STATISTICAL PART OF SPEECH TAGGER FOR URDU

MS Thesis

Submitted in Partial Fulfillment
Of the Requirements of the
Degree of

Master of Science (Computer Science)

AT
NATIONAL UNIVERSITY OF COMPUTER & EMERGING SCIENCES
LAHORE, PAKISTAN
DEPARTMENT OF COMPUTER SCIENCE

By
Hassan Sajjad
August 2007

Approved:

____

Head
(Department of Computer Science)

Approved by Committee Members:

Advisor

____
Dr. Sarmad Hussain
Professor
FAST - National University

Other Members:

____
Mr. Shafiq-ur-Rahman
Associate Professor
FAST - National University

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Statistical Part of Speech Tagger for Urdu" by Hassan Sajjad in partial fulfillment of the requirements for the degree of Master of Science.

Dated: August 2007

Dedicated to my Parents

Table of Contents

1 Introduction

Part of speech tagging system can be viewed as consisted of two main phases which are tagset design and implementation of disambiguation technique. This report will discuss each of these phases in detail. Section 2 will discuss the parts of speech proposed by Urdu grammarians. Urdu shares its large vocabulary from Arabic and Persian and shares its morphology and syntactic structure from Hindi. However, there are standard tagging guidelines provided which aims at standardizing the tagsets of all languages of the world. The tagset of English can also be used as guideline for tagset. In section 3, tagset of related languages and earlier work on Urdu tagset will be discussed.

Section 4 will discuss the previous work on major disambiguation technologies. It will discuss the rule based, statistical and transformational based approaches for part of speech disambiguation. Machine learning approach i.e. neural network, and hybrid approaches for disambiguation will also be discussed. Redesigning of tagset on the basis of literature review will be done in section 5. A discussion on ambiguous issues of tagset is also discussed in section 5. Markov Model for disambiguation is chosen in section 6.

Section 7 will discuss the methodology of part of speech tagging process. A manual check was made on the corpus to separate the words by space. Corpus was prepared by applying normalization, and by removing diacritics and non-Urdu words. The process of manual tagging was done on 100,000 words. Various issues related to suffixation, compounding, degree of adjective and adverb, etc. were observed. A statistical part of speech tagger was implemented. It was decided that the tag of a word only depends on its own tag and a tag depends only on its previous tag. Problem of unknown word was solved by making it a candidate for a list of open class words. Disambiguation of tags was left on the tagger. Add Lambda smoothing was applied to calculate the probability of unknown word. Beam search was applied to reduce the search space.

The results of tagger are shown in section 8. Tagger showed an accuracy of 97.2% while testing on the data of 10,000 words. Tagger finds problem in disambiguating between the tags of noun and proper noun. Tagger was unable to detect the features of language based on phrase analysis. Tagger shows low accuracy while disambiguating between demonstratives and pronouns. In the end, it was concluded that the standard disambiguation techniques can be used for Urdu language.

2 Part of Speech Analysis of Urdu

The preparation of tagset may require the computational analysis of parts of speech of the language. Considering the work of Urdu grammarian in this context, their work can be viewed as influenced from two different languages. Many Urdu grammar writers use Arabic language as base line and proposed three main parts of speech for Urdu i.e. noun, verb and particle (Platts 1909, Javed 1981, Haq 1987). However, there are other Urdu grammarians which proposed nearly ten independent parts of speech for Urdu (Schmidt 1999). In this section, parts of speech proposed by Urdu grammarians will be discussed. The list of parts of speech of each grammarian can be found in appendix. However, list of parts of speech in appendix is covering tags up to two levels i.e. starting from the basic part of speech to second level distribution.

In 1909, Platts proposed a part of speech tagset for Urdu. The tagset contains three main parts of speech i.e. noun, verb and particle. Articles were not included under any part of speech. However, it was discussed separately as determiner of noun. Noun was divided into thirteen categories including three categories of adjective and ten categories for pronoun. Nouns and proper nouns were handled under one category of substantive noun. Discussion on noun was based on three features i.e. gender, number and declension. Cardinals, ordinals, collective numerals, distributives and multiplicatives, numeral adverbs, fractional numbers and RAKAM were handled under the category of numerals. In the categories of pronoun, words with marking like "اس نے" were considered as one word. A separate part of speech of reciprocal pronoun was given to the words like "ایک دوسرے". Platts did not propose any subcategory of verb. However, all properties and forms of verb were discussed as its features. Particles were divided into four categories i.e. adverb, postposition, conjunction and interjection (Platts 1909). A complete list of parts of speech proposed by Platts can be found in appendix.

In 1971, Siddiqi provides an analysis of Urdu grammar and proposed six parts of speech for Urdu. In addition to three categories proposed by Platts, Siddiqui defined a separate category for adjective and pronoun. The adverbs were also kept separate from particles. A new category named distinct was introduced. Adverbs and negative particles were catered inside the category of distinct. Noun was distributed on the basis of its structure and nature. Some semantic distributions of noun were also provided e.g. sound noun. Indefinite pronoun and relative pronoun were also distributed under noun. Numerals were also catered under nouns. On the basis of structure, noun was divided into three sub categories. Common nouns were catered under original noun. Infinitive verbs were categorized under verbal noun. On the basis of the nature of noun, it was divided into three types. Substantive noun were used to cater proper nouns. Adjectives were further divided into comparative and exaggeration. At the first level, particle was divided into construction, conjunction, تخصیص and فجائیہ. Conjunction was further divided into seven types. The details of parts of speech proposed by Siddiqi can be found in appendix (Siddiqi 1971).

Javed (1981) analyzed parts of speech of Urdu under two categories. The first category contains four parts of speech and second category contains the subtypes of particles. First category was divided into noun, verb, adjective and pronoun. Apparently, the four parts of speech look similar to those proposed by Siddiqi. But the sub types under these categories were quite different. Noun was divided into common noun, proper noun, collective noun, abstract noun and un-count noun. Most of the distributions of noun were done on semantic grounds. Adjective was divided into personal, numeral, quantitative, emphatic and pronoun. The distribution of adjective was also done on the basis of semantics differences. Verb was divided into seven types. Adverb was taken as sub-category of both verb and pronoun. Words of verbal nature were categorized under verb. Adverbial particles were also considered as sub type of verb. Ker particle (see section 5) was also categorized under verb. Pronoun was divided in ten parts of speech. Pronouns of respect were separately catered under pronoun. Particles were divided into six categories. Particles were consisted of case markers, interjection, conjunction, negative particles and intensifier. Conjunction was further divided into six types. The interjection was semantically divided into the interjection of happiness and sorrow (Javed 1981). List of parts of speech proposed by Javed can be found in appendix.

In 1987, Haq provides an analysis of Urdu grammar and proposes parts of speech based on two features i.e. consistent and non-consistent. The consistent categories were those that have some meaning attached with them. Consistent categories were divided into noun, pronoun, adjective, and verb. Non-consistent categories were those categories that alone have no meaning but they add meaning to consistent categories. Non-consistent is divided into ربط، عطف، تخصیص، فجائیہ. In consistent categories, noun was divided into common noun and proper noun. Pronoun was divided into personal, relative, interrogative, indefinite and demonstratives. Adjective was divided into personal, numeral, quantitative, نسبتی and pronoun. Adverb was catered under the category of verb. Ker particle (see section 5) was also handled under verb as separate part of speech. In comparison with Javed (1981), categories of interjection were merged into one category and no separate category for intensifier was defined (Haq 1987).

In 1999, Schmidt provides an analysis of Urdu grammar. Rather than analyzing the language as consisted of three parts of speech, Schmidt proposed ten basic parts of speech of Urdu. Schmidt analysis was very different from other grammar writers. The tagset includes noun, pronoun, adjective, adverb, postposition, verb, particle, interjection, conjunction and number as main parts of speech of Urdu. Pronouns were divided into seven types which were demonstrative, personal, reflexive, interrogative, indefinite, relative and repeated. Pronouns used as adjective were analyzed under the category of adjective. Adverbs were analyzed as time, place, manner, degree and modal. Postpositions were divided into grammatical, spatial-temporal and compound postpositions. Grammatical postpositions include کو and the inflections of کا. میں, پر, سے, تک were handled under grammatical postpositions. Verb was analyzed as based on their forms. The words with relative nature are handled inside each category. Another difference between Schmidt's tagset and other grammarian's tagset was of particles. Schmidt has included only intensifiers under particles. All other types of particles were defined as separate category. Conjunction was divided into coordinating, correlative, causal, concessive and subordinating conjunctions. The category of number was divided into cardinal, ordinal, fractional, multiplicatives, money and time. A list of parts of speech proposed by Schmidt can be found in appendix (Schmidt 1999).

3 Definition of Tagset

"The computational division of syntactic, morpho-syntactic and semantic features of a language into separate categories"

"Computational part of speech categories of a language"

Natural language processing may require building a part of speech tagset which should cover required depth of morphological and derivational categories of the language. There are three types of information that may be considered as guideline for generating a tagset. First type is the tagset of languages that are related in their morphological or morpho-syntactic or syntactic nature with source language. Previous tagsets of Urdu, Persian, Arabic and Hindi may be considered in this context. However, there are morpho-syntactic and syntactic tagsets of English language. Their analysis may also be used for the tagset of Urdu. There are general tagging guidelines provided which aims at standardizing tagsets of all languages of the world (Halteren 2005). In the following section, computational work on Urdu tagset and tagset of related languages will be discussed.

3.1 Earlier Computational Work on Tagset of Urdu

In 2003, Hardie implemented a POS tagger for Urdu. The tagset used by Hardie was based on the analysis of Schmidt and was following EAGLES guidelines of tagset. EAGLES guidelines aims at generalizing the design of the tagset. In EAGLES guidelines, general design of tagset was divided into three parts. First and compulsory part contains thirteen tags which are noun, adjective, pronoun, adverb, verb, article, adposition, numeral, conjunction, interjection, unique, residual and punctuation (Hardie 2003). The recommended attributes include number, gender, case, finiteness and other features. The optional part consists of similar attributes with lesser applicability and depends upon the language under observation. Recommended and optional attributes of EAGLES guidelines increase morpho-syntactic depth of the tagset. That's why; their inclusion in the tagset is optional.

Urdu tagset proposed by Hardie make use of all three levels of EAGLES guidelines. The tagset was based on morpho-syntactic categories of Urdu. A total of 350 tags were provided. In the tagset, noun was divided into 48 tags. Features of noun i.e. gender, number, case were explicitly handled in the tagset. All forms of verb i.e. infinitive, participles, subjunctives, imperatives were handled with separate tags. Verb was divided into 115 tags. The auxiliaries were divided into general and special auxiliaries. Special auxiliary verbs contain گا, رہا, چاہیے and ہو. Adjective was categorized as simple, determiner and Y-V-K-J determiners. The determiner adjective was used to define the categories of number, fraction, indefinite determiner. All inflection forms of ایسا, ویسا, کیسا, جیسا were handled in the tag of Y-V-K-J determiner. Multiplicative marker, adjectival particles and WALA was handled inside adjective. However, all of them and their inflectional forms get separate tag. Pronouns were divided into five categories i.e. personal, personal possessive adjective, Y-V-K-J, reflexive and other pronouns.

According to Hardie, some pronouns take adjective markings. That's why they were named as adjective. The tag Y-V-K-J represents the demonstrative nature of a category. This nature was observed in pronoun, adjective and adverb, and was handled as separate category in each distribution.

Hardie tagset contains 350 tags. All inflectional forms of a word are handled as separate category. The distribution of tags like noun, proper noun and acronym are based on semantic differences. Words with izafat are handled in two separate ways. If izafat is written then it will get a separate tag of zz. However, if izafat is not written then the two words will be handled separately. A complete list of tags can be found in appendix.

In 2007, Ijaz and Hussain proposed a tagset for Urdu. Tagset was divided into eleven parts of speech i.e. verb, adjective, common noun, adverb, numeral, conjunction, auxiliary, postposition, case marker, harf and pronoun. Each tag of the tagset contains a parameters i.e. features of the tag. The properties of each tag i.e. gender, number, case, etc. were handled inside the feature parameter of a tag. (Ijaz et al. 2007).

3.2 Tagset of Related Languages

Urdu is a language of Indo-European family. Major part of Urdu is influenced from Persian and Arabic. The vocabulary of Urdu is also loaned from these languages. The script in which Urdu is written in is based on Arabic alphabets. Urdu and Hindi are closely related languages and share their phonology, morphology, and syntax with each other. In this section, tagsets of Arabic, Hindi and English will be discussed. The detailed tagset can be found in appendix.

The Arabic grammar writers have provided morpho-syntactic tagset for Arabic which consists of 177 tags including 103 tags for noun, 57 tags for verb, 9 tags for particle, 7 tags for residual and 1 tag of punctuation. However, all Arabic grammarian sticks to main three parts of speech i.e. noun, verb, particle. All entities that include in a noun phrase are considered as types of noun i.e. common noun, proper noun, pronoun, adjective and numeral are types of noun. Verb is divided into perfective, imperfective and imperative. All other types are considered under the category of particle (Khoja, et al.).

Urdu shares its morphological and structural information from Hindi. The standard tagset for Hindi is based on the tagset of Penn Treebank. Some categories from Penn Tree are directly taken. The discussion on Penn Treebank can be found later in this chapter. In Hindi tagset, some categories are slightly changed in the tagset. New tags are also proposed according to the nature of language. The basic structure of tagset was based on syntactic categories of the language. The tagset was aimed at less number of tags and was not focusing on finer details of the language. Hindi tagset contains noun, proper noun, pronoun, verb, adjective, adverb, postposition, particles, conjunct, question word, quantifier, negative, interjection and special as main parts of speech[1]. The detail tagset can be found in appendix.

The earliest work on tagset was conducted in US and focus was on English language. Major milestone in the history of tagset was proposed by Klein and Simmons (1963).

[1] A part of speech tagger for Indian languages, available at http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

After that, Greene and Rubin (1971:1) proposed a tagset influenced from the Klein and Simmons tagset. These tagsets were based on the syntactic nature of the text. For example, verbal participles are not described with the verbal elements but with noun, adjectives and determiners.

Ellegård (1978: 96-98) used a tagset to parse text of Brown Corpus. Tagset was defined in decomposable[2] fashion. There were 25 single character tags for major word classes. However, each tag contains inflectional information about the word. The tagset was based on flat structure such that tags of noun and pronoun were having no relation between them. Penn Treebank tagset contains 48 tags. Out of them, 36 tags consist of main part of speech and rest of the 12 tags is for punctuation marks (Taylor, A., et al.). The tagset was aimed at reducing the number of tags and increasing the accuracy of the system. Tagset neglects those features of language which are recoverable at later stage. The complete list of Penn TreeBank tagset can be found in appendix.

4 Review of Part of Speech Tagging Technologies

This section will discuss different part of speech tagging technologies and the analysis of their results. At the end, technique for the tagging of Urdu will be decided on the basis of the efficiency and available resources.

A part of speech tagging system can be viewed as consisting of three main parts i.e. tokenization, assigning potential tags to each token, disambiguation by choosing most appropriate tag for a word or tagging unknown words (van Halteren and Voutilainen 1999:110). The task of assigning potential tags to a word can be done either by looking from the lexicon or by extracting some morphological information from the word and then tag it accordingly. Next phase is to remove the ambiguity and to assign the most appropriate tag to that word. Several methods are used to remove the ambiguity between the tags.

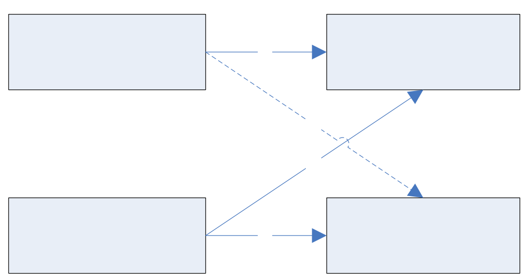


Figure 1: Methodologies for part of speech tagging (Hardie 2003)

Figure 1 describes generally used methodologies for part of speech tagging (Hardie 2003). However, hybrid approaches can also be used by combining different methodologies. Considering figure 1.1, linguist's knowledge is used to define the rules for disambiguation of tag. Corpus of text provides different types of words with their appropriate tags. Type B takes tagged corpus and on the basis of the frequency of the word with a particular tag, annotates the un-tag text. The most recent approaches to disambiguation are machine learning techniques like neural networks. Neural networks technique uses corpus data to extract linguistic information. Thus lies in category B.

[2] According to Hardie (2003: 48), if the string representing a tag having more than one character and its shorter string represent some other tag then that tagset is called decomposable. For example, tag N is used to represent a noun and some other character with N to show some additional properties of the noun.

Type C extracts the contextual information from the corpus of text and defines the rules to disambiguate the tag. Most recent work on this type of technique is done by Eric Brill in 1999. Section 4.3 will describe some work done under type C.

No work has been found under type D (Hardie 2003). This may be due to the reason that different human beings have different level of knowledge about the language. Thus, generating probabilities on the frequency of the occurrence of a word may differ from person to person.

Following section describes different approaches to disambiguation. There are three approaches that are commonly used i.e. rule based, statistical and transformational based. However, there are other approaches like finite state intersection grammar, finite state morphology, hybrid approaches to part of speech tagging, etc. (Torbjőrn Lager and Joakim Nivre, Bryan Jurish). This document will focus on the most commonly used techniques for part of speech tagging.

4.1 Rule Based Approaches to Disambiguation

Rule based approaches to disambiguation consist of a rule containing word and its contextual information. The application of rule on a particular word reduces the number of potential tags attached to that word to single tag. According to Jurafsky et al. (2005: 327), ideally rule based part of speech tagging system consists of two stages. First stage assigns each word a list of potential parts of speech by using a dictionary. Second stage uses hand written disambiguation rules to cut down the list to a single part of speech for each word.

One of the earliest works on rule based part of speech tagging was done by Klein and Simmons (1963). Their program, computational grammar coder (CGC), tags the word using lexicon and the suffix information. Set of rules are defined to remove the ambiguity. Klein and Simmons use a tagset of 30 tags and achieve accuracy rate of 90%. Greene and Rubin also use rule base approach to tag the word (Greene and Rubin, 1971). Their program, TAGGIT, follows same steps, using lexicon and the suffix information, to tag the word. However, TAGGIT was able to handle exceptions like capitalized words, words having apostrophes, etc. Greene and Rubin's disambiguation method was different from CGC. Rules were applied in order i.e. from most specific to least specific. Their first rule was based on instinct. Hardie (2003) explains it by an example that Greene and Rubin write a rule that a verb following modal auxiliary verb is infinitive rather than having present tense. Greene and Rubin then use a program to add rules by manually disambiguating the tags. These rules introduce errors of incorrectly tagging a word. TAGGIT was reported to have a disambiguation rate of 77%. Remaining ambiguity was removed manually. Later work on CG approach was done by Voutilainen (1995) and Karlsson (1995). Voutilainen (1995) made ENGTWOL tagger which was based on early rule based systems of two stage architecture, although both lexicon and rules were much complicated than early once. Hindle (1989) works on disambiguating words in a deterministic parser and analyzes rule based tagger without giving any information of the syntax. Other work on rule based tagger was done by Brodda (1982), Paulussen and Martin (1992) and Brill et al (1990).

4.2 Statistical Approaches to Disambiguation

Statistical approaches are based on the information from the corpus of text. Corpus of text provides the frequency of the sequence of tag which will help in disambiguating the sentence by choosing the sequence of tag with highest frequency. The work on statistical part of speech tagging started in late 1970's. Some initial work was done by Bahl and Mercer (1976) and Debili (1977). However, significant work on probabilistic part of speech tagging started when Garside and Leech (1985), and Beale (1985) provide the probabilistic formulation of disambiguation problem in part of speech tagging. In 1986, Derouault and Merialdo did some significant work for the training of statistical parameters. Derouault and Merialdo (1986) manually tag a small amount of text and then use a bootstrap method to tag large corpus. Church (1988) and Kempe (1993) use second order Markov Models for disambiguation. Training of their system is done by using a large hand tagged corpora. Using this method, Church (1988) and Kempe (1993) are able to tag 96% of words correctly. The problem arises for languages that are not having any training data available. Jelinek (1985) and Cutting et al. (1992) overcome the problem of tag training data and train their taggers on untagged data using Baum-Welch algorithm. The results provided by Jelinek (1985) and Cutting et al. (1992) were comparable with Church (1988) and Kempe (1993).

4.3 Transformational Based Approach

Transformation based approach for tagging is a machine learning approach (Brill, 1995). It was inspired from both rule based and stochastic taggers. Like rule based systems, transformational based learning is based on rules. Like probabilistic approach, rules are automatically induced from the data (Jurafsky et al. 2005, 333).

Transformational approach for tagging, called Brill tagging, is not a disambiguation technique. It is a learning or improvement technique. It takes an unambiguously tagged text to learn from it. Pre-tagged corpus is used to evaluate the results of the rules. System starts by running an initial state annotator on an un-tagged corpus. This process assigns a single tag to each word based on the lexicon in which frequency of word with the tag is given. This tagged corpus is compared against pre-tagged corpus and list of rules are learned. These rules are applied on the output taken from state annotator. After applying these rules, success of transformation is measured by comparing it with the reduction in errors. The list of transformations is ordered from most effective to least effective. The process of adding rules ends when no more transformations can be found that improve the tagging (Hardie 2003, 271).

Brill (1992) argues about the advantages of transformational based approach over rule based and stochastic approaches. According to Brill (1992), in rule based approaches, it is difficult to construct rules and in probabilistic approaches much space is required to store the tables of frequencies. Transformational based approach overcomes these issues by providing an automatic extraction of rules. Space required to store these rules is less than storing the probabilistic information. Other advantages describe by Brill (1992) is that it is easy to use Brill's tagger with other tagsets or with different languages.

4.4 Other Approaches to Disambiguation

Neural Networks Approach:

According to Hardie (2003: 280), neural network approach to disambiguation is a machine learning approach. It consists of interconnected layers where each layer works as a processing unit. On activation of a layer, it connects with other layers with weighted links. Weights given to the links and the activation values of the units are the parameters of the network. Figure 2 provides an overview of 3-layer structure of neural network (Schmid 1994).

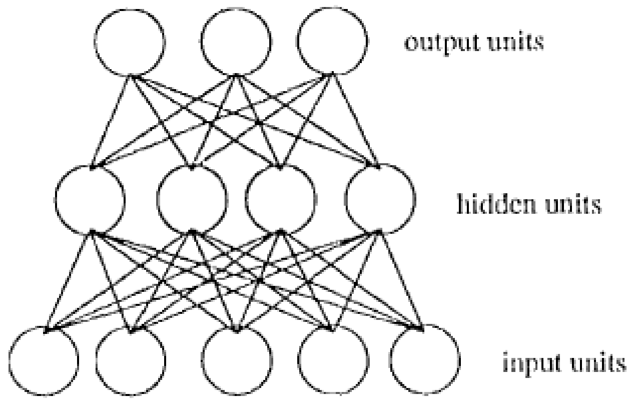


Figure 2: A 3-layer structure of neural network

The bottom layer is called the input layer and top layer is called the output layer. Layers between input and output layers are called Hidden layers as only the input and the output layers are visible. The training of neural network can be done by adjusting the weights of the links and the activation values of the layers or units (Hardie, 2003: 281).

Neural network system takes ambiguously tagged word and its contextual information as input. Input layer consists of a set of units equal to the number of tags in the tagset. For each word, all tags with which a word was marked are activated. Network knows about the correct tag due to the training and deactivates other output units. The use of contextual information varies from system to system. Schmid (1994) takes three preceding words and 2 following words as contextual information of a word. According to Schmid (1994), reducing contextual information from three preceding words and one following word to two preceding words and one following word decreases the accuracy only by 0.1%. Increasing the contextual information to three preceding words and two following words showed no improvement in accuracy.

Hardie (2003: 283) finds the performance of neural network taggers comparable with the performances of rule based and probabilistic approaches. Schmid (1994) reported an accuracy rate of 96.22% and found it better than Markov model tagger.

4.5 Hybrid Approaches to Disambiguation

A hybrid tagger can be defined as a combination of disambiguation techniques use to serve the purpose of a single disambiguation technique. Hybrid methods are ideally be used to increase the accuracy of the system.

CLAWS system is a good example of hybrid approach. In CLAWS1, the WORDTAG lexical analysis component has initially assigned potential tags which were altered by rule based component IDIOMTAG. After that a stochastical disambiguator was applied (Hardie 2003).

CLAWS system gives an example of hybrid approach in which both rule based system and stochastic system were developed together. Tapanainen and Voutilanien (1994) do an experiment to combine rule based system, EngCG, and stochastic disambiguation system, Xerox tagger, initially developed as separate systems. These two taggers were having complementary strength i.e. EngCG is rarely wrong but does not disambiguate fully whereas Xerox tagger is less reliable but disambiguate fully (Hardie 2003: 292). Tapanainen and Voutilanien run both taggers parallel on same text and then combine both outputs by allowing Xerox tagger to resolve the ambiguities left by the EngCG tagger. Results were found to have accuracy rate of 98.5% which were better than any of the tagger.

5 Redesigning of Urdu Tagset

Tagset of a language caters main parts of speech as well as morphological information of the language. There are various issues that need to consider for the efficient design of tagset. First problem is about the level of categorical distribution that the tagset should contain. A tagset may be consisted either of syntactic categories or it may be consisted of morpho-syntactic categories. Considering the efficiency in machine learning process and to reduce lexical and syntactic ambiguity, it was decided to concentrate on the syntactic categories of language. The syntactic categories lead to less number of tags which also improves accuracy of manual tagging[3] (Taylor, A., et al.).

Considering the work of Urdu grammar writers, most of the categories were based on semantic differences. The morphological information of the categories was either handled through separate parts of speech or was considered as features of the language. Most of the categories were lacking their computational side. However, the detailed analysis of these grammar writers really helps in covering the depth of the language. The tagset of Hardie was properly covering the features of the language. However, Hardie tagset was based on morpho-syntactic categories of Urdu. Some of the tags were divided on the basis of semantic differences (see section 3.1). For a syntactic tagset, the features of Urdu language need to be analyzed on the basis of the structure of the language. It was also mentioned in the literature that smaller tagset improves the accuracy of the tagger. Following is the redesigning of tagset on the basis of the work of Urdu grammarians and earlier tagsets of Urdu.

There were three types of corpus available for analysis i.e. literature, news and poetry corpus. For the design of tagset, only literature and news corpus was analyzed. The corpus was based on the most recent available vocabulary used by local people.

Following is the proposed list of POS tags followed by some of their examples. The syntactic analysis on the tags is done in discussion section.

Demonstrative: Demonstratives are divided into four categories. All four categories of demonstratives have ambiguity with four categories of pronoun. Phrase level analysis was done to distinguish between demonstrative and pronoun. The detailed comparison of demonstrative and pronoun can be found in discussion section. Following are some examples of demonstratives.

Personal demonstrative (PD)	This category includes the elements of demonstrative and personal demonstratives. Following is an example of it.

[3] A part of speech tagger for Indian languages, available at http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf.

هم، تم، آپ، یہ، وہ، اس

یہ <PD> مسجدیں <NN> ہماری <G> پہچان<NN> ہیں <VB>.<SM>

Relative demonstrative (RD)
جو، جن، جنہوں

جو<RD> لڑکا<NN> صبح<NN> آیا<VB> تھا<TA> وہ<PP> میرا<G> دوست<NN> ہے<VB>.<SM>

Kaf demonstrative (KD)
کن، کوئی

کن<KD> لوگوں<NN> کو<P> آم<NN> اچھا<ADJ> لگتا<VB> ہے<TA>.

کمرے<NN> میں<P> کوئی<KD> لڑکا<NN> نہیں<NEG> ہے<VB>.

Adverbial demonstrative (AD)
اب، تب، ادھر، یہاں

میں <PP> ایسا <AD> کام <NN> نہیں<NEG>کر <VB>سکتا<AA>.

Nouns: Nouns are divided into two categories. First category consists of simple nouns which are represented by NN in the tagset. However, there are other nouns that show adverbial nature like time, place, manner, etc. These are also catered under noun. The proper nouns are kept in a separate category. Following are some examples of different types of nouns.

Noun (NN)
جہاز، زمین، درخت، لڑکا،
اوپر، اندر، سمیت، طرح، طرف

یہ <PD> مسجدیں <NN> ہماری <G> پہچان<NN> ہیں <VB>.

چھت<NN> کے <P> اوپر <NA> حامد<PN> ہے<VB>.

Proper noun (PN)
لاہور، پشاور، پاکستان

لاہور<PN> باغات <NN> کا <P> شہر <NN> ہے<VB>.

Pronouns: Pronouns are divided into six categories based on their syntactic structure. Most of the categories are consistent with the types provided by Urdu grammarians. The analysis and justification of the newly proposed categories can be found in discussion section. Following are some examples of the types of pronouns.

Personal pronoun (PP)
میں، ہم، تم، آپ، یہ، وہ، اس

میں <PP> تمھارا <G> دوست <NN> ہوں <VB>.

Reflexive pronoun (RP)
خود، آپ

میں <PP> اپنا <GR> کام<NN> خود<RP> کروں<VB> گا<TA>.

Relative pronoun (REP)
جو، جن، جنہوں

علی<PN> جو <REP> حامد<PN> کا <P> بھائی <NN> ہے<VB> میرا <G>دوست <NN> ہے<VB>.

Adverbial pronoun (AP): The adverbial pronouns occur at the place of nouns with adverbial nature and show the property of time, place, manner, etc. They are represented by AP in the tagset. Consider the following examples:

Example:
اب، تب، ادھر، یہاں

علی<PN> نے <P> اب <AP> کھانا <NN> کھایا <VB> ہے<TA>.

Kaf pronoun (KP): Kaf pronouns add interrogative property in the sentence. They are divided into two categories. Kaf pronouns, represented by KP, are used to ask question about a noun. The second category includes adverbial kaf pronouns which are used at the place of nouns with adverbial nature. Following are their examples:

Kaf pronoun (KP)
کون، کوئی، ، کن

کمرے<NN> میں<P> کون <KP> ہے<VB>۔

Adverbial kaf pro (AKP)
کدھر، کب، کیسا

علی <PN> کدھر <AKP> گیا <VB> ہے<TA>۔

Genitive reflexive (GR)
اپنا

اپنا<GR> کام<NN> خود<RP> کرنا<VB> میرا<G> فرض<NN> ہے<VB>۔

Genitives (G) Consider the above example of genitive reflexive.
میرا، تمھارا، ہمارا، تیرا

Verb (VB): At sentence level, any word showing action in any form is considered as verb. No further categorization is done. Consider the following examples of verb:

Example:
لکھنا، کھاتا، جاتا، کرنا

وہ <PP> روٹی<NN> کھا <VB>رہا<AA> ہے<TA>۔

Auxiliaries: Based on the syntactic nature of language, auxiliaries are divided into two categories. Aspectual auxiliaries always occur after main verb of the sentence. Tense auxiliaries are used to show the time of the action. They occurred at the end of the verb phrase. Consider the examples of aspectual and tense auxiliaries:

Aspectual auxiliary (AA) Consider the example of verb.
رہا، کرنا، چکہ

Tense auxiliary (TA)
ہے، ہیں، ہوں، تھا، تھے، تھیں، گا، گی، گے، ہو، ہوں

Consider the above describe examples.

Adjective (ADJ): Adjectives are catered as one category. The information related to the degree of adjective is not taken into account. Following are given some examples of adjectives.

ظالم، خوبصورت، کمزور، بیکار، سمجھدار، نفیس

حامد <PN> بہت <ADV> ظالم <ADJ> لڑکا <NN> ہے<VB>۔

Adverb (ADV): Adverbs are handled as one category in the tagset. Consider the following examples of adverbs.

Example:
بہت، نہایت، بڑا

وہ <PP> بڑا <ADV> محنتی <ADJ> لڑکا <NN> ہے<VB>۔

Quantifier (Q): Consider following examples of quantifier:

Example:
کچھ ، چند،تمام، اتنے، سب، تھوڑا، تھوڑے، کئی، بعض، کل

سب <Q> لوگ<NN> تھوڑا <Q> انتظار<NN> کریں <VB>۔

Numerals: Numerals are divided into four categories based on their syntactic structure. Cardinal (CA), ordinal (OR), fractional (FR) and multiplicative (MUL) are types included in the tagset. Following are the examples of each category.

Cardinal (CA)
ایک، دو، تین، چار بیالیس،
انسٹھ، ننانوے، ہزار، دو ہزار

پہلے <OR> دو <CA> لڑکوں <NN> کو<P> بلاؤ <VB> .

Ordinal (OR) Consider the example of cardinal.
پہلا، دوسرا، تیسرا، چوتھا، پانچواں، چھٹا، ساتواں، آٹھواں، آخری

Fractional (FR)
چوتھائی، ڈھائی، اڑھائی

ڈھائی <FR> کلو<U> دودھ<NN> دینا<VB>-

Multiplicative (MUL)
گنا، دگنا، دہرا، تہرا

علی <PN>حامد<PN> سے<P> دگنا<MUL> موٹا<ADJ> ہے<VB>-

Measuring unit (U): They are frequently used with numerals. However, they have a different syntactic structure than numerals. Consider the example of fractional to see the occurrence of measuring units.

Example:
پون، پائو، کلو، سیر

ڈھائی <FR> کلو<U> دودھ<NN> دینا<VB>-

Conjunction: Conjunctions are divided into coordinating and subordinating conjunctions. Following are their examples:

Coordinating (CC)
اور, یا

حامد<PN> اور<CC> علی<PN> اچھے<ADJ> دوست<NN> ہیں<VB>-

Subordinating (SC)
کیونکہ, کہ

حامد <PN> سے<P>کہو<VB>کہ<SC>مجھ <PP> سے<P>ملے<VB>-

Intensifier (I): There are only three words in this category. Consider their following examples:

Example:
ہی، بھی، تو

میں <PP> بھی <I> آؤں <VB>گا.<TA>-

Adjectival particle (A): This category includes only one word sa with its two inflection forms. This particle is normally used for comparison. Consider the following examples of adjectival particle.

Example:
سا، سے، سی

مینڈک<NN> ایک<CA> عجیب<ADJ> سا<A> جانور<NN> ہے<VB>-

KER particle (KER): These particles normally occur in verb phrase. There are only two entities in this class. Consider the following examples:

Example:
کے، کر

گھر<NN> پہنچ<VB> کر<KER> فون<NN> کر<VB> دینا<AA>-

Title: Titles are divided into two categories based on their pre and post occurrence around a proper noun. Consider their examples below.

Pre-title (PRT)
حضرت، میاں

میاں<PRT> سرمد<PN> صاحب<POT> اچھے<ADJ> انسان<NN> ہیں<VB>۔

Post-title (POT)
جی، صاحب

Consider the example of pre-title above.

Semantic Marker (P): Following are the list of particles included into this category. However, the entity سے is kept as separate category due to its ambiguous usage.

کا ، کو ، کی ، کے ، نے، میں، تلک ، پر، تک

حامد<PN> کو<P> علی<PN> نے<P> چھڑی<NN> سے <SE> مارا<VB>۔

SE (SE): سے Consider the above example

Wala (WALA): This category contains one word wala and its inflections. Consider its examples:

Example:
والا، والی، والے

پھل<NN> بیچنے<VB> والا<WALA> آدمی<NN> آیا<VB> ہے<TA>۔

Negation (NEG): Consider the following examples of negation.

Example:
نہ، نہیں

میں <PP> ایسا <AD> کام <NN> نہیں<NEG> کر <VB> سکتا<AA>۔

Interjection (INT): Interjections normally occur at the start of the sentence. They are kept as separate category in the tagset. Following are its examples:

Example:
واہ, سبحان اللہ، اچھا

واہ<INT> کیا<ADV> اچھی<ADJ> بات<NN> کی<VB> ہے<TA>۔

Question words (QW): There are some words instead of kaf pronouns that are used for the interrogation in the sentence. However, these words cannot be replaced by a noun or pronoun. A separate category of question words has been formed for these words. Consider their examples below:

Example:
کیا، کیوں

کیا <QW> علی <PN> سکول <NN> جائے <VB> گا.<TA>۔

Punctuation marks: In this tagset, punctuation marks are divided into two categories. Sentence markers mark the boundary of the sentence. Phrase markers are used inside the sentence but never used at the end of sentence. Consider their examples below:

Sentence marker (SM) '.', '?'

Phrase marker (PM) ',' , ';'

DATE 2007, 1999

Expression (Exp): Any word or symbol which is not handled in this tagset will be catered under expression. It can be mathematical symbols, digits, etc.

5.1 Discussion

Considering above tag set, noun is divided into noun and proper noun. However, in the tagset, it is mentioned that nouns with adverbial nature are also kept under noun. These nouns contained information about time and place. Due to this reason, most of the grammar writers categorize them as noun of time and place (Platts 1909, Javed 1981, Haq 1987). However, some grammar writers also consider them under adverbs (Schmidt 1999). Looking at the language syntactically, these elements with adverbial nature occur at the place of noun. To make syntactic structure of language consistent, it was decided to consider them under noun. Following are some examples of it.

صبح<NN> اٹھنا<VB> اچھی<ADJ> عادت<NN> ہے<VB>۔

چھت<NN> کے<P> اوپر<NN> حامد<PN> ہے<VB>۔

Pronouns are divided into six types based on their syntactic nature in the sentence. The adverbial pronouns are of same nature like nouns with adverbial features. That's why, they are categorized under pronoun.

حامد <PN> نے<P> صبح <NN> کھانا <NN> کھایا<VB>۔

حامد <PN> نے<P> تب <AP> کھانا <NN> کھایا<VB>۔

Usage of Adverbial pronoun تب

Most of the categories involved in pronouns are similar with demonstratives. Difference was analyzed on the basis of their phrase boundary. It was observed that pronouns occur as standalone unit in a phrase or occur without having a noun as its neighbor in a phrase whereas demonstratives make phrase boundary with the next noun. The adverbial pronouns are also showing similar behavior. Consider following examples:

یہ <PP> حامد<PN> کا<P> بھائی<NN> ہے<VB>۔

یہ <PD> مسجدیں <NN> ہماری <G> پہچان<NN> ہیں <VB>۔

In case of pro-drop, demonstrator becomes the pronoun. Consider the example below; if word لوگ (people) is dropped here then وہ will become the pronoun here.

Without pro-drop

وہ <PD> لوگ<NN> گانا<NN> گاہیں<VB> گے<TA>۔

After pro-drop

وہ<PP> گانا<NN> گاہیں<VB> گے<TA>۔

Kaf pronouns are divided into two categories. Both are actually question words that can be replaced by a noun. However, syntactic structure of adverbial kaf pronoun is different from other kaf pronouns. While observing kaf pronouns in general, the ambiguity was found with the demonstratives. Phrase level analysis as explain above is used to distinguish between kaf pronoun and demonstratives. The demonstrators with interrogative nature are kept inside demonstrative category. Consider following examples of kaf pronouns, adverbial kaf pronouns and kaf demonstratives.

Kaf pronoun

کمرے<NN> میں<P> کون<KP> ہے<VB>۔

کن<KP> لوگوں<NN> کو<P> آم<NN> اچھے<ADJ> لگتے<VB> ہیں<TA>۔

Adverbial kaf pronoun

حامد<PN> کدھر<AKP> گیا<VB> ہے<TA>۔

Demonstrative

تم <PP>کن <KD> سے<SE> ملنے<VB> جا<VB> رہے<AA> ہو<TA>۔

KER tag contains two elements کر ، کے (Javed 1981). These particles occur in verb phrase and semantically show the completion of verb. Following are there examples:

میں<PP> کام<NN> کر<VB> کے<KER> تھک<VB> گیا<AA> ہوں<TA>۔

میں<PP> کام<NN> کر<VB> کر<VB> کے<KER> تھک<VB> گیا<AA> ہوں<TA>۔

میں<PP> وہاں<AP> جا<VB> کر<KER> تھک<VB> گیا<AA> ہوں<TA>۔

Semantic marker is containing particles that show the semantic marking of subject, object and indirect object, etc. (Butt et al. 2001). The marking objects are also called semantically motivated cases as they are used to express semantic motivations (Butt et al. 2001). Due to this reason, they are not separated under more than one category. However, SE is kept separate under unique category due to its ambiguous usage.

WALA والا is considered a unique entity due to its different morpho-syntactic nature. It is categorized under adjective and noun by Urdu grammar writers (Javed 1981, Schmidt 1999). However, it is still considered as an issue due to varied usage. For this tagset, it is decided to handle it as a separate tag.

Expression includes symbols, mathematical formulae, digits, etc. In general, this tag caters any exceptional word or character that occurs in the text. There might be a case when two exceptional characters or words are occurring consecutive. In that case, only one expression tag will be assigned.

6 Selecting Disambiguation Approach for Urdu

Literature review of disambiguation approaches can be summarized as follows:

- Rule based approach
- Probabilistic approach
  - Markov model
- Transformational based learning
- Other approaches like neural networks
- Hybrid approaches

There are many factors that play an important role while selecting a disambiguation approach. Performance of disambiguation approach, properties of the language, nature of the tagset, available resources, and time limitations, all played an important role in the selection of an approach.

According to Daelemans (1999: 303-304), methods like neural networks have several advantages over statistical methods such as requiring less training data, fewer parameters and fast training procedure. However, Daelemans provides some counter arguments in support of statistical methods such as the effectiveness of new technologies has not been evaluated fully.

Considering the performance of the systems, Markov model taggers generally achieve an accuracy of 97% (Hardie 2003: 295). Brill (1995) reports a similar accuracy rate. Voutilainen (1995: 186-187) reports an accuracy rate of 99.7%-100% using rule based CG methodology. For comparability, these are small performance differences. Thus, choosing the methodology on the basis of performance of the system is difficult.

Consider language; Urdu is written in Perso-Arabic text, the texts in question are coded in Unicode. Brill (1995) and Cutting et al. (1992)'s tagger require ASCII text. So, it is possible to rule out these two taggers.

Urdu is a highly inflected language and having SOV word order. Sánchez León and Nieto Serrano (1997: 163-164) suggest that the potentially free order of language could lead to greater ambiguity i.e. it becomes harder to guess the tag of a word on the basis of its context. This might suggest that for a language like Urdu, probabilistic model would be unsuitable. Dandapat et al. (2006) implemented a Markov model for Bengali which is a free order language and reported an accuracy of 89%. Brill (1995: 544) reports that all disambiguation techniques utilize the same kind of information. Thus probabilistic model can not be ruled out by just arguing that the language is free order.

The nature of the tagset may affect the performance of disambiguation method. Tapanainen and Voutilainen (1994) suggest that Markov model taggers operate better with small tagsets, whereas rule based approaches operate better with larger tagsets. Sánchez León and Nieto Serrano (1997) work on Spanish tagsets ranging from 40 to 475 tags and use them with Markov model and report that larger tagset improves performance if the model has appropriate biases. Thus size of the tagset may not help in deciding the disambiguation technique.

Let's consider the practical benefits and drawbacks of the probabilistic approach, rule based approach and hybrid approach. Hybrid approach uses the best features of several methodologies. Tapanainen and Voutilainen (1994) create a hybrid tagger from two pre-existing taggers. In case of Urdu, one rule based tagger is available (Hardie 2003). Hybrid approach requires at least one more tagger for Urdu. Considering the time limitation of the thesis, only one approach can be implemented and hybrid work can be left for future research. Therefore, hybrid approach can be ruled out.

According to Weischedel et al. (1993), having a corpus of limited vocabulary; the probabilistic models offer a mathematically grounded means of predicting the most likely tag. In case of unknown words, probabilistic models provide the best solution. Weischedel et al. (1993) also mention that for a given vocabulary size, it is difficult to provide full syntactic and semantic features by handcrafted rules. Probabilistic models overcome this limitation by considering contextual information from the corpus. Another point mentioned by Weischedel et al. (1993) was that rule based approach do not perform well on long sentences on which probabilistic approach can effectively operate.

Now considering Urdu, a corpus of approx eighty million words is available. The number of unique words in the corpus is about 52,000[4]. Thus, shows a good frequency of the words in the corpus. Making the rules of 52,000 words over the corpus of 10,000,000 words seem cumbersome and much time consuming. Here, after considering the resources and the analysis of different writers, rule based approach can be ruled out. Hence, for the current work, statistical approach can be used for part of speech tagging.

7 Methodology

This section will discuss the steps followed in the implementation of part of speech tagger. The availability of training data is the first step towards the automatic annotation of text. A corpus of 110,000 words was selected from two domains. After applying normalization and removing diacritics, data of 100,000 words was manually annotated for training. In the implementation of part of speech tagger, Hidden Markov model was implemented. Add Lambda smoothing was applied to avoid zero probabilities. In order to shorten the search space and to speed up the time, beam search was applied. The detail discussion of each step is as followed.

7.1 Preparation of Corpus

The accuracy of a tagger also depends on the corpus. The inclusion of foreign words, free orderness in the corpus significantly affects the results of the tagger. A corpus of amount eighty million words was taken from Jang (www.jang.com.pk). The available eighty million corpus was based on six domains i.e. games, news, finance, culture entertainment, consumer information and personal information. At start, it was decided to drop the corpus of games, finance and consumer information due to the excess of foreign words in the corpus. At later stage, personal information was also dropped due to the lack of structure of the corpus. Out of the domain of news and cultural entertainment, 110,000 words were selected as corpus. Before actually starting the annotation, corpus was gone through various steps in order to maintain the consistency of the text.

7.1.1 Normalization

Urdu shares its character set with Arabic. There are characters in Urdu that can be represented by more than one Unicode. This problem of inconsistency was frequently seen in the corpus. In order to keep the characters consistent, normalization was applied before doing any processing on the corpus. Following is the list of normalizations applied.

Table 1: Normalization

Problem words	Unicode	Normalized words	Unicode
ة	0629	ۃ	06C3
ك	0643	ک	06A9
ه	0647	ہ	06C1

4 Corpus of Urdu is available with Centre for research in Urdu language processing (CRULP). Further detailed about the corpus can be found in section 7.1

ى	0649	ی	06CC
ي	064A	ی	06CC
ۀ	06C0	ۂ	06C2
,	002C	،	060C
.	002E	۔	06D4
;	003B	؛	061B
?	003F	؟	061F
آ	0622	آ	0627 + 0653
أ	0623	أ	0627 + 0654
ؤ	0624	ؤ	0648 + 0654
ۂ	06C2	ۂ	06C1 + 0654
ۓ	06D3	ئے	0626 + 06D2

7.1.2 Other Issues

In Urdu, most of the diacritics are considered optional. Due to optionality of diacritics, two similar words one with diacritics and other without diacritics do exist in the corpus. Therefore, it was decided to remove the diacritics from the corpus. It was also observed that there occur some non Urdu characters in the corpus. These words were also deleted from the corpus. A List of diacritics and non-Urdu words is given below.

Table 2: Diacritics and non-Urdu words

Diacritics	Non-Urdu words
(0650) ِ	"
(064B) ً	*
(064F) ُ	#
(064D) ٍ	$
(064C) ٌ	%
(0670) ٰ	&
(0652) ْ	'
(0656) □	*
(0654) ٔ	+
(060C) ،	-
(0651) ّ	/ \
(0657) □	<>
(0659) □	=
(0640) ـ	@
(0653) ٓ	( )
(FDFA) □	^
(064E) َ	|
	~
	`
	”
	‘
	’
	“

7.2 Manual Tagging

A corpus of 100,000 words was selected for manual tagging. After applying normalization and by removing diacritics, test corpus was divided into 10 equal parts. A word list of the corpus was generated and each word was given its expected tag. This lexicon was further use to speed up the annotation process. Each part of the corpus was first annotated with the generated lexicon. All potential tags of each word were assigned. The errors were manually removed from the corpus. Same procedure was repeated up to 50,000 words. Rest of the 50,000 words was automatically annotated from the tagger and was manually checked for errors. This procedure speeds up the manual tagging process and helps in analyzing the issues of the tagger and the corpus. Following section will discuss some linguistic issues faced while manually annotating the corpus.

7.2.1 Suffixation

The problem of considering suffixes as one word or considering it as part of its root word was faced during annotation. Considering suffix as separate word may create the problem of including a non-word in the lexicon. Some suffixes like ناک do exist as separate word but their usage as suffix makes it an adjective rather than a noun. This way of handling suffixation may also disturb the learning of statistical tagger and increase the ambiguity for the tagger. Consider the following example:

Table 3: Three ways of tagging the word having a suffix

(a)	(b)	(c)
خوفناک<ADJ> انسان<NN>	خوف<NN> ناک<NN> انسان<NN>	خوف<ADJ> ناک<ADJ> انسان<NN>

In the above example, the word with suffix can be tagged in three ways. Part b is lexically assigning the tags to the words. This will tag the word independent of its context. Thus, lose the actual feature of the word. Part c is separately tagging the word and suffix but assigning the tag according to the context of the word. This will wrongly guide the machine learning process as in this way noun is followed by two adjectives rather than one. The ambiguity for word ناک will also be increased. For these reasons, it was decided to consider the root and suffix as one word.

7.2.2 Words with Zer-e-izafat

In Urdu, combining words with zer-e-izafat is a very common phenomenon. Sometimes these words cannot be separated as two words or can be replaced by having semantic marker in it. Consider the following example:

Table 4: Two cases of words with zer-e-izafat

(a)	(b)
وزیراعظم<NN>	وزیر<NN> صحت<NN>
* اعظم کا وزیر	صحت کا وزیر

Here, it is clear that part (a) of the example becomes ungrammatical when replaced. That’s why, it was decided to consider (a) as on word and consider (b) as two separate words.

7.2.3 Verb Phrases Acting as Adjective

It was observed that the occurrence of verb phrase at the place of adjective is very frequent in corpus. Consider the following example:

Table 5: Verb phrase acting as adjective

(a)	(b)	(c)
روتے<VB> ہوئے<AA> بچے<NN>	روتے<ADJ> ہوئے<ADJ> بچے<NN>	روتےہوئے<ADJ> بچے<NN>

There are three possible ways of tagging this problem. However, example (c) is not appropriate as we are considering two words as one word. Example (b) is again causing problem to machine learning process. At the end, it was decided to treat verb and auxiliaries independent of its context.

7.2.4 Complex Predicate

There are some words which are noun and adjective, and occur in a verb phrase. These words are called complex predicates (Butt 2003). When these words were analyzed separately, it becomes very difficult to distinguish them either noun or adjective. For the current work, it was decided to keep word and its tag consistent throughout the training corpus. However, a practical solution to this problem is discussed later.

7.3 Computational Modeling

This section will discuss the techniques used in the implementation of a part of speech tagger. Hidden Markov Model was used as disambiguation technique. In order to reduce the search space of the tagger, beam search was applied. Frequency of unknown words is handled by applying Add-Lambda smoothing. Following is the detailed discussion of each technique.

7.3.1 Design

Design of application was divided into three components. Pre-processor and training database works as standalone unit. Output of pre-processor and training database is used by tagger to annotate the text. Pre-processor takes input in the form of text file. After applying normalization rules, diacritics and symbols were removed from the input. Training database takes annotated text in the form of a text file and calculates the unigram word tag probabilities and the probability of a tag t_i given its previous tag t_{i-1}. The words from the list of word tag probability will be used as lexicon by the tagger. Tagger takes two inputs, one the output of pre-processor and second the output of training database and outputs the annotated text. The detail discussion on the working of each module can be found in next section. Following is the design diagram of the tagger.

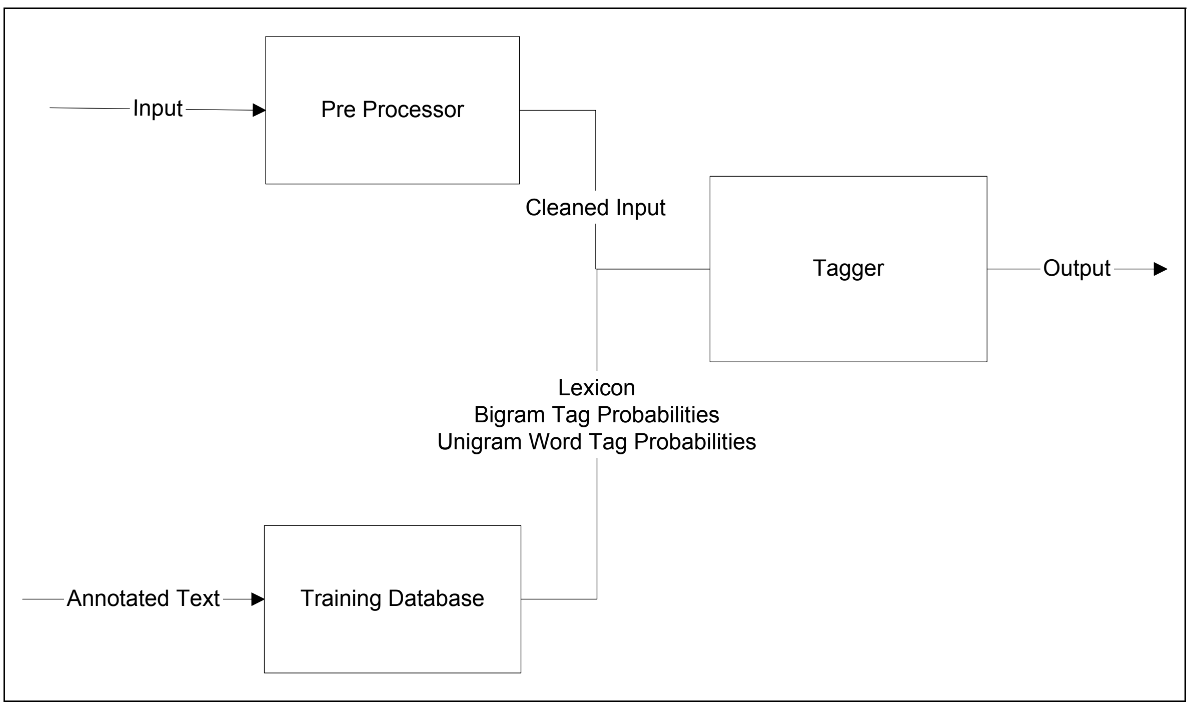


Figure 3: Design of the tagger

7.3.2 Pre-processor

In order to control the consistency between training data and input text, a separate module called pre-processor was build. Pre-processor module takes input in the form of a text file and normalize the text. Diacritics were also removed from the text. Following is the algorithm of pre-processor.

- Take input from a text file
- Load normalization rules
- Load a list of diacritics
- Load a list of symbols
- Apply normalization
- Remove diacritics
- Remove symbols from the corpus
- Save the output in a text file

A list of normalization rules, symbols and diacritics can be found in section 7.1.

7.3.3 Training Database

Part of speech tagger takes information from three databases i.e. lexicon, word tag probabilities and tag tag probabilities. These information sources are built by training database by using annotated text or training text as input. In the implementation, a separate module was built for each database. Following is the discussion on each algorithm.

General Algorithm of training database

- Take annotated text from a text file
- Calculate total counts of each word tag pair i.e. total number of occurrences of each word w with tag t
- Calculate total counts of each tag tag pair i.e. total occurrences of each tag t_i having previous tag t_{i-1}
- Calculate total counts of each tag i.e. total occurrences of each tag t_{i-1}
- Apply smoothing (next section)
- Calculate probabilities
- Save the probabilities of word tag pair and tag tag pair in separate files

Probability calculation was done using following formula:

Word tag probability $P(w_i \mid t_j) = C(w_i\, t_j) / C(t_j)$
Tag tag probability $P(t_i \mid t_{i-1}) = C(t_i\, t_{i-1}) / C(t_{i-1})$

In order to calculate the probability of unknown word, smoothing was applied by introducing an unknown pair in the database. The smoothing algorithm can be found in next section.

7.3.4 Tagger

Application of part of speech tagger takes two inputs i.e. cleaned input text from pre-processor and other is the databases. Input text is observed sentence by sentence by the tagger. Tagger creates the annotated output of each sentence. Following is algorithm of the tagger.

- Read input from text file
- Load databases
- Divide input on the basis of sentence marker
- Continue until input ends
    - Take a sentence

  - Repeat until sentence end
    - Take a word from sentence
    - Assign its potential tags from lexicon or assign potential tags for an unknown word (Make branches if potential tag > 1)
    - Assign word tag probabilities to the pair
    - Assign tag tag bigram probabilities
    - If number of branches are more than Beam size (say 10), sort all branches on the basis of cumulative score up to the current word and take top 10 branches (highest score)
  - Save the output sentence by sentence
- Write output in a text file

At sentence level, file containing probabilities of word tag pair was used as lexicon for the tagger. Hidden Markov model was used as disambiguation technique. Problem of unknown word was handled by assigning a list of candidate tags to that word. Zero probability of unknown word was handled by applying Add Lambda smoothing. Following are the details on Hidden Markov model, Add Lambda smoothing and unknown word handling.

7.4 Implementation Techniques

7.4.1 Markov Model for Part of Speech Tagging

Hidden Markov model is used to estimate the best sequence of tags for a sentence. It utilizes a tagged corpus to estimate the frequency of the occurrence of a tag with a word. It is called Hidden as the actual sequence of states i.e. tag generated for a sentence is unknown. According to Rabiner (1989), Hidden Markov model has five parameters (Scott M. Thede et al).

1. Total number of states in the model is represented by N. For part of speech tagger, N is the total number of tags used by the system. One tag consists of one state.
2. Total number of output symbols and is represented by M. For part of speech tagging, M will be the number of words in the lexicon of the system.
3. Probability of moving from state i to state j and is represented by a_{ij}. It is called transition probability of the states. For part of speech tagging, state transition probability will be the probability of moving from tag i to tag j in other words, probability that tag j will follow tag i. This probability is normally estimated from the corpus.
4. Observation probability $b_j(k)$ will be the probability of having symbol k on state j. For part of speech tagging, it will be the probability of word having tag j.
5. Initial state distribution π_j is the probability that model will start in state i. For part of speech tagging, this is the probability that the sentence will start with tag i.

Choosing HMM for part of speech tagging will determine the most likely tag sequence that generates the words in the sentence. Following formula provides an overview of the basic HMM part of speech tagging (Jurafsky et al. 2005, 329).

$$P(\text{word} \mid \text{tag}) * P(\text{tag} \mid \text{previous n tags}) \qquad (1)$$

Equation 1 represents a tag sequence for the whole sentence. According to 1, the tag of a word depends on the probability of a word tag pair multiply by the probability of the sequence of tags from the start of the sentence. Dependency of a tag on previous n tags is called N-gram model. In order to tag a single word, bigram HMM tagger has to use. The bigram model of tagging a word w_i with a tag t_i is given by the maximum probability of tag t_i with previous tag t_{i-1} and the probability of the word w_i having tag t_i i.e. (Jurafsky et al. 2005, 329).

$$t_i = \text{argmax } P(t_i \mid t_{i-1}) P(w_i \mid t_i) \quad (2)$$

Consider a sequence of words $W = w_1 w_2 \ldots w_n$, and a sequence of tags $T = t_1 t_2 \ldots t_n$. The maximum probable solution for a sequence of tags given that the sequence of words can be represented as follows:

$$\text{Max } P(t_1 t_2 \ldots t_n \mid w_1 w_2 \ldots w_n) \quad (3)$$

Taking $T = t_1 t_2 \ldots t_n$ and $W = w_1 w_2 \ldots w_n$, equation 3 becomes;

$$\text{Max } P(T \mid W) \quad (4)$$

According to Bayes theorem;

$$P(A|B) P(B) = P(B|A) P(A) \quad (5)$$

$$P(T|W) P(W) = P(W|T) P(T) \quad (6)$$

Here, P(W|T) can be expressed as the probability of the sequence of words W given that the tag sequence T. P(W) is the probability of the sequence of words which will remain constant for a sentence so neglecting P(W) for further calculations. P(T) is the probability of the tag sequence. P(T|W) can be expressed as the probability of the sequence of tags given that the sequence of observation symbols W.
The equation becomes;

$$P(T|W) = \max P(W|T) P(T) \quad (7)$$

$$P(t_1 t_2 \ldots t_n \mid w_1 w_2 \ldots w_n) = \max P(w_1 w_2 \ldots w_n \mid t_1 t_2 \ldots t_n) P(t_1 t_2 \ldots t_n) \quad (8)$$

Taking the simplifying assumption to reduce the complexity and dependency of the equation (Jurafsky et al. 2005, 332; Charniak et al. 1993);

1. Words are independent of each other
2. Words identity only depends on its own tag
3. A tag depends only on its previous tag

Applying the first assumption will reduce the sequence of words to one word i.e. the word of a tag depends on the maximum probability of the sequence of tags of the previous words plus its own tag.

$$\max P(w_i \mid t_1 t_2 \ldots t_n) P(t_1 t_2 \ldots t_n) \text{ where } i = 1 \ldots n \quad (9)$$

Applying the second assumption, that a words depend only on its own tag;

$$\max P(w_i \mid t_1\, t_2 \ldots t_n)\, P(t_1\, t_2 \ldots t_n) \text{ where } i = 1\ldots n \quad (10)$$

$$\max P(w_i \mid t_i)\, P(t_1\, t_2 \ldots t_n) \text{ where } i = 1\ldots n \quad (11)$$

Third assumption will change the dependency of a tag on the previous tag.

$$\max P(w_i \mid t_i)\, P(t_i \mid t_{i-1}) \text{ where } i = 2\ldots n \quad (12)$$

The dependency of a tag only on its previous tag is called the first order Hidden Markov model as shown in equation 11. In second order HMM, the current tag depend on two previous tags can be formulated as:

$$\max P(w_i \mid t_i)\, P(t_i \mid t_{i-1}\, t_{i-2}) \text{ where } i = 3\ldots n \quad (13)$$

For the current tagger, it was decided to limit the probability of tag sequence to bigram. Thus following formula will be implemented for part of speech tagger.

$$\text{Max } P(w_i \mid t_i)\, P(t_i \mid t_{i-1}) \text{ where } i = 2\ldots n \quad (14)$$

7.4.2 Unknown Word Problem

A training corpus of 100,000 words is used to train Hidden Markov Model. Length of the corpus is always finite. It is not possible to cover all words of the language. Also due to high inclusion of foreign words, new words are entering into the language day by day. These new words and the words which are not part of the corpus are known as Unknown word. Every tag of the word has some probability to be the tag of that word. This means, whenever an unknown word occurs, number of branches will exceed by the total number of tags. And if consecutive unknown words occur then the number of branches will exceed exponentially. The time to calculate these branches will also increased exponentially. The number of candidate tags for new word can be reduced if training corpus is covering all words of closed class. However, currently it was not that case. However, analysis was done on the training corpus and those closed class tags were removed from the list of candidate tags which were completely covered by the training corpus. A list of potential tags for a new word is given in the following table.

Table 6: Candidate tags for unknown words

NN	ADJ
ADV	CA
VB	OR
AA	U
TA	DATE
Q	

The probability of new words is handled by smoothing and reduction in search space is done by beam search. Next two sections will discuss them.

7.4.3 Smoothing

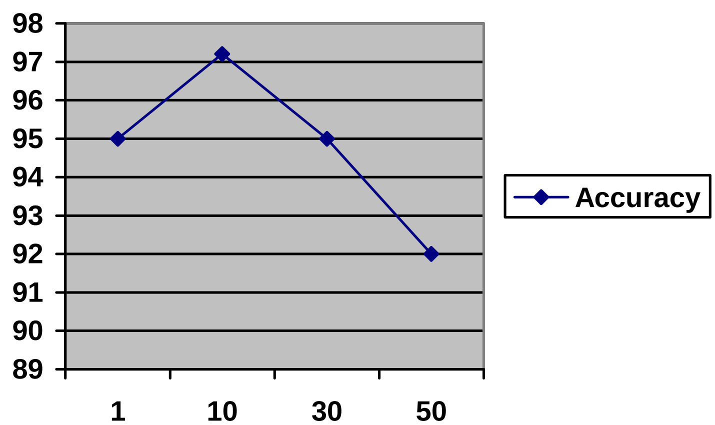
Due to the high productivity of language, there may occur words that have not seen before by the tagger. These unknown words will be assigned zero probability by the tagger. Thus makes the probability of whole sentence zero. Smoothing is used to assign these unknown words some probability other than zero. For part of speech tagger, Add Lambda smoothing was applied. A value of 0.5 was taken for lambda[5]. For unknown word, a new word tag pair was added in the list of word tag probabilities. For new tag sequence, a new tag tag pair was introduced in the list. Following algorithm was applied on each case.

- Add all counts to a variable say "All"
- Add unknown word pair with count equal to zero
- Add 0.5 to each count
- Add all counts after adding 0.5 say "All0.5"
- Multiply each count with the result of ("All" / "All0.5")

Smoothing was applied in the training database module. After calculating the frequency of each pair, smoothing was applied on each count. The probability of each pair was calculated with the help of smoothed counts.

7.4.4 Beam Search

Part of speech tagger process the input in the chunks of sentence. While working at sentence level, if an unknown word occurs, there will be 11 candidate tags for it. If a sentence is having five unknown words then the branches for these five words will be 11^5 i.e. exponential increase in branches. Processing so many branches may cause loss of memory and time. In order to control the number of branches, a threshold of 50, 30, 10 and 1 was selected. The accuracy of the tagger was observed on these thresholds. It was found that tagger shows relatively high accuracy at threshold of 10 i.e. number of branches should not increase 10. Whenever, number of branches exceeds 10, first ten branches with relatively high cumulative score were selected. Following graph is showing the rise and fall of accuracy curve over the change in threshold.



[5]The information about the value of lambda is taken from: http://www-rohan.sdsu.edu/~gawron/stat/discounting.htm

8 Results

Accuracy of tagger was checked over test corpus of 10,000 words. Test data was randomly selected from same domain. After applying normalization and by removing diacritics, test data was automatically tagged through tagger. Same test data was manually tagged in order to compare the accuracy of tagger. An application was build which takes automatically tagged test data and manually tagged data as input. In order to see the percentage of error over test corpus, tag of a word in test corpus was compared against the tag of manually tagged corpus. Tagger showed an accuracy of 97.2% i.e. an error rate of 2.8% over the test corpus of 10,000 words. Error rate over each tag was also calculated and analyzed to further improve the accuracy of tagger.

Results of the tagger are sorted over the accuracy rate of tags. In order to see the effect of each tag over the accuracy of tagger, total occurrences of each tag in test corpus are also calculated. All those tags that have an occurrence of below 10 are neglected from the analysis. Looking at the accuracies, tags can be divided into various clusters. The tags of accuracy 96% to 100% can be considered as satisfactory. The tags of accuracy between 84% and 94% can be considered as second cluster. It is interesting to see that most of the tags of demonstratives and pronouns lie in second cluster. Discussion on low accuracy rate of these categories can be found in next section. Last cluster contains two frequently occurring tags i.e. proper noun and KER tag. The high frequency and low accuracy rate of these tags significantly affect the results of the tagger. Following table is summarizing the results of the tagger.

Table 7: Results of tagger over test corpus of 10,000 words

Tag	Total occurrences in test corpus	Accuracy
FR	-	-
MUL	-	-
POT	-	-
NEG	-	-
SM	404	100
RP	3	100
GR	56	100
G	7	100
Q	82	100
CC	171	100
SE	190	100
WALA	50	100
INT	2	100
SC	188	100
CA	185	100
AD	112	100
AP	63	100
DATE	20	100
OR	32	100

KD	14	100
PRT	8	100
U	14	100
P	1978	99
I	96	99
TA	293	98
NN	2600	98
AA	379	97
ADV	131	97
REP	43	97
KER	72	97
RD	18	96
ADJ	487	96
VB	1008	96
PP	248	96
PD	112	92
PN	384	83
KP	7	80
QW	9	75
A	7	62
AKP	4	33

9 Analysis of Tagset on the Basis of Results

Manual annotation requires linguist to analyze corpus on the basis of phrase level analysis. Results of the tagger help in analyzing the practicality of tagset. Various points that may need a change in the tagset were observed in the process of manual annotation and in the analysis of the output. However, due to time limitation, only some changes were made in the tagset and other changes were left for future work. Following is the discussion on each issue.

9.1 Noun

While observing language, linguist finds problem in disambiguating the part of speech of a word as adjective or noun. Situation becomes worst when handling the words of complex predicates. It was observed that noun can be analyzed under these parameters:

- Nouns accept an adjective in their noun phrase other does not
- Noun can occur as complex predicates other not
- Nouns accept an adverb behind them other not
- Some nouns are derived from adjectives

These parameters were observed in the corpus and it was found that in the category of noun, there are different syntactic structure exist. However, due to time limitation, these were not properly observed.

9.2 Infinitive Verbs

In manual annotation, verbs acting as noun (infinitive verbs) are treated as verb. Analyzing syntactic structures of these words, it was observed that these words occur at the place of noun. Due to small training data, occurrence of unknown word is very frequent in test corpus. Whenever an unknown word occurs at the place of noun, the most probable tag for that word will be noun which is wrong in our case. The accuracy of KER tag is also affected by considering infinitives as verb. KER tag takes a verb behind it. The tagger needs to disambiguate KER tag with the کے word of semantic marker. Major distinction between KER tag and semantic marker can be made by considering the tag of one previous word. But infinitive verbs nullify this distinction. Consider following example:

Table 8: Comparison of KER tag and semantic marker

(a)	(b)	(c)
کام<NN> کرنے<VB> کے<P> بعد<NN>	کام<NN> کر<VB> کے<KER>	کام<NN> کرنے<NN> کے<P> بعد<NN>
Handling of infinitive in manual tagging	Syntactic structure of KER	Future work

There were 72 words of KER tag in the test corpus. Out of these 72 words, 3% words of kER tag were wrongly detected by the tagger. The accuracy of verb is also due to infinitive verbs. It was observed that accuracy of KER tag can be improved if infinitive verbs are handled separate from verb.

9.3 Noun vs. Other Tags

Tagger confuses the category of pre-title and post-title with nouns. Syntactically, the behavior of pre-title and post-title is same as that of noun. Difference was made on semantic grounds. For an unknown word, it is not possible for the tagger to get a higher probability of pre-title tag.

10 Analysis of Statistical approach on the Basis of Results

Statistical approaches to disambiguation require training data to model the language. The analysis on input data is based on the statistical technique and training data. While observing Urdu language and analyzing the results of the tagger, it was observed that statistical approach is finding problem in disambiguating between some particular pairs of tags. Following is the discussion on these categories.

10.1 Demonstratives vs. Pronouns

Demonstratives are divided into four types. All these types are ambiguous with the four types of pronoun. Difference between pronouns and demonstratives is based on phrase boundary analysis which is discussed in the section of tagset. Looking at tagger practically, it analyses the language in a flat structure. In flat structure, there is an equal probability of getting a noun after pronoun and demonstratives. Consider the following example:

Table 10: Examples of demonstratives and pronoun

وہ <PD> لوگ<NN> گانا<NN> گاہیں<VB> گے<TA>۔	وہ <PP> گانا<NN> گاہیں<VB> گے<TA>۔

In the above example of demonstrative, it is taking a noun inside its phrase and pronoun is not having any noun inside its phrase. But in flat structure, both demonstratives and pronouns are having noun after them thus confusing the tagger. This issue can be quoted as deficiency of statistical approach in handling phrase level ambiguities of Urdu language.

10.2 Noun vs. Proper noun

In the tagset, noun is divided into two categories i.e. noun and proper noun. Most of the distinction between nouns and proper nouns is based on semantics. However, there are structural differences as well. Nouns take pre-nominal elements i.e. adjectives, cardinal, ordinal, etc. behind them whereas proper nouns only take some pre-nominal elements in special cases. Consider the following example:

Table 11: Examples of nouns and proper nouns

پہلے<OR> دو<CA> آدمیوں<NN> کو<P> بلائو<VB>	پہلے<OR> دو<CA> حامد<PN> کو<P> بلائو<VB>

The example of proper noun taking pre-nominal elements is very rare in normal Urdu. However, probability of having Noun and proper noun at the start of a sentence is nearly equal. Due to these structural similarities, tagger confuses while handling unknown words as noun or proper noun.

11 Future Work

Part of speech tagger implemented above gives an accuracy of 97.2%. An obvious extension is to improve the accuracy up to 99%. An analysis of tagset on the basis of results is given in section 9. For future work, further analysis on the tagset can be done and implemented. Analysis of statistical technique is also given in section 11. A good future work is to analyze the implemented statistical technique and add heuristics to help the tagger in disambiguating the tags.

Words from the corpus of 100,000 words were used as lexicon for the tagger. For future work, larger lexicon can also be build which will significantly improve the accuracy of the tagger. Training data of 100,000 words was not sufficient to get a very high accuracy from the tagger. For future work, training data up to 1000,000 words can also be built and statistical technique can also be extended to bigram word probabilities.

12 Conclusion

Thesis was aimed at designing a syntactic tagset of Urdu and implementing a standard statistical approach to compare its results with other languages. In the thesis, Hidden Markov Model was implemented. Over the training corpus of 100,000 words, tagger showed an accuracy of 97.2%. By applying a standard statistical technique and achieving a relatively good accuracy are the answers to these questions. On the basis of the results, it can be concluded that standard statistical approach can be used for Urdu language. It was also observed that free orderness is not very frequent in writing. Thus does not significantly affect the accuracy of the tagger. It was also observed that tagger finds problems while disambiguating at phrase level. High accuracy can be achieved by merging the problematic categories of the tagset or by adding some heuristics which will help the tagger in disambiguating the tags.

Reference

Bahl, LR and Mercer, RL (1976) Part of speech assignment by a statistical decision algorithm. In: *IEEE International Symposium on Information Theory*, 88-89. Ronneby.

Beale, A. D. (1985). "A probabilistic approach to grammatical analysis of written English by computer." In Proceedings, *Second Conference of the European Chapter of the ACL,* Geneva, Switzerland, 159-165.

Brill, E.; Magerman, D.; Marcus, M.; and Santorini, B. (1990). "Deducing linguistic structure from the statistics of large corpora." In Proceedings, *DARPA Speech and Natural Language Workshop, Hidden Valley PA*. 275-282.

Brill, E (1995) Transformation-based error-driven learning and Natural Language Processing: a case study in part-of-speech tagging. In: *Computational Linguistics*, 21 (4): 543-565.

Brodda, Benny (1982). "Problems with tagging and a solution." *Nordic Journal of Linguistics,* 93-116.

Butt, M., et al (2001). Non.nominative subjects in Urdu: A computational analysis.

Butt, M. (2003). "Tense and aspect in Urdu", Konstanz University, Germany

Charniak, E, Hendrick, C, Jacobson, N and Perkowitz, M (1993) Equations for part of speech tagging. In: *Proceedings of the Eleventh National Conference on Artificial Intelligence*. Menlo Park: AAAI Press/MIT Press.

Chanod, J-P and Tapanainen, P (1995) Tagging French – comparing a statistical and a constraint-based method. In: *Proceedings of the Seventh Conference of the European Chapter of the ACL*. Dublin: Association for Computational Linguistics.

Chernokova, Sonia. (1989) "اردو افعال", Taraqqi Urdu Bureau, New Delhi.

Church, K (1988) A stochastic parts program and noun phrase parser for unrestricted text. In: *Proceedings of the second conference on Applied Natural Language Processig, ACL.*

Cutting, D, Kupiec, J, Pederson, J, and Sibun, P (1992) A practical part-of-speech tagger. In: *Proceedings of the third conference on Applied Natural Language Processing*, ACL.

Debili, Fathi (1977). "Traitements syntaciques utilisant des matrices de precedence frequentielles construites precedence frequentielles construites automatiquement par automatiquement par apprentissage." Doctoral dissertation Engineering Department, Universite Paris 7, France.

Daelemans, W (1999) Machine learning approaches. In: van Halteren (1999a).

Dandapat, S and Sarkar, S (2006) Part of speech tagging for Bengali with Hidden Markov Model, *department of computer science and engineering, IIT, Kharagpur, India.*

De Marcken, CG (1990) Parsing the LOB corpus. In: *Proceedings of the 1990 Conference of the Association for Computational Linguistics*, 243-251.

Derouault, Anne-Marie, and Merialdo, Bernard (1986). “Natural Language modeling for phoneme-to-text transcription.” In *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 8(6), 742-749.

DeRose, SJ (1988) Grammatical category disambiguation by statistical optimization. Im: *Computational Linguistics*, 14(1): 31-39.

E1-Beze, M and Merialdo, B (1999) Hidden Markov models. In: van Halteren (1999a)

Ellegård, A (1978) *The syntactic structure of English texts : a computer-based study of four kinds of texts in the Brown University Corpus*. Gothenburg Studies in English, 43. Gothenburg: Gothenburg University.

Garside, R (1987) The CLAWS word-tagging system. In: Garside, Leech and Sampson (1987).

Garside, R and Smith, N (1997) A hybrid grammatical tagger: CLAWS4. In: Garside, Leech and McEnery (1997).

Garside, R., and Leech, F. (1985). “A probabilistic parser.” In Proceedings, *Second Conference of the European Chapter of the ACL,* Geneva, Switzerland, 166-170.

Greenbaum, S and Yibin, N (1996) About the ICE Tagset. In: Greenbaum (1996).

Haq, M. Abdul. (1987) “اردو صرف و نحو”, Amjuman-e-Taraqqi Urdu (Hind).

Hardie, A (2003) The computational analysis of morphosyntactic categories in Urdu. PhD thesis, Lancaster University.

van Halteren, H, (2005) “Syntactic word class tagging”.

van Halteren, H and Oostdijk, N (1993) The TOSCA analysis system. In: Aarts, J, de Haan, P and Oostdijk, N (1993) *English language corpora: design, analysis and exploitation. Papers form the thirteenth International Conference on English Language Research on Computerised Corpora, Nijmegen 1992.* Amsterdam: Rodopi.

van Halteren, H and Voutilainen, A (1999) Automatic taggers: an introduction. In: van Halteren (1999a).

Harris, ZS (1962) *String analysis of sentence structure*. The Hague: Mouton.

Heikkilä, J (1995) A TWOL-based lexicon and feature system for English. In: Karlsson et al. (1995).

Hindle, D. Acquiring disambiguation rules from text. *Proceedings of the 27th Annual Meeting of the Association for Computation Linguistics*, 1989.

Ijaz, M., Hussain, S. (2007). " Corpus based Urdu lexicon development", In the proceedings of Conference of Language Technology 2007

Jurish, Bryan,. Part of speech tagging with finite state morphology.

Javed, Ismat. "نئ اردو قواعد", 1981. Taraqqi Urdu Bureau, New Delhi.

Jelinek, F (1985) Markov source modeling of text generation. In: Skwirzinski, JK (ed.) (1983) I*mpact of Processing Techniques on CommunicationL Proceedings of the NATO Advanced Study Institute 1983.* Dordrecht: Nijhoff.

Jurafsky, D and Martin H. James, (2000), Speech and Language Processing, Prentice Hall.

Karlsson, F (1995) The formalism and environment of Constraint Grammar Parsing. In: Karlsson et al. (1995).

Karlsson, F, Voutilainen, A, Heikkilä, J and Anttila, A (eds.) (1995) *Constraint Grammar: a language-independent system for parsing unrestricted text*. Berlin: Mouton de Gruyter.

Kempe, A. (1993). A stochastic Tagger and an Analysis of Tagging Errors. *Internal paper. Institute for Computational Linguistic,* University of Stuttgart.

Khoja, S., Garside, R., Knowles, G., " A tagset for morphosyntactic tagging of Arabic", Lancaster University.

Klein, S and Simmons, RF (1963) A computational approach to grammatical coding of English words. In: *Journal of the Association for Computing Machinery*, 10: 334-347.

Lager, Torbjorn and Nivre, Joakim,. Part of speech tagging from a logical point of view.

Leech, G (1997) Introducing corpus annotation. In: Garside, Leech and McEnery (1997).

Leech, G and Wilson, A (1999) Standards for tagsets. In: van Halteren (1999a). (Edited version of *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora* (1996): available on the internet at http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html .)

Marshall, I. (1987) Tag selection using probabilistic methods. In: Garside, Leech and Sampson (1987).

Marcus, M, Santorini, B and Marcinkiewicz, MA (1993) Building a large annotated corpus of English: the Penn Treebank. In: *Computational Linguistics*, 19(2): 313-330.

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

Paulussen, H., and Martin, W. (1992). "A lemmatizer-tagger for medical abstracts." In Proceedings, *Third Conference on Applied Language Processing,* Trento, Italy, 141-146.

Platts, John T (1909). "A Grammar of the Hindustani or Urdu Language", London.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE,* 77 (2), 257-286

Sánchez León, F and Nieto Serrano, AF (1997) Retargeting a tagger. In: Garside, Leech and McEnery (1997).

Sampson, G (1987) Alternative grammatical coding systems. In: Garside, Leech and Sampson (1987).

Scott M. Thede and Mary P. Harper. "A second order Hidden Markov model for part of speech tagging".

Schmid, H. (1994) Part-of-speech tagging with neural networks. In *Proceedings of COLING-94, Kyoto*.

Siddiqi, Dr. Abu-ul-Lais. (1971) "جامع القواعد", Markazi Urdu Board, Lahore.

Stolz, Ws, Tannenbaum, PH and Carstensen, FV (1965) A stochastic approach to the grammatical coding of English. In: *Communications of ACM,* 8 (6): 399-405.

Tapanainen, P and Voutilainen, A (1994) Tagging accurately – don't guess if you know. In: *Proceedings of the Fourth Conference on Applied Natural Language Processing*. Stuttgart.

Taylor, A., et al. The PENN TREEBANK: an overview.

Voutilainen, A (1995) Morphological disambiguation. In: Karlsson et al. (1995).

Weischedel, R, Meteer, M, Schwartz, R, Ramshaw, L and Palmuzzi, J (1993) Coping with ambiguity and unknown words through probabilistic models. In: *Computational Linguistics*, 19 (2): 359-382.

Appendix

Parts of Speech Proposed by Platts

Table 12: Analysis of Platts (Platts 1909)

Main category	**Sub category**	**Example**
Noun	Substantive noun	All common and proper noun e.g. ماں، احسان، لڑکا
	Adjective	اچھا، عمد ہ، بہادر
	Numeral adjective	ایک، سترہ، پانچواں، دونوں، سیکڑوں، دوکنا، دو دو، ڈھائ، ایک بار
	Personal pronoun	میں، تو، مجہ، تجہ، میرا، همارا
	Demonstrative pronoun	یہ، وہ، اس، انھوں، اس سے، ان میں
	Relative pronoun	جو، جس
	Correlative pronoun	جو کرے گا سو بھرے گا۔
	Interrogative pronoun	کون، کس، (who, what, which)
	Indefinite pronoun	کو□□ئی، کسی، جو کوئی، کوئی، یور
	Reflexive pronoun	اپنا، اپ سے، اپ کو
	Reciprocal pronoun	ایک دوسرا
	Possessive pronoun	Genitive case of personal pronoun e.g. میرا، اس کا، اپنا
	Pronominal adjective	دونوں، بہت، بعض، سب، فلاں
Verb	Conjunctive participle	Platts dd not propose any type of verb under its subcategory. However, all the properties and forms of verb are discussed as its features.
Particle	Adverb	نہیں، کر تو سہی، میں کہاں تو کہاں
	Postposition	اگے، سااته، طرف، نزدیک
	Conjunction	اور، نہ نفع ہو نہ نقصان، یا، کہ، مگر، لیکن،ورنہ، پر، اس لیے، لہزا
	Interjection	واہ، کاش

Parts of Speech Proposed by Siddiqi

Table 13: Analysis of Siddiqui (Siddiqui 1971)

Main category	Sub category	Example
Noun	With respect to structure	--
	With respect to nature	--
	Sound	میائوں، بھن بھن، گڑگڑاہٹ
	Indefinite	فلانا، ایسا، تیسا
	Relative	جو
	Interrogative	کیا، کون، کون سی، کب، کیسا، ادھر، کتنا، کیسے
	Numerals	ایک، دو، پاؤ، بعض، کچھ، بہت
Adjective	Personal	لمبا، خوبصورت، بیمار
	نسبتی	فارسی، مدنی، مردانہ، عیسوی
	Numeral	پہلا، دوسرا
Pronoun	Demonstrative	یہ، وہ
	Personal	میں، ہم، تم، أپ
	Relative	احمد کے پاس ہے وہ میری ہےجووہ کتاب
	Interrogative	کیا، کون
	Indefinite	کوئی، کچھ
	Reflexive	خود، أپ (For emphasis)
Verb	Intransitive	ہےبیٹھا رہا ہے، اسلم کھااحمد آم
	Transitive	دیاحمد نے اسلم کو کتاب
	Predicative	ہو، ہوں، ہے، تھا
Distinct		اب، تب، ادھر، تو، بلکل، یکایک، ایک بار، دو بار، کتنا، جی ہاں، نہیں تو، کبھی نہ کبھی، أگے أگے
Particle	Construction	نے، کو، سے، میں، تک، پر، کا، کے، کی
	Conjunction	--
	تخصیص	ہی، تو، بھی، تنہا، تو، صرف، اکیلا، کبھی، کہاں، یوں
	فجاءیہ	واہ، کاش، خدا کرے، سبحان اللہ

Table 14: Subcategories proposed by Siddiqui (Siddiqui 1971)

Main category	Sub category		Example
Noun with respect to structure	Original		اونٹ، تلوار، قلم
	Verbal		اٹھنا، بیٹھنا، جاگنا، سونا
	Morphed		پڑھتا تھا، پڑھنا سے سرخیسرخ سے
Noun with respect to nature	Substantive		قلم، کاغذ، لاہور، پاکستان
	Adjective	Comparative	کم، کم تر، خوب، خوب تر، بہتر، اکبر
		Exaggeration	بڑا۔ بہت، خوب، نہایت، نہایت ہی
	Pronoun		یہ، وہ
Personal pronoun	Courteous		تم، أپ، وہ
	Possessive		میرا، تمھارا
Conjunction particle	شرط		جب، جو، اگر، جو، جس وقت، جوں جوں، کیوں، ورنہ
	استثنا		لیکن، مگر، کبھی، سوا
	استدراک		میں نہیں مانا، بلکہ، البتہ، مگر مگرلیکن،اس نے بہت کہا

		ہاں
	تردید	نہ نہ، خواہ، چاہو، کہ، یا، یا تو
	وصل	وقار أیا پھر یا، و، احمد أیا اور,
	بیانیہ	کہ
	علت	اس لیے، اس واسطے، تاکہ، لہزا، کہ

Parts of Speech Proposed by Javaid

Table 15: Analysis of Javaid (Javaid 1981)

Main category	Sub category	Example
Noun	Common	کتا، بلی، قلم، کاغذ
	Proper	لاہور، پاکستان
	Collective	فوج، جھنڈ، ریوڑ
	Abstract	وقت، فاصلہ، جذبہ
	Un-count	پانی، چاندی
Adjective	Personal	لمبا، خوبصورت، بیمار
	Numeral	ایک، دو، پاؤ، بعض، کچھ، بہت
	Quantitative	تھوڑا پانی، دودھ، کچھ
	Emphatic	کافی، بہت، بڑا شریر (To show intensity)
	Pronoun	یہ، وہ
Verb	Predicative	ہو، ہوں، ہے، تھا
	Intransitive	احمد آم کھا رہا ہے، اسلم بیٹھا ہے
	Transitive	احمد نے اسلم کو کتاب دی
	Verbal	جانا، کھانا، شرمانا، مسکرانا
	حالیہ	ڈرتے ڈرتے، آتے آتے، بہتا ہوا پانی
	حالیہ معطوف	سو کر اٹھا، میں کام کر کر کے تھک گیا
	Adverb	یں، آس پاس، اردگرد، گھر م تر ، زیادہ سے، تیزی سے، غلطی سے، آگے، پرسوں سے
Pronoun	Demonstrative	یہ، وہ
	Personal	میں، ہم، تم، أپ
	Relative	جو کتاب احمد کے پاس ہے وہ میری ہے
	Interrogative	کیا، کون
	Courteous	تم، أپ، وہ
	Possessive	میرا، تمھارا، أپ کا، تیرا
	Reflexive	خود، أپ (For emphasis)
	Common	کوئی، کچھ، فلاں، بعض، تو کرو
	اضافی مشترکہ	تم اپنا کام کرو، وہ اس کا کام کرے
	Adverb	کب، ادھر، یوں، ایسے، کیسے
جار		نے، کو، سے، میں، تک، پر، کا، کے، کی، کے پاس، کے پہلے، کی کے ساتھ، کے لیے، کی وجہ سے
عطف		جو، جہاں، حالانکہ، تاوقتیکہ ، بھی، پھربھی، اس لیے، یایا
	وصل	کیا کیا، یا,اور
	تردید	نہ نہ، خواہ
	استثنا	لیکن، مگر
	ترقی	بلکہ، پھر بھی، تا ہم
	علت	اس لیے، اس واسطے، تاکہ، لہزا
فجاءیہ		واہ، کاش، خدا کرے، سبحان اللہ

نداءیہ		اے، ارے، او
تاکید		ہی، تو، بھی، سہی، ہر گز
اثبات و نفی		ہاں، نہیں، جی ہاں

Parts of Speech Proposed by Haq

Table 16: Analysis of HAQ

Main category	Sub category	Example
Noun	Proper	لاہور، پاکستان
	Common	کتا، بلی، قلم، کاغذ
Pronoun	Personal	میں، ہم، تم، أپ، اس کا
	Relative	احمد کے پاس ہے وہ میری ہے، جنہوں نے، جن سےجووہ کتاب
	Interrogative	کیا، کون
	Indefinite	کوئی، کچھ
	Demonstrative	یہ، وہ
Adjective	Personal	لمبا، خوبصورت، بیمار، ٹھوس، ہلکا
	Numeral	ایک، دو، پاؤ، بعض، کچھ، بہت، دوگنا، اتنا، سب، کئی، پون
	Quantitative	چار سیر، پانچ گز
	نسبتی	فارسی، مدنی، مردانہ، عیسوی
	Pronoun	وہ، یہ، کون، جو، کیا
Verb	Predicative	ہونا، دینا، دکھائی دینا
	Intransitive	احمد أیا
	Transitive	احمد نے اسلم کو کتاب دی
	معطوف	اٹھا، وہ خبر سنا کر چلا گیا کرسو
	Adverb	اب، کب، أج، اچانک، یکا یک، ہمیشہ، یہاں، باہر، یوں، سچ مچ، جتنا، کتنا
ربط		نے، کو، سے، میں، تک، پر، کا، کے، کی، پیچھے، باہر، درمیان، طرح
عطف	وصل	اور، و
	تردید	نہیںکہیایا، خواہ خواہ، نہ نہ، أتے ہو
	استدراک	میں نہیں مانا، بلکہمگرلیکن،اس نے بہت کہا
	استثنا	وہ نہیں أیامگرسب أئے
	شرط	جب، جو، اگر
	علت	کیونکہ، اس لیے، لہزا
	بیانیہ	کہ
تخصیص		ہی، تو، بھی
فجاءیہ		واہ، کاش، خدا کرے، سبحان اللہ

Parts of Speech Proposed by Schmidt

Table 17: Analysis of Schmidt (Schmidt1999)

Main category	Sub category	Example
Noun		لڑکا، گھر، کنواں، لڑکپن
Pronoun	Demonstrative	کو گھوڑے کہتے ان کا نام کیا ہے،ہم اس گھڑا ہے، یہ ایک لڑکا ہے، وہ تھے
	Personal	میں، تو، وہ پاس رہتا ہے ، یہ علی کے پاس جانا چاہتا ہے، تم، مجھ، اس میں کوئی شک نہیں، ان، ہم
	Reflexive	اپنا، خود، آپس میں، خود بخود
	Interrogative	کیا، کون، کس، کنہوں نے

	Indefinite	کسی، کوئی، کچھ یار کرو،
	Relative	جو، کون کون، کوئی نہ کوئی، کچھ کچھ، کچھ نہ کچھ
Adjective	--	اچھا، دلچسپ، معلوم ہونا، مبتلا ہونا، ایسا، کیسا، ویسا، سا، سے، والا
Adverb	Time	ہمیشہ، کل، اکثر، اب، تب، کب، کس وقت، جس وقت
	Place	وہاں، ادھر، اس جگہ، اس طرف، اندر، باہر، قریب، دور
	Manner	کرو ایسایوں، اس طرح، کیوں،
	Degree	ذہینبڑابہت، زیادہ،
	Modal	نہیں، نہ، مت، شاید، ضرور، پھر، صرف
Postposition	Grammatical	تار بھیہجا کوکا، کے، کی، نے، والدہ
	Spatial-temporal	سے، تک، میں، پر
	Compound	کی وجہ سے، کے ساتھ، کے بعد، کے نیچے
Verb	Root	جا، کر، دے، سن
	Imperfective participle	آتا، جاتا،
	Perfective participle	آیا، گیا، سنا
	Infinitive	جانا، کرنا، دینا، سننا
Particle	Contrastive emphatic	پڑھے گا، نہیں توتووہ اردو
	Exclusive emphatic	ہی
	Inclusive emphatic	بھی
	Adjectival	سا، سی، سے
Interjection	Vocative	او ، ارے
	Free	واہ، ہائے، اوہو
Conjunction	Coordinating	اور، یا، مگر، لیکن، بلکہ، جب سے
	Correlative	(بھی بھی)، یا یا، نہ نہبھی جاؤں گا اور تم بھیمیں
	Causal	کیونکہ، چونکہ
	Concessive	اگرچہ، حالانکہ
	Subordinating	اگر، تاکہ، بشرطیکہ، کہ
Number	Cardinal	ایک، سترہ، لاکھ
	Ordinal	پہلا، دوسرا، اکیسویں
	Fraction	پون، سوا، چوتھائی، بٹا، تین بار، دوگنا، دفعہ، مرتبہ

Urdu Tagset Proposed by Hardie

Tag	Example	Description
AL	ال	Arabic definite article
AU	واہ	Interjection
CC	مگر	Coordinating conjunction
CCC	یا	Correlative coordinating conjunction
CS	کہ	Subordinating conjunction
FF		Foreign word
FX		Non-Perso-Arabic string
FO		Formula (e.g. mathematical)
FZ		Letter of the alphabet
FS		Other symbol
FA		Acronym
FB		Abbreviation
FU		Other unclassifiable non-Urdu element
IB	از، فی	Preposition
II	پر، میں	Unmarked postposition
IIC	ے، یں، ہیں	Clitic postposition *ē*, *ē~*, *hē~*
IIM1N	کا	Marked masculine singular nominative postposition *kā*
IIM1O	کے	Marked masculine singular oblique postposition *kē*
IIM2N	کے	Marked masculine plural nominative postposition *kē*
IIM2O	کے	Marked masculine plural oblique postposition *kē*
IIF1N	کی	Marked feminine singular nominative postposition *kī*
IIF1O	کی	Marked feminine singular oblique postposition *kī*
IIF2N	کی	Marked feminine plural nominative postposition *kī*

IIF2O	کی	Marked feminine plural oblique postposition *kī*
IV	کے	Verbal postposition *kē*
JJM1N	بڑا	Marked masculine singular nominative adjective
JJM1O	بڑے	Marked masculine singular oblique adjective
JJM2N	بڑے	Marked masculine plural nominative adjective
JJM2O	بڑے	Marked masculine plural oblique adjective
JJF1N	بڑی	Marked feminine singular nominative adjective
JJF1O	بڑی	Marked feminine singular oblique adjective
JJF2N	بڑی	Marked feminine plural nominative adjective
JJF2O	بڑی	Marked feminine plural oblique adjective
JJU	خطرناک	Unmarked adjective
JD	زیادہ ، کافی	Indefinite determiner
JDNU	ایک ، اٹھارہ	Cardinal number
JDNUO	دونوں	Oblique cardinal number
JDNUC	چو(گنا)	Pre-multiplicative clitic cardinal number *du–*, *ti–*, *cau–*
JDNM1N	تسرا	Masculine singular nominative ordinal number
JDNM1O	تسرے	Masculine singular oblique ordinal number
JDNM2N	تسرے	Masculine plural nominative ordinal number
JDNM2O	تسرے	Masculine plural oblique ordinal number
JDNF1N	تسری	Feminine singular nominative ordinal number
JDNF1O	تسری	Feminine singular oblique ordinal number
JDNF2N	تسری	Feminine plural nominative ordinal number
JDNF2O	تسری	Feminine plural oblique ordinal number
JDFU	سوا	Unmarked fraction
JDFM1N	پونا	Masculine singular nominative fraction
JDFM1O	پونے	Masculine singular oblique fraction

JDFM2N	پونے	Masculine plural nominative fraction
JDFM2O	پونے	Masculine plural oblique fraction
JDFF1N	پونی	Feminine singular nominative fraction
JDFF1O	پونی	Feminine singular oblique fraction
JDFF2N	پونی	Feminine plural nominative fraction
JDFF2O	پونی	Feminine plural oblique fraction
JDYM1N	اِتنا ، ایسا	Masculine singular nominative proximal demonstrative adjective (*itnā, aisā*)
JDYM1O	اِتنے ، ایسے	Masculine singular oblique proximal demonstrative adjective (*itnē, aisē*)
JDYM2N	اِتنے ، ایسے	Masculine plural nominative proximal demonstrative adjective (*itnē, aisē*)
JDYM2O	اِتنے ، ایسے	Masculine plural oblique proximal demonstrative adjective (*itnē, aisē*)
JDYF1N	اِتنی ، ایسی	Feminine singular nominative proximal demonstrative adjective (*itnī, aisī*)
JDYF1O	اِتنی ، ایسی	Feminine singular oblique proximal demonstrative adjective (*itnī, aisī*)
JDYF2N	اِتنی ، ایسی	Feminine plural nominative proximal demonstrative adjective (*itnī, aisī*)
JDYF2O	اِتنی ، ایسی	Feminine plural oblique proximal demonstrative adjective (*itnī, aisī*)
JDVM1N	اُتنا ، ویسا	Masculine singular nominative distal demonstrative adjective (*utnā, vaisā*)
JDVM1O	اُتنے ، ویسے	Masculine singular oblique distal demonstrative adjective (*utnē, vaisē*)
JDVM2N	اُتنے ، ویسے	Masculine plural nominative distal demonstrative adjective (*utnē, vaisē*)
JDVM2O	اُتنے ، ویسے	Masculine plural oblique distal demonstrative adjective (*utnē, vaisē*)
JDVF1N	اُتنی ، ویسی	Feminine singular nominative distal demonstrative adjective (*utnī, vaisī*)
JDVF1O	اُتنی ، ویسی	Feminine singular oblique distal demonstrative adjective (*utnī, vaisī*)
JDVF2N	اُتنی ، ویسی	Feminine plural nominative distal demonstrative adjective (*utnī, vaisī*)
JDVF2O	اُتنی ، ویسی	Feminine plural oblique distal demonstrative adjective (*utnī, vaisī*)
JDKM1N	کتنا ، کیسا	Masculine singular nominative interrogative adjective (*kitnā, kaisā*)
JDKM1O	کتنے ، کیسے	Masculine singular oblique interrogative adjective (*kitnē, kaisē*)
JDKM2N	کتنے ، کیسے	Masculine plural nominative interrogative adjective (*kitnē, kaisē*)

JDKM2O	کتنے ، کیسے	Masculine plural oblique interrogative adjective (*kitnē, kaisē*)
JDKF1N	کتنی ، کیسی	Feminine singular nominative interrogative adjective (*kitnī, kaisī*)
JDKF1O	کتنی ، کیسی	Feminine singular oblique interrogative adjective (*kitnī, kaisī*)
JDKF2N	کتنی ، کیسی	Feminine plural nominative interrogative adjective (*kitnī, kaisī*)
JDKF2O	کتنی ، کیسی	Feminine plural oblique interrogative adjective (*kitnī, kaisī*)
JDJM1N	جتنا ، جیسا	Masculine singular nominative relative adjective (*jitnā, jaisā*)
JDJM1O	جتنے ، جیسے	Masculine singular oblique relative adjective (*jitnē, jaisē*)
JDJM2N	جتنے ، جیسے	Masculine plural nominative relative adjective (*jitnē, jaisē*)
JDJM2O	جتنے ، جیسے	Masculine plural oblique relative adjective (*jitnē, jaisē*)
JDJF1N	جتنی ، جیسی	Feminine singular nominative relative adjective (*jitnī, jaisī*)
JDJF1O	جتنی ، جیسی	Feminine singular oblique relative adjective (*jitnī, jaisī*)
JDJF2N	جتنی ، جیسی	Feminine plural nominative relative adjective (*jitnī, jaisī*)
JDJF2O	جتنی ، جیسی	Feminine plural oblique relative adjective (*jitnī, jaisī*)
JXGM1N	گنا	Masculine singular nominative multiplicative marker *gunā*
JXGM1O	گنے	Masculine singular oblique multiplicative marker *gunē*
JXGM2N	گنے	Masculine plural nominative multiplicative marker *gunē*
JXGM2O	گنے	Masculine plural oblique multiplicative marker *gunē*
JXGF1N	گنی	Feminine singular nominative multiplicative marker *gunī*
JXGF1O	گنی	Feminine singular oblique multiplicative marker *gunī*
JXGF2N	گنی	Feminine plural nominative multiplicative marker *gunī*
JXGF2O	گنی	Feminine plural oblique multiplicative marker *gunī*
JXSM1N	سا	Masculine singular nominative adjectival particle *sā*
JXSM1O	سے	Masculine singular oblique adjectival particle *sē*
JXSM2N	سے	Masculine plural nominative adjectival particle *sē*

JXSM2O	سے	Masculine plural oblique adjectival particle *sē*
JXSF1N	سی	Feminine singular nominative adjectival particle *sī*
JXSF1O	سی	Feminine singular oblique adjectival particle *sī*
JXSF2N	سی	Feminine plural nominative adjectival particle *sī*
JXSF2O	سی	Feminine plural oblique adjectival particle *sī*
JXVM1N	والا	Masculine singular nominative adjectival / occupational particle *vālā*
JXVM1O	والے	Masculine singular oblique adjectival / occupational particle *vālē*
JXVM2N	والے	Masculine plural nominative adjectival / occupational particle *vālē*
JXVM2O	والے	Masculine plural oblique adjectival / occupational particle *vālē*
JXVF1N	والی	Feminine singular nominative adjectival / occupational particle *vālī*
JXVF1O	والی	Feminine singular oblique adjectival / occupational particle *vālī*
JXVF2N	والی	Feminine plural nominative adjectival / occupational particle *vālī*
JXVF2O	والی	Feminine plural oblique adjectival / occupational particle *vālī*
LL	زمہ	Nongrammatical lexical element
NNMM1N	لڑکا	Common marked masculine singular nominative noun
NNMM1O	لڑکے	Common marked masculine singular oblique noun
NNMM1V	لڑکے	Common marked masculine singular vocative noun
NNMM2N	لڑکے	Common marked masculine plural nominative noun
NNMM2O	لڑکوں	Common marked masculine plural oblique noun
NNMM2V	لڑکو	Common marked masculine plural vocative noun
NNMF1N	چڑیا	Common marked feminine singular nominative noun
NNMF1O	چڑیا	Common marked feminine singular oblique noun
NNMF1V	چڑیا	Common marked feminine singular vocative noun
NNMF2N	چڑیاں	Common marked feminine plural nominative noun
NNMF2O	چڑیوں	Common marked feminine plural oblique

		noun
NNMF2V	چڑیو	Common marked feminine plural vocative noun
NNUM1N	بھائی	Common unmarked masculine singular nominative noun
NNUM1O	بھائی	Common unmarked masculine singular oblique noun
NNUM1V	بھائی	Common unmarked masculine singular vocative noun
NNUM2N	بھائی	Common unmarked masculine plural nominative noun
NNUM2O	بھائیوں	Common unmarked masculine plural oblique noun
NNUM2V	بھائیو	Common unmarked masculine plural vocative noun
NNUF1N	بہن	Common unmarked feminine singular nominative noun
NNUF1O	بہن	Common unmarked feminine singular oblique noun
NNUF1V	بہن	Common unmarked feminine singular vocative noun
NNUF2N	بہنیں	Common unmarked feminine plural nominative noun
NNUF2O	بہنوں	Common unmarked feminine plural oblique noun
NNUF2V	بہنو	Common unmarked feminine plural vocative noun
NPMM1N		Proper marked masculine singular nominative noun
NPMM1O		Proper marked masculine singular oblique noun
NPMM1V		Proper marked masculine singular vocative noun
NPMM2N		Proper marked masculine plural nominative noun
NPMM2O		Proper marked masculine plural oblique noun
NPMM2V		Proper marked masculine plural vocative noun
NPMF1N		Proper marked feminine singular nominative noun
NPMF1O		Proper marked feminine singular oblique noun
NPMF1V		Proper marked feminine singular vocative noun
NPMF2N		Proper marked feminine plural nominative noun
NPMF2O		Proper marked feminine plural oblique noun

NPMF2V		Proper marked feminine plural vocative noun
NPUM1N		Proper unmarked masculine singular nominative noun
NPUM1O		Proper unmarked masculine singular oblique noun
NPUM1V		Proper unmarked masculine singular vocative noun
NPUM2N		Proper unmarked masculine plural nominative noun
NPUM2O		Proper unmarked masculine plural oblique noun
NPUM2V		Proper unmarked masculine plural vocative noun
NPUF1N		Proper unmarked feminine singular nominative noun
NPUF1O		Proper unmarked feminine singular oblique noun
NPUF1V		Proper unmarked feminine singular vocative noun
NPUF2N		Proper unmarked feminine plural nominative noun
NPUF2O		Proper unmarked feminine plural oblique noun
NPUF2V		Proper unmarked feminine plural vocative noun
OO	و	Persian compound-forming conjunction *ō*
PPM1N	میں	First person singular nominative personal pronoun (*mai~*)
PPM1O	مجھ	First person singular oblique personal pronoun (*mujh*)
PPM2N	ہم	First person plural nominative personal pronoun (*ham*)
PPM2O	ہم	First person plural oblique personal pronoun (*ham*)
PPT1N	تو	Second person singular nominative personal pronoun (*tū*)
PPT1O	تجھ	Second person singular oblique personal pronoun (*tujh*)
PPT2N	تم	Second person plural nominative personal pronoun (*tum*)
PPT2O	تم	Second person plural oblique personal pronoun (*tum*)
PGM1M1N	میرا	First person singular masculine singular nominative possessive adjective (*mērā*)
PGM1M1O	میرے	First person singular masculine singular oblique possessive adjective (*mērē*)
PGM1M2N	میرے	First person singular masculine plural

		nominative possessive adjective (*mērē*)
PGM1M2O	میرے	First person singular masculine plural oblique possessive adjective (*mērē*)
PGM1F1N	میری	First person singular feminine singular nominative possessive adjective (*mērī*)
PGM1F1O	میری	First person singular feminine singular oblique possessive adjective (*mērī*)
PGM1F2N	میری	First person singular feminine plural nominative possessive adjective (*mērī*)
PGM1F2O	میری	First person singular feminine plural oblique possessive adjective (*mērī*)
PGM2M1N	ہمارا	First person plural masculine singular nominative possessive adjective (*hamārā*)
PGM2M1O	ہمارے	First person singular masculine singular oblique possessive adjective (*hamārē*)
PGM2M2N	ہمارے	First person singular masculine plural nominative possessive adjective (*hamārē*)
PGM2M2O	ہمارے	First person singular masculine plural oblique possessive adjective (*hamārē*)
PGM2F1N	ہماری	First person singular feminine singular nominative possessive adjective (*hamārī*)
PGM2F1O	ہماری	First person singular feminine singular oblique possessive adjective (*hamārī*)
PGM2F2N	ہماری	First person singular feminine plural nominative possessive adjective (*hamārī*)
PGM2F2O	ہماری	First person singular feminine plural oblique possessive adjective (*hamārī*)
PGT1M1N	تیرا	Second person singular masculine singular nominative possessive adjective (*tērā*)
PGT1M1O	تیرے	Second person singular masculine singular oblique possessive adjective (*tērē*)
PGT1M2N	تیرے	Second person singular masculine plural nominative possessive adjective (*tērē*)
PGT1M2O	تیرے	Second person singular masculine plural oblique possessive adjective (*tērē*)
PGT1F1N	تیری	Second person singular feminine singular nominative possessive adjective (*tērī*)
PGT1F1O	تیری	Second person singular feminine singular oblique possessive adjective (*tērī*)
PGT1F2N	تیری	Second person singular feminine plural nominative possessive adjective (*tērī*)
PGT1F2O	تیری	Second person singular feminine plural oblique possessive adjective (*tērī*)
PGT2M1N	تمہارا	Second person plural masculine singular nominative possessive adjective (*tumhārā*)

PGT2M1O	تمہارے	Second person singular masculine singular oblique possessive adjective (*tumhārē*)
PGT2M2N	تمہارے	Second person singular masculine plural nominative possessive adjective (*tumhārē*)
PGT2M2O	تمہارے	Second person singular masculine plural oblique possessive adjective (*tumhārē*)
PGT2F1N	تمہاری	Second person singular feminine singular nominative possessive adjective (*tumhārī*)
PGT2F1O	تمہاری	Second person singular feminine singular oblique possessive adjective (*tumhārī*)
PGT2F2N	تمہاری	Second person singular feminine plural nominative possessive adjective (*tumhārī*)
PGT2F2O	تمہاری	Second person singular feminine plural oblique possessive adjective (*tumhārī*)
PY1N	یہ	Singular nominative proximal demonstrative pronoun (*yah*)
PY1O	اِس	Singular oblique proximal demonstrative pronoun (*is*)
PY2N	یہ	Plural nominative proximal demonstrative pronoun (*yah*)
PY2O	اِن	Plural oblique proximal demonstrative pronoun (*in*)
PY2E	اِنہوں	Plural oblique proximal demonstrative pronoun before *nē* (*inhō~*)
PV1N	وہ	Singular nominative distal demonstrative pronoun (*vah*)
PV1O	اُس	Singular oblique distal demonstrative pronoun (*us*)
PV2N	وہ	Plural nominative distal demonstrative pronoun (*vah*)
PV2O	اُن	Plural oblique distal demonstrative pronoun (*un*)
PV2E	اُنہوں	Plural oblique distal demonstrative pronoun before *nē* (*unhō~*)
PK1N	کیا، کون	Singular nominative interrogative pronoun (*kyā, kaun*)
PK1O	کس	Singular oblique interrogative pronoun (*kis*)
PK2N	کیا، کون	Plural nominative interrogative pronoun (*kyā, kaun*)
PK2O	کن	Plural oblique interrogative pronoun (*kin*)
PK2E	کنہوں	Plural oblique interrogative pronoun before *nē* (*kinhō~*)

PJ1N	جو	Singular nominative relative pronoun (*jō*)
PJ1O	جس	Singular oblique relative pronoun (*jis*)
PJ2N	جو	Plural nominative relative pronoun (*jō*)
PJ2O	جن	Plural oblique relative pronoun (*jin*)
PJ2E	جنہوں	Plural oblique relative pronoun before *nē* (*jinhō~*)
PRF	خود, آپ	Reflexive pronoun (*āp, xud*)
PRC	آپس	Reciprocal pronoun (*āpas*)
PGRM1N	اپنا	Masculine singular nominative reflexive possessive adjective (*apnā*)
PGRM1O	اپنے	Masculine singular oblique reflexive possessive adjective (*apnē*)
PGRM2N	اپنے	Masculine plural nominative reflexive possessive adjective (*apnē*)
PGRM2O	اپنے	Masculine plural oblique reflexive possessive adjective (*apnē*)
PGRF1N	اپنی	Feminine singular nominative reflexive possessive adjective (*apnī*)
PGRF1O	اپنی	Feminine singular oblique reflexive possessive adjective (*apnī*)
PGRF2N	اپنی	Feminine plural nominative reflexive possessive adjective (*apnī*)
PGRF2O	اپنی	Feminine plural oblique reflexive possessive adjective (*apnī*)
PNN	کچھ, کوئی	Nominative indefinite pronoun (*kōī, kuch, sab*)
PNO	کسی	Oblique indefinite pronoun (*kīsī, kuch, sabhō~*)
PA	آپ	Honorific pronoun (*āp*)
QQ	کیا	Question marker *kyā*
RR	ہمیشہ	General adverb
RRJ	بائیں	General adverb derived from adjective
RD	زیادہ	Degree adverb
RM	ضرور	Modal adverb
RMN	نہیں، نہ، مت	Negative modal adverb (*nahī~, nah, mat*)
RY	اب	Proximal demonstrative adverb (*ab, yahā~, idhar, yū~*)
RYXHC	یہیں	Fused proximal demonstrative adverb and exclusive emphatic particle: *yahā~* + *hī* =

		yahī~
RYJ	ایسے	Proximal demonstrative adverb derived from adjective (*aisē)*
RV	وہاں	Distal demonstrative adverb (*tab*, *vahā~*, *udhar*, *tyū~*)
RVXHC	وہیں	Fused distal demonstrative adverb and exclusive emphatic particle: *vahā~* + *hī* = *vahī~*
RVJ	ویسے	Distal demonstrative adverb derived from adjective (*vaisē)*
RK	کیوں	Interrogative adverb (*kab*, *kahā~*, *kidhar*, *kyō~*)
RKXHC	کہیں	Fused interrogative adverb and exclusive emphatic particle: *kahā~* + *hī* = *kahī~*
RKJ	کیسے	Interrogative adverb derived from adjective (*kaisē)*
RJ	جدھر	Relative adverb (*jab*, *jahā~*, *jidhar*, *jū~*)
RJXHC	جہیں	Fused relative adverb and exclusive emphatic particle: *jahā~* + *hī* = *jahī~*
RJJ	جیسے	Relative adverb derived from adjective (*jaisē)*
TT	سہی	Sentence tag-word
VV0	سن	Root form lexical verb
VVNM1N	سننا	Infinitive lexical verb, masculine singular nominative
VVNM1O	سننے	Infinitive lexical verb, masculine singular oblique
VVNM2	سننے	Infinitive lexical verb, masculine plural nominative
VVNF1	سننی	Infinitive lexical verb, feminine singular nominative
VVNF2	سننی	Infinitive lexical verb, feminine plural nominative
VVTM1N	سنتا	Masculine singular (nominative) imperfective participle lexical verb
VVTM1O	سنتے	Masculine singular oblique imperfective participle lexical verb
VVTM2N	سنتے	Masculine plural (nominative) imperfective participle lexical verb
VVTM2O	سنتے	Masculine plural oblique imperfective participle lexical verb
VVTF1N	سنتی	Feminine singular (nominative) imperfective participle lexical verb
VVTF1O	سنتی	Feminine singular oblique imperfective participle lexical verb
VVTF2N	سنتی ,	Feminine plural (nominative) imperfective participle lexical verb

	سنتیں	
VVTF2O	سنتی	Feminine plural oblique imperfective participle lexical verb
VVYM1N	سنا	Masculine singular (nominative) perfective participle lexical verb
VVYM1O	سنے	Masculine singular oblique perfective participle lexical verb
VVYM2N	سنے	Masculine plural (nominative) perfective participle lexical verb
VVYM2O	سنے	Masculine plural oblique perfective participle lexical verb
VVYF1N	سنی	Feminine singular (nominative) perfective participle lexical verb
VVYF1O	سنی	Feminine singular oblique perfective participle lexical verb
VVYF2N	سنیں , سنی	Feminine plural (nominative) perfective participle lexical verb
VVYF2O	سنی	Feminine plural oblique perfective participle lexical verb
VVSM1	سنوں	First person singular subjunctive lexical verb
VVSM2	سنیں	First person plural subjunctive lexical verb
VVST1	سنے	Second person singular subjunctive lexical verb
VVST2	سنو	Second person plural subjunctive lexical verb
VVSV1	سنے	Third person singular subjunctive lexical verb
VVSV2	سنیں	Third person plural subjunctive lexical verb
VVIT1	سن	Second person singular imperative lexical verb
VVIT2	سنو	Second person singular imperative lexical verb
VVIA	سنئے	Second person honorific imperative lexical verb
VX0	پڑ	Root form general auxiliary verb
VXNM1N	پڑنا	Infinitive general auxiliary verb, masculine singular nominative
VXNM1O	پڑنے	Infinitive general auxiliary verb, masculine singular oblique
VXNM2	پڑنے	Infinitive general auxiliary verb, masculine plural nominative
VXNF1	پڑنی	Infinitive general auxiliary verb, feminine singular nominative
VXNF2	پڑنی	Infinitive general auxiliary verb, feminine plural nominative

VXTM1N	پڑتا	Masculine singular (nominative) imperfective participle general auxiliary verb
VXTM1O	پڑتے	Masculine singular oblique imperfective participle general auxiliary verb
VXTM2N	پڑتے	Masculine plural (nominative) imperfective participle general auxiliary verb
VXTM2O	پڑتے	Masculine plural oblique imperfective participle general auxiliary verb
VXTF1N	پڑتی	Feminine singular (nominative) imperfective participle general auxiliary verb
VXTF1O	پڑتی	Feminine singular oblique imperfective participle general auxiliary verb
VXTF2N	پڑتی , پڑتیں	Feminine plural (nominative) imperfective participle general auxiliary verb
VXTF2O	پڑتی	Feminine plural oblique imperfective participle general auxiliary verb
VXYM1N	پڑا	Masculine singular (nominative) perfective participle general auxiliary verb
VXYM1O	پڑے	Masculine singular oblique perfective participle general auxiliary verb
VXYM2N	پڑے	Masculine plural (nominative) perfective participle general auxiliary verb
VXYM2O	پڑے	Masculine plural oblique perfective participle general auxiliary verb
VXYF1N	پڑی	Feminine singular (nominative) perfective participle general auxiliary verb
VXYF1O	پڑی	Feminine singular oblique perfective participle general auxiliary verb
VXYF2N	پڑی , پڑیں	Feminine plural (nominative) perfective participle general auxiliary verb
VXYF2O	پڑی	Feminine plural oblique perfective participle general auxiliary verb
VXSM1	پڑوں	First person singular subjunctive general auxiliary verb
VXSM2	پڑیں	First person plural subjunctive general auxiliary verb
VXST1	پڑے	Second person singular subjunctive general auxiliary verb
VXST2	پڑو	Second person plural subjunctive general auxiliary verb
VXSV1	پڑے	Third person singular subjunctive general auxiliary verb

VXSV2	پڑیں	Third person plural subjunctive general auxiliary verb
VXIT1	پڑ	Second person singular imperative general auxiliary verb
VXIT2	پڑو	Second person singular imperative general auxiliary verb
VXIA	پڑنے	Second person honorific imperative general auxiliary verb
VGM1	گا	Masculine singular future auxiliary *gā*
VGM2	گے	Masculine plural future auxiliary *gē*
VGF1	گی	Feminine singular future auxiliary *gī*
VGF2	گی	Feminine plural future auxiliary *gī*
VRM1	رہا	Masculine singular durative auxiliary *rahā*
VRM2	رہے	Masculine plural durative auxiliary *rahē*
VRF1	رہی	Feminine singular durative auxiliary *rahī*
VRF2	رہی	Feminine plural durative auxiliary *rahī*
VC1	چاہئے	Singular cāhiē-type auxiliary
VC2	چاہئیں	Plural cāhiē-type auxiliary
VH0	ہو	Root form *hō*
VHNM1N	ہونا	Infinitive *hōnā*, masculine singular nominative
VHNM1O	ہونے	Infinitive *hōnē*, masculine singular oblique
VHNM2	ہونے	Infinitive *hōnē*, masculine plural nominative
VHNF1	ہونی	Infinitive *hōnī*, feminine singular nominative
VHNF2	ہونی	Infinitive *hōnī*, feminine plural nominative
VHTM1N	ہوتا	Masculine singular (nominative) imperfective participle *hōtā*
VHTM1O	ہوتے	Masculine singular oblique imperfective participle *hōtē*
VHTM2N	ہوتے	Masculine plural (nominative) imperfective participle *hōtē*
VHTM2O	ہوتے	Masculine plural oblique imperfective participle *hōtē*
VHTF1N	ہوتی	Feminine singular (nominative) imperfective participle *hōtī*
VHTF1O	ہوتی	Feminine singular oblique imperfective

		participle *hōtī*
VHTF2N	ہوتی, ہوتیں	Feminine plural (nominative) imperfective participle *hōtī / hōtī~*
VHTF2O	ہوتی	Feminine plural oblique imperfective participle *hōtī*
VHYM1N	ہوا	Masculine singular (nominative) perfective participle *hūā*
VHYM1O	ہوئے	Masculine singular oblique perfective participle *hūē*
VHYM2N	ہوئے	Masculine plural (nominative) perfective participle *hūē*
VHYM2O	ہوئے	Masculine plural oblique perfective participle *hūē*
VHYF1N	ہوئی	Feminine singular (nominative) perfective participle *hūī*
VHYF1O	ہوئی	Feminine singular oblique perfective participle *hūī*
VHYF2N	ہوئی, ہوئیں	Feminine plural (nominative) perfective participle *hūī / hūī~*
VHYF2O	ہوئی	Feminine plural oblique perfective participle *hūī*
VHSM1	ہوں	First person singular subjunctive *hū~*
VHSM2	ہوں	First person plural subjunctive *hō~*
VHST1	ہو	Second person singular subjunctive *hō*
VHST2	ہو	Second person plural subjunctive *hō*
VHSV1	ہو	Third person singular subjunctive *hō*
VHSV2	ہوں	Third person plural subjunctive *hō~*
VHIT1	ہو	Second person singular imperative *hō*
VHIT2	ہو	Second person plural imperative *hō*
VHIA		Second person honorific imperative
VHHM1	ہوں	First person singular indicative present *hū~*
VHHM2	ہیں	First person plural indicative present *hai~*
VHHT1	ہے	Second person singular indicative present *hai*
VHHT2	ہو	Second person plural indicative present *hō*
VHHV1	ہے	Third person singular indicative present *hai*
VHHV2	ہیں	Third person plural indicative present *hai~*

VHPM1	تھا	Masculine singular indicative past *thā*
VHPM2	تھے	Masculine plural indicative past *thē*
VHPF1	تھی	Feminine singular indicative past *thī*
VHPF2	تھیں	Feminine plural indicative past *thī~*
XT	تو	Contrastive emphatic particle *tō*
XH	ہی	Exclusive emphatic particle *hī*
XHC	ی، یں، ہیں	Clitic exclusive emphatic particle *ī*, *ī~*, *hī~*
ZZ	ئے	izāfat
.	۔	Full stop (U+06D4)
,	،	Comma (U+060C)
?	؟	Question mark (U+061F)
!	!	Exclamation mark (U+0021)
:	:	Colon (U+003A)
;	؛	Semi-colon (U+061B)
"	"	Neutral quotation mark (U+0022)
(	)	Open parenthesis (U+0028)
)	(	Close parenthesis (U+0029)
[	[	Open square bracket (U+005B)
]	[	Close square bracket (U+005D)
~	/	Other punctuation

Arabic Tagset

Tag	Description of word category	Example (Arabic)	Transcription	Translation
NCSgMNI	Singular, masculine, nominative, indefinite common noun	كتابٌ	*kitabun*	book
NCSgMAI	Singular, masculine, accusative, indefinite common noun	كتاباً	*kitaban*	book
NCSgMGI	Singular, masculine, genitive, indefinite common noun	كتابٍ	*kitabin*	book
NCSgMND	Singular, masculine, nominative, definite common noun	الكتابُ	*alkitabu*	the book
NCSgMAD	Singular, masculine, accusative, definite common noun	الكتابَ	*alkitaba*	the book
NCSgMGD	Singular, masculine, genitive, definite common noun	الكتابِ	*alkitabi*	the book
NCSgFNI	Singular, feminine, nominative, indefinite common noun	مدرسةٌ	*madrasatun*	school
NCSgFAI	Singular, feminine, accusative, indefinite common noun	مدرسةً	*madrasatan*	school
NCSgFGI	Singular, feminine, genitive, indefinite common noun	مدرسةٍ	*madrasatin*	school
NCSgFND	Singular, feminine, nominative, definite common noun	المدرسةُ	*almadrasatu*	the school

NCSgFAD	Singular, feminine, accusative, definite common noun	المدرسةَ	*almadrasata*	the school
NCSgFGD	Singular, feminine, genitive, definite common noun	المدرسةِ	*aladrasati*	the school
NCDuMNI	Dual, masculine, nominative, indefinite common noun	كتابان	*kitaban*	two books
NCDuMAI	Dual, masculine, accusative, indefinite common noun	كتابين	*kitabain*	two books
NCDuMGI	Dual, masculine, genitive, indefinite common noun	كتابين	*kitabain*	two books
NCDuMND	Dual, masculine, nominative, definite common noun	الكتابان	*alkitaban*	the two books
NCDuMAD	Dual, masculine, accusative, definite common noun	الكتابين	*alkitabain*	the two books
NCDuMGD	Dual, masculine, genitive, definite common noun	الكتابين	*alkitabain*	the two books
NCDuFNI	Dual, feminine, nominative, indefinite common noun	مدرستان	*madrasatan*	two books
NCDuFAI	Dual, feminine, accusative, indefinite common noun	مدرستين	*madrasatain*	two schools
NCDuFGI	Dual, feminine, genitive, indefinite common noun	مدرستين	*madrasatain*	two schools
NCDuFND	Dual, feminine, nominative, definite common noun	المدرستان	*almadrasatan*	the two schools
NCDuFAD	Dual, feminine, accusative, definite common noun	المدرستين	*almadrasatain*	the two schools
NCDuFGD	Dual, feminine, genitive, definite common noun	المدرستين	*almadrasatain*	the two schools
NCPlMNI	Plural, masculine, nominative, indefinite common noun	كتبٌ - مسلمون	*muslimoon – kutubun*	Muslims – books
NCPlMAI	Plural, masculine, accusative, indefinite common noun	كتباً - مسلمين	*muslimeen – kutuban*	Muslims – books
NCPlMGI	Plural, masculine, genitive, indefinite common noun	كتبٍ - مسلمين	*muslimeen – kutubin*	Muslims – books
NCPlMND	Plural, masculine, nominative, definite common noun	الكتبُ - المسلمون	*almuslimoon – alkutubu*	the Muslims – the books
NCPlMAD	Plural, masculine, accusative, definite common noun	الكتبَ - المسلمين	*aluslimeen – alkutuba*	the Muslims – the books
NCPlMGD	Plural, masculine, genitive, definite common noun	الكتبِ - المسلمين	*almuslimeen – alkutubi*	the Muslims – the books
NCPlFNI	Plural, feminine, nominative, indefinite common noun	مسلماتٌ–مدارسٌ	*madarisun – muslimaatun*	schools – Muslims
NCPlFAI	Plural, feminine, accusative, indefinite common noun	مسلماتاً–مدارساً	*madarisan – muslimaatan*	schools – Muslims
NCPlFGI	Plural, feminine, genitive, indefinite, common noun	مسلماتٍ–مدارسٍ	*madarisin – muslimaatin*	schools – Muslims
NCPlFND	Plural, feminine, nominative, definite common noun	المسلماتُ–المدارسُ	*almadarisu – almuslimaatu*	the schools – the Muslims
NCPlFAD	Plural, feminine, accusative, definite common noun	المسلماتَ–المدارسَ	*almadarisa – almuslimaata*	the schools – the Muslims
NCPlFGD	Plural, feminine, genitive, definite common noun	المسلماتِ–المدارسِ	*almadarisi – almuslimaati*	the schools – the Muslims
NP	Proper noun	شيرين–جده	*Jiddah – Shyryn*	Jeddah – Shereen
NPrPSg1	First person, singular, neuter, personal pronoun	كتابي - ضربني–أنا	*ana- kitaabee – ḍarabanee*	Me – my book – he hit me
NPrPSg2M	Second person, singular, masculine, personal pronoun	كتابكَ–أنتَ	*anta – kitaabuka*	You – your book
NPrPSg2F	Second person, singular, feminine, personal pronoun	كتابكِ–أنتِ	*anti – kitaabuki*	You – your book
NPrPSg3M	Third person, singular, masculine, personal pronoun	هو - كتابه	*kitaabahu – huwa*	His book – him
NPrPSg3F	Third person, singular, feminine, personal	هي - كتابها	*kitaabuhaa – hiya*	Her book –

	pronoun			her
NPrPDu2	Second person, dual, neuter, personal pronoun	كتابكما–أنتما	*antumaa – kitaabakumaa*	You two – your book
NPrPDu3	Third person, dual, neuter, personal pronoun	كتابهما–هما	*humaa – kitaabahumaa*	Those two – their book
NPrPP11	First person, plural, neuter, personal pronoun	كتابنا–نحن	*naḥnu – kitaabunaa*	Us – our book
NPrPP12M	Second person, plural, masculine, personal pronoun	كتابكم–أنتم	*antum – kitaabakum*	You – your book
NPrPP12F	Second person, plural, feminine, personal pronoun	كتابكن–أنتن	*antunna – kitaabakunna*	You – your book
NPrPP13M	Third person, plural, masculine, personal pronoun	كتابهم–هم	*hum – kitaabahum*	Them – their book
NPrPP13F	Third person, plural, feminine, personal pronoun	هن - كتابهن	*kitaabahunna – hunna*	Their book – them
NPrRSSgM	Singular, masculine, specific, relative pronoun	الذي	*allathi*	Who
NPrRSSgF	Singular, feminines, specific, relative pronoun	التي	*allati*	Who
NPrRSDuM	Dual, masculine, specific, relative pronoun	اللذين–اللذان	*alladhaani – alladhaini*	Who
NPrRSDuF	Dual, feminine, specific, relative pronoun	اللتين–اللتان	*allataani – allataini*	Who
NPrRSPlM	Plural, masculine, specific, relative pronoun	اللذين - اللائي	*allaiy – alladheena*	Who
NPrRSPlf	Plural, feminine, specific, relative pronoun	اللاتي - اللائي	*allaaiy - allatee*	Who
NPrRC	Common, relative pronoun	مهما– ما –من	*men – maa – mahmaa*	Who – what
NPrDSgM	Singular, masculine, demonstrative pronoun	ذلك– ذاك – ذا –هذا	*hadhaa – dhaa – dhaaka – dhaalika*	This – that
NPrDSgF	Singular, feminine, demonstrative pronoun	– تلك – ذي – ذه – هذي –هذه تيك–تاك	*haadhihi – haadhee – dhih – dhy – tilka – taaka – teeka*	This – that
NprDDuM	Dual, masculine, demonstrative pronoun	– هذين – ذانك – ذان –هذان ذينك–ذين	*haadhani – dhaani – dhaanika – hadhaini –dhaini*	This – that
NprDDuF	Dual, feminine, demonstrative pronoun	تين – هتين – تانك – تان –هتان تينك–	*haatani – taani- taanika – haataini*	This – that
NPrDPl	Plural, neutral, demonstrative pronoun	أولائك – أولاء – أولى –هؤلاء أولاك– أولالك –	*haaolaai – olaa- olaaika- olaalika – olaaka*	Those
NNuCaSgM	Singular, masculine, cardinal number	أربع	*arba'*	Four
NNuCaSgF	Singular, feminine, cardinal number	أربعة	*arba'a*	Four
NNuOrSgM	Singular, masculine, ordinal number	رابع	*raabi'*	Fourth
NNuOrSgF	Singular, feminine, ordinal number	رابعة	*raabia*	Fourth
NNuNaSgM	Singular, masculine, numerical adjective	رباعي	*rubaa'y*	Of four
NNuNaSgF	Singular, feminine, numerical adjective	رباعية	*rubaa'iya*	Of four
NACSgMNI	Singular, masculine, nominative, indefinite adjective	سعيدٌ	*sa'ydun*	happy
NACSgMAI	Singular, masculine, accusative, indefinite adjective	سعيداً	*sa'ydan*	happy
NACSgMGI	Singular, masculine, genitive, indefinite adjective	سعيدٍ	*sa'ydin*	happy
NACSgMND	Singular, masculine, nominative, definite adjective	السعيدُ	*alsa'ydu*	the happy
NACSgMAD	Singular, masculine, accusative, definite adjective	السعيدَ	*alsa'yda*	the happy
NACSgMGD	Singular, masculine, genitive, definite adjective	السعيدِ	*alsa'ydi*	the happy
NACSgFNI	Singular, feminine, nominative, indefinite adjective	سعيدةٌ	*sa'ydatun*	happy

NACSgFAI	Singular, feminine, accusative, indefinite adjective	سعيدتاً	*sa'ydatan*	happy
NACSgFGI	Singular, feminine, genitive, indefinite adjective	سعيدةٍ	*sa'ydatin*	happy
NACSgFND	Singular, feminine, nominative, definite adjective	السعيدةُ	*alsa'ydatu*	the happy
NACSgFAD	Singular, feminine, accusative, definite adjective	السعيدةَ	*alsa'ydata*	the happy
NACSgFGD	Singular, feminine, genitive, definite adjective	السعيدةِ	*alsa'ydati*	the happy
NACDuMNI	Dual, masculine, nominative, indefinite adjective	سعيدان	*sa'ydan*	two happy
NACDuMAI	Dual, masculine, accusative, indefinite adjective	سعيدين	*sa'ydain*	two happy
NACDuMGI	Dual, masculine, genitive, indefinite adjective	سعيدين	*sa'ydain*	two happy
NACDuMND	Dual, masculine, nominative, definite adjective	السعيدان	*alkitaban*	the two happy
NACDuMAD	Dual, masculine, accusative, definite adjective	السعيدين	*alsa'ydain*	the two happy
NACDuMGD	Dual, masculine, genitive, definite adjective	السعيدين	*alsa'ydain*	the two happy
NACDuFNI	Dual, feminine, nominative, indefinite adjective	سعيدتان	*sa'ydatan*	two happy
NACDuFAI	Dual, feminine, accusative, indefinite adjective	سعيدتين	*sa'ydatain*	two happy
NACDuFGI	Dual, feminine, genitive, indefinite adjective	سعيدتين	*sa'ydatain*	two happy
NACDuFND	Dual, feminine, nominative, definite adjective	السعيدتان	*alsa'ydatan*	the two happy
NACDuFAD	Dual, feminine, accusative, definite adjective	السعيدتين	*alsa'ydatain*	the two happy
NACDuFGD	Dual, feminine, genitive, definite adjective	السعيدتين	*alsa'ydatain*	the two happy
NACPlMNI	Plural, masculine, nominative, indefinite adjective	سعيدون	*sa'ydoon*	happy
NACPlMAI	Plural, masculine, accusative, indefinite adjective	سعيدين	*sa'ydeen*	happy
NACPlMGI	Plural, masculine, genitive, indefinite adjective	سعيدين	*sa'ydeen*	happy
NACPlMND	Plural, masculine, nominative, definite adjective	السعيدون	*alsa'ydoon*	the happy
NACPlMAD	Plural, masculine, accusative, definite adjective	السعيدين	*alsa'ydeen*	the happy
NACPlMGD	Plural, masculine, genitive, definite adjective	السعيدين	*alsa'ydeen*	the happy
NACPlFNI	Plural, feminine, nominative, indefinite adjective	سعيداتٌ	*sa'ydaatun*	happy
NACPlFAI	Plural, feminine, accusative, indefinite adjective	سعيداتاً	*sa'ydaatan*	happy
NACPlFGI	Plural, feminine, genitive, indefinite, adjective	سعيداتٍ	*sa'ydaatin*	happy
NACPlFND	Plural, feminine, nominative, definite adjective	السعيداتُ	*alsa'ydaatu*	the happy
NACPlFAD	Plural, feminine, accusative, definite adjective	السعيداتَ	*alsa'ydaata*	the happy
NACPlFGD	Plural, feminine, genitive, definite adjective	السعيداتِ	*alsa'ydaati*	the happy
VPSg1	First person, singular, neuter, perfect verb	كسرتُ	*kasartu*	I broke
VPSg2M	Second person, singular, masculine, perfect verb	كسرتَ	*kasarta*	You broke
VPSg2F	Second person, singular, feminine, perfect verb	كسرتِ	*kasarti*	You broke
VPSg3M	Third person, singular, masculine, perfect verb	كسر	*kasara*	He broke
VPSg3F	Third person, singular, feminine, perfect verb	كسرتْ	*kasarat*	She broke
VPDu2	Second person, dual, neuter, perfect verb	كسرتما	*kasartumaa*	You (two) broke

VPDu3M	Third person, dual, masculine, perfect verb	كسرا	*kasaraa*	They (two) broke
VPDu3F	Third person, dual, feminine, perfect verb	كسرتا	*kasarataa*	They (two) broke
VPPl1	First person, plural, neuter, perfect verb	كسرنا	*kasarnaa*	We broke
VPPl2M	Second person, plural, masculine, perfect verb	كسرتم	*kasartum*	You broke
VPPl2F	Second person, plural, feminine, perfect verb	كسرتن	*kasartunna*	You broke
VPPl3M	Third person, plural, masculine, perfect verb	كسروا	*kasaroo*	They broke
VPPl3F	Third person, plural, feminine, perfect verb	كسرن	*kasarna*	They broke
VISg1I	First person, singular, neuter, indicative, imperfect verb	أكسرُ	*aksiru*	I break
VISg1S	First person, singular, neuter, subjunctive, imperfect verb	أكسرَ	*aksira*	I break
VISg1J	First person, singular, neuter, jussive, imperfect verb	أكسرْ	*aksir*	I break
VISg2MI	Second person, singular, masculine, indicative, imperfect verb	تكسرُ	*taksiru*	You break
VISg2MS	Second person, singular, masculine, subjunctive, imperfect verb	تكسرَ	*taksira*	You break
VISg2MJ	Second person, singular, masculine, jussive, imperfect verb	تكسرْ	*taksir*	You break
VISg2FI	Second person, singular, feminine, indicative, imperfect verb	تكسرين	*taksiryna*	You break
VISg2FS	Second person, singular, feminine, subjunctive, imperfect verb	تكسري	*taksiry*	You break
VISg2FJ	Second person, singular, feminine, jussive, imperfect verb	تكسري	*taksiry*	You break
VISg3MI	Third person, singular, masculine, indicative, imperfect verb	يكسرُ	*yaksiru*	He breaks
VISg3MS	Third person, singular, masculine, subjunctive, imperfect verb	يكسرَ	*yaksira*	He breaks
VISg3MJ	Third person, singular, masculine, jussive, imperfect verb	يكسرْ	*yaksir*	He breaks
VISg3FI	Third person, singular, feminine, indicative, imperfect verb	تكسرُ	*taksiru*	She breaks
VISg3FS	Third person, singular, feminine, subjunctive, imperfect verb	تكسرَ	*taksira*	She breaks
VISg3FJ	Third person, singular, feminine, jussive, imperfect verb	تكسرْ	*taksir*	She breaks
VIDu2I	Second person, dual, neuter, indicative, imperfect verb	تكسران	*taksiraani*	You break
VIDu2S	Second person, dual, neuter, subjunctive, imperfect verb	تكسرا	*taksiraa*	You break
VIDu2J	Second person, dual, neuter, jussive, imperfect verb	تكسرا	*taksiraa*	You break
VIDu3MI	Third person, dual, masculine, indicative, imperfect verb	يكسران	*yaksiraani*	They break
VIDu3MS	Third person, dual, masculine, subjunctive, imperfect verb	يكسرا	*yaksiraa*	They break
VIDu3MJ	Third person, dual, masculine, jussive, imperfect verb	يكسرا	*yaksiraa*	They break
VIDu3FI	Third person, dual, feminine, indicative, imperfect verb	يكسران	*yaksiraan*	They break
VIDu3FS	Third person, dual, feminine, subjunctive, imperfect verb	يكسرا	*yaksiraa*	They break
VIDu3FJ	Third person, dual, feminine, jussive, imperfect verb	يكسرا	*yaksiraa*	They break
VIPl1I	First person, plural, neuter, indicative, imperfect verb	نكسرُ	*naksiru*	We break
VIPl1S	First person, plural, neuter, subjunctive, imperfect verb	نكسرَ	*naksira*	We break
VIPl1J	First person, plural, neuter, jussive, imperfect verb	نكسرْ	*naksir*	We break
VIPl2MI	Second person, plural, masculine, indicative, imperfect verb	تكسرون	*taksiroon*	You break

Hindi Tagset

Table 18: Tagset for Hindi language[6]

Main Category	**Sub category**	**Example**
Noun	Noun	Boy, river, thought, hardness
	Location	Up, down, front, back
	Compound	
Proper noun	*Compound*	RAM, BJP
Pronoun		Who, that, he, the boy who
Verb	Verb finite main	He drinks, the boy is
	Auxiliary	Has
	Nonfinite adjectival	Eating
	Nonfinite adverbial	After eating, drinking
	Nonfinite nominal	Drinking
Adjective		
Adverb.		Slowly, fast
Postposition		By, for
Particle		ہی، بھی، تو
Conjunct		And, or , that
Question words		What, how
Quantifer		More, little, all, much
Number quatifier		Third, three
Intensifier		Too much, much more
Negative		No, not
Interjection words		
Special		

[6] A part of speech tagger for Indian languages, available at http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

Tagset of Penn Treebank

Table 19: Pen TreeBank tagset for English[7]

Category	Sub category
Coordinating conjunction	
Cardinal number	
Determiner	
Existential there	
Foreign word	
Preposition or subordinating conjunction	
Adjective	Comparative, superlative
List item marker	
Modal	
Noun	Singular, plural, proper singular, proper plural
Pre-determiner	
Pronoun	Personal , possessive
Adverb	Comparative, superlative
Particle	
Symbol	
To	
Interjection	
Verb	Root, past tense, gerund, past participle, non-3rd person singular present, 3rd person singular present
Question words	Wh-determiner, wh-pronoun, possessive wh-pronoun, wh-adverb
Punctuation marks	

[7] The information about Penn TreeBank is taken from the following document: http://www.ling.ohio-state.edu/~dm/02/spring/795K/casden-treebank-4up.pdf