Developing a Part of Speech Tagset for Sindhi

Mutee U Rahman
Department of Electrical Engineering and Computer Science, Isra University, Hyderabad Sindh 71000, Pakistan
muteeurahman@gmail.com

Abstract

Part-of-Speech (POS) tagging is process of assigning unique grammatical tags to every word in a sentence. POS tagset is primary requirement of POS tagging process. This research paper discusses various grammatical classes of Sindhi with reference to POS tagset design and tagging. Various issues like tagset design considerations, tagset size and granularity, part of speech types, subtypes and their attributes for tagging are discussed in detail. General guidelines for designing Sindhi POS tagset of any possible size and granularity are given. Obligatory and proposed tagsets for Sindhi are presented which provide basis for further research in part of speech tagging, tagged corpus, chunking, syntax analysis, information retrieval, part of speech usage analysis and other natural language processing applications.

1. Introduction

Part-of-Speech tagging is key research area in natural language processing and computational linguistics. POS tagging is prerequisite of chunking and parsing of natural language text. It is an essential requirement to develop computational grammar of any language. Information retrieval systems make extensive use of POS tags for text indexing. Text to speech systems use these tags for pronunciation of words. POS taggers are used to tag huge amount of text to develop POS tagged corpus which is key resource for text analytics and intelligent text processing. Basic requirement of POS tagging is a properly defined part of speech tagset.

Despite of few research initiatives of NLP resource development for Sindhi language [1], [2], [3], [4] the development of Sindhi POS tagset is still an open research area. Neither the published work regarding tagset design nor the design guidelines are available.

Following sections discuss tagsets, existing work in Sindhi POS tagging, Sindhi word classes and their division with reference to POS tagset design, possible attributes of word classes, obligatory tagset, and recommended tagset of Sindhi in detail.

2. Tagsets

Tagset is a list of lexical entries with their grammatical categories or tags (POS tags). Language specific tagset is primary requirement of any tagging algorithm. Tagset design issues include representation of linguistic information needed, size, and granularity of the tagset. Generally, larger tagsets are more useful but result in lower accuracy and smaller tagsets are less useful but result in more accuracy.

Some example tagsets for English include London Lund Corpus [5] with 197 tags, Lancaster UCEREL [6] with 165 tags, LOB corpus [7] with 135 tags, Penn POS tagset [8] with 48 tags.

Apart from English, various tagsets are also available for other languages including; Urdu [9], [10], Hindi [11], Pashto [12], Sindhi [1] and Punjabi [13].

3. Existing Work in Sindhi POS Tagging

Sindhi is one of the less resourced languages in NLP studies. Only few published research papers are available on POS tagging of Sindhi. Which discuss rule based and wordnet based POS tagging in Sindhi. A rule based POS tagger [1] is first ever published work on POS tagging of Sindhi which presents a rule based tagging algorithm with a tagset of 67 tags. Few disambiguation rules for tags and some tokenization issues with tokenization scheme are discussed.

Sindhi POS tagging using wordnet [2] is another published work available. The work is almost identical to the above discussed research work. Instead of rules wordnet is used for tag disambiguation.

A comparison of rule based and wordnet based tagging algorithms is given in [3]. The paper concludes that wordnet based approach gives more accurate results compared to rule based approach.

Problems with the tagset proposed in above papers include un-necessary granularity (for example nouns are given twenty different tags), ambiguity in tags (generic as well as specific tags are present) and use of Preposition tag (in Sindhi prepositions don't exist instead postpositions are used). Also, the tagset does not follow any standard guidelines for POS tagset design.

POS tagging research work in Sindhi is still subject to research and major areas of concentration are POS tagset design and comprehensive approach of POS tagging that can serve as the basis of further research on parsing, machine translation and other NLP applications.

4. Sindhi Word Classes

Sindhi word classes are divided into eight different parts of speech which include: noun (اسم), pronoun (ضمير), postposition (حرف جر), adjective (صفت), adverb (ظرف), verb (فعل), conjunction (حرف جملو) and interjection (حرف ندا). Most of the classes are further divided into subclasses. While defining the POS tagset word classes and their subclass considerations directly affect the size of the tagset. Various issues related to the selection of subclasses for defining tagset need to be considered before tagset finalization.

Generally, word classes are divided into two broad categories: closed word classes and open word classes. Following sections discuss open and closed Sindhi word classes in detail.

4. 1 Closed Word Classes in Sindhi

Closed word classes in a language are those classes which have relatively fixed membership. These classes usually contain small number of words compared to open word classes. Postposition (حرف جر) in Sindhi (preposition in English) is an example of closed word class. New postpositions are rarely added. Table-1 shows closed word classes of Sindhi and their examples.

4.1.1. Pronouns (ضمير). Like other languages pronouns in Sindhi are forms that act as a kind for referring a noun and are considered closed word class.

Sindhi pronouns are divided into seven different types which include: demonstrative pronoun, wh-pronoun, reflexive pronoun, relative pronoun, co-relative pronoun, and indefinite pronoun. Personal pronouns are further divided into three categories namely: first person, second person and third person pronouns.

Table 1: Closed word classes in Sindhi.

S.No	Word Class	Examples
1.	Pronoun (ضمير)	تون، توهان، اسين، مان، آئون، اُهي، اِهي
2.	Postposition (حرف جر)	کي، تي، ۾، جو، جا، جي، مان، تان، تائين
3.	Conjunction (حرف جملو)	۽، يا، پر، ته به، پر، چاڪاڻ ته
4.	Intransitive (فعل لازمي), Transitive (فعل متعدي) Auxiliaries (فعل معاون)	آهي، ٿو، ها، ٻيو، هلان، ڊوڙيو
5.	Numerals, Cardinals, Ordinals, Fractals, Multipliers	هڪ، پنج، ڇهون، پنجوٽو، اڌ، منو، پاءُ
6.	Negative, Affirmative	نه، ڪونه، ڪونهي، ها، ٻلي
7.	Articles	به، پڻ، ئي، ته

Demonstrative pronouns are also divided into two categories. Table-2 shows different types of pronouns in Sindhi along-with their examples.

Third person pronouns have ambiguity with demonstrative pronouns; for example: third person pronouns ho:a هو (that/he) and uhe: اُهي (they) can also be demonstrative pronouns. In the sentence ' اُهي چوڪرا اچن پيا' (those boys are coming) 'اُهي' is remote demonstrative pronoun and in sentence 'اُهي اچن پيا' (they are coming) 'اُهي' is third person pronoun. This ambiguity needs to be resolved during POS tagging process.

Table 2: Sindhi Pronouns and their subclasses

S.No.	Type	Subtype(s)	Example
1.	Personal Pronoun	1st Person	اسين، آئون، مان،
		2nd Person	توهان، اوهين، تون،
		3rd Person	اهي، انهن
2.	Demonstrative Pronoun	Proximate	هيءُ، اِهو
		Remote	هو، اُهو
3.	Wh-Pronoun	-	ڪهڙو، ڇا
4.	Reflexive Pronoun	-	پاڻ، خود
5.	Relative Pronoun	-	جهڙو
6.	Co-relative Pronoun	-	تهڙو
7.	Indefinite Pronoun	-	ڪونه ڪو

4.1.2. Postpositions (حرف جر). Another closed word class in Sindhi is postposition. Postpositions usually come after nouns, pronouns, adjectives and adverbs within their own syntactic position. They show relationship between two nouns, noun and pronoun or nouns and adjectives. There are two types of postpositions in Sindhi: simple postpositions and compound postpositions. Another type of postpositions

may also be considered as hidden postpositions which may be tagged as postposition or postpositional/ ablative case of nouns or adverbs. Table-3 shows three types of postpositions and their examples.

4.1.3. Conjunction (حرف جملو). Conjunctions in Sindhi are divided into copulative, concessive, adversative, conditional, interrogative, casual and final categories [14] but syntactically all these categories fall into two main types coordinate and subordinate conjunctions [15]. Table-4 shows these two types with examples.

Table 3: Types of Sindhi Postpositions

S.No.	Postposition Type	Example
1.	Simple	۾، تي، کان، تان
3.	Compound	جي اڳيان، جي وچ ۾، کانسواءِ
4.	Hidden	ڳوڻان، گهران

Table 4: Conjunction types in Sindhi

S.No.	Conjunction Type	Example
1.	Coordinate	۽، يا، پر
2.	Subordinate	جيڪڏهن، جيتوڻيڪ، تنهن هوندي به

4.1.4. Transitive verb (فعل لازمي), Intransitive verb (فعل متعدي) and Auxiliaries (فعل معاون). Sindhi verbs are divided into four major categories: intransitive, transitive, auxiliary and compound verbs. Intransitive verbs (فعل لازمي) are verbs without object in a sentence. For example: in sentence "آئون ڊوڙان ٿو" the word "ڊوڙان" is intransitive verb.

Transitive verbs (فعل متعدي) are verbs with subject and object in a sentence. For example: in sentence "احمد خط لکي ٿو" the word "لکي" is a transitive verb as it has subject "احمد" and object "خط". Transitive and intransitive verbs are further divided into active and passive types.

Auxiliary or helping verbs in Sindhi are used to complete the sentence in different tenses. Auxiliaries are used with main verbs, adverbs and nouns. Examples of auxiliary verb include: آهي، پيو، ٿو etc.

All three types of verbs discussed above are closed class type verbs in Sindhi.

Table 5: Participles in Sindhi

S.No.	Participle Type	Example
1.	Present Participle (اسم حاليه)	ڊوڙندو، ڏيندي، ڏسڻندا
2.	Past Participle (اسم مفعول)	پڙهيل، ڊوڙيو، لڏل، ڳاتو
3.	Future Participle(اسم استقبال)	وڃڻو، ورڻهڻو، ونڊيون
4.	Verbal Noun (اسم فاعل)	اپائيندڙ، اپاڻڻهار،ڪاشيگر
5.	Conjunctive Participle(اسم معطوفي)	کائي، پڙهائي، وڻيو

4.1.5. Participles (کردنت يا مشتق). Participles are derived from verb roots. As root forms of verbs are closed classes in Sindhi so is the case with participles. Five participle types of Sindhi and their examples are shown in Table-5. Past participle and future participle sometimes act as adjectives in sentences and can create ambiguity during POS tagging.

4.1.6. Adjectives (صفت). Few types of adjectives in Sindhi which include cardinals, ordinals, multipliers, fractals and pronominal adjectives belong to closed class and usually new words are not added in these types. Table-6 shows examples of such types of adjectives.

Table 6: Closed class adjectives in Sindhi

S.No.	Adjective Type	Example(s)
1.	Cardinal	هڪ، ٻه، چار
2.	Ordinal	پهريون، پنجون
3.	Multiplier	بيڻو، پنجوڻو
4.	Fractal	اڌ، منو
5.	Pronominal Adjective	اسين ماڻهو

4.1.7. Adverbs (ظرف). Subset of adverbs also belongs to closed class which includes: negative, affirmative, temporal, manner, quantity, and pronoun adverbs. Table-7 shows different types of closed class adverbs in Sindhi.

Table 7: Closed class adverb types in Sindhi

S.No.	Adverb Type	Example(s)
1.	Temporal Adverb	هينئر، ڪڏهن
2.	Manner Adverb	آهستي، ڏاڍو
3.	Negation Adverb	نه، ڪونه
4.	Quantity Adverb	ڪيترا، گهٽا
5.	Affirmative Adverb	ها، ٻلي
6.	Pronoun Adverb	هيئن، هوڏانڊن

4.1.8. Articles (حرف). Articles also belong to closed word class. Examples of articles include: ٻه، ٻڙ، ڪي، ته.

4.2 Open Word Classes in Sindhi

Major open classes in most of the languages are: nouns, verbs, adjectives, adverbs and interjections. Sindhi also has these open classes. Following sections discuss open Sindhi classes in detail.

4.2.1. Nouns (اسم). Three categories of Sindhi Nouns are: proper noun (اسم خاص), common noun (اسم عام) and abstract noun (اسم ذات). Like all languages of the world, noun in Sindhi is open word class. Table-8 shows examples of three types of noun along-with newly included words in these classes.

Table 8: Types of Sindhi Nouns

S.No.	Noun Type	Example	Newly Included Word(s)
1.	Proper Noun	احمد، ڪراچي، پاڪستان	انٽرنيٽ، مائڪروسافٽ، نوڪيا
3.	Common Noun	ملڪ، شهر، ڪتاب	اي ميل، اٽارني، ڪمپيوٽر
4.	Abstract Noun	اچاڻ، ڏاڻ، شاهوڪار	ڪنيڪٽيوٽي

4.2.2. Adjectives (صفت). Adjectives define properties of nouns. Most of the native Sindhi adjectives belong to closed class as cardinals, ordinals, multipliers, fractals and pronominal adjectives discussed in section 4.1.6. But there are examples of adjectives which are adopted from other languages. For example: in the sentence "احمد هڪ اينيميٽيڊ ويب سائٽ ناهي" the word "اينيميٽيڊ" is an adjective which is adopted and is transliterated form of English word "animated". The word خوبصورت and شاندار are also examples of adopted words in Sindhi adjectives.

4.2.3. Adverbs (ظرف). Temporal, manner, negation, quantity, affirmative, and pronoun adverbs usually belong to closed word classes as discussed in section 4.1.7. Space, noun and adjective adverbs belong to open class types. As nouns and subset of adjectives themselves are open class types so is the case with noun and adjective adverbs. For example: in the sentence "آن لائن هليو اڄ" the word "آن لائن" can be considered as space adverb or noun adverb.

4.2.4. Compound Verbs (مرڪب فعل). The only open class verb types in Sindhi are compound verbs. Compound verbs in Sindhi are formed by combining nouns, adjectives/adverbs and participles with verbs. Table-9 shows examples of compound verbs and newly formed compound verbs.

Table 9: Sindhi Compound Verbs

S.No.	Compound Verb	Description	Newly adopted Compound Verb(s)
1.	راند ڪرڻ، شادي ڪرڻ، عرض ڪرڻ	Formed by combining nouns with verbs	چيڪ ڪرڻ، اي ميل ڪرڻ
3.	خوش ٿيڻ، ناراض ڪرڻ	Formed by combining adjectives/adverbs with verbs	آن لائن ٿيڻ، آف لائن ڪرڻ
4.	ڪري پوڻ، بکيو ڇڏڻ، ٻڌي سگهڻ	Formed by Combining verbs and participles	ڪنيڪٽ ٿيڻ

4.2.5. Interjection (حرف ندا). Interjections convey emotions in sentences. There are a small number of interjections used in Sindhi. Few examples are: واه واه، هاءِهاءِ!، افسوس!، گهوڙا!، ڪاش!، اڙي! and شاباس. However, new emotions can be included at any time therefore, interjections belong to open class type.

5. Developing a Sindhi POS Tagset

According to EAGLES (Expert Advisory Groups on Language Engineering Standards) [16] guidelines for morpho-syntactic tagging of languages three different levels of constraints may be considered for POS tagging are: obligatory, recommended and optional. Following sections discuss and present obligatory and recommended tagsets for Sindhi according to EAGLES guidelines.

5.1. Obligatory POS Tagset for Sindhi

EAGLES guidelines recommend only one attribute as obligatory tag for every grammatical category. Therefore, the obligatory POS tagset for Sindhi according to EAGLES guidelines will be as given in Table-10.

The NU (numeral) tag is used to tag numeric values in text, PU (punctuation) is for tagging punctuation marks, as punctuation marks are treated as words in POS tagging. The tag 'R' (residual) is used for foreign words or mathematical formulae.

Table 10: Obligatory Tagset for Sindhi

N	[noun]	V	[verb]	AR	[article]
PN	[pronoun]	AV	[adverb]	NU	[numeral]
PP	[postposition]	C	[conjunction]	PU	[punctuation]
AJ	[adjective]	I	[interjection]	R	[residual]

5.2. Recommended POS Tagset for Sindhi

The recommended POS tags for widely recognized grammatical categories of Sindhi according to EAGLES guidelines are discussed below. Appendix-A shows recommended attributes and their values for various word classes. At some places due to language specific features of Sindhi these recommendations may differ from EAGLES guidelines.

Noun (N) can have three different tags: common noun (NC), proper noun (NP), and abstract noun (NA). These basic tags can be further extended to recommended and optional tags as per EAGLES guidelines. For example: an intermediate tag for common noun, masculine, plural, nominative can be N1121, in which every position represents the attribute;

number at that position represents one of the attribute values given in Appendix-A. The intermediate tag can have an equivalent final tag NCmsn. This method can be extended for every attribute and value of noun class. However, this will result in 36 different tags for nouns which is theoretically more complete and useful but practically infeasible. Due to increase in number of tags automatic tagging of corpus will be exponentially complex as this increase directly affects the accuracy of tagging algorithms [17]. Therefore three basic tags for nouns are considered. However, according to attribute values of nouns given in Appendix-A there can be any number of possible tags. Table-11 shows proposed tags for nouns. The intermediate tags shown in the table are generated according to EAGLES guidelines from subtype attribute values of Appendix-A. For instance: in N2000; N refers to noun, 2 is type of noun (proper noun), and three 0's show that none of the gender, number and case attributes is considered for tagging.

Table 11: Proposed tags for Sindhi nouns

S.No.	[TAG] Word Class	Intermediate TAG	Example
1.	[NC] Common Noun اسم عام	N1000	هي گهر <NC> اسانجو آهي.
2.	[NP] Proper Noun اسم خاص	N2000	ڪراچي<NP> تمام وڏو شهر آهي.
3.	[NA] Abstract Noun اسم صفاتي	N3000	هي سيٽ <NC> شاهوڪار<NA> ماڻهو آهي.

Seven different types of pronouns and their basic tags are: personal pronoun (PP), demonstrative pronoun (PD), wh-pronoun (PWh), reflexive pronoun (PRF), relative pronoun (PRL), co- relative pronoun (PCR), and indefinite pronoun (PI). Personal pronoun (PP) is further divided into first person (PP1), second person (PP2) and third person (PP3) types. In Appendix-A attribute DemType at (vi) is used only for demonstrative pronouns, which are further divided into proximate (PDP) and remote (PDR) types and are proposed to be considered for basic tagset. Total ten different tags are proposed for pronouns. In Sindhi demonstrative pronouns are also used as third person personal pronouns and this ambiguity needs to be sorted out in tagging algorithm. Table-12 shows different tags for pronouns with examples.

Verb (V) attributes are divided into ten different categories. Auxiliary and main verbs are considered in tagset design. According to transitivity type main verbs are divided into: transitive (VT), intransitive (VI), casual transitive (VC), casual transitive double (VCTD) and casual transitive double twisted (VCTDT). All transitive and intransitive verbs are further divided into: active and passive types. These active and passive forms are handled via voice attribute in recommended attributes given in Appendix-A. Verb participles are divided into five different types and are tagged separately in the tagset. Tags of these five participles are: present participle (VPPrs), past participle (VPPst), future participle (VPFutr), verbal noun (VPVNn) and conjunctive participle (VPConj). Table-13 shows different verb type tags.

Table 12: Proposed tags for Sindh pronouns

S.No.	[TAG] Word Class Example
1.	[PP1] 1st Person Pronoun ضمير متڪلم آئون <PP1> هلان ٿو.
2.	[PP2] 2nd Person PN ضمير حاضر اوهين <PP2> اچو ها ته سنو ٿئي ها.
3.	[PP3] 3rd Person PN ضمير غائب اهي <PP3> ڇوڪرا ڏاڍا تڪڙا آهن.
4.	[PDP] Demonstrative Pronoun Proximate ضمير اشارو ويجهو هيءَ <PDP> ڇوڪرو آهي.
5.	[PDR] Demonstrative Pronoun Remote ضمير اشارو ڏور هو <PDR> ڇوڪرو آهي.
6.	[PWh] Wh-Pronoun ضمير استفهام تنهنجو <PP2> هٿي ڪهڙو<PWh> ڪم.
7.	[PRF] Reflexive Pronoun ضمير مشترڪ مون <PP1>پاڻ<PRF>هي<PDP>ڪم ڪيو آهي.
8.	[PRL] Relative Pronoun ضمير موصول جهڙي <REP>ڪرٽي تهڙي <CRP> پرٽي.
9.	[PCR] Co-relative Pronoun ضمير جواب موصول جهڙي <PRL>ڪرٽي تهڙي <PCR> پرٽي.
10.	[PI] Indefinite Pronoun ضمير مبهم ڪونه ڪو <PI> ته ايندو.

Adjective (AJ) tags can be divided into: characteristic (AJ1), cardinal (AJC), ordinal (AJO), pronominal (AJP), aggregate (AJA), quantifier (AJQ), fractal (AJF) and multiplier (AJM). Adjective types and their tags are shown in Table-14.

Adverb (AV) tags include: temporal adverb (AVT), space adverb (AVS), manner adverb (AVM), negation (AVNEG/NEG), quantity adverb (AVQ), affirmative adverb (AVA), noun adverb (AVN), adjective adverb (AVAJ) and pronoun adverb (AVP). Table-15 shows adverb types and their tags.

Two types of postpositions (prepositions are not used in Sindhi) are: simple and compound and have (PP) and (PPC) tags respectively. Table-16 shows postposition types and their tags.

Conjunctions are syntactically divided into: coordinate (CC) and subordinate (CS) types.

Interjection has tag (I) and is not further divided into any type. Conjunction and Interjection types with tags are shown in Table-17.

Table 13: POS tags for Sindhi verbs

S.No.	[TAG] Verb Class Example
1.	[VIA] Active Intransitive Verb فعل لازمي معروف اُئون ڏوڙان <VIA> ٿو.
2.	[VIP] Passive Intransitive Verb فعل لازمي مجهول گاڏي ۾ چڙهجي <VIP> ٿو.
3.	[VTA] Active Transitive Verb فعل متعدي معروف مان خط لکان <ATR> ٿو.
4.	[VTP] Passive Transitive Verb فعل متعدي مجهول خط لکجي <PTR> ٿو.
5.	[VCTA] Active Casual Transitive Verb فعل متعدي بالواسطه معروف مان خط لکايان <ACT> ٿو.
6.	[VCTP] Passive Casual Transitive Verb فعل متعدي بالواسطه مجهول خط لکائجي <PCT> ٿو.
7.	[VCTDA] Active Casual Transitive Double Verb فعل متعدي بالواسطه ٻٽومعروف مان خط لکارايان <DCT> ٿو
8.	[VCTDP] Passive Casual Transitive Double Verb فعل متعدي بالواسطه ٻٽو مجهول خط <VCTDA> لکارائجي ٿو.
9.	[VCTDTA] Active Casual Transitive Double Twisted Verb فعل متعدي بالواسطه ٻٽودهرو مان خط لکارارايان < VCTDTA > ٿو
10.	[VCTDTP] Passive Casual Transitive Double Twisted Verb فعل متعدي بالواسطه ٻٽودهرومجهول خط لکارارائجي < VCTDTP > ٿو.
11.	[VAux] Auxiliary Verb فعل معاون مون خط لکيو آهي <VAux>
12.	[VPPrs] Present Participle اسم حاليه هوڏوڙندو < VPPrs > گهر ويو.
13.	[VPPst] Past Participle اسم مفعول اهو ڪتاب ڪاتل < VPPst > آهي
14.	[VPFutr] Future Participle اسم استقبال مون کي خط لکڻو <VPFutr> آهي.
15.	[VPVNn] Verbal Noun اسم فاعل گاڏي هلائيندڙ <VPVNn> کي روڪيو.
16.	[VPConj] Conjunctive Participle اسم معطوفي استاد ٻارن کي پڙهائي <VPConj> گهر ويندو.

Table 14: Adjective types and tags

S.No.	[TAG] Word Class
1.	[AJ1] Characteristic Adjective صفت عارف سلجٹو <AJ1> چوڪر آهي.
2.	[AJC] Cardinal عددي صفت اڪبر مون کان پنج <AJC> سئو اڏارا ورتا.
3.	[AJO] Ordinal قطاري صفت شمس ڏوڙ ۾ پنجون <AJO> نمبر آيو.
4.	[AJP] Pronominal Adjective ضميري صفت اسين <AJP> ماڻهو لاڙ جا.
5.	[AJA] Aggregative Adjective صفت مطلق اوهان مان هڪ به <AJA> انعام جي لائق ڪونهي.
6.	[AJQ] Quantifier صفت مقداري گهٽا <AJQ> ماڻهو گوڙ ڪندا.
7.	[AJF] Fractal عدد جا جزا پاءُ <AJF> ڪير وٺي آ.
8.	[AJM] Multiplier عدد جو پيرو تيل جي قيمت پنجوٽي <AJM> ٽي وڻي.

Table 15: Adverb types and tags

S.No.	[TAG] Word Class Example
1.	[AVT] Temporal Adverb ظرف زمان احمد هينئر < AVT > آيو آهي.
2.	[AVS] Space Adverb ظرف مڪان هيٺ < AVS > نه ويهو.
3.	[AVM] Manner Adverb ظرف تميز هي همراه ڪم ڪار جو چڱو < AVM > آهي.
4.	[AVNeg] Negation Adverb ظرف نفي سليم ڪونه < AVNeg > ايندو.
5.	[AVQ] Quantity Adverb ظرف مقدار ماڻهو گهٽا < AVQ > هوندا ته گوڙ ٿيندو.
6.	[AVA] Affirmative Adverb ظرف اثباتي تون پلي < AVA > هليو اچ.
7.	[AVN] Noun Adverb اسميه ظرف اڳيان < AVN > هليو اچ.
8.	[AVAJ] Adjective Adverb صفتي ظرف ڏاڍيان < AVAJ > نه ڳالهاءِ.
9.	[AVP] Pronoun Adverb ضميري ظرف پوءِ جيئن < AVP > توهين چئو.

Table 16: Postposition types and tags

S.No.	[TAG] Word Class
1.	[PP]Post Position حرف جر وڃي ڪڏ ۾ <PP> پئو.
2.	[PPC]Compound Post Position مرڪب حرف جر احمد اسان ڪانسواءِ <PPC> هليو ويو.

Table 17: Conjunction, Interjection types and tags

S.No.	[TAG] Word Class
1.	[CC] Coordinate Conjunction احمد ۽ <CC> سليم گڏجي آيا.
2.	[CS] Subordinate Conjunction ڏاڍو چيو مانس تنهن هوندي به <CS> هو ڪونه مڙيو.
3.	[I] Interjection حرف ندا ڪاش! <INJ> تون اچين ها.

Pronominal suffixes (PSx) are commonly used in Sindhi like few other south Asian languages. Due to complex nature, important morpho-syntactic structure and major role in semantics pronominal suffixes are considered as a different word class for POS-tagging purpose.

Three types of pronominal suffixes are nominal (PSxN), postpositional (PSxP) and verbal (PSxV). Other attributes are not considered in tagset being proposed; however, one can consider other attributes of Appendix-A and generate tags accordingly. Table-18 shows different types of pronominal suffixes and their tags.

Articles are assigned article (A) tag. Only few articles are there in Sindhi as discussed in section 1.1.8.

Residual (R) is considered for foreign words, formulas, acronyms etc. Residuals can be assigned

many tags according to their types but only one tag (R) is considered for all residuals.

Numerals are numbers which occur in text. These numbers are tagged as numeral (NU) tags.

Other tags considered for proposed tagset include DATE, Title (divided into Pretitle (PRT) and posttitle (POT)) and sentence marker (SM). Table-19 shows various tags and examples including article, residual, numeral, title, sentence maker and date.

Table 18: Pronominal suffix types and their tags

S.No.	[TAG] Word Class	Example(s)
1.	[PSxN] Pronominal Suffix with Nouns (Nominal Suffix) اسميه ضمير متصل	پٹم، پٹس، چاچهين
2.	[PSxV] Verbal Pronominal Suffix فعليه ضمير متصل	اٹم، اٹئون، لڪندم
3.	[PSxP] Postpositional Pronominal Suffix جري ضمير متصل	ڪين، ساڻس، وٽئون

Table 19: Miscellaneous Tags

S.No.	[TAG] Word Class	Example
1.	[AR] Article	به، ، ته، ئي، پڻ
	[R] Residual	IBM, ن.ڪ , Device
	[NU] Numerals	12, 445، ٢٣
	[PRT]Pre title	محترم، جناب، سائين
	[POT]Post title	صاحب، خان
	SM	<SM>.اٿي ته هلون
	DATE	12/12/2011, 2012 فيبروري 13

As discussed earlier, POS tagset design considerations depend on the purpose for which tagset is being designed. The intension of proposed tagset usage is to tag a corpus with necessary tags, so that it can be used for syntactic analysis and parts of speech usage analysis in Sindhi text. Therefore, necessary level of granularity is considered. For example: only two tags for verbs are considered: main verb (V) and auxiliary verb (VAux); all verb types other than auxiliary are tagged as main verbs. Verb participles are also tagged separately as discussed in section 1.1.5. Compound verbs are not treated separately but their individual parts are separately tagged.

The proposed tagset is shown in Table-20. The tagset contains total 52 tags. The table shows proposed tags with intermediate tags; these intermediate tags are generated by using attributes and values given in Appendix-A as per EAGLES guidelines.

6. Conclusion

Discussion on various issues about the open and closed Sindhi word classes, tagset design issues in general, and Sindhi tagset design in particular, proposed tagset and attribute values of Sindhi word classes will provide basis for further research in various fields of Sindhi NLP. The proposed tagset discussed and presented provides basis for POS tagging and tagged Sindhi corpus construction. While designing the tagset necessary granularity level is considered to cope with the basic word level syntactic information and is therefore useful for parts of speech usage analysis in Sindhi corpus. The syntactic analysis of tagged corpus with these tags is also possible. The tagged corpus can also be used as training corpus for automatic grammar learning applications. By using the EAGLES guidelines and the word class attribute values of Appendix-A automatic tagset generation can also be implemented depending on the NLP application requirements.

Acknowledgement

Special thanks go to Dr. Ghulam Ali Allana and Dr. Hameedullah Kazi for their guidance, suggestions and valuable comments.

References

[1] J. A. Mahar, , G.Q. Memon. "Rule Based Part of Speech Tagging of Sindhi Language," in proc. *International Conference on Signal Acquisition and Processing, ICSAP 2010, Bangalore, India*, February 9-10, 2010, pp.101-106.

[2] J. A. Mahar, , G.Q. Memon. “Sindhi Part of Speech Tagging System using WordNet”, *International Journal of Computer Theory and Engineering*. 2010. 2(4): 538-545.

[3] J.A. Mahar, H. Shaikh, A.R. Solangi. “Comparative Analysis of Rule Based, Syntactic and Semantic Sindhi Parts of Speech Tagging Systems”. *International Journal of Academic Research*. 2011. Vol. 3. No. 5.

[4] M. Rahman. “Towards Sindhi Corpus Construction In *Linguistics and Literature Review* 63 Vol. 1 No 1, 2011 pp. 74-85. UMT Lahore Pakistan.

[5] S. Jan, ed. *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press. 1990.

[6] L. Löfberg. , D. Archer., S. Piao., P., Rayson., T. McEnery., K., Varantola., J-P. Juntunen., “Porting an English semantic tagger to the Finnish language”. In *Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) Proceedings of the Corpus Linguistics 2003 conference*. UCREL technical paper number 16. UCREL, Lancaster University. 2003. pp. 457- 464.

[7] G. Leech., R. Garside and E. Atwell. *The automatic grammatical tagging of the LOB corpus*. Newsletter of the International Computer Archive of Modern English. 1983. 7, 13-33.

[8] M. P. Marcus, M. A. Marcinkiewicz, B. Santorini. "Building a large annotated corpus of English: the penn Treebank", *Computational Linguistics*. 1993. v.19 n.2.

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

[10] CRULP, *Urdu Part of Speech Tagset*. Center for Research in Urdu Language Processing. National University of Computer and Emerging Sciences. 2007. Lahore Pakistan.

[11] S. Manish. and P. Bhattacharyya. "Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information Without Extensive Linguistic Knowledge". In *Proceedings of the International Conference on NLP (ICON08)*, Pune, India. 2008.

[12] I. Rabbi., M. A. Khan, and R. Ali. "Developing a tagset for Pashto part of speech tagging". In *Proceedings of the International Conference on Electrical Engineering*.2008. pp. 1-6.

[13] Aglsoft. "Punjabi Part of Speech Tagger". 2009. Retrieved (February 2012).
Available: http://http://punjabi.aglsoft.com/?show=tagger

[14] E. Trumpp. *Grammar of the Sindhi Language*. London-Leipzig. 1872.

[15] G. A. Allana "Sindhi boli jo tashrehi grammar (Descriptive Grammar of Sindhi Language)". Sindhi Language Authority, Hyderabad, Pakistan. 2010.

[16] G. Leech., A. Wilson. *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora.* Istituto di Linguistica Computazionale, Pisa, Italy. 1996.

[17] Meyer, Chalres. F. *English corpus linguistics: An introduction*. Cambridge University Press. 2002. pp. 91-91.

Table 20: Proposed Sindhi POS Tagset

S.No.	TAG	Intermediate Tag	Word Class
1	NC	N1000	Common Noun
2	NP	N2000	Proper Noun
3	NA	N3000	Abstract Noun
4	PP1	P100010	1st Person Pronoun
5	PP2	P100020	2nd Person PN
6	PP3	P100030	3rd Person PN
7	PDP	P200001	Demonstrative PN Proximate
8	PDR	P200002	Demonstrative Pronoun Remote
9	PWh	P300000	Wh-Pronoun
10	PRF	P400000	Reflexive Pronoun
11	PRL	P500000	Relative Pronoun
12	PCR	P600000	Co-relative Pronoun
13	PI	P700000	Indefinite Pronoun
14	V	V1000000000	Verb
15	Vaux	V2000000000	Passive Intransitive Verb
16	VPPrs	V1000000101	Present Participle
17	VPPst	V1000000102	Past Participle
18	VPFutr	V1000000103	Future Participle
19	VPVNn	V10000002000	Verbal Noun
20	VPConj	V10000003000	Conjunctive Participle
21	AJD	AJ1000	(Characteristic) Adjective
22	AJC	AJ2000	Cardinal
23	AJO	AJ3000	Ordinal
24	AJP	AJ4000	Pronominal Adjectives
25	AJA	AJ5000	Aggregative Adjectives
26	AJQ	AJ6000	Quantifier
27	AJF	AJ7000	Fractal
28	AJM	AJ8000	Multiplier
29	AVT	AV1	Temporal Adverb
30	AVS	AV2	Space Adverb
31	AVM	AV3	Manner Adverb
32	AVNeg	AV4	Negation Adverb
33	AVQ	AV5	Quantity Adverb
34	AVA	AV6	Affirmative Adverb
35	AVN	AV7	Noun Adverb
36	AVAJ	AV8	Adjective Adverb
37	AVP	AV9	Pronoun Adverb
38	PP	PP1	Post Position
39	PPC	PP2	Compound Post Position
40	CC	CC1	Coordinate Conjunction
41	CS	CC2	Subordinate Conjunction
42	I	I	Interjection
43	PSxN	PS10000	Pronominal Suffix with Nouns (Nominal Suffix)
44	PSxV	PS20000	Verbal Pronominal Suffix
45	PSxP	PS30000	Postpositional Pronominal Suffix
46	AR	AR	Article
47	R	R0	Residual
48	NU	NU	Numerals
49	PRT	PRT	Pre title
50	POT	POT	Post title
51	SM	SM	Sentence Maker
52	DATE	DATE	DATE

Appendix-A
Sindhi Part-of-Speech Tags and Attributes

Noun (N) Attributes	**Value(s)**		
(i) Type:	1. Common	2. Proper	3. Abstract
(ii) Gender:	1. Masculine	2. Feminine	
(iii) Number:	1. Singular	2. Plural	
(iv) Case:	1. Nominative	2. Oblique	3. Vocative

Verb (V) Attributes	**Value(s)**		
(i) Type:	1. Main	2. Auxiliary	
(ii) Gender:	1. Masculine	2. Feminine	
(iii) Number:	1. Singular	2. Plural	
(iv) Person:	1. First	2. Second	3. Third
(v) Transitivity	1. Transitive 2. Intransitive 3. Casual Transitive 4. Casual Transitive Double 5. Casual Transitive Twisted		
(vi) Finiteness:	1.Finite	2. Non Finite	
(vii) Participle Type:	1. Tense Participle 2. Verbal Noun 3. Conjunctive		
(viii) Voice:	1. Active	2. Passive	
(ix) Mood / Word Form:	1. Subjunctive	2. Imperative	3. Presumptive 4. Counter Factual
(x) Tense:	1.Present	2. Past	3. Future

Adjective (AJ) Attributes	**Value(s)**		
(i) Type:	1. Descriptive 4. Pronominal 7. Fractal	2. Cardinal 5. Aggregate 8. Multiplier	3. Ordinal 6. Quantifier
(ii) Gender:	1. Masculine	2. Feminine	
(iii) Number:	1. Singular	2. Plural	
(iv) Case:	1. Nominative	2. Oblique	3. Vocative

Pronoun (PN) Attributes	**Value(s)**		
(i) Type:	1. Personal 4. Reflexive 7. Indefinite	2. Demonstrative 5. Relative	3. Wh 6. Co-relative
(ii) Gender:	1. Masculine	2. Feminine	
(iii) Number:	1. Singular	2. Plural	
(iv) Case:	1. Nominative	2. Oblique	
(vi) Person:	1. First	2. Second	3. Third
(v) DemType:	1. Proximate	2. Remote	

Adverb (AV) Attributes	**Value(s)**		
(i) Type:	1. Temporal 4. Negation 7. Noun	2. Space 5. Quantity 8. Adjective	3. Manner 6. Affirmative 9. Pronoun

Postposition (PP) Attributes	**Values(s)**	
(i) Type:	1. Simple	2. Compound

Conjunction (C) Attributes	**Value(s)**	
(i) Type:	1. Coordinate	2. Subordinate

Interjection (I)	

Pronominal Suffix Attributes	**Value(s)**		
(i) Type:	1. Nominal	2. Postpositional	3. Verbal
(ii) Gender:	1. Masculine	2. Feminine	
(iii) Number:	1. Singular	2. Plural	
(iv) Case:	1. Nominative	2. Oblique	3. Agentive
(v) Tense:	1. Present	2. Past	3. Future

Article (AR)	

Punctuation Attributes	**Value(s)**
(i) Type:	1. Period 2. Comma 3. Semicolon 3. Colon 4. Dash (Long) Hyphen) 5. Hyphen 6. Ellipsis 7. Question Mark 8. Exclamation 9.Opening Inverted Comma 10. Closing Inverted 12. Opening Bracket 13. Closing Bracket

Residual (R) Attributes	**Value(s)**
(i) Type:	1. Foreign Word 2. Formula 3. Symbol 4. Acronym 5. Abbreviation 6.Unclassified
(ii) Gender:	1. Masculine 2. Feminine
(iii) Number:	1. Singular 2. Plural

Numeral (NU)	

Title Attributes	**Value(s)**
(i) Type:	1. Pre Title 2. Post Title

Date	

Sentence	