Universal Dependency Treebank for a low-resource Dardic Language: Torwali

Naeem Uddin, Daniel Zeman
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics, Prague
naeemuddinhadi@gmail.com, zeman@ufal.mff.cuni.cz

Abstract

This paper presents and discuss the linguistic phenomena encountered in the development of the ongoing first ever universal dependency treebank for the Torwali Language. Torwali belongs to the Kohistani sub-group of Dardic Indo-Aryan languages, and is considered an endangered (Moseley, 2010) and indigenous language, which makes it extremely low-resourced in terms of linguistic and computational resources. With the aim of including Torwali in Universal Dependencies (UD) (de Marneffe et al. 2021), we are annotating a diverse set of example sentences for POS tags, features and dependency relations.

Keywords: Torwali, Universal Dependencies, treebank, POS tags

1 Introduction

Northern Pakistan is characterized as one of the most linguistically diverse regions with around 30 indigenous language communities (Liljegren and Akhunzada, 2017), among which many are endangered languages. Torwali is one of such communities, located in Swat Kohistan, which belongs to the Kohistani sub-group of the Dardic Indo-Aryan languages (Ullah, 2004). It has two dialects (the Bahrain and Chail dialects), with a total of over 100,000 speakers approximately.

Torwali [ISO 639-3: trw], is a marginalized and low-resource language written in right-to-left Perso-Arabic script. There is a glossonymic variation between Torwalik (Biddulph, 1880), Torwali (Grierson, 1929), Kohistani (Raverty, 1862) and Torwali-Kohistani (Rensch 1992) across time, literature and communities. There are some resources of Torwali language description (Grierson, 1929), a study of linguistic features (Lunsford, 2001) and a structured lexical database (Ullah, 2004). But, when it comes to resources for computational processing of Torwali, there is a lack of robust resources of morphology and grammar.

UD treebank data is useful for downstream tasks in NLP, including semantic parsing (Reddy et al., 2017) and natural language understanding (Schuster and Manning, 2016), and syntactic corpora are useful for linguists studying language typology and change (Levshina, 2019). As there is no coverage in UD to date for Dardic languages, this work upon completion and inclusion in UD, will help in creation of treebanks for other Dardic languages as well, and will help mitigate the bias towards the “big” Indo-European languages (Nivre et al. 2020) in the UD.

This paper discusses the approach to build a Torwali dependency treebank as well as syntactic constructions and grammatical structure of an extremely low-resourced and less-studied language. The treebank is currently under development with a target to cover 500 sentences taken from Inam Ullah’s Torwali-English-Urdu Dictionary**.**

2 Data collection and annotation

The data for this study was extracted from the lexical database of Torwali in Toolbox format created by Inam Ullah (Ullah, 2004), which encodes linguistic information using tagged field markers. Only example sentences were extracted, as they provide naturalistic usage data for lexical items in context. A custom script was developed to parse the Toolbox file and isolate fields containing example sentences, while discarding other lexical metadata. Minor formatting inconsistencies, such as irregular punctuation or spacing, were corrected to ensure uniformity across sentences.

After preprocessing, the extracted sentences were converted into the CoNLL-U format, which consists of a structured 10-column format designed to facilitate manual annotation. This format enables the inclusion of tokenization, morphological features, and syntactic

Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025), pages 140–147
August 28-29, 2025 ©2025 Association for Computational Linguistics

dependencies. The last column of the CoNLL-U format is used, among other things for transliteration, allowing for a consistent representation of the Torwali text in a Latin-based alphabet for easier analysis and cross-linguistic comparison.

Since we have extracted almost all the sentences from Toolbox file, the current workflow involves a rather large amount of manual work. The treebank has been annotated by a native speaker of Torwali, the first author of this paper.

3 Morphology

Torwali has a complex morphology because it is basically a fusional language which uses several strategies like stem modification, reduplication and existence of words in inflected form, derived form, compound form and root form.

In Torwali, nouns are inflected for number and case and the stem can be joined by an optional plural suffix and an optional oblique case marker. Torwali uses several strategies to mark plurality, but the primary morphological method is tone along with verb agreement like for most of the singular nouns have a tone with rising pitch from low-to-high and their plural counterparts have a tone with low pitch.

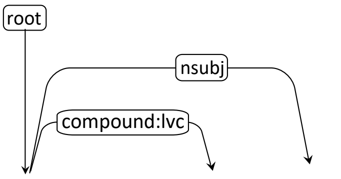
Torwali verbs inflect for tense, aspect, mood and gender and most of the verb forms make gender and number distinction only, no distinction for person. Torwali has three tenses: present, past and future. We did not record any distinction between simple present and present continuous tense in Torwali. It also encodes elevation in motion verbs, distinguishing 'going up' and 'going down' based on real altitude. This reflects the speakers' mountainous environment, with multiple verbs (e.g., **واد**/wad*/came down-went down/*, and **اوگھاد**/ughād*/came up-wen*t *up*) conveying nuanced direction, though all translate as 'go' or 'come' in English.

4 Syntactic features

4.1 Word order and head position

Torwali language exhibits a SOV (Subject-Object-Verb) word order, where the verb consistently appears in clause-final position, so it is a head-final language; in accord with that, it uses pospositions rather prepositions. This structure is further complicated by gender agreement, in which the verb cross-references the gender and number of its preceding subject or object, creating a tight sytactic bond between the verb and its arguments. Consider the following example sentence (1). [1]

(1)



کوییٔ **باٹ** **جوَوٗی**

koi *bat* *ǰəwəi*

do.PRES.SG.F talk woman

VERB NOUN NOUN

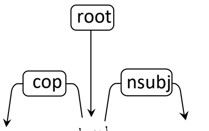
“ The woman talks”

In (1), the gender is marked by attaching **یئ** /i/ to verb /**کو**/ko/, if the preceding subject is feminine and /**دو**/du/ is attached to verb if the preceding subject is masculine.

4.2 Copula

In Torwali, the copula is morphologically rich and obligatorily present in all major predicate structures, including predicate nominals, locative clauses, and possessive constructions. The copula inflects for gender and number, typically distinguishing masculine singular, feminine singular, and plural but it does not mark future tense, which is expressed through other verbal strategies. We annotated the copula AUX and attached it to the nominal or adjectival predicate via the cop relation as shown in (2), while the predicate itself serves as the head of the clause, and the subject is attached with it.

(2)



تھو ڈاکٹر سے

thu *ḍakṭər* *se*

is.M doctor he.3SG

AUX NOUN PRON

“He is a doctor”

For present tense the copula is تھو.M.SG/*thu,* چھی.F.SG/*čhi* and تھی.PL/*thi,* and for past tense

[1] The words in all tree diagrams are ordered right-to-left in accord with the Torwali original. This has to be borne in mind when reading the English glosses.

اشو.M.SG/*ašu* , أشی.F.SG/*æši* and اشی.PL/*aši*.

4.3 Number

As mentioned in section 3, nouns in Torwali are inflected for number and use a mix strategy which includes tone, joining an optional plural suffix and stem modification. The nouns نار/*nar*/dance.SG and نأر/*nær*/dance.PL demonstrate a singular-plural distinction, which shows modification of stem. Interestingly, نأر/*nær* can also function in contexts when combined with verbs describing continuous or interrupted actions:

(3) جوَؤی نأر مے لار گأ
gæ *lār* *me* *nær* *ǰəwəi*
go.PAST.F fall in dance woman.SG
“ The woman fell while dancing”

This dual usage suggests that نأر may also mark imperfective aspect or dynamic action rather than strict plurality. Now, below is a sentence where نأر marks plurality.

(4) جوَؤی نأر کوؤدد
kowdud *nær* *ǰəwəi*
do.PAST dances.PL woman.SG
“ The woman used to perform dances”

Furthermore, the use of tone may also be used to mark plurality along with the argument the verb takes. From the example below in (5), to mark plurality, there is a change in tone, usually from high-to-low for vowel ending plurals like جوَؤی/ *ǰəwəi* /women, along with the verb /کو/ ko/ taking an argument /دی /di/ i-e:

(5) جوَؤی نار کودی
kodi *nar* *ǰəwəi*
do.PRES.PL dance woman.PL[tone=HL]
“ The women dance”

There is no orthographic distinction between these tonal items such as in case of جوَؤی/ *ǰəwəi*, which can be plural or singfular depending on tone. Such constructions can lead to ambiguities while annotating them in CoNLL-U format and can confuse the parser as well, which possibly could be addressed by adding a custom feature like Tone=”_” in the `FEATS` column.

4.4 Tonal distinctions and minimal pairs

In Torwali minimal pairs exist via tonal contrasts, with identical forms differing in pitch, following are some of the examples:

ژاد/*ẓàd*/morning/[HL] vs ژاد/*ẓād*/blood/[H]

ماش /*màš* /uncle[HL] vs ماش/*māš*/fish.sg[H]

پھاپ/*phāp*/lung[H] vs پھاپ/*phàp*/uncle[HL]

Agreement patterns further complicate this; there are some variants of lexical items which the native speakers do not treate as separate lexical items such as (e.g., ژات/ژاد/*ẓād*/*ẓāt*^{h}/blood) and there are other such examples as well. But, as of now there is no evidence that such items are regionally, or phonologically conditioned, and this is early to say that such words show dialectal or contextual allomorphy.

Apart from tonal minimal pairs, interesting observation regarding phoneme minimal pairs is the accusative forms of:

تیس /*tes*/him/her/, کیس /*kes*/whom, میس /*mes*/this

and their oblique counterparts,

تِس /*tis*/him/her, کِس/*kis*/who/whom, مِس/ *mis*/this

the only difference between them is the change from /e/ to /i/.

As the tone plays a very significant role in the language, it affects many areas of Torwali grammar, therefore, there is a need to analyze tone in Torwali from different angles (Lunsford, 2001). Furthermore, this tone-driven ambiguity complicates parsing if two words are only differentiated by tone, and that tone is not encoded, parsers can easily confuse them. We have not yet encoded pitch or tone in our annotation.

4.5 Gender

As presented in examples sentences (1) from section 4.1, Gender in Torwali is not marked on nouns but is instead determined by verbs. The verb agrees with the gender of the preceding noun or argument, as seen in (6) and (7).

(6) سے او پودو
pudu *u* *se*
drink.PRES.M water 3rd.SG
VERB NOUN PRON
“He drinks water”

(7) پویئ او سے
pui *u* *se*
drink.PRES.F water 3^{rd}.SG
VERB NOUN PRON
"She drinks water"

This alignment suggests a head-marking pattern where verbs encode gender information. In order to annotate such a sentence, the gender of almost all 3^{rd} person pronouns must be determined by verb ending and/or suffix, as shown in (6) and (7), because سے*/se = he/she/it/they*, which can either be masculine, feminine, singular or plural.

Other examples where the adjective agrees with the noun it modifies:

(8) تھو گھن باڈ
thu *ghən* *bad*
is.M large/big.M stone
VERB ADJ NOUN
"Stone is big/large"

(9) چھی گھین نہیت
čhi *ghen* *nhet*
is.F large/big.F river
VERB ADJ NOUN
"River is big/large"

Although, grammatical gender can be distinguished by biological gender such as /داد/ *dad/grandfather/* and /دأت */dæt/ grandmother/.*
It is also noted that there is no gender marker for 3^{rd} person in future tense.

(10) بن-نین کتاب تی
bən-nin kitab ti
read.FUT.M.F book 3SG
VERB NOUN PRON
"He will read the book"

In above example, neither the 3^{rd} person /تی/*ti*/ nor the verb gives any information about the gender of the subject. Verb here has a suffix نین /*-nin*, for future tense.

4.6 Gender and number neutralization

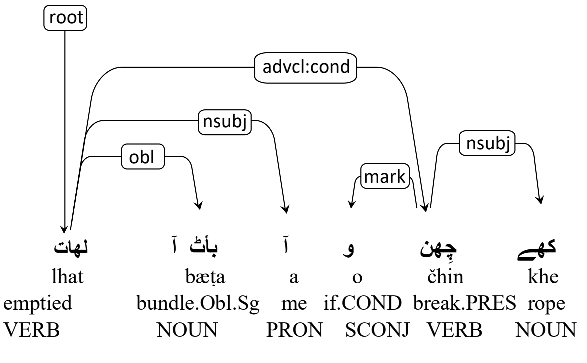
As discussed in sections 4.1, 4.2, gender of nouns and pronouns is marked by verbs and auxiliaries. But the invariant past continuous tense suffix /دود/*dud*/ i-e from لَھنگُودُود/*lhəŋudud*/(was/were entering) shows complete gender and number neutrality. Which contrasts sharply with Urdu's gender and number-sensitive past auxiliaries (/تھا/*tha*/تھی/*thi*/تھے/*the*/, indicating simpler inflectional morphology.
A strange phenomenon arises when the gender and number-invariant third-person pronoun (سے/*se*/ = he/she/it/they) combines with an invariant past continuous tense verb with the suffix /دود/*dud*/ such as لَھنگُودُود /*lhəŋudud*/ ("was/were entering"). This combination makes it particularly difficult to encode explicit grammatical information, like gender and number within the UD framework. While UD can accurately represent this lack of marking, it highlights the degree to which some languages rely on context rather than morphology to convey these fundamental grammatical categories. Below example illustrates the behavior.

(11)
لَھنگُودُود یے دکان سے
lhəŋudud ye dukan se
entering.PST to shop 3SG/3PL.MASC/FEM
"He was entering the shop"

4.7 Conditional constructions

Torwali encodes conditionals morphologically as well as well syntactically. Here is an example of morphological marking of conditional mood with the conditional suffix: -و

(12)


root
advcl:cond
nsubj
obl
mark
nsubj
لھات باٹ آ آ و چِھن کھے
lhat bæṭa a o čhin khe
emptied bundle.Obl.Sg me if.COND break.PRES rope
VERB NOUN PRON SCONJ VERB NOUN

"I would get rid of the bundle if the rope breaks"

چِھن و/*čhin-o*/if break(s) , in above example we have a conditional marker و/o which is dependent upon the head of the phrase as `mark` and the verb چِھن/*čhin* is dependent on the `root` as `advcl:cond`

In Torwali, sentence can also have both morphological and syntactic marking of conditionals, by adding an optioal کو/*ko* to the right of the clause with an already present mandatory conditional marker و/o at the leftmost end of the clause.

(14)

لهات	باأٹ آ	آ	و	چِهن	کهے	کو
lhat	bæṭa	a	o	čhin	khe	ko
emptied	bundle.Obl.SG	me	if.COND	break	rope	if
VERB	NOUN	PRON	SCONJ	VERB	NOUN	SCONJ

Such constructions show that in Torwali, "کو/*ko*" is somewhat optional which typically serve to reinforce the conditional meaning, clarify clause boundaries, or emphasize tense/aspect/discourse nuances. Somewhat like Urdu and Pashto in the following examples (both meaning "If he had come, I would have gone"):

Urdu:

اگر وہ آتا تو میں جاتا۔
Agar voh ātā to māi jātā

Pashto:

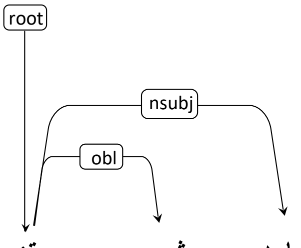
که هغه راغلی وای، نو زه تللی وم
ka haġa rāġlī wāy, no za talelī wom

4.8 Case

Based on Sir Aurel Stein's collection of three historical texts from 1926, Grierson's 1929 manuscript examines key grammatical features of Torwali. He also outlines the noun case system by identifying eight cases: nominative, accusative, ergative, instrumental, dative, ablative, genitive, and locative.

In Torwali, case suffixes are attached to nouns, which sometimes are phonologically bound and cannot stand alone. We treated the case markers in Torwali like the ergative suffix a/ ا *or e/*ے in annotation based on their phonological properties. Since the case suffixes we encountered were phonologically bound, we chose to treat them as a single token because noun+case-suffix behaved more like a single unit due to tight phonological bonding and lack of syntactic separability between the two elements. For example, as shown in (12) from section 4.7, باٹ-آ/bæṭ-a, bæṭ/bundle is noun and bæṭ-a/bundle.obl is the oblique case. Which usually is pronounced as a single unit. Other examples (15) and (16) below shows how we treated شیرے / *šire*/ in the house, and شیر مے/*šir me*/ in the house/, in the sentences below:

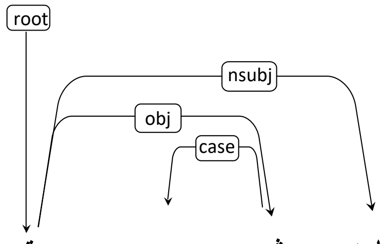
(15) *Šir-e:*



تهو	شیرے	ساجد
thu	šir-e	Sajid
is.aux	house.obl.sg	Sajid
VERB	NOUN	NOUN

"Sajid is in the house"

(16) *Šir me:*



تهو	مے	شیر	ساجد
thu	me	šir	Sajid
is.AUX	in.ADP	house	Sajid
VERB	ADP	NOUN	NOUN

"Sajid is in the house"

In (16),we treated مے /*me*/ as a postpostion based on the fact that مے/ *me*/ , is usually used in torwali as a separate word meaning "in".

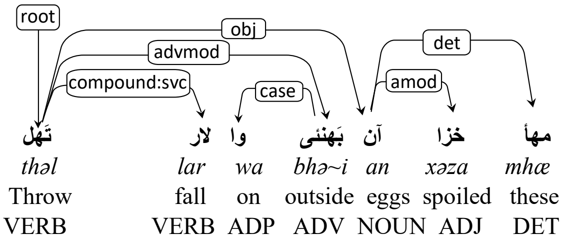
4.9 Compound verb constructions

Torwali exhibits productive compound verb formations, which involves a sizable number of verb-noun and verb-verb concatenations exemplified by the following verb-verb compound.

لار تَهل
thəl lar
throw fall

تَهل */thəl/*throw functions as the main verb and لار */lar/*fall functions as the aspectual light verb modifying the main verb. To annotate such constructions, the `compound:svc` relation is used as shown in the following example.

(17) # text = مهأ خزا آن بَهنئی وا لار تَهل



“Throw these spoiled eggs outside”

If we jump back to example sentence from (4.3)

گأ	لار	مے	نأر	جوَؤی
gæ	*lār*	*me*	*nær*	*ǰəwəi*
go.PST.F	fall	in	dance	woman.SG

“ the woman fell while dancing”

Here again, لار گأ/ *lar gae* is a compound verb where گأ /gæ act as a Feminine past auxiliary verb (from بیو /*bəyu*/ to go) and shows perfective aspect (completed action), and مے/*me*/during is a subordinating conjunction.

Similarly, for noun-verb concatenation, we used treated the verb as the head and we used the relation `compound:lvc` which is used in other Indo-Aryan languages as well, in which such construction exists. Example sentence (1) from section 4.1 shows a noun-verb complex.

4.10 Multiword tokens and reduplication

We also encountered idiomatic adverbial phrases like the one given below:

بهیس پہ بهیسا/ bhes pə bhesa/ for no reason

In this case, "بهیسا/bhesa" might have developed a pragmatic extension and used standalone for *for no reason* or *without reason.* But, "بهیس پہ بهیسا/bhes pə bhesa" as a whole is a reinforced idiom, a structure that emphasizes the meaning by repeating or echoing. The entire unit behaves as syntactically atomic and for now we have annotated and treated "بهیس پہ بهیسا/ bhes pə bhesa " and other such phrases as a multiword token with relation to the head as `advmod`.

In Torwali, reduplication is also attested, likely serving derivational or intensifying functions (e.g., pluralization, aspectual marking) and we treated them as multiword tokens. Below are some examples of such words.

/گیل میل/	gel mel/	bread-and-all/
/چُن چُن/	čun čun/	very small/
/پہٹ پہٹ/	phiṭ phiṭ/	pieces/
/چئ مئ/	čəi məi/	tea and such/
/دستی دستی/	dəsti dəsti/	very quickly/

There is also verb repetition in Torwali, like the phrase "بهییل بهییل" /*bhəyel bhəyel*/ is a reduplicated verb form that functions as an adverbial phrase meaning "while sitting" or "in the midst of sitting". We treated such reduplicated verbs as a single token.

Conclusion

Torwali displays mixed characteristics, but the majority of its features are those of a fusional language, employing multiple strategies to convey grammatical and semantic information. This includes stem modification, suprasegmental changes, and reduplication, all of which are used to modify the meanings of words (Lunsford, 2001). In addition, there is extensive use of pitch and tone as grammatical markers, playing a crucial role in distinguishing between word forms and meanings.

This work presents an analysis of the basic grammatical and linguistic features of the Torwali, documented during morphosyntactic annotation and the development of a treebank. As the annotation process continues and the treebank expands, we expect to encounter additional morphosyntactic features, contributing further to our understanding of the language’s structure and complexity.

Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology); and from the grants "Language Understanding: from Syntax to Discourse of the Czech Science Foundation (Project No. 20-

16819X)"; and LINDAT/CLARIAH-CZ (Project No. LM2023062) of the Ministry of Education, Youth, and Sports of the Czech Republic.

We extend our thanks to Mr.Allauddin for his contributions in linguistic analysis of Torwali texts and for guiding us with his linguistic knowledge of Torwali, and to Mr. Inam Ullah for his support in sharing his valuable Torwali language data.

References

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

Ullah, Inam. 2004. ‘Lexical Database of the Torwali Dictionary.’ In The Asia lexicography conference. Chiangmai: Payap University

Lunsford, Wayne. 2001. An Overview of Linguistic Structures in Torwali, A Language of Northern Pakistan. M.A. Thesis, University of Texas at Arlington.

Grierson, George A. 1929. Torwali: An Account of a Dardic Language of the Swat Kohistan. Royal Asiatic Society. London.

Biddulph, John. 1880. *Tribes of the Hindoo Koosh*. Calcutta: Office of the Superintendent of Government Printing.

Moseley, Christopher. 2010. *UNESCO Atlas of the World's Languages in Danger*. UNESCO. Available online at: http://www.unesco.org/culture/languages-atlas.

Liljegren, H., & Akhunzada, F. (2017). Linguistic diversity, vitality and maintenance : A case study on the language situation in northern Pakistan. Multiethnica. Meddelande Från Centrum För Multietnisk Forskning, Uppsala Universitet, (36–37), 61–79. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-148722

Schuster, Sebastian and Manning, Christopher D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 2371–2378. European Language Resources Association (ELRA), Portorož, Slovenia.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. Linguistic Typology, 23(3):533–572.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Ullah, Inam (2019) "Digital Dictionary Development for Torwali, A Less-studied Language: Process and Challenges," Proceedings of the Workshop on Computational Methods for Endangered Languages: Vol. 2 , Article

Uddin, Naeem., & Uddin, Jalal. (2019). A step towards Torwali machine translation: an analysis of morphosyntactic challenges in a low-resource language. In Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, 6-10.

Hyunji Hayley Park, Lane Schwartz, and Francis Ty ers. 2021. Expanding Universal Dependencies for polysynthetic languages: A case of St. Lawrence Is land Yupik. In Proceedings of the First Workshop on Natural Language Processing for Indigenous Lan guages of the Americas, pages 131–142, Online. As sociation for Computational Linguistics.

Aryaman Arora. 2022. Universal Dependencies for Punjabi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5705–5711, Marseille, France. European Language Resources Association.

Alexey Koshevoy, Anastasia Panova, and Ilya Makarchuk. 2023. Building a Universal Dependencies Treebank for a Polysynthetic Language: the Case of Abaza. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 1–6, Washington, D.C.. Association for Computational Linguistics.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal Semantic Parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark. Association for Computational Linguistics.

Raverty, Henry George. *An Account of Upper and Lower Suwat, and the Kohistan, to the Source of the Suwat River; with an Account of the Tribes Inhabiting those Valleys*. Journal of the Asiatic Society of Bengal, vol. 31, no. III, 1862, pp. 227–281.

Rensch, Calvin R. 1992. Patterns of language use among the Kohistani of the Swat valley.

Sociolinguistic survey of northern Pakistan, Volume 1, Languages of Kohistan, by Calvin R. Rensch, Sandra J. Decker, and Daniel G. Hallberg. Islamabad: National Institute of Pakistan Studies and Summer Institute of Linguistics.