Word Segmentation for UD: a Comparison of IsiZulu and Sepedi

Laurette Marais[1], Laurette Pretorius[2]
[1]NLP Research Group, CSIR, Pretoria, South Africa
[2]Division of Computer Science, Mathematical Sciences, Stellenbosch University, South Africa
[1]laurette@loerie.org.za, [2]lpretorius@sun.ac.za

Abstract

The Southern Bantu language family contains languages with so-called conjunctive orthographies and disjunctive orthographies. In languages with conjunctive orthographies, such as isiZulu, orthographic words correspond to linguistic words, whereas in languages with disjunctive orthographies, prefix morphemes of verbs and other predicates are written as disjunct, orthographic words. When developing Universal Dependencies treebanks, the basic principle is to consider syntactic (linguistic) words, but for languages with agglutinating morphology, it has been argued that this reduces the informativeness of the treebank. In this paper we investigate this claim by analysing and measuring the effects of annotating universal dependencies on the basis of orthographic words on two morphosyntactically parallel treebanks for isiZulu and Sepedi.

Keywords: disjunctive orthography, conjunctive orthography, token, syntactic word, annotation, UD treebank

1. Introduction

The question of *wordhood* (What is a word?) is central to describing and annotating the grammatical structure (morphology and syntax) of any language, including Southern Bantu languages[1] such as Sepedi and isiZulu. By distinguishing between an *orthographic word*[2] and a *linguistic word*[3], the two writing systems prevalent in the Southern Bantu languages may be described as *disjunctive*, where a single linguistic word may be represented by a number of orthographic words, and *conjunctive*, where a single linguistic word is represented by a single orthographic word (Kosch, 2006, pp.2-4).

Sepedi has a disjunctive orthography and isiZulu a conjunctive one. Moreover, these languages exhibit a rich agglutinative morphology and are closely related in terms of grammatical structure. We illustrate this by means of the example in Table 1:

In the context of the Universal Dependencies[4] (UD) framework, an orthographic word is usually referred to as a *token* and a linguistic word as a *syntactic word*.

In principle, UD annotation is based on a lexicalist view of syntax, which means that dependency relations hold between words and the basic units of annotation are syntactic words. However, the example in Table 1 raises the question of marking boundaries of syntactic words in UD. It is noted[5] that "[t]here are also situations where multiple surface tokens act as one (morpho)syntactic word. The standard solution is that these tokens are kept as independent nodes in the dependency structure and they are connected using specialized relations (`compound`, `flat`, `fixed`, `goeswith`)".

Language	Sentence POS	# Words (orth,ling)
English	I see them PRON VERB PRON	3,3
isiZulu	*ngiyababona* VERB	1,1
Sepedi	*ke a ba bona* VERB	4,1

Table 1: Orthographic words vs. linguistic words.

However, in the case of Sepedi, these specialized relations are less reflective of the morphosyntactic structure of multi-token words such as in Table 1. For example, *ke* (I) and *ba* (them) play the role of the subject and object of the sentence, and is therefore more accurately annotated with the syntactic relations `nsubj` and `obj`, respectively.

It remains an open question which annotation strategy in UD is best for morphologically rich languages (see also (Goldman et al., 2025)). The Southern Bantu languages provide a unique angle from which to investigate this question due to their divergent orthographies while being highly similar

[1]These languages include the Nguni group of languages, viz. isiZulu, isiXhosa, Siswati and isiNdebele, and the Sotho-Tswana group, viz. Sepedi, Setswana and Southern Sotho.

[2]Defined by Kosch (2006) as "a sound [grapheme] or a sequence of sounds seperated from other sounds or sequences of sounds by means of spaces".

[3]Defined by Kosch (2006) as "a unit which has its own independent meaning".

[4]https://universaldependencies.org/

[5]https://universaldependencies.org/workgroups/newdoc/word_segmentation.html

morphosyntactically.

Gaustad et al. (2024) annotated a Setswana corpus of 20 sentences by using syntactic relations to avoid reducing "the granularity and informativeness of the treebank". Setswana belongs to the same language group as Sepedi and they share a disjunctive orthography.

Given this decision, the question then for a language like isiZulu (also closely related but with a conjunctive orthography) is whether some kind of sub-word tokenisation would be beneficial to combat the same loss of informativeness. However, morphological surface tokenisation for isiZulu is non-trivial due to widespread morphophonological alternation (Pretorius and Bosch, 2003). How then to approach this question?

In this paper, we exploit the high structural similarity between Sepedi and isiZulu as well as their respective, standardised orthographies to investigate and to some degree quantify the effect of tokenising at the sub-syntactic-word level.

Since neither language is yet represented within the UD project, we develop and utilise two parallel treebanks of 94 sentences each as the basis for our experiment, annotated at the token level.

In Section 2 we introduce our parallel treebank and describe how it was designed to provide a truly parallel setup via the Grammatical Framework formalism.

Sections 3 and 4 attempt to shed light on the question of informativeness of a treebank from an intrinsic and extrinsic perspective, respectively. Section 3 considers the differences between the treebanks from a linguistic and complexity point of view, while Section 4 shows how these differences affect downstream NLP applications, with our example being a comparison between UDPipe parsers trained on the two treebanks.

We provide some conclusions in Section 5 and note directions for future work.

2. Designing Parallel Treebanks

One way of approaching our question about informativeness might be to compare an isiZulu treebank annotated on orthographic words (tokens) to an isiZulu treebank annotated on some systematic subword segmentation. (Or, conversely a Sepedi treebank annotated on tokens compared to a Sepedi treebank where tokens are grouped together into linguistic words.) However, as mentioned above, morphological surface segmentation is not a trivial task for isiZulu. Hence, it is not always obvious where morphological surface boundaries should be inserted into a token. Moreover, it is not always clear what granularity of segmentation is best – should all morpheme boundaries be marked or does it make sense to consider some sequences of morphemes together, such as ignoring the morpheme boundary between the pre-prefix and basic prefix of nouns?

Instead of speculating about alternatives for a specific language, we exploit the fact that isiZulu and Sepedi are morphosyntactically highly similar, but have different orthographies. These standardised orthographies provide a natural, linguistically informed basis for considering the question. Our experiment is performed on a treebank designed to be highly parallel in terms of the morphosyntax of the sentences.

2.1. GF Resource Grammars for IsiZulu and Sepedi

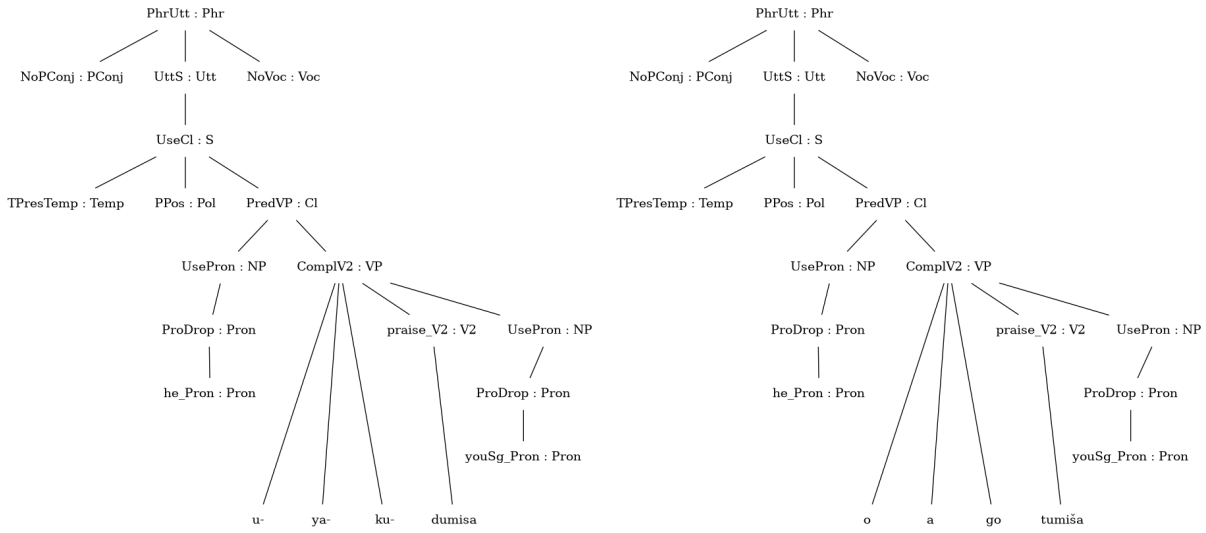
Grammatical Framework (GF) is a formalism, programming language and software ecosystem that enables multilingual grammar engineering. One of the key features of GF is its distinction between abstract syntax and concrete syntax, where an abstract syntax tree (AST) is typically a language independent representation of a sentence, with one or more corresponding concrete syntax trees (CSTs) that describe how the AST is realised, or linearised, in different natural languages.

The GF Resource Grammar Library (RGL) (Ranta, 2009) consists of an abstract syntax core of linguistic functions that attempt to model universal syntax functions of languages, such as predication, determination and modification. The concrete implementation of this for a specific language is called a GF Resource Grammar.

GF Resource Grammars (RGs) for isiZulu and Sepedi have been developed[6]. In addition to implementing concrete syntax linearisation functions for syntax functions in the existing RGL, a significant number of Bantu-language specific abstract syntax functions were defined in so-called extension modules. These functions were first developed for use by the isiZulu RG, and were refined when the Sepedi RG was developed. These two RGs therefore share an abstract syntax, and due to the close similarity between the languages, their concrete syntaxes share many similarities reflecting the shared characteristics of nominal classification and concordial agreement (Faaß et al., 2012).

For example, consider the two GF trees in Figure 1 representing the sentence 'He praises you' in both languages. The two trees not only show the identical abstract syntax tree shared by the two sentences, but also the striking similarities in how the sentences are linearised in the two languages.

[6]Available at https://github.com/LauretteM/gf-rgl



(a) isiZulu (b) Sepedi

Figure 1: GF ASTs for the sentence 'He praises you.'

2.2. A Parallel GF Treebank

Our starting point for setting up our experiment was a GF treebank of 100 isiZulu sentences. These sentences were captured from an isiZulu textbook (Taljaard and Bosch, 1988), and a large variety of morphosyntactic constructions are represented in the treebank. The original purpose of the treebank was to serve as a regression test treebank during development of the isiZulu Resource Grammar. In previous work (Marais et al., 2024), the abstract syntax trees of the treebank also served as a regression test during development of a Siswati Resource Grammar, where the trees would be linearised in Siswati using the Siswati RG, and compared to the translations of the original isiZulu sentences in a parallel Siswati textbook by the same authors.

To obtain a parallel isiZulu-Sepedi treebank, i.e. a multilingual GF treebank, we followed a similar approach by linearising the abstract syntax trees using the Sepedi RG to obtain Sepedi sentences. These generated sentences were then manually evaluated. Since isiZulu and Sepedi belong to different language groups within the Southern Bantu family, it was expected that some abstract syntax trees would not linearise to valid Sepedi sentences. Indeed, this was the case for 6 of the 100 sentences, primarily due to lexical mismatches between the languages and the fact that Sepedi only has one basic past tense, while isiZulu has two.

In any case, 94 of the abstract syntax trees originally obtained via isiZulu sentences did linearise into correct Sepedi sentences, and these 94 pairs of sentences that share an abstract syntax tree form the basis of our experiment.

The significance of the shared abstract syntax tree is that it allows us to define in which way the sentences are identical and therefore in what sense the resulting isiZulu and Sepedi treebanks are parallel. Equivalence at the level of abtract syntax trees can be considered as equivalence of the deep structure, while the high degree of similarity between the implementations of the two RGs contributes to near-equivalence of the surface structure.

2.3. Parallel UD Treebanks

Before we discuss our annotation approach, we briefly return to two aspects of Bantu grammar that are important when discussing UD annotation.

Firstly, the grammatical (morphological and syntactic) structure is based on two systems: nominal classification (the noun class system) and concordial agreement (the system of concords). In most Bantu languages, each noun belongs to one of between 15 and 18 noun classes (Marten, 2021). IsiZulu has 17 noun classes and Sepedi has 18. A concord is a structural element (agreement marker/morpheme) which formally marks the relationship between a noun and other words in a sentence, which must be observed in all parts of the utterance which are linked to the noun. Word categories such as verbs, pronouns, adjectives, relatives, possessives etc. are brought into concordial (i.e. grammatical) agreement by means of these concords[7]. It should be clear that such information is essential for useful UD annotation. Therefore, we comprehensively include noun class and agreement features in our annotations.

Secondly, pronominalisation in the languages under discussion is, in general, not a process of *sub-*

[7]Detailed expositions may be found in (Poulos and Msimang, 1998) and (Poulos and Louwrens, 1994).

stitution, as in English[8], but a process of *deletion* (Louwrens, 1991, pp.91-95). In particular, the subject and object concord are not *primarily* pronominal forms, but agreement morphemes which mark the syntactic relationship between subject and object nouns and verbs. Only when the noun is deleted, does the concord acquire the *secondary* status of being a pronominal form due to the co-referential relationship which exists between the concord and the deleted noun.

For example, *monna o a sepela* (the man walks) becomes *o a sepela* (he walks) by *deleting* the noun from the sentence. There is no substitution taking place and the subject concord *o* now acquires a pronominal function.

For this reason, we follow Gaustad et al. (2024) by annotating Sepedi subject and object concords with the UPOS `PRON`, even in sentences where they do not perform a pronominal function. Our reason for this is that the UPOS tagset offers no better alternative. The only other possible option is `PART`, which is already used for various other morphemes. Using `PRON` throughout is consistent.

The two treebanks were annotated manually by adding dependency relations and UPOS (Table 2), using as point of departure the approach proposed by Gaustad et al. (2024). We also list the morphological information and features that we employ (Table 3).

Ann. types	Labels
Relations	advmod, conj, cop, det, nmod, nsubj, obj, obl, root, vocative, xcomp, case, compound, expl, mark
Subrelations	det:emph, obl:lmod, nmod:poss, advmod:lmod, obl:tmod
UPOS	ADJ, ADV, AUX, NOUN, PRON, VERB, DET, PART

Table 2: IsiZulu and Sepedi: UD annotation labels for relations and UPOS (only Sepedi).

A closer look at feature annotation shows that in the Setswana treebank the only universal feature (and value) that was employed is `Polarity=Neg`. The only miscellaneous feature (and values) was `NounClass` and `Bantu1, ..., Bantu17`.

We decided to separate annotation of inherent agreement information (using universal features `Person` and `Number` and miscellaneous feature `NounClass`, in line with Gaustad et al. (2024)) from concordial agreement information (new miscellaneous features `Agreement` and `AgreementObj`).

[8]For example, the noun in the sentence 'The man walks.' is pronominalised to 'He walks.' by *substituting* the noun with the pronoun.

Feature	Value
Universal	
Case	Com, Loc
Mood	Imp, Sub
Number	Sing, Plur
Person	1, 2
Polarity	Neg
Tense	Past, Pres
VerbForm	Inf
Miscellaneous	
Agreement	1ps, 2ps, 1pp, 2pp Bantu1, ..., Bantu17, Bantu18
AgreementObj	1ps, 2ps, 1pp, 2pp Bantu1, ..., Bantu17, Bantu18
NounClass	Bantu1, ..., Bantu17, Bantu18

Table 3: IsiZulu and Sepedi: UD annotation labels for features and their values (only Sepedi).

Examples showing the reasoning for this are given in Appendix 8.1.

In terms of dependency relations and UPOS, the annotation for the Sepedi and isiZulu treebanks followed as closely as possible the approach of the Setswana treebank (Gaustad et al., 2024), as summarised in Appendix 8.2. However, regarding morphological features, our annotation was more extensive, as shown in Table 3[9].

3. Intrinsic Comparison

We start our comparison of the intrinsic differences between the treebanks by noting some token statistics, shown in Table 4. The low type-token ratio (TTR) of Sepedi relative to isiZulu is expected, given that frequently occurring morphemes occur as separate tokens, whereas these same morphemes contribute to unique tokens (types) in the isiZulu treebank. We also note that for Sepedi the average tokens per sentence (which is equal to the average relations per sentence) is higher than that for isiZulu, which is to be expected.

In comparing the two treebanks, we follow two approaches. We first highlight the way in which these differences manifest themselves from a linguistic point of view by walking through a number of example sentences.

Then we consider the issue from a quantitative perspective, in which we measure the entropy – and as such the informativeness – of the dependency relations. This serves to confirm and to some

[9]In the Bantu languages the third person is made up of a large variety of noun classes where each noun class has its own "he/she/it" and "they" (Louwrens, 1994, p.198).

Metric	isiZulu	Sepedi
Sentences	94	94
Tokens	221	352
Types	167	158
Type-Token Ratio	0.76	0.45
Avg Tokens/Sentence	2.35	3.74

Table 4: Token statistics for isiZulu and Sepedi treebanks.

u-	*ya-*	*ku-*	*dumisa*
SC1	LongPres	OC2ps	VStem
o	*a*	*go*	*tumiša*
SC1	LongPres	OC2ps	VStem

Table 5: Comparative morphological analysis of 'He praises you'.

degree quantify the intuition that annotating on the basis of the orthography (i.e. tokens) of Sepedi and isiZulu would lead the Sepedi treebank to be more informative.

3.1. Linguistic Considerations

Consider the trees shown in Figure 2, which are the isiZulu and Sepedi renderings for 'He praises you'.

It is immediately obvious that the isiZulu tree encodes significantly less information in its dependency relations than the Sepedi tree, since the isiZulu tree consists of a single node. This is typical of the way the orthography affects how verbs are annotated. However, as shown in Table 5, the morphosyntactic structure of the two sentences is identical.

To mark the morphological features of the sentences, we use the custom `NounClass`, `Agreement` and `AgreementObj` features.

For the Sepedi sentence, the `PRON` element marked by `nsubj` is annotated with `NounClass=Bantu1` and the `PRON` element marked by `obj` is annotated with `Person=2|Number=Plur`. This reflects the fact that pronouns typically supply agreement information in a sentence.

For the isiZulu sentence, we can only annotate the verb, and hence we annotate it using `Agreement=Bantu1` and `AgreementObj=2ps`. This is necessary since two agreement values are present in the verb, and none of them are inherent features of the verb *to praise*.

We may simply note, at this point, that our decision to follow Gaustad et al. (2024) in assigning `PRON` to subject and object morphemes now causes the morphological annotation for Sepedi and isiZulu to diverge significantly on near identical morpheme sequences.

More importantly, essential information is lost in the case of isiZulu, since the treebank does not communicate which morphs are present in the verb and have given rise to the morphological annotation. Clearly, a language model trained on the Sepedi sentence would conceivably be better able to generalise to the new sentence 'He praises him (the boy)' than a model trained on the isiZulu sentence. The new Sepedi sentence would be *o a o tumiša*, while the isiZulu sentence would be the token *uyamdumisa*.

The inability in the isiZulu treebank to isolate the morphemes that function as pronouns in the sentence has some other strange side-effects. Consider the sentence in Figure 3. We again give a (slightly simplified) morphological analysis in Table 6, showing the high degree of similarity between the sentences.

Here, in the Sepedi sentence, the subject concord *ke* ('I') functions as the unmarked subject of the sentence, while the addition of *nna* has a determinative and emphatic function. This can easily be accommodated in the Sepedi treebank as shown in Figure 3. However, the isiZulu sentence mirrors the morpheme sequence of the Sepedi sentence exactly, and *mina* performs the exact same function in the isiZulu sentence as *nna* in the Sepedi. An annotation that captures this leads to the tree in Figure 3, where the relation between *mina* and the verb *ngihamba* is marked as `det:emph`. At best, this strikes us as linguistically awkward.

3.2. Entropy with Respect to Dependency Relations

We turn now to comparing the information contained in the dependency relations of the two treebanks. Specifically, we want to know how informative dependency relation annotations are for each of the tokenisation strategies.

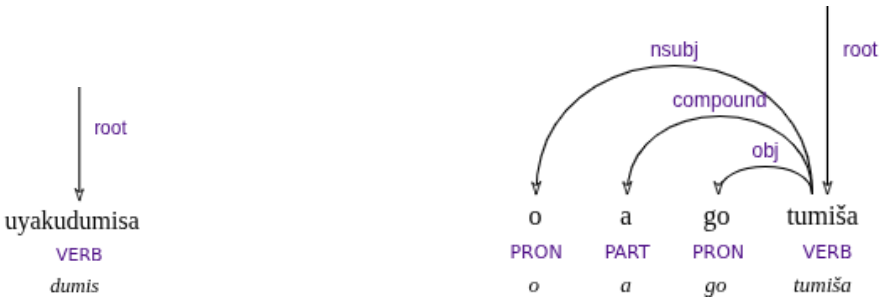
To measure informativeness, we calculate the Shannon entropy over dependency relations for each treebank. This measures on average how surprising, or stated otherwise, informative, dependency relations are within each treebank.

Shannon entropy H over dependency relations X is defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i)$$

where $X = \{x_1, x_2, ..., x_n\}$ is the set of dependency relations for a given treebank, and $p(x_i)$ is the relative frequency of x_i.

The resulting number for H represents the average number of bits needed to encode a dependency relation, and hence the higher the number,



(a) isiZulu (b) Sepedi

Figure 2: Trees for the sentence *He praises you.*

mina	*ngi-*	*hamba*	*ne-*	*ntombi*
Pron1ps	SC1ps	VStem	AdvPre	N9
nna	*ke*	*sepela*	*le*	*mosetsana*
Pron1ps	SC1ps	VStem	AdvPre	N1

Table 6: Comparative morphological analysis of 'I myself walk with a girl'.

the more informative, on average, dependency relation annotations are within the treebank.

Finally, to get a sense of how informative the dependency relation annotations are per sentence, we can multiply H with the average number of relations per sentence.

The results are given Table 7 and clearly show that, per sentence on average, the Sepedi treebank is more than twice as informative as the isiZulu treebank.

4. Extrinsic Comparison

Given the significant difference in informativeness of the annotation approaches for the two languages, the next question to consider is what practical effects this may have in using the treebanks for downstream tasks. One such task is to train a parser on the treebanks.

UDPipe is an open source trainable pipeline for training dependency parsers on treebanks in the CoNLL-U format (Straka et al., 2016). In order to investigate the two different annotation approaches on training dependency parsers, we run the UDPipe training script on both treebanks, following a 10-fold cross-validation approach to mitigate the small sizes of the treebanks.

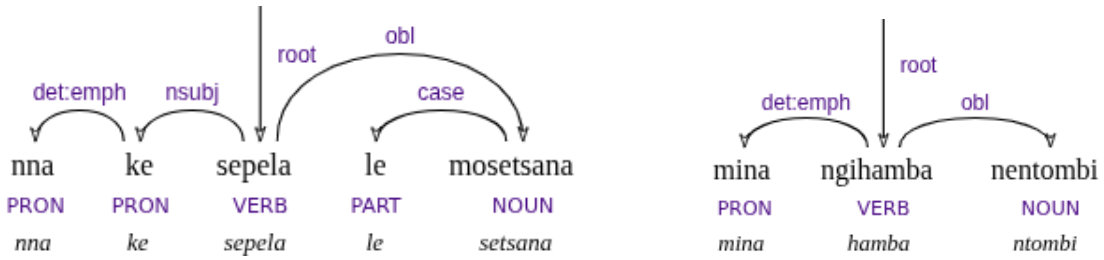
Since the Sepedi treebank is significantly more informative, annotating a Sepedi sentence is a harder task than annotating an isiZulu sentence. On the other hand, the sentences in the parallel treebanks are essentially equivalent, and hence a learning algorithm may benefit from a more informative annotation scheme.

Our experiment consists of training four different models on each treebank. We train two baseline models in which both a tagger and parser are trained, and two models where the gold standard UPOS tags are provided and only a parser is trained. For both the baseline and gold UPOS models, we train one version each that includes the morphological feature annotations, and one model each in which the features are removed before training. Since much of our annotations were added to the miscellaneous column (which is ignored by UDPipe), for the models trained on data that includes morphological features, we merge the miscellaneous features with the universal features before training.

The results for labeled attachment scores (LAS) and unlabeled attachment scores (UAS) are given in Table 8. Table 9 shows the performance of the tagger trained on the baseline models.

The first thing we notice is that the unlabeled attachment score (UAS) is always better for Sepedi than for isiZulu. In addition, the labeled attachment scores (LAS) provide some useful insights into the relative informativeness of the treebanks, visualised in Figure 4.

For the Sepedi, model performance improves in an intuitive way. Performance on the baseline model is better when the features are supplied. If we compare the baseline model with features to the gold UPOS model without features, we only see a small improvement, suggesting that the UPOS tagger trained when features are present produces a tagger that is about as useful as having gold standard UPOS tags (without features). Measuring the UPOS tagger directly, we indeed find that it has an accuracy of 88.39% (Table 9). The biggest gain comes from providing gold standard morphological features in addition to gold standard UPOS



(a) Sepedi (b) isiZulu

Figure 3: Trees for the sentence 'I myself walk with a girl'.

Measure	isiZulu	Sepedi
Relations/sentence	2.351	3.744
Avg bits/relation	2.477	3.222
Avg bits/sentence	5.824	12.067

Table 7: A comparison of the informativeness of dependency relation annotations for both treebanks.

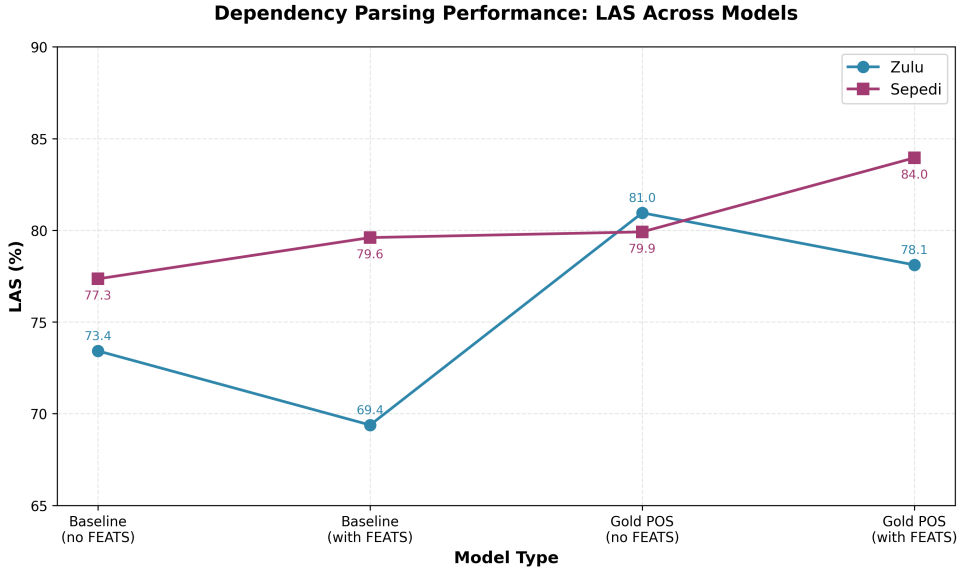


Figure 4: Labeled attachment score across model types.

Model	isiZulu	Sepedi
LAS (%)		
Baseline (no FEATS)	73.42	**77.35**
Baseline (with FEATS)	69.38	**79.60**
Gold UPOS (no FEATS)	**80.95**	79.91
Gold UPOS (with FEATS)	78.11	**83.95**
UAS (%)		
Baseline (no FEATS)	83.77	**89.99**
Baseline (with FEATS)	77.41	**93.53**
Gold UPOS (no FEATS)	89.96	**93.00**
Gold UPOS (with FEATS)	88.83	**95.61**

Table 8: Dependency parsing performance (LAS and UAS) across model configurations for isiZulu and Sepedi.

Metric	isiZulu	Sepedi
Baseline (no FEATS)		
POS Accuracy	74.63	**88.39**
FEATS F1	—	—
Baseline (with FEATS)		
POS Accuracy	71.22	**85.59**
FEATS F1	21.86	**44.78**

Table 9: POS accuracy and FEATS F1 scores for baseline models on isiZulu and Sepedi.

tags. This shows that UDPipe is not learning to produce all relevant features and is helped when they are provided. This is confirmed by the measured F1 score of the predicted features, which is only

44.78%.

The picture for isiZulu is very different. Apart from the gold UPOS model without features (where the isiZulu model is slightly better), isiZulu models consistently perform worse than their Sepedi counterparts. What is also notable is that model performance degrades whenever features are added. Indeed, we see the baseline model's tagger struggling to learn the features, with an F1 score of 21.86%. The fact that adding features degrades performance could suggest overfitting or a search space that is too large given the small amount of training data. The latter is very likely on such a small dataset as ours, given the well-known problem of data sparsity when it comes to a conjunctively written, agglutinative language such as isiZulu. This has been shown for sequence-to-sequence models to be mitigated by performing morphological segmentation (Mkhwanazi and Marais, 2024). The superior performance of the Sepedi model, which is in effect morphologically segmented to a large degree, confirms this.

5. Conclusion

We have presented parallel UD treebanks for isiZulu and Sepedi that have their origin in a multilingual GF treebank[10]. This means that each pair of sentences in the treebanks are linearisations of the same GF abstract syntax tree. Given the high linguistic similarity between the two languages, we have what is in effect treebanks that are near-identical from a morphosyntactic point of view. This has allowed us to compare the informativeness of the two treebanks when they are annotated as UD treebanks *at the token level*.

Our intrinsic comparison consisted of considering linguistic effects of the two annotation strategies and measuring the relative entropy of the relations in the treebanks. The latter showed that Sepedi token annotations are twice as informative as isiZulu token annotations. Our extrinsic comparison centred around the performance of dependency parsers trained on the treebanks, where Sepedi parsers outperform isiZulu parsers in a variety of model configurations.

From these comparisons, it is clear that there are significant advantages in annotating a Sepedi treebank at the token (and hence sub-syntactic) level, while annotation of an isiZulu treebank may share the same benefits if morphological surface segmentation of tokens was performed.

We relied on the standardised orthography of Sepedi to determine sub-syntactic elements for annotation. Given the non-triviality of morphological surface segmentation for isiZulu, future work might involve determining what systematic for segmentation would best serve isiZulu within the UD framework. However, our results indicate that a segmentation strategy that mirrors the standardised Sepedi orthography would provide significant increases in informativeness.

[10]The data is available at https://github.com/LauretteM/gf-bantu-resources.

6. Future work

A full contextualisation of this comparative analysis of two Bantu languages in terms of similar work for other languages (see, for example, (Goldman et al., 2025)) is beyond the scope of this paper, but forms part of important future work.

In this paper we performed an analysis on a small, high quality dataset. Although the sentences in our corpus are relatively short, the investigation centred around the interaction of the morphology and orthography of isiZulu and Sepedi. As such, our analysis would generalise to any corpus containing longer sentences. Future work might entail a similar analysis of a larger corpus.

7. Bibliographical References

Gertrud Faaß, Sonja Bosch, and Elsabé Taljard. 2012. Towards a Part-of-Speech Ontology: Encoding Morphemic Units of Two South African Bantu Languages. *Nordic Journal of African Studies*, 21(3):23–23.

Tanja Gaustad, Ansu Berg, Rigardt Pretorius, and Roald Eiselen. 2024. The first Universal Dependency Treebank for Tswana: Tswana-Popapolelo. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@ LREC-COLING 2024*, pages 55–65.

Omer Goldman, Leonie Weissweiler, and Reut Tsarfaty. 2025. Workshop of The UniDive 2025 Shared Task on Multilingual Morpho-Syntactic Parsing: Proceedings of the Workshop.

Inge M. Kosch. 2006. *Topics in morphology in the African language context*. Unisa Press.

Louis J. Louwrens. 1991. *Aspects of Northern Sotho grammar*. Via Afrika Ltd.

Louis J. Louwrens. 1994. *Dictionary of Northern Sotho grammatical terms*. Via Afrika, Pretoria, South Africa.

Laurette Marais, Laurette Pretorius, and Lionel Clive Posthumus. 2024. Bootstrapping syntactic resources from isiZulu to Siswati. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@ LREC-COLING 2024*, pages 77–85.

Lutz Marten. 2021. Noun Classes and Plurality in Bantu Languages. In *The Oxford Handbook of Grammatical Number*. Oxford University Press.

Sthembiso Mkhwanazi and Laurette Marais. 2024. Generation of segmented isiZulu text. *Journal of the Digital Humanities Association of Southern Africa*, 5(1).

George Poulos and Louis J. Louwrens. 1994. *A linguistic analysis of Northern Sotho*. Via Afrika Ltd.

George Poulos and Christian T. Msimang. 1998. *A linguistic analysis of Zulu*. Via Afrika Ltd.

Laurette Pretorius and Sonja E Bosch. 2003. Computational aids for Zulu natural language processing. *Southern African Linguistics and Applied Language Studies*, 21(4):267–282.

Aarne Ranta. 2009. The GF resource grammar library. *Linguistic Issues in Language Technology*, 2.

Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

P.C. Taljaard and S.E. Bosch. 1988. *Handbook of isiZulu*. J.L. Van Schaik, Pretoria.

Sepedi		
Token	**UPOS**	**Features**
yo	PRON	Agreement=Bantu1
monna	NOUN	NounClass=Bantu1
ga	ADV	Polarity=Neg
a	PRON	Agreement=Bantu1
na	VERB	_
mogatša	NOUN	NounClass=Bantu1

isiZulu		
Token	**UPOS**	**Features**
le	PRON	Agreement=Bantu9
ndoda	NOUN	NounClass=Bantu9
ayinamfazi	NOUN	Polarity=Neg
		Agreement=Bantu9
		NounClass=Bantu1

Table 10: Morphological features in the sentence 'This man has no wife' in Sepedi and isiZulu.

8. Appendices

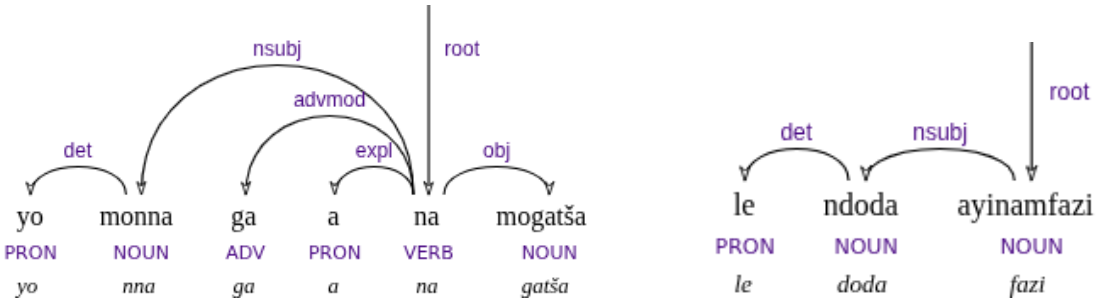
8.1. Appendix A: Morphological Annotation Examples

In Table 10, we give an example that contains both inherent morphological features and concordial features of tokens. In the subject noun phrases of the sentences (*yo monna* and *le ndoda*), the nouns are annotated using the `NounClass` feature, indicating that it is an inherent feature, and the demonstrative pronouns preceding them are annotated using the `Agreement` feature, indicating that the feature if concordial.

In the predicates of the sentences (*ga a na mogatša* and *ayinamfazi*), the annotation reflects the different orthographies, with each token in the Sepedi annotated with a single feature, and the single token in isiZulu annotated with multiple features.

In this case, a distinction between inherent and concordial features may not be strictly necessary for the Sepedi, since each token has a single feature annotation. However, the morphologically complex token *ayinamfazi*, which includes a subject agreement morpheme (*-yi-*), as well as the copulative base noun (*-mfazi* belonging to noun class 1), requires that we annotate two kinds of agreement information.

In addition, utilising two sets of features `NounClass` and `Agreement` reflects accurately in the Sepedi which tokens provide the agreement information, and which tokens reflect the agreement information of other elements of the sentence (see Figure 5).



(a) Sepedi (b) isiZulu

Figure 5: Trees for the sentence 'This man has no wife' in Sepedi and isiZulu.

Table 11 provides another example of how inherent and concordial features are used in both languages (see Figure 2 for their trees).

In the Sepedi sentence, the subject and object agreement morphemes (*o* and *go*) are annotated using inherent features `Person`, `Number` and `NounClass`. However, this annotation does not make sense for isiZulu. Although the subject and object agreement morphemes (*u-* and *ku-*) fulfill exactly the same role in the sentences, it makes no linguistic sense to annotate a verb token as though it has inherent agreement features, and two different inherent agreement features to boot. Our schema for morphological annotation therefore diverges for the two languages on this point, but we feel this strikes the correct balance between consistency across closely related languages and linguistic integrity for each language on its own terms.

8.2. Appendix B: Summary of Annotations

Table 12 lists the *morphemes* and *word categories* that occur as *tokens* in the Sepedi UD treebank, together with the dependency relations and UPOS with which they are annotated. Similarly, Table 13

Sepedi		
Token	**UPOS**	**Features**
o	PRON	NounClass=Bantu1
a	PART	Tense=Pres
go	PRON	Person=2
		Number=Sing
tumiša	VERB	

isiZulu		
Token	**UPOS**	**Features**
uyakudumisa	VERB	Agreement=Bantu1
		Tense=Pres
		AgreementObj=2ps

Table 11: Morphological features in the sentence 'He praises you' in Sepedi and isiZulu.

Token	**Relation**	**UPOS**
noun	nsubj, obj, nmod:poss, obj:lmod,	NOUN
	obl, root, vocative, obl:tmod	NOUN
absolute pronoun	det:emph, root	PRON
demonstrative	det, obj	PRON
possessive pronoun	nmod:poss	PRON
quantitative pronoun	det	PRON
adjective stem	root (when copulative complement)	ADJ
verb stem	root	VERB
infinite verb stem	xcomp	VERB
subject agreement morpheme	expl or nsubj (when pronominalised)	PRON
subject agreement morpheme class 15 (*go*)	compound	PRON
object agreement morpheme	expl or obj (when pronominalised)	PRON
negative morpheme	advmod	ADV
potential morpheme	compound	PART
progressive morpheme	compound	PART
present tense morpheme	compound	PART
future tense morpheme	compound	PART
adjective agreement morpheme	mark	PART
possessive agreement morpheme	case	PART
associative, instrumental morpheme	case	PART
auxiliary verb stem	aux	AUX
identifying copulative	cop	AUX
descriptive copulative	cop	AUX
associative copulative	root	VERB
infinite verb (*go*)	mark	PART

Table 12: Sepedi: Common tokens, their dependency relations and UPOS.

lists the *word categories*[11] that occur as *tokens* in the isiZulu UD treebank, together with the dependency relations and UPOS with which they are annotated. We follow the word categorisation of Poulos and Louwrens (1994) and Poulos and Msimang (1998).

[11] Due to the conjunctive orthography of isiZulu there are no morphemes that can occur as tokens.

Token	Relation	UPOS
noun	nsubj, obj, nmod:poss, obj:lmod,	NOUN
	obl, root, vocative	NOUN
absolute pronoun	det:emph, root	PRON
demonstrative	det, obj	PRON
possessive pronoun	nmod:poss	PRON
quantitative pronoun	det	PRON
verb	root, nsubj, obj	VERB
infinite verb	xcomp	VERB
adjective	root (when copulative complement)	ADJ
adverb	advmod, obl	ADV
auxiliary verb	root	ADV
auxiliary verb	cop	AUX

Table 13: IsiZulu: Common tokens, their dependency relations and UPOS.