The first Universal Dependency Treebank for Tswana: Tswana-Popapolelo

Tanja Gaustad[†], **Ansu Berg**[‡], **Rigardt Pretorius**[‡], **Roald Eiselen**[†]
[†]Centre for Text Technology (CTexT), [‡]Setswana
North-West University, Potchefstroom, South Africa
{FirstName.LastName}@nwu.ac.za

Abstract

This paper presents the first publicly available UD treebank for Tswana, *Tswana-Popapolelo*. The data used consists of the 20 Cairo CICLing sentences translated to Tswana. After preprocessing these sentences with detailed POS (XPOS) and converting them to universal POS (UPOS), we proceeded to annotate the data with dependency relations, documenting decisions for the language specific constructions. Linguistic issues encountered are described in detail as this is the first application of the UD framework to produce a dependency treebank for the Bantu language family in general and for Tswana specifically.

Keywords: Dependency treebank, annotation, Bantu languages, Tswana

1. Introduction

Along with a recent push to broaden the linguistic diversity in Natural Language Processing (NLP) research (Joshi et al., 2020), there has been an increased interest in syntactic annotations for under-resourced languages from Sub-Saharan Africa resulting in treebanks for Bambara (Aplonova and Tyers, 2017; Dione, 2021), Beja (Kahane et al., 2021), Wolof, and Yoruba (Dione, 2021). The only such resource currently available for Tswana is a treebank based on Lexical Functional Grammar (Berg, 2018) consisting of phrases and simple sentences (LR Berg, 2018).

This paper describes the first publicly available Tswana[1] treebank *Tswana-Popapolelo* annotated in the Universal Dependency (UD) framework (de Marneffe et al., 2021). As a proof of concept, we chose to annotate a small data set in UD as well as document linguistic annotation issues and decisions when applying UD to Tswana so that going forward more data can be annotated more easily.

In this paper, we will focus on the building of a UD treebank for Tswana (see section 2 for the necessary background information), describing the data (section 3) and preprocessing (section 4), the annotation process (section 5) and, most importantly, issues we encountered (section 6) when trying to apply the UD framework to a novel language (family).

[1]We will be using Tswana as this is the preferred term in an international setting rather than Setswana which is generally used in South Africa (as outlined in the South African Constitution of 1996 and the Use of Official Languages Act 12 of 2012). The same decision to not use prefixes applies to the names of the other official South African languages in this article.

2. Background

2.1. Linguistic Background

Tswana (ISO-639-3 tsn) is a Bantu language spoken in the north western parts of South Africa, the eastern parts of Namibia which border on Botswana and in Botswana, where it is the national language and most of the people are first language speakers. It is one of the 12 official languages of South Africa and is spoken by 8,3% of the population (Statistics South Africa, 2023), making it the 6th most frequent home language. Next to sign language and two Germanic languages (Afrikaans and English), the official languages comprise nine Bantu languages: Four Nguni languages (Ndebele, Xhosa, Zulu, and Swati), three Sotho languages (Northern Sotho, Southern Sotho, Tswana), as well as Venḓa, and Tsonga. Tswana is classified in the South-Eastern Zone of Bantu languages. These Bantu languages are divided in language groups and Tswana is included in the Sotho language group (group S31) (Maho, 2003).

Bantu languages have a number of linguistic characteristics that make them substantially different from most Indo-European languages (van der Velde et al., 2022): all of them are tone languages; they use an elaborate system of noun classes (Katamba, 2003) and their nominal and verbal morphology is highly agglutinative and very productive (Katamba, 1993). Especially the last two characteristics are important in the context of syntactic annotation.

The selection of the orthographic writing style adopted for Tswana was influenced by historical and phonological reasons (Taljard and Bosch, 2006, 433). Phonologically, the strong homographic character of the verbal prefixes of the Sotho languages

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 55–65
25 May, 2024.

(including Tswana) has led to the adoption of a disjunctive orthography regarding verbal prefixes. Nguni languages, on the other hand, have adopted a conjunctive writing system (Louwrens and Poulos, 2006). In this context, a distinction is made between *linguistic* words and *orthographic* words. For languages like English or Afrikaans, a linguistic word and an orthographic word largely coincide. For the conjunctively written languages, one orthographic word corresponds to one or more linguistic words. For disjunctively written languages like Tswana, however, several orthographic words can correspond to one linguistic word. The following example illustrates the disjunctive (Tswana) versus conjunctive (Zulu) writing styles:

Tswana	*ke a mo rata*			
	ke	a	mo	rata
	I	[pres]	him/her	love
	'I love him/her'			
Zulu	*ngiyamthanda*			
	ngi-	-ya-	-m-	-thanda
	I	[pres]	him/her	love
	'I love him/her'			

The implications of the disjunctively written verbal prefixes in the syntactic description of Tswana will be discussed in more detail in section 6.

2.2. Universal Dependencies (UD)

Universal dependencies (UD) is an international, collaborative project with two main aims: to develop a common framework describing the grammatical structure of the world's languages (de Marneffe et al., 2021) and to create treebanks for various languages applying this framework (Nivre et al., 2020). The project strives to produce cross-linguistically consistent treebanks (with language-specific extensions where necessary) describing syntactic structures as well as morphological features. This framework allows for comparisons between languages (including languages with free word order), research from a language typology perspective as well as the development of multilingual parsers. As stated in the introduction to the UD project: "The annotation scheme is based on an evolution of (universal) Stanford dependencies (de Marneffe et al., 2021), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tag sets (Zeman, 2008)."[2] The syntactic relations in UD are represented as dependency trees rather than phrase structure trees which makes the annotated data easier to use and interpret in downstream tasks. Currently, there are over 250 treebanks in more than 140 different languages available[3].

Our choice of using UD stems from the intended use of UD treebanks that benefits both the computational and linguistic research communities. As de Marneffe et al. (2021) point out, UD needs to comply with a number of (competing) criteria which include a) linguistic requirements, such as achieving a satisfactory level of annotation for linguistic analysis of individual languages and being good for highlighting structural similarities across related languages, b) computational needs, i.e. being suitable for parsing with high accuracy and supporting downstream natural language processing tasks, and, last but not least, c) pragmatic requisites, namely being suitable for rapid, consistent annotation by a human annotator and easily comprehensible by non-linguist users.

An added benefit of joining an endeavour like UD is the available infrastructure in terms of how-tos on contributing, validation scripts, support from the UD community and the visibility, availability and re-usability of the annotated data through official releases.

3. Data: Cairo CICLing Corpus

For a UD treebank to be included in an official release, it has to contain at least 20 sentences and 100 words. This can most easily be achieved by translating the 20 example sentences in the Cairo CICLing Corpus[4] to the desired language, in our case Tswana. Using these sentences has the added advantage that they contain different linguistic constructions, making it a good first test case for discussing how to annotate these constructions in the targeted language.

After procuring the 20 Cairo CICLing sentences, three Tswana native speakers[5] translated all the data without consulting each other. In a second step, the Tswana team decided on a consensus translation where consensus on the final translated sentences was attained after considering free and word-for-word translations. This was then the input to the preprocessing described next.

4. Preprocessing: Tokenisation, XPOS and UPOS

The tokenisation for the 20 translated sentences corresponds to orthographic words as used in the official orthography of Tswana (Cole, 1955; Krüger,

[2] https://universaldependencies.org/introduction.html, retrieved 12-02-2024.

[3] See https://universaldependencies.org/ for an overview.

[4] https://github.com/UniversalDependencies/cairo/blob/master/translations.txt

[5] These were graduate students paid for their time.

PRON	43	20%	ADV	12	6%
VERB	34	16%	AUX	12	6%
PART	32	15%	CCONJ	8	4%
NOUN	26	12%	SCONJ	5	2%
PUNCT	23	11%	ADJ	4	2%
PROPN	15	7%			
Total tokens: 214			Type-Token ratio: 0,47		

Table 1: Overview of UPOS tags assigned in the 20 Tswana sentences.

nsubj	25	12%	nmod	10	5%
punct	23	11%	aux	8	4%
root	20	9%	obj	8	4%
case	16	7%	compound	6	3%
expl	16	7%	fixed	6	3%
mark	11	5%	ccomp	5	2%
advmod	10	5%	xcomp	5	2%
cc	10	5%	obl	4	2%
conj	10	5%	others	21	10%

Table 2: Overview of dependency relations assigned in the 20 Tswana sentences.

2006). We will discuss ramifications and potential different choices in more detail in section 6.

An important premise when annotating universal dependencies is the presence of parts-of-speech (POS), more specifically universal POS (UPOS) (Petrov et al., 2012). The UPOS tag set contains 17 tags: 6 for open classes (nouns, verbs, etc.), 8 for closed classes (e.g. pronouns, conjunctions) and 3 for miscellaneous items (such as punctuation and symbols).

In the very limited work done on Tswana, there is not yet consensus on how to accommodate traditional Tswana POS in UPOS[6] and the application of UPOS tags is not always straightforward (Dione et al., 2023). However, there are Tswana POS taggers with more extensive tag sets (Eiselen and Puttkammer, 2014; Puttkammer et al., 2018; Malema et al., 2020; Dibitso et al., 2022). For the purposes of the Tswana UD annotations, the NCHLT tokeniser and POS tagger were used to annotate the data with detailed POS (typically referred to as XPOS in UD). This tagger includes 26 main tags, and 188 tags when including class information. The detailed POS tags were subsequently converted to UPOS tags based on a conversion table. Table 1 provides an overview of the distribution of the assigned UPOS tags in the data.

Even with the seemingly simple task of reducing the XPOS tag set to UPOS, there were a few difficult decisions during the conversion, especially to not overload one particular UPOS tag (mostly PART particle) with too many distinct XPOS categories.

The main problem concerned verbal prefixes in Tswana. As mentioned earlier, due to the disjunctive writing style, several classes of verbal morphemes are written as separate words. These morphemes are also separately tagged in the XPOS schema and include concordial morphemes (subject and object morphemes), possessive, negative, aspectual, and tense morphemes. Subject and object morphemes were treated as subject and object concords respectively in XPOS. With no direct equivalents available in UPOS, they were tagged as PRON (pronoun), while possessive concords are tagged as PART (particle). Tense and aspect morphemes (assigned the tag MORPH in XPOS) were also converted to PART (particle) in UPOS, while the negative markers were converted to ADV.

Additionally, in our detailed XPOS tag set, ideophones received their own tag IDEO as they are considered a separate word class expressing an action, manner or property through sound imitation, but not always exhibiting the same syntactic function in a sentence. However, there is no such tag in UPOS and no equivalence in other language families in the UD catalogue was found. As there were no ideophones in the Tswana Cairo CICLing sentences, the tag was not needed, but in further annotations we would consider the ADV (adverb) tag for ideophones in Tswana.

The advantage of having both XPOS and UPOS tags at our disposal during syntactic annotation is that highly ambiguous tokens (homophonic and morphosyntactically ambiguous, e.g. *ka* 'with, on, through', *go* 'copulative verb in different moods, at, to+verb, to+location') can more easily be linked correctly and that both the automatically assigned XPOS tags as well as the converted UPOS tags could be corrected if needed.

5. Annotation Process

The syntactic annotation as well as corrections to the UPOS tags was done in Arborator Grew (Guibon et al., 2020)[7]. In a first step, the annotators checked the UPOS and XPOS tags and corrected them where necessary. Then, the syntactic structure was added incrementally: start by identifying the root for each sentence, link the subject and object(s), then proceed to link and label the remaining dependencies. During the annotation, vari-

[6] Tswana has been included in the UDMorph Tagger https://lindat.mff.cuni.cz/services/teitok-live/udmorph/index.php?action=tag, but the conversion from the detailed POS tag set to UPOS has not been checked by linguists (yet).

[7] https://arboratorgrew.elizia.net/

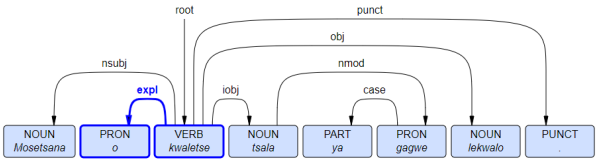

Figure 1: Sentence 1 with an overt subject.

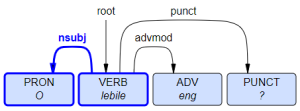

Figure 2: Sentence 2 with a covert subject.

ous options for certain syntactic constructions were discussed and documented until the annotators agreed on the preferred linking and associated dependency labels. Table 2 shows an overview of the assigned dependency relations.

The *Tswana-Popapolelo* treebank will be available in the next UD release[8], along with all language specific documentation.

6. Issues Encountered

Some of the annotations were straightforward for the current small data set, but specific linguistic idiosyncrasies of Tswana as an example of a disjunctively written Bantu language required more in-depth discussions on how to use existing UD relations. These are detailed below.

6.1. Disjunctive Orthography and Verbal Constructions in Tswana

As has been described in section 2.1, a disjunctive writing style has been adopted for Tswana. Especially for verbal prefixes, this means that a large number of orthographic tokens preceding the verb would in traditional linguistics be seen more as morphemes rather than "proper" words. The proper identification of words is generally taken as an imperative preprocessing step for syntactic description. In this regard, the disjunctively written verbal prefixes of Tswana cause orthographic tokens which are part of linguistic words (Taljard and Bosch, 2006) and therefore compromise the lexical integrity of verbs. Tswana verbal prefixes carry inflectional information while suffixes carry inflectional as well as derivational information (Krüger, 2006, 268). These verbal prefixes also carry both morphological and syntactic information which makes it difficult to assign UPOS tags and syntactic relations to them.

One obvious solution to address this problem is to adjust the tokenisation to reflect linguistic words, rather than orthographic words. Although this would certainly simplify and more closely align the data with the UD framework, this would also reduce the granularity and informativeness of the treebank. With this in mind, we opted to annotate the relations between all orthographic words. More details on the implication of tokenisation is provided in section 6.5.

In the UD annotation of Tswana, the disjunctively written verbal prefixes are linked to the verbal root via arcs. We will now describe how the different parts of verbal constructions have been handled.

6.1.1. Subject Concords

In instances where an overt subject is realised in a sentence, the subject concord is an agreement marker which marks the relation between the overt subject and the verb. In these cases we opted for the `expl` relation. This relation is used in UD for phenomena such as clitic doubling (e.g. in Romance languages) or the doubling of a lexical nominal and a pronominal clitic (e.g. in Greek and Bulgarian). Even though subject concords are not the same as clitics, they behave in a similar fashion in that they are a type of "pronominal" copy without its own semantic role. An example for Tswana can be seen in figure 1 of sentence 1.[9]

(1) Mosetsana **o** kwaletse
girl she[SubjConc] write[appl-perf]
tsala ya gagwe lekwalo
friend of her letter
'The girl wrote a letter to her friend.'

In instances where the overt subject is not realised, the (covert) subject concord acquires a

[8] https://github.com/UniversalDependencies/UD_Tswana-Popapolelo

[9] All figures were produced with http://www.let.rug.nl/kleiweg/conllu/.

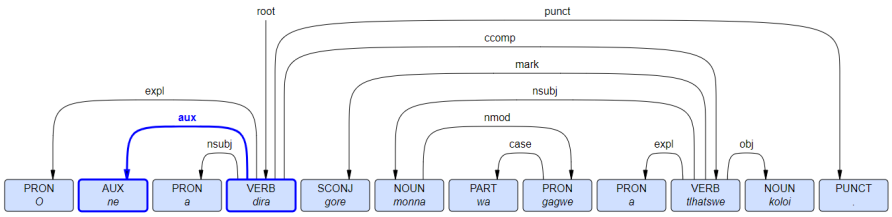

Figure 3: Sentence 3 showcasing the use of the `aux` relation.

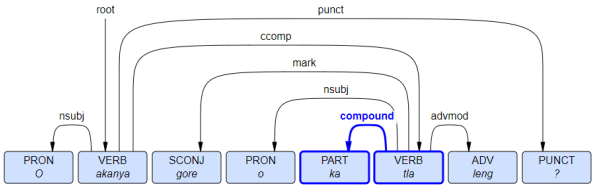

Figure 4: Sentence 4 showcasing the use of the `compound` relation for certain TAME morphemes.

pronominal status (as it would be a pronoun in translation) and becomes the actual subject, hence is annotated with the `nsubj` relation as shown in figure 2 of sentence 2.

(2) **O** lebile eng
you[SubjConc] see[perf] what
'What are you looking at?'

6.1.2. Auxiliary Verbs

An auxiliary verb enriches the meaning of the complementary main verb, copulative verb or another auxiliary verb phrase and can add semantic information regarding the mood, tense, aspect and/or polarity of a verb. It also adds information on the progression or completion of an action: It expresses a certain type of duration of the action or it expresses the logical time at which the action is executed. For example, the auxiliary verb *ne* expresses a relative past tense indicating that the action was taking place or had taken place at some point in the past. If the complementary verb is in the present tense then it indicates an action that is incomplete and continuing at a certain moment in the past. If the complementary verb is in the perfect it indicates that the action had been completed at the point of reference (Pretorius, 1997; Krüger, 2013a).

In UD, auxiliary verbs are a closed class that cannot have any children. The `aux` relation is used in Tswana to indicate the relation between a verb and the preceding auxiliary verb, as with other languages. However, we encountered the issue of auxiliaries taking a (doubled) subject concord as a dependent. In sentence 3, the subject concord occurs twice: once (realised as *o*) with the auxiliary *ne* and once (realised as *a*) with the verb *dira*[10], but both referring to the subject 'she' and both needed for the sentence to be grammatical.

(3) **O** **ne**
she[SubjConc] aux[past-indef
a dira gore monna wa
she[SubjConc] make that husband of
gagwe a tlhatswe koloi
her he[SubjConc] wash[pass] car
'She made her husband wash the car.'

At this stage, we have chosen to annotate the subject concord with the auxiliary verb with a `expl` relation, while the subject concord with the main verb becomes the `nsubj` and the relation between the auxiliary and the main verb is tagged `aux` as can be seen in figure 3.

6.1.3. TAME Morphemes

The disjunctively written Tense-Aspect-Mood-Evidentiality (TAME) morphemes in the morphological structure of a verb always occur in a fixed order and are not morphosyntactically flexible. For the TAME morphemes including the present tense morpheme *a*, the progressive morpheme *sa* 'still', the potential morpheme *ka* 'can, may' and the future tense morpheme *tla* 'will, shall' the `compound` relation is applied to express that this is a combination of lexemes that morphosyntactically behave as single words. See the example in sentence 4 and figure 4.

(4) O akanya gore
you[SubjConc] think that
o ka **tla** leng
you[SubjConc] can come when
'When do you think you can come?'

[10]The meaning of the auxiliary verb *ne* requires the consecutive form of the subject agreement morpheme following it.

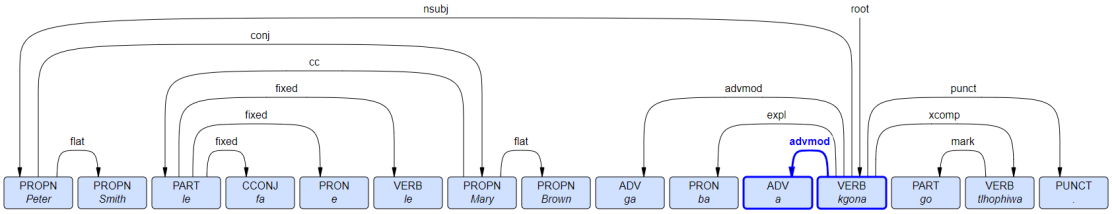

Figure 5: Sentence 5 showcasing the use of the `advmod` relation for negation TAME morphemes.

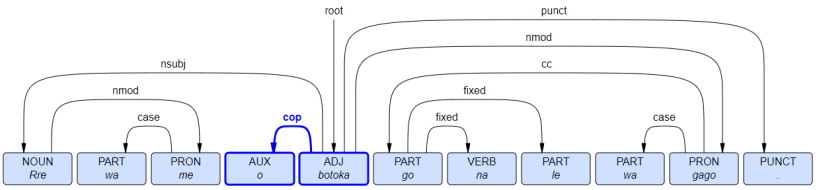

Figure 6: Sentence 6 with a describing copulative verb.

For the negative morphemes *ga*, *sa*, *se*, we have opted to use the combination of ADV and `advmod`, parallel to English, as illustrated in sentence 5 and figure 5.

(5) Peter Smith le fa e le Mary Brown **ga**
Peter Smith neither nor Mary Brown not
ba **a** kgona go
they[SubjConc] not can to[InfMarker]
tlhophiwa
select[pass]
'Neither Peter Smith nor Mary Brown could be selected.'

6.2. Copulatives

A copula is the relation of a function word used to link a subject to a nonverbal predicate. In Tswana three types of copulative verbs are distinguished:

- identifying copulative: identifies a subject with regards to type, status or profession or to predicate the existence or presence of a thing, e.g. *Lekwalo lê ke la gago* 'This letter is yours';
- describing copulative: establishes some quality, characteristic or state of a subject, or its situation or locality, e.g. *Ditlhako tsa me di dintšha* 'My shoes are new';
- associative copulative: expresses the idea of the English have or be with and indicates possession or association, e.g. *Sediba sê se na le metsi* 'This well has water'.

The morphological structure of these verbs may include tense, aspect, mood and polarity information.

When the verb in Tswana is an identifying or describing copulative verb, the root of the clause is the complement of the copulative verb. These two types of copulative verbs are POS tagged as AUX, and the `cop` relation is used between the root and the preceding copulative verb. See sentence 6 and figure 6 for an example of a describing copulative.

(6) Rre wa me **o** **botoka** go na le wa
father of me is[cop] cooler than of
gago
you
'My dad is cooler than yours.'

In the case of an associative copulative verb, the root of the clause is the copulative verb. The associative copulative verbs in Tswana are POS tagged as VERB, and the `obj` relation is used between the root and the complement that follows it, as showcased in sentence 7 and figure 7. This analysis differs from traditional Tswana linguistic descriptions (Cole, 1955; Krüger, 2006, 2013b).

(7) Ga ba **na** **kakanyo** epe
not they[SubjConc] have idea none
gore e kwadilwe ke
that it[SubjConc] write[perf-pass] by
mang.
who
'They have no idea who wrote it.'

6.3. Use of the `mark` Relation

Conjunctions that mark a clause as subordinate to another clause are annotated as `mark` in UD. In Tswana, the marker is an introductory member of a clause that includes an action in the subjunctive or participial mood. For the subjunctive, the conjunction *gore* 'that, so that' is used, as shown in sentence 8 and figure 8. For the participial, a conjunction such as *fa* 'as, while, when, if', *le fa* 'even if, although, while' and *ka* 'since' are used.

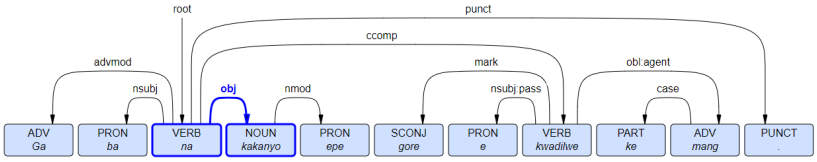

Figure 7: Sentence 7 with an associative copulative verb.

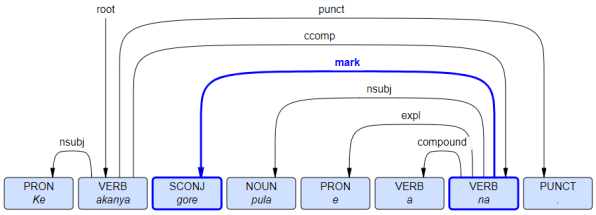

Figure 8: Sentence 8 showcasing the use of the `mark` relation for a subordinate clause.

(8) Ke akanya **gore** pula
I[SubjConc] think that rain
e a na
it[SubjConc] [pres] falls
'I think that it is raining.'

The `mark` relation is also used in Tswana for infinitive verbs, analogue to English and German, for example *go tlogela* 'to stop' where the marker is *go* 'to'. See sentence 9 and figure 9 for an illustration.

(9) O ne
he[SubjConc] aux[past-indef]
a leka **go** **tlogela**
he[SubjConc] try to[InfMarker] quit
go goga le go
to[InfMarker] smoke and to[InfMarker]
nwa
drink
'He tried to stop smoking and drinking.'

Furthermore, `mark` is used in a relative clause where the qualificative particle is the marker as seen in sentence 10 and figure 10: The qualificative particle always agrees with a specific noun class in Tswana, for example in *e kgolo* '[part] big' the marker is the qualificative particle *e* that indicates noun class 9 agreement.

(10) A Iguazu ke naga **e**
[InterPart] Iguazu is[cop] country [part]
kgolo kgotsa ke **e** nnye
big or is[cop] [part] small
'Is Iguazu a big or a small country?'

6.4. Interrogative Particle *a*

In Tswana there is an interrogative particle *a* added at the beginning of a sentence to change an indicative sentence to an interrogative one. After consultation with the UD community, we have decided to assign the UPOS tag PART (particle) to *a* as well as link it directly to the root of the sentence with a `discourse` relation (following the Latin example of *ne*). As this particle works on a more pragmatic level, the `discourse` relation "used for interjections and other discourse particles and elements (which are not clearly linked to the structure of the sentence, except in an expressive way)" as described in the UD overview of relations[11] seemed the best choice. Sentence 11 and figure 11 show an example of this for Tswana.

(11) **A** o batla
[InterPart] you[SubjConc] want
go tsamaya?
to[InfMarker] leave/go
Do you want to leave/go?

6.5. Tokenisation in Tswana

An issue that we definitely have not solved yet and that is connected to the previous section 6.1 is the tokenisation of Tswana. Traditionally, computational analyses for disjunctively written South African Bantu languages, i.e. Northern Sotho, Southern Sotho, Tswana, Venda and Tsonga, have been done on orthographic words as then no conversions are needed from the original text. The implications of choosing to use orthographic words rather than linguistic words, however, will be felt at various levels when working on the syntactic analysis of Tswana applying UD dependencies:

- When doing annotation: Working on the orthographic word means more time and effort will be spent on getting the UPOS as well as

[11] https://universaldependencies.org/u/dep/all.html#al-u-dep/discourse.

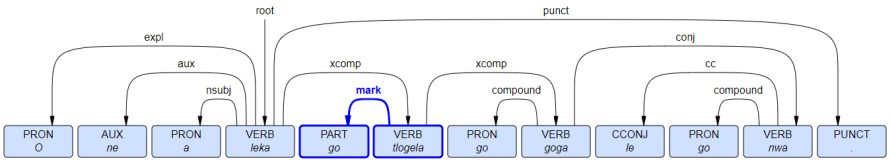

Figure 9: Sentence 9 showcasing the use of the `mark` relation in infinitives.

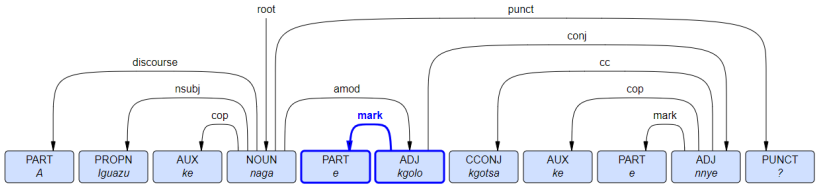

Figure 10: Sentence 10 showcasing the use of the `mark` relation for qualificative particles.

the dependency relations right (both pertaining more to the syntactic level). With linguistic words, the syntactic structure becomes more straight forward (simpler?), but at the same time more care needs to be given to adding morphological information to retain the necessary detail to be able to disambiguate.

- From a computational linguistics view point: Specifically in the UD framework, if Tswana text is analysed using orthographic words, the resulting annotations make it more directly comparable with European languages and the syntactic annotations will be more diverse and informative. On the other hand, using linguistic words will put more emphasis on the similarities with other, especially conjunctively written Bantu languages, but the syntactic structure will be simpler as a lot of information will only be contained on the morphological (sub-word) level.

- In relation to linguistic analyses: In traditional (structural) grammatical descriptions, the left hand boundary of Tswana verbs is considered to be the first prefix of such a verb, even if it is written disjunctively. This implies that verbs such as *ke a mo rata* in 2.1 would be tokenised as one word, namely a verb. This verb would constitute a sentence in itself and would be the predicate of the sentence. The syntactic analysis of the sentence would thus not indicate the pronominal value of the subject and object concords so as to indicate that the sentence contains a subject and object (Taljard and Bosch, 2006; Louwrens and Poulos, 2006; Krüger, 2006; Cole, 1955; Pretorius et al., 2015). In later descriptions (Berg, 2018), the lexical integrity of the verb is maintained but the argument status of these concords is indicated on the functional level.

So, if we were to decide to "attach" verbal prefixes to the verb, the original structure in 12 based on orthographic words would change to the representation in 13.

(12) **Ga ke a kgona**
not I[SubjConc] [pres] able
go tshwarelela ka gore
to[infMarker] keep up because
o ne
he[SubjConc] aux[past-indef]
a taboga ka lebelo thata
he[SubjConc] run with speed much

(13) **[Ga ke a kgona]** go tshwarelela ka gore
[I wasn't able] to keep up because
[o ne] **[a taboga]** ka lebelo thata
[he aux] [he ran] with speed much
'I wasn't able to keep up, because he ran too fast.'

We feel more work is needed to explore where to draw the boundaries when "attaching" verbal prefixes as well as to fully understand the consequences of such an approach.

7. Conclusion and Future Work

This paper contains the description of the first publicly available UD treebank for Tswana, based on the 20 translated Cairo CICLing sentences. The resulting treebank shows that this was a successful first endeavour to apply UD to a Bantu language and forms the basis for further annotation of Tswana to a more extensive set. The main benefit of starting with such a small data set is that many of the most problematic annotations can be discussed in detail, and the corresponding outcomes can be documented without needing a substantial reannotation of the data at a later stage. As would be expected, not all issues have been resolved yet and

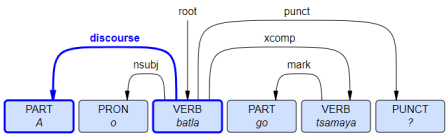

Figure 11: Sentence 11 with an interrogative particle.

some decisions had to be made on how to best apply the existing framework to a novel language with unique linguistic characteristics. We hope the detailed report on the issues encountered will also help others when annotating new Sotho and Bantu languages in UD.

With the *Tswana-Popapolelo* treebank now available, we plan to annotate extra data with the help of student assistants. The current annotations are based on our understanding of the literature and feedback we received from the UD community, but the choices made thus far will definitely be further refined and the available annotated data for Tswana will be expanded by adding it to *Tswana-Popapolelo*. This includes experimenting with different tokenisation strategies for the same data to study the repercussions on the dependency analyses.

Once a larger set of treebank data is available, we will also train automatic parsers to pre-annotate data to assist and simplify the annotation process. Ultimately we aim to have enough data to train accurate full dependecy parsers, including XPOS, UPOS, lemma and morphological taggers, while at the same time leveraging the work of others that use UD treebanks to train various NLP tools.

8. Acknowledgements

We would like to thank Kevin Mavalela and Kaboentle Maibi for their contribution to the translations and initial discussions and annotations. Furthermore, we are grateful to the UD community for their responses to our queries.

9. Bibliographical References

Ekaterina Aplonova and Francis M. Tyers. 2017. Towards a dependency-annotated treebank for Bambara. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 138–145, Prague, Czech Republic.

Ansu Berg. 2018. *A computational syntactic analysis of Setswana*. Ph.D. thesis, North-West University, Potchefstroom, South Africa.

Desmond T. Cole. 1955. *An introduction to Tswana grammar*. Longman, Cape Town.

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

Mary Dibitso, Pius A. Owolawi, and Sunday O. Ojo. 2022. An hybrid part of speech tagger for Setswana language using a voting method. In *International Conference on Intelligent and Innovative Computing Applications*, pages 245–253.

Cheikh M. Bamba Dione. 2021. Multilingual dependency parsing for low-resource African languages: Case studies on Bambara, Wolof, and Yoruba. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 84–92, Online. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba O. Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, pages 10883–10900.

Roald Eiselen and Martin J. Puttkammer. 2014. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3698–3703.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.

Francis Katamba. 1993. *Morphology*. Springer, New York.

Francis Katamba. 2003. Bantu nominal morphology. In Derek Nurse and Gérard Philippson, editors, *The Bantu Languages*, pages 103–120. Routledge, London & New York.

Caspar J.H. Krüger. 2006. *Introduction to the morphology of Setswana*. Lincom Europe, München.

Caspar J.H. Krüger. 2013a. *Setswana syntax: a survey of word group structures: Volume 1*. Lincom Europe, München.

Caspar J.H. Krüger. 2013b. *Setswana syntax: a survey of word group structures: Volume 2*. Lincom Europe, München.

Louis J. Louwrens and George Poulos. 2006. The status of the word in selected conventional writing systems - the case of disjunctive writing. *Southern African Linguistics and Applied Language Studies*, 24(3):389–401.

Jouni Maho. 2003. A classification of the Bantu languages: an update of Guthrie's referential system. In Derek Nurse and Gérard Philippson, editors, *The Bantu Languages*, pages 639–651. Routledge, London & New York.

Gabofetswe Malema, Boago Okgetheng, Bopaki Tebalo, Moffat Motlhanka, and Goaletsa Rammidi. 2020. Complex Setswana parts of speech tagging. In *Proceedings of the first workshop on Resources for African Indigenous Languages (RAIL)*, pages 21–24.

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 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Laurette Pretorius, Biffie Viljoen, Ansu Berg, and Rigardt Pretorius. 2015. Tswana finite state tokenisation. *Language Resources and Evaluation*, 49(4):831–856.

Rigardt Pretorius. 1997. *Auxiliary Verbs as a Subcategory of the Verb in Tswana*. Ph.D. thesis, PU for CHE, Potchefstroom, South Africa.

Martin Puttkammer, Roald Eiselen, Justin Hocking, and Frederik Koen. 2018. NLP web services for resource-scarce languages. In *Proceedings of ACL 2018, System Demonstrations*, pages 43–49, Melbourne, Australia. Association for Computational Linguistics.

Department of Statistics South Africa. 2023. Census 2022: Statistical release. Technical report, Department of Statistics, Republic of South Africa.

Elsabé Taljard and Sonja E. Bosch. 2006. A comparison of approaches to word classtagging: Disjunctively vs. conjunctively written Bantu languages. *Nordic Journal of African Studies*, 15(4):428–442.

Mark van der Velde, Koen Bostoen, Derek Nurse, and Gérard Philippson, editors. 2022. *The Bantu Languages*, 2nd edition. Routledge, London & New York.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

10. Language Resource References

LR Berg, Ansu. 2018. *Setswana Test suite and Treebank*. South African Centre for Digital Language Resources (SADiLaR). PID https://hdl.handle.net/20.500.12185/478.