NCHLT Project: Text Resources

2010-2012

Protocol: Morphological Analysis (Afrikaans)

# Introduction

## Aim of this protocol

The aim of this protocol is to guide annotators in the process of annotating the Afrikaans corpus on the morphological analysis layer for the NCHLT project: Text Resources. This document explicates some general principles to be followed in annotating the corpus, explains the process of annotation, and provides a detailed explanation of the tag set to be used.

This is a dynamic, “living” document, and should be adapted following the procedure as explicated in the section “Changing the protocol”.

## Concept and definitions

The aim of annotation on the morphological analysis layer is to identify the morphemes of each token (specifically morphologically complex words), as identified on the tokenisation layer (i.e. usually words, multi-word expressions, and punctuation). The following definitions apply:

* **Morphological analysis**: The identification and description of morphemes in complex words.
* **Morpheme**: A simplex symbolic (i.e. (grammatically) meaningful) unit in the language system. It is simplex in the sense that it does not contain smaller symbolic units as subparts; e.g. a morpheme, like the plural morpheme [[pl]/[s]], specifies through convention plurality, ‘more than one’-ness on the semantic pole. This definition also implies that simplex words like *tafel* ‘table’, *groot* ‘big’ or *speel* ‘play’ are considered morphemes. (See Van Huyssteen, 2010.) Figure 1 details the main categories of morphemes that are identified for this project (for detailed categories, see tag set below).



* + **Examples:**
    - Reduced stem: ***ATKV****-lid;* ***WNNR****-projek*
    - Full stem – Allomorphic minus: ***bot****e;* ***pal****e*
    - Full stem – Allomorphic plus: ***katt****e;* ***sakk****ie*
    - Full stem – Allomorphic: ***stow****e;* ***wolv****in*
    - Full stem – Unchanged: ***hond****e;* ***tafel****tjie*
    - Dependent stem: ***erde****werk;* ***denk****rigting;* ***gast****ehuis;* ***kontrakt****e;* ***elektr****isiteit;* ***administr****asie;* ***sekretar****iaat;* ***bio****diesel; hidrologie: [****hidr]*** *[o]* ***[logie]***
    - Suffix: *vergeet****agtig****; hond****jie***
    - Prefix: ***ge****speel;* ***ver****kleur*
    - Ambifix: ***ge****been****te****;* ***ge****speel****d****e*
    - Inflectional:
      * Plural (PL): *hond****e****; katt****e***
      * Diminutive (DIM): *hond****jie****; kat****jie***
      * Comparative (CMPR): *mooi****er****; lelik****er***
      * Superlative (SUP): *mooi****ste****; lelik****ste***
      * Attributive (ATTR): *mooi****e****; lelik****e***
      * Partitive Genitive (PRTT-GEN): *iets mooi****s****; iets lelik****s***
      * Past (PST): ***ge****speel;* ***ge****klim*
      * Participle (PART): ***ge****meganiseer****d****; meganiser****end***
      * Special (SPEC):
        + *Infinitive (INF): iets te drink****e****; iets te et****e***
        + *Inchoative (INCH):* ***ont****vlam;* ***ont****lok*
        + *Intensive (INTS):* ***ver****slaap;* ***ver****drink*
        + *Iterative (ITER):* ***her****doop;* ***her****skryf*
        + *Transitive (TR):* ***ver****dink;* ***be****vaar*
        + *Negative (NEG):* ***on****moontlik;* ***wan****hoop*
    - Derivational:
      * Nominaliser (NR): *rooi****e****; belangrik****heid***
      * Verbaliser (VR): ***be****moeilik;* ***ver****afrikaans*
      * Adjectivaliser (ADJR): *gel****erig****; skouspel****agtig***
      * Adverbialiser (ADVR): *sagg****ies****; wys****lik***
    - Cran morpheme: ***boe****pens;* ***tjoep****stil*
    - Linking morpheme: *besigheid****s****plan; wild****e****bees; elektr****isi****teit*

## General principles

The following general principles apply:

* + Annotation on the morphological analysis layer should be in excess of 98%
  + The tag set below is based on:
    - Van Huyssteen, GB. 2010. (Re)defining Component Structures in Morphological Constructions: A Cognitive Grammar Perspective. In: Michel, S & Onysko, A (eds.). *Cognitive Approaches to Word-Formation*. ISBN 978-3-11-022359-0. [http://www.degruyter.de/cont/fb/sk/detailEn.cfm?id=IS-9783110223590-1] Berlin: Mouton de Gruyter. pp. 97-126.
  + The general principles of the Text Encoding Initiative P5 Guidelines (TEI P5; <http://www.tei-c.org/Guidelines/>) should be followed when in doubt.
  + In accordance with CELEX’s Dutch Linguistic Guide (http://www.ldc.upenn.edu/Catalog/docs/LDC96L14/dug\_a4.pdf; section 3.1.1) a distinction is made between immediate segmentation; flat, complete segmentation; and hierarchical, complete segmentation. For purposes of the current project, we provide flat, complete segmentations.
    - Immediate segmentation: [besigheid] [s] [besluit]
    - Flat, complete segmentation: [besig] [heid] [s] [be] [sluit]
    - Hierarchical, complete segmentation: [ [ [besig] [heid] ] [s] [ [be] [sluit] ] ]
  + For the sake of computational processing, we follow a morpheme-based (i.e. so-called Item-and-Arrangement) approach to morphology. Words are therefore analysed as arrangements of morphemes.
  + As a general principle we attempt to identify the smallest, meaningful unit, synchronically speaking (i.e. even if an analysis might have been possible from a diachronic perspective, but would not currently be recognised by the average, informed Afrikaans speaker, that analysis is not assigned). Therefore, words like *emansipeer* ‘emancipate’, gelagsaal ‘saloon’, and *gelukkig* ‘lucky’ are not analysed, but *geëmansipeer* ‘emancipated’, *gelagsaaltafel* ‘saloon table’, and *gelukkigheid* ‘luckiness’ are analysed as *[ge] [ëmansipeer]*, *[gelagsaal] [tafel]*, and *[gelukkig] [heid]*.
  + In principle and broadly speaking, only the following words are analysed:
    - Compounds, including:
      * Non-lexicalised compounds, i.e. compounds that are fully compositional, and where the meaning is therefore fully predictable from the constituents
        + tafelbladhout: [tafel] [blad] [hout] ‘table top wood’
      * Lexicalised compounds that are still fully compositional
      * Lexicalised and non-lexicalised particle verbs (also occurring in compounds or derivations)
        + afdwing: [af] [dwing]; afdwingbaar: [af] [dwing] [baar]
        + This also applies to particle verbs whose meanings are not fully compositional: aftree: [af] [tree]
        + This also includes cases that are splitable, although the verb part does not occur independently in other contexts: navors: [na] [vors], afsonder: [af] [sonder] (where [vors] and [sonder]=**RFV**).
        + **Note**: In cases where it is not a seperable verb, the string is not analysed: aanvaar, aansoek
    - Derivations, including:
      * Non-lexicalised derivations
      * Note: Derivations with non-compositional meanings are not analysed: *besigheid*
    - Inflected word-forms, limited to:
      * Lexicalised present and regular past participles where the meaning is fully compositional (i.e. the meaning can be predicted easily based on the verb and the participle forming morpheme(s)).
        + *bemoedigend: [bemoedig] [end];*
        + Not fully compositional cases are not analysed: *afgelope, bekend, besorgd/besorg, bykomend, geletterd, gemiddeld, gesyferd, getiteld, ingewikkeld, omvattend, ooglopend, vergange, volkome, volslae*
        + Highly lexicalised borderline cases are not analysed: *bestaande, lewend*
        + Lexicalised synthetic compoundsare left unanalysed: *belanghebbend, bogenoemde, bykomend, doeltreffend, gelyklopend, indrukwekkend, kommerwekkend, laasgenoemde, meegaande, ondergenoemde, onderliggend, onderstaande, ooreengekome, oorkoepelend, uitgelese, uitstaande, uitstekend, vanselfsprekend, voldoende, volgehoue, vooraanstaande, voorafgaande, weersinwekkend*
        + Suppletion forms are considered lexicalised, and are therefore not analysed. They are tagged as **RFA**.

*goed* ‘good’ *beter* *beste*

*naby* ‘near’ *nader* *naaste*

*baie* ‘many’ *meer* *meeste*

* + - * + Strong past participles are not analysed and are tagged as a whole as **Miscellaneous**.

*aangewese, aangeskrewe, bedorwe, beskore, gebroke, gestorwe, geswore, onafgebroke, onbesonge*

Note also that some compounds with strong past participles as constituent can also not be analysed if the strong past participle does not occur as independent word. Such words are also tagged as a whole as **Miscellaneous** (i.e. **MM**).

*agtergeblewene: [agtergeblewene]* (*geblewene* does not occur as an independent word)

*aangewese: [aangewese]* (in this context, *gewese* does not mean ‘ex’ as in *gewese eggenoot* ‘ex-husband’, but is rather the strong past participle form of *aanwys* ‘designate’)

*vasbeslote: [vasbeslote]*

*hoogaangeskrewe* (*aangeskrewe* does not exist without *hoog-*; it is therefore not analysed, but marked as **MM**)

*voorgeskrewe: [voorgeskrewe]*

* + - * Non-lexicalised inflected forms of nouns, verbs and adjectives:
        + Plurals (PL):
        + Diminutives (DIM):
        + Comparative forms of adjectives (CMPR):
        + Superlative forms of adjectives (SUP):
        + The attributive -e (ATTR):
        + The partitive genitive (PRTT-GEN):
        + The infinitive -e (INF):
        + The past tense (PST):
        + Participles (present and past) (PART):
        + Relicts from archaic expressions (tagged as **SFF**): *ten volle: [voll] [e], steen des aanstoots: [aanstoot] [s]*
  + Reduplications
  + Given the above, all lexicalised words (except the cases listed above) are therefore marked as **Lexicalised** and not analysed. A word is considered lexicalised when it appears as a defined or undefined lemma in:

Labuschagne, F.J. & Eksteen, L.C. 1993. *Verklarende Afrikaanse woordeboek*. Agtste uitgawe. Pretoria: J.L. van Schaik.

* + Where possible, give preference to analyses with stem allomorphy; affix allomorphy is only allowed in a restricted number of cases, as indicated in the tag set.
    - *rotte: [rott] [e]* (instead of *[rot] [te]*)
    - *administratiewe: [administratiew][e]* (where *[administratiew]*=**RDAA**)
    - *daklose: [dak]****[los]****[e]* (where [*los*]=**SDAM**)
    - *suksesvolle: [sukses]****[voll]****[e]* (where [*voll*]=**SDAP**)
    - *geëet: [ge] [ëet]* (where [*ëet*]=**RFVA**)
    - Note a few **exceptional cases**: *individueel: [individu] [eel]* (where preference is given to affix allomorphy (*[ueel] > [eel]*), due to amalgamation of identical phones).
    - In the case of allomorphs of DIM, all are analysed as different morphemes (and not as allomorphs): *-ie, -jie, -tjie, -etjie, -pie, -kie, -‘tjie,* etc.
    - In the case of allomorphs of PL, all are analysed as different morphemes (and not as allomorphs): *-e, -‘e, -s, -‘s,* etc.
    - Note analyses of cases like the following, where one might have given preference to analyses with affix allomorphy, but they are analysed as linking morphemes plus affixes:
      * *-der: bestuurder: [be][stuur][d] [er], verder: [ver] [d] [er]* (where *d*=**MLG**)
      * *-eling: drenkeling: [drenk][e][ling]* (where *e*=**MLG**)
      * *-elik: onhebbelik: [on][hebb][e][lik]* (where *e*= **MLG**)
      * *-erig: mooierig: [mooi][e][rig]* (where *e*= **MLG**)

## Guidelines for some specific cases

* + A small subset of arbitrary words is changed unsystematically when inflected (or in a limited number of cases in compounds), although the stem is still recognisable. In such cases, an allomorphic stem is identified, despite the fact that such a stem does not comply with the definition of allomorphic stems.
    - *breë: [bre] [ë]* (where *[bre]*=**RFAA**)
    - *breër: [bre] [ër]* (where *[bre]*=**RFAA**)
    - *droër: [dro] [ër]* (where *[dro]*=**RFAA**)
    - *hoë: [ho] [ë]* (where *[ho]*=**RFAA**)
    - *hoër: [ho] [ër]* (where *[ho]*=**RFAA**)
    - *lae: [la] [e]* (where *[la]*=**RFAA**)
    - *moeë: [moe] [ë]* (where *[moe]*=**RFAA**)
    - *vroeë: [vroe] [ë]* (where *[vroe]*=**RFAA**)
    - *goeie: [goei] [e]* (where *[goei]*=**RFAA**, and *[e]*=**SFA**)
    - *wye: [wy] [e]* (where *[wy]*=**RFAA**)
    - *daagliks: [daag] [liks]* (where *[daag]*=**RFNA**)
    - *daeliks: [da] [e] [liks]* (where *[da]*=**RFNA**)
    - *bedrae: [bedra] [e]* (where *[bedra]*=**RFNA**)
    - *lange: [lang] [e]* (where *[lang]*=**RFAA**; *[lank]* and *[lang]*=**RFA** when used independently)
    - *jonger: [jong] [er]* (where *[jong]*=**RFAA**; *[jonk]* and *[jong]*=**RFA** whenused independently)
    - *ouer: [ou] [er]* (where *[ou]*=**RFAA**)
    - *eerder: [eer] [d] [er]* (where *[eer]*=**RFAA**, and *[er]*=**SDA**)
    - *eers: [eer] [s]* (where *[eer]*=**RFAA**, and *[s]*=**SDB**)
    - *eerste: [eer] [ste]* (where *[eer]*=**RFAA**, and *[ste]*=**SDA[[1]](#footnote-1)**)
    - *derde: [der] [de]* (where *[der]*=**RFAA**, and *[de]*=**SDA**)
    - *dertien: [der] [tien]* (where *[der]*=**RFAA**)
    - *veertien: [veer] [tien]* (where *[veer]*=**RFAA**)
    - **NOTE**: The majority of cases (like *vroeë:* ***[vroe]*** *[ë]*) are considered allomorphs of their Germanic independent word correlates (i.e. *vroeg* in this example).
  + Given the general principles above, stems from the Germanic stratum that cannot occur as independent words, are analysed as allomorphs of the stems of the POS category where the stems occur most, or otherwise priority is given to **Verb**, then **Noun**, then **Adjective,** then **Adverb**, then **Other**.[[2]](#footnote-2)
    - *gaste: [gast] [e]* (where *gast*=**RFNA**)
    - *beduidend: [be] [duid] [end]* (where *duid*=**RFVA**)
    - *waarskuwing: [waarskuw] [ing]* (where *waarskuw*=**RFVA**) (also ***ruw****er,* ***stuw****ing,* ***sluw****er*)
    - *-moei-: bemoei: [be] [moei], vermoei: [ver] [moei], moeite: [moei] [te], moeilik: [moei] [lik]* (where *[moei]*=**RFVA**)
    - *-heem-: uitheems: [uit] [heem] [s]*
    - *-vred-: vrede: [vred] [e], bevredig: [be] [vred] [ig], vreedsaam: [vreed] [saam], tevrede: [te] [vred] [e]* (where *[vreed]* and *[vred]*=**RFNA**)
  + **NOTE**: In a small set of verbs, it is seemingly difficult to determine whether a string should be analysed as consisting of participle affixes, or otherwise as an allomorphic stem. In such cases, preference is given to the latter interpretation.
    - *uitgebreide: [uit] [ge] [breid][[3]](#footnote-3) [e]* (instead of *[uit] [ge] [brei] [d] [e]*, the reason being that the original verb in Dutch is *uitbreiden*, with its participle form *breidde uit*). Note that *[ge]*=**PFP**.
  + **NOTE**: Stems like *[hank]* in *afhanklik*, *[spronk]* in *oorspronklik, [gank]* in *toeganklik* and *[konink]* in *koninkie* are considered allomorphs; hence *[hank]*=**RFVA**, *[spronk]*=**RFNA ,** *[gank]*=**RFNA**, and *[konink]*=**RFNA**.
  + Only a limited set of linking morphemes from the Germanic stratum are recognised – see list below. Note that the hyphen (“-“) is seen as a linking morpheme from the Germanic stratum (i.e. **MLG**).
    - *-n-: buitengewone: [buite] [n] [gewon] [e]*
    - *-s-: fakulteitsraad: [fakul] [teit] [s] [raad]*
    - *-e-: perdekar: [perd] [e] [kar]*
    - *-en-: vertrouenswendel: [ver] [trou] [en] [swendel]*
    - *-es-: geestesoog: [geest] [es] [oog]*
    - *-er-: kinderwaentjie: [kind] [er] [wa] [en] [tjie]*
    - *-ens-: nooiensperskes: [nooi] [ens] [perske]; lydensgebeure: [lyd][ens] [gebeur] [e]*
    - *-der-: beendergestel: [been] [der] [ge] [stel]*
    - *-ere-: goederetrein: [goed] [ere] [trein]*
    - *-dere-: beendere-urn: [been] [dere] [-] [urn]*
    - *hyphen: see-eend: [see] [-] [eend]*
  + Compounds and derivations from the Classical stratum are not analysed into morphemes.
    - *administrasie*
    - *administratief*
    - *administreer*
    - *aktief*
    - *aktiwiteit*
    - *ambisieus*
    - *fusillade*
    - *kollegiaal*
    - *konsulaat*
    - *kontinent*
    - *kosmopolitaans*
    - *ministerieel*
    - *organies*
    - *organisasie*
    - *organisatories*
    - *planetarium*
    - *registreer*
    - *sentra*
    - *sentrum*
    - *stimulans*
    - *variable*
    - *variansie*
    - *variant*
* With regard to so-called confixes (also called **combining forms** or **neoclassical stems**; see Meesters, 2004: 61-62, or **tegnostamme** in Afrikaans (Combrink, 1990: 27)) (e.g. ***hidro****dinamika* or ***tegno****park* or***elektro****tegnies*), the question is whether such constituents should be analysed as affixes or as dependent stems. See Meesters (2004) for a lengthy discussion of all the issues related to these and other similar forms.
* When a confix (mostly from the Classical stratum) combines with an independent word (mostly from the Germanic stratum), the confix is analysed as a dependent stem(i.e. thelinking morpheme is analysed as part of the dependent stem).
  + *neurochirurg: [neuro] [chirurg]* (where *[neuro]*=**RD**)
  + *hidromeganika: [hidro] [meganika]* (where *[hidro]*=**RD**)
  + *psigometrie: [psigo] [metrie]* (where both parts are **RD**)
* Confixes are distinguished from affixes according to the following criteria (note that these are more tendencies than absolutes – Meesters, 2004: 63):
  + The status of the string in the source language (mostly Greek or Latin) determines the status in Afrikaans (i.e. if it is a stem in Latin, it is a dependent stem (i.e. confix) in Afrikaans; if it is an affix or particle in Greek, it is an affix in Afrikaans).
    - *proaktief: [pro] [akt] [ief]* (waar *[pro]*=**PDA**)
    - *elektrotegnies: [elektro] [tegnies]* (waar *[elektro]*=**RDA**)
  + Confixes combine with confixes and affixes to form independent words; affixes don’t combine with other affixes to form independent words.
    - *fonologie: [fon] [o] [logie]* (unanalysed in this corpus)
  + Generally speaking, confixes have more semantic weight than affixes. For example, a final confix form the semantic head of a neoclassical compound; it is not that easy to identify the same kind of criterium for initial confixes.
    - *hidrofobie: [hidro] [fobie]* (**fear** of water, where *[fobie]* is the semantic head)
  + As a test, one could also try to translate a potential confix/affix with a word/affix from the Germanic stratum. If the string is translated as a word, then the string is a confix; if it is translated as an affix, it is an affix.
    - In *neurofisiologie*, *[neuro]* is translated as *brein* (an independent word), and we can thus conclude that *[neuro]* is a confix.
    - In *sirkumposisie*, *[sirkum]* cannot be translated with an independent word, and we can thus conclude that *[sirkum]* is a prefix.
  + Initial confixes end on /i/, /o/ or schwa, while it is not a prerequisite for affixes.
    - Confixes: *Afro-, Euro-, Anglo-, Indo-, bio-, elektro-, etno-, gastro-, maksillo-, morfo-, neo-, neuro-, outo-, palato-, petro-, sosio-, strato-, tegno-, termo-*, etc.
    - Affixes: *ambi-, ante-, anti-, dis-, hemi-, hiper-, hipo-, i-/im-/in-/-ir-, inter-, intra-, kata-, post-,* etc.
  + When in doubt, use the list in Addendum (based on Combrink, 1990: 37-50, as well as HAT) as a guideline. Also consider the lists for English at:
    - http://en.wikipedia.org/wiki/List\_of\_Greek\_and\_Latin\_roots\_in\_English
    - http://www.prefixsuffix.com/rootchart.php
  + Words that are formed using zero derivation (i.e. cases where there is a change in POS category without any explicit morpheme) are analysed.
    - *bedoel: [be] [doel]* (as in “*wat bedoel is*”) (where *[be]*=**PFU**)
    - *bedoelde: [be] [doel] [d] [e]* (as in “*die bedoelde land*”) (where *[be]*=**PFU** and *[d]***=SFP**)
    - *verkorte: [ver] [kort] [e]* (where *[ver]*=**PDV** and *[e]***=SFA**)
  + Named entities (i.e. proper nouns, including names of languages) are not analysed, unless it is clear from the context that a morphological analysis is possible (e.g. compounds (including compounds of geographical locations – like dams, streets, plains, etc.), derivations and inflections).
    - *Johannesburg: [Johannesburg]*
    - *Rietvlei: [Rietvlei]*
    - **BUT:***Rietvleidam: [Rietvlei] [dam]*
    - **BUT:** *Kerkstraat: [kerk] [straat]*
    - *Suid-Amerika: [Suid-Amerika]*
    - *Afrikaans: [Afrikaans]* (name of language)
    - **BUT:***Nederlandse: [Nederland] [s] [e]*(e.g. *Nederlandse kaas* ‘Dutch cheese’), and *Nederlandse: [Nederlands] [e]* (e.g. *Nederlandse woorde* ‘Dutch words’)
    - *Jannie: [Jannie]* (diminutive form of a name, which is used frequently as proper name)
    - **BUT:***Gerhardjie: [Gerhard] [jie]*
    - **BUT:***Gerhards: [Gerhard] [s]*
  + Compounds with named entities as determiners are analysed, but the named entities are not analysed.
    - *Rietvlei-kruising: [Rietvlei] [-] [kruis] [ing]*
    - *Johannesburg-busboikot: [Johannesburg] [-] [bus] [boikot]*
  + With regard to words derived from proper names (person and place names):
    - Give preference to affix allomorphy (i.e. try to preserve the proper name).
      * *Afrikaans: [Afrika] [ans]*
      * *Amerikaanse: [Amerika] [ans] [e]*
      * **BUT**: *Comskiaanse: [Chomski] [aan] [s] [e]*(where *[Chomski]*=**RFNA**)
      * **BUT**: *Antarktiese: [Antarkt] [ies] [e]* (where *[Antarkt]*=**RFNA**)
      * **BUT***: Indiese: [Ind] [ies] [e]* vs. *Indiër: [Indië] [r]* (where *[Ind]*=**RFNA**)
    - Language names derived from place names (e.g. *Nederlands*, *Afrikaans*, *Yslands*) should be interpreted as ‘language spoken in X’, rather than ‘language spoken by X living in Y”. Hence, the language name *Afrikaans* should be analysed as *[Afrika] [ans]* (where *[ans]*=**SDAM**) , and not as *[Afrika] [an] [s]*.
    - Adjectives derived from proper names (e.g. *Suid-Afrikaans, Lutheraans*) follow the same principle as for language names (e.g. *[Suid] [-] [Afrika] [ans]* and *[Luther] [aans]*).
  + Words from other languages (e.g. *administration* (instead of *administrasie*)), including loan words (e.g. *gentleman, township*) are left unanalysed, and are tagged as a whole as **Miscellaneous** (i.e. **MM**).
  + Back-formations are tagged as a whole as **Miscellaneous** (i.e. **MM**).
    - *kompromie: [kompromie]* (also *kompromis*)
    - *demokraat: [demokraat]* (back-formation from *demokrasie*)
    - *boskaas: [boskaas]*
    - *rassis: [rassis]* (back-formation from *rassisme*)
    - *seksis: [seksis]* (back-formation from *seksisme*)
  + Particles in lexicalised and non-lexicalised particle verbs are annotated as independent stems.
    - *opskryf: [op] [skryf]* (where *[op]*=**RFO**)
    - **Note** that a couple of particle verbs should not be analysed, in cases where the verb cannot be interpreted as a proper verb in the context.
    - *afsonderlike: [afsonder] [lik] [e]* (in this context, *sonder* is not a verb with the meaning ‘without’)
    - *aanvaar* (here *vaar* does not refer to the regular senses of the verb)
  + The past tense morpheme in particle verbs is analysed as a prefix.
    - *op****ge****skryf: [op] [ge] [skryf]* (where *ge*=**PFV**)
  + Afrikaans has a large number of fixed expressions where relicts of inflection processes have been fossilised. Such relict morphemes are tagged as **SFF**.
    - *Engels des lig****s***
    - *God****s*** *water*
    - *Ten voll****e***
    - *Om den brod****e***
  + The class PDS can only be assigned to prefixes that are not category determining, e.g. *ge-be-her-er-ont-ver* as **PDS** can only attach to verbs, or *on-anti-a-il-in-ir*as **PDS**  can only attach to adjectives.
    - *besoek* (V)*: [be] [soek]*
    - *behulp* (N) (soos in *met behulp van* of *behulpsaam*): *[be] [hulp]*
    - *onmoontlik* (A): *[on] [moontlik]*
    - *ondier* (N): *[on] [dier]*
    - **NOT**: *beoog: [be] [oog]*, since *oog* is a noun; in this case, *be*=**SDV**)
  + With the exception of the class of affixes marked as **PDS**, if an affix (mostly prefix) is not category determining, assign the POS category within the context.
    - *Mikro-biologie* (*[mikro]*=**PDN**)
    - *Mikro-biologiese* (*[mikro]*=**PDA**)
  + Where a stem could be interpreted as either an adjective or adverb (typically indicated in a dictionary like HAT as ***b.nw., bw.***), an adjective interpretation is assigned; the same applies suffixes that could be either an adjectivaliser or adverbialiser (e.g. in cases of conversion), where preference is given to an adjectivaliser interpretation.
    - *besonder*=**RFA**, even if the POS tag is **Adverb** in the context of the sentence, since besonder can be either **Adjective** or **Adverb**.[[4]](#footnote-4)
    - *weekliks: [week] [liks]* (where *liks*=**SDA**)
    - *aansienlik: [aan] [sien] [lik]* (as in *Hy is aansienlik langer as sy* ‘He is significantly taller than she’, where *[lik]*=**SDA,** even though the POS of *aansienlik=***Adverb**)
    - Note that polysemous cases in a dictionary, i.e. where meaning 1 is marked as an adjective, and meaning 2 as an adverb, the appropriate tag (**RFA** or **RFB**) in the context of the sentence should be assigned, e.g. in “*Die ewige bakleiery*” *[ewig]*=**RFA**; in “*Vir ewig en altyd*” *[ewig]*=**RFB**.
  + Numerals (e.g. *een* ‘one’ and *tien* ‘ten’ and “10” and “234”) are considered adjectives (i.e. **RFA**).
    - *miljoen* (‘million’) and *miljard* (‘milliard’) are considered stems (i.e. **RFA**), since they form the basis for the formation of other large-number names. Hence, and rather oddly, *[ljoen]* (in *biljoen*) and *[ljard]* (in *triljard*) are considered allomorphs of *miljoen* and *miljard* respectively (hence **RFAA**, and *[bi]* and *[tri]* are **RD**).
  + When names of directions (e.g. *noord* ‘north’ and *suid* ‘south’) are used as first constituent in place names (e.g. *Noord-Korea* ‘North Korea’ or *Suid-Afrika* ‘South Africa’), they are considered adverbs (i.e. **RFB**).
  + Affixoids should be treated with care, since it is quite a contentious category. If a string is deemed an affixoid, it should be tagged as an affix (e.g. **SDA**).
    - *loodarm: [lood] [arm]* (where *[arm]*=**SDA**)
    - *hondsiek: [hond] [siek]* (where *[hond]*=**PDA**)
  + With regards to acronyms:
    - When in doubt whether an acronym is a letter name word or a letter sound word, choose **letter sound word** as default.
    - In mixed cases, mark it as **letter name word**: *AsgiSA, AgriSA*

### Technical specifications

* All documents and annotation layers should be saved in UTF-8 format.

### Authoritative references

The following language-specific works of reference should be considered authoritative when in doubt:

* Odendal, F.F. & Gouws, R.H. (reds.). 2005. *Verklarende handwoordeboek van die Afrikaanse taal*. Vyfde uitgawe. Pinelands: Maskew Miller Longman.
* Taalkommissie. (samest.). 2009. *Afrikaanse Woordelys en Spelreëls*. Tiende uitgawe. Kaapstad: Pharos Woordeboeke.
* www.etymonline.com (mostly for analyses of words from the Classical stratum)
* etymologiebank.nl (mostly for analyses of words from the Germanic stratum)

# Tag set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **POS** | **LEVEL1** | **LEVEL2** | **CLASS** | **EXPLANATION** |
|  | stemDep |  | RD | Dependent stem: ***elektr****isiteit,* ***administr****asie,* ***sekretar****iaat,* ***hidro****-****logie****,* ***bio****diesel* |
|  | stemDepAllo |  | RDA | Dependent stem - Allomorph: ***aks****ie, demo****krat****ies, leksiko****graf****iese* |
| **Noun** | stemFull | N | RFN | Full stem - Unchanged: ***hond****e;* ***tafel****tjie* |
| **Noun** | stemAllo | N | RFNA | Full stem - Allomorphic: ***stow****e;* ***wolv****in;* ***gast****e;* ***Ind****iese* |
| **Noun** | stemMin | N | RFNM | Full stem - Allomorphic minus: ***bot****e;* ***pal****e* |
| **Noun** | stemPlus | N | RFNP | Full stem - Allomorphic plus: ***katt****e;* ***sakk****ie* |
| **Verb** | stemFull | V | RFV | Full stem - Unchanged: *ge****werk****;* ***werk****loos* |
| **Verb** | stemAllo | V | RFVA | Full stem - Allomorphic: *naam****gew****ing; aan****duid****ing; onder****ligg****end* |
| **Verb** | stemMin | V | RFVM | Full stem - Allomorphic minus: *ont****nem****ing* |
| **Verb** | stemPlus | V | RFVP | Full stem - Allomorphic plus: ***plakk****er* |
| **Adjective** | stemFull | A | RFA | Full stem - Unchanged: ***mooi****ste;* ***lelik****e* |
| **Adjective** | stemAllo | A | RFAA | Full stem - Allomorphic: ***styw****e;* ***groww****e;* ***dro****ë,* ***gou****e,* ***sagt****e* |
| **Adjective** | stemMin | A | RFAM | Full stem - Allomorphic minus: ***final****e* |
| **Adjective** | stemPlus | A | RFAP | Full stem - Allomorphic plus: ***sagg****ies* |
| **Other** | stemFull | O | RFO | Full stem - Unchanged: ***aan****beveel*; ***oor****gee;* ***25****-jarige* |
| **Other** | stemAllo | O | RFOA | Full stem - Allomorphic: *daar****mee, ower****priester* |
| **Other** | stemMin | O | RFOM | Full stem - Allomorphic minus: ??? |
| **Other** | stemPlus | O | RFOP | Full stem - Allomorphic plus: ***inn****ig* |
| **Adverb** | stemFull | B | RFB | Full stem - Unchanged: ***saam****bind,* ***Suid****-Afrika* |
| **Adverb** | stemAllo | B | RFBA | Full stem - Allomorphic: *ge****sament****lik, ge****noeg*** |
| **Adverb** | stemMin | B | RFBM | Full stem - Allomorphic minus: ***sam****espel* |
| **Adverb** | stemPlus | B | RFBP | Full stem - Allomorphic plus: ??? |
| **Acronym: Letter sound word (RK)** | stemRed | RK | RRRK | Reduced stem: ***Sanlam****-gebou;* ***Avbob****-werknemer;* ***AgriSA****;* ***AsgiSA*** |
| **Acronym: Letter name word (RL)** | stemRed | RL | RRRL | Reduced stem: ***ATKV****-lid;* ***WNNR****-projek* |
| **Abbreviation (RA)** | stemRed | RA | RRRA | Reduced stem: ***b.v.p.****-beslissing* |
|  | sufDer | NR | SDN | Derivation suffix - Nominaliser: *rooi****e****; belangrik****heid****; verkop****ing*** |
|  | sufDer | VR | SDV | Derivation suffix - Verbaliser: *kontrakt****eer*** |
|  | sufDer | ADJR | SDA | Derivation suffix - Adjectivaliser: *gele****rig****; skouspel****agtig*** |
|  | sufDer | ADVR | SDB | Derivation suffix - Adverbialiser: *sagg****ies****; wys****lik****; eer****s*** |
|  | sufDer | FixedX | SDF | Derivation suffix - Fixed Expressions: *sy****ne****, jou****ne****, my****ne****, lende****ne*** |
|  | sufDerAllo | ADJR | SDAA | Derivation suffix - Allomorphic: *administrat****iew****e* |
|  | sufDerAllo | NR | SDNA | Derivation suffix - Allomorphic: *Afrika****n****er* |
|  | sufDerPlus | NR | SDNP | Derivation suffix - Allomorphic plus: *held****inn****e; agent****skapp****e* |
|  | sufDerPlus | VR | SDVP | Derivation suffix - Allomorphic plus: ??? |
|  | sufDerPlus | ADJR | SDAP | Derivation suffix - Allomporphic plus: *sukses****voll****e,* |
|  | sufDerPlus | ADVR | SDBP | Derivation suffix - Allomorphic plus: ??? |
|  | sufDerMin | NR | SDNM | Derivation suffix - Allomorphic minus: *onaangenaam****hed****e; ampten****ar****e; Luther****an****e; fabrik****at****e* |
|  | sufDerMin | VR | SDVM | Derivation suffix - Allomorphic minus: *assess****er****ing* |
|  | sufDerMin | ADJR | SDAM | Derivation suffix - Allomorphic minus: *geweten****los****e, aanvaar****bar****e, provinsi****al****e, finansi****ël****e, buig****sam****e* |
|  | sufDerMin | ADVR | SDBM | Derivation suffix - Allomorphic minus: ??? |
|  | sufFlex | PL | SFM | Inflection suffix - Plural: *hond****e****; katt****e*** |
|  | sufFlex | DIM | SFD | Inflection suffix - Diminutive: *hond****jie****; kat****jie*** |
|  | sufFlex | F | SFV | Inflection suffix – Feminine: kelner**in**; bestuurder**es** |
|  | sufFlex | CMPR | SFC | Inflection suffix - Comparative: *mooi****er****; lelik****er*** |
|  | sufFlex | SUP | SFS | Inflection suffix - Superlative: *mooi****ste****; lelik****ste*** |
|  | sufFlex | ATTR | SFA | Inflection suffix - Attributive: *interessant****e***; *lelik****e*** |
|  | sufFlex | PRTT-GEN | SFG | Inflection suffix - Partitive Genitive: *iets mooi****s****; iets lelik****s,*** *onses insien****s*** |
|  | sufFlex | PART | SFP | Inflection suffix - Participle: *meganiser****end,*** *bepaal****d****, beperk****t****e* |
|  | sufFlex | INF | SFI | Inflection suffix - Infinitive: *iets te drink****e****; iets te et****e*** |
|  | sufFlex | FixedX | SFF | Inflection suffix - Fixed Expressions: *ten volle, ter will****e****, ten aanhor****e****, ons****es*** *insiens* |
|  | prefDer | NR | PDN | Derivation prefix - Nominaliser: ***eks****vrou,* ***aarts****biskop,* ***sub****kategorie,* ***super****vrou* |
|  | prefDer | VR | PDV | Derivation prefix - Verbaliser: ***be****moeilik;* ***ver****afrikaans* |
|  | prefDer | ADJR | PDA | Derivation prefix - Adjectivaliser: ***semi****-produktief,* ***hiper****sensitief,* ***intra****molekulêr* |
|  | prefDer | ADVR | PDB | Derivation prefix - Adverbialiser: ***ge****noeg,* ***n****ooit* |
|  | prefFlex | PST | PFV | Inflection prefix - Past: ***ge****speel;* ***ge****klim* |
|  | prefFlex | PART | PFP | Inflection prefix - Participle: ***ge****trou,* ***ge****kruis* |
|  | prefFlex | UNIQ | PFU | Inflection prefix - (***ge-be-her-er-ont-ver-wel*** *and* ***anti-a-on-il-in-ir-wan-nie***): Inchoative: ***ont****vlam*; Intensive: ***ver****slaap*; Iterative: ***her****doop*; Transitive: ***ver****dink;* Negative: ***a****sosiaal* |
|  | ambiDer | NR | ADN | Derivation ambifix - Nominaliser: ***ge****berg****te, be****send****ing, be****ampt****e*** |
|  | ambiDer | VR | ADV | Derivation ambifix - Verbaliser: ***be****vred****ig****,* ***ver****ou****der****,* ***ver****klein****eer, veront****skuld****ig*** |
|  | ambiDer | ADJR | ADA | Derivation ambifix - Adjectivaliser: ***ge****dienst****ig****,* ***be****hoor****lik****,* ***on****will****ig****,* ***on****verset****lik****,* ***veront****waard****ig*** |
|  | ambiDer | ADVR | ADB | Derivation ambifix - Adverbialiser: ***be****tyd****s****,* ***on****verwyl****d****,* ***on****verwag****s*** |
|  | ambiFlex | PART | AFP | Inflection ambifix - Participle: ***ge****meganiseer****d*** *,* ***ge****rig****t****e* |
|  | morphCran |  | MC | Cran morpheme: ***boe****pens;* ***tjoep****stil* |
|  | morphLink | GER | MLG | Linking morpheme: *besigheid****s****plan; wild****e****bees; bestuur****d****er; drenk****e****ling; mooi****e****rig* |
|  | morphLink | CLAS | MLC | Linking morpheme: *akt****iwi****teit; elektr****isi****teit; elektr****o****magneet; elektr****ifi****seer; kur****at****or; profession****al****isme* |
|  | morphSym |  | MS | Symbols/characters: *$, R, &* |
|  | morphMisc |  | MM | Used for any string that is not covered by the above set of tags. |

As an example, the word *indiensnemingskontrakte* will be analysed and tagged as follows:

**in**<RFO>

**dien**<RVF>

**s**<SDN>

**nem**< RMV >

**ing**<SDN>

**s**<ML>

**kontrakt**<SD>

**e**<SFM>

# Changing the protocol

Since this protocol is considered to be a dynamic document, it is subject to change from time to time. For purposes of good governance, the following principles and procedures should be followed:

* The language expert (in consultation with the project leader) has the final say in changing this procedure document, and is the only authoritative person who might change the protocol.
* If an annotator (or other project member) wishes change a specification, s/he should log a request on Bugzilla, indicating the nature of the change request. The language expert then considers the request, and adds the change to the protocol document. The protocol document is saved with a new version number. All other annotators are then informed of the change, and requested to change previous annotations, where applicable.
* If an annotator wishes to add specific examples to the protocol, s/he sends an email to the language expert with the suggested examples. The language expert then adds the examples to this document, and when a significant number of examples have been added, saves the document with a new version number. All other annotators are then informed of the new version of the protocol document.

## Roles and contact details

The following roles in this project can be identified:

* Project leader: Responsible for successful technical delivery of project.
  + Martin Puttkammer  
    [martin.puttkammer@nwu.ac.za](mailto:martin.puttkammer@nwu.ac.za)
* Project manager: Responsible for timeous delivery of project, within budget. If you suspect that you will not be able to deliver on a set milestone, contact the project manager.
  + Ané Bekker  
    [ane.bekker@nwu.ac.za](mailto:ane.bekker@nwu.ac.za)
* Language expert: Responsible for language specific requirements and specifications. This person also controls the protocol document.
  + Afrikaans: Gerhard van Huyssteen  
    gerhard@trifonius.co.za

## References

Booij, G. 2002. *The Morphology of Dutch*. Oxford: Oxford University Press.

Combrink, JGH. 1990. *Afrikaanse morfologie*. [Afrikaans morphology]. Pretoria: Academica.

Meesters, G. 2004. *Marginale morfologie in het Nederlands: paradigmatisches samestellingen, neoklassieke composite en splintercomposita*. [Marginal morphology in Dutch: paradigmatic compounds, neo-classical compounds, and portmanteau words/blends]. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.

Smeaton, W. 1843. *Etymological Manual of the English and French Languages*. New Haven: JP Hart.

[WNT] Woordenboek der Nederlandsche Taal. Available at: http://gtb.inl.nl.

1. Also in *oorkant****ste****, linkerkant****ste****, onder****ste****, middel****ste***, etc. [↑](#footnote-ref-1)
2. To broad test whether a certain string is a potential allomorphic stem from the Germanic stratum, add *-en* to the string; if the resultant string is found as a verb in a Dutch dictionary, the original stem could be considered a dependent stem, e.g. *duid+en=duiden* (which is a verb in Dutch), or *gast+en=gasten* (which is a plural noun in Dutch). [↑](#footnote-ref-2)
3. Also in *uitbreiding, uitgebreidheid, verbreiding*, etc. [↑](#footnote-ref-3)
4. The rationale for this decision is based on the need to easily find cases of conversion: by searching in the corpus for words with POS tag X but morphologically determining tag Y, one could find cases of conversion. [↑](#footnote-ref-4)