Verbal extension sequencing: an examination from a computational perspective

W.N. Anderson
School of Computing
University of South Africa
PRETORIA
E-mail: wanderson@acm.org

A.E. Kotzé
Department of Learner Support
University of South Africa
PRETORIA
E-mail: kotzeae@unisa.ac.za

Abstract

Verbal extension sequencing: an examination from a computational perspective

Lexical transducers utilise a two-level finite-state network to simultaneously code morphological analysis and morphological generation rewrite rules. Multiple extensions following the verb root can be morphologically analysed as a closed morpheme class using different computational techniques. Analysis of a multiple extension sequence is achieved by trivial analysis, based on any combination of the closed class members, but this produces unnecessary over-generation of lexical items, many of which may not occur in a lexicon. Limiting the extension combinations, in an attempt to represent examples that may actually exist – in terms of both the possible number of extensions in a sequence and the relative ordering of the extensions – leads to a radical reduction in the generation of lexical items while the ability to analyse adequately is maintained. The article highlights details of an investigation based on both trivial analysis and an approach that prevents dramatic over-generation. The article is based on test data reflecting possible extension sequences and the morphophonemic alternations of these extensions for Northern Sotho, garnered from literature research, lexicographic investigation and the computational morphological analysis of texts.

Opsomming

Werkwoordekstensieopeenvolging: 'n ondersoek vanuit 'n rekenaarmatige perspektief

Leksikale oorvormers maak gebruik van 'n tweevlak-eindigendetoestandnetwerk om tegelykertyd morfologiese analiserings- en morfologiese genereringsherskrywingsreëls te enkodeer. Veelvuldige werkwoordelike ekstensies kan morfologies as 'n geslote morfeemklas geanaliseer word deur van verskillende rekenaarmatige tegnieke gebruik te maak. Die analise van 'n veelvuldige opeenvolging van ekstensies word verwesenlik deur triviale analise, gebaseer op enige kombinasie van lede van die geslote klas, maar dit gee aanleiding tot onnodige oorgenerering van leksikale items, waarvan baie items nie in 'n woordeboek voorkom nie. Indien ekstensiekombinasies beperk sou word in 'n strewe na die verteenwoordiging van voorbeelde wat werklik bestaan, sowel ten opsigte van die moontlike aantal ekstensies in 'n reeks as die relatiewe ordening van die ekstensies, het dit 'n radikale vermindering in die generering van leksikale items tot gevolg, terwyl die vermoë om te analiseer voldoende gehandhaaf word. Die artikel belig besonderhede van triviale analise en 'n benadering wat dramatiese oorgenerering voorkom. Dit is gebaseer op toetsdata wat die moontlike ekstensievolgordes asook die morfofonemiese alternasies van hierdie ekstensies in Noord-Sotho weerspieël. Dit is bekom deur literatuurnavorsing, leksikografiese ondersoeke en die rekenaarmatige morfologiese analise van tekste.

1. Introduction

The research on which this article is based is part of a project funded by the National Research Foundation (NRF) in South Africa.[1] The aim of the project is to develop computational morphological analysers for a number of selected African languages. This article is based on work in progress in respect of Northern Sotho. Xerox finite-state lexical transducer software is used to design and build the analyser.

Due to the agglutinating nature of Northern Sotho and the possibility to concatenate up to six verbal extensions, a morphological analyser has to be designed to deal with an intricate morphology such as that of Northern Sotho. Not only does the concatenation of verbal

1 NRF project no. FA 2005033000044 entitled "Computational morphological analysis".

extensions offer a challenge to the design of the analyser, but the added complexity of accommodating different realisations of extensions in different environments is a further requirement that has to be satisfied. The identification of all the rules that are in operation with reference to verbal extensions in Northern Sotho, their analysis and the subsequent design of the lexical transducer for verbal extensions has been a significant frontier that needed to be conquered. The investigation on which this article is based was undertaken in the absence of a substantial textbook or other source material in respect of not only Northern Sotho but also the other Sotho languages. This being the case, the article should bring new insights into aspects of verb stem morphology in Northern Sotho specifically and also regarding the other Sotho languages.

2. Source investigation

The three major Northern Sotho textbooks, Ziervogel *et al*. (1976), Lombard *et al*. (1988) and Poulos and Louwrens (1994) were consulted for information regarding extension sequencing and combinations. Ziervogel *et al*. (1976) do not cover this aspect of verbal morphology at all, while Lombard *et al*. (1988:130-131) devote one page in total to combined extensions. Poulos and Louwrens (1994:151-152) clearly have the intension of being more informative on this topic than the other two mentioned textbooks, but, as this is a complex matter, two pages cannot be expected to do much justice to it. They briefly explain the relative positioning of the causative and the applied, the applied and the reciprocal, and the neutral and reversive extensions versus the causative, applied and reciprocal extensions. Their observation (Poulos & Louwrens, 1994:151-152) that the neutral and reversive extensions precede the causative, applied and reciprocal extensions is of real importance to our research but to this can be added the contactive, positional, iterative, and denominative extensions. This information we had to establish for ourselves by analysing linguistic data. They cite a verb stem that contains four extensions (Poulos & Louwrens, 1994:151-152) although the *Comprehensive Northern Sotho Dictionary* (CNSD) contains examples with six extensions, which could be increased to seven should the passive extension be appended to verb stems having six extensions. It has to be conceded that verb stems with that many extensions are probably somewhat unnatural because

neither the Pretoria Sepedi Corpus (PSC)[2] data nor the Northern Sotho Bible (Bibele: Taba Ye Botse, 2002)[3] contain verb stems with that many extensions. Anderson and Kotzé (2006) discuss the sequencing of prefixes from a finite-state perspective. They, however, do not cover the morphology of verbal extensions. Kotzé (2007) addresses the question as to why the verbal extensions of Northern Sotho are ordered in a specific sequence. He argues, based on statistical evidence of extension combinations, that their relative positioning corresponds with an important criterion for inflection or derivation, namely productivity. The data indicates that when the more productively used affixes are used in a string they will occur furthest from the stem. Conversely, the less productively used affixes will occur closer to the stem in a string. The conclusion is also drawn that widely distributed extensions are more inflectional than extensions with a more limited distribution. Although notice was taken of the detailed analysis of extension sequencing in the Setswana verb (cf. Krüger, 2006) the authors decided to undertake their own investigation about extension sequencing aimed at specifically Northern Sotho, seeing that this is not dealt with in detail in any Northern Sotho source encountered.

3. Trivial analysis of multiple verbal extensions

Lexical transducers utilise a two-level finite-state network to simultaneously code morphological analysis and morphological generation rewrite rules. Multiple extensions following the verbal root can be morphologically analysed as a closed morpheme class using different computational techniques. Analysis of multiple verbal extensions can, for instance, be done by *trivial analysis* based on any combination of verbal extensions as a closed class of morphemes. In the case of trivial analysis the sequencing of the morphemes does not matter and a transducer based on this principle will unnecessarily over-generate lexical items. Many of the morpheme combinations generated by such a lexical transducer will, in fact, not appear in the lexicon. The script appearing in example (1) is an example of a Perl script that could be used to isolate all the possible Northern Sotho morpheme sequences from any text. Note that the script searches for words that conform to both the scientific

2 We owe thanks to Prof. Danie Prinsloo of the University of Pretoria for making the data available to us.

3 Our sincere thanks to Dr. Eric Hermanson of the Bible Society who gave us access to this document for research purposes.

and practical orthography so that it could be used on dictionaries as well as texts written in the practical orthography. The scientific orthography is used in, for instance, the CNSD. It makes use of diacritics to distinguish between vowel phonemes that appear identical without the diacritics.

Example (1):

```
print "Started looking for verbal extension
sequences...\n";
while ($line = <> )
{
    if ($line =~ m/ (\s)
        (
            ([a-zA-Z|šŠÊêÔô]+)
            (
            ((ag)((a(l|tš|d))|((ê|e)tš)))|
            ((ol)(l|(o(g|l|tš|d|š))))|
            (o((tš)|d|g|š))|
            (ala(l|tš|d))|
            (al(ê|e)tš)|
            ((a(g|k)an)(y))|
            ((i|e)(w))|
            ((ê|e)(l|(tš)|d|w|g|š|r))|
            (a(l|d|k|r|m|n|(tš))(y))|
            (iš)|
            ((i)(l|tš|m))|
            (š)|
            ((t)šh)|
            (y|j)
            )
            {1,}
            ([a-zA-Z|šŠÊêÔô]+)
        )
        (\s)
        /gx )
    {
    print "<$2>\n";
}

}
print "Finished looking!\n";
```

Some of the results that were obtained from the application of this script (with at least four possible extensions) are listed under example (2). The examples were obtained from the Northern Sotho Bible (Bibele: Taba Ye Botse, 2002).

Example (2):

mošetšakeletšo	*one who takes advice*
ngwadišetšwa	*has been registered for*
ngwaletšwego	*(that) was written to*
ngwalolotšwego	*(that) was rewritten*
nyakišišwe	*cause to be investigated*
rarakanyetša	*make complicated for*
šwalalanyeditšwego	*(that) was scattered/dispersed for*
šwalalanyetša	*scatter/disperse for*
šwalalanyetšago	*(that) scatters/disperses for*
šwalalanyetšwa	*is scattered for*
swantšhetšwe	*take as an example*
swaraganywa	*being put/tied/stapled together*
swaretšwego	*(that) is forgiven for*
tswalanywa	*match (things)*
tswalelelwa	*lock up, imprison, confine*
tswaleletšego	*(that) locked up*

The usefulness in creating such initial scripts (where extension sequence order is irrelevant) was in the value of obtaining initial test data to establish the veracity of the extension sequencing rules in the literature.

Part of our exercise in determining rules in respect of extension sequencing was to survey all of the extensions that we were able to document in all their realisations in different phonological and morphological environments. This is a similar methodological approach to that followed by Anderson and Kotzé (2006). We took cognisance of the notes of Westphal *et al.* (1974) and the numbering systems of both Guthrie (1971) and Cole (s.a.) with reference to verbal extensions in Bantu, but decided to adopt a different numbering system. Hence, the numbers we assigned to extensions have been done on purely arbitrary grounds for use in the lexical transducer and not to reflect any grammatical aspects associated with the extensions.

Table 1: Verbal extension[4] numbering by Guthrie and Cole

Extension	Norm variant	Allomorphs	Guthrie Number	Cole Number[5]
Applied	-êl-	-êtš- -êd-	2 188	1
Passive	-w-	-iw- -ew-	2 194	2
Neuter	-agal-	-agêtš- -agatš- -agad-	2 198	9+4?
Neuter	-êg-	-êš-	2 190	3
Neuter	-al-	-atš- -êtš- -ad-	2 197b	4
Causative	-y-		2 193	6
Assistative (previously Causative)	-iš-	-tšh-	2 187	7
Reversive (transitive)	-ol-	-oll- -olol- -otš- -olotš- -od- -olod-	2 195 2 195a	8
Reversive (intransitive)	-og-	-og- -olog- -oš- -ološ-	2 197 2 197a	9
Reciprocal	-an-	-any-	2 185	11
Contactative	-ar-	-êr- -êš-	2 199a	13

4 The past tense suffix *-ilê* (and variations thereof) is not considered to be an extension and is therefore not included in this table.

5 In the case of extensions consisting of more than one syllable, Cole used a question mark to suggest the historical combination of more than one extension as the possible elements responsible for the polysyllabic extension. Guthrie also expresses his doubts about the origins of *-agan-*/*-akan-* by means of a question mark.

Positional	-am-		2 184/2 184a/2 191b	10
Dispersive	-alal-	-alêtš- -alatš- -alad-	2 196	4+4?
Associative	-agan-/ -akan-	-agany-/ -akany-	2 197b+2 186?	9+11?**6**
Iterative	-ak-		(portion of) 2 198	Variant of 9?

A numbering system makes it possible to document rule-related information to the lexical transducer in a simplified manner. A single number is used to represent the norm variant as well as allomorphs of extensions. This is useful since some of the extensions appear in quite a number of different forms, as is illustrated below with reference to the applied extension.

Example (3): Phonologically conditioned variants of the applied extension

-êl-	norm variant
-êtš-	result of fusion with causative *-y-*, for instance *bilêtša* (*invite*)
-êtš-	result of fusion with past tense suffix *-ilê*, full form being *-êtše*, as in for instance *rêkêtše* (*bought for*)
-êtš-	realises in environment of preceding s, tsw, š, tš, tšh, ny
-êd-	realises when followed by causative *-iš-*

Using unique numbers that had been assigned to individual extensions, the sequences and by implication the combinations of extensions, were formalised as linguistic rules. Data was primarily obtained from the Comprehensive Northern Sotho Dictionary (CNSD), the Pretoria Sepedi Corpus (PSC) and an electronic version of the Northern Sotho Bible (Bibele: Taba Ye Botse, 2002). The CNSD was inspected in order to find verb stems to which a significant variety of extensions can be affixed. This process was not pre-

6 The possibility that this could be a combination of the neuter and reciprocal which would make it 3 + 11 and not 9 + 11 was mentioned by an adjudicator. The exact composition, if indeed so, of bisyllabic extensions is and has been a matter of uncertainty.

scribed by a scientific methodology *per se*. Selected roots were included in the experimental corpus purely because it was obvious upon inspection that the number of extensions that could be affixed to them were higher than a perceived "average". In some cases roots were chosen because they combined with a particular extension that is not productive in order to at least include such extensions in the data. After investigation and analysis of a number of extended stems, certain patterns as regards extension combinations and sequences could be discerned, and because of this observation many of the selected verb stems were eventually excluded from the data list in order to eliminate unnecessary duplication. In addition, the 600 longest, mostly guessed, Northern Sotho verb stems, were extracted from the PSC, and were analysed in terms of the extension combinations and sequences they contained.

After this initial phase, possible extension sequences were searched for in an electronic version of the CNSD[7] to establish more environments within which they occur, to seek confirmation about their combinations with other extensions and to complete the list of their positions relative to other extensions. Extensions often occur in clusters of two, three or more. By extension sequencing we have the particular sequence of these extensions in a cluster of morphemes in mind.

A similar generic script, as defined in the Perl example above (example (1)), was also run against the CNSD and an electronic version of the Northern Sotho Bible (Bibele: Taba Ye Botse, 2002) in order to extract all extension combinations, regardless of sequence, as a test sample.

Finally, possible extension sequences were created by the authors by combining extensions without adding them to stems, and searched for in the electronic version of the CNSD and the sample from the PSC, in case certain combinations might not yet have been encountered.

Based on all of the data collected and inspected, as described, and after assigning unique numbers from 1 to 19 to the verbal extensions occurring in Northern Sotho, rules such as the following were completed in respect of all the extensions:

---

7 We would like to thank the publishers and copyright owners of this invaluable source, Van Schaik, for giving us permission to have it scanned for use in our research.

Example (4): Extension sequences: the applied and causative *-iš-*

Extension # **1 = -êl-**

Preceded by R (R = root), 1, 2, 3, 4, 7, 8, 9, 10, 11, 12, 14, 15

Followed by 1, 2, 3, 4, 16

Extension # 4 = -iš-

Preceded by R (R = root), 1, 2, 3, 4, 6, 7, 8, 9, 12, 13, 14, 15

Followed by 1, 2, 4, 6, 11, 16

An omission of possible extension combinations from the lexical transducer would at this stage not have posed any problems as testing of the transducer against selected texts was, among others, aimed at identifying any finite-state rule deficiencies. These can be easily added subsequently in order to add to the comprehensiveness of the transducer.

4. Probabilistic combination of sequences

Many of the extension combinations may be regarded as irrelevant since they do not conform to any possible surface form. Based on data from the CNSD we made the assumption that a sequence of five extensions as a maximum would seldom be surpassed. As mentioned previously, Poulos and Louwrens (1994:152) list a sequence of four as significant, although we have encountered as many as six. A sequence of five extensions renders a total of 195 possibilities (nineteen extensions in five extensions maximum per sequence), viewed from a probabilistic perspective. To generate all 195 possibilities is not a workable solution, and in fact will over-generate as described by Beesley and Karttunen (2003:237). Moreover, these rules will produce a finite-state machine that is *circular*, meaning that it can create an infinite number of possible extension sequences.

5. The Xerox finite-state lexicon for verbal extension sequencing

5.1 Rules in the Xerox finite-state lexicon

The finite-state lexicon used by Xerox, or *lexc*, is a finite state implementation of a given lexicon. According to Beesley and Karttunen (2003:203-204) it is a “kind of right recursive phrase-structure

grammar" that "compiles into a standard Xerox finite-state network", in other words, a transducer. It consists of what is referred to as *continuation classes* which are inherited from Koskenniemi's Two-Level Morphology notation which are the basic mechanisms for describing morphotactics. The continuation classes translate into concatenation. The *Oxford Concise Dictionary of Linguistics* defines *concatenation* as "The mathematical operation of juxtaposing units to form strings" (Matthews, 2005:67) and is intrinsic to the process of agglutination where "words are easily divided into separate segments with separate grammatical functions" (Matthews, 2005:12) or where "a word is easily segmentable into its constituent morphemes" (Booij, 2005:42).

Booij (2005:71) expresses himself as follows as regards the order of affixes in complex words:

> Complex words may contain more than one … suffix, and we would therefore like to know which principles govern the *order* of affixes. Given the distinction between derivation and inflection (derivation creates lexemes, inflection creates forms of lexemes) we expect the following scheme to apply, and this is indeed basically correct ...
>
> (23) Inflectional prefixes – Derivational prefixes – Root – Derivational suffixes – Inflectional suffixes

Example (5) is an example from the Xerox *lexc* lexicon that illustrates what the applied, causative and reciprocal extensions would look like using continuation classes based on the rules cited under example (4) above. Continuation classes express combinations explicitly in the Xerox lexicon rules, like those represented in example (4) above. Example (5) is presented in isolation – as such it is an illustrative subset example only to exemplify the continuation class process for verbal extension sequences. It illustrates what affixes explicitly succeed the particular verbal extension. Thus example (5) is a computational representation of aspects of the two examples shown in example (4) above. These rules encode the left (precedes) and right (follows) environments of each of these extensions with respect to other extensions, including also another instance of the extension being defined. These rules encode how the causative relates to the applied and the reciprocal; how the applied relates to the causative and the reciprocal and how the reciprocal relates to the causative and the applied. Also note that some of the extensions, for instance the causative can be reduplicated and the same is true of the applied.

Example (5): Continuation classes

```
LEXICON Vcausative
+Causative:iš                    VApplied;
+Causative:iš                    VReciprocal;
+Causative+Reduplicated:iš       VCausative;
                                 EndWord;

LEXICON Vapplied
+Applied:êl                      VApplied;
+Applied:êl                      VReciprocal;
+Applied:êl                      VCausative;
                                 EndWord;

LEXICON Vreciprocal
+Reciprocal:an                   VApplied;
+Reciprocal:an                   VCausative;
+Reciprocal:an                   VPassive;
                                 EndWord;
```

5.2 Use of the *empty language* to prevent cyclic over-generation

The possibility of an extension repeating itself or following upon a similar extension produces a *cycle*. Within a finite-state machine a cycle becomes a circular finite-state machine that will produce an infinite path whereby extension combinations can be built up in an infinite sequence of extensions. This is, however, not a true reflection of the realities of the surface language, which results in ungrammatical morpheme combinations.

There are two mechanisms to resolve cycles or circular finite-state machines in the Xerox toolset, namely by using flag diacritics or by filtering the finite-state machine using composition. Flag diacritics are significantly more useful for long distance dependencies but are not as obvious to the linguist to use as filtering in terms of reading the rule in the lexicon. Filtering easily adds rewrite rules that are well understood by the linguist; in other words filtering applies extra rewrite rules which implement the rules in the surface language so that invalid combinations cannot be allowed. For this reason we have initially designed these as filtered rules for ease of understanding. In final implementation either method could be utilised. Examples of rewrite rules that can be written to filter are:

Example (6a), (6b) and (6c):

- Make sure the applied extension occurs at most only three times in a verb stem.
- Make sure that the (sensory) neuter extension occurs only twice in a verb stem.
- Make sure the passive extension occurs as the last extension in any combination of extensions.

Filtering reduces or prevents over-generation well (Beesley & Karttunen, 2003:293). As an example of filtering consider the rules under example (7). The examples show how the *empty language* is used to prevent cyclic over-generation. The empty language is symbolised by ~[?*], a regular expression that symbolises the complement of any language, which is not to be confused with the empty string (often symbolised by epsilon ε to computer scientists or Φ to linguists). The empty language contains nothing; not even the empty string.

Example (7a): Applied

```
~[?*] <- $[ [ %+Applied ]{4,} ];
```

-bofelelela -el-el-el- *tie securely by winding around*

Example (7b): Causative

```
~[?*] <- $[ [ %+Causative ]{4,} ];
```

-bontšhišiša -iš-iš-iš- *grasp/comprehend fully*

The filtering rule in example (7a) is a rewrite rule that takes any combination of four or more applied extensions and converts these into the empty language. Hence example 7(a) will be accepted as a valid example, but another instance of the applied extension will be marked as invalid. Similarly, (7b) will convert four or more causative extensions into the empty language, so the word in the example will be regarded as valid but any word in the language that has more causative extensions will be marked as invalid. As a consequence of these rewrite rules, all the unwanted examples will be deleted from the language, and four or more occurrences of either the applied or causative extensions in any words will be prevented.

Over-generation occurs when, from a given verb root, a verb stem based on a valid combination of sequences can be created, but that actual verb stem does not exist as a surface form. Some verb roots are highly productive and can take many extension combinations, while others are not as productive or not productive at all. As an example of a highly productive verb the basic stem *-kgoma* (*touch*, *reach*) is quoted in example (8), taken from the CNSD. (Only selected translations are included.)

Example (8):

-kgoma		*touch, reach*
-kgomela	-el-	*touch for*
-kgomelana	-el-an-	*touch for each other*
-kgomiša	-iš-	*make (someone) touch*
-kgomišana	-iš-an-	*make each other touch*
-kgomega	-eg-	*be touchable*
-kgomana	-an-	*joined to one another*
-kgomantša	-an-iš-	*join together*
-kgomantšhetša	-an-iš-el-y	*join together on behalf of*
-kgomanya	-an-y-	*stick/glue/join together*
-kgomanyetša	-an-y-el-y-	*join together on behalf of*
-kgomagana	-agan-	*be joined together*
-kgomagantša	-agan-iš-	*join together*
-kgomagantšhetša	-agan-iš-el-y-	*join together for*
-kgomagantšhetšana	-agan-iš-el-y-an-	*join together for one another*
-kgomaganya	-agan-y-	*join together*
-kgomaganyetša	-agan-y-el-y-	*join together for*
-kgomaganyetšana	-agan-y-el-y-an-	*join together for one another*
-kgomara	-ar-	*stick to*
-kgomarela	-ar-el-	*adhere/cling to*
-kgomarelana	-ar-el-an-	*stick to one another*
-kgomaretša	-ar-el-y-	*stick onto*
-kgomaretšana	-ar-el-y-an-	*cause to stick onto one another*
-kgomaragana	-ar-agan-	*sticking together*
-kgomaragantšhetša	-ar-agan-iš-el-y-	*join together for*
-kgomaragantšhetšana	-ar-agan-iš-el-y-an-	*join together for one another*
-kgomaraganya	-ar-agan-y-	*join together*
-kgomaraganyetša	-ar-agan-y-el-y-	*join together for*

-kgomaraganyetšana	-ar-agan-y-el-y-an-	*join together for one another*
-kgomarolla	-ar-oll-	*loosen what sticks*
-kgomarollela	-ar-oll-el-	*loosen for/on behalf of*
-kgomarollelana	-ar-oll-el-an-	*loosen for/on behalf of one another*
-kgomarolliša	-ar-oll-iš-	*force (someone) to loosen*
-kgomarollišana	-ar-oll-iš-an-	*force to loosen for one another*
-kgomarologa	-ar-olog-	*become loose*

We have taken this complete set of continuation classes, as the subset illustrates in example (5), and the complete set of rules, as the subset illustrates in example (7a) and example (7b), and have used filtering to remove invalid extension combination sequences so that our final lexicon only produces the number of results for correct combinations. The consequence of this is that the infinite and circular combination of all possible extensions has been narrowed to reflect the true examples from the data, such as in example (8), thus radically reducing over-generation.

A final lexicon that curbs over-generation is the result of the focused continuation classes with the filters applied to the transducer. This final filtered lexicon can produce a number of results for combinations, some of these to the level of five or six extensions. The phonological alternations of the verbal extensions are consistent across any combination. This was achieved by not encoding the phonological alternation rules in the lexicon but rather by encoding them as phonological rules using *xfst*, another Xerox finite-state tool used for constructing alternations. Examples of roots (roots are underlined according to diachronic considerations) with five or more extensions from the PSC and the CNSD respectively, appear under (9).

Example (9):

Pretoria Sepedi Corpus

-itshwaretšetša R+ar+el+y+el+y

Comprehensive Northern Sotho Dictionary

-kgomaragantšhetšana R+ar+agan+iš+el+(y)+an

The passive extension can furthermore be added to both of these stems. These stems may furthermore also appear in the past tense, requiring the affixation of the past tense suffix.

6. Lexicalised extensions

In many instances certain of the nineteen verbal extensions of Northern Sotho are clearly recognisable in verb stems, and their presence confirmed by the meaning of the stems, although a corresponding *basic* verb stem, in other words a verb stem without the extension in question, does not exist (any longer). The majority of basic verb stems may, however, be used without any extensions.

Extended verb stems from which an extension cannot be removed without rendering a non-existing basic verb stem, contain a *lexicalised* extension. The term *lexicalised* as it has been used here is in accordance with a description in Booij (2005:17): "When a possible word has become an established word, we say that it has lexicalized. An important effect of *lexicalization* of complex words is that one of its constituent words may get lost, whereas the complex word survives." Where this has happened the extended root, in other words the root plus the extension in question as such is entered in the lexicon as a lexeme as, for instance, in the case of *-apara* (*put on clothes*) (obviously including the contactive extension *-ra*) which does not have a basic constituent verb stem **-apa*. In cases such as these, the extension in question can be shown to have a reduced productivity, in other words its distribution is restricted because it cannot be appended to all verb roots, and, as has already been argued, it can no longer be removed from certain verb stems. None of Northern Sotho's nineteen verbal extensions are totally non-productive,[8] although some are certainly much more productive than others. Example (10) contains examples with lexicalised extensions from the CNSD and the PSC. The lexicalised extension appears next to the translation, marked by an asterisk.

Example (10): Lexicalised extensions

-šarakana	*be entangled*	*akan	CNSD
-ahlama	*open (mouth)*	*am	CNSD
-ahloga	*separate*	*og	CNSD

8 *-apara* (*put on* (clothes)) loses the non-productive contactive *-ar-* in *-apola* (*take off* (clothes)) but takes on *-ol-* instead, and *-ama* (*touch*/*concern*) is a recognised verb stem, but it can also take the contactive and become *-amara* (*pursue*). Ample examples exist of specific verb stems either with a particular non-productive extension or with another extension having replaced a particular non-productive extension or totally without a particular non-productive extension.

-bitša	*call*	*y	CNSD
-bapola	*stretch out (as a skin)*	*ol	CNSD
-hlaramolla	*flutter/spread the wings*	*am + *oll	CNSD
-thathamologile	*became undone*	*olog	PSC
-iphatlalaletša	*adjourn themselves*	*alal	PSC
-phatlaladitšwe	*was adjourned*	*alal	PSC

For all the extensions, apart from the long causative (namely *-iš-*), applied, reciprocal, passive, denominative, neutro-passive and iterative extensions, significant numbers of lexicalised examples can be cited. Because extensions that are more inclined to appear in lexicalised stems are also still detachable from certain verb stems we have taken the stance of dealing with such extensions as if they were productive. Stems in which such extensions have become lexicalised can at a later stage be indicated in our data corpus as not containing the extension in question. This can be implemented utilising the flag diacritics discussed above.

7. Relative sequencing as a general pattern

Based on our examination and analysis of the data as described in this article, we have been able to identify slots – indicated by P1 to P4 in Table 2 – where individual extensions are positioned relative to one another in extended verb stems. These positions are relative because they are based on the position any extension would assume in a maximally extended verb stem, in other words a verb stem containing all possible extensions. In practice this means that

- if an extended verb stem contains only one extension, that extension will be attached to the root, regardless of the extension;
- if an extended verb stem contains two extensions, one of these will be attached to the root and the other to the extension which has been attached to the root;
- the positioning of extensions relative to one another is not random,[9] thus enabling positions to be provided for in a lexical

9 The relative positions of certain extensions are variable, with semantic implications. One of the variant sequences occurs much more frequently than the other. This sequence we regard as basic to Northern Sotho and is the one that finds its way into the lexical transducer. The less frequent sequence is dealt with as an alternation in *xfst*.

transducer, making it possible to analyse or compose any sequence of verbal extensions.

Having investigated a significant number of maximally extended verb stems we propose the relative extension position tendencies as reflected in Table 2. We have used these tendencies to make our finite-state rules more rigorous. The tendencies imply, among others the following as regards extension sequencing:

- Certain extensions are always attached to the root, despite the co-occurrence of other extensions. These extensions are referred to as *Root attached* for lack of a more suitable term. If an extended stem contains, apart from a *Root attached* extension, also another extension, then the latter will be preceded by the *Root attached* extension.
- If an extended stem contains four extensions, each belonging to what has been designated as *Root attached*, *Medial*, *Penultimate* and *Pre-final vowel*, then the extensions will be sequenced as indicated, viz. P1-P2-P3-P4. Within the group designated as *Penultimate*, three extensions, namely the causative, applied and reciprocal are predominantly ordered as suggested in the table. If all three occur in the same extended verb stem then they would in most instances appear in the sequence *-iš-el-an-*. The group designated as *Pre-final vowel* occupies the position adjacent to the final vowel.[10]

10 Unlike other extensions in this group, *-y-* is also found elsewhere in verb stems. It is provisionally placed in this slot because there is sufficient evidence that when an extension is affixed to *-y-*, it is likely to detach from its position and be affixed to the new affix, in other words to the pre-final vowel position.

Table 2: Relative extension positions

ROOT	P1	P2	P3	P4
	Root attached	**Medial**	**Penultimate**	**Pre-final vowel**
	14. ar 12. ak 15. am 5. al 10. alal 11. agal 19. f 18. fal 17. iw	13 agan/akan 7. oll/ol 6. olog 9. og	4. iš 1. el 2. an 8. eg 3. y	16. w
	14. Contactive 12. Iterative 15. Positional 5. Neutro-active 10. Dispersive 11. Intensive-neutro-active 19. Denominative 18. Denominative 17. Passive	13. Associative 7. Transitive-reversive 6. Intransitive-reversive 9. Intransitive-reversive	4. Causative 1. Applied 2. Reciprocal 8. Neutro-passive	3. Causative 16. Passive

8. Conclusion

We have determined that there are sequencing rules in respect of verbal extensions and these rules determine which verbal extensions can co-occur with which other extensions based on linear concatenation sequencing rules. Sequencing rules can also spell out – if specific extensions are present in a verb stem – the sequence specific extensions would assume if they occurred in the verb stem. Such a set also produces a broad verbal sequencing rule for use in the lexicon of our finite-state lexical transducer. In addition, it ultimately allows us to limit the extension combinations to prevent potential problems with over-generation of extension sequences.

This was done by looking at examples that actually exist in terms of the possible number of extensions in a sequence and the relative ordering of those extensions. The computational implementation of the rules allows us to produce a radical reduction in the generation of lexical items. It has no effect on the ability to analyse any verb form, in other words the ability to analyse is maintained without expansion to all possible combinations. Hence the analysis ability is the same as a lexicon that could be written to interpret any combination of extension sequences, but there is a gain in reducing over-generation.

The main purpose of our endeavours was to establish the relative extension positions as reflected in Table 2, and in addition, to provide the lexical transducer with the phonological rules that would dictate the correct allomorph of any extension in any given extension sequence. Table 2 strongly suggests that the position of verbal extensions in the sequence is generally governed by the "productiveness" of the extension, i.e. non-productive generally followed by semi-productive and then generally followed by productive. In other words, either the less productive the verbal extension, the "closer" it attaches to the root, or alternatively, the more productive the verbal extension, the "further" it tends to occur from the root (Kotzé, 2007). Further detailed corpora research could perhaps elucidate information to make a more accurate assessment on productivity based on extension frequency and position.

A prototype lexical transducer which was built along the lines spelt out in this article is currently being tested and is delivering promising results on new corpus data.

List of references

ANDERSON, W.N. & KOTZÉ, P.M. 2006. Finite state tokenization of an orthographical disjunctive agglutinative language: the verbal segment of Northern Sotho. Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy.

BEESLEY, K.R. & KARTTUNEN, L. 2003. Finite state morphology. Stanford: CSLI Publications. (Series: CSLI Studies in Computational Linguistics.)

BIBELE: TABA YE BOTSE. 2002. Cape Town: Bible Society of South Africa.

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

COLE, D.T. s.a. Unpublished notes on comparative Bantu linguistic structures. Pretoria: University of South Africa.

GUTHRIE, M. 1971. Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages. Vol. 2. Farnborough: Gregg International.

KOTZÉ, A.E. 2007. Derivation versus inflection as a determinant of extension position in Northern Sotho. Paper presented at the International Conference on Bantu Languages: analysis, description and theory, Department of Oriental and African Languages, Göteborg, Sweden, 4-6 October 2007.

KRÜGER, C.J.H. 2006. Introduction to the morphology of Setswana. Muenchen: Lincom Europa. (Lincom Studies in African Linguistics.)

LOMBARD, D.P., VAN WYK, E.B. & MOKGOKONG, P.C. 1988. Introduction to the grammar of Northern Sotho. Pretoria: Van Schaik.

MATTHEWS, P.H. 2005. Oxford concise dictionary of linguistics. Oxford: Oxford University Press.

POULOS, G. & LOUWRENS, L.J. 1994. A linguistic analysis of Northern Sotho. Pretoria: Via Afrika.

WESTPHAL, E.O.J., MASIEA, S.M., TINDLENI, H.M., MZILENI, I.V. & MAIELA, M.T. 1974. The verbal extentions in Southern Bantu languages: a descriptive and comparative classification. *Bulletin of the School of African and Oriental Studies*, 37(1):213-222.

ZIERVOGEL, D., LOMBARD, D.P. & MOKGOKONG, P.C. 1976. Handboek van Noord-Sotho. 5e uitgawe. Pretoria: Van Schaik.

ZIERVOGEL, D. & MOKGOKONG, P.C. 1985. Comprehensive Northern Sotho dictionary. 2nd corrected edition. Pretoria: Van Schaik.

Key concepts:

continuation classes
empty language
extension sequencing rules: Northern Sotho verbs
trivial analysis: multiple verbal extensions

Kernbegrippe:

ekstensie opeenvolgingsreëls: Noord-Sotho werkwoorde
kontinueringsklasse
leë taal
triviale analise: veelvuldige werkwoordekstensies