Southern Journal of Arts & Humanities, volume 1 (2) 2023

A Hierarchical Part of Speech Tagset for Saraiki Language

Mubasher Hussain Malik[1], Hamid Ghous[2], Sumaira Perveen[3]

[1,3]Vision, Linguistics & Machine Intelligence Research Lab Multan, Pakistan

[2]Australian Scientific & Engineering Solutions, Sydney New South Wales, Australia

Abstract

Human languages are complex due to the diverse nature of expression whether in spoken or written forms. Natural Language Processing (NLP) combines the power of linguistics and Artificial Intelligence (AI) techniques, which enables computers to understand natural languages as humans do. Each language has its own unique set of grammar, syntax, terms, slang, and rules. Language processing involves certain tools which play a vital role in the construction, analysis, and manipulation of any language. Part of Speech (POS) tagging is one of the essential and basic processes for other applications in NLP. POS tagging is the process of classifying words into their parts of speech like nouns, verbs, prepositions, adverbs, etc. to a word. POS tagging involves the use of a proper POS tagset used to label distinct parts of the text with grammatical annotations. This helps to identify linguistic features of a word, phrase, or discourse in the text corpus during the annotation process. Saraiki language (SL) is one of the ancient regional languages of central Pakistan at present This research work focuses on the development of an SL POS tagset that will help the community to manipulate SL resources digitally. SL tagset consists of forty-seven classes of words along with sub-classes and their tags.

Keywords: Natural Language Processing, Language Processing, Saraiki Language, POS tagging, Tagset

Email: malikmubasher@gmail.com

1. Introduction

Saraiki Language (SL) is an Indo-Aryan language that is mainly spoken and understood over a large geographical area of Pakistan (Jan, 2016). Mostly around twenty-six million people spoke SL in different areas of Pakistan (Final Results (Census-2017), 2022). SL Is majorly spoken in South Punjab, Eastern Balochistan, Northern Sindh, Dera Ghazi Khan, Dera Ismail Khan, and some areas of the Indus Valley (Jumani, 2011). SL stands 61st as the world's largest language and 4th among Pakistani local languages such as Punjabi, Sindhi, and Pashto (Thomas, 2002). There are different dialects of SL available such as Rohi, Multani, Thalli, Rajanpuri, Majhi, and Shahpuri (MacKinlay, (2005). SL text is written from right to left using Arabic and Urdu Language Scripts (Jan, 2016). Most of the alphabet of Saraiki also available in Arabic and Urdu Languages (Kausar, 2015), (Abbas, 2012).

Unfortunately, in the recent few years, no attention is given to the digitization of SL resources. It is a very undeveloped language digitally and it needs time to develop a hierarchical part of speech tags for SL. Technological advancements in the area of Artificial Intelligence (AI), Natural Language Processing (NLP), and Computational Linguistics (CL) create an opportunity to make SL resources in a modern way (Skeppstedt, 2014). Machine Learning (ML) and Deep Learning (DL) contributed a lot to linguistics study (Muller, 2020), (Pfeiffer, 2020). Digital resources to store linguistics data such as words, sentences, linguistics rules, tag sets, semantics, forms, and synset are available for different digitally rich languages (Ahmed,

2015). Each language consists of alphabets, word creation rules, grammatical forms and rules, sentence formation, and other linguistic level approaches to make language understandable globally. Progress in the area of linguistics diversity and representation using NLP requires the development and implementation of resources. These resources include tools, techniques, procedures, corpus development, and development of such algorithms which can enhance the linguistics study (van Esch, 2022).

One of the more important and vital parts of linguistics research is Part of Speech (POS) tagging. POS tagging is the process of assigning grammatical tags to each word in the text (Naseem, 2017). POS tagger is a tool used to assign proper tags to each word. Choosing an appropriate tagset is a preliminary and vital task for successful POS tagging. A tagset needs to be able to encode the grammatical distinctions that are of interest for further steps in natural processing or linguistic research while allowing for efficient and accurate automatic tagging (MacKinlay, (2005). Tagset is one of the prominent and important resources which can help to mark different syntactical and semantic units of language (Ahmaed, 2014), (Rehman, 2011). As mentioned above, SL is a resource-poor language. There are several NLP applications such as text mining, information extraction and retrieval, text generation, machine translation, and text summarization, which require annotated corpus (Lewinski, 2017), (Samuel R Bowman, 2015). The most important task in the POS tagging process is the selection of an appropriate tagset (Khanam, 2013). The quality of linguistics information extraction depends upon the quality of the tagset used for annotation (Sheremetyeva, 2021)

In this research, after an extensive literature review of different language tagset, an SL tagset was introduced that could be used for POS tagging of Saraki text. The rest of
the paper is structured as follows: section 2 comprises an extensive literature review. Section 3 discussed Saraiaki Tagset Design. Finally, section 4 discussed the conclusion and future work.

2. Literature Review

Researchers contribute a lot to the field of NLP. The literature related to tagset can be divided into two categories: (i) Research related to SL resources, very limited research work contributed by researchers. In the year 2021, a hieratical SL tagset for POS tagging was introduced which consists of forty-eight tags. The tagset was developed using morphological, lexical, and syntactic annotation (Asghar, 2021) while the focus of this work is on generic tagset for annotation of SL text.

(ii) Several other tagsets of different languages were reviewed. This tagset differs from each other in terms of morphological and syntactical features, tag definitions, and granularity of the tag. Urdu POS tagsets introduced as Hardie's tagset (Baig, (2020)), Sajjad and Schmid tagset, and the CLE POS tagset (Sajjad, 2009), (Ahmed, 2015). Dogri tagset and Punjab tagset introduced in 2018 and 2019 (Kumar S. , 2018), (Kumar D. &., 2019). A Pashto tagset was introduced in 2008 (Rabbi, 2008). A hierarchal Kashmiri tagset was developed in 2009 (Kak, 2009). A Czech morphological was introduced in 2011 (Jakubcek, 2011). For the Arabic language corpus, Stanford Arabic parser tagset and POS tagset for modern standard Arabic (Diab, 2007), Arabic media system tagset (Habash, 2010), and Morphological annotations quranic Arabic tagset introduced (The Quranic Arabic Corpus). A Bulgarian POS tagset introduced by the Department of Computational Linguistics for Bulgarian corpora (The Bulgarian National Corpus, 2017-2019). Indian POS tagset developed for different Indian languages Machine Translation (Dandapat, 2009). A Nelralec POS tagset was developed for Nepali corpora for the categorization system of manual and automated analysis of morphosyntactic units in the Nepali language (Shrestha, 2021). An Irish POS tagset was developed for Irish Corpora using a POS tagger for Irish Finite State Morphology and Constraint Grammar Disambiguation (Kilgarriff, 2006). Turkish POS tagger introduced for Turkish corpora annotated by the MaltParser with the pre-trained Turkish model (Ery\ugit, 2008). A

Tajik POS tagset was introduced for Tajik Corpus. A tag entry consists of a lemma and number describing POS which have assigned the number from 01 to 16 according to 16 categories (Dovudov, 2012). A Vietnamese POS tagset was introduced for Vietnamese corpora annotated by the vnTagger (Pham, 2009). An Indian language Kannada POS tagger was introduced for the Kannada language which is a low-resource south Indian language (Swaroop, 2019). A Persian POS tagset was developed for Persian corpora annotated by a POS tagger based on Persian Syntactic Dependency Treebank (Rasooli, 2013).

The next section demonstrates the development of the SL Tagset used to annotate SL text.

3. Saraiki Tagset Design

SL is one of the resource-poor languages (Hussain, 2016). As mentioned above, very limited work has been done for the development of SL digital resources. This research contributes to the development of a hierarchical tagset annotated for the Saraiki Language corpus using a POS tagger. There are three design structures of Tagsets. The most common type of tagset design is Flat Tagset. These tagsets are easier to process. These tagsets are unable to capture a greater level of granularity without a very long number of independent labels (Sankaran, 2008). Furthermore, Hierarchical tagset design tagsets are structured relative to one another. A word can be classed as a verb or noun first. If a word is a verb, it can be checked for class (Sankaran, 2008). Finally, the Decomposability Tagset design allows distinct features to be encoded in tags by independent sub-strings (Baskaran, 2008).

The focus of this research is the development of an SL Tagset annotated for SL Corpora using a POS tagger. SL tagset is based on Perso-Arabic script. Although, most of the regional languages found similar their writing scripts may differ from each other. SL tagset consists of twelve main categories and thirty-eight subcategories tags.

3.1 Tag Structure

SL Tagset consists of twelve Word Classes (WC) also known as classes of words, such as Noun (NN), Pronoun (PN), Verb (VB), Adverb (AV), Auxiliary (AX), Nominal Modifier (NM), Adposition (AP), Particle (PT), Interjection (IJ), Symbol (SM), Foreign Fragments (FF) and Morphological Tags. WC is represented by two capital letters such as NN=Noun while Sub Classes (SC) are represented by Three capital letters such as NNP=Proper Noun. The following sections described the detail of all WC and SC developed for the SL tagset.

3.2 Noun (NN)

A Noun اسم is a term that relates to people, animals, objects, ideas, concepts, feelings, and other things. The Noun is intended for Common Noun (NNC) and Proper Noun (NNP). The MC, SC, tag, and examples of Nouns (NN) are shown in Table 1.

MC	SC	Tag
Noun (NN) اسم	Common Noun اسم نکرہ	NNC
	Proper Noun اسم معرفہ	NNP
Examples		
چھوہر، جہاز، درخت، زمین، میز، ہسپتال، والد، شہر		
ایہ/ PDM > چھوہر/NN >/چنگاJJہے/VBF		
قائدا عظم، اللہ تعالی، چناب، مینارِ پاکستان		
<قائدا عظم/NNP> پاکستان/NNP> دے/PSP بانی/JJ ہَن/AUXT		

Table 1: SL Noun and its Sub Classes

Furthermore, a single POS tag is assigned to words formed in the group while having a single meaning. For example, "قائدا عظم" has a single entity which is a combination of two words to form some meaning. For this word a tag NNP is assigned. A few standard examples of Common Noun are '**چھوہر**', '**درخت**', and '**میز**'. These adverbial nominals are considered into the common noun (NNC) tag rather than establishing a separate tag to handle the semantic variation between two groups of terms.

3.3 Pronoun (PN)

Pronouns (**ضمیر**) are words that act as nouns and can be used to replace a noun or a noun phrase.

There are eight different types of pronouns. As an alternative for the noun, the personal pronoun (PRP) develops. Some examples are **میکوں ,ایہ، او** shown in Table 2.

MC	SC	Tag
Pronoun (PN) ضمیر	Personal Pronoun **(اسم ضمیر)**	PREP
	Demonstrative Pronoun **(ضمیر اشارہ)**	PDM
	Possessive Pronoun **(ضمیراضافی)**	PRICE
	Relative Demonstrative Pronoun **(ضمیراشارہ موصولہ)**	PRD
	Relative Personal Pronoun **(اسم موصول)**	PRR
	Reflexive Pronoun **(ضمیر معکوسی)**	PRF
	Reflexive Apna Pronoun **(ضمیر اپنا معکوسی)**	APNA
	Interrogative Pronoun **(ضمیر استفہامیہ)**	INTP
Examples		
میں،میکوں، توں، تہاکوں، اساں،تیکوں، اے، اوں، اُوکوں،اِوکوں، کوڑ، کڈاں،کتھے، کیں		
<اے/PRP>کینْدی/PRPملکیت/ NNہے/AUXA		
ایہ، او، اِیں، اُوں، اِیکوں، اُکوں، اِدے، اُدے، اِیویں، اُویں، کیویں		
<ایہ/PDM>میڈی/PRPکتاب/NNہے/AUXA		
مِیڈا، تِیڈا، مِیڈی، تِیڈی، ساڈا، ساڈی، تہاڈا، تہاڈی، اِیْندی، اُنْدی، اُنھاں دا، اِنھاں دی		
<اوPDM>/تواڈےPRS/چنگےJJ/ کپڑےNN/بِن/AUX		
جہڑا،جیں،جیکوں ،جو		
<جیکوں /PRD> توںPRP/آکھاVBF/ہےAUXT/		
جہڑا، جیکوں، جنہاں، جیویں، جیں، جتھے، جڈاں		
ایہ/PDM او/PDM گھر/NN ہے/VBF <جہڑا/PRR> علی/ NN نے/PSP بٹایا /VBI ہئ/AUXT		
خود، اپنے آپ، آپ		
میں/ PRP<اپنے/ PRF آپکوں/ PRF> تیرنا/VBI		
سکھایا/VBI		
اپنا، اپنے، اپنی		
چُپ NN/کر SCK/تے< SCK/اپنا /APNA>کَم/ NN کرو/VBF		
کینڈی، کیں، کینکوں		
ایہ/PDM کھڑکی/NN<کیں/INTP> توڑی/VBF بے/VBF		

Table 2: Subcategories of Pronouns with Tags and Examples

The demonstrative pronoun ((**ضمیر اشارہ**)) comes before a noun as a specifier. As given below:

<ایہ/PDM>میڈی/PRPکتاب/NNہے/AUXA

The same form **ایہ** can be used as a personal pronoun (PRP) or demonstrative (PDM). These words can be distinguished syntactically. Possessive pronouns **(ضمیراضافی)** are pronouns that convey a relationship of ownership. Some examples are **'مِیڈا', 'تِیڈا'**. Reflexive pronouns **(ضمیر معکوسی)** are used for mention to oneself. Some examples are ' خود، اپنے آپ ' as shown in Table 3. An example is given below:

میں/PRP <اپنے/PRF آپکوں/PRF> تیرنا/VBI سکھایا/VBI

3.3 Verb (VB)

A verb (فعل) generally implies an action ("go," "eat"), an event ("to alter" (itself," "to sparkle"), or a state of being ("to modify" (itself," "to glitter") (survive "live", "stand").

MC	SC	Tag
Verb (VB) (فعل)	Main Verb Infinite	VBI
	Main Verb Finite	VBF
Examples		
کرنا، درھکھٹا، ونجٹا، آونٹا		
علی/NNP کوں/PSP <درھکٹا/VBI> پوسی/AUXT		
درھکدا،درھکدے، دھکدی، درھگا، درھکے، درھکی، درھکوں، درھک، درھکیں، درھکو		
علی/NNP تیگّھا/JJ <درھکدا/VBF>ہے/AUXA		

Table 3: Subcategories of Verbs with Tags and Examples

Sub-categories of Verb are described in detail below as shown in Table 4:

SC	Tag
Past Participle (فعل ماضی)	VBPA
Present Participle (فعل حال)	VBPR
Future Participle (فعل مستقبل)	VBPF
Examples	
درھکّا، لکھا، پڑھا	
اسلم/NNنےPSPخط/NN <لکھا/VBPA>	
درھکدا	
علی/NNP تیکّھا/JJ <درھکدا/VBPR>ہے/AUXA	
لکھ سی، پڑھ سی، کھاسی، پی سی	
او/ PDMروٹی/ NN <کھاسی/VBPF>	

Table 4: (Verb and its Types)

3.4 Adverb (AV)

An adverb **(فعل مطلق)** is a word that modifies verbs, adjectives, other adverbs, clauses, and sentences and is part of a set of words called adverbs.

There are two types of adverbs: Common adverbs (RB) and negation (NEG). Examples are shown in Table 5.

MC	SC	Tag
Adverb (AV) (فعل مطلق)	Common Adverb	RB
	Negation Adverb **(فعل نہی)**	NEG
Examples		
تقریباً، ہولے، کیّوں، ایویں، چیویں، کیویں، وَل		
<تقریبًا/RB> ڈو/CD کروڑ/CD سیاح/NN اوکوں/PRP ڈیکھنّ/VBI آنڈے/AUXA ہن/AUXT		
نئیں، نہ، مت جا		
زبان/NN دے/PSP بلاونّ/VBI اِچ/PSP تھوڑی/JJ وی/PRT پیڑ/NN <نِئیں/NEG> تھینڈی/VBF		

Table 5: (Adverb and its Types)

3.5 Auxiliary (AX)

Aspectual (AUXA), Progressive (AUXP), Tense (AUXT), and Modals are the four types of auxiliaries described here (AUXM). Examples of all these auxiliaries are shown in Table 6.

MC	SC	Tag
Auxiliary (AX) فعل معاون	Aspectual Auxiliary	AUXA
	Progressive Auxiliary	AUXP
	Tense Auxiliary	AUXT
	Modal Auxiliary	AXUM
Examples		
آنْدا، آنْدی، آنْدے، آیا، آےَ، آئ، آوّو، آ، آویں، گیا، گئ، گےَ، گِھنو		
ذہن/NN اِچ/PSP تصویرکشی/NN دا/PSP انوکھا/JJ نقشہ/NN ابھر/VBF <آیا/AUXA>		
او/PDM بندے/NN چہڑے/PRR سگریٹ نوشی/NN چھوڑ/NN گےَ/AUXA ہَن/AUXT		
کریندا‌پیا، کریندی پئ، کریندےپےَ		
کیّوں/RB ظلم/NN <کریندے پےَ/AUXP>ہوّے/AUXT		
ہے، ہئ، ہَن		
کہں/PRP نے/PSP چنگا/JJ آکھا/VBF <ہے/AUXT >		
سکدا، سکدی، سکدے، ملٹا، پوّٹا، ہوّٹا		
بیماریاں/NN توں/PSP محفوظ/JJ ریہا/VBF ونچ/VBF <سکدا/AUXM> ہے/AUXT		

Table 6: (Auxiliary and its Types)

3.6 Nominal Modifiers (NM)

The nominal modifier is a category that is used to modify nouns and pronouns in sentences. Nominal Modifier include adjectives (JJ) e.g. ‘چنگا’, quantifier e.g. ‘ہہوں’, cardinal e.g. ‘ہِک’, and ordinal (OD) e.g. ‘ڈوجھا’,

Examples are given below in Table 7:

SC	Tag
Adjective Possessive (صفت ذاتی)	JJ

Adjective of Quantity (صفت مقداری)	Q
Adjective of Number/ Cardinal	CD
Ordinal	OD
Examples	
اُچّا، بھیڑا، روڑ، سوہݨاں، چنگا، رتّا، موٹا، ذہین	
او/PDM بِک/CD <ذہین/JJ> چھوہر/NN ہے/VBF PU/۔	
بہوں، تھوڑا، چتنا، کجھ، چوکھا، اکثر	
میرے/PRS کولوں/VBI <تھوڑا/Q> چیا/PSP پاݨی/NN ہے/VBF PU/۔	
ہک، ڈُو، تَرے، ۳،5	
میز/NN تے/SCK <تَرے/CD> کتاباں/NN پیاں/PSP ہن/AUXT	
پہلا، پہلی، پہلے، ڈُوجھا، تَریجھا، ڈُوجھیs	
<پہلا/OD> چھوہر/NN کتاب/NN پڑھدا/VBI ہے/VBF PU/۔	

Table 7: Nominal modifiers and their types

3.7 Adposition (AN)

Two types of Adposition: pre-and postpositions. Some Examples are: ‘اِچ’,’ تک’, ’دا’. Examples are shown in Table 8:

MC	SC	Tag
Adposition (AN)	Preposition **(حرفِ ربط)**	PRE
	Postposition **(حرفِ اضافت)**	PSP
Examples		
سوائے، اِچ، تے، توں		
<سوائے/PRE > چُپّ/NN دے/PSP کُجھ/Q نہیں/NEG ہے/VBF PU/۔		
پیا، دا، دی، تک، بغیر/سوا		
قرطبہ/NNP دی/PSP مشہوری/NN پرے/JJ پرے/JJ <تک/PSP> ہئ/VBF		

Table 8: Adposition and their types

3.8 Particle (PT)

In the Saraiki, Particle **(حرف)** is a function word that is used in conjunction with another word or phrase to convey meaning. Particle types and examples are shown in Table 9:

Table 9: Particles and their types

MC	SC	Tag
Particle (PT) (حرف)	Conjunction **حرف عطف**	CP
	Coordinate Conjunction **حرف جار**	CC
	Subordinate Conjunction **حرف علت**	SC
Examples		
تے، کر،دے		
بازار/NN توں/PSP سبزی/NN گِھن/VBF <تے/CP> آؤ/VBF		
یا، تے، پہلے، وتّ، جہڑے ویلے، لیکن		
میکوں/PRS چاہ/NN پسنْد/JJ ہے/VBI <لیکن/CC> ڈُودھ/NN پسنْد/JJ نِہیں/NEG		
کیونکہ، جو، تاں جو، بشرط اے، تے		
زعفران/NN ہر/JJ مریض/NN استعمال/NN نِہیں/NEG کر/VBF سکدا/AUXM <کیونک/SC> اے/PRP دنیا/NN دا/PSP مہنگا ترین/JJ مسالہ/NN ہے/VBF		

3.9 Interjection (IJ)

The interjection (INJ) is usually used at the beginning of a sentence. In the tagset, it is treated as a distinct category as shown in Table 10. Examples are ‘واہ’, ‘سبحان اللہ’ etc.

MC	SC
Interjection (IJ) فجائیہ	--
Examples	
واہ، کاش، سبحان اللہ، ماشااللہ، السلام علیکم، اوّے،اوئے	
<اُف/INJ>ایہ/PDM تے/CP اتنا/Q پراݨا/Q ویلا/NN ہے/VBF	
<کاش/INJ> ایہ/PDM گھڑی/NN میڈی/PRS ہوندی/AUXM	

Table 10: Interjection and its Types

3.10 Symbol (SM)

Punctuation (PU) and other symbols (SYM) are two types of symbols as shown in Table 11.

MC	SC	Tag
Symbol (SM) علامت	Common Symbol	SYM
	Punctuation Symbol	PU
Examples		
، +، %، $، ء		
NN/اولمپکس PSP/دی <SYM/ء> CD/۱۸۹۲		
.،؟، ()، :، ؛ !		
NN/گھمڻ PSP/دا NN/شوق PRP/میکوں NN/نِکّے PSP/لا VBF/ہے <PU/.>		

Table 11: Symbols and their Types

3.11 Foreign Fragments (FF)

Foreign Fragment (FF) is a tag in Residual that covers all foreign language components. For example, ‘سبحان الله’ is an Arabic fragment to which we have assigned the interjection tag (INJ) as shown in Table 12.

MC	SC
Foreign Fragment (F)	--
Examples	
بسمِ الله الرحمن الرحيم، ماشاالله، سبحان الله، مسلمان زيستن، ويحروّمُ علهِمُ الخبائث	
جيندے PRR/تے PSP/تحرير NN/ہئ VBF/<ويحروّمُ علهِمُ الخبائث/FF >	

Table 12: Foreign Fragment

3.12 Morphological Tags

Saraiki is a morphologically rich language by its different well-structured sentences and dialects. Morphology is a subfield of computational linguistics concerned with the study of the internal structure of words (Rizvi, (2005, August)). As in English, a fixed morpheme is used to convert the present verb to the past, and there are a few more morphemes that are utilized to generate the plural. When performing morphological analysis, the difference between inflection and derivation is generally determined based on the affixations. In the context of Saraiki, the transformation from singular to plural is accomplished by morpheme variation. An example of the Saraiki nouns is given in Table 13:

Masculine/Singular	Suffix	Plular
دادا	"ا"،"ے"	دادے
ماما	"ا"،"ے"	مامے
گھوڑا	"ا"،"ے"	گھوڑے

Table 13: Masculine Singular to Plural

Saraiki has fixed morpheme /-ے / for masculine plurals but very in the case of feminine gender. For the plural morpheme of the feminine, we take examples from central dialects as shown in Table 14, and used nouns in these examples pronounced differently in different dialects (Atta, (2019)).

Feminine /Singular	Suffix	Plular
دادی	"ی"،"ں"	دادیاں
مامی	"ی"،"ں"	مامیاں
گھوڑی	"ی"،"ں"	گھوڑیاں
دھی	"ی"،"ں"	دھیاں

Table 14: Feminine Singular to Plular

For the feminine plural, Saraiki has a fixed morpheme /-ں /.

The validation of this work was performed through an expert in Saraiki language from different regions in South Punjab, Pakistan.

4. Conclusion and Future work

In conclusion, we have introduced a new part of speech (POS) tagset for the Saraiki Language. It is built on the back of a critical study of a prior tagset proposal. In this paper, we design a hierarchal tagset for the Saraiki language. This hierarchical architecture enables the creation of

interoperable and adaptable language-specific tag sets. Our tag structure defines the complete hierarchy of the tagset and this created tagset might be used as a preliminary step for more Saraiki NLP research. For future work, we will improve this existing hierarchal tagset and make a Saraiki POS tagger to automatically annotate the Saraiki corpora.

5. Saraiki POS Tagset

Table 15: Category, Types, and Tags

SC	SC	POS Tag
Noun (NN) اسم	Common Noun **اسم معرفہ**	NN
	Proper Noun **اسم نکرہ**	NNP
Pronoun ضمیر	Personal Pronoun **(اسم ضمیر)**	PRP
	Demonstrative Pronoun **(ضمیر اشارہ)**	PDM
	Possessive Pronoun **(ضمیراضافی)**	PRS
	Relative Demonstrative Pronoun **(ضمیراشارہ موصولہ)**	PRD
	Relative Personal Pronoun **(اسم موصول)**	PRR
	Reflexive Pronoun **(ضمیر معکوسی)**	PRF
	Reflexive Apna Pronoun **(ضمیر اپنا معکوسی)**	APNA
	Interrogative Pronoun **(ضمیر استفہامیہ)**	INTP
Verb (فعل)	Main Verb Infinite	VBI
	Main Verb Finite	VBF
	Past Participle (فعل ماضی)	VBPA
	Present Participle (فعل حال)	VBPR
	Future Participle (فعل مستقبل)	VBPF
Adverb (فعل مطلق)	Common Adverb	RB
	Negation Adverb (فعل نہی)	NEG
Auxiliary فعل معاون	Aspectual Auxiliary	AUXA
	Progressive Auxiliary	AUXP
	Tense Auxiliary	AUXT
	Modal Auxiliary	AUXM
Nominal Modifiers	Adjective Possessive (صفت ذاتی)	JJ
	Adjective of Quantity (صفت مقداری)	Q
	Adjective of Number/ Cardinal	CD
	Ordinal	OD
Adposition	Preposition (حرفِ ربط)	PRE
	Preposition (حرفِ ربط)	PSP
Particle (حرف)	Conjunction حرف عطف	CP
	Coordinate Conjunction حرف جار	CC
	Subordinate Conjunction حرف علت	SC
Interjection فجائیہ	--	INJ
Symbol علامت	Common Symbol	SYM
	Punctuation Symbol	PU
Foreign Fragment	--	F

References

The Bulgarian National Corpus. (2017-2019). Retrieved 7 18, 2022, from Department of Computational Linguistics: http://dcl.bas.bg/en/

Final Results (Census-2017). (2022, 07 18). Retrieved from Pakistan Bureau of Statistics, Govt of Pakistan:

https://www.pbs.gov.pk/content/final-results-census-2017

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

Adamou, E. (2016). A Corpus-driven approach to language contact:Endangerecs langauges in a comparative perspective. . 12.

Ahmaed, A. S. (2014). Hiding Based on improved Exploiting Modification Direction Method and Huffman Coding. *Journal of intelligent systems*.

Ahmed, T. U. (2015). The CLE urdu POS tagset. In *In LREC 2014, Ninth International Conference on Language Resources and Evaluation* (pp. 2920-2925).

Asghar, M. N. ((2021). A Novel Parts of Speech (POS) Tagset for morphological, syntactic and lexical annotations of Saraiki language. *Journal of Applied and Emerging Scienc*, {pp--77},.

Atta, F. &. ((2019)). Morphophonemic Variations in the Saraiki Language. *International Journal of Linguistics, Literature and Translation*, 42-53.

Baig, A. R. ((2020)). Developing a POS Tagged Corpus of Urdu Tweets. *Computers,*, 90.

Baskaran, S. a. (2008). A common parts-of-speech tagset framework for indian languages. In *In In Proc. of LREC 2008.*

Dandapat, S. a. (2009). Complex linguistic annotation--no easy way out! A case from Bangla and Hindi POS labeling tasks. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)* (pp. 10--18).

Diab, M. (2007). Towards an optimal POS tag set for Arabic processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP* (pp. 157--161).

Dovudov, G. a. (2012). POS annotated 50M corpus of Tajik language. *Language Technology for Normalisation of Less-Resourced Languages*, 93.

Ery\ugit, G. a. (2008). Erratum: Dependency Parsing of Turkish. *Computational Linguistics, 34*(4).

Habash, N. a. (2010). Mada+ tokan manual.

Hussain, S. S. (2016). "The Growth of Saraiki Language .". *Pakistan Journal of social sciences*.

Jakubcek, M. a. (2011). Czech Morphological Tagset Revisited. In *RASLAN* (pp. 29--43).

Jan, M. a. (2016). Optical Character Recognition (OCR) System For Saraiki Language Using Neural Networks. *University of Engineering and Technology Taxila. Technical Journal, 21*(3), 106.

Jumani, N. B. (2011). Effects of Native Language Saraiki on English Language Pronunciation. *International Journal of Business and Social Science, 2*(8).

Kak, A. A. (2009). TOWARDS DEVELOPING A TAGSET FOR KASHMIRI. *Nepalese Linguistics,*, 49-60.

Kausar, R. a. (2015). The History of the Urdu Language Together with Its Origin and Geographic Distribution. *nternational Journal of Innovation and Research in Educational Sciences, 2*, 5--9.

Khan, H. A. (2004). *Re-Thinking Punjab: The Construction of Siraiki Identity.* Research and Publication Centre, National College of Arts, Lahore.

Khanam, M. H. (2013). Part-Of-Speech Tagging for Urdu in Scarce Resource: Mix Maximum Entropy Modelling System . *{Proc. Int. J. Adv. Res. Comput. Commun. Eng},*, 2-9.

Kilgarriff, A. a. (2006). Efficient corpus development for lexicography: building the New Corpus for Ireland. *Language resources and evaluation, 40*(2), 127--152.

Kumar, D. &. (2019). Developing a tagset for machine learning based pos tagging in punjabi. *Indian Journals.com*.

Kumar, S. (2018). Developing POS Tagset for Dogri. *Language in India,*.

Lewinski, N. A. (2017). "An annotated corpus with nanomedicine and pharmacokinetic parameters.". *International journal of nanomadicine*.

MacKinlay, A. ((2005). The effects of part-of-speech tagsets on tagger performance.

Mughal, S. (n.d.). *Multan, Pakistan: Saraiki Isha'ati Idarah.*

Mughal, S. (n.d.). Saraiki qaidah (Saraiki primer). *Multan, Pakistan: Saraiki Isha'ati Idarah.*

Mughal, S. (n.d.). Mughal, S. Saraiki qaidah (Saraiki primer). *Multan, Pakistan: Saraiki Isha'ati Idarah.*

Muhammad Qasim, A. S. (2021). A Novel Parts of Speech (POS) Tagset for morphological, syntactic and lexical annotations of Saraiki language. *Journal of Applied and Emerging Scienc*, 77.

Muller, B. a. (2020). When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. *arXiv preprint arXiv:2010.12858*.

Naseem, A. a. (2017). Tagging Urdu Sentences from English POS Taggers. *International Journal Of Advanced Computer Science And Applications, 8*(10), 231--238.

Pfeiffer, J. a. (2020). Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Pham, D. D. (2009). A hybrid approach to vietnamese word segmentation using part of speech tags. In *2009 International Conference on Knowledge and Systems Engineering* (pp. 154--161). IEEE.

Prajapati, M. a. (2021). *A STUDY OF MATHEMATICAL MORPHOLOGY OF GUJARATI SCRIPT USING WAVELET, OPTIMIZATION AND SOFT COMPUTING TECHNIQUES.* GUJARAT TECHNOLOGICAL UNIVERSITY AHMEDABAD.

Rabbi, I. K. (2008). Developing a tagset for Pashto part of speech tagging. In *In 2008 Second International Conference on Electrical Engineering* (pp. 1-6)).

Rasooli, M. S. (2013). Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 306--314).

Rehman, A. F. (2011). An automatic approach for line dtection and removal without Smash-up characters. *The image sience journal*.

Rizvi, S. J. ((2005, August)). Analysis, design and implementation of Urdu morphological analyzer. In *In 2005 Student Conference on Engineering Sciences and Technology* (pp. (pp. 1-7)). IEEE.

Sajjad, H. &. (2009). Tagging Urdu text with parts of speech: A tagger comparison. In *In Proceedings of the 12th Conference of the European Chapter of the ACL* (pp. 692-700).

Samuel R Bowman, G. A. (2015). A Large annotated corpus for learning natural language inference.

Sankaran, B. B. (2008). Designing a common POS-tagset framework for Indian languages. In *In Proceedings of the 6th workshop on Asian language resources.*

Sheremetyeva, S. (2021). n Interoperable Platform for Multi-Grain Text Annotation. In *IMS* (pp. 56--67).

Shrestha, I. a. (2021). Fine-grained part-of-speech tagging in Nepali text. *Procedia Computer Science, 189*, 300--311.

Skeppstedt, M. e. (2014). " Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study.". *Journal of biomedical informatics*.

Swaroop, L. a. (2019). Parts of speech tagging for Kannada. In *Proceedings of the Student Research Workshop Associated with RANLP 2019* (pp. 28--31).

The Quranic Arabic Corpus. (n.d.). Retrieved 07 28, 2022, from Language Research Group: https://corpus.quran.com/documentation/tagset.jsp

Thomas, W. P. (2002). A national study of school effectiveness for language minority students' long-term academic achievement.

van Esch, D. a. (2022). Writing system and speaker metadata for 2,800+ language varieties. In *Proceedings of LREC.*