Universal Dependencies for Afrikaans

Peter Dirix\textsuperscript{1}, Liesbeth Augustinus\textsuperscript{1,2}, Daniel van Niekerk\textsuperscript{3}, and Frank Van Eynde\textsuperscript{1}

\textsuperscript{1}University of Leuven, Centre for Computational Linguistics, fname.lname@kuleuven.be
\textsuperscript{2}FWO-Vlaanderen
\textsuperscript{3}North West University, Centre for Text Technology, fname.lname@nwu.ac.za

Abstract

The Universal Dependencies (UD) project aims to develop a consistent annotation framework for treebanks across many languages. In this paper we present the UD scheme for Afrikaans and we describe the conversion of the AfriBooms treebank to this new format. We will compare the conversion to UD to the conversion of related syntactic structures in typologically similar languages.

1 Introduction

Afrikaans is a West Germanic language spoken by about 7 million people in South Africa, Namibia and a worldwide diaspora, mainly in English-speaking countries. It is one of the eleven official languages in South Africa and a main lingua franca in both South Africa and neighbouring Namibia.

Until recently, not many NLP tools were available for Afrikaans. Pilon (2005) developed a fine-grained morpho-syntactic tag set and trained a version of the TnT tagger (Brants, 2000) on a manually corrected set of ca 20K words. This tagger was used to annotate the 58M-word Taalkommissie corpus.\textsuperscript{1} The annotated corpus was subsequently put into a search tool (Augustinus and Dirix, 2013). The first small Afrikaans treebank was only created in 2015 in the context of the the AfriBooms project (Augustinus et al., 2016).

We will discuss the setup of the AfriBooms treebank in Section 2 and continue with a short overview of Universal Dependencies in Section 3. The UD language-specific description for Afrikaans as well as the conversion to the UD scheme will be given in Section 4. Section 5 concludes and discusses some plans for future work.

2 AfriBooms treebank

The basis for the development of the AfriBooms treebank is a filtered subset of the Afrikaans part of the NCHLT Annotated Text Corpora.\textsuperscript{2} It contains ca 49K tokens of PoS tagged government domain documents.

The original PoS annotation of the NCHLT corpus was based on a fine-grained tag set (Pilon, 2005). As some of the information in that tag set turned out to be superfluous for determining the sentence dependency structure, the PoS tag set was simplified to a largely universal set of PoS tags (Petrov et al., 2012). This was done in order to facilitate the syntactic annotation process for the human annotators. For example, 17 classes of verb PoS tags, distinguishing present and past tense; main verbs and auxiliaries; copular verbs, transitive verbs, intransitive verbs and verbs requiring a prepositional phrase for main verbs; separable and inseparable verbs; and finally for auxiliaries the type (modal, auxiliary of tense, auxiliary of aspect, auxiliary of mode) were all mapped to one tag \texttt{VERB}. Table 1 presents the resulting tag set.

The simplified corpus was syntactically annotated with the first version of an Afrikaans parser.\textsuperscript{3} In a next step, the annotations were manually checked by one primary annotator while a subset containing 943 words was double-checked by a second annotator. The inter-annotator agreement (IAA) is calculated in terms of labelled attachment score (LAS) and unlabelled attachment score (UAS) averaged over words (Nivre et al., 2007). The LAS is 82.5%, while the UAS is 88.9%.

\begin{table}
\centering
\begin{tabular}{|c|c|}
\hline
PoS & Tag \\
\hline
Noun & \texttt{NOUN} \\
Verb & \texttt{VERB} \\
Adjective & \texttt{ADJ} \\
Adverb & \texttt{ADV} \\
Preposition & \texttt{PREP} \\
Conjunction & \texttt{conj} \\
Pronoun & \texttt{PRON} \\
Determiner & \texttt{DET} \\
Phraze & \texttt{p} \\
\hline
\end{tabular}
\caption{The Afrikaans PoS tag set}
\end{table}
Table 1: The PoS tag set and its frequencies in the AfriBooms treebank

| AB PoS TAG | FREQUENCY | DESCRIPTION                        |
|------------|-----------|------------------------------------|
| ADJ        | 2781      | Adjectives                         |
| ADP        | 5561      | Adpositions                        |
| ADV        | 1481      | Adverbs                            |
| CONJ       | 2616      | Conjunctions                       |
| DET        | 4113      | Determiners                        |
| NOUN       | 9964      | Nouns (including proper nouns)     |
| NUM        | 635       | Numerals                           |
| PRON       | 3561      | Pronouns                           |
| PRT        | 1677      | Particles                          |
| PUNCT      | 4028      | Punctuation                        |
| VERB       | 6720      | Verbs (including auxiliary verbs)  |
| X          | 758       | Catch-all class (including abbreviations and interjections amongst others) |

Table 2: The Stanford dependency tag set and its frequencies in the AfriBooms treebank

| AB DEPENDENCY TAG | FREQ. | DESCRIPTION                      |
|-------------------|-------|----------------------------------|
| dep               | 1009  | dependent                        |
| dep:punct         | 4497  | punctuation                      |
| dep:root          | 1870  | root                             |
| dep:aux           | 2534  | auxiliary (verb)                 |
| dep:conj          | 2359  | conjunct                         |
| dep:cc            | 1886  | coordination (e.g. to conjunctions) |
| dep:arg           | 1130  | argument                         |
| dep:arg:subj      | 2605  | subject                          |
| dep:arg:comp      | 3     | complement                       |
| dep:arg:comp:obj  | 3763  | object                           |
| dep:arg:comp:obj:dobj | 111 | direct object                    |
| dep:arg:comp:obj:obj | 0 | indirect object                  |
| dep:arg:comp:obj:obj:obj | 6106 | object of preposition            |
| dep:arg:comp:compl | 0 | complementiser                   |
| dep:arg:comp:mark | 5 | marker (introducing adverbial clause) |
| dep:arg:comp:rel  | 0 | relative (introducing relative clause) |
| dep:arg:comp:acomp | 0 | adjectival complement            |
| dep:mod           | 11120 | modifier                         |
| dep:mod:advcl     | 0 | adverbial clause modifier        |
| dep:mod:tm          | 0 | temporal modifier                |
| dep:mod:amod       | 2447 | adjectival modifier              |
| dep:mod:num        | 462  | numeric modifier                 |
| dep:mod:number     | 0 | element of compound number       |
| dep:mod:appos      | 0 | appositional modifier            |
| dep:mod:abbrev     | 63   | abbreviation modifier            |
| dep:mod:adv        | 0 | adverbial modifier               |
| dep:mod:adv:neg    | 0 | negation modifier                |
| dep:mod:poss       | 830  | possession modifier              |
| dep:mod:prt        | 1375 | phrasal verb particle            |
| dep:mod:det        | 5101 | determiner                       |
| dep:mod:prep       | 0 | prepositional modifier           |

For the dependency relations, a subset of the Stanford tag set was adopted, applying the conventions of De Marneffe (2006; 2008). An overview of the dependency tags together with their frequencies is given in Table 2.

The figures in Table 2 show that the annotators often fell back onto more generic tags such as dep, dep:arg and dep:mod, resulting in a large amount of syntactic relations that could have been further specified.

All sentences in the treebank are validated according to the following principles:

- **Graph completeness**: Each sentence must form a single complete graph, i.e. all words must be reachable from the root node.
- **Dependence restriction**: Words may have multiple dependents and each phrase has at most one head.
- **Projectivity**: Connection lines between words should not cross each other.

The treebank was delivered in the Folia-XML format (van den Bosch et al., 2007). An example of a sentence from the AfriBooms treebank is given in Figure 1. It visualizes the PoS tag and dependency annotation for the sentence Die webtuiste sal ‘n nuwe deurblaai-venster oopmaak. ‘The website will open a new browser window’.

The different phases of the bootstrapping of the parsing process, as well as the details of the manual annotation and verification process are described in Augustinus et al. (2016).
Universal Dependencies

Universal Dependencies (UD) is a project developing cross-language consistent treebank annotation for as many languages as possible, aiming to facilitate multilingual or language-independent parser development, cross-lingual learning, and linguistic research from a language typology perspective (Nivre et al., 2016). The annotation scheme is based on a combination of adapted (universal) Stanford dependencies (de Marneffe et al., 2006; de Marneffe and Manning, 2008), Google universal PoS tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tag sets (Zeman, 2008). The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, and allowing language-specific extensions when necessary to encode specific features. Guidelines for version 2.0 as well as the treebanks released in this version are published on the project’s website.4

Universal dependencies describe dependency relations between words. For most languages, white space determines what a token is. Apart from contractions and clitics, words are not segmented. The use of multi-word tokens is limited to a few fixed expressions that function as adverbs or adpositions.

UD treebanks are represented in the CoNLL-U format, which is an adaptation of the older CoNLL-X format. This format is a tab-separated text file with ten columns. The first three columns respectively contain the position of the token in the sentence, the token and its lemma. Lemmas are defined as the dictionary form of the token, which depends on the language. For example verb lemmas are typically represented by the infinitive, but in Greek the indicative present first person singular is employed. Column 4 contains the universal PoS tag. The morphosyntactic annotation of the pre-converted treebank, if any, can be put in column 5. The universal and language-specific morphological features describing number, case, person, gender, mood, tense etc. in the column 6. Column 7 indicates the head of the current token in reference to its position in the sentence (column 1). Column 8 contains the universal dependency relation, while column 9 (optionally) contains an enhanced dependency graph in the form of head-depency relation pairs. Any other type of annotation can be placed in column 10. Fields should never be empty and may have an underscore as place-holder if necessary. Figure 2 presents the sentence in Figure 1 in the UD format.

Converting AfriBooms to the UDT format

4.1 Language-specific definitions for Afrikaans

In general, the structural conversion for the Afrikaans treebank aims to be in line with what was done for Dutch and German UDIs, as they are the two languages closest to Afrikaans. For those languages, UD treebanks are already available. Despite the fact that Afrikaans has a simplified morphology compared to Dutch, the languages share a lot of features, which include gen-

4https://universaldependencies.org
eral Indo-European ones, but also very specific (West) Germanic characteristics such as extensive nominal compounding, separability of compound verbs, and extensive diminutive formation.

4.1.1 PoS tags and morphological features

The Afrikaans Universal PoS tags and features are listed in Table 3.

**Nouns** For nouns (NOUN) and proper names (PROPN), we introduce a feature Degree next to number. As in in Dutch, this is in order to cover the extensive possibilities of diminutive formation, e.g. *huis* ‘house’ gets *huisie* ‘little house’ and *Jan* (‘John’) has *Jantjie* ‘little John’. Besides this language-specific feature, we need the Num feature, but not Case as Afrikaans has hardly any remainders of the old Germanic case system. The genitive is expressed by the particle *se*, which is covered by the feature PartType=Gen for particles. There are still a few fixed expressions inherited from Dutch, like *ter ere van* ‘in honour of’, which will be considered as multi-word adverbials or prepositions. We will treat fixed Latin expressions such as *ex aequo* similarly. We do not include the difference between common nouns, measurement nouns, collectives and abstract nouns, which was present in the original tag set, as most of those tags did not occur in the training set of the tagger anyway. Pluralia tanta will be represented as ‘plural’ nouns.

**Adjectives** Adjectives (ADJ) have degrees of comparison like most other Indo-European languages. We introduce the Case feature to cover for the formal, archaic genitive forms like *iets interessants* ‘something interesting’ (Donaldson, 1993). Other archaic accusative or dative forms might occur in fixed expressions (e.g. *te geleëner tyd* ‘at the proper time’), but as there are very few, these expressions are also considered multi-word adverbials. In addition to these features we need to introduce a new language-specific feature AdjType to account for the fact that most adjectives have a different form depending on whether they are used attributively or predicatively, e.g. *’n eerlike kêrel* (‘an honest guy’) vs. *dié kêrel is eerlik* (‘this guy is honest’). This is only relevant for the nominative case. The forms are indistinguishable in the comparative and superlative, but as our original tag set does make the distinction, we propose to keep it. As prescribed in the UD guidelines, ordinal numbers form part of adjectives.

**Adverbs** For adverbs (ADV), we only keep the differences in degree and we will not introduce features to describe the type of adverb (temporal, modal, etc.) at this point.

**Verbs** Verbs (VERB) have a very simple morphology in Afrikaans. Apart from a few auxiliaries and modals, verbs only have one present form, which also serves as infinitive, and a past participle, which is used in the formation of the past tense. The verb *wees* ‘to be’ has a separate infinitive next to the present form *is* ‘is’, while it also has an old preterite form to express the past (*was* ‘was’), just like some modals. Present and past participles are considered as adjectives when they behave as such. An indication of the distinction native speakers make between past participles and their (declinable) adjectival forms, is the fact that the old Dutch strong forms can only be used in adjectival positions and not to form the past tense (Donaldson, 1993). For example, one cannot say *die kind word aangenome*, only *die kind word aangeneem* ‘the child is adopted’. However, one can say *die kind is aangeneem*, in which case this is analysed as a combination of the copula and a predicative complement, next to *die kind is aangeneem*, which means ‘The child has been adopted’. The strong form often has a more figurative or abstract meaning, as in *’n gebroke hart* (‘a broken heart’) vs. *’n gebreekte bord* (‘a broken plate’) (Conradie, 2017).

For the category of auxiliaries (AUX) we intro-
duce the VerbType feature to distinguish between copular verbs, modal verbs, the passive auxiliaries word ‘be’ (present) and wees (past), and other auxiliaries. This is similar to the Dutch treatment, apart from the introduction of the passive voice category. Like Dutch and German, Afrikaans has separable verbs, i.e. verbs that are actually compounds of a particle (or sometimes an adjective or a noun) and another verb. In verb-initial clauses, the two parts get separated and the particle moves to the end of the clause. An example is Ek gaan die huis binne ‘I enter the house’ with the separable verb binnegaan ‘to enter’ (literally ‘inside go’).5

**Pronouns and determiners** Pronouns are treated in a similar way as in Dutch. Possessive and reflexive pronouns are considered a subset of the personal pronouns and have person and number features. We do not indicate the gender of third person pronouns. On top of this, we distinguish relative, interrogative, indefinite, and reciprocal pronouns as a part of the PRON class. Demonstrative pronouns are put in the DET class together with indefinite determiners and articles, as required by the UD guidelines. For articles, the distinction between definite and indefinite articles is indicated by the Definite feature.

**Adpositions** We also follow the Dutch annotations in defining three types of adpositions (ADP): prepositions, postpositions and circumpositions, encoded with the AdpType.

**Particles** We introduce three types of particles (PART): te ‘to’ introducing the infinitive (denoted by Inf), the genitive particle se ‘his/her/their’ (similarly used as ‘s in English and denoted by Gen), and the negative particle nie ‘not’ which is used in most negative sentences in addition to a negative adverb or determiner (Huddleston, 2010).

**Remaining PoS tags** The other PoS tags for numerals (NUM), coordinating conjunctions (CCONJ), subordinating conjunction (SCONJ), interjections (INTJ), punctuation (PUNCT), symbols (SYM), and the remainder class (X) do not have any additional features.

**Contracted forms** One common contraction is the colloquial *dis* for *dit* is (the expletive ‘it is”), which needs to be split. Note that this construction does not appear in AfriBooms treebank.

### 4.1.2 Dependency relations

UD represents dependency relations between words in the form of a tree. Only one word, dependent on the ROOT, can be the head of the sentence. All other words are dependent on another word in the tree. The main driving principle of the UD formalism is the primacy of content words.

---

| UD PoS Tag | Description | Morphological Features |
|------------|-------------|------------------------|
| ADJ        | Adjectives  | AdjType=Attr,Pred; Case=Nom,Gen; Degree=Cmp,Pos,Sup |
| ADP        | Adpositions | AdpType=Circ,Post,Prep |
| ADV        | Determiners | Degree=Cmp,Pos,Sup |
| AUX        | Auxiliaries | Tense=Past,Pres; VerbForm=Fin,Inf; VerbType=Aux,Cop,Mod,Pass |
| CCONJ      | Coordinating conjunctions |  |
| DET        | Determiners | Definite=Def,Ind; PronType=Art,Dem,Ind; Degree=Dim; Num=Plur,Sing |
| INTJ       | Interjections |  |
| NOUN       | Nouns | Degree=Dim; Num=Plur,Sing |
| NUM        | Numerals | PartType=Inf,Neg,Gen |
| PART       | Particles | Case=Nom,Acc; Number=Plur,Sing; Person=1,2,3; Poss=Yes; PronType=Ind,Int,Prs,Repl; Reflex=Yes |
| PRON       | Pronouns | Degree=Dim; Num=Plur,Sing |
| PROPN      | Proper names |  |
| PUNCT      | Punctuation |  |
| SCONJ      | Subordinating conjunctions |  |
| SYM        | Symbols |  |
| VERB       | Non-auxiliary verbs | Tense=Pres,Part; VerbForm=Fin,Inf,Part; Subcat=Intr,Prep,Tran |
| X          | Other |  |

Table 3: Afrikaans Universal PoS tags and their potential morphosyntactic feature values

---

5In verb-final clauses, the verb is placed at the end of the clause after the particle, and the two are (usually) treated as a single orthographic unit. Compare the verb-initial construction to a construction with a subordinate clause: Hy sien dat ek die huis binnegaan ‘He sees that I enter the house’.
Table 4: Frequency of the UD PoS tags and the morphological features in the Afrikaans UD treebank

| UD PoS TAG | MORPHOLOGICAL FEATURES | FREQ. | EXAMPLE |
|------------|-------------------------|-------|---------|
| ADJ        | AdjType=Attr|Case=Nom|Degree=Cmp | 34 | minder |
| ADJ        | AdjType=Attr|Case=Nom|Degree=Pos | 2321 | tweede |
| ADJ        | AdjType=Attr|Case=Nom|Degree=Sup | 41 | doeltreffendste |
| ADJ        | AdjType=Pred|Case=Nom|Degree=Cmp | 20 | vinniger |
| ADJ        | AdjType=Pred|Case=Nom|Degree=Pos | 419 | nuttig |
| ADJ        | AdjType=Pred|Case=Nom|Degree=Sup | 5 | hoogste |
| ADP        | AdpType=Prep | 5604 | in |
| ADV        | Degree=Cmp | 54 | beter |
| ADV        | Degree=Pos | 1728 | vandag |
| ADV        | Degree=Sup | 11 | mees |
| AUX        | Tense=PastVerbForm=Fin|VerbType=Cop | 54 | was |
| AUX        | Tense=PastVerbForm=Fin|VerbType=Mod | 20 | wou |
| AUX        | Tense=PastVerbForm=Fin|VerbType=Pass | 266 | is |
| AUX        | Tense=PresVerbForm=Fin|InfVerbType=Aux | 384 | het |
| AUX        | Tense=PresVerbForm=Fin|InfVerbType=Cop | 608 | is |
| AUX        | Tense=PresVerbForm=Fin|InfVerbType=Mod | 1049 | sal |
| AUX        | Tense=PresVerbForm=Fin|InfVerbType=Pass | 543 | word |
| CCONJ      | | 1768 | en |
| DET        | Def|PronType=Art | 3237 | die |
| DET        | Ind|PronType=Art | 876 | 'n |
| DET        | PronType=Dem | 396 | hierdie |
| DET        | PronType=Ind | 315 | baie |
| NOUN       | Degree=Dim|Number=Plur | 5 | koekies |
| NOUN       | Degree=Dim|Number=Sing | 9 | koekie |
| NOUN       | Number=Plur | 2610 | blaaiers |
| NOUN       | Number=Sing | 6784 | toegang |
| NUM        | | 197 | twee |
| PART       | PartType=Gen | 152 | se |
| PART       | PartType=Inf | 836 | te |
| PRON       | Case=Acc|Nom|Person=1|PronType=Prs | 470 | ons |
| PRON       | Case=Acc|Nom|Number=Plur|PronType=Prs | 4 | julie |
| PRON       | Case=Acc|Nom|Number=Plur|PronType=3|PronType=Prs | 96 | hulle |
| PRON       | Case=Acc|Number=Sing|Person=1|PronType=Prs | 8 | my |
| PRON       | Case=Acc|Number=Sing|Person=2|PronType=Prs | 19 | jou |
| PRON       | Case=Acc|Number=Sing|Person=3|PronType=Prs | 13 | haar |
| PRON       | Case=Nom|Number=Sing|Person=1|PronType=Prs | 66 | ek |
| PRON       | Case=Nom|Number=Sing|Person=2|PronType=Prs | 186 | u |
| PRON       | Case=Nom|Number=Sing|Person=3|PronType=Prs | 350 | dit |
| PRON       | Number=Plur|Person=1|Poss=Yes|PronType=Prs | 308 | ons |
| PRON       | Number=Plur|Person=1|PronType=Prs | 13 | ons |
| PRON       | Number=Plur|Person=3|Poss=Yes|PronType=Prs | 89 | hal |
| PRON       | Number=Plur|Person=3|PronType=Yes | 3 | huisself |
| PRON       | Number=Sing|Person=2|Poss=Yes|PronType=Prs | 10 | jou |
| PRON       | Number=Sing|Person=2|PronType=Yes | 1 | jouself |
| PRON       | Number=Sing|Person=3|Poss=Yes|PronType=Prs | 56 | sy |
| PRON       | Number=Sing|Person=3|PronType=Yes | 12 | homself |
| PRON       | PronType=Ind | 307 | enige |
| PRON       | PronType=Int | 20 | wat |
| PRON       | PronType=Rcp | 6 | mekaar |
| PROP       | PronType=Rel | 1116 | wat |
| PROP       | Number=Sing | 463 | Suid-Afrika |
| PUNCT      | | 4027 | . |
| SCONJ      | | 946 | as |
| SYM        | | 435 | R5 |
| VERB       | Subcat=Intr|Tense=PastVerbForm=Part | 64 | gedemonstreer |
| VERB       | Subcat=Intr|Tense=PresVerbForm=Fin|Inf | 547 | werk |
| VERB       | Subcat=Prep|Tense=PresVerbForm=Fin|Inf | 25 | voldoen |
| VERB       | Subcat=Tran|Tense=PastVerbForm=Part | 725 | gomeet |
| VERB       | Subcat=Tran|Tense=PresVerbForm=Fin|Inf | 2445 | ontdek |
| X          | | 385 | DRK |
This means that in general content words are
the head instead of function words, e.g. nouns
are the head of prepositional phrases. The aim
of this principle is to allow for maximal comparabil-
ity across languages. In addition to the obligatory
dependency relations, it is possible to add an en-
hanced dependency graph to this scheme with a
more complete basis for semantic interpretation.6

For instance, the regular dependency relations
lack a dependency relation between raised subjects
and an embedded verb. It is possible to encode this
kind of information in the enhanced dependency
graph.

The UD scheme defines 37 types of relations,
of which 24 are actual dependency relations. The
taxonomy for the latter is organized along two di-
mensions, which can be represented in the form
of a matrix.7 The first dimension corresponds
to functional categories in relation to the head
core arguments of clausal predicates, non-core
dependents of clausal predicates, and dependents
of nominals) whereas the second dimension cor-
responds to the structural categories of the depend-
ent (nominals, clauses, modifiers, and function
words).

Additionally, there are 13 relations that are not
dependency relations in the narrow sense. It con-
cerns relations for analyzing coordination, multi-
tword expressions, ellipsis, and special relations
for concepts such as root, punctuation and multi-
word expressions.

Afrikaans shares many syntactic features with
Dutch and German. It has, for instance, verb-
second in main clauses but verb-final in (most)
subordinate clauses (Biberauer, 2003); and there is
the occurrence of substitute infinitives, also known
as Infinitivus Pro Participio or IPP (Augustinus
and Dirix, 2013). Afrikaans also has particular
features such as double negation (Huddlestone,
2010).

In principle, all types of Universal Depen-
dency relations can be applied to Afrikaans. The
only exception is the classifier relation (clf), as
Afrikaans has no grammaticalized classifier sys-
tem. As in Dutch and German, we introduce the
compound:prt relation for compounds of which a
part has been elided, e.g. in- en uitvoer ‘import
and export’, as well as for the particle of sepa-
rate verbs. We also introduce nsubj:pass and
csubj:pass for the subjects of passive verbs, us-
ing word (present) or is (past) as auxiliary.

4.2 Conversion and issues

As described in section 2 the original NCHLT cor-
pus was available with original more fine-grained
tags of the NCHLT corpus. We reintroduced those
features in the AfriBooms treebank in order to pre-
pare for conversion to UD, as they contain mor-
phological information which is required by the
UD guidelines. In general, we kept the mor-
phological features used in UD releases for other lan-
guages. Most of the original morphological tags
have been converted to UD features, but we did
not do this for the types of adverbs or the type of
nouns, and as well as the types of symbols and
punctuation marks, as these were semantic in stead
of morphosyntactic features. The XML format of
AfriBooms treebank was converted into a tab-
separated format, which facilitates the conversion
to the CoNLL-U format considerably. The actual
conversion was done using a Perl script.

4.2.1 PoS tags and morphological features

At the level of PoS tags and features there were
hardly any disambiguation issues. Pronouns that
have the same form in their base and oblique
forms have a Case=Nom,Acc feature which could
be disambiguated manually. We have not done
this yet. This is also the case for the feature
VerbForm=Fin,Inf which is assigned to most
verbs including auxiliaries, as the form of the
present tense is identical to the infinitive.

The counts for the PoS tags and their morpho-
logical features in the automatically converted ver-
sion of AfriBooms can be found in Table 4.

As the AfriBooms treebank is relatively small,
not all possible morphological forms occur in the
treebank, e.g. the genitive form of articles and
their comparatives. This will obviously inhibit au-
tomated parser training.

4.2.2 Dependency relations

Table 5 lists the initial mapping of the dependency
relations. Compared to the conversion of the PoS
tags and morphological features, the conversion of
the dependency relations was less straightforward,
as some structural conversion was needed.

The first problem is due to the small size of the
AfriBooms treebank, as a about one third of the
Stanford dependencies tags was not used in the

6http://universaldependencies.org/u/overview/enhanced-
syntax.html
7http://universaldependencies.org/u/dep/
treebank and we can only provide mappings for dependencies occurring in that treebank.

| AB DEP. TAG | UD DEP. TAG | UD DESCRIPTION |
|-------------|-------------|----------------|
| root        | root        | root           |
| aux         | aux         | auxiliary      |
| conj        | conj        | conjunct       |
| cc          | cc          | coordinating conjunction |
| subj        | nsubj       | nominal subject |
| dobj        | obj         | object         |
| iobj        | iobj        | indirect object |
| hobj        | case        | case-marking element |
| obj         | obj         | object         |
| amod        | amod        | adjectival modifier |
| mod         | amod        | adjectival modifier |
| num         | nummod      | numeric modifier |
| appos       | appos       | appositional modifier |
| poss        | det         | determiner     |
| det         | det         | determiner     |
| prt         | mark        | marker         |
| dep         | dep         | unspecified dependency |
| arg         | dep         | unspecified dependency |
| comp        | dep         | unspecified dependency |
| mark        | dep         | unspecified dependency |
| abbrev      | appos       | appositional modifier |
| punct       | punct       | punctuation     |

Table 5: Initial mapping between the AfriBooms and UDT dependency relations

The second problem with respect to the automated conversion is the underspecification of dependency relations in the AfriBooms treebank. The UD relations only have one generic tag (dep), while the dependencies used in the AfriBooms treebank have several levels of underspecification, e.g., mod, arg, obj (see Table 2). Those relations need to be either specified automatically or manually. In Dutch, there were actually similar issues with underspecification; when possible they were resolved in an automated way (Bouma and van Noord, 2017).

The third issue is that the human annotators of the AfriBooms treebank did not consistently follow the content word primacy principle. For instance in the case of prepositional phrases they assigned the head status to the preposition, which means we had to flip the dependency relation between them and have the noun point to the governor of the phrase. The dependency relation was set to nmod or obl, depending on the PoS of the governor. A similar issue exists for copular constructions: the verb is assigned the head of the relation, while the predicative complement is identified as an object. Again, we changed the dependency relation, making the nonverbal predicate the head (mostly the root of the sentence), introducing the cop relation and switching the governor of the subject to the nonverbal predicate. Similarly, we had to fix possessive constructions like leerders se vermoë (‘learners’ ability’), to make sure the possessive particle had the case relation, and swapping the relation between the two nouns, making the second one the governor and giving it the dependency relation of the former one, while giving the former one the nmod relation. Another dependency relation that had to swap its head, are the cc types for conjunctions, which need to point to the following noun (phrase) and not to the preceding one.

The fourth problem is that some relations are not distinguished in the original AfriBooms annotation. A number of them could be (semi-)automatically introduced. For instance, as the original treebank annotations do not make a distinction between nsubj (nominal subject) and csubj (clausal subject), we converted them all to nsubj and replaced them afterwards to csubj if the governor of the subject is either a verb or an auxiliary. Furthermore, we introduced compound:prt for compounds with partial elision, as mentioned in the previous section. In order to do this, we again had to swap the dependency relation and change the governors of all the tokens depending on the partially elided compound, as the first part of the expression was treated as the head in the AfriBooms treebank. The result needs to be reviewed manually, as there are also phrases consisting of more than one compound with partial elision (e.g. klein-, medium- en mikro-ondernemings – ‘small, medium and micro-enterprises’), and phrases of the type besigheids- en ander sektore (‘business and other sectors’), which stands for besigheidssektor en ander sektore. In the latter example the first item is a partially elided compound, but the second part consists of an adjective followed by a noun, which is also the elided part of the first compound.

In addition, we introduced aux:pass for passive auxiliaries based on their morphological features, and specified nsubj:pass for the nominal subject of those verbs. We also introduced iobj for a list of ca 30 verbs for which the indirect object is introduced with the preposition aan. Finally, we also specified the flat relation in multi-word named entities.

We replaced amod with advmod for all adverbs and negative particles that had this dependency re-
lation in the original treebank.

We also fixed the dependency relation of verbs following the particle te to xcomp, as the original treebank did not distinguish between nominal and clausal constituents. Furthermore, we also had to flip the dependency relation between relative pronouns and the content verb of the relative sentence, and made sure the verb has the xcomp dependency relation.

All of this patching work was done using an additional Perl script which we ran after the initial conversion. As mentioned in section 2 the AfriBooms treebank contains many generic dep relations, which need to be further specified. Even though the patching work greatly reduced the number of generic items, many of them should be manually reviewed. This is currently in progress. An example of a converted sentence is given in Figure 3.

As a final step, the converted treebank was validated using the UDT tools available on GitHub.8 Table 6 presents the final figures of each dependency relation category after conversion and patching.

5 Conclusion and future work

We created a UD treebank for Afrikaans using an automated conversion scheme from the existing AfriBooms treebank. It is a small treebank of about 49K words consisting of governments documents. Due to its small size and the specific genre, it does not contain all possible dependency relations and morphological feature values.

As many of the dependency relations were underspecified in the original treebank, the next step consists of a manual check. We furthermore plan to train parsers for the annotation of both the dependency relations and the morphological information. Those parsers will be used to create more annotated data, for starters from Wikipedia, but possibly also for other text types, such as the Taalkommissie corpus. A (small) part of those data could be verified manually in order to improve parsing accuracy.

| UD DEP. TAG | FREQ. | DESCRIPTION |
|-------------|-------|-------------|
| advmod      | 1780  | adverbial modifier |
| amod        | 5080  | adjectival modifier |
| appos       | 63    | appositional modifier |
| aux         | 1663  | auxiliary |
| aux:pass    | 854   | passive auxiliary |
| case        | 5890  | case-marking element |
| cc          | 1886  | coordinating conjunction |
| ccomp       | 905   | clausal complement |
| compound:prt| 408   | separable verb particle / elided part of a compound |
| conj        | 2001  | conjunct |
| cop         | 149   | copula |
| csbj        | 3     | clausal subject |
| csbj:pass   | 0     | clausal subject of passive verb |
| dep         | 1668  | unspecified dependency |
| det         | 5775  | determiner |
| flat        | 231   | flat multiword expression |
| iobj        | 53    | indirect object |
| mark        | 1051  | marker |
| nmod        | 2948  | nominal modifier |
| nsbj        | 3010  | nominal subject |
| nsbj:pass   | 500   | nominal subject of passive verb |
| nummod      | 461   | numeric modifier |
| obj         | 2804  | object |
| obl         | 2728  | oblique nominal |
| punct       | 4497  | punctuation |
| root        | 1903  | root |
| xcomp       | 965   | open clausal complement |

Table 6: The UD tag set with number of occurrences in the AfriBooms treebank

---

8https://github.com/UniversalDependencies/tools
References

Liesbeth Augustinus and Peter Dirix. 2013. The IPP effect in Afrikaans: a corpus analysis. In Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013) - NEALT Proceedings Series 16, pages 213–225, Oslo.

Liesbeth Augustinus, Peter Dirix, Daniel Van Niekerk, Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde, and Gerhard Van Huyssteen. 2016. AfriBooms: An Online Treebank for Afrikaans. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 677–682, Portorož. European Language Resources Association (ELRA).

Theresa Biberauer. 2003. Verb Second (V2) in Afrikaans: a Minimalist investigation of word-order variation. Ph.D. thesis, University of Cambridge, Cambridge.

Gosse Bouma and Gertjan van Noord. 2017. Increasing return on annotation investment: the automatic construction of a Universal Dependency Treebank for Dutch. In Proceedings of the Universal Dependencies Workshop at the 21st Nordic Conference of Computational Linguistics (NODALIDA 2017), Gothenburg.

Thorsten Brants. 2000. TnT – A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000), pages 224–231, Seattle.

Jac Conradie. 2017. The re-inflecting of Afrikaans. Paper given at the Germanic Sandwich 2017, Münster.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation at COLING 2008, pages 1–8, Manchester, UK.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2006), pages 449–454, Genoa. European Language Resources Association (ELRA).

Bruce C. Donaldson. 1993. A Grammar of Afrikaans. Mouton de Gruyter, Berlin/New York.

Kate Huddleston. 2010. Negative Indefinites in Afrikaans. Ph.D. thesis, Utrecht University.

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 2016), pages 1659–1666, Portorož. European Language Resources Association (ELRA).

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pages 2089–2096, Istanbul. European Language Resources Association (ELRA).

Suléne Pilon. 2005. Outomatiese Afrikaanse woordsoortetikettering. Master’s thesis, North-West University, Potchefstroom.

Antal van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In Computational Linguistics in the Netherlands: Selected papers from the Seventeenth CLIN Meeting, pages 191–206, LOT, Utrecht.

Daniel Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), pages 213–218, Marrakesh. European Language Resources Association (ELRA).