0

Universal Dependencies for Saraiki

MWE-UD 2024

Meesum Alam, Francis M. Tyers, Emily Hanink, Sandra Kübler

Indiana University, Department of Linguistics

May 19, 2024

Introduction

- We present the first treebank of the Saraiki/Siraiki [ISO 639-3 skr] language, using the Universal Dependency annotation scheme
- Universal Dependencies (UD) is now a widely used annotation scheme for developing syntactic annotations and parsers for a language.
- 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.
- Indo-Aryan languages have received little attention in both UD and NLP applications. There currently exist Universal Dependency treebanks for Hindi, and Punjabi (in Gurmukhi script). No lesser studied Indo-Aryan languages are covered in the UD project.

Saraiki

- Saraiki is an Indo-Aryan (IA) language widely used in Pakistan and India. Saraiki is spoken by around 25 million people in Southern and Southwestern Punjab and Northern Sindh.
- Saraiki is written from right to left in Perso-Arabic script.
- Saraiki is head-final and follows a basic Subject-Object-Verb (SOV) structure within clauses
- Saraiki shares some morphological and syntactic features with neighbouring languages like Punjabi, Sindhi and khetrani.

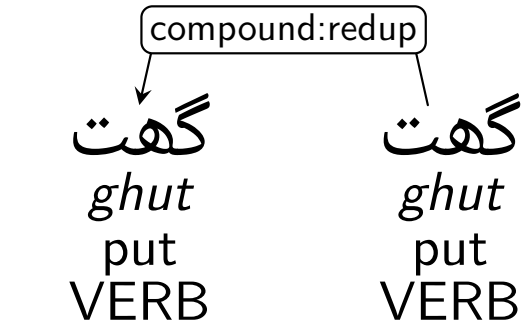
Corpus and Annotation Process

- 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, from the Jhok newspaper, [1]
- We first annotated the corpus for POS. 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.
- The dependency relationships are annotated using Annotatrix, in consultation with all co-authors and UD experts

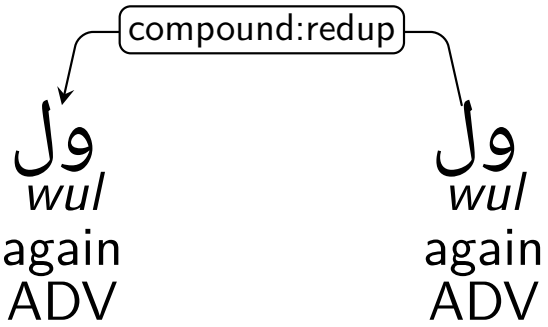
[1] These sentences are used with permission from the newspaper.

Annotation Decisions: Compounds

- Saraiki has a comprehensive system of creating multiword expressions and compounds in open and closed POS categories.
- We discuss an additional type of V-V compounding, reduplication, plus compounds involving nouns, reflexive pronouns, and adverbs



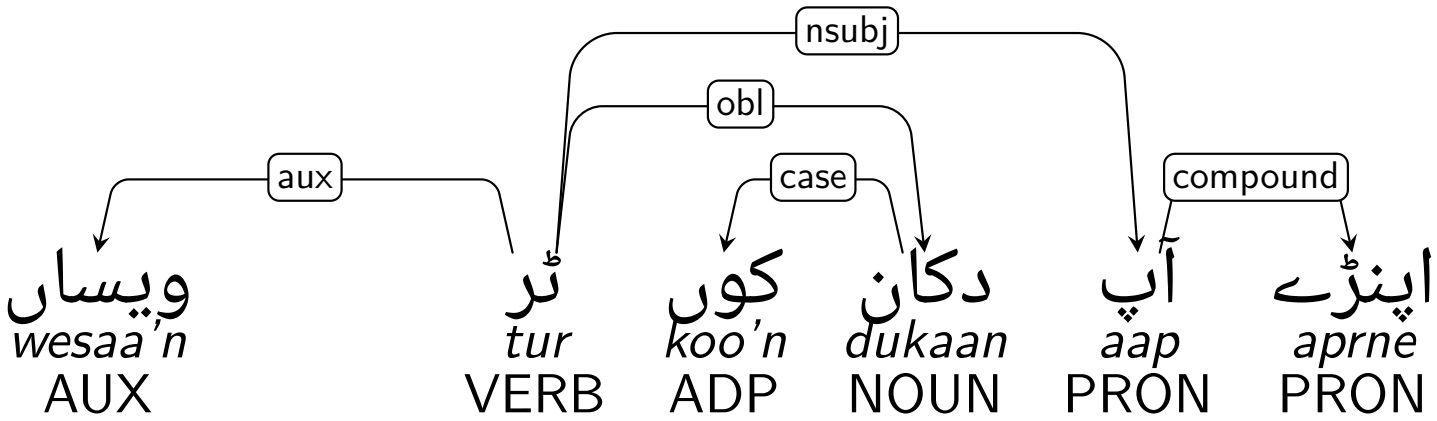
Put quickly.



Again with emphasize.

Annotation Decisions: Reflexive Pronouns

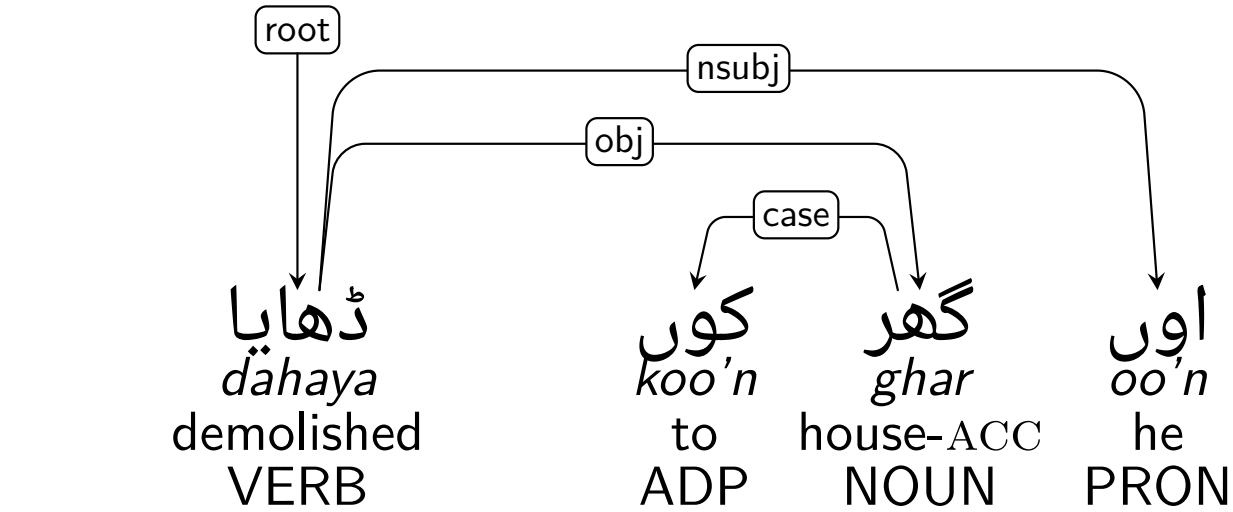
- These are constructed by combining the two words اپنڑے (*apnre* 'own') and آپ (*aap* 'self') in a multi-word expression.
- We follow the UD guidelines and use the `compound` relation to combine those two words.



I will go to shop by myself.

Annotation Decisions: Verbs

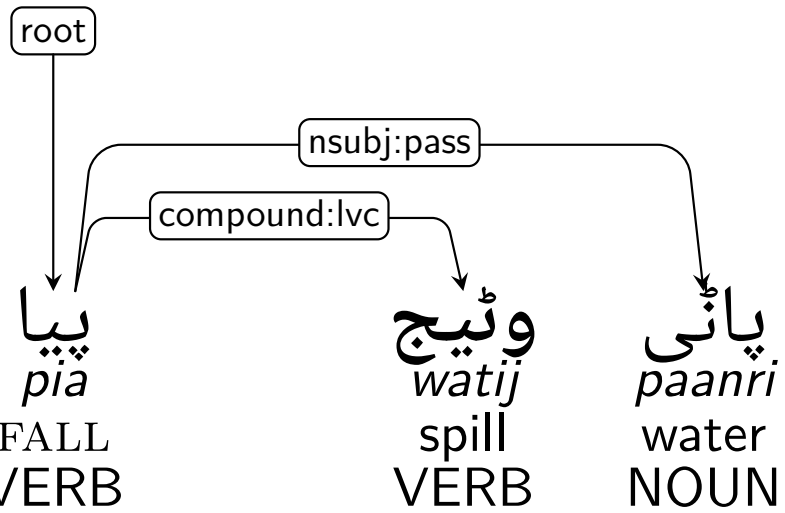
- In Saraiki, the verb system is more complex than in the neighbouring languages Punjabi, Urdu, and Hindko.
  - Split Ergative Alignment
  - Pronominal Suffixation
  - Light Verb Constructions
  - Serial Verb Constructions.



He demolished the house.

Verbs:Light Verb Constructions

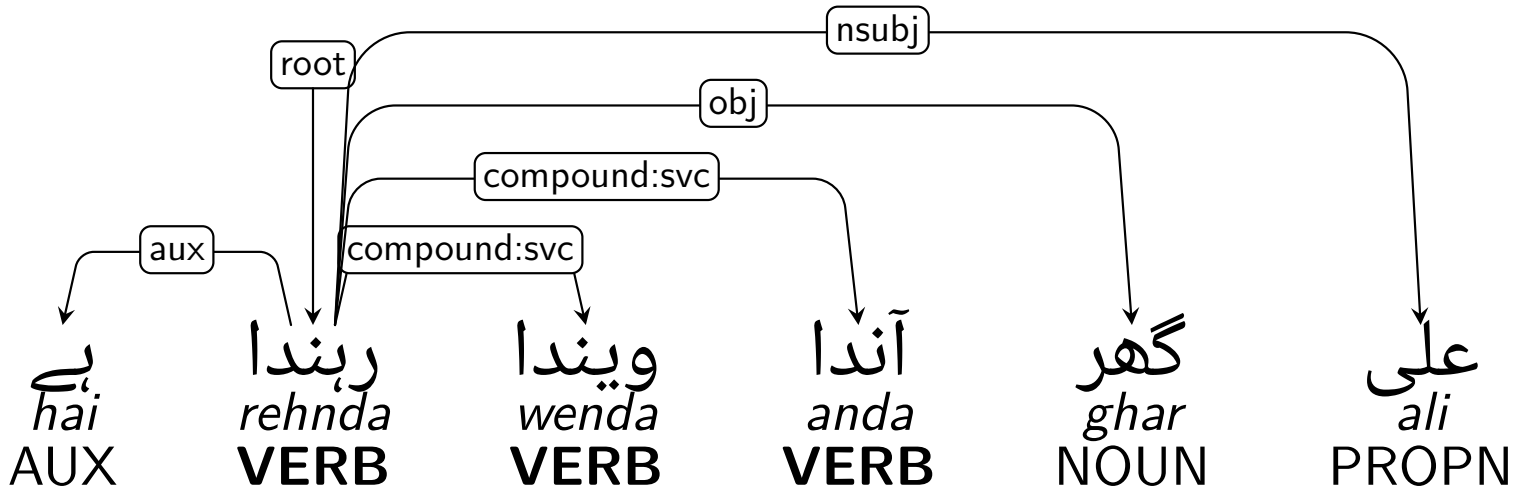
- We found sequences of verbs where the main verb is followed by another 'light' verb, in addition to constructions in which a light verb is followed by a noun or adjective.
- All such constructions have been given the dependency of `compound:lvc`



The water was spilled

Verbs:Serial Verb Constructions

- 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.



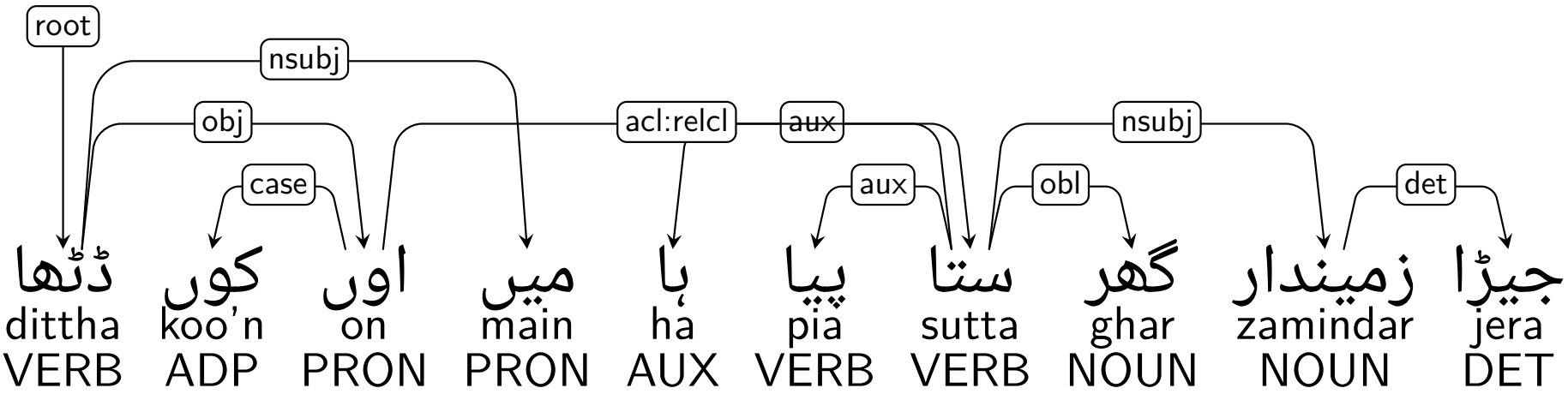
Ali used to come and go home

Annotation Decisions: Relative Clauses

- Saraiki uses جیڑا (*jera* 'that, which') as a relativizer, which agrees with its head noun in number, gender, and case.
- In the Saraiki treebank, we found both finite and non-finite relative clauses. According to Elena Bashir and Corners, both types of clauses are used freely in Saraiki.
  - Externally headed relative clauses
  - Internally headed relative clauses
  - Correlative Relative Clauses

Relative Clauses: Correlatives

- Correlative relative clauses are famous in IA languages and 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, the distal pronoun اوں (*oun* 'that') serves as the correlative.
- We annotated it as the direct object of the matrix clause.



I saw the man who was sleeping in the house.

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.
- 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.

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.