SEMI-SEMANTIC ANNOTATION: A GUIDELINE FOR THE URDU.KON-TB TREEBANK POS ANNOTATION

Qaiser ABBAS
University of Sargodha, Pakistan
qaiser.abbas@uos.edu.pk

Abstract

This work elaborates the semi-semantic part of speech annotation guidelines for the URDU.KON-TB treebank: an annotated corpus. A hierarchical annotation scheme was designed to label the part of speech and then applied on the corpus. This raw corpus was collected from the Urdu Wikipedia and the Jang newspaper and then annotated with the proposed semi-semantic part of speech labels. The corpus contains text of local & international news, social stories, sports, culture, finance, religion, traveling, etc. This exercise finally contributed a part of speech annotation to the URDU.KON-TB treebank. Twenty-two main part of speech categories are divided into subcategories, which conclude the morphological, and semantical information encoded in it. This article reports the annotation guidelines in major; however, it also briefs the development of the URDU.KON-TB treebank, which includes the raw corpus collection, designing & employment of annotation scheme and finally, its statistical evaluation and results. The guidelines presented will be useful for linguistic community to annotate sentences not only for the national language Urdu but for the other indigenous languages like Punjab, Sindhi, Pashto, etc. as well.

Keywords: semi-semantic part of speech; rich information; deep learning; parsing aid; linguistically motivated annotation; humanistic annotation

Povzetek

Rezultat tega dela so navodila za označevanje polsemantičnih besednih vrst v drevesnici URDU.KON-TB. Hierarhična označevalna shema je bila oblikovana z namenom, da razvrsti besedne vrste in jih kot take uporabi na korpusih. Tokratni koprus, ki je sestavljen iz strani Urdu Wikipedie in časopisa Jang, je bil označen s predlaganimi polsemantičnimi besednimi vrstami. Korpus vsebuje tekste lokalnih in mednarodnih novic, zgodbe s socialno temo, šport, kulturo, finance, vero, potovanja in druge teme. Uspešen poskusa označevanja je nadgradil drevesnico URDU.KON-TB. Dvaindvajset osnovnih besednih vrst je razdeljenih v podkategorije z morfološkimi in semantičnimi informacijami. Članek podaja jasne osnovne smernice označevanja. Hkrati ponuja kratek pregled razvoja drevesnice URDU.KON-TB, ki vsebuje zbirke surovih korpusov, oblikovanje in uporabo shem za označevanje ter nenazadnje tudi statistično oceno in rezultate. Predlagana navodila za označevanje so namenjena jezikovnim skupnostim, ki označujejo stavke tako v državnem jeziku Urdu kot tudi v drugih jezikih, kot so *Punjab*, *Sindhi* in drugih.

Ključne besede: polsemantična besedna vrsta; številne informacije; globoko učenje; pomoč pri razvrščanju; jezikoslovno utemeljeno označevanje; humanistično označevanje

Acta Linguistica Asiatica, 6(2), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.2.97-134

1 Introduction

A treebank or a parsed corpus is a text corpus of sentences annotated with a syntactic structure. Today, many natural language processing (NLP) and machine learning (ML) applications rely on treebanks. Treebanks are heavily used in corpus linguistics for investigating syntactic phenomena or in computational linguistics for training or testing parsers. The sentences in the treebank should be annotated according to a devised annotation scheme as presented in Figure 1 to 4 in our case. Annotation schemes can include the labeling to represent morphological forms, word classes, syntactic structures, semantics, grammatical arguments, co-references, etc. So, the corpus annotation is simply the addition of interpretative linguistic information to a corpus (Leech, 2005).

Annotation scheme that was used to develop the URDU.KON-TB treebank (Abbas, 2012, 2014a, 2014b) for the South Asian language Urdu is presented next with complete guidelines in Section 2. This annotation is actually encoded with the morphology, POS, syntactical and functional information including the handling of displaced constituents, empty categories, antecedents and anaphors, etc., but here only the case of semi-semantic part of speech (SSP) is discussed concisely. Such development of an annotation scheme is the fundamental step to build a treebank, for which the computational linguists then devise the annotation guidelines (Section 2), which is a compulsory part to build, and without which the annotation scheme has no worth at all. Annotation structure for the development of the URDU.KON-TB treebank has the combination of the PS (Phrase Structure) and the HDS (Hyper Dependency Structure) annotation detailed in Section 3.2. Annotation issues emerged during the development (Abbas, 2012) have been corrected in (Abbas, 2014a & 2014b) and the annotation guidelines presented in Section 2 is the most updated version. The corpus containing 1400 sentences[1] (discussed in Section 3.1) for the development of the URDU.KON-TB treebank was collected from the Urdu Wikipedia[2] and the Urdu Jang newspaper.[3]

[1] The size has been augmented recently to 2000 sentences by a student in her master's thesis "Annexing Treebank and the Urdu Parser"

[2] https://ur.wikipedia.org/wiki/صفحہ_اول

[3] http://jang.com.pk/index.html

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

Figure 1: A detailed version of the SSP tagset for the URDU.KON-TB treebank

The reliability of the treebank annotation or the annotation guidelines can be measured by calculating the agreement or the homogeneity among the annotators of the treebank. The reliability evaluation is a complex task for the treebank that contains rich information, but it is an essential part to play for the production of a quality treebank, so that the annotation can be readable. The annotation evaluation (Abbas,

2014a & 2014b) resolved most of our annotation issues except few. The guidelines of the URDU.KON-TB treebank are evaluated using a statistical measure known as the Krippendorf's α coefficient (Krippendorf, 2004). This can be used to evaluate the inter-annotator agreement (IAA). Randomly selected one hundred (100) sentences from the URDU.KON-TB treebank were given to five trained annotators for annotation. The annotated sentences then evaluated using the Krippendorf's α co-efficient. The α values of the IAA obtained for the part of speech (SSP) annotation is 0.964. The annotation guidelines were revised during and after this annotation evaluation. A little detailed presentation of evaluation is given in Section 4.

Section 2 describes the up to date annotation guidelines revised after the annotation evaluation (Abbas, 2014a & 2014b). Guidelines regarding the SSP annotation are detailed here in this article for easiness and simplicity along with their respective examples. To remain on track, the annotation tags are discussed according to the order of the SSP tags given in Figure 1. The discussion of the annotation guidelines is kept concise.

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 (Special Word Vala)
POSTP (Postposition)	VAUX (Verb auxiliary)

Figure 2: The main POS-tag categories for the URDU.KON-TB treebank

PROG (Progressive form)	PROG (Progressive form)
PASS (Passive form)	PASS (Passive form)
FUTR (Future tense)	FUTR (Future tense)
PAST (Past tense)	PAST (Past tense)
PRES (Present tense)	PRES (Present tense)

Figure 3: Morphological tag set to annotate subcategories of verbs and auxiliaries

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

Figure 4: Functional tag set for the URDU.KON-TB treebank

2 Semi-Semantic POS (SSP) Annotation

The term semi-semantic (partly or partially semantic) is used with the POS because some tags are encoded with semantics but not all e.g. N.SPT (a spatial noun) tag for a word *house*, ADJ.TMP (a temporal adjective) tag for a word *previous* in *previous year*, etc. There are twenty two (22) main POS tag categories, which are displayed in Figure 2. The description of the tags is given in the respective cells of the figure. These main categories are further divided into morphological and semantical subcategories according to the Figures 3 and 4, respectively. The final and detailed version of the SSP tag set is given in Figure 1. The dot "." is used to add the morphological or semantical features to the main category e.g. in V.PERF, a verb V is the main POS category like nouns, adjectives, etc., which has a perfective PERF morphology. The description of each category is as follows. It is to be noted that the Urdu script is written from right to left in coming examples. The row beneath the Urdu script is the transliteration of the sentence as proposed in Malik et al. (2010). Similarly, the row beneath the transliteration of the sentence contains the translated-word/POS-tag pair according to the SSP tag set given in Figure 1. At the end in examples next, a complete English translation of the Urdu sentence is presented. The complete guideline is going to be presented next, however, its employment procedure on the raw corpus to form the URDU.KON-TB treebank can be seen in Section 3.3. It is advised to skip this Section 2 for later reading and go to Section 3 to understand the flow of the article as this section concludes the deep design of the annotation guidelines.

2.1 Adjectives

Adjectives are used to modify a noun or pronoun (Aarts et al., 2014; Matthews, 2007; Miller et al., 1990; Stevenson, 2010). The first main category in Figure 1 is about ADJ (Adjective), which is divided into further five sub categories of tags included DEG (Degree), ECO (Echo), MNR (Manner), SPT (Spatial) and TMP (Temporal). The relevant POS annotations are provided in examples 1. Example 1(a) is the case of main POS category ADJ of adjective. There are some words like *tar* 'more' and *tarIn* 'most', which truly act as a degree adjective and not as degree adverb but there are some words which can play the role of a degree adverb or a degree adjective e.g. *ziyAdah* 'more/much', *bohat* 'more/much', etc, (Schmidt, 2013). Example 1(b) is the case of degree adjective ADJ.DEG. Example 1(c) is the case of reduplication[4] (Abbi, 1992; Boegel et al., 2007). Reduplication has two versions. First is discussed in a footnote below, while the other is the repetition of the original word e.g. *sAtH sAtH* 'with/along-with'. These two versions are named as echo reduplication and full word reduplication by Boegel et al. (2007), which are refurbished in our annotation as ECO (echo

[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'

reduplication) and REP (full word reduplication/repetition) respectively. The echo words normally start with the letters *S* or *v* or *m*. The next examples from 1(d) to 1(f) are the cases of adjectives, which have the meaning of MNR, TMP and a SPT respectively. The addition of this MNR, TMP and SPT after the POS tag ADJ represents the semantics.

(1) a. اچھا لڑکا

acHA laRkA
good/ADJ boy/N
'Good Boy'

b. اہم ترین شخصیت

aham tarIn Saxs2iat
important/ADJ most/ADJ.DEG personality/N
'Most important personality'

c. برا ورا کام

burA vurA kAm
ugly/ADJ ADJ.ECO work/N
'Ugly work'

d. جابرانہ حکومت

jaberaanah hakUmat
cruel/ADJ.MNR government/N
'Cruel Government'

e. گزشتہ سال

guzaStah sAl
previous/ADJ.TMP year/N
'Preveious Year'

f. ملتانی کھسہ

mUltAnI kHUsah
multani/ADJ.SPT shoe/N
'Multani shoe'

2.2 Adverbs

Adverbs can modify verbs, adjectives or other adverbs. They can also modify phrases, clauses and sentences (Aarts et al., 2014; Matthews, 2007; Miller et al., 1990; Stevenson, 2010). Adverbs are mostly used as a qualifier of the verbs but they can also be used independently. They are subcategorized into six forms presented in Figure 1. The annotations are given in example 2. The main category of adverbs ADV is annotated in 2(a), which is further divided into five subcategories DEG (degree), MNR (manner),

NEG (negative), SPT (spatial) and TMP (temporal). The final TMP has another subcategory REL for relative temporal adverb. In 2(b), an adverb *bohat* 'very' is used before an adjective *acHI* 'good' and it is highlighting the adjective at a certain degree, hence annotated as ADV.DEG. In 2(c), *biltartIb* 'respectively' behaves as an adverb and advocates a manner of order as ADV.MNR. The word *nah* 'not' is a negative adverb negating the action in 2(d) and it is annotated with ADV.NEG relatively. A word *sAmnE* 'front/before' is a spatial adverb and annotated as ADV.SPT in 2(e). The case of temporal adverb is displayed in 2(f), where a word *ab* 'now' is annotated as ADV.TMP. This temporal adverb is divided into another hierarchy named relative-temporal adverb, which can be seen in the last example 2(g). A word *jab* 'when' is given a POS tag as ADV.TMP.REL as follows.

(2) a. تقریبا ساری دنیا میں

taqrIban	sArI	dunIyA	mEN
almost/ADV	whole/Q	world/N.SPT	in/CM

'Almost in the whole world'

b. بہت اچھی لڑکی

bohat	acHI	laRkI
very/ADV.DEG	good/ADJ	girl/N

'Very good girl'

c. تعداد بالترتیب ۵ اور ۶ تھی

te2dAd	biltartIb	5	aor
quantity/N	respectively/ADV.MNR	5/Q.CARD	and/C.CORD
6	tHI		
6/Q.CARD	was/V.COP.PAST		

'The quantity was 5 and 6 respectively'

d. عمارت مکمل نہ ہو سکی

e2emArat	mukammal	nah	hO	sakI
building/N	complete/ADJ	not/ADV.NEG	be/V.LIGHT.ROOT	could/V.MOD.PERF

'The building could not be completed'

e. تفصیلات سامنے آئیں گی

tafs2IlAt	sAmnE	AyIN	gI
details/N	front/ADV.SPT	come/V.SUBTV	will/VAUX.FUTR

'The details will come out'

f. اب دیکھنا یہ ہے

ab	dEkHnA	yE	hE
now/ADV.TMP	to-see/V.INF	this/P.PERS	be/V.COP.PRES

'Now, this is to be seen'

g. جب یہاں کھیت ہوتے تھے

jab	yahAN	kHEt	
when/ADV.TMP.REL	here/ADV.SPT	crop-field/N.SPT	
hotE	tHE		
be/V.IMPERF	was/VAUX.PAST		

'When, there were crop fields here'

2.3 Conjunctions

Conjunctions are used to connect words, phrases, clauses or sentences (Aarts et al., 2014; Matthews, 2007; Miller et al., 1990; Stevenson, 2010). The main category of conjunction C is divided into five subcategories e.g., CAUS (causative), CONS (concessive), CORD (coordinative), CORR (correlative) and SBORD (subordinating). The last subcategory has another division of COND to represent conditional subordinate conjunction. The annotation of all divisions is presented in example 3. Words like *cUnkEh* 'since, because', *cUnAcEh* 'so, therefore', *kIUnkEh* 'because' are candidates for a causative conjunction in a clause. An example of causative conjunction is depicted in Example 3(a). The POS annotation examples of CONS and CORD are given in 3(b) and 3(c), respectively. The word *agarcEh* 'although' is acting as a concessive conjunction in the beginning of sentence in 3(b), while the other word *aor* 'and' is a coordinating conjunction in 3(c). The word *nah* 'not/neither' as a correlative conjunction is presented in 3(d), in which it is annotated with C.CORR tag. The subordinating conjunction C.SBORD is annotated in 3(e) for a word *kEh* 'that'. The C.SBORD is divided into another subcategory proposed as COND for conditional subordinating conjunction. Its annotation for a word *agar* 'if' is presented in 3(f).

(3) a.

SAyad	voh	akElA	tHA	kIUnkEh
perhaps/ADV	he/P.PERS	alone/ADJ	be/V.COP.PAST	because/C.CAUS
kHAnA	hOtEl	sE	kHAtA	tHA
meal/N	hotel/N.SPT	from,in/CM	eat/V.IMPERF	be/VAUX.PAST

'Perhaps, he was alone because he used to eat his meals in a hotel'

b.

agarcEh	AdmI	kam	tHE	magar
men/N	less/ADJ	were/V.COP.PAST	although/C.CONS	but/C.CORD
voh	pHir	bHI	jIt	gayE
they/P.PERS	then/ADV	too/PT.INTF	won/V.ROOT	V.LIGHTV.PERF

'Although the men were less but they had won either'

c.

te2dAd	biltartIb	5	aor
quantity/N	respectively/ADV.MNR	5/Q.CARD	and/C.CORD
6	tHI		
6/Q.CARD	was/V.COP.PAST		

'The quantity was 5 and 6 respectively'

d. nah tO tUm kHElE nah
neither/C.CORR PT.EMP you/P.PERS played/V.PERF nor/C.CORR
hI kHElnE diyA
PT.INTF play/V.INF gave/V.LIGHTV.PERF
'Neither you played yourself nor you allowed to play others'

e. nabI nE farmAyA kEh a2lI a2ilm
prophet/N CM said/V.PERF that/C.SBORD Ali/N.PROP knowledge/N
kA darvAzah hEN
of/CM door/N.SPT is/V.COP.PRES
'The prophet stated that Ali is the door to knowledge'

f. agar yEh mErA mAl hOtA
if/C.SBORD.COND it/P.PERS my/P.POSS property/N be/V.IMPERF
tO mEN xarc kartA
then/PT.RESULT I/P.PERS spend/V.ROOT do/V.LIGHTV.IMPERF
'If it would be my property then I will spend it'

2.4 Case markers

Case markers (CM) distinguish the grammatical functions of words, phrases, clauses, or sentences (Aarts et al., 2014; Matthews, 2007; Miller et al., 1990; Stevenson, 2010). Urdu case markers are syntactic clitics (Butt and Sadler, 2003) and divided into different forms by Butt and King (2004) e.g., ergative, accusative, dative, possessive, etc. All Urdu case markers are annotated with a simple CM tag at POS level. Four annotated examples can be seen in 3(a), 3(e) and 2(a) for instrumental case marker *sE* 'from', ergative case marker *nE*, possessive case marker *kA/kI/kE* 'of' and spatial case marker *mEN/par/tak* 'in/on/at'. The different forms of case markers play an important role in identification of argument structure like subject, object, etc. The effect of different forms and their related argument structure is discussed in Abbas (2014b).

2.5 Date

The DATE tag is used to represent dates of a month e.g. 14, 2, 31, etc. This tag is divided into three subcategories, which includes DATE.D, DATE.M and DATE.Y. Annotated examples can be seen in 4. The days of a week, month name and a year number are represented by DATE.D, DATE.M and DATE.Y, respectively.

(4) aetvAr 16 mayI 2004 kO
sunday/DATE.D 16/DATE May/DATE.M 2004/DATE.Y on/CM
'On Sunday, 16 May 2004'

2.6 Hadees

The Hadees is a report of deeds and saying of the prophet Muhammad (PBUH). These are tagged as HADEES in the URDU.KON-TB treebank. The Ahadees (plural of Hadees) in Arabic script in Urdu text are tagged only with this tag HADEES. The translated form of Ahadees in Urdu is annotated in a normal way. An example is depicted in 5 as follows. The Hadees with double quotes in the following sentence is in Arabic and hence tagged as HADEES.

(5) rasUl nE kahA “ h2UsyEno-minnI-va-anA-min-al-h2UsyEn ”
prophet/N CM said/V.PERF M.P HADEES M.P
'The prophet said, “Hussain is from me and I am from Hussain”'

2.7 Interjections

Interjections are the words or phrases used to exclaim, protest or command in a sentence. These are annotated with a tag INT. The example can be seen in 6 as follows.

(6) oE kHAnA kHAO
OE/INT food/N eat/V.SUBTV
'OE! eat the food'

2.8 Markers

The markers are used to identify the boundary of phrases, clauses, or sentences as marked by punctuation. The markers are divided into two subcategories e.g. phrase markers (M.P) and sentence markers (M.S). The punctuation within the sentence like single quotes, double quotes, colon, comma, etc., are annotated with M.P, however the boundary of the sentence like full stop and question mark is annotated with M.S. The annotated example can be seen in 7 as follows. The comma and period is marked by M.P and M.S respectively.

(7) in mEN bHakar , layIah
these/P.PERS in/CM Bhakkar/N.PROP.SPT comma/M.P Layyah/N.PROP.SPT
aOr lOdHrAN SAmil hEN .
and/C.CORD Lodhran/N.PROP.SPT include/N be/V.LIGHT.PRES full-stop/M.S
'Bhakkar, Layyah and Lodhran are included in these'

2.9 Nouns

The main noun tag N is divided into six subcategories, which includes adjectival noun (N.ADJ), noun having a manner (N.MNR), proper noun (N.PROP), repeated noun

(N.REP)[5], spatial noun (N.SPT) and temporal noun (N.TMP). The words *cHotE* 'younger' and *baRE* 'elder' are representing people having some property of young age and old age in 8(a), hence both are annotated with N.ADJ. In 8(b), the word *t2arah2* 'way, like, type' is first annotated with N.MNR but when the same word is repeated next then it gives the meaning of 'different types' and its repetition is annotated simply with N.MNR.REP. In 8(c), a subcategory N.PROP is annotated for a person name *marIyam* 'Maryam'. This subcategory is divided into two subcategories spatial and temporal, which are annotated as N.PROP.SPT and N.PROP.TMP for *panjAb* 'Punjab' and *a2Id-ul-fit2r* 'Eid festival', respectively. A common noun N is annotated in 8(b) for a word *taklIfEN* 'hardships'. There are some special common nouns, which can be repeated e.g. *kOrI kOrI* 'single penny'. When some noun is usually repeated then N.REP tag is used. So, this .REP along with the respective POS tag can be used to represent the presence of a repeated word. The annotation of N.SPT and N.TMP can be seen in 8(c) for *iz3lAa2* 'districts' and *din* 'day'. In both the subcategories, the repetition is possible for which the addition of REP with dot "." can be used accordingly.

(8) a.

cHOtE	baRE	sab	xUS	
younger/N.ADJ	elder/N.ADJ	all/Q.ADJ	happy/ADJ	
hOtE		hEN		
become/V.COP.IMPERF		be/VAUX.PRES		

'Younger and elder all become happy'

b.

UnhEN	t2arah2	t2arah2	kI	taklIfEN
they/P.PERS	type/N.MNR	type/N.MNR.REP	of/CM	hardships/N
dI	jAnE	lagIN		
give/V.PERF	go/VAUX.PASS.INF	start/VAUX.SUBTV		

'They were given hardships of different types'

c.

marIyam	panjAb	kE	ba2z	iz3lAa2
Maryam/N.PROP	Punjab/N.PROP.SPT	of/CM	some/Q	districts/N.SPT
mEN	a2Id-ul-fit2r	kE	din	gayI
into/CM	Eid-ul-Fitr/N.PROP.TMP	of/CM	day/N.TMP	went/V.PERF

'Maryam went into some districts of Punjab on the day of Eid-ul-Fitr'

2.10 Pronouns

The main category of pronoun P is divided into six subcategories P.DEM (demonstrative pronoun), P.INDF (indefinite pronoun), P.PERS (personal pronoun), P.POSS (possessive pronoun), P.REF (reflexive pronoun) and P.REL (relative pronoun). The first two subcategories P.DEM and P.INDF are annotated in 9(a) for words *yeh* 'this' and *kOI* 'any' respectively. The difference between P.PERS and P.DEM is this that when P.PERS refers to some person, place or thing, then this P.PERS behaves as a P.DEM like

[5] It lies in the category of full word reduplication as discussed in Section 2.1

in 9(a). The 3rd and 4th category P.PERS and P.POSS are annotated in 9(b) for words *mEN* 'I' and *tumhArA* 'your' respectively. P.POSS is further divided into P.POSS.REF, which is annotated for a word *apnA* 'own' in the same sentence. The repeated subcategory can be annotated after addition of .REP at the end. The fifth and sixth subcategory P.REF and P.REL are annotated in 9(c) for words *Apas* 'themselves' and *jO* 'which' respectively. The subcategory P.REL is further divided into P.REL.DEM and P.REL.PERS. These are annotated in 9(d) for words *jO* <u>*kUcH*</u> 'what ever' and *jIs* 'who' respectively.

(9) a. yeh meh2kama kOI kAm nahI kartA
this/P.DEM department/N any/P.INDF work/N not/ADV.NEG do/V.IMPERF
'This department does not do any work'

b. mEN tumhArA apnA bHAI hUN
I/P.PERS your/P.POSS own/P.POSS.REF brother/N be/V.COP.SUBTV
'I am your own brother'

c. jO Apas mEN moh2abat kI
which/P.REL themselves/P.REF among/CM love/N of/CM
mIs2Al hE
example/N be/V.COP.PRES
'Which is an example of love among themselves'

d. jIs kO jO kUcH milE
who/P.REL.PERS CM what/P.REL.DEM ever/P.INDF find/V.PERF
UTHA lEnA cAhIE
pick/V.PERF take/V.LIGHTV.INF should/VAUX.MOD.PERF
'Who finds what ever, should pick it up'

2.11 Postpositions

The postpositions are placed after a word to which it is grammatically related e.g. *sAtH* 'with' is a POSTP (postposition) in a postpositional phrase *Us kE sAtH* 'with him'. The postpositions are divided into six subcategories hierarchically as displayed in Figure 1. These include POSTP.CMP (comparative postposition), POSTP.MNR (postposition having a manner)[6], POSTP.POSS (possessive postposition), POSTP.REP (repetitive postposition), POST.SPT (spatial postposition) and POSTP.TMP (temporal post-position). The first two subcategories are annotated in 10(a) for postpositions *sE* 'than' and *t2arah2* 'like' respectively. In 10(b), the third and fourth subcategories are annotated for postpositions *pAs* 'have/has' and *sAtH* 'with' respectively. The last two

[6] The prepositions are divided into basic manner, manner by comparison and manner with a reference point by Saint-Dizier (2008) but I applied only manner in general to all related prepositions and postpositions for Urdu.

subcategories are annotated in 10(c) for postpositions *qarIb* 'near' and *ba2d* 'after' respectively.

(10) a. 25 sE z2yAdah laRkE
25/Q.CARD than/POSTP.CMP more/ADJ.DEG boys/N
aslam kI t2arah2 hEN
Aslam/N.PROP of/CM like/POSTP.MNR be/V.COP.PRES
'More than 25 boys are like Aslam'

b. xUrAk kE sAtH sAtH mErE
food/N of/CM with/POSTP with/POSTP.REP I/P.POSS
pAs pEsE bHI hEN
have/POSTP.POSS money/N also/PT.INTF be/V.COP.PRES
'I have also the money along with the food'

c. us kE qarIb h2amlE kE
him/P.PERS of/CM near/POSTP.SPT attack/N of/CM
ba2d bam pHatA
after/POSTP.TMP bomb/N exploded/V.PERF
'The bomb exploded near him after the attack'

2.12 Pray

The PRAY tag is used to annotate all types of prayers normally used in religious literature after the name of prophets, caliphs, and the righteous religious personalities e.g. the *alEh salAm* 'peace be upon him' is annotated with PRAY after the name of Jesus in 11(a) along with the other example as follows.

(11) a. h2az3rat a2lsA alEh-salAm allah
h2az3rat/TTL.REG Jesus/N.PROP AS/PRAY Allah/N.PROP
kE Ek a2z4Im nabI hEN
of/CM a/Q.CARD great/ADJ prophet/N be/V.COP.PRES
'Jesus (peach be upon him) is a great prophet of God'

b. h2az3rat mUhammad s3al-lalAhO-a2laehE-va-AlEhI-vasalam nE
TTL.REG Muhammad/N.PROP SAAWW/PRAY CM
h2az3rat a2lI kO apnA bHAI banAyA
TTL.REG Ali/N.PROP CM his/P.POSS.REF brother/N made/V.PERF
'Muhammad (peach be upon him and his descendant) made Ali his brother'

2.13 Prepositions

The prepositions are placed before a word to which it is grammatically related e.g., *bE* 'without' is a PREP (preposition) in a prepositional phrase *bE mUhAr Sutar* 'a camel without a hook). Prepositions are divided into three subcategories hierarchically as

displayed in Figure 1. These include PREP.MNR (preposition having a manner), PRET.SPT (spatial preposition) and PREP.TMP (temporal preposition). The first two subcategories are annotated in 12(a) for prepositions *bat2Or* 'as' and *andrUnE* 'in' respectively. The last subcategory is annotated in 12(b) for prepositions *dOrAnE* 'during'.

(12) a. us nE bat2Or DrAlvar andrUnE Sehar
he/P.PERS CM as/PREP.MNR driver/N in/PREP.SPT city/N.SPT
nOkrI kI
job/N do/V.PERF
'He did the job as a driver in the city'

b. voh yahAN dOrAnE taftIS
he/P.PERS here/ADV.SPT during/PREP.TMP investigation/N
A giA
come/V.ROOT go/V.LIGHTV.PERF
'He came here during the investigation'

2.14 Particles

The particles can appear after a word. These are divided into four subcategories, which include PT.ADJ (adjectival particles), PT.EMP (emphatic particles), PT.INTF (Intensifying particles) and PT.RESULT (resultant particles). All the subcategories are non-inflected except the PT.ADJ, which appears after adjective, adverb, noun or pronoun and agrees with the qualifier. The first and third subcategories are annotated in 13(a) for the particles *sA* 'like' and *bHI* 'too'. The annotation of PT.EMP is displayed in 13(b) for a word *tO*. The contrastive meaning is understood by default due to usage of PT.EMP in this sentence. In 13(c), the annotation of PT.RESULT is given for a word *tO* 'then'.

(13) a. voh Ek nAxUSgavAr sA bandah
he/P.PERS a/Q.CARD unpleasant/ADJ like/PT.ADJ man/N
bHI hE
too/PT.INTF be/V.COP.PRES
'He is like an unpleasant man too'

b. ab maslah falastIn tO
now/ADV.TMP problem/N Palestine/N.PROP.SPT PT.EMP
h2al hO gA
resolve/N be/V.LIGHT.ROOT will/VAUX.FUTR
'Now, the problem of Palestine will resolve (contrast: “the other problems will not” due to ‘tO’ effect)'

c. bAriS AyI tO mElah nahI
rain/N come/V.PERF then/PT.RESULT festival/N not/ADV.NEG
hO gA
be/V.ROOT will/VAUX.FUTR
'If the rain comes, then the festival will not hold'

2.15 Quantifiers

The quantifiers Q are used to show the amount of something. These are divided into four subcategories, which include Q.ADJ (adjectival quantifier), Q.CARD (cardinal quantifier), Q.FRAC (fractional quantifier) and Q.ORD (ordinal quantifier). In 14(a), the quantifiers *tamAm* 'all/whole', *har* 'every' and *dUsrA* 'second/other' are annotated with Q, Q.ADJ and Q.ORD, respectively. The remaining subcategories of quantifiers Q.CARD and Q.FRAC are annotated in 14(b) for words *Ek* 'one' and *cOtHAI* 'one 4th', respectively.

(14) a. tamAm mumAlik mEN har dUsrA
all/Q countries/N.SPT in/CM every/Q.ADJ second/Q.ORD
Saxs xUS hE
person/N happy/ADJ be/V.COP.PRES
'In all countries, every second person is happy'

b. mujHE Ek cOtHAI raqam dO
me/P.PERS one/Q.CARD fourth/Q.FRAC amount/N give/V.SUBTV
'Give me one fourth amount'

2.16 Questions Words

The question words QW identify a question in a sentence. These are divided into four subcategories, which include QW.REP (repeated question words), QW.TMP (temporal question words), QW.SPT (spatial question words) and QW.MNR (question words having a manner). The main category QW is depicted in 15(a) for a question word *kiyA* 'what'. If any question word is repeated then QW.REP can be used for annotation. The remaining three subcategories QW.TMP, QW.SPT and QW.MNR are annotated in a single sentence 15(b) for related question words *kab* 'when', *kidHar* 'where' and *kEsE* 'how', respectively.

(15) a. tumhArA nAm kiyA hE ?
your/P.POSS name/N what/QW be/V.COP.PRES ?/M.S
'What is your name?'

b. kab , kidHar aOr kEsE
when/QW.TMP ,/M.P where/QW.SPT and/C.CORD how/QW.MNR
jAO gE ?
go/V.SUBTV will/VAUX.FUTR ?/M. S
'When, where and how will you go?'

2.17 Symbols

The symbols SYM include brackets, parentheses, percent symbols, currency symbols, etc. All are dealt within a single category SYM as can be seen as follows.

(16) Gazvah fatah2 [fatah2e Makkah]
Gazvah/N.PROP fatah/N.PROP [/SYM fatahe/N Makkah/N.PROP.SPT]/SYM
: yeh ramz3An mEN hUI
:/M.P it/P.PERS Ramadan/DATE.M in/CM happen/V.PERF
'The Battle Conquest [conquest of Makkah]: It happened in Ramadan (a month name in Islamic Calendar)'

2.18 Titles

The titles are used to show respect or regard to personalities before addressing their names. At present it has only one subcategory TTL.REG (regard titles). Its annotation can be seen as follows.

(17) h2az3rat ImAm h2Ussyn a2lEh-salAm
his-highness/TTL.REG religious-head/N Hussain/N.PROP AS/PRAY
tIsrE ImAm hEN
third/Q.ORD religious-head/N be/V.COP.PRES
'His highness, the religious head, Hussain (PBUH) is the third religious head'

2.19 Units

The Unit U is used to represent different measuring units e.g., meter, liter, bar, grams, etc. The example of an annotation is given for the following sentence.

(18) qEsar nE 70 miliyan tan tEl
Qaiser/N.PROP CM 70/Q.CARD million/Q.CARD ton/U oil/N
darAmad kiyA
import/N do/V.LIGHT.PERF
'Qaiser imported 70 million ton oil.'

2.20 Verbs

The main verb V is divided into 11 subcategories, which are further divided into hierarchical subcategories discussed as follows. The hierarchical division of a special

word the VALA[7] and verb auxiliaries will be discussed in respective Sections 2.21 and 2.22.

2.20.1 Copula Verbs

The copula verb V.COP is used to connect the subject with the subject complement or the predicate link of a sentence (Aarts et al., 2014; Butt, 1995; Matthews, 2007; Miller et al., 1990; Stevenson, 2010). For example a sentence like *The weather is horrible* contains a subject *The weather* and a predicate link as an adjective *horrible*. The predicate of the sentence is the copula verb *is*. The copula verb connects the subject with the predicate link in this sentence. The V.COP (copula verb) is divided into six subcategories hierarchically as V.COP.IMPERF (a copula verb with imperfective morphology), V.COP.PERF (a copula verb with perfective morphology), V.COP.ROOT (a copula verb with root form), V.COP.SUBTV (a copula verb with subjunctive morphology), V.COP.PAST (copula verb with past tense) and V.COP.PRES (copula verb with present tense). The future form of copula verb itself is not possible in Urdu. In future construction, the copula verb 'be/become' always proceeds the future tense auxiliary *gA/gI/gE/gEN* 'shall/will' as can be seen in 19(d). The V.COP.IMPERF is annotated in 19(a) for a copula verb *hOtE* 'be/become' in imperfective form. Its perfective form is given in 19(b). The root form of another copula verb *ban* 'become' (Abbas and Raza, 2014; Raza, 2011) is presented in 19(c). The subjunctive form of copula verb *rahEN* 'remain' (Abbas and Raza, 2014; Raza, 2011) is annotated in 19(d). The copula verb with present and the past tense *hEN/tHE* 'are/were' is annotated in 19(e).

(19) a. SehrI parESAn nahI hOtE hEN
citizens/N worry/ADJ not/ADV.NEG be(become)/V.COP.IMPERF be/VAUX.PRES
'The citizens do not worry.'

b. xAhiS pUrI nah hUI
desire/N fulfill/ADJ not/ADV.NEG be(become)/V.COP.PERF
'The desire did not become fulfilled.'

c. sIrAj acHA ban giyA
Siraj/N.PROP good/ADJ become/V.COP.ROOT go/V.LIGHTV.PERF
'Siraj became good.'

d. Ith2AdI kAmyAb rahEN gE
allies/N successful/ADJ remain/V.COP.SUBTV will/VAUX.FUTR
'The allies will remain successful.'

e. aEsE lOg mOjOd hEN/tHE
such/ADJ people/N present/ADJ be/become/V.COP.PRES/.PAST
'Such people are/were present.'

[7] The VALA as a head word gives the phrase with adjectival property most likely and follows infinitive verb, nominal, adjectives, etc.

2.20.2 Imperfective Verbs

The imperfective verb tag V.IMPERF describes actions or states occurring generally or regularly (Schmidt, 2013). It can be identified by the inflected suffixes *tA, tI, tE, tEN* at the end of a verb. These suffixes co-exist after the root of a verb. At present, it has only one tag as V.IMPERF; however, if it is repeated then .REP can be used at the end. The example of V.IMPERF is given as follows.

(20) voh aks2ar cOrI kartE hEN
they/P.PERS often/ADV stealing/N do/V.IMPERF be/VAUX.PRES
'They often do stealing.'

2.20.3 Infinitive Verbs

An infinitive verb V.INF can be identified through its inflected suffixes *nA, nI, nE, nEN* concatenated to the root form of a verb. The annotation example is given as follows.

(21) h2akUmat kO kAm karnA hE
government/N CM work/N do/V.INF be/VAUX.PRES
'The government has to do work.'

2.20.4 Light Verbs I

A light verb V.LIGHT is a verb that contains a little semantic content of its own and it forms a predicate with some additional expression such as a noun or an adjective (Ahmed and Butt, 2011; Butt, 2003; Raza, 2011). It is subcategorized into eight different forms, which includes V.LIGHT.IMPERF (light verb with imperfective morphology), V.LIGHT.INF (light verb with infinitive morphology), V.LIGHT.PERF (light verb with perfective morphology), V.LIGHT.PROG (light verb with progressive morphology), V.LIGHT.ROOT (light verb with root form), V.LIGHT.SUBTV (light verb with subjunctive morphology), V.LIGHT.PAST (light verb with past tense) and V.LIGHT.PRES (light verb with present tense) as can be seen in 22(a)-(g) for light verbs *AtA* 'use to come', *AnA* 'to come', *AyA* 'came', *rAhA* 'remain', *A* 'come', *dORAEN* 'let to run' and *thA/hE* 'was/is' respectively. The respective subcategories can also be seen in Figure 1. All the light verbs presented in annotated sentences shared their semantic content with a preceding noun N or adjective ADJ.

(22) a. mUjE jin naz4ar AtA tHA
me/P.PERS ghost/N vision/N come/V.LIGHT.IMPERF be/VAUX.PAST
'I had used to sight the ghost.'

b. mUjE jin naz4ar AnA tHA
me/P.PERS ghost/N vision/N come/V.LIGHT.INF be/VAUX.PAST
'I had to sight the ghost.'

c. mUjE jin naz4ar AyA tHA
me/P.PERS ghost/N vision/N came/V.LIGHT.PERF be/VAUX.PAST
'I had sighted the ghost.'

d. a2li kO a2lambardArI kA a2uhdah h2Asil
Ali/N.PROP CM flag-bearer/N of/CM designation/N gain/N
rAhA
remain/V.LIGHT.PROG
'Ali had the designation of flag-bearer'

e. lOg taSadud bardASt nahI
people/N torture/N bear/N not/ADV.NEG
kar saktE
do/V.LIGHT.ROOT VAUX.MOD.IMPERF
'The people can not bear the torture'

f. ham naqSE par naz4ar dORAEN gE
we/P.PERS map/N on/CM vision/N run/V.LIGHT.SUBTV will/VAUX.FUTR
'We will look on the map'

g. SirAj pUrAnE AdmION mEN SAmil thA/hE
Siraj/N.PROP old/ADJ persons/N in/CM include/N was/is/V.LIGHT.PAST/PRES
'Siraj was/is included in old persons'

2.20.5 Light Verbs II

A light verb V.LIGHTV is a verb that contains a little semantic content of its own and it forms a predicate in the presence of an additional verb (Butt, 2003), hence is called as the verb-verb complex predicate. It is subcategorized into five different forms, which include V.LIGHTV.IMPERF, V.LIGHTV.INF, V.LIGHTV.PERF, V.LIGHTV.ROOT and V.LIGHTV.SUBTV as can be seen in 23(a)-(e) for light verbs *kartA* 'used to do', *dEnA* 'to give', *liyA* 'took', *hO* 'be/become' and *jAyE* 'should go' respectively. All the light verbs presented in the annotated sentences as follows shared their semantic content with a preceding verb V or a light verb V.LIGHT.

(23) a. voh patHar luRHkA dEtA tHA
he/P.PERS stone/N roll/V.PERF give/V.LIGHTV.IMPERF be/VAUX.PAST
'He used to roll the stone'

b. mUjHE jarmany cHOR dEnA
I/P.PERS Germany/N.PROP.SPT leave/V.ROOT give/V.LIGHTV.INF
cAhIyE
should/VAUX.MOD.PERF
'I should have to leave Germany'

c. mEN nE h2aj kar liyA
I/P.PERS CM pilgrimage/N do/V.ROOT take/V.LIGHTV.PERF
hE
be/VAUX.PRES
'I have performed the Hajj (pilgrimage)'

d. janral mUSaraf kO fOj muth2arik
general/N Musharraf/N.PROP CM army/N mobilization/N
karnA hO gI
to-do/V.LIGHT.INF be/become/V.LIGHTV.ROOT will/VAUX.FUTR
'General Musharraf will have to mobilize the army'

e. yeh saRak sitambar mEN mUkammal
this/P.DEM road/N.SPT September/DATE.M in/CM complete/ADJ
hO jAyE gI
be/become/V.LIGHT.ROOT should-go/V.LIGHTV.SUBTV VAUX.FUTR
'This road will be completed in September'

2.20.6 Modal Verbs

A modal verb V.MOD expresses a scale ranging from possibility to necessity (Abbas and Nabi Khan, 2009). It is subcategorized into three morphological forms, which includes V.MOD.IMPERF (modal verb with imperfective morphology), V.MOD.PERF (modal verb with perfective morphology) and V.MOD.SUBTV (modal verb with subjunctive morphology). This category of modal verbs is different from modal auxiliaries discussed in Section 2.22.3, in which the main verb (predicate) of the sentence is annotated with V and the modal auxiliaries are annotated with VAUX. The examples of *cAhnA* 'may want to' modified from Facchinetti et al. (2003) contain modal verb V.MOD acting as the predicate of the respective sentence and not as an auxiliary, and are presented as follows.

(24) a. voh kitAb paRHnA cAhtA hE
he/P.PERS book/N read/V.INF want/V.MOD.IMPERF be/VAUX.PRES
'He may wants to read the book.'

b. tUm nE intEqAm lEnA cAhA tHA
you/P.PERS CM revenge/N take/V.INF want/V.MOD.PERF be/VAUX.PAST
'You might have wanted to take the revenge'

c. voh intEqAm lEnA cAhEN gE
they/P.PERS revenge/N take/V.INF want/V.MOD.SUBTV will/VAUX.FUTR
'They will want to take a revenge'

2.20.7 Perfective Verbs

A verb with perfective morphology V.PERF can be identified through its inflected suffix e.g. *A, I, E, EN* concatenated to the root form of a verb. The annotation examples can be seen in 25 for a verb *kahA* 'said' in (a) and *giyA* 'went' in (b). The repetition of same verb can be annotated as V.PERF.REP. More examples can be seen in 3(d, e), 8(b), 9(d), 13(c) and 16.

(25) a.

rasUl	nE	kahA
prophet/N	CM	said/V.PERF

'The prophet said'

b.

voh	pOlIs	kE	pAs	giyA
he/P.PERS	police/N	CM	to/POSTP	went/V.PERF

'He went to the police'

2.20.8 Root Verbs

A verb with root form is a verb to which suffixes can be added (Schmidt, 2013). An annotated example can be seen in 26 for a verb *A* 'come', whose infinitive form is *AnA* 'to come'. More examples can be seen in 3(b) and 13(c). The repetition of same verb can be annotated as V.ROOT.REP.

(26)

voh	yahAN	dOrAnE	taftIS	A
he/P.PERS	here/ADV.SPT	during/PREP.TMP	investigation/N	come/V.ROOT

giA
go/V.LIGHTV.PERF

'He came here during the investigation'

2.20.9 Subjunctive Verbs

A subjunctive verb is a verb used to express hypothetical actions or conditions (Dic, 2014; Schmidt, 2013). Annotated examples can be seen in 2(e) and 15(b) for the subjunctive form of the verbs *AyIN* 'come' and *jAO* 'go' respectively.

2.20.10 Verb With Tense

There are sentences, the structures of which look like copular constructions but the argument requirement (subject and predicate link) for their predicates cannot be fulfilled. It means that either the subject or the predicate link is missing in these types of sentences. The structure of these types of sentences is closer to existential copula construction in English. For example, the sentence *There is the God* has an existential copular construction (Raza, 2011). The translation of this sentence in Urdu is *xUdA*/God

hE/is with one argument for an existential copula verb *is*. Due to incomplete arguments in these type of sentences only, the copula verb V.COP.PRES/PAST is reduced to V.PRES/PAST for present and past tense as follows.

(27) mErI xAhiS tHI/hE
my/P.POSS desire/N be/V.PAST/PRES
'It is my desire'

2.21 Special VALA

The VALA is a special word in Urdu, which normally appears in a noun or an adjective phrase. It can also express the action that is going to start in a special way as can be seen in 28(a). Another reading of the same sentence is also mentioned. A single tag VALA is used to represent all types of *vAlA* morphological forms. Example given in 28(b) has a nominal reading.

(28) a. mEN kAm karnE vAlA hUN
I/P.PERS work/N to-do/V.INF going/has/VALA be/V.COP.SUBTV
'I am going to do work or I am a working person' (two different readings)

b. mEN dUdH vAlA hUN
I/P.PERS milk/N has/VALA be/V.COP.PRES
'I am the milkman'

2.22 Verb Auxiliaries

Verb auxiliaries VAUX denote the tense, aspect, modality, voice, mood, emphasis, etc., of the sentence predicate (Aarts et al., 2014). In Urdu, a predicate or a complex predicate in the main verb phrase of the sentence precedes verb auxiliaries e.g., *hO*/V.COP.ROOT *gayI*/V.LIGHTV.PERF *hE*/VAUX.PRES 'has/have become' contains a tense auxiliary VAUX.PRES along with the complex predicate *hO gayI*. Verb auxiliaries are divided into 11 subcategories discussed as follows.

2.22.1 Imperfective Auxiliaries

The method of identification for the imperfective auxiliary VAUX.IMPERF is the same as was discussed in Section 2.20.2 of imperfective verbs V.IMPERF. It is a single sub- category with no any further divisions. An annotated example for this subcategory is given as follows.

(29) kEs invesTigESan pOlIs kE pAs calA
case/N investigation/N police/N CM to/POSTP walk/V.PERF
jAtA hE
go/VAUX.IMPERF be/VAUX.PRES
'The case (usually) goes to investigation police'

2.22.2 Infinitive Auxiliaries

The identification of infinitive auxiliaries VAUX.INF is the same as was discussed in Section 2.20.3 of infinitive verbs V.INF. It is also a single subcategory, whose annotated example is presented for *jAnE* 'to go' as follows.

(30) Ap a2rab qabAyl mEN pehcAnE jAnE
he/P.PERS Arab/N.SPT tribes/N in/CM recognize/V.INF go/VAUX.INF
lagE tHE
take/VAUX.PERF be/VAUX.PAST
'He had become known in Arab tribes'

2.22.3 Modal Auxiliaries

A modal auxiliary VAUX.MOD expresses a range from possibility to necessity (Abbas and Nabi Khan, 2009; Bhatt et al., 2011). It is subcategorized into three morphological forms, which include VAUX.MOD.IMPERF (modal auxiliary with imperfective morphology), VAUX.MOD.PERF (modal auxiliary with perfective morphology) and VAUX.MOD.SUBTV (modal auxiliary with subjunctive morphology). These modal auxiliaries are different from the modal verbs discussed in Section 2.20.6, in which the modal verbs were acting as the predicate of the sentence but here the modal auxiliaries are following the predicate of the sentence. The examples for modal auxiliaries are as follows for *saktE* 'can', *cAhIyE* 'should' and *paREN* 'has/have to'.

(31) a. voh baRE h2Ads2E kA sabab ban
they/P.PERS big/ADJ.DEG accident/N of/CM reason/N become/V.COP.ROOT
saktE hEN
can/VAUX.MOD.IMPERF be/VAUX.PRES
'They can become the reason of a big accident'

b. kAm xatam kar dEnA cAhIyE
work/N finish/N do/V.LIGHT.ROOT give/V.LIGHTV.INF should/VAUX.MOD.PERF
'The work should be finished'

c. vATar kOnsal kO qAnUnI mUSgAfIyAN
Water/N.PROP Council/N.PROP CM regulation/ADJ anomalies/N
dUr karnI paREN gIN
far/ADJ do/V.LIGHT.INF has/haveto/VAUX.MOD.SUBTV will/VAUX.FUTR
'The Water Council will have to remove the regulation anomalies'

2.22.4 Passive Auxiliaries

In sentences with passive auxiliaries VAUX.PASS, the theme/patient becomes the grammatical subject of the main verb. It is divided into five subcategories, which includes VAUX.PASS.IMPERF (passive auxiliary with imperfective morphology), VAUX.

PASS.INF (passive auxiliary with infinitive morphology), VAUX.PASS.PERF (passive auxiliary with perfective morphology), VAUX.PASS.ROOT (passive auxiliary with root form), and VAUX.PASS.SUBTV (passive auxiliary with subjunctive morphology). The given examples have a morphological annotation of passive auxiliaries *jAtA* 'use to go', *jAnA* 'to go', *giyA* 'went', *jA* 'go' and *jAyEN* 'may go' respectively. These different forms of *jA* 'go' auxiliary are considered passive only, when they are preceded by a predicate or a complex predicate with perfective morphology (Raza, 2010).

(32) a. jAnvarON kO pAnI pilAyA
animals/N CM water/N make-someone-drink/V.PERF
jAtA hE
go/VAUX.PASS.IMPERF be/VAUX.PRES
'The animals are watered'

b. kaSmIrION kA bHI sOcA
Kashmiri/N.SPT of/CM also/PT.INTF think/V.PERF
jAnA cAhIyE
go/VAUX.PASS.INF should/VAUX.MOD.PERF
'Kashmiri's should also be considered'

c. tehsIldAr ka tabAdlah kiyA giyA
Tehsil-officer/N of/CM transfer/N do/V.PERF go/VAUX.PASS.PERF
'The Tehsil officer has been transferred'

d. sUDAn mEN nasal-kUSI kI
Sudan/N.PROP.SPT in/CM genocide/N do/V.PERF
jA rahI hE
go/VAUX.PASS.ROOT continue/VAUX.PROG be/VAUX.PRES
'Genocide is being commuted in Sudan'

e. kaSmIr sE fOjEN nikAl
Kashmir/N.PROP.SPT from/CM armies/N takeout/V.ROOT
lI jAyEN
take/V.LIGHTV.PERF go/VAUX.PASS.SUBTV
'The armies may be taken out from Kashmir'

2.22.5 Perfective Auxiliaries

The identification of perfective auxiliary VAUX.PERF is the same as was discussed in Section 2.20.7 of perfective verbs. It is an independent single subcategory, whose annotation is given in the following example.

(33) lOgON kE ravaIyON mEN tabdeelI AtI gayI
people/N of/CM behavior/N in/CM change/N come/V.IMPERF go/VAUX.PERF
'The change used to come in people's behaviors'

2.22.6 Progressive Auxiliaries

The progressive auxiliary VAUX.PROG can be identified easily through its morphological form after a verb or an auxiliary. Its morphological forms include *rahA, rahE, rahI, rahIN*. An annotated example can be seen in 32(d) for a progressive auxiliary *rahI* 'continue'.

2.22.7 Root Auxiliaries

The identification of an auxiliary with a root form VAUX.ROOT is the same as discussed in Section 2.20.8 for verbs with root morphology. An annotated example is given as follows.

(34) faqIr baRHtE jA rahE hEN
beggars/N increase/V.IMPERF go/VAUX.ROOT continue/VAUX.PROG be/VAUX.PRES
'The beggars are increasing'

2.22.8 Subjunctive Auxiliaries

A subjunctive verb auxiliary VAUX.SUBTV describes an uncertain action or state contingent on something else like permission, wish, request, etc., (Schmidt, 2013). It has no further divisions. An annotated example of subjunctive auxiliary is given as follows.

(35) kiyA mEN andar AyUN ?
what/QW I/CM In/ADV.SPT come/VAUX.SUBTV ?/M.S
'May I come in?'

2.22.9 Tense Auxiliaries

The tenses of auxiliary VAUX are divided mainly into three tense divisions, which include VAUX.FUTR (future tense auxiliary e.g. *gA, gI, gE, gIN*, etc.), VAUX.PAST (past tense auxiliary e.g. *tHA, tHI, tHE, tHEN*, etc.) and VAUX.PRES (present tense auxiliary e.g. *hE, hEN*, etc.). The annotation of the future tense auxiliary can be seen in 36(a). The annotation of past tense auxiliary is presented in 36(b). Similarly, the annotation of last subcategory VAUX.PRES is annotated in 36(c).

(36) a. ham naqSE par naz4ar dORAEN gE
we/P.PERS map/N on/CM vision/N run/V.LIGHT.SUBTV will/VAUX.FUTR
'We will look on the map'

b. ham naqSE par naz4ar dORA
we/P.PERS map/N on/CM vision/N run/V.LIGHT.PERF
rahE tHE
continue/VAUX.PROG be/VAUX.PAST

'We were looking on the map'

c. ham naqSE par naz4ar dORA
we/P.PERS map/N on/CM vision/N run/V.LIGHT.PERF
rahE hEN
continue/VAUX.PROG be/VAUX.PRES
'We are looking on the map'

Tags presented in Figure 1 have been completed along with the examples. The discussion in this Section 2 concludes the SSP guidelines for the URDU.KON-TB treebank.

3 The URDU.KON-TB Treebank

The development of the URDU.KON-TB treebank was performed in three steps: the collection of sentences in the form of a corpus, manufacturing of an annotation scheme and the employment of this annotation scheme on the said corpus. These steps are overviewed as follows. In initial development of the URDU.KON-TB treebank (Abbas, 2012), a POS, a syntactic and a functional tag sets were proposed. It was an original work done after getting motivation from the Penn treebank[8] and the Urdu Lexical Functional Grammar (LFG) built during a project called PARGRAM[9]. This work has some issues, which were resolved and updated after the annotation evaluation (Abbas, 2014a & 2014b). The updated versions of the tag sets have been presented and discussed earlier in Section 2.

3.1 Corpus Construction

In initial development (Abbas, 2012), a 19 million words corpus (Ijaz and Hussain, 2007) was used that was available at CRULP/CLE.[10] This corpus was collected from the Jang[11] and the BBC[12] newspapers. This corpus had licensing constraints due to which it is not publicly available anymore (Urooj et al., 2012). One thousand (1000) sentences taken from this corpus are then extensively modified to become free from 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.[13] The data collected from the Urdu Wikipedia

[8] http://www.cis.upenn.edu/~treebank/home.html

[9] http://ling.uni-konstanz.de/pages/home/pargram_urdu/

[10] A center for language engineering in Pakistan at http://cle.org.pk/

[11] http://jang.com.pk/

[12] http://www.bbc.co.uk/urdu/

[13] https://ur.wikipedia.org

is already under that license. Thus, the size of the corpus is limited to fourteen hundred (1400) sentences. The size of corpus is kept limited within the context of doctoral work (Abbas, 2014b), however, an extension project[14] to increase the size of the treebank up to 2000 sentence is completed and will be published soon. Overall, the corpus contains the text of local & international news, social stories, sports, culture, finance, history, religion, traveling, etc.

3.2 Annotation Scheme

The annotation scheme of the URDU.KON-TB treebank consists of semi-semantic POS (SSP), semi-semantic syntactic (SSS) and functional (F) tag sets. The term semi-semantic (partly or partially semantic) is used with the POS because some tags are encoded with semantics but not all e.g. N.SPT (a spatial noun) tag for a word *house*, ADJ.TMP (a temporal adjective) tag for a word *previous* in *previous year*, etc. At the SSP level, a dot '.' is used to add morphological (Figure 3) and semantical (Figure 4) labelings of subcategories into the main categories (Figure 2) as discussed in Section 3.3. Overall, for the SSP, SSS and F annotation, a combination of phrase structure (PS) and hyper dependency structure (HDS) has been adopted. The DS is called HDS because it is not limited to make constituents on the basis of headwords, but also on the basis of the head-constituents, when you have to make a constituent from its nested constituents. The details are given in Abbas (2014b). The POS, morphological, syntactical, semantical, clausal and functional information (Abbas, 2014b) all together, makes a rich annotation scheme for the URDU.KON-TB treebank. The need for such type of schemes is highly advocated by some researchers, such as Clark et al. (2010), Skut et al.(1997), etc.

3.3 Employment of Annotation

A simple POS tag set was devised first, which contained twenty two (22) main POS-tag categories displayed in Figure 2. The description of the tags is given in the respective cells of the figure. The figure includes some non-familiar tags like HADEES and MARKER to represent the Arabic statements of prophets in Urdu text and a phrase or a sentence marker similar to punctuation marks but not all, respectively. The labels for morphological and semantic subcategories are presented in Figures 3 and Figure 4 respectively, which can be added to 22 main categories of POS tags by using a dot '.' symbol. The SSP tag set was refined during the manual annotation process of sentences and further refined after the evaluation process with the Krippendorf's α statistical model (Krippendorf, 2014) and also presented in Abbas (2014a). The final refined form of the SSP tag set is given in Figure 1. In case of morphology, if a main verb V has a perfective morphology, then the tag becomes V.PERF. Similarly, the case of spatial noun N.SPT is discussed in the beginning of Section 3.2. The semantic tags like SPT

[14] http://clsp.org/projects.html

(spatial), TMP (temporal), MNR (manner), etc. are not possible with verbs, auxiliaries, conjunctions, etc., as can be seen in Figure 1.

(37) حامد نے شیر کو جنگل میں بندوق سے مارا ۔

hAmed	nE	SEr	kO	jangal	mEN	bandUq	sE	mArA	.
N.PROP	CM	N	CM	N.SPT	CM	N	CM	V.PERF	M.S

'Hamid killed the lion in the jungle with a gun.'

An example of the SSP annotation is given in Example 37. The Urdu script is written from right to left. The row beneath the Urdu script is the transliteration of the sentence as proposed in Malik et al. (2010). The tokens of the sentence are tagged according to the SSP tag set. *hAmed* is a proper name (N.PROP). *SEr* and *bandUq* are common nouns (N), while *jangal* is a spatial common noun (N.SPT).[15] *nE, kO, mEN*, and *sE* are case markers (CM) for ergative, accusative, spatial/locative and instrumental cases, respectively. The syntactic differentiation of the case markers is done according to the studies in Butt and King (2004).

The tagset in Figure 1 represents the complete SSP tagset. The discussion on each tag is presented in Section 2. As a repeated example, consider the ADJ (Adjective) in Figure 1, which is divided into five subcategories of tags DEG (Degree), ECO (Echo), MNR (Manner), SPT (Spatial) and TMP (Temporal). Relevant examples are provided in 38.

(38) a. اچھا لڑکا

acHA	laRkA
good/ADJ	boy/N

'Good Boy'

b. اہم ترین شخصیت

aham	tarIn	Saxs2iat
important/ADJ	most/ADJ.DEG	personality/N

'Most important personality'

c. برا ورا کام

burA	vurA	kAm
ugly/ADJ	ADJ.ECO	work/N

'Ugly work'

[15] In the presence of a sense of place/location or direction to/from place/location in a word, SPT tag is used e.g. Pakistan and the country, are the two words. Pakistan is the proper name of a place (country) and is tagged as N.PROP.SPT. However, country is a common noun but having a sense of place. So, it is tagged as N.SPT. This distinction is not different from spatial adverbs e.g. there, here, etc.

d. جابرانہ حکومت

jaberaanah hakUmat
cruel/ADJ.MNR government/N
'Cruel Government'

e. گزشتہ سال

guzaStah sAl
previous/ADJ.TMP year/N
'Preveious Year'

f. ملتانی کھسہ

mUltAnI kHUsah
multani/ADJ.SPT shoe/N
'Multani shoe'

The example 38(a) is a simple case of ADJ, while 38(b) is a case of a degree adjective [16] annotated with ADJ.DEG. The comparative and superlative forms of adjectives can be made by introducing Persian suffixes *tar* 'more' and *tarIn* 'most' after the absolute form of adjectives e.g. *xUbs3Urat-tar* 'prettier' and *xUbs3Urat-tarIn* 'prettiest'. There are some words, which can play the role of a degree adverb or a degree adjective e.g. *zEyAdah* 'more/most/much', *bohat* 'more/enough', *kAfI* 'quite/too', etc. (Schmidt, 2013). If these words qualify adjectives, then this is the usage as degree adverbs, otherwise as a degree adjective. Example 38(c) is a case of reduplication (Abbi, 1992; Boegel et al., 2007). As reduplication has two versions, first 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'. Second version is the repetition of the original word e.g. *sAtH sAtH* 'with/along-with'. These two versions are named as full word reduplication and echo reduplication by Boegel et al. (2007), which are represented in our annotation as ECO (echo) and REP (repetition) respectively. The echo words normally begin with the letters *S* or *v* or *m*.

Example 38(d) 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 confirmed through ADJ.MNR e.g. *z4AlemAnah t2abdIlIyAN* 'brutal changes'. If an adjective comes individually, then its mannerism can be resolved independently through its sense or by building a sense with the predicate. If there is a sense of manner then an adjective of manner can exist like in copular construction e.g. *voh GEr-h2Az3ir hE* 'He is absent'. An exercise of adjectives and adverbs of manner for the English language can be seen at Cambridge

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

University from which this idea is taken.[17] Example 38(e) is case of an adjective having a temporal sense. Finally, example 38(f) is the case of an adjective having a spatial sense. The adjective used here is the derivational form of a city/place name Multan, which is a spatial proper noun. But it appears here as an adjective and annotated as ADJ.SPT[18] like in this sentence e.g. *voh Ek pAkistAnI laRkA hE* 'He is a pakistani boy'.

The example 38 for adjectives exploited its POS tags along with semantic tagging like TMP, SPT, MNR, etc. However, to give an introduction about morphology and verb functions, another POS category V from Figure 1 is discussed as follows. A few high quality studies were conducted on verbs for morphologically rich language (MRL) Urdu by Butt and Rizvi (2010), Butt and Ramchand (2001) and Butt (2010). The rules for identifying different forms of verbs were adopted from these studies. The V annotates the predicate/main-verb of the sentence and is divided mainly into 11 subcategories, which include COP (copula verb), IMPERF (imperfective morphological form of the 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). Their description is also given in Figure 1. These tags are further divided into subcategories depicted in Figure 1. All these tags represent different morphological forms and the function of a verb that it governs. Some annotated sentences containing different verb forms and functions are given in example 39.

(39) a. مہنگائی نے لوگوں کا جینا دوبھر کیا تھا

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

'The inflation had made the life of people hard.'

b. گرانفروشوں کے خلاف قانون حرکت میں لایا جائے

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

'The law should be practiced against inflators'

c. محمد صلی اللہ علیہ والہ وسلم نے فرمایا کہ الحسین منی و انا من الحسین
یعنی حسین مجھ سے ہے اور میں حسین سے ہوں .

mUhammad	sal-lal-la-ho-a2lEhE-va-AlEhI-salam	nE	farmAyA	keh		
N.PROP	PRAY	CM	V.PERF	C.SBORD		
“	al-hUsynON-mInnI-vA-anA-mInal-hUsyn	”	ya2nI	‘	hUsyn	
M.P	HADEES	M.P	ADV	M.P	N.PROP	

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

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

mUjH sE hE aOr mEN hUsyn sE hUN ’ .
P.PERS CM V.COP.PRES 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. تم نے حج تو کر لیا ہو گا ؟

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?'

e. جب یہاں کھیت ہوتے تھے

jab yahAN kHEt hotE tHE
ADV.TMP.REL ADV.SPT N.SPT V.IMPERF VAUX.PAST
'When, here would have been crop fields'

f. ان کا مطالبہ تھا

Un kA mUtAlibah tHA
P.PERS CM N V.PAST
'It is their demand.'

The sentence in example 39(a) is a case of noun-verb complex verb predicate, which was first proposed by Mohanan (1994). The words *dUbHar kiyA* 'made hard' is a noun-verb complex predicate. The noun *dubHar* and the verb *kiyA* with a perfective morphological form *yA* at the end are annotated as a N 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 39(d). This construction is adopted from the studies given in (Butt, 2010). The light verb after a N or an ADJ lies in the 1st category of light verbs and annotated as V.LIGHT in our annotation, while the light verb after a verb lies in the 2nd category of light verbs and annotated as V.LIGHTV. The next sentence in 39(b) is a passive sentence. A passive construction can be concluded with the inflected form of a verb *jAnA* 'to go' proceeded by another verb with perfective morphology as can be seen in 39(b). The subjunctive form of auxiliary verb tagged as VAUX.PASS.SUBTV is preceded by a perfective verb *lAyA* 'brought', which is then annotated as V.PERF. The subjunctive form of verb is acting as an aspectual auxiliary and not as a V.LIGHTV, which was discussed in (Butt and Ramchand, 2001) and adopted as it is. The rules for identification of verb function and other morphological forms can be found in Section 2.

To explore some other unusual tags, a long sentence is presented in 39(c). After the name of prophets or righteous religious-personalities, some specific and limited prayers called *s3alAvAt* 'prayers' e.g. *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 most likely in Urdu text and annotated as PRAY. Similarly, the statements of

prophet Muhammad (PBUH) called *h2adIs2* 'narration' e.g. *In-namal-aa2mAlo-bin-niyAt* 'The deeds are considered by the intensions' in Arabic is also a tradition in Urdu text and annotated as HADEES. In religious text of Urdu, this kind of phenomenon is most likely in Arabic script rather than the Urdu script. This annotation with PRAY and HADEES is performed only, when prayers or narrations appear in Arabic language in Urdu text as can be seen in 39(c). The phrase markers like comma, double quotes, single quotes, etc. are annotated with M.P and sentence marker like full stop, question mark, etc. are annotated with M.S as presented in the same example. The tense is divided into present, past and future. A predicate of the sentence with present and past tense is possible as annotated in 39(f) but not with future tense, because future tense always behaves as verb auxiliary in Urdu. The tense of verb auxiliaries like present, past and future is annotated in 39(a, d, e). A verb with imperfective morphology e.g. *tA, tI, tE, tEN* at the end of a verb is annotated with V.IMPERF as given in 39(e).

This section concludes the concept of SSP tags used in the annotation of the URDU.KON-TB treebank. There are twenty-two tags, which are divided into further subcategories as presented in Figure 1. The POS annotation evaluation via Krippendorf's Alpha α is detailed in (Abbas, 2014a & 2014b), however an overview is presented next in Section 4. This evaluation came up with POS tags issues related to readability. After evaluation, the problematic POS tags are either removed or revised and a final SSP tagset is obtained and presented.

4 Evaluation and Results

This Section describes the evaluation of the annotation guidelines of the URDU.KON-TB treebank presented in Section 2 and 3. The evaluation is the process of calculating inter-annotator agreement (IAA), which provides a quantitative answer as to the overall consistency plus feasibility of the annotation scheme. For the evaluation of the URDU.KON-TB treebank annotation, the most advanced measure known as the Krippendorf's α coefficient (Krippendorf, 2004) is used. The output of the annotators is recorded and processed. The reliability of the SSP annotations is evaluated. The issues faced in annotation evaluation are removed via respective revisions and are reported shortly in forthcoming sections.

4.1 Setup

For the reliability evaluation of annotation guidelines presented in Section 2 of the URDU.KON-TB treebank for Urdu, it was essential that annotators should be the native speakers of Urdu possessing linguistics skills. To fulfill this purpose, an undergraduate class of 25 linguistics students has been adopted in the training course of annotation at Department of English, University of Sargodha, Pakistan.[19] This training was given to

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

students as a partial part of their major course of linguistics. During this training course, thirty-two (32) 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 high caliber of understanding was selected (but not informed) secretly from each group for the final annotation. The annotation task of 100 random sentences was divided into 10 home assignments. Each assignment contained 10 sentences. After twenty days of this course, the annotation assignments were given to all students along with the selected students with an instruction not to discuss it with each other. These assignments were collected, marked and the students were awarded with grades. The annotation performed in their home assignments by the selected students was then recorded in Microsoft Excel and evaluated for the reliability of annotation or IAA by applying the Krippendorf's α coefficient. The details of the SSP annotation evaluation can be seen in Section 4.2.

4.2 SSP Tagset Evaluation & Results

The detail and definition of the SSP tagset was already described in Sections 2 and 3. The complete guidelines of the SSP tagging were given to students for annotation of sentences according to a procedure described in Section 4.1. The tagged sentences by the annotators were recorded in the form of a reliability data matrix. Details are given in the doctoral thesis by Abbas (2014b). From the reliability data matrix, values by tokens matrix was obtained which is also presented in Abbas (2014b). The α coefficient was computed according to the formula given in equation below and also described in Abbas (2014b). In this work, different variables of numerator and denominator of equation were computed and described. The value of the α coefficient obtained was 0.964 for the SSP tagging of annotators. The value of α obtained lied in the category of perfect agreement according to the Krippendorf. A perfect agreement of 0.964 has been found in case of the SSP annotation only, which means that the SSP annotation guidelines are reliable. The error analysis and discussion of issues related to SSP tag set evaluation is given in Abbas (2014b) but not discussed here due to the scope of this article.

$$metric\alpha = 1 - \frac{O_d}{E_d} = 1 - (n..-1)\frac{\sum_t \frac{1}{n_t - 1}\sum_p \sum_{q>p} n_{tp} n_{tq} metric_{pq}^{\delta^2}}{\sum_p \sum_{q>p} n_{.p} n_{.q} metric_{pq}^{\delta^2}}$$

ADV	30	30	0		100.00
ADV.DEG	10	10	0		100.00
ADV.MNR	20	20	0		100.00
ADV.NEG	10	10	0		100.00
ADV.SPT	20	20	0		100.00
ADV.TMP	70	70	0		100.00
C.CORD	30	30	0		100.00
CM	630	630	0		100.00
DATE.Y.CAL	20	8	12	DATE.Y	40.00
DIA.IZF	90	39	51	DIA	43.33
DIA.PESH	10	3	7	DIA	30.00
KER	150	71	79	CM	47.33
N	1010	1005	5	BLANK, N.SPT	99.50
N.PROP	80	80	0		100.00
N.PROP.SPT	10	10	0		100.00
N.SPT	60	56	4	N	93.33
N.TMP	60	60	0		100.00
P.DEM	50	50	0		100.00
P.INDF	10	10	0		100.00
P.PERS	200	197	3	P.PRO	98.50
P.POSS	10	8	2	P.PERS	80.00
P.POSS.REF	10	10	0		100.00
P.REL	10	10	0		100.00
POSTP	60	60	0		100.00
POSTP.CMP	30	25	5	POSTP	83.33
POSTP.SPT	20	20	0		100.00
POSTP.TMP	20	20	0		100.00
PREP	20	20	0		100.00
PT	20	9	11	PT.INTF	45.00
PT.INTF	30	13	17	PT	43.33
Q	30	30	0		100.00
Q.CARD	210	210	0		100.00
Q.FRAC	20	20	0		100.00
Q.ORD	40	36	4	Q.CORD, Q.CARD	90.00
QW	50	50	0		100.00
U	30	30	0		100.00
V.COP.IMPERF	20	20	0		100.00
V.COP.PERF	10	10	0		100.00
V.COP.ROOT	20	20	0		100.00
V.COP.TENS.PRES	110	57	53	V.COP.PRES, V.LIGHT.TENS.PRES, VAUX.TENS.PRES	51.82
V.INF	10	10	0		100.00
V.LIGHT.IMPERF	40	37	3	V.LIGHTV.IMPERF	92.50
V.LIGHT.KER	50	21	29	V.LIGHT.ROOT	42.00
V.LIGHT.PERF	40	36	4	V.LIGHTV.PERF, V.PERF	90.00
V.LIGHT.ROOT	30	30	0		100.00
V.LIGHT.TB.ROOT	30	13	17	V.LIGHT.ROOT	43.33
V.LIGHTV.PERF	60	54	6	V.LIGHT.PERF	90.00
V.LIGHTV.PROG.IMPERF	10	5	5	V.LIGHTV.IMPERF	50.00
V.PERF	200	189	11	VAUX.PASS.PERF	94.50
V.ROOT	20	20	0		100.00
V.TENS.PRES	50	21	29	V.PRES	42.00
VAUX.IMPERF	20	17	3	V.PASS.IMPERF	85.00
VAUX.LIGHTV.TENS.PRES	10	4	6	VAUX.PRES	40.00
VAUX.MOD.PERF	30	30	0		100.00
VAUX.PASS.PERF	40	40	0		100.00
VAUX.PROG.PERF	20	9	11	VAUX.PROG, V.COP.PERF	45.00
VAUX.TENS.FUTR	20	10	10	VAUX.FUTR	50.00
VAUX.TENS.PAST	10	3	7	VAUX.PAST	30.00
VAUX.TENS.PRES	220	101	119	VAUX.PRES	45.91

Figure 5: Annotators SSP tags distribution and confusion

The format of data evaluation in the Krippendorf's α was different from the data displayed in Figure 5, however, a sample of annotators' SSP tags distribution and confusion is displayed in Figure 5. The annotated data of 100 sentences that were given to the annotators contained 1281 tokens, from which the data of 904 tokens is presented. The rest of the tokens have accuracy almost more than 90% due to which they are not depicted. It is attempted to show the tags of those tokens on which the annotators were remained confused or disagreed. The tags used in the initial version of the URDU.KON-TB treebank are displayed in the first column of the figure. Adjective (ADJ) appeared 54 times in the sentences, which were then annotated by 5 annotators. The frequency of adjective annotation is depicted in the second column of the figure after multiplying 54 with 5 numbers of annotators. It concludes 270 times of annotation for adjective. Among the 270 annotations of ADJ, annotators were remained 265 times in agreement or the annotators assigned 265 times the same/identical tag ADJ. The number of times the annotators remained disagreed or confused is mentioned in the different column. Similarly, the different or confused or the disagreed tags used by the annotators are depicted in the next column. Finally, by dividing the values in the identical and the frequency columns, the percentage accuracy of each tag in the first column of the figure is calculated.

The SSP tags in the initial version of the URDU.KON-TB treebank, which are correctly annotated and have 100% accuracy include ADV and ADV with its semantic labels for adverbs, coordination conjunctions with C.CORD, case markers with CM, N.PROP, N.PROP.SPT and N.TMP for proper nouns, spatial proper nouns and temporal nouns, P.DEM and P.INDF for demonstrative and indefinite pronouns, etc. The tags contained less than or equal to 50% accuracy include tense auxiliaries e.g. VAUX.TENS.PRES mostly annotated differently with VAUX.PRES, progressive auxiliary e.g. VAUX.PROG.PERF annotated differently with VAUX.PROG and its copular behavior with V.COP.PERF, KER as a light verb e.g. V.LIGHT.KER annotated with V.LIGHT.ROOT, diacritics e.g. DIA.IZF with DIA only, etc. Annotation of some tokens with tags was left by the annotators represented with BLANK in the column 'different/confused/ disagreed tags' for each tag in the first column of the figure.

The error analysis and evaluation of tags was performed on the basis of this data depicted in Figure 5. First, the tags with less or equal to 50% of accuracy are revised with annotators decisions e.g. DATE.Y.CAL, PT.INTF, V.LIGHT.TB.ROOT, etc., and second are the tags with accuracy a little more than 50% but they have common confused pairs like V.COP.TENS.PRES and VAUX.TENS.PRES modified to V.COP.PRES and VAUX.PRES, respectively as can be seen in Figure 1. The detailed discussion on error analysis and evaluation of the SSP annotation is presented in (Abbas, 2014b).

5 Conclusion

This concludes the complete SSP guidelines of the URDU.KON-TB treebank with preliminary and essential additional information needed to explain the SSP annotation procedure in full. After introducing the URDU.KON-TB treebank (Abbas, 2012; Abbas, 2014a; Abbas 2014b) and the parser based on the URDU.KON-TB treebank (Abbas, 2014c/2015), the demand of the complete guidelines in the community was raising, due to which it is attempted to present the SSP complete guidelines as a first step. The rest of the guidelines for the semi-semantic syntactic and functional annotations will be presented soon. This effort does not only strengthen the practice of producing the guidelines for the annotation schemes but also addresses the modern issue of how to prepare and evaluate guidelines (Mikulova and Stepanek, 2010) effectively with the state of the art evaluation techniques (Krippendorf, 2004; Hayes and Krippendorf, 2007). Corpus annotated with these SSP annotation guidline can be useful to applications in this domain like natural language processing, machine learning and many language specific analysis as discussed in (Zia, et. al, 2015a/2015b; Abbas, et. al, 2009/2010/2014; Abbas 2014d).

References

Aarts, B., Chalker, S., & Weiner, E. (2014). *The Oxford Dictionary Of English Grammar*. Oxford University Press.

Abbas, Q. (2012, March). Building a hierarchical annotated corpus of urdu: the URDU. KON-TB treebank. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 66-79). Springer Berlin Heidelberg.

Abbas, Q. (2014a). Semi-semantic part of speech annotation and evaluation. *LAW VIII*, *75*.

Abbas, Q. (2014b). *Building Computational Resources: The URDU. KON-TB Treebank and the Urdu Parser* (Doctoral dissertation).

Abbas, Q. (2014c). Exploiting language variants via grammar parsing having morphologically rich information. *LT4CloseLang 2014*, 36.

Abbas, Q. (2014d). A Stochastic Prediction Interface for Urdu. *International Journal of Intelligent Systems and Applications*, *7*(1), 94.

Abbas, Q. (2015). Morphologically rich Urdu grammar parsing using Earley algorithm, Natural Language Engineering (NLE), Vol.21(2), PP.1-36, Cambridge University Press, UK

Abbas, Q., & Khan, A. N. (2009). Lexical functional grammar for Urdu modal verbs. In *Emerging Technologies, 2009. ICET 2009. International Conference on* (pp. 7-12). IEEE.

Abbas, Q., & Raza, G. (2014). A Computational Classification of Dynamic Urdu Copula Verb. *International Journal of Computer Applications*, *85*(10).

Abbas, Q., Ahmed, M. S., & Niazi, S. (2010). Language Identifier For Languages Of Pakistan Including Arabic And Persian. *International Journal of Computational Linguistics (IJCL)*, *1*(03), 27-35.

Abbas, Q., Karamat, N., & Niazi, S. (2009). Development of Tree-bank based probabilistic grammar for Urdu Language. *International Journal of Electrical & Computer Science*, *9*(09), 231-235.

Abbas, Q., Zia, T., & Khan, A. N. (2014). Syntactic and semantic analysis of Urdu modal verbs using XLE parser. *International Journal of Computer Applications*, *107*(10).

Abbi, A. (1992). Reduplication in South Asian Languages: An Areal, Typological, And Historical Study. Allied Publishers, New Delhi.

Ahmed, T., & Butt, M. (2011, January). Discovering semantic classes for Urdu NV complex predicates. In *Proceedings of the Ninth International Conference on Computational Semantics* (pp. 305-309). Association for Computational Linguistics.

Bhatt, R., Bögel, T., Butt, M., Hautli, A., Sulger, S., & King, T. H. (2011). *Urdu/Hindi modals*. Bibliothek der Universität Konstanz.

Bögel, T., Butt, M., Hautli, A., & Sulger, S. (2007). Developing a finite-state morphological analyzer for Urdu and Hindi. *Finite State Methods and Natural Language Processing*, 86.

Butt, M. (1995). *The structure of complex predicates in Urdu*. Center for the Study of Language (CSLI).

Butt, M. (2003). The light verb jungle [OL].

Butt, M. (2010). The light verb jungle: Still hacking away. *Complex predicates in cross-linguistic perspective*, 48-78.

Butt, M., & King, T. H. (2004). The status of case. In *Clause structure in South Asian languages* (pp. 153-198). Springer Netherlands.

Butt, M., & Ramchand, G. (2001). Complex aspectual structure in Hindi/Urdu. *M. Liakata, B. Jensen, & D. Maillat, Eds*, 1-30.

Butt, M., & Rizvi, J. (2010). Tense and aspect in Urdu. *Layers of aspect*, 43-66. Stanford: CSLI Publications.

Butt, M., & Sadler, L. (2003). Verbal morphology and agreement in Urdu. *Syntactic structures and morphological information. Mouton*, 57-100.

Clark, A., Fox, C., & Lappin, S. (2010). *The Handbook Of Computational Linguistics And Natural Language Processing*, 57. Wiley.com.

Facchinetti, R., Palmer, F., & Krug, M. (Eds.). (2003). *Modality in contemporary English* (Vol. 44). Walter de Gruyter.

Hayes, A. F., & Krippendorf, K. (2007). Answering The Call For A Standard Reliability Measure For Coding Data. *Communication Methods and Measures*, *1*(1), 77–89.

Hirsch, E. D., Kett, J. F., & Trefil, J. S. (2014). *The new dictionary of cultural literacy*. Houghton Mifflin Harcourt.

Ijaz, M., & Hussain, S. (2007, August). Corpus based Urdu lexicon development. In *the Proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan* (Vol. 73).

Kamran Malik, M., Ahmed, T., Sulger, S., Bögel, T., Gulzar, A., Raza, G., ... & Butt, M. (2010). Transliterating Urdu for a Broad-Coverage Urdu/Hindi LFG Grammar. In *LREC 2010, Seventh International Conference on Language Resources and Evaluation* (pp. 2921-2927).

Krippendorff, K. (2004). Reliability in content analysis. *Human communication research*, *30*(3), 411-433.

Leech, G. (2005). Adding linguistic annotation. , 17-29, Oxbow Books, Oxford.

Matthews, P. H. (2007). *The concise Oxford dictionary of linguistics*. Oxford University Press.

Mikulova, M., & Stepanek, J. (2010). Ways Of Evaluation Of The Annotators In Building The Prague Czech-English Dependency Treebank. In *LREC*.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, *3*(4), 235-244.

Mohanan, T. (1994). *Argument structure in Hindi*. Center for the Study of Language (CSLI).

Raza, G. (2010). Inferring Subcat Frames of Verbs in Urdu. In *LREC*.

Raza, G. (2011). *Subcategorization acquisition and classes of predication in Urdu* (Doctoral dissertation).

Schmidt, R. L. (2013). *Urdu, an Essential Grammar*. Psychology Press.

Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997, March). An annotation scheme for free word order languages. In *Proceedings of the fifth conference on Applied natural language processing* (pp. 88-95). Association for Computational Linguistics.

Stevenson, A. (Ed.). (2010). *Oxford dictionary of English*. Oxford University Press, USA.

Urooj, S., Hussain, S., Adeeba, F., Jabeen, F., & Parveen, R. (2012). CLE Urdu digest corpus. *LANGUAGE & TECHNOLOGY*, 47.

Zia, T, Akhtar, M. P., Abbas, Q. (2015a). Comparative Study of Feature Selection Approaches for Urdu Text Categorization. *Malaysian Journal of Computer Science*, *28*(2).

Zia, T., Abbas, Q., & Akhtar, M. P. (2015b). Evaluation of Feature Selection Approaches for Urdu Text Categorization. *International Journal of Intelligent Systems and Applications*, *7*(6), 33.