The CLE Urdu POS Tagset

[2]Tafseer Ahmed, [1]Saba Urooj, [1]Sarmad Hussain, [1]Asad Mustafa, [1]Rahila Parveen, [1]Farah Adeeba, , [3]Annette Hautli & [3]Miriam Butt

[1]Centre for Language Engineering, Al-Khawarizmi Institute of Compute Science, UET, Lahore, Pakistan
firstname.lastname@kics.edu.pk

[2]DHA Suffa University, Karachi, Pakistan
tafseer@dsu.edu.pk

[3]Univ. of Konstanz, Konstanz, Germany
firstname.lastname@uni-konstanz.de

Abstract

The paper presents a design schema and details of a new Urdu POS tagset. This tagset is designed due to challenges encountered in working with existing tagsets for Urdu. It uses tags that judiciously incorporate information about special morpho-syntactic categories found in Urdu. With respect to the overall naming schema and the basic divisions, the tagset draws on the Penn Treebank and a Common Tagset for Indian Languages. The resulting CLE Urdu POS Tagset consists of 12 major categories with subdivisions, resulting in 32 tags. The tagset has been used to tag 100k words of the CLE Urdu Digest Corpus, giving a tagging accuracy of 96.8%.

Keywords: POS Tagset, Urdu, Corpus

1. Introduction

Choosing an appropriate tagset is a preliminary and vital task for successful POS tagging. A tagset needs to be able to encode the grammatical distinctions that are of interest for further steps in natural processing or for linguistic research, while allowing for efficient and accurate automatic tagging (MacKinlay, 2005). With respect to the South Asian language Urdu (spoken mainly in Pakistan and India), several different POS tagsets have already been developed. However, in the process of POS tagging the CLE Urdu Digest corpus, the only large generally available corpus for Urdu,[1] we identified several shortcomings with the existing POS tagsets and came to the conclusion that a new revised tagset needed to be designed to: (a) provide access to the kinds of linguistic distinctions we found necessary for further natural language processing such as grammar development, machine translation and generation; (b) improve the automatic tagging.

This paper discusses the existing tagsets for Urdu (Muaz, Ali & Hussain, 2009; Sajjad, 2007; Sajjad & Schmid, 2009; Schmid, 1995) and presents a new POS tagset that has been used to tag the CLE Urdu Digest Corpus.

2. Literature Review

POS tagsets have been reviewed and revised for a variety of languages due to a variety of motivations. Lüdeling & Kytö (2008) provides a detailed comparison of a range of English POS tagsets (including tagsets for the Brown, LOB, UPENN, BNC-C5, BNC-C6, ICE, PoW and LLc corpora) along with their differences. Lüdeling reports that these tagsets differ in accordance to the requirement of the target application of the tagged corpus as well as according to the underlying linguistic theory. For example, the ICE tagging scheme differs from other tagsets mainly due to the fact that it was developed at the time when syntactic theories like Generalized Phrase Structure Grammar and Lexical-Functional Grammar had proposed the notion that a category is composed of a bundle of features. Therefore, this tagging scheme was more useful for feature-based parsers.

It is not uncommon to experiment with different tagset designs and to repeatedly revise an existing tagset in order to capture typological properties in a more linguistically adequate and computationally efficient manner. Some examples come from work on Vietnamese (Tran et al., 2009), Slovene (Dzeroski, Erjavec & Zavrel, 2000), Swedish (Carlberger & Kann, 1999) and Persian (Oroumchian et al., 2006).

2.1 South Asian POS tagsets

With respect to South Asian languages, several different tagsets have been designed. These differ in terms of morpho-syntactic features, tag definition and tag granularity. However, South Asian languages form a common linguistic area and therefore share many structural characteristics. This realization is reflected in Baskaran et al. (2008), which contains a proposal for a framework that defines an overall common POS tagset for the languages of India (see also Chandrashekar (2007) on Sanskrit). The framework follows certain principles, i.e., a tagset should be hierarchically organized and include reference to morpho-syntactic features. Further, a balanced approach should be followed in using the form vs. function as criteria for the classification of tags. This framework ensures that common categories across Indian languages are annotated in the same way.

2.2 Urdu POS tagsets

The search for a good Urdu POS tagset has already gone through multiple iterations. In 2003, Hardie designed the first POS tagset for Urdu. He followed the EAGLES guidelines (Hardie, 2003). This tagset was based on morpho-syntactic categories of Urdu and contained 350 tags. As a large number of tags is difficult to handle for computational processing (with a small-sized corpus), there has been limited follow up work based on this tagset, beyond the initial POS tagger through the EMILLE project (Lüdeling & Kytö, 2008).

Sajjad (2007) & Sajjad & Schmid (2009) designed a

[1] See http://www.cle.org.pk/clestore/.

tagset consisting of 42 tags (after analyzing the grammatical categories based on grammars of Urdu) and the tagset proposed through the EMILLE project (Hardie, 2003). This tagset contains finer distinctive categories for pronouns and demonstratives, but does not do sufficient justice to the Urdu verbal and tense/aspect system.

In 2008, another tagset[2] was developed by the Center for Research on Urdu Language Processing (CRULP), following the guidelines of the Penn Treebank, which contains 46 tags. In this tagset, a verb category has multiple tags based on the morphology of the verbs. Similarly, common nouns were also classified with finer distinctions than previously available. Muaz, Ali & Hussain (2009) make a comparison of these tagsets and propose a new tagset with 32 tags. 17 tags are the same as in the previous two tagsets, but differences among types of nouns, for example, (with or without case, compounding) were eliminated as their syntactic distribution is identical.

3. Tagset Design

As part of a larger effort whose aim it is to develop and tag a balanced corpus of Urdu (Ijaz & Hussain, 2007; Urooj et al., 2012) for use in Urdu linguistic and computational research, a revision of (Muaz, Ali & Hussain, 2009), the most recent tagset, has been undertaken.

We analyzed design principles and individual tags of the currently available tagsets, and provide a new tagset which combines qualities of all of them. The new CLE Urdu POS Tagset is logically hierarchical i.e. it provides 12 primary POS categories and then 35 subcategories. For the design of individual tags, our primary inspiration is the tagset by Muaz, Ali & Hussain (2009). However, we added, deleted and merged different tags on the basis of: (a) comparison with other tagsets, (b) syntactic distribution and other linguistic issues (examples provided below in the discussion of the tags) and, (c) the tagging of 100K words of the CLE Urdu Digest balanced corpus. Mainly, we improved the tagset by proposing tags that are motivated by a readily identifiable morpho-syntactic pattern and distribution. The following is a brief description of the tags. The detailed tagset is available via the CLE website.[3]

3.1 Noun

Nouns are divided into two sub-categories, common noun (NN) and proper noun (NNP). We decided that a single POS tag will be assigned to multiwords and name entities. For example, "*islAm AbAd*[4]" (having a space or zero-width-joiner) is tagged as NNP.

Some canonical examples of common nouns are *kitAb* 'book', *pAnI* 'water' and *yAd* 'memory'. However, the category also includes other nouns that display an adverbial nature like time, place, manner, etc. Some examples of these are: *andar* 'inside' and *yahAN* 'here' etc. These adverbal nominals can occur with or without specifiers/modifiers.

(1)

vuh	***andar***	AI
PRP	***NN***	VB
3Sg	inside	come.Perf.Sg

'She came inside.'

(2)

vuh	[ghar	kE	***andar***]	AI
PRP	NN	PSP	***NN***	VB
3Sg	inside	of	inside	come

'She came inside the house.'

There was a disagreement in previous tagsets about these adverbial nominals. The Hindi/Indian language Tagset (Bharati et al., 2006) introduced a new tag category NST (Noun Spatial Temporal) for these words. The previous Urdu tagset (Muaz, Ali & Hussain, 2009) classifies these words as postpositions. We differ from both of these approaches for the following reason.

These words allow specifiers/modifiers (cf. example (2) above) and so are different from the case markers and simple postpositions that have a noun or pronoun preceding them. Hence, we do not classify adverbial nominals with the postpositions. The other choice was to create a separate noun (sub-)tag for these words. However, we found that their syntactic behavior is similar to that of common nouns. Hence, we did not create a new tag to cater to the semantic difference between two sets of words and instead subsumed these adverbial nominals under the common noun (NN) tag.

3.2 Pronoun

Pronouns are divided into 7 subcategories. The personal pronoun (PRP) appears as a replacement of the noun. Some examples are *meN* (1Sg.Nom/Erg), *mujHE* (1Sg.Acc/Dat), *vuh* (3Sg.Nom) and *usE* (3Sg.Acc/Dat). The demonstrative (PDM) appears before a noun as its specifier., as in (3).

(3)

vuh	laRkI	AI
PDM	NN	VB
3Sg	girl	come.Perf.F.Sg

'That girl came.'

Note that the same form *vuh* acts as personal pronoun (PRP) or demonstrative (PDM). They can be differentiated on the basis of syntactic context. In (6), *vuh* is the head of noun phrase, hence it is tagged as PRP. The possessive pronouns (PRS) are the pronouns used to show the relation of ownership. Some examples are *mErA* 'my', *tumhArA* 'your' and *hamArA* 'our'.

The reflexive pronouns (PRF) are used for referring to oneself. The examples are *xud* 'self' and *apnE Ap* 'self'. The reflexive apna (APNA) is used to show self's relation with the noun. An example is given in (4).

[2] See http://www.cle.org.pk/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm.

[3] See http://www.cle.org.pk/software/langproc/POStagset.htm.

[4] Urdu is written in a modified Persio-Arabic script. In this paper, we present a Latin script transliteration of the Urdu words . The transliteration scheme followed is described in http://www.lrec-conf.org/proceedings/lrec2010/pdf/194_Paper.pdf.

(4) mErI apnI gHaRI
PRS APNA NN
my own watch
'my own watch'

There are two separate subcategories for relative pronouns: Relative Personal (PRR) and Relative Demostrative (PRD). The syntactic behaviour of these pronouns is different from personal pronouns and demonstrative. The following example demonstrates the relative personal (PRR) *jo* 'who'.

(5) vuh laRkI jo AI
PRP NN PRR VB
3Sg girl who come.Perf.F.Sg
'The girl who came.'

It was discussed whether we should create separate categories for interrogative pronouns. We found that the interrogative pronoun can replace other related POS tags e.g. pronoun, adverb and quantifier etc. Hence no special tag for interrogative pronouns is created, and the interrogative words are merged into the relevant POS category. For example, *kon* 'who' is personal pronoun (PRP) and kitnA 'how much' is quantifier (Q).

3.3 Verb

Urdu verbs can be differentiated into canonical main verbs (6) light verbs appearing with a noun or adjective (7), and copular verbs (8).

(6) vuh AI
PRP VB
3Sg come.Perf.F.Sg
'She came.' (canonical main verb)

(7) usE [yAd AI]
PRP NN VB
3Sg memory come.Perf.F.Sg
'He/She remembered.' (noun + light verb)

(8) vuh xuS he
PRP JJ VB
3Sg happy be.Pres.3.M.Sg
'He/She is happy.' (copula verb)

We followed the decision of Muaz, Ali & Hussain (2009) and Bharati et al. (2006) and did not create a separate tag for these categories because all of these verbs show a similar syntactic behavior. However, we differ with the decision of Muaz, Ali & Hussain (2009) in which the copula 'be' is merged with the tense auxiliary simply because both have the same surface form (cf. (8) vs. (9)).

(9) vuh ghar AI he
PRP NN VB AUXT
3Sg home come be.Pres.3.M.Sg
'She came home.'

In (9), *he* comes after the main verb and expresses tense information, hence it is a tense auxiliary. However, in (8) it is functioning as a main verb. For this reason, it is tagged as VB.

There are different morphological forms of Urdu verbs. The root *A* 'come' has the morphological forms *A-tE* (imperfective masculine plural), *A-tI* (imperfective feminine singular), *A-ON* (subjunctive first person singular) etc. Unlike Hardie (2003) and following Muaz, Ali & Hussain (2009) and Bharati et al. (2006), we do not create separate tags to encode morphological information. There is a single tag VB for all forms of Urdu main verbs.

However, there is an exception to this rule. The verb in the infinitive form is tagged as VBI. We provide a special tag for verbal infinitives because these act as verbal nouns and therefore display a syntactic distribution that differs from that of main verbs. We have also found that we would have liked to have been able to conduct a targeted extraction of instances of verbal infinitives in our previous work within Urdu NLP. This has not been possible with existing tagsets.

(10) sigrET pInA burA he
NN VBI JJ VB
cigarette drink.Inf.M.Sg bad be.Pres
'Smoking cigarettes is bad.'

3.4 Auxiliary

The tagset encodes the fine distinctions necessary for the complex nature the verbal complex in Urdu. There are 4 types of auxiliaries; Aspectual (AUXA), Progressive (AUXP), Tense (AUXT) and Modals (AUXM). An example of a tense auxiliary (AUXT) is given in (9). The examples of the other tags are as follows:

(11) vuh ghar A rahI he
PRP NN VB AUXP AUXT
3Sg home come prog be.Pres
'She is coming home.'

(12) vuh ghar A saktI he
PRP NN VB AUXM AUXT
3Sg home come can be.Pres
'She can come home.'

(13) vuh ghar A gaI
PRP NN VB AUXA
3Sg home come completion
'She came home.'

(14) kitAb paRhI gaI
PRP VB AUXA
3Sg read.Perf.F.Sg passive
'A/the book was read.'

3.5 Nominal Modifiers

Nominal modifiers convey information about a noun. This include adjectives (JJ) e.g. *accHA* 'good', quantifiers (Q) e.g. *kucH* 'some', cardinal (CD) e.g. dO 'two', ordinal (OD) e.g. *dUsrA* 'second', fraction (FR) e.g. *AdHA* 'half' and multiplicative (QM) e.g. *gunA* 'times'.

We found that there are many adjectives that also appear as a noun. We decided to assign the POS according to the syntactical function. For example, GulAm 'slave' appears as an adjective in (15) and as a noun in (16).

(15) GulAm mulk
JJ NN
slave country
'slave country'

(16) GulAm AyA
N VB
slave come.Perf.M.Sg
'The slave come.'

As discussed in section 3.1, we consider multiwords as a single token. The superlative and comparative forms of some borrowed adjectives have Persian suffixes *tarIn* and *tar* respectively. A space occurs between the adjective and the suffix e.g "*AzIm tar*" 'greater'and "sust tarIn" 'slowest'. We consider these as multiwords and assign the tag JJ.

3.6 Adverb

There are two sub-categories of adverbs: general adverb (RB) and negation (NEG). The adverbs expressing negative e.g. *nahIN*, *na*, *mat* are tagged as NEG. The negatives have a different (more restricted) syntactic distribution than other adverbs and have therefore received a special tag. Other adverbs e.g. manner adverbs are tagged as RB. The examples are given below.

(17) vuh AhistA calI
PRP RB VB
3Sg home walk.Perf.3.F.Sg
'She walked slowly.'

(18) vuh nahIn AI
PRP NEG VB
3Sg not come.Perf.3.F.Sg
'She did not come.'

We discussed in section 3.5 that spatial and temporal adverbials e.g. *andar* 'ínside', *ab* 'now', *kal* 'tomorrow' are tagged as common noun (NN) because of their syntactic behavior.

3.7 Adposition

There are two subcategories of adpositions:. re- and postpositions. Some examples of Urdu prepositions are: *fI* 'in'/'per', *az* 'from' , *sivAE* 'except' and *bajuz* 'except' etc. (Raza, 2011). An example with *fI* (borrowed from Arabic) is given below.

(19) 50 rupe [fI kilogram]
CD NN PRE NN
50 rupees per kilogram
'50 rupees per kilogran'

Examples of postpositions are nE (the ergative marker), kO (the accusative and dative), tak 'till', liE 'for' and bin 'without'. As discussed in section 3.1, we consider adverbial nominals e.g. andar 'inside', Upar 'above'/'over' etc. as common nouns.

3.8 Conjunction

The category conjunction is divided into the usual coordinate and subordinate conjunction, but also provides for two Urdu specific categories.
The examples of co-ordinating conjunction (CC) are *or* 'and' and *lEkin* 'but'/'however' etc. The examples of sub-ordinating conjunctions (SC) are *kiyUnkah* 'because' and *tO* 'then' etc. An example of a SC is given below.

(20) agar mahnat karO gE
SCP NN VB AUXT
If hard-work do.Sub. future

to kAmyAb ho gE
SC JJ VB AUXT
then successful be future
'if (you) will work hard then (you) will be successful.'

The above example have *agar* 'if' as pre-sentential (SCP). These words appear before the first clause in subordinating constructions.
Following Bharati et al. (2006), we introduced the tag subordinating-conjunction-*kar* (SCK) for the verb *kar*(/*kE*) 'do' appearing at the the end of embedded non-finite clauses. An example of this construction is given below.

(21) vuh [ghar bEc kar] AI
PRP NN VB SCK VB
3SG house sell do come.Perf
'She came after selling the house.'

3.9 Interjection

The interjection (INJ) normally occurs at the start of the sentence. It is kept as a separate category in the tagset. Some examples are *vAh* 'bravo'/'well done', *arE* 'O'/'hey' and *subh2An Allah* 'glory to Allah' etc. It is important to note that the multiword *subh2An Allah* gets a single tag INJ.

3.10 Particle

Particles are divided into two subcategories: a general particle tag (PRT) and a VALA tag for a language specific category ('the X one').
The general particle tag (PRT) includes emphatic particles e.g. *bHI* 'also' and *hI* 'even'.

(22) [vuh bHI] AE gI
PRP PRT VB AUXT
3SG too come.Perf future
'She too will come.'

The usages of the particle *vAl-* are described in detail in Muaz & Khan (2009). An example of is given below.

(23) sabzI valA
NN VALA
vegetable one
'The thing (e.g. meal) that has vegetables'/
'the person who sells vegetable.'

3.11 Symbol

Symbol has two categories: Punctuation (PU) and other symbols (SYM).

3.12 Residual

Residual contains one tag for Foreign Fragment (FF) covering all foreign language elements. This tag is assigned only in that situation when we cannot assign an Urdu POS tag to that word (or multiword). For example, *subh2An Allah* 'glory to Allah' is an Arabic fragment, but we assign the interjection tag (INJ) to it. Similarly, the English noun *book* in the following example is treated as noun because it has been absorbed into standard Urdu usage via intensive language contact with English.

(24) us nE buk paRHI
PRP PSP NN VB
3Sg Erg book read.Perf.F.Sg
'He/She read the book.

If we cannot assign an Urdu POS tag to a foreign fragment, then we consider it as a foreign fragment (FF).

4. Tagging the CLE Urdu Digest Corpus

The updated tagset was used to tag the CLE Urdu Digest Corpus, covering an 80% training corpus and a 20% testing corpus. The files were selected randomly. The Tree Tagger (Schmid, 1994; Schmid, 1995) was used for automatic tagging, with a machine learning technique of Decision Trees and smoothing technique of Class Equivalence. The results are given in table 1. It shows a tagging accuracy of 96.8%, indicating that our tagset is performing well.

5. Discussion and Conclusion

In analyzing the results of the tagger, it was observed that the tagger encounters problems in disambiguating between some particular pairs of tags.

While there are two tags for nouns (noun vs. proper noun), Urdu does not make a clear distributional distinction between these nouns. We have decided to nevertheless keep both tags since information about proper nouns is generally important for further natural language processing.

Nouns are confused with adjectives when they occur adjacent to one another. The same issue was found by Muaz, Ali & Hussain (2009).

Due to the fact that the postposition 'in' and the personal pronoun 'I' are written the same in Urdu (میں), the tagger confuses the two when they occur in syntactic positions where both options are possible. Similarly, the tagger finds the Urdu word وت 'to' confusing, as it can act both as a discourse particle and as introducing a subordinate clause.

On the other hand, the results of the newly added tag Foreign Fragment (FF) has shown a good accuracy as compared to the previous tagsets where this category was dealt with under expressions (Exp) (Sajjad, 2007, Sajjad & Schmid, 2009) or was ignored (Muaz, Ali & Hussain, 2009).

In conclusion, we have presented a new POS tagset for Urdu. It is based on a critical analysis of several previous iterations of tagset proposals and builds on these. The new CLE Urdu POS Tagset has been used to tag 100k words of the publically available balanced CLE Urdu Digest corpus. Work is continuing to extend the tagged corpus to 1 million words.

Tag	Total Tokens	Error	Error %	Maximum Misclassification	
VBF	2602	119	4.57	30	AUXT/NN
AUXA	760	102	13.42	98	VBF
PDM	428	77	17.99	69	PRP
PRP	1091	72	6.60	53	PDM
NN	6266	65	1.04	11	JJ
JJ	1820	54	2.97	30	NN
PSP	3844	53	1.38	30	PRP
SC	454	52	11.45	35	PRT
AUXT	704	43	6.11	28	AUXA
NNP	1014	40	3.94	37	NN
Q	291	20	6.87	15	NN
RB	462	19	4.11	9	NN
CC	502	17	3.39	6	NN
PRR	139	14	10.07	5	PRP
PRT	395	13	3.29	9	PSP
AUXP	121	9	7.44	7	VBF
PRS	115	7	6.09	6	PDM
AUXM	104	6	5.77	5	AUXA
INJ	17	6	35.29	6	NN
SCK	154	6	3.90	3	RB
SCP	65	6	9.23	5	SC
CD	622	4	0.64	2	PU
PU	2536	4	0.16	2	VBF
VBI	438	4	0.91	2	VBF
FF	72	3	4.17	3	PU
OD	150	3	2.00	2	CD
PRF	14	2	14.29	2	NN

Table 1: Results and Error Analysis

6. Acknowledgement

This work has been supported by a DAAD Research Grant, Essential Urdu Linguistic Resources.[5]

7. References

Baskaran S., Bali K., Bhattacharya T., Bhattacharyya P., Jha G. N., Rajendran S., Saravanan K., Sobha L. and Subbarao K. V. (2008). Designing a Common POS-

[5] See http://cle.org.pk/eulr/.

Tagset Framework for Indian Languages. In *Proceedings of the 6th Workshop on Asian Language Resources, 2008*.

Bharati A., Sangal R., Sharma D. M. and Bai L. (2006). *Anncorra: Annotating corpora guidelines for POS and chunk annotation for Indian languages*. LTRC-TR31.

Carlberger J. and Kann V. (1999). Implementing an efficient part-of-speech tagger. *Software-Practice and Experience*. pp. 815-832.

Chandrashekar R. (2007). POS tagger for Sanskrit. Ph.D. thesis. Jawaharlal Nehru University, New Delhi.

Dzeroski S., Erjavec T. and Zavrel J. (2000). Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. In *Proceedings of the Second International Conference on Language Resources and Evaluation, 2000*.

Hardie A. (2003). Developing a tag-set for automated part-of-speech tagging in Urdu. In Archer D., Rayson P., Wilson A., and McEnery T. (eds.), *Proceedings of the Corpus Linguistics 2003 Conference*.

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

Leech G. (1997). Grammatical Tagging. In Garsire R., Leech G. and McEnery A. (eds.), *Corpus Annotation: Linguistic Information for Computer Text Corpora*. Longman, London.

Lüdeling A., Kytö M. (eds.). (2008). *Corpus Linguistics: An International Handbook*. Berlin:Walter de Gruyter.

MacKinlay A. (2005). *The effects of part-of-speech tagsets on tagger performance*. Honours thesis. University of Melbourne.

Muaz A., Ali A. and Hussain S. (2009). Analysis and development of Urdu POS tagged corpus. In *Proceedings of the 7th Workshop on Asian Language Resources, IJCNLP'09*. Suntec City, Singapore.

Muaz A. and Khan A. N. (2009). The Morphosyntactic Behavior of 'Wala' in Urdu Language. In *Proceedings of 28th Annual Meeting of the South Asian Language Analysis Roundtable, SALA'09*. University of North Texas, US.

Oroumchian F., Tasharofi S., Amiri H., Hojjat H. and Raja F. (2006). *Creating a Feasible Corpus for Persian POS Tagging*. Department of Electrical and Computer Engineering, University of Tehran.

Raza G. (2011). *Subcategorization Acquisition and Classes of Predication in Urdu*. PhD Thesis. University of Konstanz. Germany.

Sajjad H. (2007). *Statistical Part of Speech Tagger for Urdu.* 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. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*,. Manchester, UK.

Schmid H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop,*. Dublin, Ireland.

Tran O. T., Le C. A., Ha T. Q. and Le Q. H. (2009). An Experimental Study on Vietnamese POS Tagging. In *Proceedings of Asian Language Processing,* pp. 23-27.

Urooj S., Hussain S., Adeeba F., Jabeen F. and Perveen R. (2012). CLE Urdu Digest Corpus. In *Proceedings of Conference on Language and Technology 2012 (CLT12)*. Lahore, Pakistan.