Analysis and Development of Urdu POS Tagged Corpus

Ahmed Muaz
Center for Research in Urdu Language Processing
NUCES, Pakistan
ahmed.muaz@nu.edu.pk

Aasim Ali
Center for Research in Urdu Language Processing,
NUCES, Pakistan
aasim.ali@nu.edu.pk

Sarmad Hussain
Center for Research in Urdu Language Processing
NUCES, Pakistan
sarmad.hussain@nu.edu.pk

Abstract

In this paper, two corpora of Urdu (with 110K and 120K words) tagged with different POS tagsets are used to train TnT and Tree taggers. Error analysis of both taggers is done to identify frequent confusions in tagging. Based on the analysis of tagging, and syntactic structure of Urdu, a more refined tagset is derived. The existing tagged corpora are tagged with the new tagset to develop a single corpus of 230K words and the TnT tagger is retrained. The results show improvement in tagging accuracy for individual corpora to 94.2% and also for the merged corpus to 91%. Implications of these results are discussed.

1 Introduction

There is increasing amount of work on computational modeling of Urdu language. As various groups work on the language, diversity in analysis is also developed. In this context, there has been some work on Urdu part of speech (POS) tagging, which has caused multiple tagsets to appear. Thus, there is also need to converge these efforts.

Current work compares the existing tagsets of Urdu being used for tagging corpora in an attempt to look at the differences, and understand the reasons for the variation. The work then undertakes experiments to develop a common tagset, which is syntactically and computationally coherent. The aim is to make a robust tagset and then to port the differently tagged Urdu corpora onto the same tagset. As Urdu already has very few annotated corpora, this will help consolidating them for better modeling.

The next sections present the existing tagsets and accuracies of the POS taggers reported using them. Sections 4 and 5 present baseline experiment and the methodology used for the analysis for updating the tagset. Section 6 describes the proposed tagset. Section 7 reports experiments comparing the new tagset with existing ones. Section 8 discusses the results achieved and future directions.

2 Relevant Resources of Urdu

2.1 Urdu Corpora

Several annotated corpora have been built during last few years to facilitate computational processing for Urdu language. The initial work was undertaken through EMILLE project to build multi-lingual corpora for South Asian languages (McEnery et al., 2000). They released 200,000 words parallel corpus of English, Urdu, Bengali, Gujarati, Hindi and Punjabi. In addition, there are 1,640,000 words of Urdu text in this corpus. These text collections are also annotated with part of speech tags (Hardie 2003).

Center for Research in Urdu Language Processing (CRULP[1]) gathered 18 million words corpus in order to build a lexicon. It has cleaned text from news websites from multiple domains (Ijaz et.al. 2007). Following this work, a syntactic tagset was developed based on work by existing grammarians and a corpus of 110,000 words was manually tagged. This annotated corpus is available through the center (Sajjad 2007, Hussain 2008).

Recently an English-Urdu parallel corpus has also been developed by CRULP, by translating the first 140,000 words of PENN Treebank corpus. In addition, a tagset has also been designed following the PENN Treebank guidelines. These words have been tagged manually with this new tagset. This collection is also available from CRULP, and the tagset is still unpublished.

2.2 Urdu Part of Speech tagsets

Hardie (2003) developed the first POS tagset for Urdu using EAGLES guidelines for computational processing. The tagset contains 282 morpho-syntactic tags, differentiating on the basis of number, gender and other morphological details

[1] www.crulp.org

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 24–31,
Suntec, Singapore, 6-7 August 2009.

in addition to the syntactic categories. Punctuation marks are tagged as they are, and not included in 282 tags. The tags include the gender and number agreement, in addition to syntactic information.

The complications of Urdu tagset design are also discussed. One of these complexities is word segmentation issue of the language. Suffixes in Urdu are written with an orthographic space. Words are separated on the basis of space and so suffixes are treated same as lexical words. Hence it is hard to assign accurate tag for an automatic tagger. Although the tagset is designed considering details, but due to larger number of tags it is hard to get a high accuracy with a small sized corpus. Due to its morphological dependence and its large size, this tagset is not considered in our analysis.

Two much smaller tagsets are considered for this work. They are compared in detail in Section 6. The first tagset, containing 42 tags, is designed by Sajjad (2007), based on the work of Urdu grammarians (e.g. Schmidt 1999, Haq 1987, Javed 1981, Platts 1909) and computational work by Hardie (2003). The main features of the tagset include multiple pronouns (PP, KP, AKP, AP, RP, REP, G, and GR) and demonstratives (PD, KD, AD, and RD). It has only one tag for all forms of verbs (VB), except for auxiliaries to show aspect (AA) and tense (TA) information about the verb. All noun types are assigned single tag (NN) except for Proper Nouns (PN). It also has a special tag NEG to mark any occurrence negation words (نہیں "not" and نہ "no" or "neither") regardless of context. It also has a tag SE to mark every occurrence of سے ("from") without considering the context. Another example of such a context-free lexical tag is WALA to mark every occurrence (including all the inflections) of the word والا. This tagset is referred to as T1 subsequently in this paper.

Recently Sajjad and Schmid (2009) used the tagged data of 107,514 words and carried out an experiment for tagger comparison. A total of 100,000 words are used as training set and rest as test data. Four taggers (TnT, Tree, RF and SVM) are trained using training corpus and then tested accordingly. Reported results of this work show that SVM tagger is the most accurate, showing 94.15% correct prediction of tags. Remaining three taggers have accuracies of 93.02% (Tree tagger), 93.28% (RF tagger) and 93.40% (TnT tagger).

Another tagset has recently been developed as a part of a project to develop English-Urdu parallel corpus at CRULP, following the Penn Treebank guidelines (Santorini 1990). It contains 46 tags, with fewer grades of pronouns (PR, PRP$, PRRF, PRRFP$, and PRRL) and demonstratives (DM and DMRL), as compared to T1. It has several tags for verbs on the basis of their forms and semantics (VB, VBI, VBL, VBLI, and VBT) in addition to the tags for auxiliaries showing aspect (AUXA) and tense (AUXT). The NN tag is assigned for both singular and plural nouns and includes adverbial *kaf* pronoun, *kaf* pronoun, and adverbial pronoun categories of T1. Yet, it has several other grades of common nouns (NNC, NNCR, NNCM). It also has two shades of Proper Nouns (NNP, NNPC), which are helpful in identifying phrase boundary of compound proper nouns. It also has a tag WALA that is assigned to every occurrence (and inflection) of word والا (wala). However, marking of token سے ("from") is context dependent: either it is CM when marking case or it is RBRP when occurring as an adverbial particle. This tagset is referred to as T2 subsequently in this paper.

3 Tools and Resource Selection

The decision of selecting the tagger, the tagset, and the data is the starting point for the task of POS tagging. This section gives details of the taggers chosen and the corpora used for the experiments conducted.

3.1 Selection of taggers

There are a number of existing taggers available for tagging. Two POS taggers are used in the initial step of this work to compare the initial tagging accuracies.

One of the selected taggers is Trigram-and-Tag (TnT). It is a trigram based HMM tagger in which two preceding tags are used to find the transition probability of a tag. Brants (2000) tested PENN Treebank (English) and NEGRA (German) corpora and reported 96-97% accuracy of the tagger.

Schmid (1994) proposed probabilistic POS tagger that uses decision trees to store the transition probabilities. The trained decision tree is used for identification of highest probable tags. Schmid reported an accuracy of 95-96% on PENN Treebank for this tagger.

Both taggers give good accuracy for Urdu tagging, as reported by Sajjad and Schmid (2009).

3.2 Data Used for Experimentation

Corpora annotated with the different tagsets are acquired from CRULP. The corpus originally tagged with T1 tagset is referred to as C1 (news from non-business domain) and the corpus initially annotated with T2 tagset is referred to as C2 (news from business domain), subsequently in the current work. Both C1 and C2 are taken and cleaned. The data is re-counted and approximately 100,000 words are separated for training and rest are kept for testing. The details of data are given in Tables 1 and 2 below.

Table 1. Number of tokens in Urdu corpora

Tokens	C1	C2
Training	101,428	102,454
Testing	8,670	21,181
Total	110,098	123,635

Table 2. Number of sentences in Urdu corpora

Sentences	C1	C2
Training	4,584	3,509
Testing	404	755
Total	4,988	4,264

4 Baseline Estimation

The comparison is initiated with training of existing tagsets on their respective annotated data (T1 on C1 and T2 on C2). Both corpora are tested on TnT and Tree Tagger to obtain the confusion matrices for errors. These confusion matrices are used to analyze misclassification of tags. TnT tagger shows that overall accuracy of using T1 with C1 is 93.01% and is significantly better than using T2 with C2, which gives 88.13% accuracy. Tree tagger is also trained on the corpora. The overall accuracy of T1 on C1 (93.37%) is better than that of T2 on C2 (90.49%). The results are shown in Table 3.

Table 3. Results of both tagsets on their respective corpora with TnT and Tree taggers

	T1 on C1	T2 on C2
TnT Tagger	93.01%	88.13%
Tree Tagger	93.37%	90.49%

The accuracies reported (for T1 on C1) by Sajjad and Schmid (2009) are comparable to these accuracies. They have reported 93.40% for TnT Tagger and 93.02% for Tree Tagger.

Further experimentation is performed only using TnT tagger.

5 Methodology

The current work aims to build a larger corpus of around 230,000 manually tagged words for Urdu by combining C1 and C2. These collections are initially annotated with two different tagsets (T1 and T2 respectively, and as described above). For this unification, it was necessary to indentify the differences in the tagsets on which these corpora are annotated, analyzed the differences and then port them to unified tagset.

The work starts with the baseline estimation (described in Section 4 above). The results of baseline estimation are used to derive a new tagset (detailed in Section 6 below), referred to as T3 in this paper. Then a series of experiments are executed to compare the performance of three tagsets (T1, T2, and T3) on data from two different domains (C1 and C2), as reported in Section 7 below and summarized in Table 4.

Table 4. Summary of experiments conducted

	Experiment	Tagset	Corpus
0	Baseline Estimation: Original tagsets with respective corpora	T1	C1
		T2	C2
1	Experiment1: For comparison of results of T1 and T3 on C1	T3	C1
2	Experiment2: For comparison of T1, T2 and T3 on C2	T3	C2
		T1	C2
3	Experiment3: Comparison of T1 and T3 with no unknowns	T3	C2
		T1	C2
4	Experiment4: Comparison of T1 and T3 over complete corpus	T3	C1+C2
		T1	C1+C2

The performance of T1 on C1 is already better than T2 on C2, so the first comparison for the merged tagset T3 is with T1 on C1, which is the basis of the first experiment. Then the performance of better performing tagsets (T1 and T3) are compared on the corpus C2 in the second

experiment to compare them with T2. One possible reason of relatively better performance could be the difference in application of open classes for unknown words in the test data. Therefore, the third experiment is performed using the same data as in second experiment (i.e. corpus C2) with combined lexicon of training and test data (i.e. no unknown words). Finally, an experiment is conducted with the merged corpus. Following table summarizes these experiments.

6 Tagset design

After establishing the baseline, the existing tagsets are reviewed with the following guidelines:

- Focus on the syntactic variation (instead of morphological or semantic motivation) to either collapse existing tags or introduce new ones
- Focus on word level tagging and not try to accommodate phrase level tagging (e.g. to support chunking, compounding or other similar tasks)
- Tag according to the syntactic role instead of having a fixed tag for a string, where possible
- Use PENN Treebank nomenclature to keep the tagset easy to follow and share

Comparison of T1 and T2 showed that there are 33 tags in both tagsets which represent same syntactic categories, as shown in Appendix A. The tag I (Intensifier) in T2 labels the words which are marked as ADV in T1. The words annotated as NNC, NNCR and NNCM (under T2) are all labeled as NN under T1. The words tagged as VBL, VBLI, VBI, and VBLI (under T2) are all labeled as VB under T1. Range of distinct tags for demonstratives of T1 are all mapped to DM in T2 except RD (of T1) which maps to DMRL (of T2).

In order to identify the issues in tagging, a detailed error analysis of existing tagsets is performed. Following tables represent the major tag confusions for tagging C2 with T2 using Tree and TnT taggers.

Table 5. Major misclassifications in C2 with T2 tagset using Tree tagger

Tag	Total tokens	Errors	Maximum misclassification	
VB	888	214	183	VBL
VBL	328	168	151	VB
VBI	202	47	38	VBLI
VBLI	173	52	46	VBI
AUXT	806	145	121	VBT

Table 6. Major misclassifications in C2 with T2 tagset using TnT-tagger

Tag	Total tokens	Error	Maximum misclassification	
VB	888	240	181	VBL
VBL	328	154	135	VB
VBI	202	46	34	VBLI
VBLI	173	61	55	VBI
AUXT	806	136	111	VBT

The proposed tagset for Urdu part-of-speech tagging contains 32 tags. The construction of new tagset (T3) is initiated by adopting T2 as the baseline, because T2 uses the tagging conventions of PENN Treebank. There are 17 tags in T3 that are same as in T1 and T2. These tags (CC, CD, DM, DMRL, JJ, NN, OD, PM, PRP, PRP$, PRRF, PRRF$, PRRL, Q, RB, SM, SYM) are not discussed in detail. The complete tagset along with short description and examples of each tag is given in Appendix B.

RBRP (Adverbial Particle) and CM (Case Marker) are merged to make up a new tag PP (Postposition), so every postposition particle comes under this new tag ignoring semantic context. I (Intensifier) is used to mark the intensification of an adjective, which is a semantic gradation, and syntactically merged with Q (Quantifier). NNCM (Noun after Case Marker), NNC (Noun Continuation), NNCR (Continuing Noun Termination) are merged into NN (Noun) because syntactically they always behave similarly and the difference is motivated by phrase level marking. U (Unit) is also merged with NN because the difference is semantically motivated.

DATE is not syntactic, and may be either treated as NN (Noun) or CD (Cardinal), depending upon the context. Similarly, R (Reduplication), MOPE (Meaningless Pre-word), and MOPO (Meaningless Post-word) always occur in pair with NN, JJ, or another tag. Thus they are phrasal level tags, and can be replaced by relevant word level tag in context. NNPC (Proper Noun Continuation) tag identifies compounding but syntactically behaves as NNP (Proper Noun), and is not used.

VBL (Light Verb) is used in complex predicates (Butt 1995), but its syntactic similarity with VB (Verb) is a major source of confusion in automatic tagging. It is collapsed with VB (Verb). Similarly, VBLI (Light Verb Infinitive) is merged with VBI (Verb Infinitive). AUXT (Tense Auxiliary) is highly misclassified as VBT (To be Verb) because both occur as last token in a clause or sentence, and both include tense in-

formation. The word is labeled as VBT only when there is no other verb in the sentence or clause, otherwise these words are tagged as AUXT. The syntactic similarity of both tags is also evident from statistically misclassifying AUXT as VBT. Therefore both are collapsed into single tag VBT (Tense Verb).

In T1, NEG (Negation) is used to mark all the negation words without context, but they mostly occur as adverbs. Therefore, NEG tag is removed. Similarly, SE (Postposition سے , "from") is not separated from postpositions and marked accordingly. PRT (Pre-Title) and POT (Post-Title) always occur before or after Proper Noun, respectively. Therefore, they behave as Proper Nouns, hence proposed to be labeled as NNP (Proper Noun).

7 Experiments

After designing a new tagset, a series of experiments are conducted to investigate the proposed changes. The rationale of the sequence of experiments has been discussed in Section 5 above, however the reasoning for each experiment is also given below. As T2 tags have much more semantic and phrasal information, and C2 tagged with T2 shows lower accuracy than T1 on C1, therefore further experiments are conducted to compare the performance of T1 and T3 only. Comparisons on C2 with T3 may also be drawn.

7.1 Experiment 1

As baseline estimation shows that T1 on C1 outperforms T2 on C2, the first experiment is to compare the performance of T3 on C1. In this experiment C1 is semi-automatically tagged with T3. TnT tagger is then trained and tested. T3 gives 93.44% accuracy, which is slightly better than the results already obtained for T1 (93.01%). The results are summarized in Table 7.

Table 7. Accuracies of T3 and T1 on C1

Corpus	Tagset	Accuracy
C1	T3	93.44%
C1	T1	93.01%

7.2 Experiment 2

Now to test the effect of change in domain of the corpus, the performance T1 and T3 on C2 is compared in this experiment. C2 is manually tagged with T3, then trained and tested using TnT tagger. The results obtained with T3 are 91.98%, which are significantly better than the results already obtained for T2 on C2 (88.13%).

C2 is also semi-automatically re-tagged with T1. T1 shows better performance (91.31%) than T2 (88.13%). However, the accuracy of using T3 (on C2) is still slightly higher. The results are summarized in Table 8.

Table 8. Accuracies of T3 on C1, and accuracies of T3 and T1 on C2

Corpus	Tagset	Accuracy
C2	T3	91.98%
C2	T1	91.31%

7.3 Experiment 3

Due to the change in open class set there may be a difference of performance on unknown words, therefore in this experiment, all the unknown words of test set are also included in the vocabulary. This experiment again involves T3 and T1 with C2. Combined lexica are built using testing and training parts of the corpus, to eliminate the factor of unknown words. This experiment also shows that T3 performs better than T1, as shown in Table 9.

Table 9. Accuracies of T3 and T1 with ALL known words in test data

Corpus	Tagset	Accuracy
C2	T3	94.21%
C2	T1	93.47%

7.4 Experiment 4

Finally both corpora (C1 and C2) were combined, forming a training set of 203,882 words and a test set of 29,851 words. The lexica are generated only from the training set. Then TnT tagger is trained separately for both T1 and T3 tagsets and the accuracies are compared. The results show that T3 gives better tagging accuracy, as shown in Table 10.

Table 10. Accuracies of T3 and T1 using combined C1 and C2 corpora

Corpus	Tagset	Accuracy
C1+C2	T3	90.99%
C1+C2	T1	90.00%

Partial confusion matrices for both the tagsets are given in Tables 11 and 12.

The error analysis shows that the accuracy drops for both tagsets when trained on multi-domain corpus, which is expected. The highest error count is for the confusion between noun and adjective. There is also confusion between proper and common nouns. T3 also gives significant confusion between personal pronouns and demonstratives, as they represent the same lexical entries.

Table 11. Major misclassifications in merged corpus with T1 using TnT tagger

Tag	Total tokens	Error	Maximum misclassification	
A	18	5	3	ADJ
AD	18	7	4	ADJ
ADJ	2510	551	371	NN
ADV	431	165	59	ADJ
INT	8	6	6	ADV
KD	16	9	6	Q
KER	77	28	19	P
NN	7642	548	218	PN
OR	75	24	9	Q
PD	205	55	12	PP
PN	2246	385	264	NN
PP	239	51	11	PD
Q	324	119	53	ADJ
QW	24	12	11	VB
RD	71	62	61	RP
RP	11	5	2	NN
U	24	8	8	NN

Table 12. Major misclassifications in merged corpus with T3 using TnT tagger

Tag	Total tokens	Error	Maximum misclassification	
CVRP	77	24	15	PP
DM	242	77	58	PRP
DMRL	71	64	63	PRRL
INJ	8	6	6	RB
JJ	2510	547	376	NN
JJRP	18	4	4	JJ
NN	7830	589	234	NNP
NNP	2339	390	267	NN
OD	75	23	8	JJ
PRP	642	119	33	DM

8 Discussion and Conclusion

The current work looks at the existing tagsets of Urdu being used for tagging corpora and analyzes them from two perspectives. First, the tagsets are analyzed to see their linguistic level differences. Second, they are compared based on their inter-tag confusion after training with two different POS taggers. These analyses are used to derive a more robust tagset.

The results show that collapsing categories which are not syntactically motivated improves the tagging accuracy in general. Specifically, light and regular verbs are merged, because they may come in similar syntactic frames. Reduplicated categories are given the same category tag (instead of a special repetition tag). Units and dates are also not considered separately as the differences have been semantically motivated and they can be categorized with existing tags at syntactic level.

Though, the measuring unit is currently treated as a noun, it could be collapsed as an adjective as well. The difference is sometimes lexical, where *kilogram* is more adjectival, vs. *minute* is more nominal in nature in Urdu, though both are units.

NNP (Proper Noun) tag could also have been collapsed with NN (Common Noun), as Urdu does not make clear between them at syntactic level. However, these two tags are kept separate due to their cross-linguistic importance.

One may expect that extending the genre or domain of corpus reduces accuracy of tagging because of increase in the variety in the syntactic patterns and diverse use of lexical items. One may also expect more accuracy with increase in size. The current results show that effect on additional domain (when C1 and C2 are mixed) is more pronounced than the increase in size (from approximately 100k to 200k), reducing accuracy from 94.21% (T3 with C2) to 90.99% (T3 with C1 + C2). The increase in accuracy for T3 vs. T1 may be caused by reduced size of T3. However, the proposed reduction does not compromise the syntactic word level information, as the collapsed categories are where they were either semantically motivated or motivated due to phrasal level tags.

The work has been motivated to consolidate the existing Urdu corpora annotated with different tagsets. This consolidation will help build more robust computational models for Urdu.

References

Brants, T. 2000. TnT – A statistical part-of-speech tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000* Seattle, WA, USA.

Butt, M. 1995. *The structure of complex predicates in Urdu.* CSLI, USA. ISBN: 1881526585.

Haq, A,. 1987. اردو صرف و نحو Amjuman-e-Taraqqi Urdu.

Hardie, A. 2003. Developing a tag-set for automated part-of-speech tagging in Urdu. Archer, D, Rayson, P, Wilson, A, and McEnery, T (eds.) *Proceedings of the Corpus Linguistics 2003 conference. UCREL Technical Papers Volume 16. Department of Linguistics, Lancaster University, UK.*

Hussain, S. 2008. Resources for Urdu Language Processing. *The Proceedings of the 6th Workshop on Asian Language Resources, IJCNLP'08*, IIIT Hyderabad, India.

Ijaz, M. and Hussain, S. 2007. Corpus Based Urdu Lexicon Development. *The Proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan.*

Javed, I. 1981. نئی اردو قوائد Taraqqi Urdu Bureau, New Delhi, India.

Platts, J. 1909. *A grammar of the Hindustani or Urdu language.* Reprinted by Sang-e-Meel Publishers, Lahore, Pakistan.

Sajjad, H. 2007. Statistical Part of Speech Tagger for Urdu. Unpublished MS Thesis, National University of Computer and Emerging Sciences, Lahore, Pakistan.

Sajjad, H. and Schmid, H. 2009. Tagging Urdu Text with Parts Of Speech: A Tagger Comparison.*12th conference of the European chapter of the association for computational Linguistic*s

Santorini, B. 1990. Part_of_Speech Tagging Guidelines for the Penn Treebank Project (3rd printing, 2nd revision). Accessed from ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz on 3rd May, 2009.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany.

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

McEnery, A., Baker, J. Gaizauskas, R. and Cunningham, H. 2000. EMILLE: towards a corpus of South Asian languages, British Computing Society Machine Translation Specialist Group, London, UK.

Appendix A: Mappings of tags between Tagsets T1 and T2.

	Tagset T1	Tagset T2
1.	A	JJRP
2.	AA	AUXA
3.	ADJ	JJ
4.	ADV	RB
5.	CA	CD
6.	CC	CC
7.	DATE	DATE
8.	EXP	SYM
9.	FR	FR
10.	G	PRP$
11.	GR	PRRFP$
12.	I	ITRP
13.	INT	INJ
14.	KER	KER
15.	MUL	MUL
16.	NN	NN
17.	OR	OD
18.	P	CM
19.	PD	DM
20.	PM	PM
21.	PN	NNP
22.	PP	PR
23.	Q	Q
24.	QW	QW
25.	RD	DMRL
26.	REP	PRRL
27.	RP	PRRF
28.	SC	SC
29.	SE	RBRP
30.	SM	SM
31.	TA	AUXT
32.	U	U
33.	WALA	WALA

Appendix B: New Tagset T3.

	Tag	Meaning	Example	
1.	AUX	Auxiliary	منتقل کر **سکتے** ہو	May
2.	CC	Coordinate Conjunction	ملازمین **یا** حکومتی عہدہ داروں کے ذریعے	Or
3.	CD	Cardinal	**ایک** موجودہ اداکار کو	One
4.	CVRP	Conjunctive Verb Particle	فرانسیسی قلعے بیچ **کر** بھی فنڈز بڑھانے پر راضی نہ	After
5.	DM	Demonstrative	پہلے **ایسے** واقعات نہ ہونے کے برابر تھے	Like this
6.	DMRL	Demonstrative Relative	وہ اشاعتی ادارہ سے **جو** وہ 23 سال تک چلا چکے ہیں	That
7.	FR	Fraction	**آدھے** گھنٹے میں	Half
8.	INJ	Interjection	**واہ!** کیا بات ہے	Hurrah
9.	ITRP	Intensive Particle	نہ گہرا تھا نہ **ہی** باقی رہنے والا	Too
10.	JJ	Adjective	**بلند تر** لاگتوں کے ساتھ	Taller
11.	JJRP	Adjective Particle	باہر رہنے کی بہت **سی** وجوہات کو سوچ سکتے ہیں	As
12.	MRP	Multiplicative Particle	**دگنی** رقم	Double
13.	NN	Noun	**سال** کے **آغاز** میں **افواہوں** پر	Year
14.	NNP	Proper Noun	**رابرٹ** نے کہا	Robert
15.	OD	Ordinal	**پہلا** ریٹائرمنٹ منصوبہ	First
16.	PM	Phrase Marker	**،**	,
17.	PP	Postposition	بورڈ رکنیت نو تک بڑھاتے ہوئے	To
18.	PRP	Pronoun Personal	**وہ** طریق کار کو استعمال کے اہل ہونا پسند کریں گے	They
19.	PRP$	Pronoun Personal Possessive	**میری** تیز گیند اچھی ہے	My
20.	PRRF	Pronoun Reflexive	کمپنی نے اپنے **آپ** کو بخوبی	Oneself
21.	PRRF$	Pronoun Reflexive Possessive	**اپنے** اجتماعی دفاتر	Own
22.	PRRL	Pronoun Relative	وہ اشاعتی ادارہ سے **جو** وہ 23 سال تک چلا چکے ہیں	That
23.	Q	Quantitative	**چند** لوگ	Some
24.	QW	Question Word	ایک مصنف **کیوں** یقین کرے گا	Why
25.	RB	Adverb	**ہمیشہ** بیچی گئی	Always
26.	SC	Subordinate Conjunction	کتنا رکھے گی **کیونکہ** کچھ نوکریاں	Because
27.	SM	Sentence Marker	**؟**	?
28.	SYM	any Symbol	**$**	$
29.	VB	Verb	مہنگے کپڑے **چاہتے** تھے	Wanted
30.	VBI	Verb Infinitive form	اسے لے **جانے** کے لیے	To go
31.	VBT	Verb Tense	تصور قابل عمل **ہے**	Is
32.	WALA	Association Marking Morpheme	رکھنے **والے** جاری کرنے **والا**	Associated Bearing