Development of reusable resources for Human Language Technologies (HLT) applications: practice and experience

Jackie Jones
Department of African Languages,
University of South Africa, PO Box 392,UNISA 0003, Pretoria, South Africa
jonescjj@unisa.ac.za

Sonja E. Bosch
Department of African Languages,
University of South Africa, PO Box 392,UNISA 0003, Pretoria, South Africa
boschse@unisa.ac.za

Laurette Pretorius
School of Computing, University of South Africa, PO Box 392,
UNISA 0003, Pretoria, South Africa
pretol@unisa.ac.za

Danie Prinsloo
Department of African Languages,
University of Pretoria, Pretoria, South Africa, 0002
danie.prinsloo@up.ac.za

Language resources, by their very nature, serve as a repository of linguistic knowledge. They are therefore essential in the building and improvement of natural language applications. The aim of this paper is to elaborate on the practice and the experience gained in the development, maintenance and management of such resources with specific reference to African languages. The focus is on the methods of collection and the formats concerning word lists, morphological analysis and lemma lists. The resources discussed, are those developed in collaborative research with North-West University's Spelling Checker Project. As a broader perspective, the reusability of such resources is highlighted. Recommendations are also made regarding the way forward nationally in developing a resource centre to facilitate the technological development of South African Bantu languages.

Introduction

Language resources serve as a repository of linguistic knowledge and are acknowledged as central to the development and improvement of natural language processing. Due to the prohibitive cost involved in developing resources, the coordination and reusability of such resources becomes vital to future research in this field. With an understanding of these statements it is the aim of this paper to elaborate on the practice and the experience gained in the development, maintenance and management of such resources with specific reference to African languages. The resources which are discussed are those accumulated in collaborative research with North-West University's Spelling Checker Project. This interdisciplinary project comprised research teams that included linguists representing the various African languages – Zulu, Xhosa, Northern Sotho and Tswana together with computer scientists from the University of South Africa, and North West University.

The African continent boasts some 2035 African languages (Heine and Nurse, 2000:1) that represent nearly one third of the world's languages. The African languages are grouped into four major groups: Afroasiatic, Nilo-Saharan, Niger-Congo and Khoisan. The Bantu language family, comprising an estimated 1436 languages (Williamson &

Blench, 2000), falls within the Niger-Congo Group. In South Africa four groups of Bantu languages are recognized, namely Nguni, Sotho, Tsonga and Venda. These languages share structural similarities and linguistic devices such as grammatical agreement, tonal distinctiveness and the use of ideophones to name but a few. All Bantu languages have a concatenative morphology and are therefore agglutinative by nature. This means that the majority of words comprise a number of morphemes, each of which carries meaning, and may never occur independently. Furthermore the concept 'word' is much more encompassing than entries found in a dictionary.

It should be noted that orthographically, the Sotho group of languages differs substantially from the Nguni group in that these languages use a disjunctive as opposed to a conjunctive style of writing. This may be exemplified as follows:

(1)	Zulu	Northern Sotho	
	Uyahamba	*O a sepela*	'He/she/it is going'
	Bayahamba	*Ba a sepela*	'They are going'
	Uyathanda	*O a rata*	'He/she/it loves'
	Bayathanda	*Ba a rata*	'They love'

The agglutinative qualities as well as conjunctive as opposed to disjunctive writing styles are both significant factors contributing to the complexity of spell checking in the Bantu languages.

This paper traces and explains the procedure followed in the process of collection, development and collation of resources compiled from varying diverse sources for each of the abovementioned languages. The emphasis is on decisions made and approaches followed that are practical, and based on experience gained with the development of language corpora for these languages. These decisions and approaches may be viewed as best practices with respect to appropriate standards and protocols. As the focus of this paper is on the preparation, collection and management of resources to ensure reusability, the term 'resources' including their uses and the specific resources required for this project will be discussed. Furthermore the methods used in the collection and collation of resources used are explained.

The volume of data collected during the compilation necessitated the management and maintenance according to certain standards and protocols. This was essential to ensure correctly validated and quality controlled lists that could in turn be preserved for possible reuse. The practical procedures followed in this project are exemplified with recommendations for future storage of data, guaranteeing accessibility to all for research purposes in the African languages.

Having documented the procedures followed in the development of lexicons and the management and maintenance of the data, recommendations are then made on 'the way forward'.

Language Resources[1]

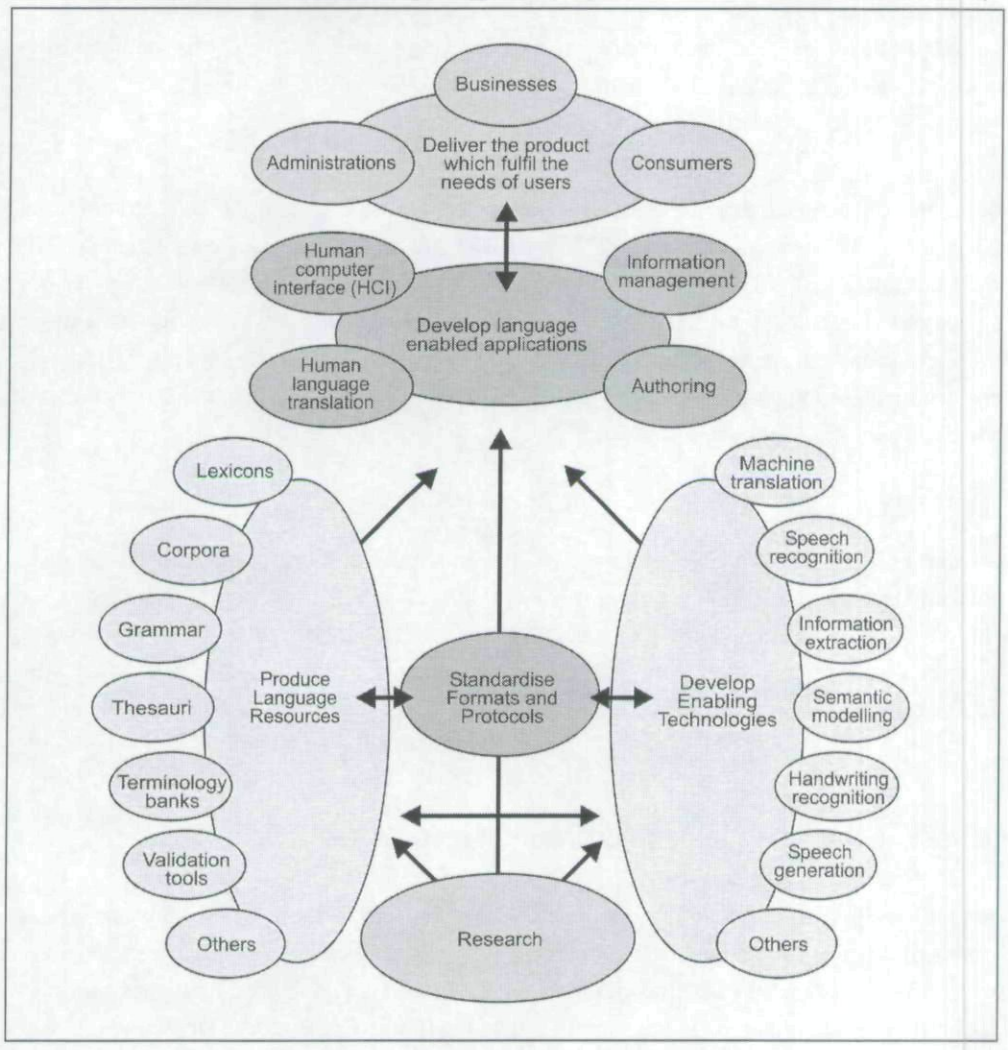
What are Language Resources?

According to Godfrey and Zampolli (1997:381) the 'term *linguistic resources* refers to (usually large) sets of language data and descriptions in machine readable form, to be used in building, improving, or evaluating natural language (NL) and speech algorithms or systems'.

Linguistic resources include written and speech corpora, lexical databases, grammars and terminologies to name but a few. From the schematic representation in Figure 1 it is clear that resources are a major component involved in the process of the development of enabling technologies. However, in order for these developments to take place, the critical issue is the adherence to the standardization of formats and protocols. From Figure 1 it also becomes very clear how central this component of standardization is in the management and maintenance of resources for purposes of reusability. Godfrey and Zampolli (1997:381) are of the opinion that while there seems

to be an increasing awareness of the significant impact both social and economic of natural language and speech systems, these need to be economically viable and have 'real life' uses. It is also interesting to note that the same authors propose that the lack of adequate resources for the majority of languages could be attributed to the tendency to test linguistic hypotheses using small amounts of critical data and the high cost involved in creating linguistic resources. Another controversial issue raised was the reusability and multifunctional aspect required of linguistic resources. It was suggested that the solution would be to attempt a consensus among different theoretical perspectives and systems design approaches (Godfrey & Zampolli, 1997:381).

Figure 1: Language engineering harnessing the power of language (http://www.hltcentral.org/usr_docs/project-ource/en/broch/harness.html)



What are the uses of Language Resources?

Language resources are a cornerstone in the development of language enabling technologies and it is for this reason that we stress the need for serious consideration of a national resource centre where resources compiled according to standardized formats and protocols are stored and entrenched for reusability and developmental purposes. Figure 1 provides some insight into the enabling technologies that can be developed from using the essential initial component namely language resources. These resources are used both for research and commercial applications. Written language corpora for example may be used for the development of spelling checkers, for hyphenation purposes, grammar checkers, word use and word collocations, building of glossaries for translation purposes, part of speech taggers and partial parsers. These resources therefore provide the products to the end users, be they in research or commerce and industry.

What resources are required in the development of spelling checkers?

For the purposes of this project, which included Bantu languages, the resources differed according to the linguistic characteristics of the languages concerned, specifically the disjunctive vs. conjunctive writing styles. Overall, the three types of resources required were word lists, morphological analysis and lemma lists.

Word lists

What was required was tens of thousands or preferably hundreds of thousands of orthographic words for each of the languages Northern Sotho, Tswana, Xhosa and Zulu. These lists could be alphabetical or frequency word lists also known as **types** in corpus linguistics or simply the **lexicon** of each language. Word lists played a decisive role in all of the spelling checkers compiled for these languages. For the disjunctively written languages Northern Sotho and Tswana more than 95% lexical recall is achieved by simply using word lists and for the conjunctively written languages Zulu and Xhosa, up to 80% lexical recall.

Morphological analysis

The Bantu languages, being mainly of an agglutinating nature, entail extensive use of prefixes as well as suffixes in the formation of words. The root is the constant core element from which words or word forms are constructed while the rest is inflection and derivation. The Nguni languages (Zulu, Xhosa, Swazi and Ndebele) follow the convention of a conjunctive writing system, that is a system in which the morphemes constituting a word appear as a single token. Each linguistic word consists of a number of bound parts or morphemes that can never occur independently as separate words. Therefore in these languages, morphological analysis is essential for the development of any kind of text processing such as spelling checking.

Lemma lists[2]

Furthermore, machine-readable lemma lists are needed as a basic resource for the morphological analyses. In order to ensure portability and reusability of this resource, it is developed as an XML document. XML is the *de facto* standard for machine-readable text documents and makes such a lexicon suitable to function as a crucial, integral language resource for a wide range of applications. A present limitation in the natural language processing of languages such as Zulu and Xhosa is the fact that machine-readable lemma lists are not readily available in any form.

What methods were used in the collection and collation of resources in this project?

Methods used in the collection of word lists

For the North West University's Spelling Checker Project existing word lists of Northern Sotho, Tswana, Xhosa and Zulu of the Department of African Languages at the University of Pretoria dating back as far as 1930 were taken as a point of departure. These lists were extended mainly with words used in class notes, dialogues and short stories in the language laboratory manuals and magazines published by the former Government Department of Co-operation and Development such as the *Tšwelopele*-series for Northern Sotho, *Tswelelopele* for Tswana, *Intuthuko* for Zulu, *Inkqubela* for Xhosa, etc. Words culled from transcriptions of news bulletins and video material such as the *Lafata*-series donated to the mentioned Department of African Languages by the *SABC* were added. The lists were also supplemented by words from dictionaries, books, magazines and electronically available texts such as dissertations completed in the Department of African Languages. The lists were further extended by means of spelling checking texts on the Internet and by word lists generated from electronic corpora for these languages. The latter strategy was used mainly to sort the word lists according to frequency of use in Northern Sotho, Tswana, Zulu and Xhosa and to confirm or validate their existence and spelling in these languages. For future applications corpora are likely to be used as the prime sources for word list generation. The (re)usability of such corpora and word lists will be discussed in more detail below.

Methods used for limited morphological analysis

The conjunctive writing system within the Nguni group of languages becomes significant in the development of spelling checkers since Nguni languages, in addition to word lists, need automatic morphological analysis in deciding whether a word is spelled correctly or not.

Spelling checking in agglutinative languages such as the Bantu languages differs significantly from that in languages such as English, in the sense that the former are characterised by words consisting of more than one morpheme. Words are formed by productive affixations of derivational and inflectional morphemes to roots or stems like 'beads-on-a-string' (Oflazer & Güzey, 1994:1). The concept of a word in such languages is therefore much wider than simply the entries found in a dictionary. Experience has shown that for a language such as Northern Sotho (a disjunctively written Bantu language) satisfactory results are obtained with relatively small lexicons (i.e. fewer than 100,000 words). A lexical recall of more than 98% is obtained with little or no additional morphological analysis. However, in the case of a Zulu spelling checker with a lexicon of more than 200,000 words, the lexical recall amounts to a mere 89% (cf. Bosch & Eiselen, 2005).

It is for this reason that word lists are not the (sole) answer to spelling checking in the conjunctively written Bantu languages. Beesley and Karttunen (2003:451) confirm that

> Many traditional spelling checkers are based on simple word lists extracted from corpora, and this approach is known to be problematic for highly inflecting languages.

Morphological analysis is essential for a spelling checker in a conjunctively written agglutinative language since one can only start looking up a word in a dictionary once one has reached the constant core element, namely the root. The identification of each morpheme in a morphological analysis deals with endless possibilities due to the productive system of word formation by means of affixation to roots or stems. The morphemes that make up words cannot combine at random but are restricted to certain combinations and orders. A morphological analyser tests the validity of combinations of morphemes (morphotactics, also known as word-formation rules).

One and the same morpheme may be realized in different ways depending on the environment in which it occurs. Morphemes may undergo changes (morphological alternations) at boundaries due to various phonetic interactions. Again, a morphological analyser recognizes the correct form of each morpheme. It is for instance specified in the analyser that palatalization can occur with certain verbal extensions (passive), noun suffixes (diminutives), etc. Other morphophonological processes that have to be provided for in the analyser are vowel elision, vowel coalescence and consonantalization, which often occur across morpheme boundaries.

For applications such as these specific spelling checkers, tight project schedules did not allow adequate time to develop sophisticated morphological analysers. Instead of using finite-state morphological analysers for Zulu and Xhosa[3], regular expressions, based on lemmatization or stemming through lexical decomposition, as well as rewrite rules were implemented (cf. Bosch & Eiselen, 2005, for a detailed discussion). Basically the surface form of a word as it appears in a text, is associated with its relevant morphological information. Although the number of derivational and inflectional morphemes is finite, the word formation rules of Zulu result in a combinatorial explosion in the number of words that can be formed. Therefore the morphological analysis largely depends on **lemma lists** in order to perform the basic lookup procedures required in the morphological analysis up to stem and root identification level.

Methods used in creating lemma lists

Machine-readable lexicons are an invaluable fundamental resource for language processing. Basic lexicons typically include details of word-specific information such as the sound structure (phonology) and the grammatical structure of each word or word form (morphology), complement structures of each word (syntax) and also the meaning of the word in different contexts (semantics). Electronic versions of published dictionaries are examples of lexicons, which have become available as resources in machine-readable form, e.g. *Longman's dictionary of*

contemporary English and the *Oxford advanced learner's dictionary*. However, this type of machine-readable lexicon is not readily available for languages such as Zulu or Xhosa.

A lemma list in electronic format extracted from a Zulu paper dictionary (Doke & Vilakazi, 1964), containing a total of over 28,000 entries was available for use in this project. For the purposes of the spelling checker, a list of about 13,000 nouns and approximately 9,000 verb stems, based on the abovementioned list proved an invaluable resource in the morphological analysis based on lemmatization and affix stripping. The lemmatization strategy removes all possible affix combinations, and then looks up the remainder in the list of possible lemmas.

For Xhosa however, the resources available in terms of lemmas were limited. This therefore demanded a time consuming exercise of extraction of lemmas from existing Xhosa paper dictionaries. Lemmas were retyped from *The English-Xhosa/Xhosa English dictionary* (Via Afrika), *A new concise Xhosa English dictionary* (McLaren) and also the *Kafir-English dictionary* by Kropf and Godfrey (1915). Subsequently lemmas from a scanned version of *The greater dictionary of Xhosa* (Vol 3) were included, while volumes 1 and 2 of this dictionary eventually became available electronically. These yielded data in largely varying formats and forms containing many inaccuracies and errors. The non-existence, but urgent need for lemma lists for Xhosa afforded the authors the opportunity 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 scrutinized 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 format was designed for human readability to make it easy for the linguists to check the generated results and verify that the generated nouns were valid Xhosa nouns, singular and plural. The second format was an XML document, constituting a rudimentary machine-readable lexicon. The benefit obtained from the latter format is of particular significance. XML has become a *de facto* standard in the creation of reusable lexical resources and offers an efficient, portable and reusable way of capturing language data and preserving the integrity thereof. A similar process was followed for the verb stems.

Some 27,240 Xhosa lemmas were collated of which 20,845 nouns and 6093 verb stems were used in the morphological analysis using the same method as employed for the Zulu analysis.

The practical procedure

We briefly discuss the raw data that formed the basis for the creation of the required Xhosa noun list after which the process followed in the generation of these lists is outlined. Examples are provided for illustrative purposes. A similar process was followed for the generation of a Xhosa verb stem list. The automation was done by means of pattern matching with Perl.

The justification for **generating** the noun list may be found in the fact that the available data typically included only the singular form of the noun together with its class number information. However, the plural form of the noun may in general be consistently obtained from this information by means of the nominal classification system that governs Xhosa noun morphology.

Finally we focus on the outcomes of this process and their potential for (re)use as language resources. The guiding principle is to present *every* word (contiguous sequence of characters) in the data to the Xhosa linguists for consideration in order to maximise completeness and correctness of the lists, while also minimizing human, particularly linguists' effort and time involved.

The data

All the available data, obtained from various sources by various means, were used. This data exhibited largely varying formats and contained numerous types of inconsistencies and errors due to typing, scanning, etc. Depending on the source, standard formats varied, as shown by the examples in Table 1:

Table 1: Variation of standard formats

Raw data file	Example of generic entry in file	Number of entries in file
1	*umlala* 3/4	8638
2	*umfana* 1/2	677
3	*um - qathango* 3	1893
4	*i - vazi* 9/10	1868
5	*i-baba* 5/6	906
6	*isi-camango* 7/8	418
7	*um-gcini* 1/2	3487
8	*umthi* 2	2329
9	*umbhalo* 3	311

A closer look at these generic entries shows that in some files

- the noun prefix was not separated from the root (files 1, 2, 8, and 9)
- the prefix and root were separated by a '-' with blank spaces on either side (files 3 and 4)
- the prefix and root were separated by a '-' without blank spaces (files 5, 6, and 7)
- only the singular class number was given (files 3 and 9)
- the singular and plural class numbers were given (files 1, 2, 4, 5, 6, and 7)
- the class numbers were given in terms of Meinhof's classification system (files 1–7 and 9)
- the class numbers were given in terms of Doke's classification system (file 8).

Common errors and inconsistencies in the data included the following:

- Strange/unexpected characters, for example

2) `umbikanye [*4] 1a/2a`[4]

```
iphiko *-piko 5/6
ilektsha / ileksha 9/10
ilakana/ilakane/ilakani 5/6
isilaza2 7/8
ubu - qololwane/n 14
i - qwela ( - e) 5
um - thula - ntabeni: 3
isi - tshixo [t f ?]
um - ty6ngololo 3/4
```

umagxa 1a/2a;

kwaMlebese ltd/loc

amanyukunyezi -/6 = amanyumnyezi

incamisa-ntliziyo 9/- / isincamisa-ntliziyo 7/- b/n

- More than one item on a line, for example

(3) ilanga lamaluluwe 5/6

ubulawu obubomvu 14/-

indlu yo - wiso - mithetho 11

u - Ziqu - zithatu - zinguThixo

ukunginkxoza 15/- 1/2 umnginkxozi; 11/- unginkxozo)

- No class information, for example

(4) intatyana

iapokrifa

u - qekele

isi - qhanyongo

i - veranda

ama - vandlakanya

- Doke's classification in file 8: The automated conversion to the corresponding Meinhof classification was done before the validation of this data took place.

For the purposes of this article a **data entry** is considered **valid** if it conforms to the generic entry associated with the data file containing it.

The process: data validation

The basic idea is to automatically reformat the valid data into one standard format suitable for the ultimate automatic generation of the noun list. The chosen format is `cp-root sc/pc` where `cp` denotes the class prefix, `root` the root, `sc` the class number (singular) and `pc` the class number (plural). If a noun does not occur in either the singular or the plural, the (non-occurring) class number is replaced by `-`. For example, `cp-root sc/-` means that the noun does not occur in the plural.

The valid reformatted data are automatically accumulated in a 'good data file'.

The invalid data are written to a separate file for electronic correction by the Xhosa linguists. The cycle is then repeated: The newly corrected data are again validated, the valid entries reformatted and likewise written to the 'good data file'. The correction cycle continues until all entries have been handled, validated and written to the 'good data file'. Correction typically consists of scrapping wrong or non-Xhosa entries, adding class numbers, correcting scanning/typing errors, etc.

It is worth noting that the validation phase is an iterative one and that only the invalid entries, the number of which significantly reduces in number after each subsequent iteration, are subjected to scrutiny by the linguists. Typically only two to three iterations were necessary.

Reformatted examples from the table of valid data above are as follows:

(5)
```
um-lala 3/4
um-fana 1/2
um-qathango 3/4
i-vazi 9/10
i-baba 5/6
isi-camango 7/8
um-gcini 1/2
um-thi 3/4
um-bhalo 3/4
```

The data validation process is concluded by sorting the 'good data file' alphabetically and then removing all duplicate entries. It should be noted that the data validation process aims at standardizing the **format** of the data, **not the contents**. The latter issue (consistency of information content in the data entries) is addressed in the generation process.

The process: noun generation

The automated generation process is only executed after the data validation phase is complete and is also an iterative procedure. It only makes use of reformatted standardized data as accumulated in the 'good data file' and is based on the valid class prefix – class number combinations as governed by Xhosa noun morphology. Class prefixes and class numbers are checked and compared for consistency. The rather detailed validity check makes provision for a variety of variant forms resulting from alternation/sound changes between morphemes.

Typically, the input data entry `cp-root sc/pc` is transformed into the following generated output with the class prefix `cp` corresponding to the singular class number `sc` and `pc` is the appropriate plural class number associated with `sc`:

```
-----------------------------------
sss n sc -root
ppp n pc
-----------------------------------
```

The string `sss` is the singular form of the noun with the root `root`, `ppp` is the associated plural form of the noun, `n` indicates the part of speech (in this case the noun), `sc` is the class number (singular) and `pc` the class number (plural).

No generation takes place if either the prefix and the class number or the class number (singular) and the class number (plural) do not agree. An entry containing any such inconsistencies is marked as 'unprocessed' and results in the following entry in the output file:

```
-----------------------------------
*** UNPROCESSED ***: cp-root sc/pc
-----------------------------------
```

Specific examples are

(6)
```
-----------------------------------
umfana n 1 -fana
abafana n 2
-----------------------------------
```

for a correct data entry and

(7)
```
-----------------------------------
*** UNPROCESSED ***: um-fana 5/6
-----------------------------------
```

for an inconsistent entry where the class prefix um- is incompatible with the class number 5.

At this stage human readable information is produced in a visually suitable way for the human linguists to see and correct any incorrect information, as shown above. This is the time consuming quality assurance part of the process in which all output generated is scrutinized and corrected by the linguist. Correct information is accumulated in a 'good generated information file', while unprocessed corrected data are subjected to a further generation and checking iteration. This process continues until all the standardized input data have been correctly used in the generation process.

The results as reusable resources

The generation process produces two different kinds of output. Firstly, the noun list for use in the spelling checker project and consisting of the singular and plural nouns are extracted from the 'good generated information file'. This noun list contains entries such as, for example

(8) umfana

abafana

Secondly, a machine-readable version of the generated information is created as an XML document. This is not presented to the human linguist, but is generated from the 'good generated information file' once the generation process is complete. The XML fragment below forms part of an extensive DTD (Document Type Definition) designed for a Xhosa lexicon on the basis of various Xhosa paper dictionary entries. In addition to the fragment of the DTD we also show a typical entry in Figure 2.

Figure 2: Fragment of DTD

```
...
<!ELEMENT body (entry+)>
<!ELEMENT entry (word-root,word-cat)+>
<!ELEMENT word-root (#PCDATA)>
<!ELEMENT word-cat (noun)>
<!ELEMENT noun (noun-prefixes,label)>
<!ELEMENT noun-prefixes ((class-pf-s,class-pf-p)|
          class-pf-s|class-pf-p)>
<!ELEMENT label (#PCDATA)>
<!ELEMENT class-pf-s (#PCDATA)>
<!ELEMENT class-pf-p (#PCDATA)>
...
```

An example of a Xhosa XML lexicon entry built from the generated information above is illustrated in Figure 3.

Figure 3: Xhosa XML lexicon entry

```
<entry>
   <word-root>fana</word-root>
   <word-cat>
      <noun>
         <noun-prefixes>
            <class-pf-s>um</class-pf-s>
            <class-pf-p>aba</class-pf-p>
         </noun-prefixes>
          <class-no>1-2</class-no>
         <label>n</label>
      </noun>
    </word-cat>
</entry>
```

In conclusion it is noted that the process described above has the following characteristics:

- It maximizes the use of data provided.
- It makes maximum use of possible automation.
- The human linguist scrutinizes and quality assures every generated entry individually.
- The human linguist does not recheck already correctly generated entries since the focus is on non-standard data, unprocessed data and incorrectly generated information.
- The format for checking on the computer monitor is designed for easy and clear viewing and electronic on-the-screen correction of the information.
- An important 'by-product' is the reusable machine-readable Xhosa lexicon as an XML document.
- The development of a prototype of a finite-state morphological analyser for Xhosa has already benefited from the reusability of this Xhosa XML lexicon.

Reusability of resources

It is not by mere coincidence that standards and protocols occupy a central position in any comprehensive model of HLT research and development (see, for example, Figure 1). The importance of standards and protocols for HLT, and in particular for the development of reusable HLT resources, becomes evident by briefly revisiting the meaning of 'reusable' in this context.

Reusability may be defined as the 'characteristic of a [software] component that (by deliberate design) allows it to be used in more than the application for which it was created, with or without modification (Reusability, 1997), while **portability** is the ability of a data format to 'transcend computer environments, scholarly communities, domains of application and passage of time' (E-MELD, 2005). If we agree that HLT resources ideally need to be reusable and portable, then it follows that the main purpose of HLT **standards** should be to 'enable HLT resources to be interoperable', that is, to facilitate reusability and portability (Standards, 2004). A rather specific kind of standard is the notion of '**protocol**', 'a convention or standard that controls or enables the connection,

communication, and data transfer between two computing endpoints. Protocols may be implemented by hardware, software, or a combination of the two. At the lowest level, a protocol defines a hardware connection' (Protocol, 2005).

By definition, standards need to be agreed upon, adopted and applied in order to succeed, a responsibility usually assumed by international standards bodies such as EAGLES (Expert Advisory Group on Language Engineering Standards), ELRA (European Language Resources Association), TEI (Text Encoding Initiative), ISO TC37 SC WG1-1 – Linguistic annotation framework, various other ISO/IEC standards (International Organization for Standardization/International Electrotechnical Commission), etc.

The obvious question now is: What standards govern lexicon compilation, morphological analysis and the extraction of root and stem lists, and how are they adhered to in the work reported on in this article?

Conceptually, we may identify three kinds of standards that should form the basis of the development of reusable HLT resources.

- Standards for structuring and representing information (e.g. coding, markup (XML) and tagging)
- Standards for accessing information (protocols)
- Standards for linguistic quality, completeness and correctness of information (integral role of the human linguist).

Reuseability relating to word lists

The importance of word lists was thus far discussed with specific reference to their utilization in the compilation of the North-West University's Spelling Checker Project. In this section the focus will be somewhat broader, illustrating how word lists and the corpora from which they are generated can be reused for a variety of basic as well as advanced applications of Human Language Technologies (HLT).

Reusability of word lists and in fact, of corpora, is virtually unlimited and new (re)uses are constantly discovered for instance the learning of new vocabulary on frequency priority, compilation of lemma lists of new or revised dictionaries, corpus stability tests, designing of lexicographic rulers and block systems, tagger lexica, custom dictionaries, morphological analysis and even physical word generation. Some of these issues will be briefly discussed in the following paragraphs. In all instances mentioned above corpora form the basis or point of departure. Firstly, large electronic corpora are created that run into millions and even hundreds of millions of words. The current approximate sizes (number of tokens) of the text corpora built in the Department of African Languages at the University of Pretoria are given in Table 2:

Table 2: Text corpora of the Department of African Languages (University of Pretoria)

Southern Sotho	4 Million
Northern Sotho	6 Million
Zulu	6 Million
Xhosa	3 Million
Ndebele	4 Million
Swazi	3 Million
Tsonga	3 Million
Venda	2 Million
Tswana	6 Million
Afrikaans	8 Million

For purposes of building the Afrikaans corpus, the said Department also uses the massive Media 24 Afrikaans archive of circa 800 million tokens and 2.6 million types.

The power of a frequency list and its potential use for various applications becomes apparent in the statistical analysis of the English language, where the values of the frequency bands in Collins COBUILD English Dictionary (COBUILD 2, 1995) show that the 17,500 most frequently used words in English represent 95% of the vocabulary of all spoken and written English.

The top 10,000 Northern Sotho words represent more than 90% of the vocabulary of the language. There are multiple ways of utilizing even basic non-marked up and non-POS tagged word lists and raw corpora. Within Microsoft Office for example, frequency lists can be used in their basic format as custom dictionaries with any major language such as English, French, German, etc. for which spelling checkers are available. We consider in this regard the results of experiments for Southern Sotho, Tsonga, Venda, Swazi and Ndebele.

Table 3: Lexical recall obtained with word lists run as custom dictionaries

Language	Number of words in test text	Words unknown to the custom dictionary (not in the word list)	Percentage of lexical recall
South Sotho	811	8	99%
Venda	302	3	99%
Tsonga	401	8	98%
Ndebele	132	20	85%
Swazi	358	57	84.1%

A new method of testing corpus stability for Northern Sotho and Tsonga using word lists was introduced by Prinsloo and De Schryver (2001). Different sections of frequency lists culled from different sections of the organic Northern Sotho corpus were compared with each other to study changes in frequency ranking of words when corpora are extended by simply adding data in a non-structured way.

Subsections of word lists, for example nouns and verbs can be used to generate additional nouns and verbs in order to extend the word lists to improve lexical recall in lexicon-based spelling checkers. This process is described in detail in Prinsloo and Eiselen (2005).

Word frequency lists also form a primary source for morphological analysis and part-of-speech (POS) tagging of corpora. For a language such as Northern Sotho it simply means that if the top 10,000 words are correctly analysed and POS-tagged, such a word list can be used to automatically analyse and POS-tag more than 90% of the words in any given corpus. For example, the ten most frequently used words in the 6 million-word Pretoria Sepedi Corpus (PSC), as indicated in brackets, are *motho* 'human being' (16,353), *batho* 'human beings' (13,720), *moka* 'the whole' (10,910), *morena* 'mister' (8419), *monna* 'man' (8271), *taba* 'issue' (8262), *kgoši* 'king' (7843), *bana* 'children' (7740), *morago* 'behind' (6145) and *mokgwa* 'manner' (6099). If these ten words are POS-tagged, the resulting tagger lexicon consisting of the ten POS-tagged words can be used to automatically tag the entire corpus. Thus 93,762 or 1.6% of all tokens in the corpus can be automatically POS-tagged when a tagger lexicon consisting of only the top 10 nouns is used.

Lexica can be stored as independent entities – e.g. a word list culled from a corpus that has been proofread, POS-tagged, morphologically analysed and updated in terms of the latest orthography. Alternatively such lexica can be freshly generated at any given time from a clean, correctly annotated corpus e.g. by means of standard corpus manipulation programs such as WordSmith Tools (Scott, 1999).

Reusability relating to morphological analysis

A great deal of the reusability relating to morphological analysis in this particular project depends on detailed documentation. Documentation for instance of what the purpose of a specific rule is, together with an example, is very useful for practical aspects such as:

- fast-tracking the development of spelling checkers for similar languages, e.g. other Nguni languages. The rules for the morphological analysis of the Xhosa spelling checker were to a large extent based on the rules for Zulu.
- debugging of the morphological analyser as rules are being implemented incrementally. By documenting all rules it is relatively simple to determine why certain words which should be correct, are flagged as incorrect and vice versa.
- adding new words to be recognised by the spelling checker. This becomes a simple exercise if the morphological analyser is well designed. By adding a new lemma, the recognition of the spelling checker expands automatically by thousands of new forms.

Table 4 and Figure 5 are examples of the documentation of the morphology of the verb in Zulu:

Table 4: Morphology of the verb

Subj Conc	Aspect	Tense Prefix	Obj Conc	Verb Root	Extensions	Terminative
i	*sa*			*buy* 'return'	*is*	*a*
ba		*ya*	*ku*	*phek* 'cook'	*el*	*a*
si				*akh* 'build'	*an*	*a*
ngi		*zo*	*ni*	*siz* 'help'		*a*

Table 5: Excerpt from documentation of the morphology of the verb in Zulu

OBJECT CONCORD [OC]

Notes: Object concord is not compulsory, but if used, it appears in the slot directly before the verb root.

Vowel is elided in each of these cases (except *-lu-* and *-ku-* which become *-lw-* and *-kw-* resp.), if object concord appears before vowel verb root.

OTHER PREFIXES

Notes: the following prefixes may be used between the subject concord and object concord, or between subject concord and verb root. These prefixes are not class bound and can be used in conjunction with any subject concord or object concord.

-ya- (may follow present tense subject concords, may also precede object concords, or both. May not combine with other prefixes below)

-sa- (may not combine with *-ya-*, but may follow compound tense prefixes)

-sazo-, *-sayo-*, *-sazoku-*, *-sayoku-* (may follow compound tense prefixes)

-zo-, *-yo-*, *-zoku-*, *-yoku-* (may follow compound tense prefixes)

Table 5 (continued): Excerpt from documentation of the morphology of the verb in Zulu

VERBAL EXTENSIONS
Notes: these are the most commonly used extensions.

-is-

-el-

-an-

-ek-

-isis-

-elel-

-ezel-

-w- / -iw- (this is the passive extension and causes some morphophonological changes, cf. below)

MORPHOPHONOLOGICAL ALTERNATIONS
Vowel elision:
si+akh+a > sakha (if *a-* or *i-* precedes a vowel verb root, the first vowel is elided)

Passive extension:
If verb root ends in ..... followed by *-w-*, then the change that takes place is....

b h > *j*

p h > *s h*

m p > *n t s h*

m b > *n j*

m > *n y*

b > *t s h*

Reusability relating to lemma lists

In the previous section a brief description of a practical, efficient and effective semi-automated process for developing lemma lists for Xhosa was given. The approach was primarily devised to not only meet the practical requirements of the project, but also to produce basic reusable lemma lists that satisfy appropriate standards.

Since no such lists were electronically available, they had to be built from a number of available paper dictionaries. As mentioned previously, the data contained many errors of different kinds. In short, the original data were **non-standard** in a number of ways. Limited time and human resources, as well as stringent project schedules and deadlines for delivery posed additional challenges. Nevertheless, the output required was a list of only and as many as possible valid Xhosa nouns and verb stems.

The process followed was specifically aimed at complying, albeit informally, with the standards mentioned earlier:

- a basic machine-readable XML lexicon was built concurrently with the compilation of the lemma list. This lexicon was validated against its DTD and is suitable for reuse in forthcoming applications.
- all lemmas, together with the generated singular and plural noun forms, were subjected to a linguistic verification and validation cycle involving human linguists.

A thorough investigation into the formulation and application of **formal** standards for HLT resources for the South African languages forms part of future work.

The way forward – National Resource Centre

Due to the complexity and information-richness of human language, software tools designed to process human language need to be based on vast amounts of diverse linguistic data such as speech, text, lexicons and grammars, in order to be robust and effective. However, the cost of developing, maintaining and distributing such databases can be prohibitive, even for large companies.

Cole (1997:382) remarks that most languages still lack adequate linguistic resources, and proposes the following two reasons for this state of affairs:

> During the 1970's and first half of the 1980's, linguists preferred to test their hypotheses with relatively small amounts of (allegedly) critical data, instead of resorting to extensive corpora based on language occurring in communicative contexts;
>
> The creation of language resources is a very costly undertaking and requires co-operation from companies, research institutions, government as well as sponsors.

In the international arena, this challenge has been addressed to a large extent by the establishment of consortia, which provide a mechanism for large-scale development, and widespread sharing of language resources. One such example of a consortium is the **Linguistic Data Consortium** (LDC) (cf. Introduction to the Linguistic Data Consortium, 1999) founded in 1992 and based at the University of Pennsylvania, in the USA.

Several parallel initiatives are under way in Europe and the Far East. For example in 1995 the **European Language Resource Association** (ELRA) (cf. European Language Resources, 2001) was established, with the specific purpose of providing a basis for central co-ordination of corpora creation, management and distribution in Europe.

In South Africa, the co-ordination of language resources is still in its infancy stages. There are pockets of expertise throughout the country developing language resources in relative isolation and in an uncoordinated way. However, a strategy for Human Language Technologies (HLT) which was approved by the Ministry of Arts, Culture, Science and Technology in 2001, has as one of its main aims, the setting of particular standards for the development of language and speech resources, as well as the creation of an appropriate mechanism to co-ordinate the activities of the main role players in the field. (The development of HLT in South Africa, 2000).

It is unacceptable that electronic language sources are destroyed on a daily basis and on a massive scale in South Africa. Interim strategies should be considered *en route* to a well-structured and coordinated national strategy and facility. A first step could be simply to encourage all generators of language in electronic format not to destroy data but to archive it in whatever way is possible and affordable to the individual, company, institution or government. Archive initiatives like *Media24* (Media24, 2005) should be encouraged and supported. Thus 'saving' the data in all senses of the word is the first prerequisite.

Secondly, it is important that such data should be portable, that is filed in a format that is readable by as many other standard programs as possible. Publishers, for example are often prepared to make data available but it can only be read by dedicated in-house or expensive antiquated and even commercially unavailable programs.

Thirdly, we should strive towards compatibility on different levels of which the first is the protocol level facilitating mutual readability/compatibility between major programs. Compare for example the situation regarding dictionary compilation programs of the National Lexicography Units (NLUs) of the Pan South African Language Board. Three programs are used by the different units. One uses *Onoma*, three started on *TshwaneLex* and the rest still use *Microsoft Word*. It is imperative that these three programs should be able to read each other's data.

Moreover, they should be compatible in structure and representation so that for instance data created in *Onoma* could be imported in *TshwaneLex* and further processed in the latter, and vice versa. A minimum requirement should be that developers and buyers of software should ensure that exportation to standard formats like XML, ASCII, PDF, etc. is possible.

Finally, proactive steps should be taken in terms of the hardware medium in which data is stored. Many valuable (re)sources are 'lost' that were stored on huge floppy disks in the era preceding the 4¼" disks and even 3½" disks. Even CDs are becoming outdated in favour of DVDs and different formats complicate data storage and retrieval even in the latter format. Maintaining machines and systems to read archived data in the years to come is a major challenge.

Conclusion

It is apparent that the development of resources for the various indigenous languages in South Africa is, at present, at different levels. This is the result of isolated projects where resources have now been developed for languages such as Zulu, Xhosa, Northern Sotho and Tswana, as in this project. The research already completed and the resources already compiled should therefore form a blueprint for similar developments within the other languages not yet developed in these areas without having to reinvent the wheel.

This paper has highlighted the critical importance of resources that may be developed during research projects and the necessity for adherence to standards and protocols in the developmental phase to ensure the reusability of such data for future research in the indigenous languages of South Africa. It is hoped that the envisaged HLT centre will not only promote further research, but also encourage closer collaboration between institutions and researchers and be instrumental in the development of a national resource centre from which all researchers in African languages may benefit in their quests to develop these languages.

Notes

1 Note that the terms *language resources* and *linguistic resources* are used interchangeably.

2 The use of the term *lemma* in this paper is preferred to *head word* because these canonical forms used in dictionaries of the Bantu languages are often not words, but may be suffixes, prefixes, stems, roots or even multiple words. For example, lemmas in Sotho dictionaries are in most cases orthographic words while lemmas in Nguni dictionaries are usually stems.

3 Morphological analysers for a number of Bantu languages are being developed by means of finite-state technology within an NRF Focus Area project entitled 'Computational Morphological Analysis'.

4 The use of `Courier New` font for displaying computer input/output and for distinguishing it from normal text has become common practice.

References

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

Bosch, S.E. & Eiselen, R. 2005. The effectiveness of morphological rules for an isiZulu spelling checker. *South African Journal of African Languages* 25(1):25–36.

(COBUILD 2) Sinclair, J. (ed). 1995. *Collins COBUILD English Dictionary.* London: HarperCollins.

Cole, R. (ed). 1997. *Survey of the state of the art in human language technology.* Cambridge: Cambridge University Press.

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

E-MELD. 2005. *E-MELD School of Best Practice.* Portability – Glossary. [O]. Available: http://www.emeld.org/school/glossary.html
Accessed on 2005/07/29

English-Xhosa/Xhosa-English Dictionary. [Sa]. Elsies River, Cape: Via Afrika.

European Language Resources Association. 2001. [O]. Available:
http://www.icp.inpg.fr/ELRA/whatis.html
Accessed on 2002/02/04

Godfrey, J.J. & Zampolli, A. 1997. Language Resources, in *Survey of the state of the art in human language technology*, edited by R. Cole. Cambridge: Cambridge University Press: 381–408.

Heine, B. & Nurse, D. (eds). 2000. *African languages: An introduction*. Cambridge: Cambridge University Press.

Introduction to the Linguistic Data Consortium. 1999. [O]. Available:
http://www.ldc.upenn.edu/ldc/about/ldc_intro.html
Accessed on 2002/02/04

Kropf, A. & Godfrey, R. (eds). 1915. *A Kafir-English dictionary*. South Africa: Lovedale Mission Press.

Language Engineering, Harnessing the Power of Language. 2005. [O]. Available:
http://www.hltcentral.org/usr_docs/project-source/en/broch/harness.html
Accessed on 2005/06/28

Longman Dictionary of Contemporary English Online. 2006. [O]. Available:
http://www.ldoceonline.com
Accessed on 2006/03/20

McLaren, J. 1963. *A new concise Xhosa-English dictionary*. Cape Town: Maskew Miller Longman.

Media 24. 2005. [O]. Available:
http://www.media24.com
Accessed on 2005/06/30

Oflazer, K. & Güzey, C. 1994. Spelling correction in agglutinative languages. *4th Conference on Applied Natural Language Processing*. Stuttgart, Germany:194–195.

Oxford advanced learner's dictionary on CD-ROM. 1997. Oxford: Oxford University Press.

Pahl, H.W. (ed. in chief). 1989. *The greater dictionary of Xhosa* (Vol. 1,2,3). Alice: University of Fort Hare.

Prinsloo, D.J. & de Schryver, G-M. 2001. Monitoring the stability of a growing organic corpus, with special reference to Sepedi and Xitsonga. *Dictionaries: Journal of The Dictionary Society of North America* 22: 85–129.

Prinsloo, D.J. & Eiselen, R. 2005. Improving a lexicon-based spelling checker for Sesotho sa Leboa. *South African Journal of African Languages* 25(1):11–24.

Protocol (computing) - Wikipedia, the free encyclopedia. 2005. [O] Available:
http//en.wikipedia.org/wiki/Protocol_(computing)
Accessed on 2005/07/29

Reusability - Component-Speak: A Glossary. 1997. [O]. Available:
http://www.cio.com/archive/030197/glossary.html
Accessed on 2005/07/29

Scott, M. 1999. *WordSmith tools* version 3. Oxford: Oxford University Press.

Standards (software) - Wikipedia, the free encyclopedia. 2004. [O]. Available:
http://en.wikipedia.org/wiki/Standards_(software)
Accessed on 2005/07/29

The development of HLT in South Africa. 2000. [O]. Available:
http://www.dacst.gov.za/arts_culture/index.htm
Accessed on 2002/02/05

Williamson, K. & Blench, R. 2000. Niger-Congo, in *African languages: An introduction*, edited by B. Heine & D. Nurse. Cambridge: Cambridge University Press:11–42.