Universal Dependencies for Saraiki

Meesum Alam, Francis M. Tyers, Emily Hanink, Sandra Kübler
Indiana University
{meealam,ftyers,emhanink,skuebler}@iu.edu

Abstract

We present the first treebank of the Saraiki/Siraiki [ISO 639-3 skr] language, using the Universal Dependency annotation scheme (de Marneffe et al., 2021). The treebank currently comprises 587 annotated sentences and 7 597 tokens. We explain the most relevant syntactic and morphological features of Saraiki, along with the decision we have made for a range of language specific constructions, namely compounds, verbal structures including light verb and serial verb constructions, along with different types of relative clauses.

Keywords: Saraiki, Universal Dependencies, Indo-Aryan Languages

1. Introduction

Universal Dependencies (UD) is now a widely used annotation scheme for developing syntactic annotations and parsers for a language (de Marneffe et al., 2021; Nivre and Zeman, 2020). It already covers around 220 languages around the world and is growing rapidly. These linguistically annotated corpora are crucial sources for NLP projects of any language. However, Indo-Aryan languages have received little attention in both UD and NLP applications. There currently exist Universal Dependency treebanks for Hindi (Ravishankar, 2017), Urdu (Ehsan and Butt, 2020), and Punjabi (in Gurmukhi script) (Arora, 2022). No lesser studied Indo-Aryan languages are covered in the UD project.

We present a UD treebank for Saraiki, a language of 25 million speakers, which is considered a neglected language in Pakistan. We follow the existing UD guidelines for the annotation where possible. Here, we describe our decisions for phenomena specific to the Saraiki language.

The remaining sections are as follows: Section 2 provides background on the Saraiki language, Section 3 discusses work on treebank construction for related languages, and Section 4 describes the corpus and annotation process. Section 5 discusses part of speech and morphological characteristics of those word classes necessary to understand the discussion of language specific phenomena, and Section 6 discusses the decisions made for language specific phenomena, namely compounds, verbal structures including light verb and serial verb constructions, as well as different types of relative clauses.

2. Saraiki

Saraiki is an Indo-Aryan language widely used in Pakistan and India. The language is one of the ancient languages of the region. Saraiki is spoken by around 25 million people in Southern and Southwestern Punjab and Northern Sindh (see the map in Figure 1). Saraiki is also known as Jataki, Multani, Thali, Riasti and Deraywal in various regions of the Punjab. Saraiki, also spelled *Siraiki*, is counted among the widely-spoken languages in the Pakistani provinces of Punjab and Khyber Pakhtunkhwa (KPK). It is the sister language of Punjabi and Sindhi but has not received much attention in linguistics research.

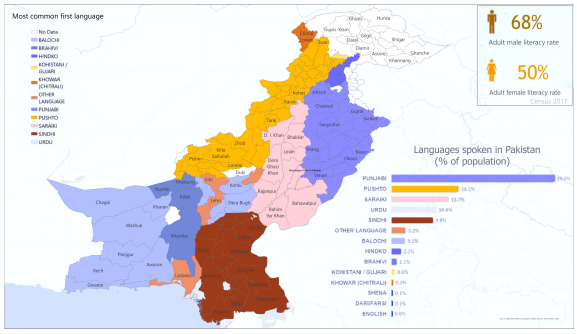


Figure 1: Map showing the percentage and distribution of languages in Pakistan. The region where Saraiki is spoken is shown in pink.

Saraiki is written from right to left in Perso-Arabic script. It is head-final and follows a basic Subject-Object-Verb (SOV) structure within clauses. According to Bashir and Conners (2019), Saraiki word order is relatively free: Topic and focus marking are generally achieved by changes in word order. Saraiki does not have definite or indefinite markers, but it does have numeric ہک (*hik* 'one') to mark indefiniteness. Saraiki is a pro-drop language, it uses clitics/pronomial suffixes in perfective transitive sentences to mark the subjects on verbs. Saraiki has split ergative alignment in addition nominative-absolutive alignment. For more details, see section 6.2.1.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 188–197. May 25, 2024.

Source	Sentences		Tokens	
	Untagged	Tagged	Untagged	Tagged
Common Voice (Ardila et al., 2020)	5 712	288	52 300	17 500
Jhok Newspaper (Dhareja, 2017–2022)	56 000	177	1.15M	5 700
Linguistic examples	—	122	1 851	1 851

Table 1: Textual basis of the Saraiki Treebank.

Saraiki shares morphological and syntactic features with Punjabi but differs on the phonological level, which has allowed it to evolve into a distinct but related language (Bashir and Conners, 2019). As the language has been spoken in different regions of Pakistan for a long time, multiple dialects have emerged over time. Shackle (1976) distinguishes six varieties: Southern Sararik, Northern Saraiki, Sindhi Saraiki, Jhangi Saraiki.

3. Related Work

NLP applications heavily rely on linguistically annotated resources; these resources have multiple functions as they test the linguistic theories, are used to train and evaluate parsing technologies, and provide insights into specific linguistic phenomena of a language (Nivre and Zeman, 2020). However, the Indo-Aryan (IA) languages lack good digital tools because of the scarcity of available corpora. This is also true for Universal Dependency treebanks; we find some IA languages added to the repository. These treebanks cover the major languages: Hindi (Tandon et al., 2016), Urdu (Bhat and Sharma, 2012), Marathi (Ravishankar, 2017), and Punjabi (Arora, 2022). Additionally, there are automated conversions of Urdu (Ehsan and Butt, 2020) and Hindi (Bhat et al., 2018) treebanks from constituent annotations.

For Saraiki, there is little research in the area of NLP. Alam et al. (2023) have developed a morphological analyzer for Saraiki, and Asghar et al. (2021) created a part of speech (POS) tagger. There is also ongoing work on a Saraiki wordnet under Higher Education of Pakistan's Funding at Sarghoda University (Gul et al., 2021), but the system has not been released yet. For the development of NLP related tools, it is equally important to understand the linguistics phenomenon of a language; Bashir and Conners (2019) have published a descriptive grammar for Saraiki, which we used as the basis for our treebank annotations.

4. Corpus and Annotation Process

The Saraiki treebank currently consists of 587 sentences, corresponding to 7 597 tokens in total.

Our treebank is based on sentences from three different sources: from the Saraiki Common Voice corpus (Ardila et al., 2020), from the Jhok newspaper (Dhareja, 2017–2022)[1], and sentences generated during the annotations discussions, to clarify decisions on specific syntactic phenomena in Saraiki. Table 1 shows the distribution of the different text types. Saraiki is under-resourced language and it is difficult to find digital texts in this language, thus limiting our options in creating a diverse textual basis for the treebank.

In a first step, the data was converted into CoNLL-U format and manually segmented. The data have been shared with Saraiki speakers and linguistics scholars in Pakistan. This helped in making decisions on parts of speech (POS) tagging. We manually annotated the corpus for parts of speech. Since there does not exist a standard POS tagging scheme for Saraiki, we left the XPOS category for future work. The POS tagged text was used for the development of a Saraiki morphological analyzer (Alam et al., 2023). Then we started annotating the corpus for universal dependencies. We currently have 587 sentences fully annotated, and will add more annotations in the future. Once we reach 1 000 sentences, the treebank will be published via the UD project.

The annotation is carried out in two steps by the first author, a native speaker of Saraiki, in consultation with the other authors. For part of speech tagging, difficult cases are resolved based on information from the the Saraiki dictionary (Jukes, 2019), along with consulting Saraiki speakers and experts from the Urdu Universal Dependency Treebank to validate decisions. The dependency relationships are annotated using Annotatrix (Tyers et al., 2017), in consultation with all co-authors and UD experts.

5. Saraiki Parts of Speech and Morphology

As of today, there does not exist a language specific part of speech tagging scheme for Saraiki. Even though there are schemes for Punjabi (Gill et al., 2009) and Urdu (Hardie, 2003), we forcused on the Universal POS tagset (Petrov et al., 2012), leaving the XPOS category for future work. All of the UD POS tags occur in our corpus; Table 2

[1]These sentences are used with permission from the newspaper.

POS Tag	Count	Percent
NOUN	1314	17.3
VERB	1231	16.2
PUNCT	759	10.1
ADJ	714	9.4
ADP	630	8.3
PRON	569	7.5
ADV	501	6.6
PROPN	417	5.5
AUX	387	5.1
CCONJ	386	5.1
DET	258	3.4
SCONJ	190	2.5
PART	188	2.5
INTJ	22	0.3

Table 2: Distribution of Universal Dependency parts of speech tags in the Saraiki Treebank.

gives a detailed picture of the distribution of the tags in the Saraiki Treebank.

Verbs Similar to other Indo-Aryan languages, Saraiki verbs undergo derivational and inflectional processes. Saraiki verbs inflect for number, gender, tense, aspect, and mood. Adverbs, compounds, and reflexives can be derived from verbs via derivational verbal morphology. Additionally, Saraiki uses verb stem alteration. To describe those, we use work by Bashir and Conners (2019) on the eight different verb stem alterations as the basis for our annotations.

In Saraiki, certain verbs play a dual role. When occurring within a light verb construction, they take the role of auxiliaries, providing information on the verb's aspect. Consequently, we distinguish between VERB and AUX, according to the structure. For infinitives, we follow decisions in the Punjabi treebank (Arora, 2022): We mark them as VERB in all instances, regardless of their semantic interpretation.

Nouns We found three types of nouns in our treebank: case-marked nouns, non case-marked nouns, and uninflected nouns. Most nouns are case-marked in addition to being inflected for gender and number. Saraiki uses four cases: direct, oblique, vocative, and ablative. Examples of nouns that can be case-marked are ماں (*maa'n* 'mother') and چھاں (*chaa'n* 'shade'). The second type of nouns are non case-marked nouns. These nouns are borrowed from neighboring languages, and are adapted to suit Saraiki morphology. Examples of this type are بال (*baal* 'male child') and ذات (*zaat* 'caste'). The last category of nouns does not take any kind of inflections; these nouns are mostly borrowed from Urdu or Persian, such as ایمان (*emaan* 'faith') and رب (*Rub* 'God').

Adjectives In Saraiki, adjectives take the case and inflection of the nouns that they modify. If a noun is not case-marked, modifying adjectives agree with it in gender and number only.

Pronouns and demonstratives Saraiki does not distinguish between third person proximal and distal pronouns and demonstratives. Instead, the distal forms for *he, she, that, those* اوں (*oo'n*) are used for both expressions alongside their proximal forms اے (*ay* 'he, she, it, this, these').

Following Bashir and Conners (2019), who identify a morphological difference between relative pronouns that stand alone or immediately precede a noun, we annotated relative pronouns as PRON where they function as independent pronouns and DET where they function as determining adjectives. The adjectival forms, unlike the stand-alone pronominal forms, inflect robustly for number, gender, and case of the noun they precede and modify.

6. Annotation Decisions

In this section, we focus on language specific constructions, focusing on the treatment of (split) ergative sentences, serial and light verbs, as well as compounds and relative clauses. Remember that Saraiki is head-final and written right to left.

6.1. Compounds

Saraiki has a comprehensive system of creating multiword expressions and compounds in open and closed POS categories. In section 6.2, we will focus on the V-V compound in serial verb and light verb constructions. Here, we discuss an additional type of V-V compounding, reduplication, plus compounds involving nouns, reflexive pronouns, and adverbs.

Reduplication This is common for emphasis, for noun compounding and pluralization. In these cases, we annotate the verbs using `compound:redup`, with the first verb as the head. Interestingly, reduplication can occur with all open class categories. Verb reduplication is different from light or serial verb constructions. These verbs do not provide tense, aspect, and modality information, and they are not part of complex serial verb predicates. In example (1), گھت (*ghut* 'put') is reduplicated, either for emphasis or to indicate a quick action. As described above, reduplication can be used with almost all open categories of the

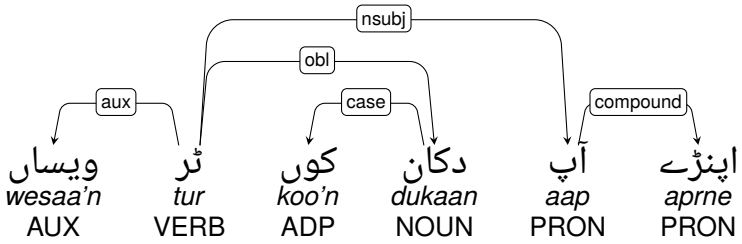
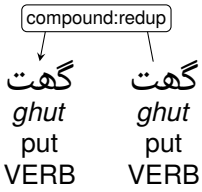
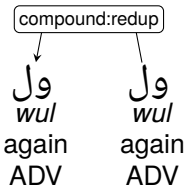


Figure 2: The annotation of the example in (4).

grammar in Saraiki. In example (2), reduplication is used to emphasize the adverb ول (*wul* 'again').

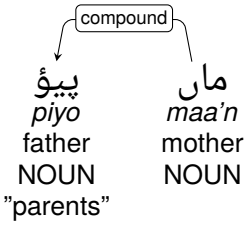
(1)


"put quickly"

(2)


"**Again**" (emphasized)

Noun-Noun Compounds In Saraiki, there are a wide range of concepts that are expressed as noun-noun compounds. We use the `compound` relation in these cases. Example 3 shows a combination of ماں (*maa'n* 'mother') and پیؤ (*piyo* 'father') meaning "parents".

(3)


"parents"

Reflexive Pronouns These are constructed by combining the two words اپنڑے (*apnre* 'own') and آپ (*aap* 'self') in a multi-word expression (see example 4 and Figure 2). We follow the UD guidelines and use the `compound` relation to combine those two words.

(4) ویساں ٹر کوں دکان آپ اپنڑے
wesaa'n tur koo'n dukaan aap aprne
AUX.FUT go to shop-ACC PRON PRON
"I will go to the shop by myself"

6.2. Verbs

In Saraiki, the verb system is more complex than in the neighbouring languages Punjabi, Urdu, and Hindko (Bashir and Conners, 2019). Syntactically, Saraiki exhibits split ergativity in addition to pronominal suffixation onto verbs in some contexts. It uses two types of light verb constructions: one consisting of two verbs where one verb acts as an auxiliary, contributing only tense, aspect and modality information, and another consisting of a noun or adjective in addition to the light verb. Additionally, Saraiki employs serial verb constructions. We will discuss all these phenomena and annotation decisions in more detail below. In the Common Voice corpus by Ardila et al. (2020), out of all the verbs construction we found approximately 21% light verb constructions; interestingly, half of these light verb constructions use the verb تھیونڑ (*thivaṇ* 'to become'). These numbers are based on the current treebank, but we expect the percentages to remain stable as we add more sentences.

6.2.1. Syntactic Split Ergativity

Saraiki belongs to the group of languages that have both nominative–accusative and ergative-absolutive alignment (see Dixon (1994) for an overview). According to Bashir and Conners (2019), Saraiki shows an ergative-absolutive pattern only in perfective contexts, a pattern common across Indo-Aryan languages. It is important to know that unlike Urdu, Punjabi, and Hindi, Saraiki lacks a dedicated ergative morpheme. Consequently, the effects of this split are observable only in verbal agreement patterns. The generalization is that verbs agree with agents of transitive verbs and subjects of intransitive verbs in the same way in the imperfective aspect, but do not agree with agents of transitive verbs in the perfective aspect. Thus, while patients are oblique in imperfective contexts, it is agents that are oblique in perfective contexts. Table 3 lays out the case alignment pattern across imperfective and perfective contexts.

The aspectual contrast giving rise to this split is exemplified below. The imperfective sentence in example (5) shows a typical nominative-accusative agreement pattern, in which the verb

	Intransitive Subject	Transitive Agent	Patient
Perfective	Nom	Obl	Nom
Imperfective	Nom	Nom	Obl

Table 3: Split-ergative alignment in Saraiki. Subjects of intransitive verbs are always nominative, while agents and objects of transitive verbs depend on the aspect of the verb. In perfective aspect, the oblique encodes the agent, while in imperfective aspect the oblique encodes the patient.

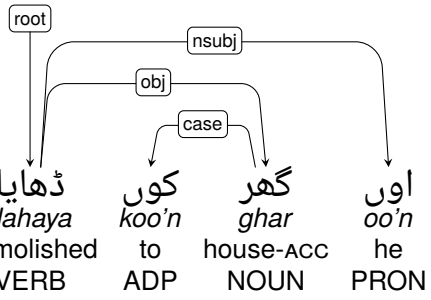
agrees with the nominative argument قاسم (*Qasim* 'Qasim'). The same case and agreement pattern is found with intransitive verbs, which agree with their nominative subject.

In the perfective sentence in example (6) in contrast, the agent of the transitive verb پڑھی (*parhi* 'read'), قاسم (*Qasim* 'Qasim') carries the oblique case, while the direct object کتاب (*kitaab* 'book') carries nominative case. Notably, the verb in this context agrees with its direct object rather than its subject. The generalization is thus that, in perfective contexts only, agents of transitive verbs i) are oblique arguments ii) may not control subject agreement.

(5) اے پڑھدا کتاب
ay parhda kitaab
AUX read-PRES-SG-M book-OBL-SG-F
قاسم
qasim
Qasim-NOM-SG-M
"Qasim reads the/a book"

(6) ہا پڑھی کتاب
ha parhi kitaab
AUX read-PP-SG-F book-SG-F
قاسم
qasim
Qasim-OBL-SG-M
"Qasim read the/a book"

In our treebank, both patterns are present. For the ergative sentences, we decided to follow the Urdu (Ehsan and Butt, 2020) and the Hindi treebank (Bhat and Sharma, 2012), we annotate agents as `nsubj` and patients and other non-agents as `obj`. We are aware that this does not agree with the decisions made in the Basque treebank (Aduriz et al., 2003), which uses `subj` for such arguments in the ergative.

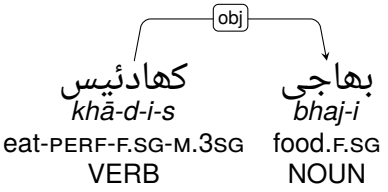


(7) "He demolished the house"

Example (7) shows an example of an ergative sentence, where we annotate the agent اوں (*oo'n* 'he'), which is in the oblique case, is the subject, and گھر (*ghar* 'house') is the direct object in ergative case.

We note that another type of agent marking is also available. This strategy uses pronominal suffixes (clitics) on the verb to mark the grammatical features of the agent. In this type of structure, the transitive verb in the perfective form shows object agreement, with the pronominal agent cliticized onto the end of the verb. In example (8), the verb پیتم (*pita-m* 'I drank') agrees with the noun پاٹی (*paanri* 'water'), and the agent 1.M.SG is added to the end of the verb پیتم. In example (9), the verb کھادئیس (*khād-i-s* 'he ate') agrees with بھاجی (*bhaj-i* 'food-F.SG'), and the agent is marked on verb.

(8) پیتم پاٹی
pita-m *paanri*
drink-PST-1.M.SG water.M.SG
"I drank water"



(9) "He ate food"

These constructions are possible only in the perfective forms. Note that while Bashir and Conners (2019) call these pronominal suffixes, Syed and Raza (2019) call them clitics. On either treatment, this type of construction is sensitive to the morphological features of the agent, which are marked on the verb. Following the UD guidelines, we annotate the argument as direct object `obj`.

This morphologically embedded ergativity (differential case marking) is also found in Hebrew (Glinert, 2004) and Hungarian (Bárány, 2012).

6.2.2. Serial Verb Construction

Serial verbs mostly conceptualize one event and are realized as one linear, complex predicate with-

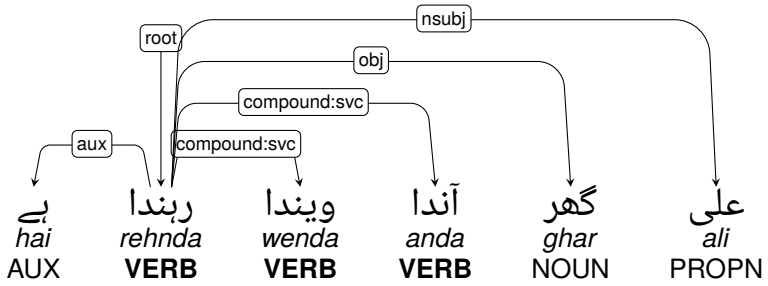


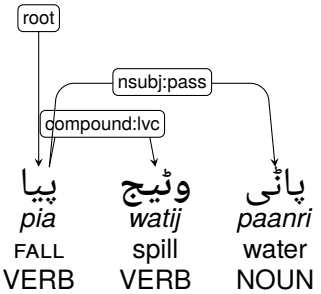
Figure 3: The annotation for the serial verb construction (POS of serial verbs in bold) of example (10).

out explicit coordination or subordination markers. This feature is common in many IA languages. Example (10) shows a sentence from our treebank, and Figure 3 shows our annotation. Since we do not yet know enough about the constraints on this construction, we decided to annotate the involved verbs serially. As Saraiki is a head final language (written from right to left), we mark the last verb as the head of the clause and create `compound:lvc` relations with other verbs. We anticipate changes to these annotations in the future once we have a better understanding of this construction.

(10) ہے رہندا ویندا آندا گھر علی
hai rehnda wenda anda ghar ali
AUX keeps go come home Ali-NOM
'Ali keeps coming and going home"

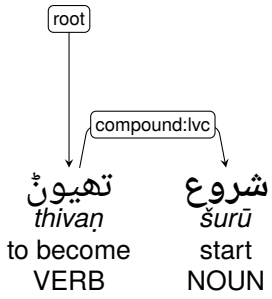
6.2.3. Light Verb Constructions

In Saraiki, we find sequences of verbs where the main verb is followed by another 'light' verb, in addition to constructions in which a light verb is followed by a noun or adjective. In both cases, the light verb has little semantic content. In V-V LVCs, the second verb mostly contributes information about aspect or modality. All such constructions have been given the dependency of `compound:lvc`. We show an example in (11): وٹیج (*watij* 'spill') is the main verb in the structure, and پیا (*pia* 'fall') provides aspectual information about the main verb, indicating that the action is completed.

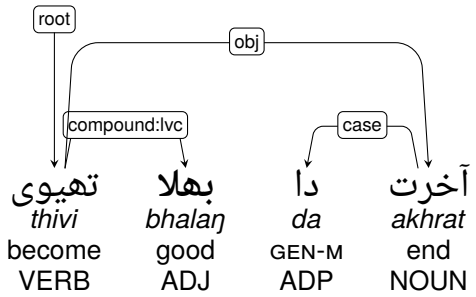


(11) "The water was spilled"

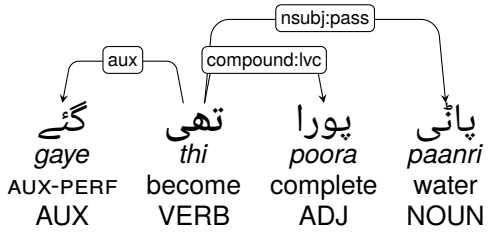
In the treebank, we also found the verb تھیوݨ (*thivaṇ* 'become'), a change of state verb (Bashir and Conners, 2019) in Saraiki, which, unlike ہوݨ (*hovaṇ* 'be'), appears in SVCs, LVCs, and as an auxiliary. تھیوݨ (*thivaṇ* 'become') can also be followed by another light verb construction. Where it occurs in a light verb construction, we mark it as a root with a `compound:lvc` dependency to the noun or verb (see examples (12) and (13)); when تھیوݨ (*thivaṇ* 'become') is not part of the light verb construction, we mark it as an auxiliary AUX (see example (14)).



(12) "to start"



(13) "may (you) have a better end"



(14) "The (land) filled (with) water" (lit.: water full become go-PERF)

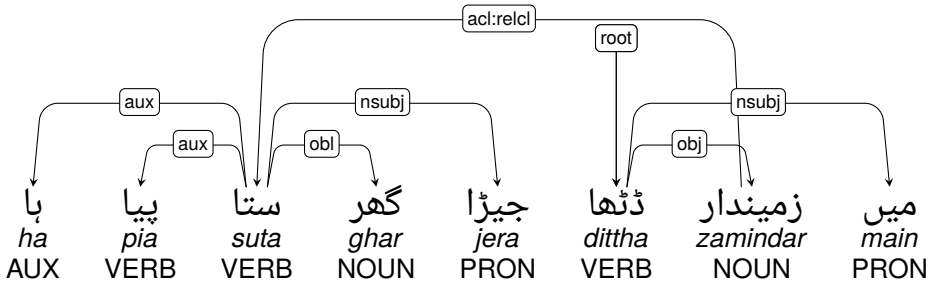


Figure 4: The annotation of the example of an externally headed relative clause in (15).

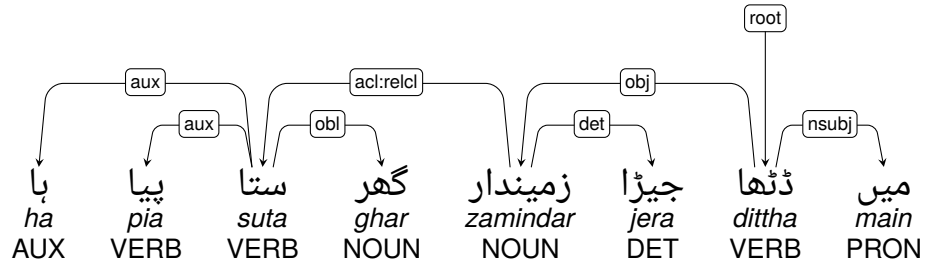


Figure 5: The annotation of the example of an internally headed relative clause in (16).

6.3. Relative Clauses

In the Saraiki treebank, we found both finite and non-finite relative clauses. According to Bashir and Conners (2019), both types of clauses are used freely in Saraiki. While Saraiki uses externally headed relative clauses, it also uses internally headed and correlative forms. Saraiki uses جیڑا (*jera* 'that, which') as a relativizer, which agrees with its head noun in number, gender, and case. These types of constructions are also available in Urdu (Ehsan and Butt, 2020; Bhat and Sharma, 2012) and Punjabi (Arora, 2022).

The examples discussed here are part of the sentences created for analyzing specific constructions in Saraiki. We use those examples so that we can focus on the relevant construction without interference from other syntactic phenomena.

Example (15) shows an externally headed relative clause, the annotation is shown in Figure 4. In such cases, جیڑا (*jera* 'which') functions as relative pronoun; here it modifies زمیندار (*zamindar* 'farmer'). We annotate the relative pronoun as `nsubj` of the verb of the relative clause, ستا (*sutta* 'sleep-pst'), which in turn is dependent on the noun in the matrix clause via the `acl:relcl` relation.

(15) ہا پیا ستا گھر جیڑا
ha pia suta ghar jera
AUX PROG sleep-PST house REL.M.SG
ڈٹھا زمیندار میں
dittha zamindar main
see-PST farmer DIR.1.M.SG
"I saw the farmer who was sleeping in the house"

Example (16) shows a version of the sentence with an internally headed relative clause, the annotation is shown in Figure 5. Here, the head noun زمیندار (*zamindar* 'farmer') occurs inside the relative clause, i.e., between the relative pronoun and the object of the relative clause (گھر *ghar* 'house'). Since this means that the relative clause has a relativizer and the noun it refers to, we have decided that the head noun زمیندار (*zamindar* 'farmer-m-sg') serves as the direct object (`obj`) in the matrix clause, and the relativizer serves as its determiner in a `det` relation. Consequently, the verb of the relative clause is dependent on the head noun via a `acl:relcl` relation. This analysis means that we do not consider the head noun to be part of the relative clause, since it provides the only "attachment site" for the relative clause.

(16) ہا پیا ستا گھر زمیندار
ha pia suta ghar zamindar
AUX PROG sleep-PST house farmer
جیڑا ڈٹھا میں
jera dittha main
REL-M-SG see-PST DIR.1.M.SG
"I saw the farmer who was sleeping in the house"

Example (17) shows the same internally headed version, but in a different word order, with a fronted relative clause. The annotation is shown in Figure 6. Based on our current understanding, we

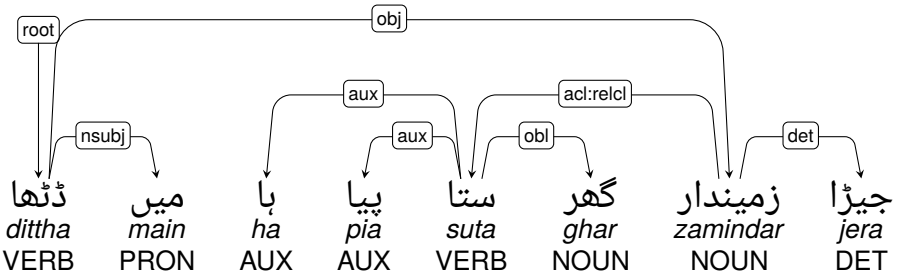


Figure 6: The annotation of the example of an internally headed, fronted relative clause in (17).

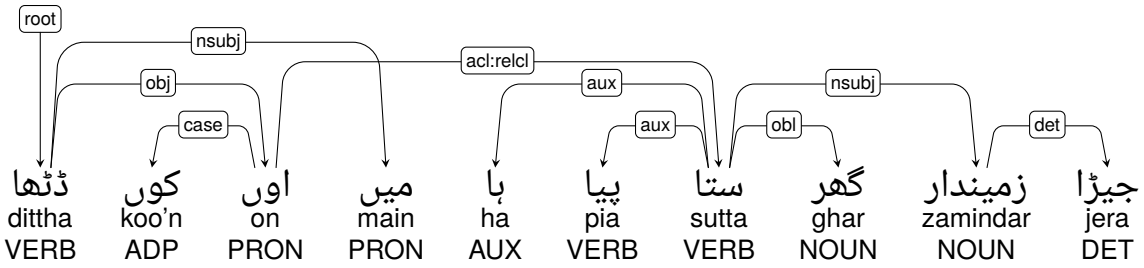


Figure 7: The annotation of the example of a correlative relative clause in (18).

assume that the only difference between all three variants is in information structure.

(17) ڈٹھا میں ہا پیا ستا
dittha main ha pia suta
see-PST DIR.1.M.SG AUX PROG sleep-PST
گھر زمیندار جیڑا
ghar zamindar jera
home farmer REL-M-SG
"I saw the farmer who was sleeping in the house"

Example (18) shows the same sentence, but uses a correlative. The annotation is shown in Figure 7. Correlative relative clauses are a variant of internally headed relative clauses where the relative clause is dependent on, and in an anaphoric relation to, a pronoun in the matrix clause. In example (18), the distal pronoun اوں (*oun* 'that') serves as the correlative. Consequently, we annotate it as the direct object of the matrix clause. The fronted relative clause is dependent on this pronoun. Parallel to the internally headed examples in (16) and (17), we analyze the relativizer as a determiner dependent on the subject of the relative clause.

(18) ڈٹھا اوں کوں میں ہا
dittha koo'n on main ha
see-PST to ACC.3.M.SG DIR.1.M.SG AUX
پیا ستا گھر زمیندار
pia sutta ghar zamindar
PROG sleep-PST house farmer
جیڑا
jera
REL.SG.M
"I saw the farmer who was sleeping in the house"

7. Conclusion and Future Work

We have presented a treebank for Saraiki, annotated using Universal Dependencies. We discussed the textual basis of the treebank and a range of language specific syntactic phenomena. The treebank is work in progress, it currently comprises 587 sentences. We will we will keep extending it and release it once we reach 1 000 sentences.

For future work, we will need to have a closer look at the relative clauses. Additionally, we plan to automatically annotate the morphological features using the Apertium morphological analyzer for Saraiki (Alam et al., 2023). We hope that this treebank will spur deeper investigations of Saraiki as well as the creation of NLP tools for the language. We also plan to train a syntactic parser, and investigate zero-shot techniques to extend our work to other regional languages such as Punjabi (Shahmukhi), Hindko, and Khetrani.

8. Acknowledgements

We would like to thanks Pervaiz Qadir for developing the Saraiki corpus in Mozilla Common Voice and Zahoor Dhareja for giving permission for us to use data from Jhok newspaper for the treebank. We would also like to thanks Daniel Swanson and Daniel Zeman for their help with annotation decisions.

9. Bibliographical References

I. Aduriz, M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. Díaz de Ilarraza, A. Garmendia, and M. Oronoz. 2003. Construction of a Basque dependency treebank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, pages 201–204, Växjö, Sweden.

Meesum Alam, Alexandra O'Neil, Daniel Swanson, and Francis Tyers. 2023. A finite-state morphological analyzer for Saraiki. In *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL)*, pages 9–13.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 4218–4222, Marseille, France.

Aryaman Arora. 2022. Universal Dependencies for Punjabi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 5705–5711.

Muhammad Nabeel Asghar, Farrukh Javed Saleemi, Sajid Iqbal, Muhammad Umar Chaudhry, Muhammad Yasir, Sibghat Ullah Bazai, and Muhammad Qasim Khan. 2021. A novel parts of speech (POS) tagset for morphological, syntactic and lexical annotations of Saraiki language. *Journal of Applied and Emerging Sciences*, 11(1):pp–77.

András Bárány. 2012. Hungarian conjugations and differential object marking. In *Proceedings of the First Central European Conference in Linguistics for Postgraduate Students*, pages 3–25.

Elena Bashir and Thomas J Conners. 2019. *A Descriptive Grammar of Hindko, Panjabi, and Saraiki*, volume 4. Walter de Gruyter.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, pages 987–998, New Orleans, LA.

Riyaz Ahmad Bhat and Dipti Misra Sharma. 2012. Dependency treebank of Urdu and its evaluation. In *Proceedings of the Sixth Linguistic Annotation Workshop (LAW)*, pages 157–165.

Marie-Catherine de Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Zahoor Dhareja. 2017–2022. Jhok Multan. (Daily newspaper in Pakistan).

Robert MW Dixon. 1994. *Ergativity*. Cambridge University Press.

Toqeer Ehsan and Miriam Butt. 2020. Dependency parsing for Urdu: Resources, conversions and learning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 5202–5207, Marseille, France.

Mandeep Singh Gill, Gurpreet Singh Lehal, and Shiv Sharma Joshi. 2009. Part of speech tagging for grammar checking of Punjabi. *Linguistic Journal*, 4(1):6–21.

Lewis Glinert. 2004. *The Grammar of Modern Hebrew*. Cambridge University Press.

Sarah Gul, Musarrat Azher, and Sana Nawaz. 2021. Development of saraiki wordnet by mapping of word senses: A corpus-based approach. *Linguistics and Literature Review*, 7(2):46–66.

Andrew Hardie. 2003. Developing a tagset for automated part-of-speech tagging in Urdu. In *Corpus Linguistics*.

Andrew John Jukes. 2019. *Dictionary of the Jatki or Western Panjábi language*. Routledge.

de Marneffe M.-C. Ginter F. Hajič J. Manning C. D. Pyysalo S. Schuster S. Tyers F. M. Nivre, J. and D. Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*, Istanbul, Turkey.

Vinit Ravishankar. 2017. A universal dependencies treebank for Marathi. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 190–200.

Christopher Shackle. 1976. *The Siraiki Language of Central Pakistan: A Reference Grammar*. School of Oriental and African Studies, Univ. of London.

Nasir Abbas Syed and Ghulam Raza. 2019. Pronominal suffixes and clitics in Saraiki. *Pakistan Journal of Languages and Translation Studies*, (1):148–173.

Juhi Tandon, Himani Chaudhry, Riyaz Ahmad Bhat, and Dipti Misra Sharma. 2016. Conversion from Paninian karakas to Universal Dependencies for Hindi dependency treebank. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X)*, pages 141–150.

Francis M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2017. UD annotatrix: An annotation tool for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 10–17, Prague, Czech Republic.