Universal Dependencies Treebank for Khoekhoe (KDT)

Kira Tulchynska[1], Sylvanus Job[2], Alena Witzlack-Makarevich[1,3]

[1]Hebrew University of Jerusalem, [2]University of Namibia, [3]University of Cologne,
Correspondence: kira.tulchynska@mail.huji.ac.il

Abstract

This paper reports on the development of the first dependency treebank for Khoekhoe (KDT). Khoekhoe (Khoe-Kwadi, Namibia) is a low-resource language with few linguistic and computational resources available publicly. This treebank consists of 29k words across six texts taken from various registers. It includes a substantial portion of spoken conversational data. These sentences were annotated manually according to the Universal Dependencies framework. In this paper, apart from presenting the strategies that have been followed to create the treebank, we also discussed some challenging morphological features and syntactic constructions found in the corpus and outlined how we have handled them using the current Universal Dependencies specification.

1 Introduction

Universal Dependencies (UD) is a cross-linguistically consistent framework for annotating various aspects of morphosyntax (McDonald et al., 2013; Nivre et al., 2016; de Marneffe et al., 2021). It provides a set of guidelines that enable comparison between treebanks across diverse linguistic typologies.

This paper reports the development of the Khoekhoe Dependency Treebank (KDT) a UD treebank for Khoekhoe, a Khoe-Kwadi language spoken primarily in Namibia. The treebank has been released as UD_Khoekhoe-KDT, an official Universal Dependencies (UD) treebank, available at `https://universaldependencies.org/treebanks/naq_kdt/` under the CC BY-SA 4.0 license. This represents the first UD treebank not only for this language family but also for any of the so-called Khoisan languages, a group of about a dozen languages of Southern Africa characterized by an extensive use of click consonants in their phonology (Güldemann, 2014; Witzlack-Makarevich and Nakagawa, 2019).

The annotation process highlighted several morphosyntactic properties of Khoekhoe that required careful consideration within the UD framework. These include the treatment of mood particles, valency-changing suffixes (e.g. applicative, reflexive), the use of nominalization as a clausal embedding strategy, and reported speech constructions involving quotative particles. In each case, we propose UD-compliant solutions and, where appropriate, introduce new feature values or relation subtypes.

2 Khoekhoe

Khoekhoe (`ISO 639-3: naq`; `Glottocode: nama1264`), also known as Nama-Damara, is a Khoe language of the Khoe-Kwadi family (Güldemann, 2014). It is spoken mainly in central and southern Namibia. Khoekhoe represents a dialect continuum (Haacke, 2018), it includes the northern varieties of Hai‖om and ǁĀkhoe, which differ considerably from the central and southern varieties. This paper is based on the corpus which includes only the latter varieties, which are closer to what is considered as standardized Khoekhoe.

With approximately 245,000 speakers (11.8% of the total population of Namibia), Khoekhoe is the second most spoken language in the country after Oshiwambo (Namibia Statistics Agency, 2011). Although Khoekhoe is well-documented and receives official recognition as a language of instruction in Namibia's educational system (Brenzinger, 2013), and despite being the largest non-Bantu click language in Africa, Haacke and Eiseb (2002) consider it to be an 'endangered language'.

The majority of Khoekhoe speakers are multilingual: many possess at least a functional proficiency in Afrikaans, with English fluency steadily increasing since Namibia's independence (Haacke, 2018, p. 142). Consequently, code-switching and word borrowing from Afrikaans and English to Khoe-

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 119–128

khoe are very frequent.

3 Corpus

The Khoekhoe corpus used for the treebank currently contains approximately 29k words across six texts. The selection was primarily determined by practical considerations (availability and need of interlinear glossed texts for other research purposes). We balanced the corpus to include texts from multiple modes and genres. Currently, the corpus is skewed towards translated texts. We plan to expand its original spoken and written component in later releases.

One text (5.5k words) is a transcription of a conversation in Khoekhoe between two friends. It was transcribed and translated into English by native speakers, then supplemented with interlinear morphosyntactic glosses in ELAN (Max Planck Institute for Psycholinguistics, 2024). All the sentences containing personal information that could identify the speakers were removed. The remaining sentences were shuffled to ensure further anonymization.

All the other five texts are written. They include translations of the Cairo CICLing Corpus[1] (215 words) and the BivalTyp dataset (Witzlack-Makarevich and Job, 2024) (750 words); subtitles of the films *Bridge of Spies* (2015) (14k words) and a section of *Titanic* (1997) (6k words) translated from English by a native speaker; and chapters from the schoolbook *Khomai da ra I* (1971) with sentences shuffled (2k words). Also these texts were supplemented with interlinear morphosyntactic glosses in ELAN (Max Planck Institute for Psycholinguistics, 2024).

Text	Mode	Text origin	Size
conversation_Windhoek5	Spoken	Original	5.5k
grammar_Cairo	Written	Translation	215
grammar_BivalTyp	Written	Translation	750
film_Bridge	Written	Translation	14k
film_Titanic	Written	Translation	6k
book_Khomai	Written	Original	2k

Table 1: Corpus information

4 Methodology and Overview

The treebank annotations were performed manually by the first author on the basis interlinear morphosyntactic glosses. We plan to cross-check the annotations in the future versions of the treebank. Ambiguous structures were resolved through consensus among all authors.

For the first 60 sentences (330 tokens), a script was created for tokenization and conversion of interlinear morphosyntactic glosses into CoNLL-U format. This annotated sample was then used to train a UDPipe 1 (Straka and Straková, 2017) pipeline. This pipeline was subsequently applied to parse the next batch of sentences. The parsed sentences were manually reviewed, and this process was repeated iteratively until the entire corpus was annotated. UDPipe 1 was selected over UDPipe 2 (Straka, 2018) for this process because it is more user-friendly, and any potential differences in annotation quality would become negligible after manual review.

Annotation guidelines were developed alongside the annotation process, with all newly encountered issues being addressed before proceeding with the next batch.

Of the 17 universal part-of-speech tags recognized by the UD framework, only `SYM` is not currently used in the KDT treebank. Table 2 provides an overview of the frequencies of the universal part-of-speech tags in KDT.

Class	UPOS	Count	%
Open	ADJ	775	2.67
	ADV	1680	5.79
	INTJ	635	2.19
	NOUN	3537	12.19
	PROPN	820	2.83
	VERB	3941	13.59
Closed	ADP	1405	4.84
	AUX	5694	19.63
	CCONJ	652	2.25
	DET	827	2.85
	NUM	191	0.66
	PART	477	1.64
	PRON	3269	11.27
	SCONJ	917	3.16
Other	PUNCT	4144	14.29
	X	44	0.15

Table 2: UPOS tags and their frequencies in KDT

Table 3 presents the features, their values, and counts in KDT, with features not currently recognized by UD marked with †. Some features (e.g. `Assoc`) are introduced here for the first time, while others (e.g. `Voice=Rfl`) have been used in other treebanks. The new features and values, as well as language-specific decisions regarding the feature used with the declarative auxiliary *ge*, are discussed in Section 5.

[1] https://github.com/UniversalDependencies/cairo/tree/master

Feature	Values	Count	%
Abbr	Yes	21	0.05
Aspect	Imp, Perf, Punct†	1335	2.96
Assoc†	Yes†	17	0.04
Case	Acc, Nom, Voc	7240	16.07
Clusivity	Ex, In	135	0.30
Degree	Dim	86	0.19
Deixis	Contr†, Prox, Remt	860	1.91
Evident	Nfh	8	0.02
ExtPos	ADP, ADV, CCONJ, DET, SCONJ	56	0.12
Foreign	Yes	31	0.07
Gender	Fem, Masc, Neut	6286	13.96
Mood	Ass†, Imp, Ind, Int, Pot, Prh†	2694	5.98
Number	Dual, Plur, Sing	7453	16.55
NumType	Card, Ord	225	0.50
Person	1, 2, 3	7421	16.48
Polarity	Pos, Neg	758	1.68
Poss	Yes	238	0.53
PronType	Dem, Emp, Ind, Int, Neg, Prs, Rel, Tot	4410	9.79
Tense	Fut, Past, Pres, RecPast†	1758	3.90
Typo	Yes	77	0.17
Voice	Act, Appl†, ApplPass†, ApplRefl†, Pass, Rcp, Refl†	3930	8.73

Table 3: Features, values and their frequencies in KDT. † marks features and values introduced in KDT

Finally, the 15 most frequent dependencies in KDT are listed in Table 4. Syntactic analysis and annotation decisions for language-specific structures are discussed in Section 6.

Dependency	Count	%
aux	5299	18.27
punct	4144	14.29
root	3589	12.37
nsubj	2859	9.86
advmod	1714	5.91
case	1455	5.02
obj	1270	4.38
obl	1016	3.50
mark	975	3.36
cc	662	2.28
nmod:poss	660	2.28
det	564	1.94
amod	455	1.57
conj	419	1.44
advcl	386	1.33
...		
other	3541	12.21

Table 4: Dependencies and their frequencies in KDT

5 Morphology

5.1 Associative Plural

Associative plurals are found in most languages across Australia, Asia, and Africa. They are almost entirely absent in Western Europe. This structure conveys the meaning of 'and others associated with it' (Daniel and Moravcsik, 2013).

In Khoekhoe, the associative plural is marked by the suffix *-hâ*, which precedes the person-gender-number suffix. For example, *Jackhân* means 'Jack and company'; *saruhân* (derived from *saru* 'cigarette') means 'cigarettes and other things associated with smoking' (see Hagman 1977, p. 29). The associative plural suffix precedes either a dual or a plural suffix (e.g. the common plural suffix *-n* in the two examples). Thus, it cannot be analyzed as a value of the `Number` feature. Instead, we introduce an `Assoc` feature with the value `Yes` to account for its function.

5.2 Mood Values

The values which are coded as `Mood` in the UD annotation scheme include both the typical mood categories (sometimes referred to as *verbal mood*), such as potential, as well as values which could be better analyzed as *sentence mood*, such as interrogative and imperative (see Portner 2018, pp. 4–5). These two sub-types of mood are not distinguished in UD.

For Khoekhoe we use seven mood values. In addition to the familiar imperative (`Imp`), potential (`Pot`), and interrogative (`Int`) moods, Khoekhoe has four further moods, for which dedicated values had to be introduced to UD. These are the indicative (`Ind`, apprehensive (`App`), assertive (`Ass`), and prohibitive (`Imp`) moods. All grammatical mood categories in Khoekhoe are expressed by grammatical particles and are coded as `AUX` in the UD annotation scheme. The individual mood auxiliaries are in the majority of cases in complementary distribution. With some exceptions, they occur in the clause-second position.

5.2.1 Indicative vs. declarative

In Khoekhoe, declarative sentences can and are often marked with the particle *ge*, as in (1). The function of this particle can be best captured with the sentence mood label *declarative* (Portner 2018, pp. 4–5), as it does not occur in interrogative and imperative sentences. However, we decided against introducing a new `Mood` value and use instead the existing verbal mood value `Ind` in the annotation to enhance the cross-linguistic comparability of the annotations.

(1) *Sara-s* *ge* *ǁkhai-s-a* *ūhâ.*
Sara-3F.SG.SBJ DECL flu-3F.SG-OBL have
'Sara has the fly.'

5.2.2 Prohibitive

Whereas some languages use their regular sentential negation strategy to express prohibitions directed at second person (also known as negative imperatives), a substantial proportion of world's languages use to this end dedicated negation strategies not found in declaratives (see van der Auwera et al. 2013). In Khoekhoe, the regular negative particles are *tama* in the present and past and *tide* in the future. By contrast, in the prohibitive mood a dedicated particle *tā* is used instead. We introduce the `Mood` value `Prh` to annotate the auxiliary *tā* in this function, as illustrated in (2).

(2) *Tā* *ti* *ôa-b-a* *ǂnau.*
PROH my child-3M.SG-OBL beat
'Do not beat my child.'

5.2.3 Apprehensive

The term *apprehensive construction* is used to refer to constructions which conventionally encode, or pragmatically implicate, that the situation described by the clause is an undesirable possibility (Vuillermet et al., 2025+). As such, apprehensionality is a subtype of modality (AnderBois and Dąbkowski, 2025). In some languages, such as Khoekhoe, apprehensives are part of a grammatical mood paradigm, while in others, they are part of a different syntactic class and can include e.g. markers of negation (Vuillermet et al., 2025+).

We introduce a `Mood` value `App` to annotate the apprehensive mood marker *tā*, as in (3). This is the same particle used to mark the prohibitive (see Section 5.2.2), as is not uncommon cross-linguistically. However, the two contexts are easily distinguished: whereas the prohibitive is used in the imperative construction directed at the second person, which does not have an over expression of subject in the singular, the apprehensive is used in sentences which have overtly expressed subjects, such as *aorob* 'man' in (3).

(3) *Nēsisa* *aoro-b* *ge* *tā* *ti-ta*
now man-3M.SG.SBJ DECL APPR 1SG-1SG.OBL
nî *|gū.*
FUT come_near
'Now the man should not come near me.'

5.2.4 Assertive

Cross-linguistically, there is a range of forms and constructions whose relation to core mood is not currently well understood. Among them Portner (2018, p. 7) lists assertive. The details of the use of the assertive auxiliary in Khoekhoe are still understudied. It seems that speakers uses this form to emphasize the fact that they do not take responsibility or authority over the assertion. Thus, it might be more appropriately characterized as a member of the evidentiality system and not of the mood system, though cross-linguistically the two can share the same slot and be in complementary distribution, as is the case in Khoekhoe (see Aikhenvald 2003).

In Khoekhoe, the assertive marker consist of two parts *kom(o)* ... *o*, as in (4). The first part *kom(o)* occurs in the clause-second position, the second element *o* is clause-final. Since the assertive marker is discontnious and does not enclose the whole clause, the structure could not be analyzed as a fixed multi-word expression. Thus, both parts are coded as auxiliaries with `Mood=Ass`.

(4) *ǁNā-n* *komo* *awoxa-n*
DIST-3C.PL.SBJ ASSERT1 ancestor-3C.PL
ǂû-n-a *o.*
food-3C.PL-OBL ASSERT2
'Those are the foods of the ancestors.'

5.3 Voice Values

Khoekhoe has four valency-changing verbal suffixes: passive, reciprocal, reflexive, and applicative (Hagman, 1977, pp. 77–82). Two of these (passive and reciprocal) are already recognized as possible `Voice` values in UD, while the other two (reflexive and applicative) are not.

In UD, reflexivity is recognized as a possible feature of pronouns and determiners and is annotated with `Reflex`. However, in Khoekhoe, as in many other languages of the world (see e.g. 'Feature GB114: Is there a phonologically bound reflexive marker on the verb?' in Skirgård et al. 2023), reflexive events are encoded with a dedicated reflexive voice marked by the verbal suffix *-sen* (Hagman, 1977, pp. 81–82). It indicates that the agent and the patient are coreferential and decreases the syntactic valency of the predicate by one (Zúñiga and Kittilä, 2019, pp. 154–155), as in (5).

(5) *Hui-sen=ta* *ge.*
help-REFL=1SG DECL
'I help myself.'

This type of reflexive construction is not yet covered by the UD guidelines. To capture this typologically common feature, we use the feature value `Voice=Rfl`, which is already in use in the Abaza, Turkish, and Turkish-German treebanks.

The applicative construction increases the syntactic valency of the verb by one: it introduces a new direct object that corresponds to a non-core argument in the non-applicative voice (Zúñiga and Kittilä, 2019, p. 53). The applicative voice in Khoekhoe is marked by the suffix *-ba* (Hagman, 1977, pp. 78–79), as in (6). To capture this morphosyntactic behavior, we use `Appl` as a value of the `Voice` feature.

(6) *Ti-ta* *ge* *ǂhanu-b-a*
1SG-1SG.SBJ DECL government-3M.SG-OBL
sîsen-ba *tama* *hâ.*
work-APPL NEG PFV
'I don't **work for** the government.'

Additionally, Khoekhoe verbal morphology allows the applicative voice to combine with the other voice categories (Hagman, 1977, p. 77). This is typologically common, e.g. Zúñiga and Kittilä (2019, p. 75) suggest that in many languages that have both applicative and passive morphological markers, these categories co-occur, as they do in Khoekhoe, as in (7). To account for such combinations, we introduce the following complex values: `ApplPass`, `ApplRcp`, and `ApplRfl`.

(7) *Sîsenao-gu* *ge* *kausa* *tsaugoma-b-a*
worker-3M.PL.SBJ DECL fat ox-3M.SG-OBL
go *ǂā-ba-he.*
RPST slaughter-APPL-PASS
'One fat ox was slaughtered for the workers.
(lit. The workers were slaughtered one fat ox for.)'

As in some other languages, Khoekhoe has a number of morphologically reflexive, reciprocal, and applicative verbs which are arguably lexicalized. Geniušienė (1987, p. 31) discusses this common issue with respect to reflexives and proposes a distinction between the familiar reversible reflexive verbs and the less studied class of non-reversible reflexive verbs. She suggests the following criteria of reversibility (Geniušienė, 1987, pp. 145–148) to distinguish between the two: (1) morphological reversibility, i.e. a situation when a derived unit is formally related to a base word, morphological non-reversibles are traditionally known as *reflexiva tantum*; (2) syntactic reversibility, viz. a change of reversible reflexive properties according to one of the regular patterns; (3) lexical reversibility, viz. the identity of lexical distribution relative to the corresponding syntactic positions in a non-reflexive construction and related reflexive construction; (4) semantic reversibility, viz. a regular, standard change of the meaning of a reflexive, thus, semantic non-reversible reflexive verbs have the meaning which is related to that of the base non-reflexive way in some idiosyncratic way. Individual morphologically reflexive verbs can be non-reversible according to several of the criteria (2) to (4). Similar observations apply to reciprocal and applicative verbs, see e.g. Peterson (2006, pp. 169–170) and individual contributions in Zuniga and Creissels (2024) on the lexicalization of applicatives.

Though in theory the distinction between lexicalized and productive cases might be clear and there are straightforward examples of lexicalization in Khoekhoe, such as e.g. *mû* 'to see' vs. *mûsen* 'to appear, to look like', in other cases the application of the criteria in (2) to (4) can be admittedly challenging. In KDT we do not annotate for voice unambiguous cases of lexicalized reflexives, reciprocals, and applicatives. The decision is made on a lexeme-by-lexeme basis.

6 Syntax

6.1 Quotative marker

Many African languages have particle-like function words dedicated to marking reported discourse (Güldemann, 2008, pp. 122–124). In Khoekhoe, *ti* is used as the quotative marker. It directly follows the quoted material, as in (8) (Hagman, 1977, pp. 136–138).

(8) *[|Owesa=ta* *a]* ***ti*=*n*** *ge*
lazy=1SG.SBJ COP.PRS QUOT=3C.PL.SBJ DECL
ra *mî.*
IPFV say
'They say [I am lazy].'

The complementizer *!khais* is sometimes used in place of the quotative marker *ti* to express indirect speech, as in (9). This usage of *ti* can therefore be

analyzed and annotated as a subordinating conjunction marker (mark), as in Figure 1.

(9) *Mî-ba* *tsi=ta* *go* *[‖nā*
say-APPL 2M.SG.OBJ=1SG.SBJ RPST DIST
‖khami-s-a=ts *nî* *ū* *ǂgao]*
phone-3F.SG-OBL=2M.SG.SBJ FUT take want
!khai-s-a
COMPL-3F.SG-OBL
'I told you that you wanted to take the call.'

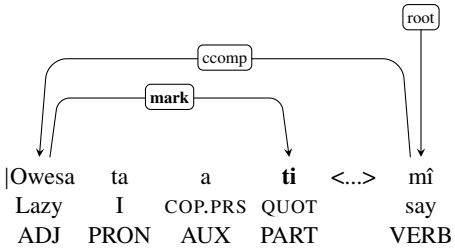

Figure 1: Annotation of (8)

This quotative marker is also used to introduce noun phrases expressing names and labels, as in (10). This usage is annotated with the case relation, as in Figure 2. The marker was preliminary tagged as PART, but it will be changed to SCONJ in the next version of the treebank.

(10) *Nē* *‖gaba-s* *ge* *[kompas]* ***ti***
PROX tool-3F.SG.SBJ DECL compass QUOT
ra *‖gai-he.*
IPFV call-PASS
'This tool is called a compass.'

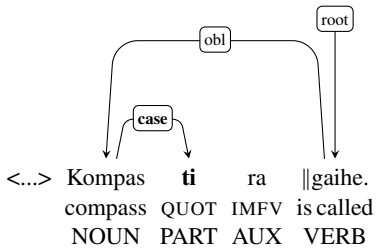

Figure 2: Annotation of (10)

6.2 Nominalization

In Khoekhoe, one way to embed a clause is through nominalization. This process involves the use of the nominalizer enclitic *=s*, which is formally identical to the third-person feminine singular suffix and enclitic. The nominalized clause then functions as a nominal within the sentence, that is, it can take the oblique case suffix, it can be followed by a postposition, and it can functions as a clausal argument, e.g. object (Hagman, 1977, pp. 123–135), as in (11).

(11) *ǂÂihō=ta* *ge* *ra* *‖î-n*
remeber=1SG.SBJ DECL IPFV 3-3PL.SBJ
gere *‖nāti* *mî=s-a*
PST.IPFV like_that say=NMLZ-OBL
'I remember them saying like that.'

6.2.1 Analysis of nominalized constructions

Nominalized constructions can be analyzed in two ways: either with the nominalizer as the head of the structure or with it depending on the root of the nominalized clause.

The first approach treats the nominalizer as an *empty noun*, with the nominalized clause acting as a relative clause clause modifier. Under this analysis, the literal translation of the nominalized structure in (11) would be 'one/the thing which is them saying'. This annotation is illustrated in Figure 3.

The second approach treats the nominalizer as a subordination marker dependent on the root of the nominalized clause. In this case, the literal translation of the nominalized structure in (11) would be 'that they said'. This annotation is illustrated in Figure 4.

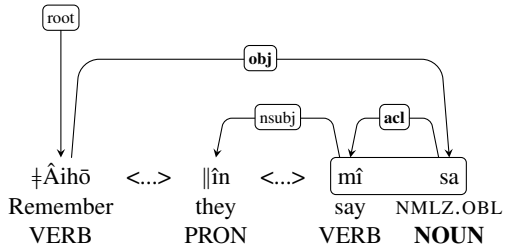

Figure 3: Head-nominalizer analysis of (11)

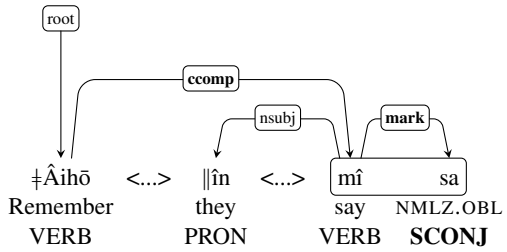

Figure 4: Dependent-nominalizer analysis of (11)

A key advantage of the head-nominalizer analysis is that it highlights the fully nominal function of the nominalized constituent: the root of the structure is a noun, it depends on the main clause's root through a nominal dependency (e.g. `obj` vs. `ccomp`), and it can directly take a `case` dependency from a postposition.

On the other hand, the dependent-nominalizer analysis more directly represents the embedded clause and aligns better with cross-linguistic annotation practices. This approach focuses on the syntactic relation between the subordinate and main clauses, rather than the language-specific realization of it. Given its greater cross-linguistic consistency, we adopt the dependent-nominalizer analysis, treating the nominalizer as a subordination marker.

6.2.2 Nominalization with additive focus

Another usage of the nominalizer *=s* in Khoekhoe is on the constituents modified by the additive focus adverb *tsîn* 'also', when the focused element that precedes the adverb is not a nominal. The nominalized focused element is not necessarily a clause – as in (12) – but can be a constituent of any type, e.g. a postpositional phrase, as in (13). The combination of the nominalizer and the additive focus particle, *=s tsîn*, is analyzed as a fixed multiword expression (and thus annotated with `fixed`), which functions as an emphasizing adverbial (`advmod:emph`, with `ExtPos=ADV` on the first element, *s*). The annotation of the structure is illustrated in Figure 13.

(12)
Tsî=s	*ge*	*!hāsara*	*te=s*	
and=2F.SG.SBJ	DECL	insult	1SG.OBJ=NMLZ	
tsîn-a	*go*	*dī.*		
also-OBL	RPST	do		

'And you've also insulted me.
(Lit. And you've also done the insulting of me.)'

(13)
‖î-s	*!nâ=s*	*tsîn-a=ta*		*ge*
3-3F.SG	in=NMLZ	also-OBL=1SG.SBJ		DECL
‖khaisa	*kurixa*	*|gôa-b-a*	*ūhâ.*	
eight	year_old	son-3M.SG-OBL	have	

'Also with (lit. in) her I have an eight-year-old son.'

6.3 Flat Structures

UD already recognizes two subtypes of the `flat` relation: `flat:foreign` and `flat:name`. For Khoekhoe, we introduce three additional subtypes: `flat:num`, `flat:reparandum`, and `flat:title`.

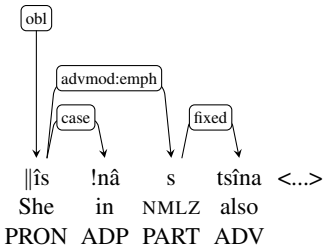

Figure 5: Annotation of *=s tsîn* structure in (13)

6.3.1 Numerals

In Khoekhoe, larger numbers are expressed using separate words for different place values (e.g. *‖khaisadīsi haka kurigu* 'eighty-four years'). These multi-word numerals may include a coordinating conjunction (e.g. *|gui|oadīsi khoesekaidīsi **tsî** hakadīs |gui|a‖î kuri* 'one thousand nine hundred **and** forty-first year'), but otherwise, the structure lacks a clear syntactic head.

The final word in the numeral can be either a cardinal (annotated as `NUM` with `NumType=Card`) or an ordinal (annotated as `ADJ` with `NumType=Ord`). All preceding parts of the numeral are always cardinals and are annotated accordingly.

One approach to annotating such numerals is to use the `compound` relation, as is done in languages such as English and Russian. However, it has two drawbacks. First, the `compound` relation is not used elsewhere in Khoekhoe, making this an isolated and atypical application. Second, it is unclear which word in the multi-word numeral should be treated as the root.

An alternative approach is to use a numeral-specific subtype of the `flat` relation. The `flat:num(ber)` relation is already used for similar cases in Komi Zyrian, Persian, and Vietnamese.

Given these considerations, we adopt an analysis where multi-word numerals are treated as (mostly) flat structures. If a coordinating conjunction is present, the `conj` and `cc` relations are used to connect coordinating elements and the conjunct *tsî* 'and.' Within these elements, numerals are linked to the first word using the `flat:num` relation. Thus, the first word in the structure is treated as a technical root of the numeral.

Depending on whether the numeral expresses a cardinal or an ordinal number (determined by the final word), the first word in the structure attaches to the modified noun using either the `nummod` relation (Figure 6) or the `amod` relation (Figure 7).

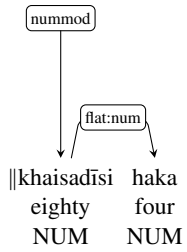

Figure 6: Cardinal multi-word numeral

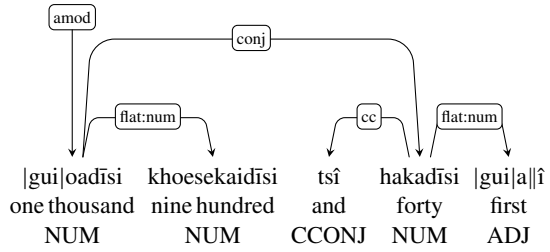

Figure 7: Ordinal multi-word numeral

6.3.2 False start

Part of KDT include various speech disfluencies, such as false starts. In the transcription conventions applied to compile the corpus, false starts are marked with square brackets.

If a false start contains an incomplete word, it is tagged as `X` and left unannotated for morphological features (see Figure 8). Otherwise, if the word is complete, it is fully morphologically annotated (see Figure 9). In all cases, a false start is treated as an overridden disfluency and therefore depends on the repair using the `reparandum` relation.

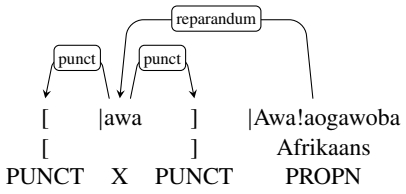

Figure 8: False start with an incomplete word

In many instances, speakers produce multiple false starts of the same word (e.g. *[ti ti] ti zulas* '[my my] my struggle'). One possible approach would be to attach each false start separately to the repair using the `reparandum` relation. However, this would prevent the paired punctuation marks surrounding the false start from attaching to the same word without creating a non-projective structure. To avoid this issue, we instead analyze the entire false start as a flat structure and annotate it using the newly introduced `flat:reparandum` relation, as in Figure 10.

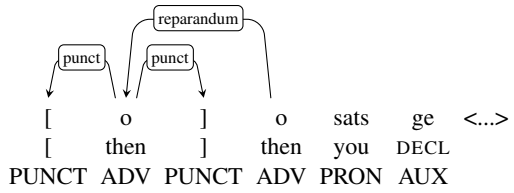

Figure 9: False start with a complete word

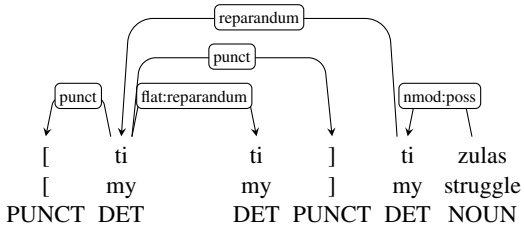

Figure 10: Multiple false starts

6.3.3 Titles

Another common type of flat structure consists of noun phrases that include a title and a proper name, such as *Mr. Smith*, *painter Picasso*, and *brother Sam*. In fact, more than a third (80 out of 218) of the `flat` relations in the Khoekhoe treebank involve a `NOUN` representing a title, profession, or kinship term of a following `PROPN`. Given the frequency of this pattern, we introduce the `flat:title` relation subtype to account for these structures, as in Figure 11.

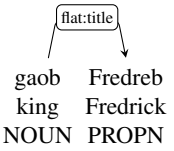

Figure 11: Flat title relation

7 Conclusion

The Khoekhoe treebank presented in this paper is the first case of a Khoisan language added to UD. Khoekhoe has a range of typologically common mood and grammatical voice features so far underrepresented in UD treebanks. We furthermore present several solutions to issues present primarily in corpora of spontaneous spoken language, such as false starts.

Acknowledgments

This research has been funded by Israel Science Foundation, project "Peripheral Khoekhoe varieties: A comprehensive documentation and description" (Personal research grant no. 2892/20); Global Strategy and Partnerships Funding Scheme of the University of Zurich, project "Event Packaging in Language"; the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 281511265 – SFB 1252 "Prominence in Language" at the University of Cologne.

References

Alexandra Y. Aikhenvald. 2003. Evidentiality in typological perspective. In Alexandra Y. Aikhenvald and R. M.W. Dixon, editors, *Studies in evidentiality*, pages 1–33. John Benjamins, Amsterdam.

Scott AnderBois and Maksymilian Dąbkowski. 2025. The semantics and expression of apprehensional modality. *Language and Linguistics Compass*, 19(1):e70002.

Matthias Brenzinger. 2013. The twelve modern Khoisan languages. In Alena Witzlack-Makarevich and Martina Ernszt, editors, *Khoisan languages and linguistics: Proceedings of the 3rd International Symposium, July 6-10, 2008, Riezlern/Kleinwalsertal*, Research in Khoisan Studies. Cologne: Rüdiger Köppe.

Michael Daniel and Edith Moravcsik. 2013. The associative plural (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

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

Emma Geniušienė. 1987. *The typology of reflexives*. Mouton de Gruyter, Berlin.

Tom Güldemann. 2008. *Quotative indexes in African Languages*. Mouton de Gruyter, Berlin.

Tom Güldemann. 2014. 'Khoisan' linguistic classification today. In Tom Güldemann and Anne-Maria Fehn, editors, *Beyond 'Khoisan': Historical relations in the Kalahari Basin*, pages 1–41. John Benjamins, Amsterdam.

Wilfrid H. G. Haacke. 2018. Khoekhoegowab (nama/damara). In Tomasz Kamusella and Finex Ndhlovu, editors, *The social and political history of Southern Africa's languages*, pages 133–158. Palgrave Macmillan, London.

Wilfrid H. G. Haacke and Eliphas Eiseb. 2002. *A Khoekhoegowab dictionary with an English-Khoekhoegowab index*. Gamsberg Macmillan, Windhoek.

Roy Stephen Hagman. 1977. *Nama Hottentot Grammar*. Indiana University publications. Research Center for Language and Semiotic Studies, Indiana University.

The Language Archive Max Planck Institute for Psycholinguistics. 2024. ELAN. Computer program.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Namibia Statistics Agency. 2011. Namibia household income & expenditure survey 2009/2010. Technical report, Namibia Statistics Agency, Windhoek.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

David A. Peterson. 2006. *Applicative constructions*. Oxford University Press, Oxford.

Paul Portner. 2018. *Mood*. Oxford University Press, Oxford.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino,

Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Johan van der Auwera, Ludo Lejeune, and Valentin Goussev. 2013. The prohibitive (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Marine Vuillermet, Eva Schultze-Berndt, and Martina Faller. 2025+. Introduction. In *Apprehensional constructions in a cross-linguistic perspective*. Language Science Press, Berlin.

Alena Witzlack-Makarevich and Sylvanus Job. 2024. Bivalent patterns in Khoekhoe. In *BivalTyp: Typological database of bivalent verbs and their encoding frames*.

Alena Witzlack-Makarevich and Hirosi Nakagawa. 2019. Linguistic features and typologies in languages commonly referred to as 'Khoisan'. In Ekkehard Wolff, editor, *The Cambridge handbook of African linguistics*, pages 382–416. Cambridge University Press, Cambridge.

Fernando Zuniga and Denis Creissels, editors. 2024. *Applicative Constructions in the World's Languages*. De Gruyter Mouton, Berlin.

Fernando Zúñiga and Seppo Kittilä. 2019. *Grammatical Voice*. Cambridge University Press, Cambridge.