Semi-Semantic Part of Speech Annotation and Evaluation

Qaiser Abbas
Fachbereich Sprachwissenschaft
Universität Konstanz
78457 Konstanz, Germany
qaiser.abbas@uni-konstanz.de

Abstract

This paper presents the semi-semantic part of speech annotation and its evaluation via Krippendorff's α for the URDU.KON-TB treebank developed for the South Asian language Urdu. The part of speech annotation with the additional subcategories of morphology and semantics provides a treebank with sufficient encoded information. The corpus used is collected from the Urdu Wikipedia and news papers. The sentences were annotated manually to ensure a high annotational quality. The inter-annotator agreement obtained after evaluation is 0.964, which lies in the range of perfect agreement on a scale. Urdu is comparatively an under-resourced language and the development of the treebank with rich part of speech annotation will have significant impact on the state-of-the-art for Urdu language processing.

1 Introduction

Urdu, an invariant of *Hindavi* came into existence during the muslim rule from 1206 AD to 1858 AD (Khan, 2006). They used Persian/Urdu script for Urdu in contrast of the Devanagari script for Hindavi. Urdu became a literary language after existence of an increasing number of literature during 18th and 19th century (McLane, 1970). Hindi/Hindavi is a close language to Urdu except the script writing style and the differences in the formal and informal versions. Urdu is the national language of Pakistan and an official language in India. According to a report by SIL Ethnologue (Lewis, 2013), Urdu/Hindi has 456.2 million speakers in the whole world. Urdu is a morphologically rich language (MRL) and in need of a number of resources to compete in the race of computational resources.

The design of the part of speech (POS) annotation scheme depends upon the need. If the people want to do text processing, text mining, etc., then they might be interested in a limited POS annotation scheme. However, the people who are interested in language parsing, then a POS annotation scheme with rich information is needed. Getting state-of-the-art parsing results for a MRL is a challenge till to date. According to Tsarfaty et. al. (2013; 2010), without proper handling of morphological entities in the sentences, promising results for MRLs can not be achieved and the depth of information encoded in an annotation correlates with the parsing performance. The best broad coverage and robust parsers to date have grammars extracted from the treebanks, which are a collection of syntactically annotated sentences by humans. The problem statement described requires an explicit encoding of morphological information at the POS level and the treebanks with sufficient encoding of morphology, POS, syntactic and functional information are the best candidates to provide the state-of-the-art parsing results in case of MRLs. The work presented here is the part of a large effort made for the construction of the URDU.KON-TB treebank, which was built by considering the parsing needs of Urdu. The annotation scheme of the treebank contains semi-semantic POS (SSP), semi-semantic syntactic (SSS) and functional (F) tag sets, from which only the SSP tag set is presented here along with its annotation evaluation.

The relevant resources of Urdu are now growing but most of the resources lack in morphological and functional information. The initial corpus developed in the EMILLE project (McEnery et al., 2000) comprised multi-lingual corpora for the South Asian languages. Its Urdu part was annotated according to a

LAW VIII - The 8th Linguistic Annotation Workshop, pages 75–81,
Dublin, Ireland, August 23-24 2014.

POS annotation scheme devised by Hardie (2003), which contained 350 morpho-syntactic tags based on the gender, number agreement. It was so detailed that the Urdu computational linguists avoided it to practice in statistical parsing, even it was a good effort. However, now the computational linguists are realizing and attempting morphological information in their annotation (Manning, 2011). In (2007), Urdu ParGram project introduced a resource that lied in the domain of tree-banking. In this project, Urdu lexical functional grammar (LFG) was encoded, which is still in progress. The LFG grammar encoded has rich morphological information, but unfortunately, the annotation scheme is not published yet due to their different motives towards the parallel treebank development. Similarly, in (2009), Sajjad and Schmid presented a new POS annotation scheme, which lacks in morphological, syntactical and functional information. Due to which, it can only be used for the training of POS taggers and is not suitable for the parsing purpose. Moreover, the explicit annotation evaluation was not performed. Another POS tag set was devised by Muaz et. al. in (2009), which contained 32 general POS tags. The devised scheme has the same issues as mentioned in the work of Sajjad and Schmid (2009). In (2009), Abbas et. al. built the first NU-FAST treebank for Urdu with the POS and syntactic labels only. The design of that treebank neither contained detailed morphological and functional information nor any information about the displaced constituents, empty arguments, etc. Another Hindi-Urdu tree-banking (HUTB) (Bhatt et al., 2009; Palmer et al., 2009) effort was done in a collaborative project[1]. However, the Urdu treebank being developed was comparatively small and was being done as a part of a larger effort at establishing a treebank for Hindi. Moreover, many of the issues with respect to Urdu were not quite addressed and the project is still in progress. To continue this effort, another treebank for Urdu was designed by Abbas in (2012), which comprised of 600 annotated sentences and it was done without the annotation evaluation.

The current work presented in this paper, not only enhances the size of the proposed treebank by Abbas (2012), but also resolves the annotation issues along with the complete annotation guidelines and its evaluation. The development of the URDU.KON-TB treebank starts with the collection of a corpus discussed briefly in Section 2. The semi-semantic (partly or partially semantic) POS (SSP) annotation scheme is described in Section 3. Similarly, the evaluation of the SSP annotation is presented in Section 4 along with a brief presentation of annotation issues. Finally, the conclusion is given in Section 5 and the detailed version of the SSP tag set is given in Appendix.

2 Corpus Collection

One thousand (1000) sentences taken from the corpus (Ijaz and Hussain, 2007) are extensively modified to get rid of licensing constraints, because we want to share our corpus freely under a Creative-Commons-Attribution/Share-Alike License 3.0 or higher. The next four hundred (400) sentences are collected from the Urdu Wikipedia[2], which is already under the same license. Thus the size of the corpus is limited to fourteen hundred (1400) sentences. The corpus contains text of local & international news, social stories, sports, culture, finance, history, religion, traveling, etc.

3 Semi-Semantic POS (SSP) Annotation

After the annotation evaluation presented in Section 4, the revised annotation scheme of the URDU.KON-TB treebank has a semi-semantic POS (SSP), semi-semantic syntactic (SSS) and a functional (F) tag set. The term semi-semantic (partly or partially semantic) is used with the POS because the tags are compounded with the semantic tags partially e.g. a noun *house* with spatial semantics tagged as N.SPT, an adjective *previous* in the *previous year* with temporal semantics tagged as ADJ.TMP, etc. The same concept is applied on the SSS annotation. The details of SSS and F labeling is beyond the scope of this paper. At POS level, a dot '.' is used to add morphological and semantical subcategories into the main POS categories displayed in Table 1 of Appendix. The POS, morphological and semantical information all together, make a rich SSP annotation scheme for the URDU.KON-TB treebank. The need for such type of schemes is highly advocated in (Clark et al., 2010; Skut et al., 1997), etc.

[1] `http://verbs.colorado.edu/hindiurdu/`

[2] http://ur.wikipedia.org/wiki/صفحۂ اول

A simple POS tag set was devised first, which had twenty two (22) main POS-tag categories described in Table 1 of Appendix, which includes some non-familiar tags like HADEES and M to represent the Arabic statements of prophets in Urdu text and a phrase or a sentence marker, respectively. The labels for morphological and semantic subcategories are presented in Tables 2 and 3 of Appendix, respectively, which can be added to the 22 main POS tag categories by using a dot '.' symbol in the form of compound tags like N.SPT and ADJ.TMP mentioned earlier. In case of morphology, if a verb V has a perfective morphology, then the compound tag becomes V.PERF. The SSP tag set was refined during the manual annotation process of the sentences and further refined after the annotation evaluation process discussed in Section 4. The final refined form of the SSP tag set depicted in Table 4 of Appendix is the revised form of the POS tag set presented in the initial version of the URDU.KON-TB treebank by Abbas in (2012).

As an example, consider the ADJ (adjective) from the final refined form of the SSP tag set given in Appendix, which is divided into five subcategories of tags DEG (Degree), ECO (Echo), MNR (Manner), SPT (Spatial) and TMP (Temporal). Relevant examples are provided in 1 of Appendix. The example 1(a) of Appendix is a simple case of ADJ, while 1(b) of Appendix is the case of a degree adjective[3] annotated with ADJ.DEG. The example 1(c) of Appendix is the case of reduplication[4] (Abbi, 1992; Bögel et al., 2007). Reduplication has two versions. First *Echo Reduplication* is discussed in the footnote, while the other *Full Word Reduplication* is the repetition of the original word e.g. *sAtH sAtH* 'with/alongwith'. These are adopted in our annotation as ECO (echo) and the REP (repetition), respectively. The example 1(d) of Appendix is the case of adjective having a sense of manner annotated as ADJ.MNR. If an adjective qualifies an action noun, then a sense of action or something is produced, whose behavior or the way to do that action is exploited through ADJ.MNR e.g. *z4AlemAnah t2abdIlIyAN* 'brutal changes'. An exercise of manner adjectives and manner adverbs for English can be seen at Cambridge University[5]. The example 1(e) of Appendix is the case of an adjective having a temporal sense discussed earlier. Finally, the example 1(f) of Appendix is the case of an adjective having a spatial sense. The adjective used here is the derivational form of a city name 'Multan', but it appears here as an adjective and annotated as ADJ.SPT[6] like in this sentence e.g. *voh Ek pAkistAnI laRkA hE* 'He is a pakistani boy'.

Example 1 of Appendix exploited the POS tags for adjectives along with the semantic tagging like TMP, SPT, MNR, etc. However, to give an introduction about morphology and verb functions, another POS category of verb V given in Appendix is presented. It is divided into 11 subcategories, which include COP (copula verb), IMPERF (imperfective morphological form of verb), INF (infinitive form of verb), LIGHT (1st light verb with nouns and adjectives), LIGHTV (2nd light verb with verbs), MOD (modal verb), PERF (perfective morphology), ROOT (root form), SUBTV (subjunctive form), PAST (past tense of a verb) and PRES (present tense of a verb). These tags have further subcategories. All tags represents different morphological forms and the function of a verb that it governs. A few high quality studies were adopted to identify different forms and functions of Urdu verbs (Butt, 2003; Butt, 1995; Butt and Rizvi, 2010; Butt and Ramchand, 2001; Butt, 2010; Abbas and Raza, 2014; Abbas and Nabi Khan, 2009) and some annotated sentences from the URDU.KON-TB treebank are given in example 2 of Appendix.

The sentence in example 2(a) of Appendix is the case of adjective-verb complex verb predicate. These adjective/noun-verb complex predicates were first proposed by Ahmed and Butt (2011). The adjective *dubHar* 'hard' and the verb *kiyA* 'did' with a perfective morphology *yA* at the end are annotated as a ADJ and a V.LIGHT.PERF, respectively. Similarly, a perfective verb *liyA* 'took' after a root form of verb *kar* 'do' is an example of the verb-verb complex predicate depicted in 2(d) of Appendix. This construction is adopted from the studies given in (Butt, 2010). The next sentence in 2(b) of Appendix has a passive construction, which can be inferred from the inflected form of a verb or a verb auxiliary *jAnA* 'to go' preceded by another verb with perfective morphology. To explore some unusual tags, a long sentence

[3] This division is used to represent absolute, comparative and superlative degree in adjectives and adverbs.

[4] In Urdu like other South Asian languages, the reduplication of a content word is frequent. Its effect is only to strengthen the proceeding word or to expand the specific idea of a proceeding word into a general form e.g. *kAm THIk-THAk karnA* 'Do the work right' or *kOI kapRE-vapRE dE dO* 'Give me the clothes or something like those'.

[5] `http://www.cambridge.org/grammarandbeyond/wp-content/uploads/2012/09/Communicative_Activity_Hi-BegIntermediate-Adjectives_and_Adverbs.pdf`

[6] Spatial adjectives are used to describe a place/location, direction or distance e.g. *multAnI* 'Multani', *aglI* 'next', and *dUr* 'far' respectively.

is presented in 2(c) of Appendix. After the name of prophets or righteous religious-personalities, some specific and limited prayers called *s3alAvAt* 'prayers' like *sal-lal-la-ho-a2lEhE-va-AlEhI-salam* 'May Allah grant peace and honor on him and his family', *a2lEh salAm* 'peace be upon him', etc., in Arabic is the most likely in Urdu text and annotated as the PRAY. Similarly, the statements of prophet Muhammad (PBUH) known as *h2adIs2* 'narration' like *In-namal-aa2mAlo-bin-niyAt* 'The deeds are considered by the intensions' in Arabic script is also a tradition in Urdu text and annotated as the HADEES. The phrase markers like comma, double quotes, single quotes, etc. are annotated with the M.P and sentence marker like full-stop, question mark, etc., are annotated with the M.S as presented in the same example.

4 SSP Annotation Evaluation

The SSP annotation evaluation was performed via Krippendorff's α coefficient (Krippendorff, 2004), which is a statistical measure to evaluate the reliability annotation or the inter-annotator agreement (IAA). Krippendorff's α (Krippendorff, 1970; Krippendorff, 2004) satisfies all our needs including random nominal data and five number of annotators in contrast to multi-π (Fleiss, 1971) and multi-κ (Cohen and others, 1960), which can handle only fixed nominal data and they are basically not designed for more than two annotators (Artstein and Poesio, 2008; Carletta et al., 1997). The nominal data given to annotators for the SSP annotation was not fixed. In this situation, the general form of the Krippendorff's α coefficient was selected to meet this requirement.

For the reliability evaluation of the SSP annotation guidelines, it was essential that the annotators should be the native speakers of Urdu along with the linguistics skills. To fulfill this purpose, an undergraduate class of 25 linguistic students was trained at the Department of English, University of Sargodha[7], Pakistan. During this training, thirty two lectures on annotation guidelines with practical sessions were delivered. The duration of each lecture was of 3 hours. The class was further divided into five groups and during their initial practical sessions, one student with a high caliber of understanding from each group was selected (but not informed) secretly for the final annotation. The annotation task of 100 random sentences was divided into 10 home assignments, which were then given to all students (including 5 secret students) periodically with an instruction not to discuss it with each other. The annotation performed by the selected 5 students was then recorded and evaluated. The value of α coefficient obtained after evaluation is 0.964 for the SSP annotation, which is narrated as a good reliability in (Krippendorff, 2004) and lies in the category of perfect agreement according to a scale in (Landis and Koch, 1977). It also means that the IAA is 0.964 and the SSP annotation guidelines are reliable.

The issues found before and after the annotation evaluation concludes the addition, deletion or revision of several tags. For example, the continuous auxiliary *rahA*/VAUX.PROG.PERF and its inflected forms can behave as a copula verb as V.COP.PERF, which was not considered in the initial work. The annotators did not respond well during the annotation of complex predicates, so their identification rules are revised which includes tense, passive, modal, etc., auxiliaries or verbs can not behave as complex predicate e.g. VAUX.LIGHT.MOD is not possible in the updated version. Similarly, the KER tag for identification of a special clause ending with *kar/V.KER kE/KER* 'after doing', was found to be ambiguous and deleted. It was updated with their genuine tags as *kar/V.ROOT kE/CM*.

5 Conclusion

Sufficient rich information in the SSP annotation was encoded to meet the parsing needs of MRL Urdu. The α coefficient value obtained advocates the quality of the SSP annotation along with the complete annotation guidelines for the URDU.KON-TB treebank. Such kind of annotated corpus with rich morphology and semantics is not only useful for the parsing purpose but can be used for the training of POS taggers, text mining, language identification (Abbas et al., 2010) and in many other applications as well.

Acknowledgements

Special thanks to Professor Miriam Butt, due to which linguistically rich the URDU.KON-TB treebank is existing and she participated overwhelmingly in the annotation process of this treebank.

[7] http://uos.edu.pk/

References

Qaiser Abbas and A Nabi Khan. 2009. Lexical Functional Grammar For Urdu Modal Verbs. In *Emerging Technologies, 2009. ICET 2009. International Conference on*, pages 7–12. IEEE.

Qaiser Abbas and Ghulam Raza. 2014. A Computational Classification Of Urdu Dynamic Copula Verb. *International Journal of Computer Applications*, 85(10):1–12, January.

Qaiser Abbas, Nayyara Karamat, and Sadia Niazi. 2009. Development Of Tree-Bank Based Probabilistic Grammar For Urdu Language. *International Journal of Electrical & Computer Science*, 9(09):231–235.

Qaiser Abbas, MS Ahmed, and Sadia Niazi. 2010. Language Identifier For Languages Of Pakistan Including Arabic And Persian. *International Journal of Computational Linguistics (IJCL)*, 1(03):27–35.

Qaiser Abbas. 2012. Building A Hierarchical Annotated Corpus Of Urdu: The URDU.KON-TB Treebank. *Lecture Notes in Computer Science*, 7181(1):66–79.

Anvita Abbi. 1992. *Reduplication In South Asian Languages: An Areal, Typological, And Historical Study*. Allied Publishers New Delhi.

Tafseer Ahmed and Miriam Butt. 2011. Discovering Semantic Classes For Urdu NV Complex Predicates. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 305–309. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement For Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A Multi-Representational And Multi-Layered Treebank For Hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 186–189. Association for Computational Linguistics.

Tina Bögel, Miriam Butt, Annette Hautli, and Sebastian Sulger. 2007. Developing A Finite-State Morphological Analyzer For Urdu And Hindi. *Finite State Methods and Natural Language Processing*, page 86.

Miriam Butt and Tracy Holloway King. 2007. Urdu In A Parallel Grammar Development Environment. *Language Resources and Evaluation*, 41(2):191–207.

Miriam Butt and Gillian Ramchand. 2001. Complex Aspectual Structure In Hindi/Urdu. *M. Liakata, B. Jensen, & D. Maillat, Eds*, pages 1–30.

Miriam Butt and Jafar Rizvi. 2010. Tense And Aspect In Urdu. *Layers of Aspect. Stanford: CSLI Publications*.

Miriam Butt. 1995. *The Structure Of Complex Predicates In Urdu*. Center for the Study of Language (CSLI).

Miriam Butt. 2003. The Light Verb Jungle. In *Workshop on Multi-Verb Constructions*.

Miriam Butt. 2010. The Light Verb Jungle: Still Hacking Away. *Complex predicates: cross-linguistic perspectives on event structure*, page 48.

Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C Kowtko, and Anne H Anderson. 1997. The Reliability Of A Dialogue Structure Coding Scheme. *Computational linguistics*, 23(1):13–31.

Alexander Clark, Chris Fox, and Shalom Lappin. 2010. *The Handbook Of Computational Linguistics And Natural Language Processing*, volume 57. Wiley. com.

Jacob Cohen et al. 1960. A Coefficient Of Agreement For Nominal Scales. *Educational and psychological measurement*, 20(1):37–46.

Joseph L Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin*, 76(5):378.

Andrew Hardie. 2003. Developing A Tagset For Automated Part-Of-Speech Tagging In Urdu. In *Corpus Linguistics 2003*.

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

Abdul Jamil Khan. 2006. *Urdu/Hindi: An Artificial Divide: African Heritage, Mesopotamian Roots, Indian Culture & Britiah Colonialism*. Algora Pub.

Klaus Krippendorff. 1970. Estimating The Reliability, Systematic Error And Random Error Of Interval Data. *Educational and Psychological Measurement*, 30(1):61–70.

Klaus Krippendorff. 2004. Reliability In Content Analysis. *Human Communication Research*, 30(3):411–433.

J Richard Landis and Gary G Koch. 1977. The Measurement Of Observer Agreement For Categorical Data. *biometrics*, pages 159–174.

Gary F. Simons & Charles D. Fennig Lewis, M. Paul. 2013. *Ethnologue: Languages Of The World, 17th Edition*. Dallas: SIL International.

Christopher D Manning. 2011. Part-Of-Speech Tagging From 97% To 100%: Is It Time For Some Linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189. Springer.

Anthony McEnery, Paul Baker, Rob Gaizauskas, and Hamish Cunningham. 2000. EMILLE: Building A Corpus Of South Asian Languages. *VIVEK-BOMBAY-*, 13(3):22–28.

John R McLane. 1970. *The Political Awakening In India*. Prentice Hall.

Ahmed Muaz, Aasim Ali, and Sarmad Hussain. 2009. Analysis And Development Of Urdu POS Tagged Corpus. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 24–29. Association for Computational Linguistics.

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi Syntax: Annotating Dependency, Lexical Predicate-Argument Structure, And Phrase Structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

Hassan Sajjad and Helmut Schmid. 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*, pages 692–700. Association for Computational Linguistics.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An Annotation Scheme For Free Word Order Languages. In *Proceedings of the fifth conference on Applied natural language processing*, pages 88–95. Association for Computational Linguistics.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. 2010. Statistical Parsing Of Morphologically Rich Languages (SPMRL): What, How And Whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12. Association for Computational Linguistics.

Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. 2013. Parsing Morphologically Rich Languages: Introduction To The Special Issue. *Computational Linguistics*, 39(1):15–22.

Appendix

(1)

(a)

acHA	*laRkA*
ADJ	N

‘good boy’

(b)

aham	*tarIn*	*Saxs2iat*
ADJ	ADJ.DEG	N

‘most important personality’

(c)

burA	*vurA*	*kAm*
ADJ	ADJ.ECO	N

‘ugly work’

(d)

jaberaanah	*hakUmat*
ADJ.MNR	N

‘forceful government’

(e)

guzaStah	*sAl*
ADJ.TMP	N

‘previous year’

(f)

mUltAnI	*kHUsah*
ADJ.SPT	N

‘multani shoe’

(2)

(a)

mehangAI	*nE*	*lOgON*	*kA*	*jInA*	*dUbHar*	*kiyA*	*tHA*
N	CM	N	CM	N	ADJ	V.LIGHT.PERF	VAUX.PAST

‘The inflation had made the life of people hard’

(b)

giraN-faroSoN	*kE*	*xilAf*	*qAnUn*	*harkat*	*mEN*	*lAyA*	*jAyE*
N	CM	POSTP.MNR	N	N	CM	V.PERF	VAUX.PASS.SUBTV

‘The law would be practiced against inflators’

(c)

mUhammad	*sal-lal-la-ho-a2lEhE-va-AlEhI-salam*	*nE*	*farmAyA*	*keh*	“
N.PROP	PRAY	CM	V.PERF	C.SBORD	M.P

al-hUsynON-mInnI-vA-anA-mInal-hUsyn	”	*ya2nI*	‘	*hUsyn*	*mUjH*	*sE*	*hE*
HADEES	M.P	ADV	M.P	N.PROP	P.PERS	CM	V.COP.PRES

aOr	*mEN*	*hUsyn*	*sE*	*hUN*	’	.
C.CORD	P.PERS	N.PROP	CM	V.SUBTV	M.P	M.S

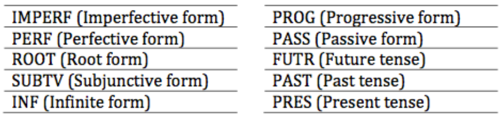
‘Muhammad (May Allah grant peace and honor on him and his family) said that “al-hUsynON-mInnI-vA-anA-mInal-hUsyn” means ‘Hussain is from me and I am from Hussain’ . ’

(d)

tUm	*nE*	*haj*	*tO*	*kar*	*liyA*	*hO*	*gA*	*?*
P.PERS	CM	N	PT.EMP	V.ROOT	V.LIGHTV.PERF	VAUX.SUBTV	VAUX.FUTR	M.S

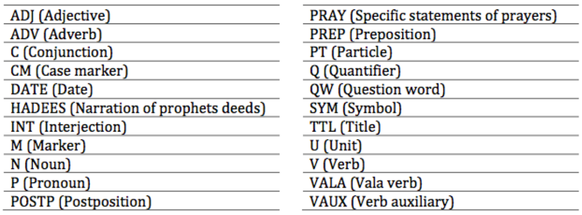
‘You will have made the pilgrimage?’

Table 2: Morphological tag set subcategories



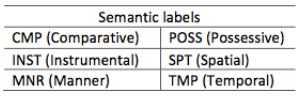
IMPERF (Imperfective form)	PROG (Progressive form)
PERF (Perfective form)	PASS (Passive form)
ROOT (Root form)	FUTR (Future tense)
SUBTV (Subjunctive form)	PAST (Past tense)
INF (Infinite form)	PRES (Present tense)

Table 1: The main POS-Tag categories



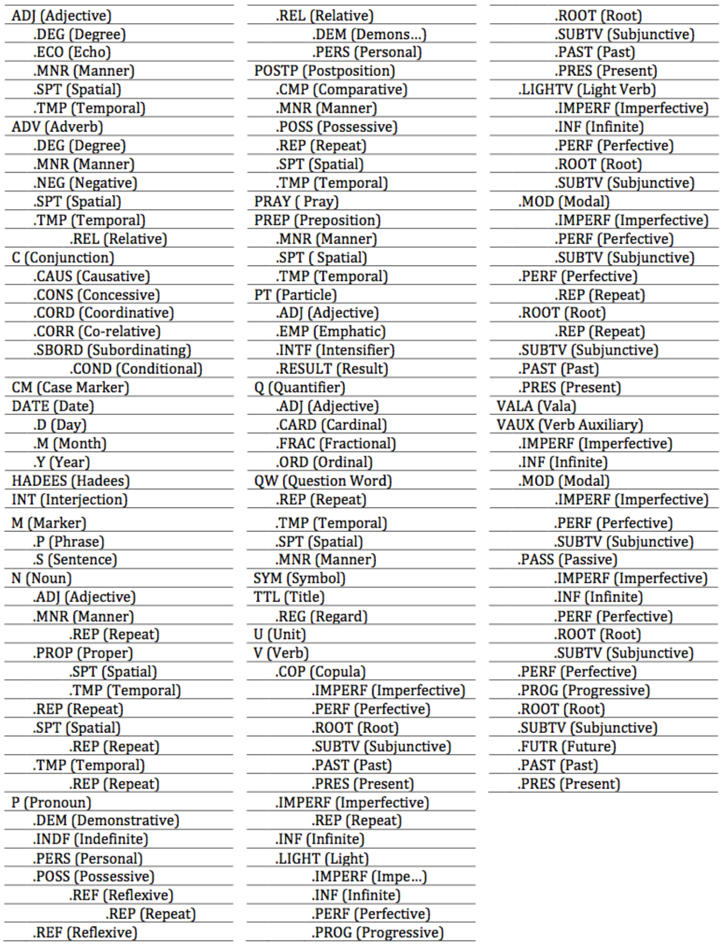
ADJ (Adjective)	PRAY (Specific statements of prayers)
ADV (Adverb)	PREP (Preposition)
C (Conjunction)	PT (Particle)
CM (Case marker)	Q (Quantifier)
DATE (Date)	QW (Question word)
HADEES (Narration of prophets deeds)	SYM (Symbol)
INT (Interjection)	TTL (Title)
M (Marker)	U (Unit)
N (Noun)	V (Verb)
P (Pronoun)	VALA (Vala verb)
POSTP (Postposition)	VAUX (Verb auxiliary)

Table 3: Semantical tagset.



Semantic labels	
CMP (Comparative)	POSS (Possessive)
INST (Instrumental)	SPT (Spatial)
MNR (Manner)	TMP (Temporal)

Table 4: A detailed version of the SSP tagset for the URDU.KON-TB treebank



ADJ (Adjective)	.REL (Relative)	.ROOT (Root)
.DEG (Degree)	.DEM (Demons...)	.SUBTV (Subjunctive)
.ECO (Echo)	.PERS (Personal)	.PAST (Past)
.MNR (Manner)	POSTP (Postposition)	.PRES (Present)
.SPT (Spatial)	.CMP (Comparative)	.LIGHTV (Light Verb)
.TMP (Temporal)	.MNR (Manner)	.IMPERF (Imperfective)
ADV (Adverb)	.POSS (Possessive)	.INF (Infinite)
.DEG (Degree)	.REP (Repeat)	.PERF (Perfective)
.MNR (Manner)	.SPT (Spatial)	.ROOT (Root)
.NEG (Negative)	.TMP (Temporal)	.SUBTV (Subjunctive)
.SPT (Spatial)	PRAY (Pray)	.MOD (Modal)
.TMP (Temporal)	PREP (Preposition)	.IMPERF (Imperfective)
.REL (Relative)	.MNR (Manner)	.PERF (Perfective)
C (Conjunction)	.SPT (Spatial)	.SUBTV (Subjunctive)
.CAUS (Causative)	.TMP (Temporal)	.PERF (Perfective)
.CONS (Concessive)	PT (Particle)	.REP (Repeat)
.CORD (Coordinative)	.ADJ (Adjective)	.ROOT (Root)
.CORR (Co-relative)	.EMP (Emphatic)	.REP (Repeat)
.SBORD (Subordinating)	.INTF (Intensifier)	.SUBTV (Subjunctive)
.COND (Conditional)	.RESULT (Result)	.PAST (Past)
CM (Case Marker)	Q (Quantifier)	.PRES (Present)
DATE (Date)	.ADJ (Adjective)	VALA (Vala)
.D (Day)	.CARD (Cardinal)	VAUX (Verb Auxiliary)
.M (Month)	.FRAC (Fractional)	.IMPERF (Imperfective)
.Y (Year)	.ORD (Ordinal)	.INF (Infinite)
HADEES (Hadees)	QW (Question Word)	.MOD (Modal)
INT (Interjection)	.REP (Repeat)	.IMPERF (Imperfective)
M (Marker)	.TMP (Temporal)	.PERF (Perfective)
.P (Phrase)	.SPT (Spatial)	.SUBTV (Subjunctive)
.S (Sentence)	.MNR (Manner)	.PASS (Passive)
N (Noun)	SYM (Symbol)	.IMPERF (Imperfective)
.ADJ (Adjective)	TTL (Title)	.INF (Infinite)
.MNR (Manner)	.REG (Regard)	.PERF (Perfective)
.REP (Repeat)	U (Unit)	.ROOT (Root)
.PROP (Proper)	V (Verb)	.SUBTV (Subjunctive)
.SPT (Spatial)	.COP (Copula)	.PERF (Perfective)
.TMP (Temporal)	.IMPERF (Imperfective)	.PROG (Progressive)
.REP (Repeat)	.PERF (Perfective)	.ROOT (Root)
.SPT (Spatial)	.ROOT (Root)	.SUBTV (Subjunctive)
.REP (Repeat)	.SUBTV (Subjunctive)	.FUTR (Future)
.TMP (Temporal)	.PAST (Past)	.PAST (Past)
.REP (Repeat)	.PRES (Present)	.PRES (Present)
P (Pronoun)	.IMPERF (Imperfective)	
.DEM (Demonstrative)	.REP (Repeat)	
.INDF (Indefinite)	.INF (Infinite)	
.PERS (Personal)	.LIGHT (Light)	
.POSS (Possessive)	.IMPERF (Impe...)	
.REF (Reflexive)	.INF (Infinite)	
.REP (Repeat)	.PERF (Perfective)	
.REF (Reflexive)	.PROG (Progressive)	