Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele

Laurette Pretorius, Sonja Bosch
University of South Africa
PO Box 392, UNISA, 0003, Pretoria, South Africa
E-mail: pretol@unisa.ac.za, boschse@unisa.ac.za

Abstract

A finite-state morphological grammar for Southern Ndebele, a seriously under-resourced language, has been semi-automatically obtained from a general Nguni morphological analyser, which was bootstrapped from a mature hand-written morphological analyser for Zulu. The results for Southern Ndebele morphological analysis, using the Nguni analyser, are surprisingly good, showing that the Nguni languages (Zulu, Xhosa, Swati and Southern Ndebele) display significant cross-linguistic similarities that can be exploited to accelerate documentation, resource-building and software development. The project embraces recognized best practices for the encoding of resources to ensure sustainability, access, and easy adaptability to future formats, lingware packages and development platforms.

1. Introduction

The normalisation and technological development of under-resourced languages not only involves the creation of basic resources, tools and technologies for processing these languages electronically, but also requires that such work be sustainable. Sustainability means that resources will remain accessible and available into the future even if the formats, software systems and development platforms on which they were developed become obsolete. To be sustainable they must even transcend communities of practice, domains of applications and the passage of time (Simons & Bird, 2008). Maxwell's (2012) approach to the sustainability of (morphological) grammars is of significance. He makes the point that grammatical descriptions must be reproducible (testable) and archivable, and shows that by making a grammar, including machine-processible rules, archivable, it also becomes reproducible. By using literate programming a descriptive and a formal grammar are interwoven. Furthermore, the formal grammar is a structured XML version of the descriptive grammar and is also parsing technology agnostic. In other words, it is archivable, (re)producible, human and machine-readable.

The problem that we address in this article is in some sense the converse one – we perform a semi-automated corpus-based extraction of a novel formal morphological grammar for Southern Ndebele from an existing morphological parser for Nguni. Our notion of an automatically extracted morphological grammar (as defined in section 4) is suitable for both human and machine processing. On the one hand, it employs a well-established grammar formalism that facilitates human readability and enables linguists to understand, evaluate and enhance the current descriptive status of the language. On the other hand, the machine processability makes it suitable as a resource from which new parsers can be (semi-)automatically constructed using other sustainable formalisms as well. In cases where groupings of under-resourced languages exist, developments towards sustainability of language resources for the group on the basis of collective information is of specific significance for the least resourced members of the group, both in terms of sustainable human-readable and machine-readable resources.

We examine the case of Southern Ndebele (ISO 639-3: `ndl`)[1], a resource-scarce language, for which we have semi-automatically obtained a morphological grammar from a general Nguni morphological analyser, which was bootstrapped from a mature hand-written Zulu morphological analyser called ZulMorph (Pretorius & Bosch, 2010). Due to the lack of Southern Ndebele language resources and data for inclusion in the bootstrapped Nguni analyser, Southern Ndebele morphological analysis relies heavily on the morphological structure of the other languages in the Nguni group.

The surprisingly good results obtained for Southern Ndebele morphological analysis suggest that a significant fragment of Southern Ndebele morphological grammar is covered by the Nguni analyser via the cross-linguistic similarities within the group. In this paper we describe a procedure for extracting this fragment and making it explicit with the purpose of creating a purely Southern Ndebele language resource that could be used in various ways, including the following: (a) to complement and extend the existing Southern Ndebele (paper-based) documentation towards enhanced Southern Ndebele language description; (b) to subject the fragment of the Southern Ndebele morphological grammar to human elicitation for quality assurance, correctness and enhancement purposes (the standard formal grammar notation is amenable to both human and machine processing); (c) to create a machine-readable formal morphological grammar that could serve as basis for the automated construction of morphological parsers in the future towards sustained Southern Ndebele language technological development.

The article is structured as follows: a background on Nguni languages; a brief explanation of the bootstrapping process with regard to Nguni languages; an illustration of the automated extraction of a morphological grammar for Southern Ndebele from the Nguni analyser; evaluation and results; and finally a conclusion and future work.

[1] Southern Ndebele should be differentiated from Northern Ndebele (ISO 639-3: (`nde`), spoken in Zimbabwe.

2. Nguni languages

The Nguni group of languages (S30 according to Guthrie's classification (Nurse & Philippson, 2003:649)) belongs to the South Eastern zone of the Bantu language family, and includes the following official languages of South Africa: Xhosa (S41), Zulu (S42), Swati (S43) and Southern Ndebele (S407).

The Nguni languages have a rich and complex morphology with a noun classification system that categorises nouns into a number of noun classes, as determined by prefixal morphemes also known as noun prefixes. Noun prefixes also link nouns to other words in the sentence by means of a system of concordial agreement, which is the pivotal constituent of the whole sentence structure of these languages, and governs grammatical correlation in most parts of speech.

2.1 Southern Ndebele

Southern Ndebele was the last South African Bantu language to receive official recognition, in fact during 1985 in the previous homeland of KwaNdebele which led to Southern Ndebele being introduced as official subject in schools for the first time. The language is spoken predominantly in the Mpumalanga and Gauteng provinces by a population of approximately 640,000.

Although Southern Ndebele has been an official language for 27 years, no comprehensive grammar, suitable for use by teachers and students, exists. Two master's dissertations on certain aspects of Southern Ndebele grammar were written by Potgieter (1950) (a description of the Ndzundza dialect) and Jiyane (1994). The first dictionary (isiNdebele/English, 2006) appeared more than 20 years into the official status of the language. Although an open-source spell checker for Southern Ndebele, based on word lists[2], a spellchecker and hyphenator[3], and corpora of 1.0 million untagged tokens[4] exist, no dictionary in electronic format, nor a detailed formally documented linguistic description was available for the bootstrapping process.

3. Nguni finite-state morphological analysis

Finite-state technology remains a preferred approach for modelling the morphology of natural languages. While machine learning approaches have grown in use, results for Nguni remain at the proof-of-concept level, due to among others morphological complexity and severe lack of appropriate language data (see for example Spiegler et al., 2008).

3.1 ZulMorph

The development of a broad-coverage finite-state morphological analyser prototype for Zulu (ZulMorph) is based on the Xerox Finite-state Tools (Beesley & Karttunen, 2003) and is reported on in detail in several publications, e.g. Pretorius and Bosch (2010). The Xerox software tool **lexc** is used to enumerate the required and essential natural-language lexicon and to model the morphotactic structure of Zulu words in this lexicon. Subsequently **lexc** source files are produced and compiled into a finite-state network which renders morphotactically well-formed, but rather abstract morphophonemic or lexical strings. The morphophonological (phonological and orthographical) alternations are modelled with rules written in the Xerox **xfst** format. Here the changes (orthographic/spelling) that take place between lexical and surface words when morphemes are combined to form new words/word forms, are described. These lexical strings are referred to again in section 4.3 where the extraction of the morphophonological rules is discussed.

Finally, the **lexc** and **xfst** finite-state networks are composed into a single network, namely a so-called lexical transducer that includes all the morphological information about the language being analysed, and constitutes the computational morphological analyser of the language, in this case the Zulu morphological analyser ZulMorph. It is customary to refer to the analysis language as the upper language of the transducer and to the surface form language as the lower language. Table 1 shows a summary of the core components of ZulMorph:

Morphotactics
Affixes for all parts-of-speech (e.g. SC, OC, CL PREF, V SUF, N SUF, TAM morphemes etc.) **Pronouns** (e.g. absolute, demonstrative, quantitative) **Demonstrative copulatives** **Word roots** (e.g. nouns, verbs, relatives, adjectives, ideophones, conjunctions) **Rules** for legal **combinations** and **orders** of morphemes (e.g. *ba-ya-si-khomb-is-a* and not **si-ba-ya-khomb-a-is*)
Morphophonological alternations
Rules that determine the **form** of each morpheme (e.g. *ku-hamb-w-a* > *ku-hanj-w-a, u-mu-lilo* > *u-m-lilo*)

Table 1: Core components of ZulMorph

3.2 A bootstrapped Nguni analyser

Due to the lack of language resources, bootstrapping of applications for new languages, based on existing applications for closely related languages, has gone a long way to reduce development time and efforts of building morphological analysers for lesser resourced languages – spoken by relatively few people – thereby ensuring technological development for such languages as well. Antonsen et al. (2010) report on the notable gain in reusing grammatical resources when porting language technology to new languages in the Uralic language family. The use of ZulMorph to bootstrap broad-coverage finite-state morphological analysers for Xhosa, Swati and

[2] http://translate.org.za/content/view/1610/54/

[3] http://www.nwu.ac.za/content/nwu-potchefstroom-campus-context-11

[4] http://web.up.ac.za/default.asp?ipkCategoryID=1883&sub=1&parentid=482&subid=1866&ipklookid=9

Southern Ndebele is discussed extensively in Bosch et al. (2008). The results of a preliminary evaluation based on parallel test corpora of approximately 7,000 types each for the four languages, indicate that the "high degree of shared typological properties and formal similarities among the Nguni varieties warrants a modular bootstrapping approach" (Bosch et al. 2008:66). The bootstrapping process is done in various stages by reusing the core components of the Zulu analyser for the three additional Nguni languages, shown in Table 1.

The bootstrapping approach functions as semi-automatic support to human linguistic expertise that allows linguists to focus their attention on just those aspects in which the languages differ. Adapting ZulMorph to provide for affix variations in the related languages, e.g. the form of morphemes in the 'closed' classes, proved to be a trivial implementation matter. However, certain areas in the grammars of individual languages that differ substantially from those applicable to Zulu required custom modelling and were built into the analyser as additional components e.g. the copula construction and the formation of the extended noun stem of Southern Ndebele. In the latter case the noun stem may suffix morphemes signifying the diminutive, augmentative etc. In the Southern Ndebele corpus two Southern Ndebele specific constructions occur which need to be included in the morpheme sequencing: a noun stem may suffix a demonstrative pronoun `[Dem7][Pos1]` or a possessive concord followed by a pronominal stem `[PossConc3] [PronStem7]`, e.g.

isilwanesi ('this animal')
`i[NPrePre7]si[BPre7]lwana.7-8[NStem-NR]`[5]
`lesi[Dem7][Pos1]`
umsilaso ('its tail')
`u[NPrePre3]mu[BPre3]sila.3-4[NStem]`
`wa[PossConc3]so[PronStem7]`

The Nguni lexical transducer has approximately 270 000 states and 833 000 transitions and occupies 14.3 MB of memory.

Simon and Bird (2008) identify six necessary and sufficient conditions for the sustained use of such resources. In particular, a language resource must be extant, discoverable, available, interpretable, portable, and relevant. The sustainability characteristics for ZulMorph and the Nguni analyser are given in Table 2.

Sustainability characteristics of ZulMorph and Nguni analyser
Extant Yes: Xerox finite-state tools implementations; appropriately backed-up off-site; mature prototypes in an advanced state of completion.
Discoverable Not yet: has not been released yet.
Available Limited: data analysis done on request, e.g. for National Centre for HLT, South Africa[6] .
Interpretable Yes: strictly based on the finite-state formalism and tools as described in (Beesley and Karttunen, 2003); adheres to relevant encoding standards; appropriately documented.
Portable, best practices Yes: shown to be compatible with equivalent open source initiatives such as foma (Hilden, 2009) and HFST (Lindén et al., 2011). Finite-state computational morphology is well established and can be expected to survive into the future. Finite-state research agendas already make provision for certain known limitations (Wintner, 2007).
Relevant Yes: constitute essential enabling technologies for next stages in the natural language processing pipeline of the agglutinating morphologically complex Nguni languages.

Table 2: Sustainability characteristics of ZulMorph and the Nguni analyser

4. Extraction of Southern Ndebele morphological grammar

The morphological grammar for Southern Ndebele consists of two main components, viz. rules that govern the morphotactics and rules that model the morphophonological alternations of Southern Ndebele. The morphotactics component (section 4.2) is a set of rules of the form $N \rightarrow N+$ or $N \rightarrow \Sigma$ where + is the Kleene plus operator, N is the (finite) set of morphological labels/tags and Σ is the (finite) set of actual morphemes. A distinguished symbol S ε N is the start symbol. The set of ordered morphophonological alternation rules are encoded by means of **xfst** (conditional) replacement rules. The rule `A -> B || L _ R` means that any string in language `A` is replaced by any string in language `B` only if the left context of the string in `A` is in the language `L` and the right context is in the language `R`. `A`, `B`, `L` and `R` are regular languages. These are addressed in section 4.3

The extraction (reverse engineering) of a morphological grammar for Zulu would be based on the full finite-state description of the complete Zulu morphology, as implemented in ZulMorph by means of **lexc** and **xfst**. This approach is essentially different from the extraction of a Southern Ndebele morphological grammar, which is corpus-based since the Nguni analyser does not contain much by way of explicit Southern Ndebele information. The grounding of the morphological grammar in Southern Ndebele therefore takes place via appropriate attested corpora. This approach results in a partial morphological grammar from the Nguni morphological analysis of the words in the corpus. Future work will focus on bigger corpora in order to increase the coverage of the Southern Ndebele morphological grammar.

4.1 General corpus-based approach

For the purposes of demonstrating the validity of the

[5] The notation NR indicates a Southern Ndebele specific morpheme.

[6] http://www.dac.gov.za/newsletter/khariambe_3_4.html

approach, a small representative Southern Ndebele corpus was used. After standard pre-processing and tokenisation were performed, the word list of 180 types (unique words) was subjected to morphological analysis with the Nguni analyser. In order to further constrain the proof-of-concept to tractable scope for the purpose of this article, only analyses that are based on noun stems were considered. Typical analyses that form the basis of the extraction process explained in sections 4.2 and 4.3 are as follows:

wesilwana ('of the animal')

```
wa[PossConc1]i[NPrePre7]si[BPre7]lwana.7-8
[NStem-NR]
```

wabhudanga ('he dreamt')

```
wa[PTSC1]bhudang[VRoot-NR]a[VerbTerm]
```

The coverage of the morphological grammar (both morphological structure and word root lexicons) can be increased by (a) systematically including other parts of speech, for example verbs, and (b) using larger corpora.

4.2 Rules for morpheme sequencing and the lexicon

The morphological grammar rules are automatically extracted from the morphological analyses by means of a pattern-matching procedure. As is customary for finite-state approaches all possible analyses for any given form are produced. For the purposes of obtaining the morphological grammar no (context dependent) disambiguation is necessary since all analyses are assumed to be valid and are therefore relevant for the extraction of morphological grammar rules.

In general, the following main parts of speech are recognised in the Nguni languages: noun, pronoun, demonstrative, qualificative, verb, copulative, adverb, ideophone, interjection, conjunction and interrogative (cf. Poulos and Msimang, 1998:26). In the corpus, we focus on a selection of these parts of speech, which contain the tag `NStem`, and we discuss examples of morpheme sequencing rules obtained from the analyses of these words.

Noun

The noun in the Nguni languages is constructed of two main parts, namely a noun prefix and a noun stem with the annotation `[NStem]` in ZulMorph. The noun prefix is the carrier of class information[7] and is usually divided into a so-called preprefix and a basic prefix. The noun stem may suffix morphemes signifying the diminutive, augmentative etc. In the Ndebele corpus two Ndebele specific constructions occur which required an enhancement of the morpheme sequencing: a noun stem may suffix a demonstrative pronoun or a possessive concord followed by a pronominal stem, e.g.

```
Noun -> NPrePre BPre NStem AugSuf
Noun -> NPrePre BPre NStem-NR PossConc PronStem
Noun -> NPrePre BPre NStem-NR Dem
```

Copulative

The copulative is a non-verbal predicate in Nguni and can be formed with a variety of words or stems, e.g. nouns, pronouns, adverbial forms etc. The following examples demonstrate the morpheme sequencing in copulatives formed from noun stems:

```
Copulative -> CopPre BPre NStem
Copulative -> SubjSC PreLoc-s LocPre (NPrePre)
BPre NStem
```

Qualificative

The qualificative part of speech is a collective term that covers different types of qualifying or descriptive words such as the adjective, relative, possessive and enumerative, as illustrated below:

```
Qualificative -> PossConc NPrePre NStem DimSuf
Qualificative -> RelConc AdvPre NPrePre BPre NStem
```

Adverb

The adverb in Nguni languages is quite a mixed bag, involving numerous types of grammatical constructions, mainly derived from other parts of speech such as nouns, pronouns, demonstratives, qualificatives etc. Since we focus on words containing noun stems in this paper, the adverbs under discussion are formed by prefixes and/or suffixes added to a noun. Constructions include adverbs formed by using prefixes *nga-* (instrumental), *na-* (associative) etc.; the locative prefixes *ku-*, *e-*; and prefix *e-* in combination with the locative suffix *–ini.* , e.g.

```
Adverb -> AdvPre NPrePre BPre NStem DimSuf
Adverb -> LocPre NPrePre BPre NStem DimSuf
Adverb -> LocPre NPrePre BPre NStem LocSuf
```

The latter exemplifies a long distance dependency between a locative prefix and a locative suffix. In particular, it is a circumfix or a co-ordinated pair consisting of a locative prefix which requires a locative suffix. This dependency cannot be completely captured by grammar rules since it requires idiosyncratic information about the specific noun stem.

4.3 Rules for morphophonologial alternations

Morphophonological alternations are the rules that determine the form of each morpheme. The rules for morphophonological alternations extracted from words based on noun stems in the Southern Ndebele corpus, are discussed by means of examples. In each case the morphemes that have undergone change are underlined in the surface word; then the lexical form is given, followed by the alternation rule in human readable form and an **xfst** representation.

We briefly explain the use of the lexical forms in identifying the specific rules that that were applicable to Southern Ndebele. Using the **lexc** morphotactics finite-state network on its own yields lexical forms as lower language strings.

These lexical forms contain special multicharacter symbols that are used in the **xfst** finite-state network to ensure that the rules fire correctly. We mention only a few. `^BR` and `^ER` denote the beginning and end of a word root.

[7] For conciseness of the grammar the class information is removed, since the focus is on the morpheme sequencing.

This is necessary for preserving the word root and to manage alternations that take place at word root boundaries; `^U`, `^MU`, `^I`, `^N`, `^SI`, etc. are placeholders for the noun prefixes to ensure that rules that apply only to such prefixes do not fire in cases where `u`, `mu`, `i`, `n`, `si`, etc. appear as other morphemes or parts of morphemes. They are finally removed by means of auxiliary rules. The `%` symbol is used in **xfst** to literalise special **xfst** symbols.

The use of the lexical form is illustrated by means of an example. In the first example it denotes the morpheme sequence `eisirhodloini. W`hen compared to the surface form *esirhodlweni* it is clear that the rules that fired must have been those that replace `ei` with *e* and `oini` with *weni*. By inspection the **xfst** rules

```
define VowelCombs  a e -> e , a i -> e ,
a o -> o , a u -> o , e a -> e , e i -> e ,
e u -> e , u a -> a , u o -> o;
```

and

```
define oiniRule o %^ER i n i -> w e n i;
```

may be identified as relevant and appropriate for Southern Ndebele.

Consonantalisation

esirhodlweni (‘in the court yard’)

`e^LP^I^SI^BRrhodlo^ERini`

Rule: o + ini > weni

```
define oiniRule o %^ER i n i -> w e n i;
```

Vowel coalescence

nenja (‘and the dog’)

`na^I^N^BRja^ER`

Rule: a + i > e

esinombala (‘that has the colour’)

`esina^U^MU^BRbala^ND^ER`

Rule: a + u > o

See the `VowelCombs` rule above.

Vowel elision

emthini

`e^LP^U^MU^BRthi^ERini`

Rules: e + u > e (where e is a locative prefix);   mu > m (where mu is followed by more than one syllable).

```
define muRule
%^MU -> [m | 0 ] || _ %^BR m
.o. %^MU -> m || _ [%^BR Syllable Syllable %^ER
| %^BR Syllable %^ER [Vowel | Syllable] | %^BR
Syllable Syllable]
.o. %^MU -> m || _ %^BR Vowel
.o. %^MU -> m u;
```

Palatalisation

emlonyeni

`e^LP^U^MU^BRlomo^ERini`

Rule: mo + ini > nyeni

```
define locRule m o %^ER i n i -> n y e n i
```

5. Results and discussion

We obtained the morphological grammar rules (in which the non-terminal symbols are self-explanatory labels/tags) and morphemes (terminal symbols) that are applicable to the words in the corpus. The morpheme sequencing rules below have been condensed somewhat, but still reflect the automatic extraction. These rules should still be subjected to human elicitation. The | is the union operator, (and) denote optionality and [and] are used to delimit the scope of the union operator.

S → Adverb|Copulative|Noun|Qualificative

Adverb → AdvPre NPrePre BPre NStem (DimSuf)

Adverb → AdvPre NPrePre NStem (DimSuf)

Adverb → NegPre PTSC AdvPre NPrePre BPre NStem

Adverb → [PTSC|SC|SitSC|SubjSC] AdvPre NPrePre BPre NStem

Adverb → LocPre NPrePre BPre NStem (DimSuf)

Adverb → LocPre NPrePre BPre NStem LocSuf

Adverb → LocPre NPrePre NStem DimSuf

Copulative → SC PreLoc-s LocPre (NPrePre) BPre NStem

Copulative → SubjSC PreLoc-s LocPre (NPrePre) BPre NStem

Copulative → CopPre NPrePre NStem (DimSuf)

Copulative → NegPre PTSC CopPre BPre NStem DimSuf

Copulative → NegPre SC CopPre BPre NStem (DimSuf)

Copulative → ([PTSC|SC|SitSC|SubjSC]) CopPre BPre NStem (DimSuf)

Copulative → [PTSC|SC|SubjSC] CopPre NPrePre BPre NStem

Noun → (NPrePre) (BPre) NStem (DimSuf)

Noun → NPrePre BPre NStem ([AugSuf|DimSuf])

Qualificative → NPrePre BPre NStem PossConc PronStem

Qualificative → PossConc NPrePre BPre NStem (DimSuf)

Qualificative → PossConc NPrePre NStem DimSuf

Qualificative → RelConc AdvPre NPrePre BPre NStem

Qualificative → RelConc CopPre BPre NStem (DimSuf)

Qualificative → RelConcPT AdvPre NPrePre BPre NStem

Qualificative → RelConcPT CopPre BPre NStem DimSuf

The $N \rightarrow \Sigma$ grammar rules are summarised in Tables 3-5. For example, *LocSuf → ini* is the last row in Table 4.

Cl	**N-Pre-Pre**	**Bpre**	**Poss-Conc**	**Rel-Conc/ PT**	**SC/ Subj SC/ SitSC**	**PT SC**	**Pron-Stem**
1	*u*	*mu*	*wa*	*o*	*u/a/e*		
2	*a*	*ba*	*ba*				
1a	*u*						
2a							
3	*u*		*wa*	*o*	*u*		
4				*e*	*i*	*ya*	
5	*i*	*li*	*la*	*eli*	*li*		
6	*a*	*ma/me*	*a*	*a*	*a/e*	*a*	
7	*i*	*si*		*esi*	*si*		*so*
8			*za*		*zi*		
9	*i*	*n*		*e*	*i*	*ya*	
10	*i*	*zin*	*za*		*zi*		
14	*u*	*bu*	*ba*				
15			*kwa*				
1pp				*esi*	*si*		
2ps				*o*	*u*		

Table 3: Prefixes that depend on class, number and person

Prefix	**Morpheme**
AdvPre	*na, nga*
CopPre	*ngu, wu, bu, ku, li, si, zi*
LocPre	*e, ku, o*

NegPre	*a*
PreLoc-s	*s*
AugSuf	*kazi*
DimSuf	*ana*
LocSuf	*ini*

Table 4: Other affixes

Zulu	Xhosa	Southern Ndebele
bala.3-4	*cabanga.11-10*	*bhudango.5-6*
dlebe.9-10	*hle.11-10*	*bizo.5-6*
khathi.7-8	*hlolo.1-2*	*dlebe.9-10*
lomo.3-4	*hlolokazi.1-2*	*kukurumbu.9-10*
suku.10-11	*nto.9-10*	*pungutja.5-6*
thongo.14	*phapha.5-6*	*rhodlo.7-8*
vila.5-6	*qadi.5-6*	*tjhada.5-6*

Table 5: An extract of noun stems with class information

The alternation rules are manually extracted from the 180 rule Nguni **xfst** script, as explained in section 4.3.

Observations

The experiment based on 180 words, focussed on noun stems, already covers a wide spectrum of the Southern Ndebele morphological grammar. Moreover, increasing the corpus will improve the rules (morphology), the word root lexicons and the affixes for all parts of speech.

Human elicitation was responsible for the removal of Class 11 concordial elements in Table 3 since this noun class does not feature in Southern Ndebele. The adverb in Southern Ndebele is not described by Jiyane (1994), therefore the adverbial prefixes and locative prefixes identified from the corpus and listed in Table 4, also call for human elicitation.

In Table 5, noun stem cross-linguistic similarities are illustrated. A total of 88 possible noun stems are identified in the analysis of the small representative Southern Ndebele corpus. Of these 80.6% (71 noun stems) are Zulu and (11.4%) 10 noun stems are Xhosa. The 7 noun stems (8%) listed under Southern Ndebele, are noun stems with relevant class information that are not shared with either Zulu or Xhosa. Here too, human elicitation will confirm the appropriateness of the Zulu and Xhosa noun stems together with their class information, in a Southern Ndebele context.

6. Conclusion and future work

The proof-of-concept corpus-based morphological grammar extraction procedure yielded a novel prototype language resource for Southern Ndebele. The procedure scales well. This resource is human-readable, adds to the description of the language and is also machine-readable, allowing parser development, and supports sustainability. Future work includes the extension of the grammar extraction procedure to all parts of speech; the application to larger corpora; a comprehensive evaluation of the approach; the extension of the proof-of-concept to a possible evaluation procedure for existing morphological parsers for the other Nguni languages; and the representation of the extracted formal grammar in XML as a *de facto* standard for sustainability.

7. Acknowledgement

The constructive comments of Ken Beesley and two anonymous reviewers are appreciated.

8. References

Antonsen, L., Trosterud, T., Wiechetek, L. (2010). Reusing grammatical resources for new languages. In *Proceedings of LREC 2010*, pp. 2782—2789.

Beesley, K.R., Karttunen, L. (2003). *Finite state morphology*. Stanford, CA: CSLI Publications.

Bosch, S., Pretorius, L., Fleisch, A. (2008). Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies*, 17(2), pp. 66--88.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the EACL 2009 Demonstrations Session*, pp. 29--32.

isiNdebele/English Dictionary. (2006). Johannesburg: Phumelela Books.

Jiyane, D.M. (1994). Aspects of isiNdebele grammar. MA Dissertation. University of Pretoria.

Lindén, K., Silfverberg, M., Axelson, E., Hardwick, S., Pirinen, T.A. (2011). HFST - Framework for compiling and applying morphologies in systems and frameworks for computational morphology. In *Communications in Computer and Information Science*, (100).

Maxwell, M. (2012). Electronic grammar and reproducible research. *Language Documentation and Conservation.* Preprint. To appear.

Nurse, D., Philippson, G. (2003). The Bantu languages. London: Routledge.

Poulos, G., Msimang, T. (1996). *A linguistic analysis of Zulu.* Pretoria: Via Afrika Limited.

Potgieter, E.F. (1950). Inleiding tot die klank- en vormleer van isiNdzundza, 'n dialek van Suid-Transvaalse Ngoeni-Ndebele, soos gepraat in die distrikte Rayton en Pretoria. MA Dissertation. University of South Africa.

Pretorius, L., Bosch, S.E. (2010). Finite-state morphology of the Nguni language cluster: modelling and implementation Issues. In A. Yli-Jyrä, Kornai, A., Sakarovitch, J. & Watson, B. (Eds.), *Finite-State Methods and Natural Language Processing 8th International Workshop, FSMNLP 2009. Lecture Notes in Computer Science*, Vol. 6062, pp. 123--130.

Simons, G.F., Bird, S. (2008). Toward a global infrastructure for the sustainability of language resources. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*.

Spiegler, S., Gol´enia, B., Shalonova, K., Flach, P., Tucker, R. (2008). Learning the morphology of Zulu with different degrees of supervision. In *Spoken Language Technology Workshop*, SLT. IEEE, pp. 9--12.

Wintner, S. (2007). Strengths and weaknesses of finite-state technology: a case study in morphological grammar development. *Natural Language Engineering*, 14(4), pp. 457--469.