Computational Morphological Analysers and Machine-Readable Lexicons for South African Bantu Languages

Sonja Bosch, Jackie Jones, Laurette Pretorius, Winston Anderson
University of South Africa
PO Box 392, UNISA, 0003, South Africa
boschse@unisa.ac.za, jackiej@stthomas.co.za, pretol@unisa.ac.za, winston.anderson@btgroup.co.za

Abstract

In this paper the development of computational morphological analysers for six South African Bantu languages is discussed. Due to the rich agglutinating morphological structures of these languages, the morphological processing poses particular challenges. These challenges are of an orthographical, a morphological as well as of a lexical nature. The current status of the project is reported on, firstly in terms of the development of prototypes of morphological analysers for the various languages, and secondly in terms of the development of standardised XML machine-readable lexicons for the South African Bantu languages, based on an appropriate general data model.

1. Introduction

It is well known that localisation initiatives are supported by language translation, and in particular machine assisted and machine translation. However, there is much more to machine translation than meets the eye, especially in highly agglutinating languages such as the Bantu languages. The importance of morphological analysis is recognised as a basic enabling application for further kinds of natural language processing (NLP), including part-of-speech tagging, parsing, semantic analysis and information retrieval, and also for high-level applications such as spelling checking, lexicography, language teaching, text-to-speech systems, question answering and last but not least machine translation.

In order to be of practical use, such analysis needs to be automated and be based on underlying machine-readable lexicons that conform to common lexical specifications and de facto international standards to ensure their compatibility at international and multilingual level. The morphological analyser is regarded as the first in a series of text processing components.

Human language technologies (HLT) and NLP enable the electronic handling of both spoken and written language, and are aimed (amongst other things) at improving communication between humans and machines, as well as communication among humans. This is especially important in a country such as South Africa with its eleven official languages, which need to be developed technologically so that automated services can be rendered to citizens in their language of choice.

The development of computational morphological analysers for South African Bantu languages is linked to a project funded by the National Research Foundation in South Africa. The main research question in the project concerns the development of finite-state morphological analysers for five Bantu languages, namely Zulu, Xhosa Swati and Ndebele (belonging to the Nguni group of languages), and Northern Sotho and Tswana (belonging to the Sotho group of languages).

2. Challenges posed by Morphological Analysis of Bantu Languages

Automated morphological analysers exist for many European languages, but the development of morphological analysers has only been reported for a few Bantu languages, such as Swahili (Hurskainen 1992) and a few others in southern Africa (for example, Bosch & Pretorius 2003). Due to the rich agglutinating structures of these languages, the morphological processing poses particular challenges. These challenges are of an orthographical, a morphological as well as of a lexical nature.

In the case of the Nguni languages, a conjunctive system of writing is adhered to with a one-to-one correlation between orthographic words and linguistic words. For example, the Zulu orthographic word siyakuthanda (si-ya-ku-thand-a) 'we like it' is also a

[1] An earlier version of this paper was published in the Proceedings on the Workshop on Networking the Development of Language Resources for African Languages, 5th International Conference on Language Resources and Evaluation, 22 May 2006, Genoa, Italy, pp 38-43.

linguistic word. The Sotho languages on the other hand, are disjunctively written, and the above mentioned single Zulu orthographic word is written as four orthographic words in Northern Sotho, namely re a go rata (re a go rat-a) 'we like it'. These four orthographic entities constitute one linguistic word. It should be noted that, in contrast, the English orthographic words 'we like it' are three independent words that each have their own meaning and can stand alone.

The **orthographical challenge**, which lies in the writing conventions of the Bantu languages, may according to Hurskainen and Halme (2001, p.399), be ascribed to the fact that disjunctive writing systems "require a special treatment, before they can be analysed successfully". Pre-processing of the text in order to identify linguistic words, before morphological analysis takes place, is one of the options of addressing this challenge.

The morphological challenges in computational morphological analysis are twofold and comprise the modelling of two general linguistic components, namely morphotactics (word formation rules) as well as morphophonological alternations:

- The **morphotactics component** includes word formation rules, which determine the construction of words or word forms from an inventory of morphemes. This inventory of morphemes consists of word roots and affixes. Morphemes that make up words cannot combine at random, but are restricted to certain combinations and orders. A morphological analyser is required to recognise valid combinations of morphemes of the language in question.

- The morphophonological alternations component deals with the morphophonological changes between lexical and surface levels. A morphological analyser should identify the correct form of each morpheme since one and the same morpheme may feature in different ways depending on the environment in which it occurs.

The main **lexical challenge** in the building of morphological analysers for the Bantu languages is the fact that machine-readable lexicons, which are fundamental resources, are not readily available in any form. Although online dictionaries for Bantu languages are reported on by de Schryver (2003), such dictionaries available for Zulu and Xhosa for instance, contain a maximum of 2000 to 3000 lemmas and do not include explicit linguistic information, which is essential for a word root dictionary of the analyser. In the case of Northern Sotho, a bilingual online dictionary SeDiPro 1.0 (de Schryver 2003, p.10) containing over 20,000 entries is available with linguistic information. However, such online dictionaries are only accessible for look-up of individual words or word stems, and are not accessible as a whole.

3. Meeting the Challenges

3.1 Orthographical Challenges

The Sotho languages pose a pre-processing challenge in that the disjunctive orthographical tradition isolates as separate "words" what are essentially affix morphemes of a lexical unit. Thus in order to correctly analyse multi word lexical units morphologically without causing excessive ambiguity, a multi word tokeniser is required. For Northern Sotho, this was addressed by first constructing regular expressions to deal with all verb constructions and they were then extended to address all predicate constructions. These cater for the most complex multi word tokens in the Northern Sotho language.

The grammars historically cover the verbs and copulatives reasonably adequately but other research theses and more modern study grammars (for example, Louwrens 1989) had to be consulted to get consolidated views of these rules. None of the sources adequately covered what Ziervogel and Mokgokong (1985) term "deficient verbs", but in other texts are referred to as auxiliary verbs. A new linguistic research project is now under way to examine these in more detail.

There are various computational alternatives to producing a tokeniser (Hurskainen & Halme 2001). The Northern Sotho morphological analyser team chose the approach of using finite state software to construct the tokeniser (Beesley 2004). The tokeniser for Northern Sotho now adequately deals with all predicate clauses (verbs, auxiliary verbs and copulatives). For more information see Anderson and Kotze (2006).

3.2 Morphological Challenges

Since human language technology is a novel field of research in South Africa, especially in the field of Bantu languages, a team approach was decided on for the morphological analysis project. Each language team consists of a computer scientist and one or two linguists.

Morphological analysis in this project is based on a finite-state computational approach, using the natural language independent Xerox Finite-State Tools (Beesley & Karttunen 2003). This integrated set of tools is used to model and implement the complexities of word-formation rules as well as morphophonological alternations by means of finite-state networks. The latter are subsequently combined algorithmically into larger networks that perform morphological analysis.

The Xerox tools provide a declarative programming language, lexc (Lexicon Compiler) for specifying the required natural language lexicon and for modelling the morphotactic structure of the words in the language concerned.

Alternation rules are subsequently needed to map the abstract lexical strings into properly spelled surface strings, as they occur in the natural language. The alternation rules are formulated as regular expressions, and are then compiled into a finite-state network by means of the Xerox tool xfst.

In practical terms this means that all morphemes in the natural language need to be arranged in a cascade of LEXICONs (in a lexc description), while each entry in a LEXICON consists of morphological information and either a continuation class (the name of the next LEXICON in the cascade) or the end symbol #, which indicates the end of a valid morpheme sequence, as shown below in the example of a lexc description:

```
...
LEXICON NounPrefixes
...
i[NPrePre7]si[BPre7]:□I□SI
NStem;
i[NprePre8]zi[Bpre8]:□I□ZI
NStem;
...
LEXICON  NStem
...
gubhu                          NClass7-
8;
...
LEXICON NClass7-8
@U.CL.7-8@                        NomSuf;
...
LEXICON NomSuf
ana[DimSuf]:ana
#;
...
```

The **lexc** source file is then compiled into a finite-state network. This network recognises morphotactically well-formed, but still abstract morphophonemic or lexical strings such as
`i[NPrePre7]si[BPre7]gubhu[NRoot]ana[Dim].`
Alternation rules are subsequently needed to map these abstract lexical strings into properly spelled surface strings, as they occur in the natural language. The alternation rules are formulated as regular expressions, and are then compiled into a finite-state network by means of the Xerox tool **xfst.**

The orthographic changes that manifest between the lexical and surface words when morphemes are combined to form new words or word forms are described as illustrated in the following example:

b h [o|u] -> j || _ a n a

This alternation rule models the change of a bilabial sound -bh- appearing in the final syllable of a noun stem such as -gubhu to a palatal sound -j- when the diminutive suffix -ana is added to the noun stem.

The final step in the development of the morphological analyser is the combination of the **lexc** and **xfst** finite-state networks by means of composition (cf. Beesley & Karttunen 2003) into a single network, a so-called lexical transducer. This transducer constitutes the morphological analyser and represents all the morphological information about the language being analysed.

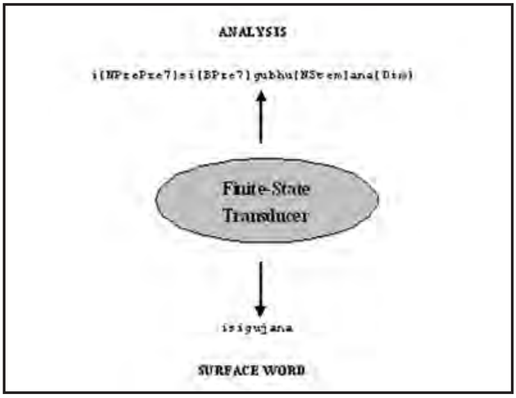

FIGURE 1 GIVES A SCHEMATIC REPRESENTATION OF THE APPLICATION OF A MORPHOLOGICAL ANALYSER.

In Figure 1 the morphological analyser maps the Zulu morphemes i-, -si-, gubhu and -ana to isigujana 'little calabash'. In other words, if the surface word isigujana constitutes the input string to the finite-state transducer, the output string is the morphological analysis which consists of the following morphemes in combination with their morphological feature tags: i[NPrePre7]si[BPre7]gubhu[NStem]ana[Dim].

The arrow in Figure 1 indicates the bidirectionality of the transducer and shows that analysis takes place in the upward direction while generation takes place in the downward direction. For more details regarding the Zulu morphological analyser prototype (ZulMorph) see Bosch and Pretorius (2003) as well as Pretorius and Bosch (2003a) and (2003b).

3.3 Lexical Challenges

In addressing this problem of unavailability particularly for Zulu and Xhosa, a lemma list in electronic format was extracted from a Zulu paper dictionary (Doke & Vilakazi, 1964). For Xhosa however, the resources available in terms of lemmas were even more limited. This therefore demanded a time consuming exercise of extraction of lemmas from existing Xhosa paper dictionaries. Lemmas were retyped from a number of dictionaries and the scanning and proof reading of these resources increased and contributed to the development of the lemma lists substantially. These various sources yielded data in largely varying formats and forms containing many inaccuracies and errors. The non-existence, but urgent need for lemma lists for Xhosa also created the opportunity for researchers to devise a practical compilation procedure in accordance with appropriate standards in order to ensure reusability. The procedure for producing a large and reliable collection of Xhosa nouns and verb stems from this data consisted of a semi automated data validation phase and, in the case of nouns, an automated generation phase. Data inconsistencies were identified by means of Perl style pattern recognition, then scrutinised and corrected by the linguists in the team in the data validation phase. The validated data formed the input to the automated generation phase. Nouns were generated in two formats. The first of these was for human readability and the second was in an XML document. The second of these is particularly important in the creation of reusable lexical resources for future applications. The only available Swati paper dictionary is being scanned and proofread in stages and then included into an electronic lemma list. Similarly for Tswana a paper dictionary has been scanned and is in the process of being proofread also to be developed into a lemma list for use in morphological analysis.

Regarding word lists for Northern Sotho, the major dictionaries were examined. The largest, the Comprehensive Northern Sotho dictionary (Ziervogel & Mokgokong 1985) includes support for the extra vowels (beyond the five standard vowels) marked with a circumflex in Northern Sotho, as well as support for the letter š and its capitalised form. Furthermore, the comprehensive dictionary includes tone markings on each main entry. In order to obtain an accurately scanned word list these characters needed to be recognised by optical character recognition (OCR) software. No standard Northern Sotho OCR packages are available, so standard language settings were used. The scanning errors are consistent with the incorrectly scanned characters. Therefore, Perl scripts were developed to automatically correct the incorrectly optically recognised text. A further process of human editing is now underway to confirm all corrections. Subsequent to the dictionary scan, many other works have been scanned to add to the Northern Sotho test corpora. Eastern European language recognisers, such as Czechoslovakian, have proved most effective in recognising characters due to their adequate handling of š and its capitalised form (cf. ABBYY FineReader 2007).

In terms of the lexical challenges our ultimate aim is to develop from these above-mentioned word lists and paper dictionaries machine-readable lexicons according to a standardised data model in XML format that would be applicable to all the languages under investigation.

4. Current Status of the Project

The current status of the project is reported on, firstly in terms of the development of prototypes of morphological analysers for the various languages, and secondly in terms of the development of machine-readable lexicons for the South African Bantu languages, based on the above-mentioned proposed data model.

4.1 Analyser Prototypes for the various Languages

The Zulu analyser prototype (ZulMorph) at present covers most of the morphotactics and morphophonological alternations required for the automated analysis/generation of nouns of all classes, the positive and negative forms of verbs (including object concords, tense morphemes, aspectual prefixes and verbal extensions), pronouns, the demonstrative and copulative demonstrative, underived adverbs, relatives and adjectives, possessives, conjunctions and ideophones. Word categories that still need to be completed are compound tenses of the verb, and derived adverbs. Preliminary testing of the current prototype was done on a test corpus consisting of 30,000 types. The application of the morphological analyser to the test corpus results in the recognition of approximately 77% of the types in the corpus. This result may be ascribed partially to morphological constructions that have not been dealt with completely, but mainly to roots that do not yet occur in the root lexicon. By design the morphological analyser includes LEXICONs (lists) of noun stems, verb roots, relative stems etc. and therefore it only analyses words based on roots or stems that feature in this so-called underlying lexicon.

Since individual words in wordlists are analysed in the morphological analysis component, the ambiguity rate is high, and each word is assigned all its possible readings or analyses. For an application such as a second generation spelling checker that has some form of automatic morphological analysis implemented as a part of the spelling checker (without grammar checking), such ambiguity poses no problems. The reason is that the emphasis is on lexical recall or the recognition of correctly spelled words by the spelling checker irrespective of their context (cf. Bosch & Eiselen 2005).

In the case of running text being analysed, the challenge in the subsequent phase of the project is the elimination of ambiguity or contextually inappropriate readings such as the following:

```
bakhe   ba[PossConc2]khe[PronStem]
bakhe   ba[PossConc14]khe[PronStem]
bakhe   ba[SC2]akh[VRoot]e[VerbTermPerf]
bakhe  bu[SC14]akh[VRoot]e[VerbTermPerf]
```

These analyses of bakhe illustrate ambiguity not only regarding class concord information, i.e. classes 2 and 14, but also stem and root information. The first two examples are analysed as possessive pronouns while the latter two are analysed as verbs.

The research aims for the other Nguni languages in the project, i.e. Xhosa, Swati and Ndebele closely follow those for Zulu, since all these languages follow a conjunctive writing system. This enables the fast-tracking of the development of the morphological analysers for the Nguni languages by adapting the Zulu continuation classes and rules. Implementation and testing of the model in terms of the Xerox finite state tools are already in progress.

Regarding Northern Sotho a framework is under development for the nominal and verbal structures of Northern Sotho, with special emphasis on establishing the order of verbal extensions, reduplication patterns in nouns and verbs, as well as formalising rules for the derivation of morphological processes that involve the phonological process palatalisation in the formation of passives and diminutives. Implementation and testing of the Northern Sotho prototype (NsoMorph) is based on a limited, though representative lexicon, while cleaning up a scanned version of a Northern Sotho dictionary is in progress. The first prototype of a morphological analyser for Tswana (TsnMorph) is being developed with nouns being treated first, while other word categories are added systematically.

Progress with the development of analyser prototypes for the various Bantu languages in the project has been reported in a number of publications (cf. University of South Africa 2007).

4.2 Development of Machine-Readable Lexicons

By definition the analyser can only recognise and analyse words of which the roots/stems have been explicitly included in its embedded lexicon. Ideally, a comprehensive machine-readable lexicon in the form of an XML document should be available for each language as a basic resource from which word roots/stems may be obtained.

As stated previously, electronic lemma lists for Xhosa and Zulu have been developed albeit on a small scale. To date lemma lists for these two languages, extracted from paper dictionaries, contain a total of over approximately 28,000 entries each.

In order to eventually arrive at an XML lexicon structure, the underlying standardised data model needs to be formulated and verified first. This is the subject of recent work in this regard (Bosch et al. 2006) where a data model towards a standardised machine-readable lexicon for all languages in the project is developed and formulated as an XML DTD. This model aims to ensure maximum inclusiveness of all linguistic information and to provide flexibility and handle the various representations applicable to Bantu languages in particular. It is therefore applicable to diverse uses of electronic lexicons ranging from research in numerous areas resulting in publication. Included in this data model are particular requirements for complete and appropriate representation of linguistic information as identified in the study of available paper dictionaries. As starting point the extent to which the Bell and Bird (2000) data model may be applied to and modified for the above-mentioned languages was investigated. It was found that changes to this data model were necessary to make provision for the specific requirements of lexical entries in the relevant languages.

Our model differs in various ways from the Bell and Bird model. The latter model was originally designed for descriptive purposes while our model is primarily for computational use, where the emphasis is on marking up lexicon and linguistic information for logical structure in order to provide essential information for the computational language processing task concisely, precisely and unambiguously. Examples are the representation of class information, singular and plural, locative formation (derivation) in the case of nouns, and verbal extensions (derivations) in the case of verbs. Further examples are the identification of specific socio-linguistic features in Xhosa and Zulu such as isiHlonipho sabafazi (married women's language of respect) and Xhosa isiKhwetha (male initiates' language) both features of which would also necessitate explicit representation in the lexicon. Another area where the Bell and Bird model seemed inadequate to accommodate the South African Bantu languages was the exclusion of the appropriate nesting of derived forms so prevalent within these languages. Other aspects of interest include our stem entry approach, the reflexive form of the verb, and the desirability or not of recursion in machine-readable lexicons.

The proposed data model seems to provide flexibility and handles the various representations applicable to Bantu languages in particular and is therefore applicable to diverse uses of machine-readable lexicons. Our hope is that it will contribute to the further discussion and development of a common scheme for storing lexical data not only for the South African Bantu languages, but for the Bantu language family as a whole.

5. Conclusion and Future Work

Morphological analysis is generally recognised as a technology that enables the development of more advanced tools and practical applications in various areas of natural language processing, such as part-of-speech tagging, syntactic parsing, text-to-speech systems, information extraction, and machine translation. Research in this project concerning the development of computational morphological analysers for South African Bantu languages has confirmed the importance of comprehensive machine-readable lexicons as fundamental resource of the morphological analysers. The current project in computational morphological analysis includes research into the development of automated morphological analysers for Zulu (ZulMorph), Xhosa (XhoMorph), Swati (SswMorph), Ndebele (NblMorph), Northern Sotho (NsoMorph) and Tswana (TsnMorph), using finite-state methods in computational morphology. It is envisaged that the project will eventually cover all South African Bantu languages.

Further aims of the project are:

- Wider distribution of the intermediate versions of the electronic lexicon for constructive feedback from the broader community of lexicographers and other users/speakers of the relevant languages.
- Investigation into a disambiguation component, the task of which is to eliminate contextually inappropriate readings.
- Research into lexicon design and development in order to contribute to the international definition of standards as envisaged by the International Standards Organisation ISO/TC37/SC4, whose goal it is to develop a platform for the design and implementation of linguistic resource formats and processes in order to facilitate the exchange of information between language processing modules (Romary and Ide 2002).
- Research into place names as occurring in the various languages of the project, for inclusion in the relevant machine-readable XML lexicons.

6. Acknowledgements

This material is based upon work supported by the National Research Foundation under grant number 2053403. Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Research Foundation.

7. References

ABBYY FineReader. (2007). [O]. Available at: http://www.abbyy.com/finereader_ocr/
[Accessed on 31 May 2007].

Anderson, W.N. & Kotze, P.M. (2006). Finite State tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of Northern Sotho. In Proceedings of the 5th International Language Resources and Evaluation Conference, Genoa, Italy.

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

Beesley, K.R. (2004). Tokenizing Transducers. Xerox Research Centre. Europe. Unpublished course notes presented in Pretoria, South Africa, September 2004.

Bell, J. & Bird, S. (2000). A Preliminary Study of the Structure of Lexicon Entries. [O] Available at: http://www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html
[Accessed on 19 September 2005].

Bosch, Sonja E & Roald Eiselen. (2005). The effectiveness of morphological rules for an isiZulu spelling checker. South African Journal of African Languages 25(1) pp. 25-36.

Bosch S.E. & Pretorius, L. (2003). Building a computational morphological analyser/generator for Zulu using the Xerox finite-state tools. In Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, 10th Conference of the European Chapter of the Association for Computational Linguistics, April 13-14 2003, Budapest, Hungary. ACL. pp. 27-34.

Bosch, S.E. & Pretorius, L. (2004). Software tools for morphological tagging of Zulu corpora and lexicon development. In Proceedings of the 4th International Language Resources and Evaluation Conference, Lisbon: ARTIPOL , vi, pp. 1251-1254.

Bosch. S.E., Pretorius, L. & Jones, J. (2006). Towards machine-readable lexicons for South African Bantu languages. In Proceedings of the 5th International Language Resources and Evaluation Conference, Genoa, Italy.

De Schryver, G-M. (2003). Online Dictionaries on the Internet: An Overview for the African languages. Lexikos, 13, pp. 1-20.

Doke, C.M. & Vilakazi, B. (1964). Zulu-English Dictionary. Johannesburg: Witwatersrand University Press.

Hurskainen, A. (1992). A two-level formalism for the analysis of Bantu morphology: an application to Swahili. Nordic Journal of African Studies, 1(1), pp. 87-122.

Hurskainen, A. & Halme, R. (2001). Mapping between Disjoining and Conjoining Writing Systems in Bantu Languages: Implementation on Kwanyama. Nordic Journal of African Studies, 10(3), pp. 399-414.

Louwrens, L.J. (1989). Northern Sotho. Study guide for Grammar. University of South Africa: Pretoria.

Pretorius, L. & Bosch, S. (2002). Finite-State Computational Morphology - Treatment of the Zulu Noun. South African Computer Journal, 28, pp. 30-38.

Pretorius, L. & Bosch, S. (2003a). Finite-State Computational Morphology: An Analyzer Prototype for Zulu. Machine Translation, 18, pp. 195-216.

Pretorius, L. & Bosch, S. (2003b). Towards technologically enabling the indigenous languages of South Africa: the central role of computational morphology. Interactions of the Association for Computing Machinery, 10(2), pp.56-63.

Romary, L. & Ide, N. (2002). Standards for Language Resources. In Proceedings of the 3th International Language Resources and Evaluation Conference, 1, pp. 59-65.

University of South Africa. Department of African Languages. (2007). [O]. Available at: http://www.unisa.ac.za/africanlanguages [Accessed on 29 May 2007].

Ziervogel, D & Mokgokong, P.C. (1985). Comprehensive Northern Sotho dictionary. Second corrected edition J.L. van Schaik: Pretoria.