Stellenbosch Papers in Linguistics Plus, Vol. 44, 2014, 49-70
doi: 10.5842/44-0-644

The sequence and productivity of Setswana verbal suffixes

Rigardt Pretorius

School of Languages, Potchefstroom Campus, North-West University
E-mail: rigardt.pretorius@nwu.ac.za

Abstract
Setswana is an agglutinative language with a rich verbal morphology, allowing for an elaborate system of verbal inflection. Until now, research on Setswana verbal morphology has largely been based on qualitative methods. This paper discusses the frequency of use and the sequencing of Setswana verbal suffixes, based on statistics extracted from the 67,284 orthographic-unit, annotated NCHLT Setswana corpus which includes 9,146 verbs. On this quantitative basis, the relationship between productivity/frequency and the position/slot of Setswana verbal suffixes is investigated. In addition, the relationship between the frequency and position of these same Setswana verbal suffixes and their inflectional or derivational nature is also considered. The data is subsequently used to evaluate and comment on existing descriptive grammars of Setswana.

Keywords: Setswana, morphology, verbal suffixes, tokenisation, sequence, productivity

1. Introduction and theoretical background

Setswana is a Bantu language that appears in the South-Eastern zone of Guthrie's (1971) zonal topogram. Guthrie numbers the three Sotho languages as Setswana (S 31), Northern Sotho (S 32) and Southern Sotho (S 33). Krüger (1994, 2006), whose views are largely based on the work of Van Wyk (1967) and Lombard, Van Wyk and Mokgokong (1985), identifies seven word classes in Setswana, where only the nouns and verbs are considered to be open classes. Setswana is generally referred to as an agglutinative language and, like the other languages in the Sotho-Tswana group, it has a rich verbal morphology allowing for the stringing together of several morphemes. Furthermore, these languages are characterised by a disjunctive orthography affecting mainly verbs; thus a single linguistic word (verb) may be represented by a number of orthographically separate units (Krüger 2006:12-28, Kosch 2006:3).

For quite some time, there has been considerable theoretical interest in multiple affixation in languages in general, and in Bantu languages like Setswana in particular, especially with regard to the verb. On a theoretical level, several explanations have been put forward in order to account for the order in which affixes occur in agglutinative languages. For example, Bybee (1985) attributes the most widely attested orders to the semantic function and scope of each affix. This implies that affixes with a greater "relevance" to the action referred to by the verbal

root will appear closer to it. The use of a particular verbal extension is thus dependent on its semantic compatibility with the entire preceding verbal stem. The use of the reciprocal suffix *-an-* in Setswana is a good example. While it is not compatible with the verbal stem *kwala* ('write'), as in **re a kwalana* ('we write each other'), when this suffix is considered in relation to the verbal stem *kwalela* ('write to/for'), then the reciprocal meaning is compatible, as in *re a kwalelana* ('we write to each other'). Other theoretical approaches include that of Baker (1985) who, in contrast to Bybee (1985), uses syntactic criteria to explain affix-ordering via a so-called "mirror principle", and Hyman (2002) who differs from both Bybee and Baker in arguing for morphology-internal reasons for such affix-ordering. Hyman relies on a Pan-Bantu template Causative, Applied, Reciprocal, Passive (CARP) for determining the order of affixes in Bantu.

Of more importance for this article is the position of Kosch (2006, 2007) who discusses the inflectional and derivational morphology of the Sotho languages, focusing mainly on Northern Sotho. She indicates several criteria for the distinction between inflectional and derivational morphemes and stresses the fact that this distinction should be treated as language-specific and not as a discreet distinction, i.e. she proposes a sliding scale between derivation and inflection. Most importantly, however, she highlights the relationship between the productivity of a suffix and its position relative to the root.

2. Affix-ordering in the Sotho languages

While the traditional grammars of Cole (1955), Krüger (1994, 2006) and other related literature indeed deal with affixation, no corpus-based study on the order of Setswana verbal affixes has yet been attempted. There have, however, been such attempts in relation to other Sotho languages. Anderson and Kotzé (2008) and Kotzé (2011), for example, present a matrix for Northern Sotho affixes based on an analysis of 458 basic verb-stem entries from the *Concise Northern Sotho Dictionary*.

The aim of this paper is thus to provide the first corpus-based analysis of the order of Setswana verbal suffixes based on their productivity (which is understood as equivalent to frequency of use) and position/slot. This analysis is based on data from the first Setswana corpus which was recently annotated. The analysis based on this corpus is then used as a critical basis from which to comment on and expand the existing literature on Setswana verbal morphology. In what follows below, existing views on Southern Bantu and Setswana verbal morphemes will be considered first. For Setswana, the focus is on the views of Krüger (1994, 2006), and for Northern Sotho the views of Kosch (2006, 2008), Anderson and Kotzé (2008) and Kotzé (2011) will be the main focus. Subsequently, and more generally, the abovementioned theoretical frameworks dealing with the sequencing of verbal suffixes will be examined. Then, a brief overview of two corpus-based projects dealing with the morphological analysis of Setswana will be provided, after which the Setswana corpus data will be presented and interpreted. The paper will conclude with an analysis of how current theories and models align with this new data, and a number of proposals for amendment will be put forward.

3. Setswana verbal morphology

Krüger (1994, 2006) elaborates on a hierarchical approach to the analysis of morphologically complex words in Setswana that he introduced as far back as the late 1960's (cf. Krüger 1967, 1968, 1976). His view is that a neutral or end-constituent analysis does not consider the mutual

relations between the different morphemes that constitute the word. He distinguishes between "grammatical morphemes", which can either be inflectional or derivational, and "lexical morphemes", which are roots and stems. He views grammatical morphemes as inherent meaningful parts of a word which exist only by the grace of the word's form and meaning. The semantic values of grammatical morphemes are dependent on the word as they can only be activated by the meaning of a stem or root.

Unlike earlier scholars, Krüger (1994, 2006) accounts for the prefixes and suffixes as determinants of the stems of words in Setswana. Thus, like Dik and Kooij (1970) and Stageberg (1971), Krüger employs morphological as well as semantic features in determining stems. For Krüger, a stem is that part of a morphologically complex word that has a word-correlate and which may include one or more grammatical morphemes. He agrees with Lyons (1990:59) when he differentiates between roots and stems as follows: "The difference between stems and roots is that roots are morphologically unanalyzable, whereas stems may include in addition to their root one or more derivational affixes"[1]. Table 1 on the next page is Krüger's (2006:257) presentation of the morphemes of the Setswana verb. Here, he predicts morpheme sequences in the linear arrangement in row A, details the morphological items in row B, and outlines the hierarchical arrangement from lowest (1) to highest rank (15) in row C.

[1] Pretorius (2000) gives a more detailed discussion of his and others' views on roots and stems.

Table 1. Krüger's (2006) schematic representation of the morphological structure of Setswana verbs

A. Linear arrangement

B. Morphological items

C. Hierarchical arrangement from lowest rank (1) to highest (15)

									Mutually exclusive					**Mutually inclusive**					
	Prefixes								**Semi-productive suffixes**					**Productive suffixes**					**Ending**
A	**Inf. morph.**	**Neg. morph.**	**Subj. conc.**	**Neg. morph.**	**Asp. morph.**	**Temp. morph.**	**Obj. conc.**	**Root**	**Rev. intr.**	**Rev. tr.**	**Den.**	**Neut.**	**Iter.**	**Caus.**	**Appl.**	**Rec.**	**Perf.**	**Pass.**	**Neg./ Mod.**
B	go	ga-	Non-loc. Loc. go- Inf. go- CL1. e-	-se- -sa- -a-	-a- -sa- -ka-	-tla-	Non-loc. Loc. -e- Inf. -go Refl. -i-	Orig. den. Idio.	-og- -olog-	-ol- -olol-	-f- -fal-	-êg- -al- -agal- -êsêg- -agan-	-ak-	-is- -y-	-êl- -êts-	-an-	-il-	-w- -iw-	-a -e -ê -ng (rel.) -ng (imp.)
C	6a	1	6 > 6a	1	2	3	4		15	14	13	12	11	10	9	8	5	7	6 (mod.) 1 (neg.)

In Krüger's hierarchical analysis of Setswana verbs, verbal morphemes are analysed/removed one-by-one on the basis of a hierarchy that he has assigned to them (refer to the numbers in brackets in Figure 1 below). The morpheme with the smallest number is reduced first, the rationale being that the remaining stem must still be a meaningful word. As the root is taken as the reference point, prefixes are removed to the left while suffixes are removed to the right.

In the example in Figure 1 below, the basic stem, which is also the verbal lemma[2] in this case, is *go apaya* ('to cook'). When the basic stem is analysed, the root *-apay-* is reached.

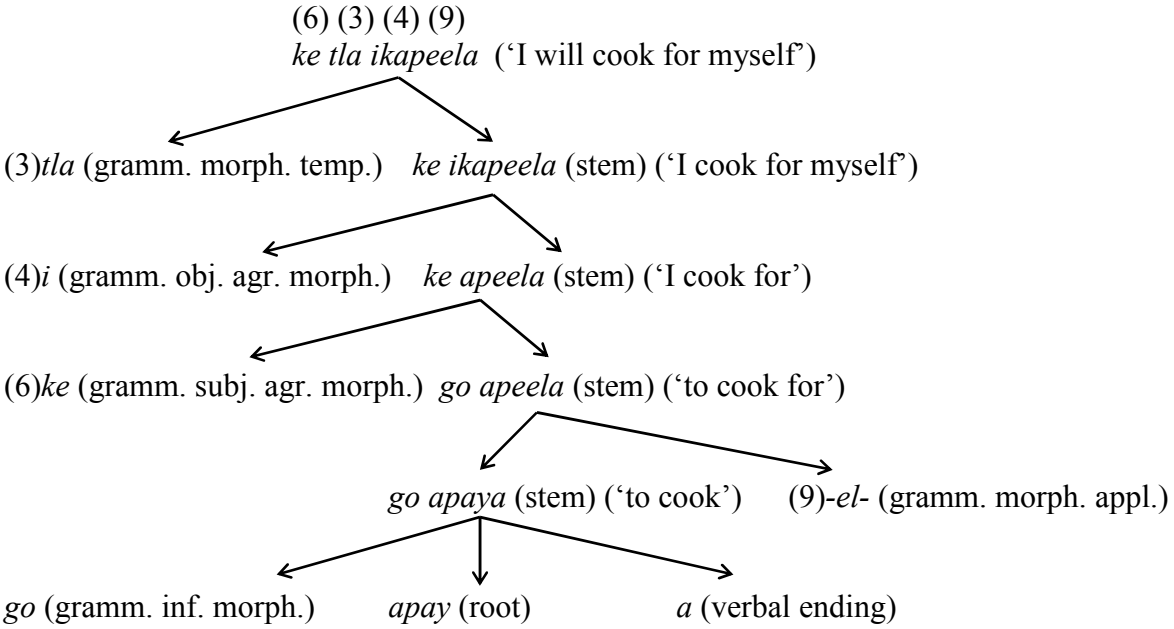

Figure 1. An example of a hierarchical analysis of a Setswana verb

Alternatively, a linear morphological analysis is possible and the glosses used in the Xerox and NCHLT Projects are shown below. More detail regarding these projects is provided in §4.1; at this point, the glosses below are simply meant to illustrate an alternative linear analysis.

In the Xerox Project, morphemes are divided by + and identified with the text directly below each morpheme, as in (1):

(1) *Ke tla ikapeela (dijo).* ('I will cook (food) for myself')
ke + tla + i + apay + el- + -a
AgrSubjP1sg + Fut + Refl + cook + Appl + VerbEnd

In the NCHLT Project, each morpheme is preceded by $ and then categorised in square brackets, as in (2):

(2) *Ke tla ikapeela (dijo).* ('I will cook (food) for myself')
$ke[csP1]$tla[temp]$i[ref]$apay[vr]$el[app]$a[ve]

[2] Brits (2006) gives a more elaborate discussion of the verbal lemma in Setswana.

These linear analyses highlight the importance of the question regarding the ordering of the relevant affixes.

3.1. Reasons for the linear order of suffixes

Two prominent views on the sequence of verbal suffixes are (i) that the order of extensions is determined by the main processes of word formation, namely inflection and derivation, and (ii) that the order of extensions in the Bantu languages is the result of a Pan-Bantu default template. Thus, some researchers favour compositional reasons while others claim a fixed order.

Hyman (2002) argues that affix-ordering, or at least certain aspects thereof, is directly determined by purely morphological reasons. Languages are therefore able to impose specific morphotactic constraints for which there is no synchronic extra-morphological explanation. Booij (2005:68) agrees with this when he states that "[t]he word formation templates of a language define the set of possible complex words of that language". Bybee (1985) and Baker (1985), on the other hand, hold the opinion that morphology-external factors result in the order of morphemes. Kosch (2006, 2008), consulting Bybee (1985), discusses the inflectional and derivational nature of morphemes and their role in word formation in Northern Sotho, while Krüger (1994, 2006) discusses word formation and analysis in some detail for Setswana. Booij (2005:67) states that the use of the notion of 'productivity' presupposes the idea of rule-governed morphological creativity and that the aforementioned notion is relevant in the domains of inflexion, derivation and compounding. For the purposes of this paper, "productivity" does not refer to the ability to create new words, but rather the ability to combine with stems with suitable syntactic and semantic features.

According to Booij (2005:71), "[g]iven the distinction between derivation and inflexion (derivation creates lexemes, inflexion creates forms of lexemes)", the following schema should apply to the order of affixes:

Inflectional prefixes	Derivational prefixes	Root	Derivational suffixes	Inflectional suffixes

For Setswana, Krüger (1994:18) agrees with this classification when he states that "structurally derivational morphemes tend to occur more centrally in the word form than inflectional morphemes". According to this postulation, the verbal extensions should then be evaluated as derivational morphemes, whereas the modal endings should be evaluated as inflectional morphemes.

3.2. Inflection versus derivation

Kosch (2007:1) indicates that "[i]nflection and derivation are two morphological processes that continue to intrigue and fascinate researchers because of their elusive nature". She also states that "[i]n investigating the interface, the obvious place to start is the characteristics which are accepted as generally valid for the two types" (Kosch 2007:1). The tests for the distinction between the processes of inflection and derivation are popularly based on the criteria described in §3.2.1 to §3.2.3.

On the differences between derivational and inflectional morphology, Kosch (2006) quotes Bybee (1985:87) when stating that "there is not necessarily a discreet distinction between inflection and derivation […] the differences between the two types of morphology are just a matter of degree". Kosch (2006) discusses the derivational and inflectional characteristics of a number of Northern Sotho verbal extensions. She states that a discussion of the collective criteria would be more reliable than exploiting only selected criteria.

In attempting to indicate reasons for the positions of verbal suffixes, three of these criteria will be attended to briefly in the following sections: change in lexical meaning (§3.2.1), productivity (§3.2.2), and proximity to the root (§3.2.3).

3.2.1. Change in lexical meaning

According to Bybee (1985:17), a form must have a meaning that is communicatively useful enough to ensure a high frequency of occurrence. A morpheme will be inflectional if it has a regular meaning and minimal semantic content. This allows it to combine with a large number of appropriate items in a particular category. For example, the Setswana past-tense suffix *-ilê* (or one of its allomorphs) can occur with most verb stems. Derivational morphemes, on the other hand, tend to have varying, somewhat inconsistent and idiosyncratic meanings in combination with different lexical items. As such, these morphemes are more likely to have lexical restrictions on their applicability; for example, the reversive suffix *-oll-* in Setswana applies only to verbs that are inherently reversible, and the reciprocal suffix *-an-* only applies to transitive verbs or lexically compatible stems as explained in the introduction in §1.

In Setswana, as in Northern Sotho, the causative suffix *-is-/-y-* is a morpheme which, it could be argued, appears at the interface between inflection and derivation. It has a predictable meaning, namely 'cause or assist someone to do something', and it can be attached to a large number of verbs. By this reasoning, this suffix would then be inflectional, yet it is described as derivational in African language morphology. This is probably because this morpheme may have a radical effect on the meaning of the resulting verb.

3.2.2. Productivity

In discussing productivity, Kosch (2006) points out that inflectional morphemes are more productive than derivational morphemes. In this regard, Bauer (1992:13) states that "if you can add an inflectional affix to one member of a class, you can add it to all members of the class, while with a derivational affix, it is not generally possible to add it to all members". Inflectional affixes are highly productive and have a relatively predictable distribution, while derivational affixes have limited productivity and a more restricted distribution. In Setswana, the modal suffix *-a*, which was categorised in Figure 1 as a verbal ending, can be attached to the majority of verb roots or extended verb roots. This suffix is therefore an inflectional affix. On the other hand, suffixes such as the positional suffix *-am-* and dispersive *-al-* are restricted by semantic considerations and only appear with a limited number of verb roots. There are also a number of suffixes which have become fossilised with the verbal root, such as the contactive suffix *-ar-* in *apara* ('to wear/put on') and *sikara* ('to carry something on your back'). Thus, according to the productivity principle, the positional and contactive suffixes would be more derivational than inflectional.

3.2.3. Proximity to the root

In cases where both inflectional and derivational affixes co-occur, the derivational affixes generally (but not necessarily) occur closer to the root than the inflectional ones. Compare the following examples:

(3) *-kwalotse* ('had rewritten')
root (*-kwal-*) + derivational morpheme (*-olol-*)(reversive transitive + inflectional morpheme (*-il-*)(perfect))

(4) *-bofologilê* ('became loose')
root (*-bof-*) + derivational morpheme (*-olog-*)(reversive intransitive + inflectional morpheme (*-il-*)(perfect))

Kotzé (2011) employs Bybee's (1985) criterion of lexical generality to seek an explanation for the ordering of Northern Sotho verbal extensions. This approach, which is also more data-centred, is one that considers the relevance of the affix to a stem. "Relevance" refers to the extent to which the meaning of an affix alters or affects the meaning of a stem. (The degree to which a morpheme will semantically affect a stem will determine its relative proximity to the stem.) Commenting on the verb, Bybee (1985:211) states that "the more a concept has to do with the content of a verb, the closer it will occur to the verb stem".

As mentioned in §3.2.2, there are a number of verbal suffixes in Setswana which affect the content of a verb to such an extent that they have become fossilised to the root, as in (5) and (6) below:

(5) *-apa* > *-ap**ar**a* ('clothe, get dressed')
(6) *-phatla* > *-phatl**alal**a* ('disperse, scatter, lie stretched out')

Note that the basic roots to which these suffixes were attached no longer exist in the language.

4. Computational morphological analysis in Setswana – the source of the data

In light of the different views on Bantu and Setswana verbal morphology detailed in the previous section, the theoretical utility of a corpus-based analysis of Setswana verbs becomes obvious. This section is focused on describing the outlines of such an attempt.

This article emanates from previous work which focused on developing a rule-based morphological analyser for Setswana using Xerox finite-state tools (this project henceforth referred to as "the Xerox Project") and, more recently, work done as part of a project for the National Centre for Human Language Technology (this work henceforth referred to as "the NCHLT Project"). The latter project, from which the data for this article is derived, aims to develop text resources for 10 South African languages. These resources are managed by the Resource Management Agency (____).

The NCHLT Project involved the manual tokenisation, lemmatisation, part-of-speech tagging and morphological analysis of a 67,284 orthographic-unit Setswana corpus, including 9,146 verbs. The corpus was compiled from all Setswana texts that could be accessed in electronic

format and consists mainly of items from government texts. The corpus domain is specific and the language in this domain may differ from that in other domains.

The software LARA (Lexical Annotation and Regulation Assistant), developed by the Centre for Text Technology at the North-West University, was used for the annotation of this corpus. This software is available free of charge for research purposes. Once the manual annotation was completed, it was possible to extract statistics on the verbal morphology. As the tokenisation of Setswana verbs is influenced by the disjunctive orthography of the verbal prefixes, it is necessary to briefly elaborate on this process and how it is handled in the two abovementioned projects.

4.1 Tokenisation

Tokenisation or word-segmentation may be defined as the process of breaking up the sequence of characters in a text at word boundaries (see, for example, Palmer 2000). The dilemma for Setswana is that its orthography sometimes presents these morphemes as *orthographic* words or units, while they are not *linguistic* words. Consider the following example: *Ke a ba thusa* ('I help them') is linguistically one word in Setswana, consisting of the subject-agreement morpheme *ke*, the present-tense morpheme *a*, the object-agreement morpheme *ba*, and the verbal root *thus*- followed by the verbal ending *-a*. The English sentence, however, consists of three linguistic words. It is therefore necessary to distinguish between orthographic words and linguistic words for Setswana.

As an example of the problems that arise as a result of the disjunctive Setswana orthography, Otlogetswe (2007:125), in a study on corpus design for Setswana lexicography, indicates that *a* is the most frequent "word" in his 1.3 million "word" Setswana corpus. However, the status of *a* in Setswana as a word or morpheme is ambiguous as it could be any of the following six linguistic words or morphemes:

(7)			
	Subject-agreement morpheme:	*Mabone **a** a tuka*.	('The lights are burning/on.')
	Object-agreement morpheme:	*Ke tla **a** tima*.	('I will turn them (lights) off.')
	Present-tense morpheme:	*Ke **a** ba itse.*	('I know them.')
	Demonstrative pronoun:	*Mabone **a***	('These lights')
	Interrogative particle:	***A** o tla nthusa?*	('Will you help me?')
	Hortative particle:	***A** re tsamaye.*	('Let's go.')

In the Xerox Project, two tokeniser transducers and a finite-state (rule-based) morphological analyser are combined to effectively solve the Setswana tokenisation problem (cf. Pretorius, Berg, Pretorius and Viljoen 2009). Verbs are thus tokenised before morphological analysis. Alternatively, in the NCHLT Project, tokenisation is initially done on white space and once verbal prefixes have been tagged for parts of speech, they are manually concatenated to their roots to ensure correct word counts.

5. Data from the Setswana corpus in the NCHLT Project

Suffixes are referred to as "tails" in the NCHLT Project. Tail positions are determined by working from the longest combination of tails. As the longest sequence of tails includes five suffixes, five slots are allocated (verbal endings are not counted). The tail closest to the root is in tail position (TP) 1 while the tail furthest away from the root would then be in TP5.

Consider the following example in (8):

(8) *gatisitsweng* ('that has been stressed/printed')
gat[vr]$is[cau]$il[per]$w[pas]$a[ve]$ng[rts]

Here, the lemma is *go gata* ('to step on, to trample, to press'), and the root *-gat-* is followed by three tails (the causative *-is-*[cau], the perfect *-il-*[per] and the passive *-w-*[pas]) and two verbal endings (the verbal ending *-a*[ve] and the relative ending *-ng*[rts]).

The following suffixes appear in the NCHLT Project's tagset. The examples below indicate the type of suffix in square brackets followed by $ and then the actual Setswana suffix. The tagset is presented in rows here in (9) for the purposes of comparison with Krüger's table (cf. Table 1).

(9) [cau] $-is, [app] $-el, [rec], $-an [per],$-il [pas] $-iw
[rtr]$-ol, [rint] $-og, [neut], [nact], [npas], [den] $-fa/-fal, [iter] $-ak
[con] $-ar, [disp] $-al [pos] $-am
[asso] $-agan

The top row shows what Krüger refers to as the "productive suffixes", which are the same as the CARP template. Krüger, however, includes the perfect suffix preceding the passive here. The second row shows Krüger's semi-productive suffixes, while the non-productive suffixes appear in the third row. Krüger omits the non-productive suffixes from his table, but he does refer to them in his book (Krüger 2006:257).

The following graphs (Figures 2-6) show data of TP1 to TP5, followed by data of the total number of tails (Figure 7).

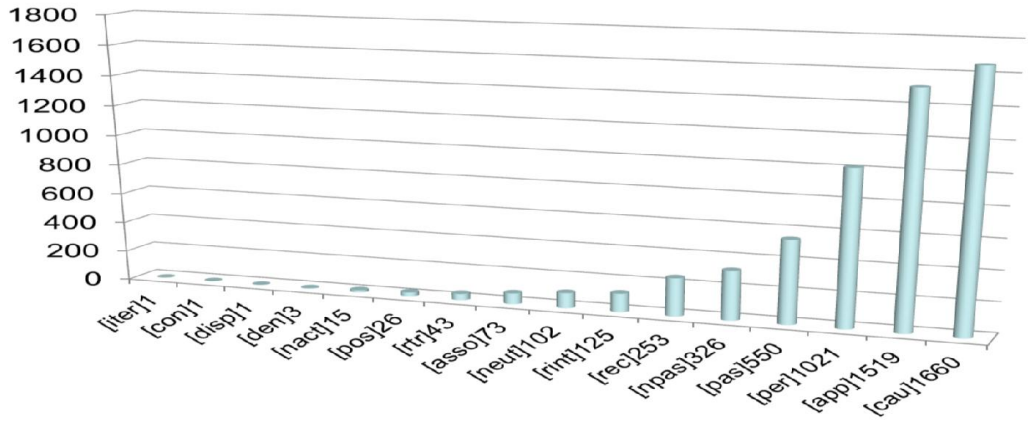

Figure 2. TP1[3]

[3] Note that these suffixes appear directly next to the root and are thus root-attached.

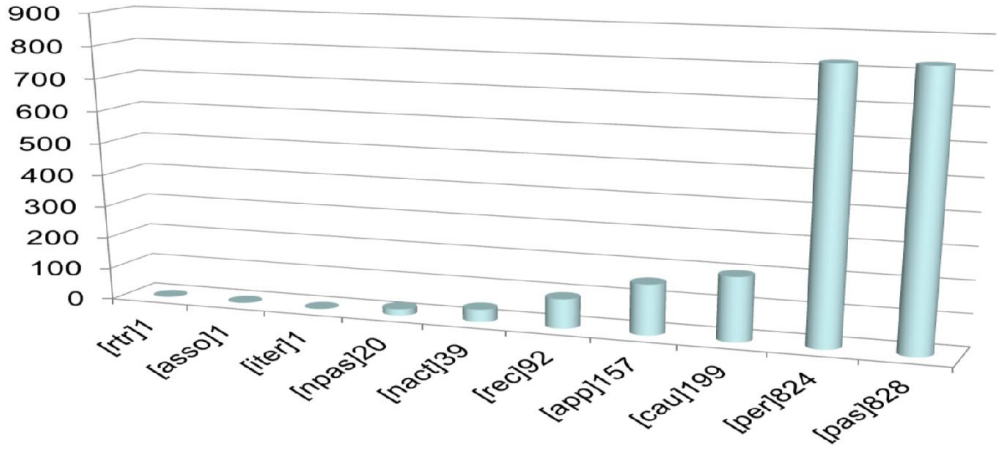

Figure 3. TP2[4]

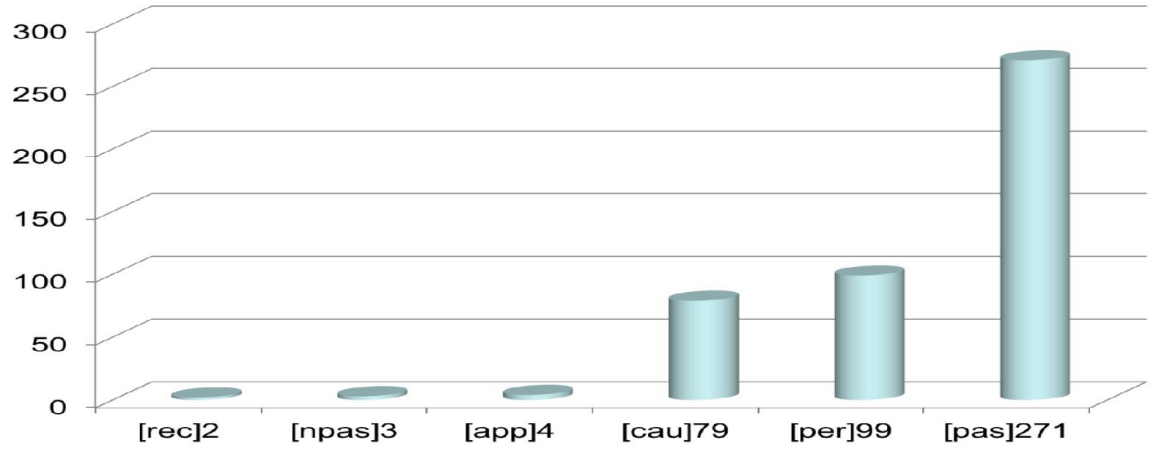

Figure 4. TP3[5]

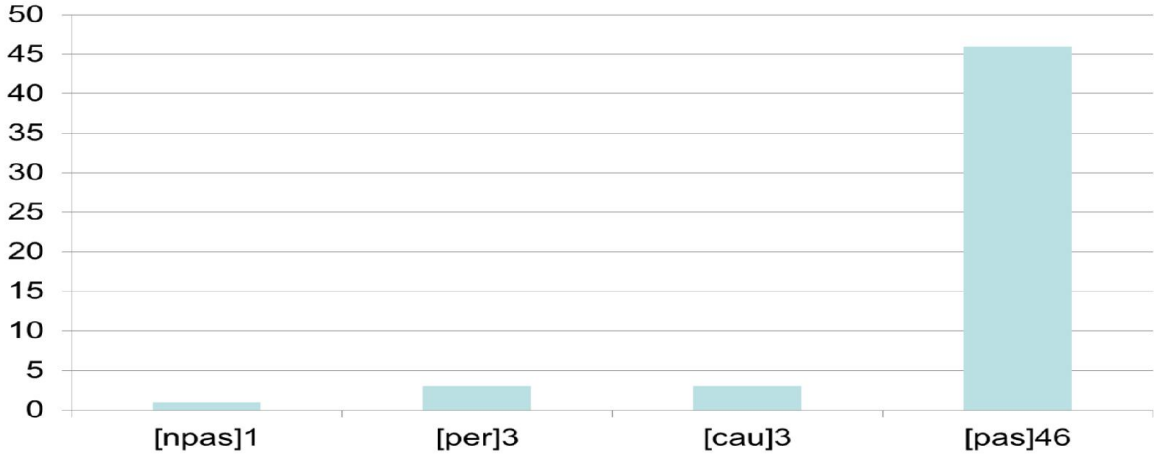

Figure 5. TP4[6]

[4] 2162 of the 9146 verbs in the corpus take at least two suffixes.
[5] 458 of the 9146 verbs in the corpus take at least three suffixes.
[6] 53 of the 9146 verbs in the corpus (0.57%) take at least four suffixes.

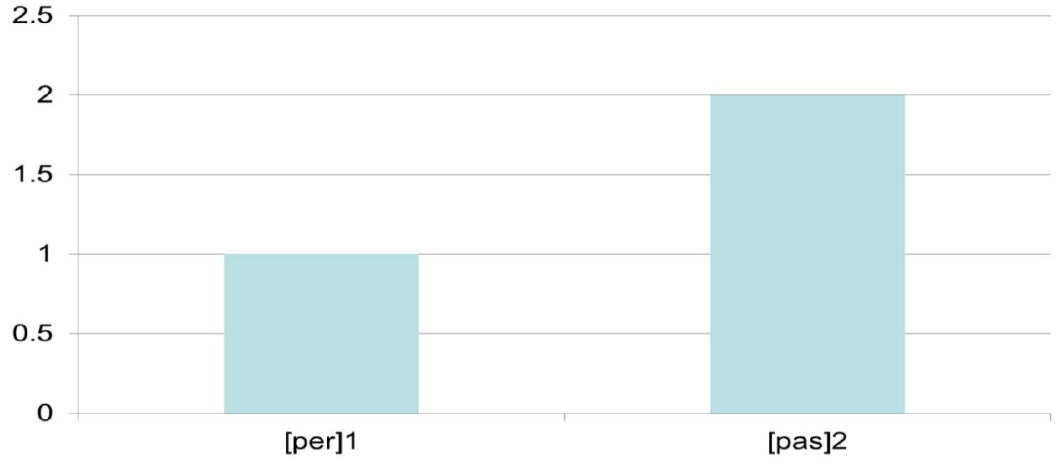

Figure 6. TP5

Only three of the verbs in the corpus take 5 suffixes. These verbs are given in (10) to (12):

(10) *tsenyeletswang*[7] ('being included')
tsen[vr]$is[cau]$el[app]ets$w[pas]$a[ve]$ng[rts]

(11) *sireletsegile*[8] ('being protected')
sir[vr]$el[app]$el[app]$is[cau]$eg[npas]$il[per]$e[ve]

(12) *lekanyediwa* ('being compared to / being measured for')
lek[vr]$an[rec]$y[cau]$el[app]$is[cau]$iw[pas]$a[ve]

Figure 7 below shows the amount of times that each morpheme appears in the corpus, while Table 2 shows the percentages for each morpheme based on these totals.

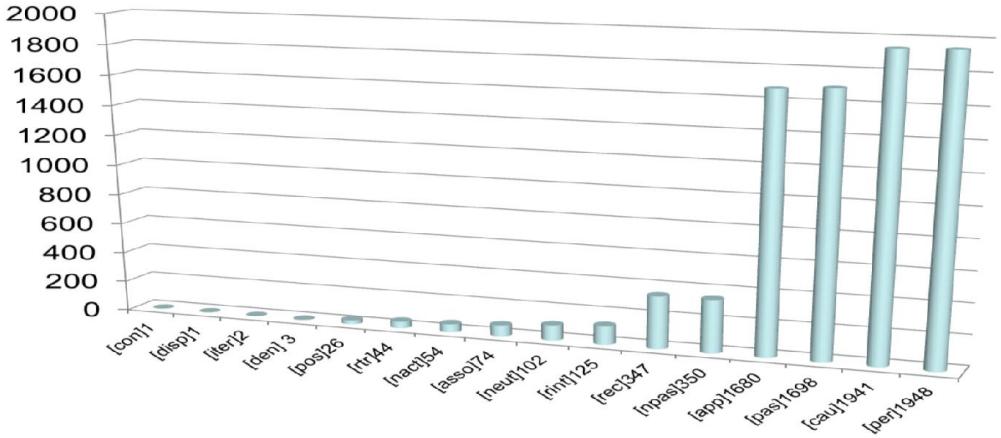

Figure 7. Tail totals

[7] *Setlamo se se* ***tsenyeletswang*** ('(the business which is) being included')
[cau][app][app][cau][pas] 2 0.02 tsen 2

[8] *Kgotlatshekelelo e tshwanetse go netefatsa gore dikgatlhegelo tsa basadi botlhe mabapi le thoto di* ***sireletsegile****.*
'(The court is supposed to ensure that the rights of all women regarding property) are protected'
[app][app][cau][npas][per] 1 0.01 sir 1

Table 2. Total tails with percentages

[per]	1948	23.1812%
[cau]	1941	23.0979%
[pas]	1698	20.2062%
[app]	1680	19.992%
[npas]	350	4.165%
[rec]	347	4.1293%
[rint]	125	1.4875%
[neut]	102	1.2138%
[asso]	74	0.8806%
[nact]	54	0.06426%
[rtr]	44	0.5236%
[pos]	26	0.3094%
[den]	3	0.0357%
[iter]	2	0.0238%
[con]	1	0.0119%
[disp]	1	0.0119%
TOTAL:	**8396**	**100%**

At this juncture, it is appropriate to elaborate on the contents and layout of the addenda of this paper. Addendum 1 lists the unique combinations of tails that occur more than once in the corpus. Note that in this addendum only one instance of a specific tail combination is counted per root, therefore adopting a wordlist approach. These totals appear in column 3 of the table in this addendum, but the percentages for the unique appearances are calculated for the 5,728 verbs with suffixes. There are 1,138 verbal suffix combinations in the corpus, therefore a single instance is 0.0878%. When this percentage is multiplied by the number of occurrences of each unique string, the percentage in brackets is the result (column 4).

Addendum 2, on the other hand, lists all instances of single tails and tail combinations with the same root in the corpus. The count is therefore corpus-oriented and not wordlist-oriented (as with the data in Addendum 1). The list shows the suffix, the amount of times that it appears (column 3), the percentage of that appearance in column 4 (calculated for all appearances including verbs with no suffixes), and the root with which the suffix appears the most (column 5). It is interesting to note that 3,418 of the 9,146 verbs in the corpus (37.4%) do not take any suffixes, only a single verbal ending or a verbal ending together with a relative ending. The order of suffixes is clear from the data in both Addenda 1 and 2.

Finally, Addendum 3 details the number of verbs in the corpus and in the wordlist which appear with 1, 2, 3, 4 or 5 suffixes, or with no suffixes at all.

The examples in (13) to (17) are of sequences from the data that do not conform to the order found in the grammars (causative, applicative, reciprocal, perfect, passive):

(13) Rec. – Caus. (7)
Ikgolaganye > *golega* (‘to bring yourself in contact with’)
$i[ref]goleg[vr]$an[rec]$is[cau]$e[ve]

(14) Rec. – Caus. – Pass. (2)
Lekanyetswang > *lekana* (‘that is being measured/quantified for/to’)
lek[vr]$an[rec]$is[cau]$el[app]$is[cau]$w[pas]$a[ve]$ng[rts]

(15) Appl. – Appl. – Caus. (6)
feleletsa > *fela* (‘to finish/finalise’)
fel[vr]$el[app]$el[app]$is[cau]$a[ve]

(16) Caus. – Npas (3)
isegang > *ya* (‘that is appropriate’)
y[vr]$is[cau]$eg[npas]$a[ve]$ng[rts]

(17) Rec. – Appl. (2)
Itekanetse > *leka* (‘healthy/fit/sound’)
$i[ref]lek[vr]$an[rec]$el[app]$il[per]$e[ve]

Having considered these statistics, deductions regarding the existing grammars of Setswana verbal suffixes can now be made, as detailed in the following section.

6. Insights gained from the data: Assessment of existing grammars

The correlation between suffix position and productivity, as established in the previous section, is presented in Table 3. These results support Krüger’s (2006) classification of suffixes into the unproductive, semi-productive and productive categories. New insights gained here are that the neutro-passive suffix is included in the penultimate group and the denominative suffix is included in the attached-to-root category. The recognition of the associative suffix and its productivity is an addition to Krüger’s categorisation (see Table 1). It has to be kept in mind that the application of the other criteria mentioned by Kosch (2006, 2007), apart from those discussed in §3.2, may influence this classification.

These results for Setswana are fairly similar to the classification that Anderson and Kotzé (2008, 2011) present for Northern Sotho. However, the percentages for the different suffixes differ substantially; this may be ascribed to the content of the corpus and the criteria for the calculation of percentages.

Table 3. The correlation between suffix position and productivity

Attached to root	Medial	Penultimate	
Contactive 1 0.0119%	Reversive transitive 44 0.5236%	Reciprocal 347 4.1293%	Applicative 1,680 19.992%
Dispersive 1 0.0119%	Neutro-active 54 0.6426%	Neutro-passive 350 4.165%	Passive 1,698 20.2062%
Iterative 2 0.0238%	Associative 74 0.8806%		Causative 1,941 23.0979%
Denominative 3 0.0357%	Neuter 102 1.2138%		Perfect 1,948 23.1812%
Positional 26 0.3094%	Reversive intransitive 125 1.4875%		

Table 4 is a combination of the classification in Table 3 and that of Krüger in Table 1. In Table 4, the morphemes in Table 3's "attached-to-root" category coincide largely with Krüger's "non-productive suffixes" (recall that the latter do not appear in Table 1), Table 3's "medial" morphemes with Krüger's "semi-productive suffixes", and Table 3's "penultimate" morphemes with Krüger's "productive suffixes".

Table 4. Morphemes of the Setswana verb

PREFIXES						ROOT	SUFFIXES				Verb-final categorial morpheme	
Neg. morph.	**Subj. agr. morph.**	**Neg. morph.**	**Asp. morph.**	**Temp. morph.**	**Obj. agr. morph.**		**Attached to root**	**Medial**	**Penultimate**		**Negative/ Modal**	
ga-	a. non-loc. b. loc. go- c. classless e- d. inf. go-	-se- -sa- (-a-)	-a- -sa- -ka-	-tla-	a. non-loc. b. loc. go- c. ideo.	a. orig. b. den. c. ideo.	Contactive 1 0.0119%	Reversive transitive 44 0.5236%	Reciprocal 347 4.1293%	Applicative 1,680 19.992%	-a -e -ê	-ng (rel.) -ng (imp.)
							Dispersive 1 0.0119%	Neutro-active 54 0.06426%	Neutro-passive 350 4.165%	Passive 1,698 20.2062%		
							Iterative 2 0.0238%	Associative 74 0.8806%		Causative 1,941 23.0979%		
							Denominative 3 0.0357%	Neuter 102 1.2138%		Perfect 1,948 23.1812%		
							Positional 26 0.3094%	Reversive intransitive 125 1.4875%				

With respect to all the data in this paper, three additional observations can be made. Firstly, as is clear from the data provided in Addendum 3, Setswana verbs with more than three suffixes do not occur as frequently as previously thought. In this corpus, only two instances occur where a verb has five suffixes. The majority of verbs in the corpus do not, in fact, contain any suffixes, while most of those that do include only one. The domain of the corpus may be a contributing factor, but the data is certainly more representative of natural language than Anderson and Kotzé's (2008, 2011) Northern Sotho data, taken from the *Concise Northern Sotho Dictionary*, where suffixes seem to have been mechanically added onto verbs.

Secondly, the frequency of the neutro-passive suffix is higher than that of the reciprocal suffix, which questions the position of the former in Krüger's (2006) categorisation schema in Table 1.

Finally, the associative suffix *-agan-* is mentioned in Cole (1955:211) as a "neuter-reciprocal form" that is a compound of the neuter *-eg-* and the reciprocal suffix *-an-*, as in *menagana* ('become folded together'). He also mentions an extensive reciprocal form which indicates reciprocal or associative participation in an extensive action, such as *thubakana* ('smash into one another'). Krüger (2006:257) does not include an associative suffix in his table, but he does mention the compound neuter *-eg-* and the reciprocal *-an-* (Krüger 2006:212). The associative morpheme appears 74 times in the corpus and should clearly be recognised as a Setswana verbal suffix. It is thus included in Table 4's "medial" column.

7. Concluding remarks

In conclusion, the results of the corpus analysis upon which this article is based show that proximity to the root and productivity are found to be in opposition for Setswana verbal suffixes. More productive suffixes appear further from the root in a string of suffixes. Generally, these results provide support for the categorisation schemes provided by Krüger (2006) for Setswana, and those of Kosch (2006, 2007) and Anderson and Kotzé (2008, 2011) for Northern Sotho. There are, however, a number of additional findings that contradict the current models. These include (i) that the degree of suffixation on the Setswana verb is not as extensive as is often assumed; (ii) that the position of the neutro-passive suffix vis-à-vis traditional grammars of Setswana may need re-thinking, and (iii) that the associative suffix does indeed appear in Setswana.

Apart from the abovementioned theoretical and empirical findings, information on the order of verbal suffixes, on an applied level, can be helpful for several computer-based applications. By way of an example, rule-based lexical transducers utilise a two-level finite-state network to simultaneously code morphological structure and generate rewrite rules. The analysis of a verb with multiple suffixes is, however, generally achieved through an analysis based on any logically possible combination of all known suffixes. This results in the unnecessary over-generation of lexical items, several of which may not occur in the relevant lexicon. Naturally, limiting the number of extensions that may occur in a string and, more importantly, pre-determining the most likely combinations in such strings should aid in restricting output to representations that have a good chance of actually existing. As such, the results of the current research can be of assistance in the future refinement of such computational tools.

References

Anderson, W.N. and A.E. Kotzé. 2008. Verbal extension sequencing: An examination from a computational perspective. *Literator* 29(1): 43-64.

Baker, M. 1985. The mirror principle and morphosyntactic explanation. *Linguistic Inquiry* 16(3): 373-414.

Bauer, L. 1992. *Introducing linguistic morphology*. Edinburgh: Edinburgh University Press.

Booij, G. 2005. *The grammar of words*. Oxford: Oxford University Press.

Brits, J.H. 2006. Automatic Setswana Lemmatisation. MA thesis, North-West University.

Bybee, J.L. 1985. *Morphology. A study of the relation between meaning and form.* Amsterdam: John Benjamins Publishing Company.

Cole, D. 1955. *An introduction to Tswana morphology*. Cape Town: Longman.

Dik, S.C. and J.G. Kooij. 1970. *Beginselen van de algemene taalwetenschap*. Utrecht/Antwerp: Uitgeverij het Spektrum.

Guthrie, M. 1971. *Comparative Bantu: An introduction to the comparative linguistics and prehistory of the Bantu languages*. Vol. 2. Farnborough: Greg International Publishers.

Hyman, L. 2002. Suffix ordering in Bantu: A morphocentric approach. In G. Booij and J. van Marle (eds.) *Yearbook of morphology*. Netherlands: Springer. pp. 245-281.

Kosch, I.M. 2006. *Topics in morphology in the African language context*. Pretoria: University of South Africa.

Kosch, I.M. 2007. Validity of criteria for inflectional and derivational processes. Paper presented at the 14th Biennial International ALASA Conference, 9-11 July 2007, Nelson Mandela Metropolitan University, Port Elizabeth.

Kotzé, A.E. 2011. Lexical generality as a determinant of extension position in Northern Sotho. *South African Journal of African Languages* 31(1): 30-40.

Krüger, C.J.H. 1967. Die Struktuur van die Woordgroep in Tswana. D.Litt. dissertation, University of Pretoria.

Krüger, C.J.H. 1968. Subkategorieë van die werkwoord in Tswana. *Taalfasette* 7(1): 26-30.

Krüger, C.J.H. 1976. *Aspekte van die woordgroepstruktuur.* Inaugural lecture. Potchefstroom: Potchefstroom University for Christian Higher Education.

Krüger, C.J.H. 1994. Notes on morphology with special reference to Tswana. *South African Journal of African Languages* 14(1): 15-23.

Krüger, C.J.H. 2006. Introduction to the morphology of Setswana. *LINCOM studies in African linguistics 69*. Munich: LINCOM Europa.

Lombard, D.P., E.B. Van Wyk and P.C. Mokgokong. 1985. *Introduction to the grammar of Northern Sotho.* Pretoria: Van Schaik.

Lyons, J. 1990. *Language and linguistics: An introduction*. Cambridge: Cambridge University Press.

Otlogetswe, T.J. 2007. Corpus Design for Setswana Lexicography. PhD thesis, University of Pretoria.

Palmer, D.D. 2000. Tokenisation and sentence segmentation. In R. Dale, H. Moisl and H. Somers (eds.) *Handbook of natural language processing*. New York: Marcel Dekker, Inc. pp. 11-35.

Pretorius, W.J. 2000. Die identifisering en beskrywing van die begrippe stam en wortel in die Afrikatale, met besondere verwysing na die Sothotale. *Tydskrif vir Taalonderrig* 34(1): 51-62.

Pretorius, R., A. Berg, L. Pretorius and B. Viljoen. 2009. Setswana tokenisation and computational verb morphology: Facing the challenge of a disjunctive orthography. In G. De Pauw, G.-M. de Schryver and L. Levin (eds.) *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages.* Athens: Tehnografia Digital Press. pp. 66-73.

Stageberg, N.C. 1971. *An introduction to English grammar.* New York: Holt, Rinehart and Winston.

Van Wyk, E.B. 1967. The word classes of Northern Sotho. *Lingua* 17(2): 230-261.

Addendum 1: Unique tail combinations – wordlist

iw	$iw[pas]	167	(14.6%)
il	$il[per]	116	(10.18%)
is	$is[cau]	110	(9.65%)
ilw	$il[per]$w[pas]	86	(7.55%)
el	$el[app]	78	(6.84%)
elw	$el[app]$w[pas]	45	(3.9%)
eg	$eg[npas]	37	(3.24%)
elil	$el[app]$il[per]	34	(2.98%)
elilw	$el[app]$il[per]$w[pas]	32	(2.80%)
isiw	$is[cau]$iw[pas]	32	(2.80%)
elel	$el[app]$el[app]	29	(2.54%)
an	$an[rec]	26	(2.28%)
egil	$eg[npas]$il[per]	19	(1.66%)
isilw	$is[cau]$il[per]$w[pas]	18	(1.58%)
isw	$is[cau]$w[pas]	16	(1.40%)
isil	$is[cau]$il[per]	13	(1.14%)
agan	$agan[asso]	10	(0.878%)
elelw	$el[app]$el[app]$w[pas]	9	(0.79%)
agal	$agal[neut]	9	(0.79%)

isel	$is[cau]$el[app]	8	(0.7%)
elis	$el[app]$is[cau]	8	(0.7%)
anis	$an[rec]$is[cau]	7	(0.61%)
olol	$olol[rtr]	7	(0.61%)
eseg	$eseg[npas]	7	(0.61%)
elelis	$el[app]$el[app]$is[cau]	6	(0.52%)
elelil	$el[app]$el[app]$il[per]	6	(0.52%)
eliw	$el[app]$iw[pas]	6	(0.52%)
og	$og[rint]	5	(0.43%)
ol	$ol[rtr]	5	(0.43%)
elelilw	$el[app]$el[app]$il[per]$w[pas]	5	(0.43%)
ogelw	$og[rint]$el[app]$w[pas]	4	(0.35%)
ologil	$olog[rint]$il[per]	4	(0.35%)
iselw	$is[cau]$el[app]$w[pas]	4	(0.35%)
iseliw	$is[cau]$el[app]$iw[pas]	4	(0.35%)
esegil	$eseg[npas]$il[per]	4	(0.35%)
iselilw	$is[cau]$el[app]$il[per]$w[pas]	4	(0.35%)
anw	$an[rec]$w[pas]	4	(0.35%)
ogel	$og[rint]$el[app]	3	(0.26%)
iseg	$is[cau]$eg[npas]	3	(0.26%)
elilw	$el[cau]$il[per]$w[pas]	3	(0.26%)
agalis	$agal[neut]$is[cau]	3	(0.26%)
egel	$eg[npas]$el[app]	3	(0.26%)
anilw	$an[rec]$il[per]$w[pas]	3	(0.26%)
al	$al[nact]	3	(0.26%)
elan	$el[app]$an[rec]	3	(0.26%)
aganisw	$agan[asso]$is[cau]$w[pas]	3	(0.26%)
agalil	$agal[neut]$il[per]	3	(0.26%)
anelil	$an[rec]$el[app]$il[per]	2	(0.17%)
isis	$is[cau]$is[cau]	2	(0.17%)
elanw	$el[app]$an[rec]$w[pas]	2	(0.17%)
am	$am[pos]	2	(0.17%)
elelisw	$el[app]$el[app]$is[cau]$w[pas]	2	(0.17%)
elelisil	$el[app]$el[app]$is[cau]$il[per]	2	(0.17%)
ololw	$olol[rint]$w[pas]	2	(0.17%)
ololw	$olol[rtr]$w[pas]	2	(0.17%)
elelan	$el[app]$el[app]$an[rec]	2	(0.17%)
elil	$el[cau]$il[per]	2	(0.17%)
ogan	$og[rint]$an[rec]	2	(0.17%)
isisiw	$is[cau]$is[cau]$iw[pas]	2	(0.17%)
agaliw	$agal[neut]$iw[pas]	2	(0.17%)
aganil	$agan[asso]$il[per]	2	(0.17%)
anisw	$an[rec]$is[cau]$w[pas]	2	(0.17%)
anil	$an[rec]$il[per]	2	(0.17%)
anel	$an[rec]$el[app]	2	(0.17%)
aganis	$agan[asso]$is[cau]	2	(0.17%)
amisilw	$am[pos]$is[cau]$il[per]$w[pas]	2	(0.17%)
anisel	$an[rec]$is[cau]$el[app]	1	(0.0878%)

Addendum 2: Unique tail combinations – corpus appearance

–	NoTailSlot	3418	37.40%	–	–
is	$is[cau]	909	9.95%	akarel	126
il	$il[per]	652	7.14%	nn	126
el	$el[app]	523	5.72%	lat	76
ilw	$il[per]$w[pas]	366	4.01%	bay	21
w	$w[pas]	332	3.63%	dir	65
elil	$el[cau]$il[per]	301	3.29%	tshwan	292
elel	$el[app]$el[app]	290	3.17%	tsw	123
iw	$iw[pas]	216	2.36%	tlal	14
eg	$eg[npas]	199	2.18%	tlhok	70
elw	$el[app]$w[pas]	182	1.99%	rom	37
elil	$el[app]$il[per]	177	1.94%	tshwan	89
isiw	$is[cau]$iw[pas]	165	1.81%	dir	65
an	$an[rec]	153	1.67%	farolog	62
elilw	$el[app]$il[per]$w[pas]	89	0.97%	nay	12
egil	$eg[npas]$il[per]	77	0.84%	kgeth	37
isw	$is[cau]$w[pas]	61	0.67%	tsen	27
isilw	$is[cau]$il[per]$w[pas]	58	0.63%	kwal	19
elis	$el[app]$is[cau]	56	0.61%	tsw	32
elan	$el[app]$an[rec]	53	0.58%	nay	32
agan	$agan[asso]	52	0.57%	sam	27
isil	$is[cau]$il[per]	38	0.42%	kwal	6
ogel	$og[rint]$el[app]	34	0.37%	am	25
elelis	$el[app]$el[app]$is[cau]	33	0.36%	sir	12
agalalis	$agal[neut]$al[nact]$is[cau]	32	0.35%	dir	32
anis	$an[rec]$is[cau]	30	0.33%	goleg	17
elelw	$el[app]$el[app]$w[pas]	27	0.30%	fitlh	14

Addendum 3: Tail counts

Corpus		**Wordlist**	
0	3418	0	457
1	3154	1	802
2	2096	2	536
3	417	3	184
4	50	4	35
5	3	5	2

Addendum 4: Abbreviations and their meanings

Obj. agr. morph.	Object-agreement morpheme
Subj. agr. morph.	Subject-agreement morpheme
Appl.	Applied suffix
Asp. morph.	Aspectual morpheme
Caus.	Causative suffix
CL	Noun class

Den.	Denominative suffix
den.	Denominative (root)
Idio.	Ideophonic (root)
imp.	Imperative
Inf.	Infinitive
Inf. morph.	Infinitive morpheme
Iter.	Iterative suffix
Loc.	Locative
Neg. morph.	Negative morpheme
Neg./Mod.	Negative / modal (endings)
Neut.	Neuter suffix
Non-loc.	Non locative
NPas	Neutro-passive
Obj. conc.	Object concord
Orig.	Original (root)
Pass.	passive suffix
Perf.	perfect suffix
Rec.	reciprocal suffix
rel.	Relative
Rev. intr.	Reversive intransitive suffix
Rev. tr.	Reversive transitive suffix
Subj. conc.	Subject concord
temp.	Temporal
[app]	Applied
[asso]	Associative
[cau]	Causative
[con]	Contactive
[den]	Denominative
[disp]	Dispersive
[iter]	Iterative
[nact]	Neutro-active
[neut]	Neuter
[npas]	Neutro-passive
[pas]	Passive
[per]	Perfect
[pos]	Positional
[rec]	Reciprocal
[rint]	Reversive intransitive
[rtr]	Reversive transitive