JXGF1N	Feminine singular nominative multiplicative marker *gunī*
JXGF1O	Feminine singular oblique multiplicative marker *gunī*
JXGF2N	Feminine plural nominative multiplicative marker *gunī*
JXGF2O	Feminine plural oblique multiplicative marker *gunī*
JXSM1N	Masculine singular nominative adjectival particle *sā*
JXSM1O	Masculine singular oblique adjectival particle *sē*
JXSM2N	Masculine plural nominative adjectival particle *sē*
JXSM2O	Masculine plural oblique adjectival particle *sē*
JXSF1N	Feminine singular nominative adjectival particle *sī*
JXSF1O	Feminine singular oblique adjectival particle *sī*
JXSF2N	Feminine plural nominative adjectival particle *sī*
JXSF2O	Feminine plural oblique adjectival particle *sī*
JXVM1N	Masculine singular nominative adjectival / occupational particle *vālā*
JXVM1O	Masculine singular oblique adjectival / occupational particle *vālē*
JXVM2N	Masculine plural nominative adjectival / occupational particle *vālē*
JXVM2O	Masculine plural oblique adjectival / occupational particle *vālē*
JXVF1N	Feminine singular nominative adjectival / occupational particle *vālī*
JXVF1O	Feminine singular oblique adjectival / occupational particle *vālī*
JXVF2N	Feminine plural nominative adjectival / occupational particle *vālī*
JXVF2O	Feminine plural oblique adjectival / occupational particle *vālī*
LL	Nongrammatical lexical element
NNMM1N	Common marked masculine singular nominative noun
NNMM1O	Common marked masculine singular oblique noun
NNMM1V	Common marked masculine singular vocative noun
NNMM2N	Common marked masculine plural nominative noun
NNMM2O	Common marked masculine plural oblique noun
NNMM2V	Common marked masculine plural vocative noun
NNMF1N	Common marked feminine singular nominative noun
NNMF1O	Common marked feminine singular oblique noun
NNMF1V	Common marked feminine singular vocative noun
NNMF2N	Common marked feminine plural nominative noun
NNMF2O	Common marked feminine plural oblique noun
NNMF2V	Common marked feminine plural vocative noun
NNUM1N	Common unmarked masculine singular nominative noun
NNUM1O	Common unmarked masculine singular oblique noun
NNUM1V	Common unmarked masculine singular vocative noun
NNUM2N	Common unmarked masculine plural nominative noun
NNUM2O	Common unmarked masculine plural oblique noun
NNUM2V	Common unmarked masculine plural vocative noun
NNUF1N	Common unmarked feminine singular nominative noun
NNUF1O	Common unmarked feminine singular oblique noun
NNUF1V	Common unmarked feminine singular vocative noun
NNUF2N	Common unmarked feminine plural nominative noun
NNUF2O	Common unmarked feminine plural oblique noun
NNUF2V	Common unmarked feminine plural vocative noun
(Tags for proper nouns are as the tags for common nouns, except that they begin NP instead of NN)	
OO	Persian compound-forming conjunction *ō*
PPM1N	First person singular nominative personal pronoun (*mai~*)
PPM1O	First person singular oblique personal pronoun (*mujh*)
PPM2N	First person plural nominative personal pronoun (*ham*)
PPM2O	First person plural oblique personal pronoun (*ham*)
PPT1N	Second person singular nominative personal pronoun (*tū*)
PPT1O	Second person singular oblique personal pronoun (*tujh*)
PPT2N	Second person plural nominative personal pronoun (*tum*)
PPT2O	Second person plural oblique personal pronoun (*tum*)
PGM1M1N	First person singular masculine singular nominative possessive adjective (*mērā*)
PGM1M1O	First person singular masculine singular oblique possessive adjective (*mērē*)
PGM1M2N	First person singular masculine plural nominative possessive adjective (*mērē*)
PGM1M2O	First person singular masculine plural oblique possessive adjective (*mērē*)
PGM1F1N	First person singular feminine singular nominative possessive adjective (*mērī*)
PGM1F1O	First person singular feminine singular oblique possessive adjective (*mērī*)
PGM1F2N	First person singular feminine plural nominative possessive adjective (*mērī*)

PGM1F2O	First person singular feminine plural oblique possessive adjective (*mērī*)
PGM2M1N	First person plural masculine singular nominative possessive adjective (*hamārā*)
PGM2M1O	First person singular masculine singular oblique possessive adjective (*hamārē*)
PGM2M2N	First person singular masculine plural nominative possessive adjective (*hamārē*)
PGM2M2O	First person singular masculine plural oblique possessive adjective (*hamārē*)
PGM2F1N	First person singular feminine singular nominative possessive adjective (*hamārī*)
PGM2F1O	First person singular feminine singular oblique possessive adjective (*hamārī*)
PGM2F2N	First person singular feminine plural nominative possessive adjective (*hamārī*)
PGM2F2O	First person singular feminine plural oblique possessive adjective (*hamārī*)
PGT1M1N	Second person singular masculine singular nominative possessive adjective (*tērā*)
PGT1M1O	Second person singular masculine singular oblique possessive adjective (*tērē*)
PGT1M2N	Second person singular masculine plural nominative possessive adjective (*tērē*)
PGT1M2O	Second person singular masculine plural oblique possessive adjective (*tērē*)
PGT1F1N	Second person singular feminine singular nominative possessive adjective (*tērī*)
PGT1F1O	Second person singular feminine singular oblique possessive adjective (*tērī*)
PGT1F2N	Second person singular feminine plural nominative possessive adjective (*tērī*)
PGT1F2O	Second person singular feminine plural oblique possessive adjective (*tērī*)
PGT2M1N	Second person plural masculine singular nominative possessive adjective (*tumhārā*)
PGT2M1O	Second person singular masculine singular oblique possessive adjective (*tumhārē*)
PGT2M2N	Second person singular masculine plural nominative possessive adjective (*tumhārē*)
PGT2M2O	Second person singular masculine plural oblique possessive adjective (*tumhārē*)
PGT2F1N	Second person singular feminine singular nominative possessive adjective (*tumhārī*)
PGT2F1O	Second person singular feminine singular oblique possessive adjective (*tumhārī*)
PGT2F2N	Second person singular feminine plural nominative possessive adjective (*tumhārī*)
PGT2F2O	Second person singular feminine plural oblique possessive adjective (*tumhārī*)
PY1N	Singular nominative proximal demonstrative pronoun (*yah*)
PY1O	Singular oblique proximal demonstrative pronoun (*is*)
PY2N	Plural nominative proximal demonstrative pronoun (*yah*)
PY2O	Plural oblique proximal demonstrative pronoun (*in*)
PY2E	Plural oblique proximal demonstrative pronoun before *nē* (*inhõ~*)
PV1N	Singular nominative distal demonstrative pronoun (*vah*)
PV1O	Singular oblique distal demonstrative pronoun (*us*)
PV2N	Plural nominative distal demonstrative pronoun (*vah*)
PV2O	Plural oblique distal demonstrative pronoun (*un*)
PV2E	Plural oblique distal demonstrative pronoun before *nē* (*unhõ~*)
PK1N	Singular nominative interrogative pronoun (*kyā, kaun*)
PK1O	Singular oblique interrogative pronoun (*kis*)
PK2N	Plural nominative interrogative pronoun (*kyā, kaun*)
PK2O	Plural oblique interrogative pronoun (*kin*)
PK2E	Plural oblique interrogative pronoun before *nē* (*kinhõ~*)
PJ1N	Singular nominative relative pronoun (*jō*)
PJ1O	Singular oblique relative pronoun (*jis*)
PJ2N	Plural nominative relative pronoun (*jō*)
PJ2O	Plural oblique relative pronoun (*jin*)
PJ2E	Plural oblique relative pronoun before *nē* (*jinhõ~*)
PRF	Reflexive pronoun (*āp, xud*)
PRC	Reciprocal pronoun (*āpas*)
PGRM1N	Masculine singular nominative reflexive possessive adjective (*apnā*)
PGRM1O	Masculine singular oblique reflexive possessive adjective (*apnē*)
PGRM2N	Masculine plural nominative reflexive possessive adjective (*apnē*)
PGRM2O	Masculine plural oblique reflexive possessive adjective (*apnē*)
PGRF1N	Feminine singular nominative reflexive possessive adjective (*apnī*)
PGRF1O	Feminine singular oblique reflexive possessive adjective (*apnī*)
PGRF2N	Feminine plural nominative reflexive possessive adjective (*apnī*)
PGRF2O	Feminine plural oblique reflexive possessive adjective (*apnī*)
PNN	Nominative indefinite pronoun (*kōī, kuch, sab*)
PNO	Oblique indefinite pronoun (*kīsī, kuch, sabhõ~*)
PA	Honorific pronoun (*āp*)
QQ	Question marker *kyā*
RR	General adverb
RRJ	General adverb derived from adjective
RD	Degree adverb
RM	Modal adverb
RMN	Negative modal adverb (*nahī~, nah, mat*)
RY	Proximal demonstrative adverb (*ab, yahā~, idhar, yū~*)
RYJ	Proximal demonstrative adverb derived from adjective (*aisē)*
RV	Distal demonstrative adverb (*tab, vahā~, udhar, tyū~*)
RVJ	Distal demonstrative adverb derived from adjective (*vaisē)*
RK	Interrogative adverb (*kab, kahā~, kidhar, kyõ~*)
RKJ	Interrogative adverb derived from adjective (*kaisē)*
RJ	Relative adverb (*jab, jahā~, jidhar, jū~*)

RJJ	Relative adverb derived from adjective (*jaisē)*
TT	Sentence tag-word
VV0	Root form lexical verb
VVNM1N	Infinitive lexical verb, masculine singular nominative
VVNM1O	Infinitive lexical verb, masculine singular oblique
VVNM2	Infinitive lexical verb, masculine plural nominative
VVNF1	Infinitive lexical verb, feminine singular nominative
VVNF2	Infinitive lexical verb, feminine plural nominative
VVTM1N	Masculine singular (nominative) imperfective participle lexical verb
VVTM1O	Masculine singular oblique imperfective participle lexical verb
VVTM2N	Masculine plural (nominative) imperfective participle lexical verb
VVTM2O	Masculine plural oblique imperfective participle lexical verb
VVTF1N	Feminine singular (nominative) imperfective participle lexical verb
VVTF1O	Feminine singular oblique imperfective participle lexical verb
VVTF2N	Feminine plural (nominative) imperfective participle lexical verb
VVTF2O	Feminine plural oblique imperfective participle lexical verb
VVYM1N	Masculine singular (nominative) perfective participle lexical verb
VVYM1O	Masculine singular oblique perfective participle lexical verb
VVYM2N	Masculine plural (nominative) perfective participle lexical verb
VVYM2O	Masculine plural oblique perfective participle lexical verb
VVYF1N	Feminine singular (nominative) perfective participle lexical verb
VVYF1O	Feminine singular oblique perfective participle lexical verb
VVYF2N	Feminine plural (nominative) perfective participle lexical verb
VVYF2O	Feminine plural oblique perfective participle lexical verb
VVSM1	First person singular subjunctive lexical verb
VVSM2	First person plural subjunctive lexical verb
VVST1	Second person singular subjunctive lexical verb
VVST2	Second person plural subjunctive lexical verb
VVSV1	Third person singular subjunctive lexical verb
VVSV2	Third person plural subjunctive lexical verb
VVIT1	Second person singular imperative lexical verb
VVIT2	Second person singular imperative lexical verb
VVIA	Second person honorific imperative lexical verb
(Tags for general auxiliary verbs are as the tags for lexical verbs, except that they begin VX instead of VV; there is a third set of parallel tags, beginning VH, for forms of the auxiliary verb hōnā*, "be" – see also the tags for the irregular past and present tenses of* hōnā *below)*	
VGM1	Masculine singular future auxiliary *gā*
VGM2	Masculine plural future auxiliary *gē*
VGF1	Feminine singular future auxiliary *gī*
VGF2	Feminine plural future auxiliary *gī*
VRM1	Masculine singular durative auxiliary *rahā*
VRM2	Masculine plural durative auxiliary *rahē*
VRF1	Feminine singular durative auxiliary *rahī*
VRF2	Feminine plural durative auxiliary *rahī*
VC1	Singular cāhiē-type auxiliary
VC2	Plural cāhiē-type auxiliary
VHHM1	First person singular indicative present *hū~*
VHHM2	First person plural indicative present *hai~*
VHHT1	Second person singular indicative present *hai*
VHHT2	Second person plural indicative present *hō*
VHHV1	Third person singular indicative present *hai*
VHHV2	Third person plural indicative present *hai~*
VHPM1	Masculine singular indicative past *thā*
VHPM2	Masculine plural indicative past *thē*
VHPF1	Feminine singular indicative past *thī*
VHPF2	Feminine plural indicative past *thī~*
XT	Contrastive emphatic particle *tō*
XH	Exclusive emphatic particle *hī*
XHC	Clitic exclusive emphatic particle *ī*, *ĩ~*, *hī~*
XB	Inclusive emphatic particle *bhī*
ZZ	Izāfat
Punctuation marks are tagged as themselves.	