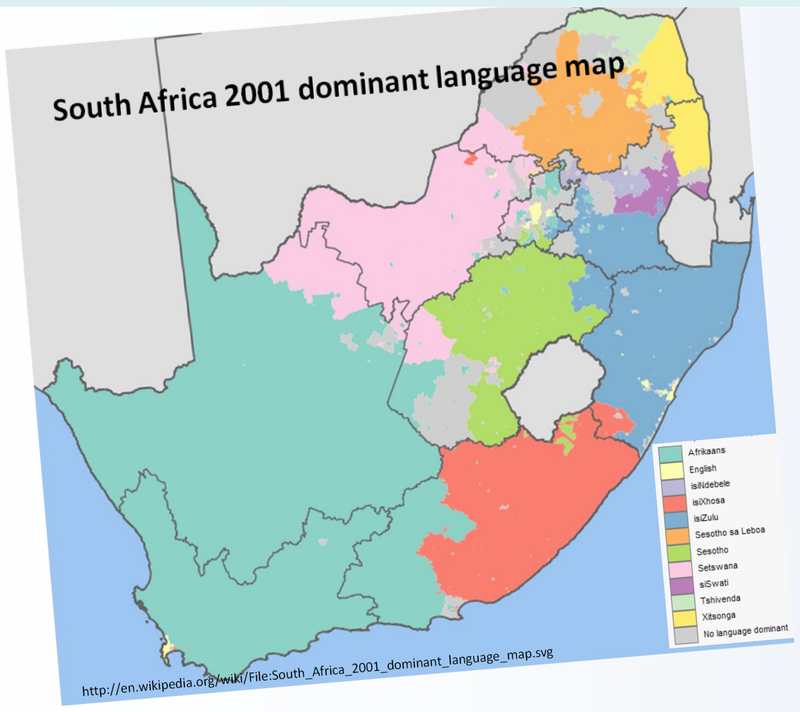
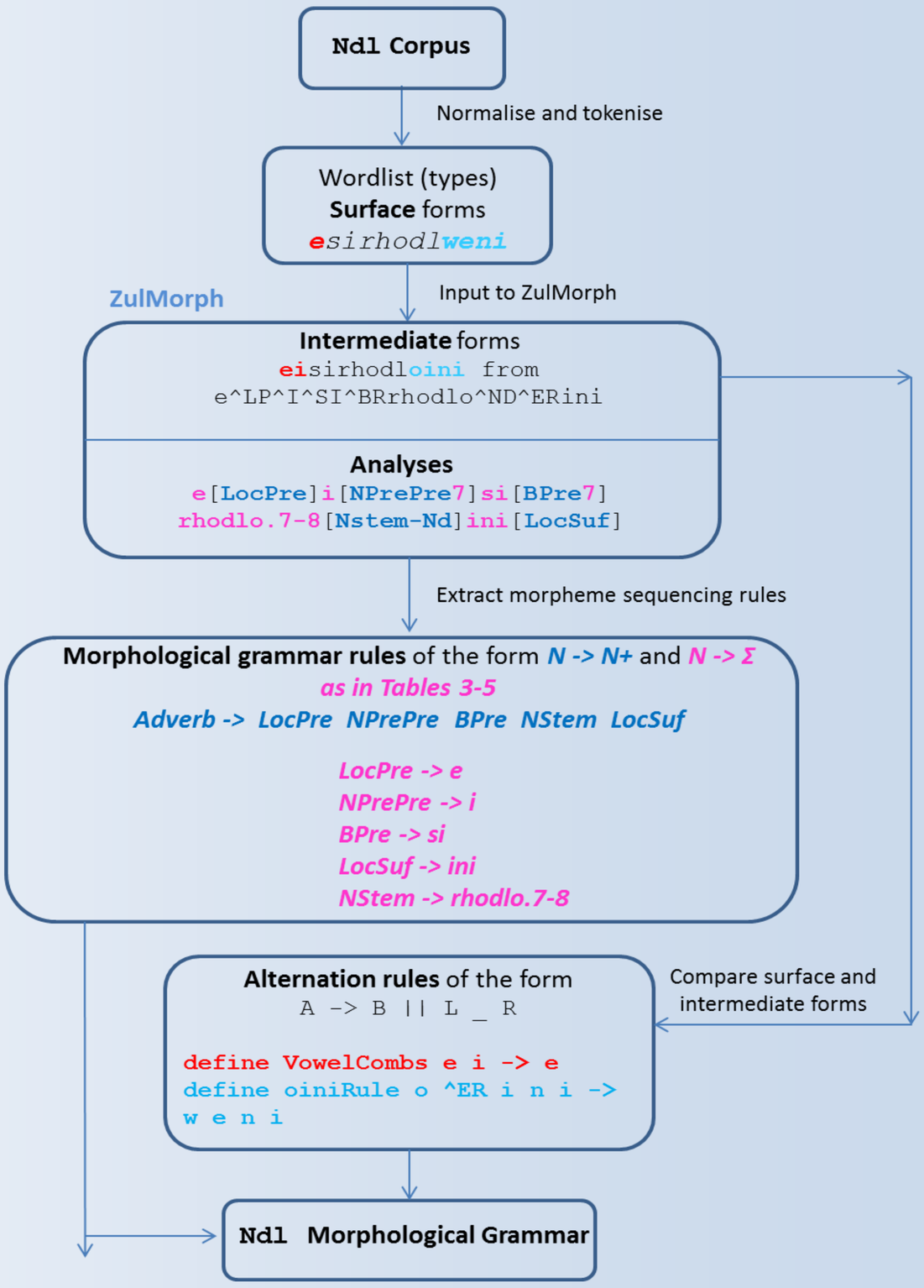
Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele

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 recognised best practices for the encoding of resources to ensure sustainability, access, and easy adaptability to future formats, lingware packages and development platforms.





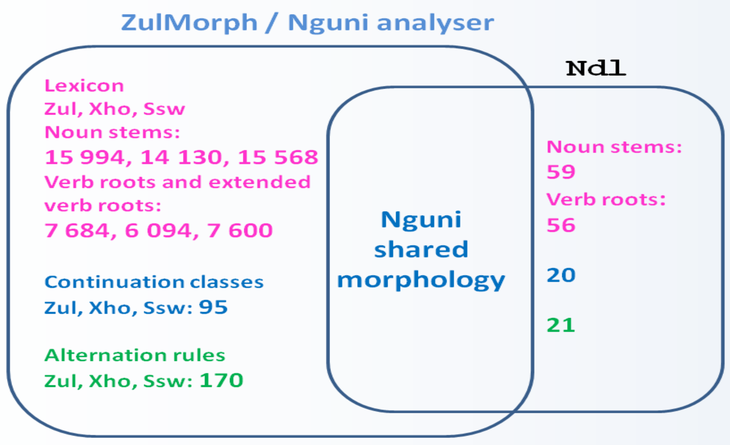
Core components of ZulMorph

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*)



Examples of grammar rules of the form *N -> N+*:

S → Adverb|Copulative|Noun|Qualificative
Adverb → [PTSC|SC|SitSC|SubjSC] AdvPre NPrePre BPre NStem
Copulative → ([PTSC|SC|SitSC|SubjSC]) CopPre BPre NStem (DimSuf)
Noun → NPrePre BPre NStem ([AugSuf|DimSuf])
Qualificative → NPrePre BPre NStem PossConc PronStem

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 http://www.dac.gov.za/newsletter/khariambe_3_4.html
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.

CONCLUSION

- Novel prototype language resource for Southern Ndebele;
- Scales well, is sustainable;
- Human-readable, descriptive, machine-readable, allows parser development.

FUTURE WORK

- Grammar extraction procedure - all parts of speech;
- Larger corpora;
- Comprehensive evaluation;
- Possible evaluation procedure for existing morphological parsers for the other Nguni languages;
- XML representation of the extracted formal grammar.