Nordic Journal of African Studies 16(2): 131–145 (2007)

Towards Machine-Readable Lexicons for South African Bantu languages[*][†]

SONJA E. BOSCH,
LAURETTE PRETORIUS
&
JACKIE JONES
University of South Africa, South Africa

ABSTRACT

Lexical information for South African Bantu languages is not readily available in the form of machine-readable lexicons. At present the availability of lexical information is restricted to a variety of paper dictionaries. These dictionaries display considerable diversity in the organisation and representation of data. In order to proceed towards the development of reusable and suitably standardised machine-readable lexicons for these languages, a data model for lexical entries becomes a prerequisite. In this study the general purpose model as developed by Bell and Bird (2000) is used as a point of departure.

Firstly, the extent to which the Bell and Bird (2000) data model may be applied to and modified for the above-mentioned languages is investigated. Initial investigations indicate that modification of this data model is necessary to make provision for the specific requirements of lexical entries in these languages. Secondly, a data model in the form of an XML DTD for the languages in question, based on our findings regarding Bell and Bird (2000) and Weber (2002) is presented. Included in this model are additional particular requirements for complete and appropriate representation of linguistic information as identified in the study of available paper dictionaries.

***Keywords**: data model, lexical entries, Bantu languages, machine-readable lexicons*

INTRODUCTION

For the purposes of this paper the term machine-readable lexicon is understood as "a lexicographic knowledge base from which lexica of all … different kinds can be derived automatically." (Van Eynde and Gibbon, 2002: 2). Natural language processing (NLP) can be considerably improved by the availability of a complete and accurate lexicon. Most NLP applications, including lexicography

[*] 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.

[†] An earlier version of this paper was published in the Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy 1722–1727.

and lexical semantics (Itai, Wintner and Yona, 2006: 19) depend in some way on a machine-readable lexicon as a basic resource. For example, machine-readable lexicons are an essential component in the development of morphological analysers.

The goal in the development of a machine-readable lexicon is to be as inclusive as possible, thus incorporating all relevant information in the most efficient and economical manner, to be reusable and to conform to suitable and appropriate international standards. In order to achieve this, the design of an appropriate data model may represent a first step. It should be noted that a conceptual data model, as proposed in this article, would certainly be of use in the building of any kind of (electronic) dictionary. For a taxonomy of such dictionaries the interested reader is referred to De Schryver (2003).

For the Bantu languages, which are lesser studied languages, lexical information is not readily available in the form of machine-readable lexicons. In most cases, such lexical resources need to be newly developed. At present the availability of lexical information is restricted to a variety of paper dictionaries. These dictionaries display considerable diversity in the organisation and representation of data. This diversity emanates from factors such as designers' decisions, user needs, intended mode of delivery and economic considerations. Bell and Bird (2000) maintain that as "we look to a future in which lexical data is increasingly deployed online, this diversity presents problems for exchanging data and for developing general purpose tools". They furthermore identify the core problem, namely that "there is no general purpose data model for lexical entries" and continue to propose a data model. Their model is based on sample entries collected from fifty-five lexicons for a large variety of languages, including a number of lesser studied languages of the world and is presented in the form of an XML (Extensible Markup Language) DTD (Document Type Definition).

This model has since been discussed from various perspectives - one being its relevance for exotic languages such as Huallaga Quechua (Weber, 2002); another is its suitability as a conceptual basis for a so-called Unified Model for Language Data, designed for use in the Semantic Web (Farrar, 2002; Farrar and Lewis, 2005). Indeed, the Bell and Bird model forms the basis for the data structures contained in the General Ontology for Linguistic Description (GOLD) (Simons et al., 2004).

The purpose of this paper therefore is to address the need for an appropriately general data model for lexical information in the context of the Bantu languages. Firstly, we use the data model proposed by Bell and Bird (2000) as a point of departure and investigate how this model may be applied and modified for the Bantu languages of South Africa. Secondly, we discuss the necessity for modification of the Bell and Bird (2000) model based on the idiosyncrasies of the Bantu languages considering also the recommendations of Weber (2002). We present an excerpt of our XML DTD for the languages in question and demonstrate its use by means of examples.

1. PROBLEM STATEMENT

In order to proceed towards the development of reusable and suitably standardised machine-readable lexicons for the South African Bantu languages, a data model for lexical entries is a prerequisite. However, in the design and development of such a data model the characteristics of the languages concerned are relevant and consideration should be given to the current representation of paper dictionary entries.

A number of paper dictionaries for the Sotho languages (Southern Sotho, Northern Sotho and Tswana) and Nguni languages (Zulu, Xhosa, Swati and Ndebele) was consulted, and examples extracted from a variety of these dictionaries are given below. These examples illustrate the diversity that had to be considered in order to ensure inclusiveness in describing all relevant information for each language.

Examples (1) and (2) represent languages belonging to the Nguni group of languages which follow a conjunctive orthography. It should be noted that the examples include similar entries in each case, i.e. the equivalent of the noun for ‘human being’ and the verb for ‘love’, and that (2c) includes an example of the socio-linguistic feature *isiHlonipho sabafazi* (married women’s language of respect).

(1) Xhosa (A New Concise Xhosa-English Dictionary, 1984)

(1a)
ntu, *um-*, n. 1, a human being, a man or woman, a person; pl. abantu, men, persons, people, esp. the native people; ...

(1b)
thanda, v. t. like, love, esteem; wish, will, desire; *i-thanda*, n. 3, a lover of, and *isi-thanda*, n. 4, a great lover of, (cattle, money, etc.)... *thandana*, love each other; ...
thandeka, be lovable, amiable; ... *thandela*, love for; wish, desire for, ... *thandisa*, cause to love, wish or desire; ...

(2) Zulu (English-Zulu Zulu-English Dictionary, 2005)

(2a)
-ntu (*umuntu*, 3.2.9, *abantu*) n. [Ur-B.*muntu*.>dim. *umntwana*; *unomuntu*; *ubuntu*; *isintu*; *u(lu)ntu*; *umuntukazana*; *bantu*]
1. Human being, person; man (not of necessity male). ...
6. Loc. forms: … *kumuntu* … *emuntwini* …

(2b)
thanda (3.9) v. [>perf. –*thandile*; pass. *thandwa*; neut. *thandeka*; ap. *thandela*; rec. *thandana*; caus. *thandisa*; int. *thandisisa*; dim. *thandathanda*; *umathandana*; *izithandani*; *umthandi*; *intando*; *u(lu)thando*; *isithandwa*.]
1. Like, love, be fond of; value, esteem, admire; prefer ....

(2c)
^{o}**-haqa** (*i(li)haqa*, 3.2.9.9, *amahaqa*) n. [<*haqa*.] hlonipha term for *i(li)bodwe*, cooking pot.

In examples (1) and (2) it is illustrated that in the case of the Nguni languages all entries are stem entries, possibly due to the conjunctive writing style.

Examples (3) and (4) represent languages belonging to the Sotho group of languages, which follow a disjunctive orthography. The examples again include similar entries in each case, i.e. the equivalent of the noun for 'human being' and the verb for 'love'.

(3) Northern Sotho (The New English-Northern Sotho Dictionary, 1976)

(3a)
motho, n., human being, a person; ...

(3b)
batho, n. pl., people; ...

This plural form is a separate entry which is found listed under the letter 'b' and is not represented under the singular *motho*.

(3c)
rata, v.t., love, like, wish, will, want to.

(3d)
i'thata, v. reflex., *rata*, love oneself, be selfish, egoistic.

(4) Southern Sotho (Sesuto-English Dictionary, 1976)

(4a)
motho, n., human being, male or female person ; ...

(4b)
rata, v.t., to love, to like, to will; *ratèha*, v.n., to be lovely, to be lovable ; *ithata*, v.t., to love oneself ; to be selfish ; *ratana*, to love one another ; *ratisa*, v.t., to cause to love ; *ratisuoa*, v.t., to be obliged to love, to crave for (of a pregnant woman desiring certain things to eat) ; *ratisana*, to cause or teach one another to

love ; *ithatisa* v.r., to cause oneself to love ; *ratéla*, v.t., to love for ; *ithatèla*, v.r., to love for oneself, to like, to prefer; *ratisisa*, v.t., to like very much.

In examples (3a) and (4a) it is noticeable that the noun *motho* ‘human being’ is entered as an orthographic word and not as a stem as in the case of the Nguni languages.

However, example (5a) differs in representation from other Northern Sotho dictionaries in that the tradition of stem entry is being followed:

(5) Northern Sotho (Comprehensive Northern Sotho Dictionary, 1975)

(5a)
-tho, mo-/ba- ... human being, person, man (in general); ...

It is evident from these paper dictionary entries that the aspect of disjunctively as opposed to conjunctively written languages, as well as the agglutinative characteristics of the Bantu languages, becomes significant in the representation of entries.

The conjunctive system of word division in the Nguni languages, including Xhosa and Zulu, has given rise to lexicographic representation according to stems while in the case of the other South African Bantu languages such as Northern Sotho, disjunctivism has given rise to lexicographic representation of orthographic words in most dictionaries (Van Wyk, 1995). This discrepancy in lexicographic representations is clearly illustrated in example (1a) in which the single Xhosa stem entry *-ntu* corresponds to the two Northern Sotho orthographic word entries in examples (3*a) motho* and (3b) *batho*.

Secondly, the concept of recursion (which results in nested entries) is found in all paper dictionaries consulted. This is due to the agglutinating morphological structure of Bantu languages according to which series of prefixes and suffixes are built around base forms such as noun stems or verb roots. It should also be noted that due to language idiosyncrasies, rules of derivation are not consistent since for example all verbal extensions are not able to combine with all verb roots. This necessitates the explicit inclusion of known occurrences as subentries under the base form.

For example, in Zulu the base form *-funda* ‘read/learn’ has the following derived forms which we accommodate in subentries under the base form:

(6)
-fundela ‘read/learn for’
-fundeka ‘readable’
-fundisa ‘teach’.

Typically the causative verbal extension *-is-*, as in *-fundisa*, may be further extended with a reciprocal extension *-an-*, which would be considered a subentry *-fundisana* ‘teach each other’ under *-fundisa* ‘teach’.

It would therefore seem appropriate to represent the relationship of a derived form as a subentry under the base form. An example of this regarding the language Huallaga Quechua may be found in Weber (2002). Note that we do not include our subentries under the Sense of the base form. We include them under the MSI of the base form.

2. OUR APPROACH

Considering the above-mentioned features of the Bantu languages we concur with the Bell and Bird (2000) basic structure of a lexical entry, namely a 4-tuple <*Form*, *Morpho-syntactic information*, *Sense(s)*, *Auxiliary information*> (Farrar, 2003), where *Form* includes pronunciation information, *Sense(s)* contain semantic information such as sense definitions, while information regarding aspects such as, for example, dialects and etymology is included under *Auxiliary information*. Moreover, the dictionary entries that we propose adhere to the rules for creating tree-like dictionary entries (Ide, Kilgariff and Romary, 2000).

However, for building some of the first machine-readable lexicons for the South African languages, the Bell and Bird (2000) model is too abstract in the sense of Farrar (2003) and 'ambiguous' in the sense that it allows the duplication of information in different places in the model. We prefer to be explicit about what we consider as auxiliary information. Furthermore, initial investigations also indicate that refining and extending this data model are necessary to make provision for the specific requirements of lexical entries in these languages. The inclusion of derived forms as subentries under the original base form, as recommended by Weber (2002) in his reaction to the Bell and Bird (2000) data model, is especially significant for the Bantu languages, as shown in the previous section.

We present an excerpt of our data model in the form of a fragment of our XML DTD for the languages in question, based on our findings regarding Bell and Bird (2000) and Weber (2002). For a complete and appropriate representation of linguistic information as identified in the study of available paper dictionaries, additional information is explicitly included in this model. Examples of this include the representation of class information, singular and plural, feminine, augmentative, diminutive and locative formation in the case of nouns, and verbal extensions 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) as illustrated in example (2c) and Xhosa *isiKhwetha* (male initiates' language) both features of which would also necessitate explicit representation in the lexicon.

For illustrative purposes, we also show the XML entry for the base form *-ntu* (noun, class1-2) and a fragment of the XML entry corresponding to *rata* (verb).

3. A MODEL FOR THE SOUTH AFRICAN BANTU LANGUAGES

In the development of a general model for the South African languages it is important to take cognisance of other models developed for other languages, and to bear in mind that the Bantu languages that were investigated do differ from those studied by Bell and Bird (2000). According to Wittenburg, Peters and Drude (2000) “It is important to assess these differences and aim at the integration of lexical resources in order to improve lexicon creation, exchange and reuse“. One of the most significant areas where this model seemed inadequate to accommodate the South African Bantu languages was the exclusion of the appropriate nesting of derived forms so prevalent in these languages.

The modifications made to the Bell and Bird (2000) model concern specific areas due to the differences in the structure and writing styles of the Bantu languages. It was also our aim at the outset to make our DTD precise and to avoid repetition of data which may result in ambiguity and redundancy in computational applications.

Before discussing the modifications, we would like to concur with Bell and Bird (2000) that the complexity of the `Head` element should be minimised by including as much information as possible in the `Body`. For our purposes and considering the entries of the paper dictionaries consulted, we came to the conclusion that the information in the `Head` should be limited to base form entries. Information such as phonetic transcriptions and tone which is occasionally included in the `Head` should be contained in the `Body` together with other linguistic information.

In an attempt to include all the relevant linguistic information as captured from a variety of paper dictionaries of the South African Bantu languages, the following modifications are implemented in our data model:

- Considering that the majority of the South African Bantu language paper dictionaries follow the stem entry approach it was decided that affix information should not appear in the `Head` but should be included in the `Body` element. Since a noun stem may often combine with a variety of different prefixes, it would seem appropriate to include affix information in the `Body` element as opposed to the `Head`, which is then reserved for the stem or base form. It is for this reason that we model noun stems, which may combine with prefixes from a number of different classes, as separate lexicon entries, one for each pair of classes (singular and plural) or each class should there not be a singular and plural pair. For example, we would have an entry for *-ntu*, class1-2 (of which we give the XML entry below), an entry for *-ntu* class 14, and so on.
- Due to the prominence of the verbal extensions in the Bantu languages, verbal extension information should appear at the level of `MSI` and not merely in a comment as is given for the example from the Bantu language Shona in the Bell and Bird (2002) model. It should also be readily extractable from the XML document and not appear as text together with other types of information. The

level at which the comment is introduced in Bell and Bird (2000) and which would equate to the level of verbal extensions would not be overtly visible. This is therefore not appropriate for verb information in the case of the Bantu languages.

• Provision should be made for the possibility of one or more nouns. The noun forms that are indicated in some paper dictionaries are the feminine, augmentative and diminutive, e.g.

(7)
indlovu n. 'elephant' > *indlovu**kazi*** n. (fem.) 'elephant cow, queen'
u(lu)tho n. 'something, anything' > *u(lu)tho**kazi*** n. (aug.) 'huge terrifying thing or affair'
inkosi n. 'king, paramount chief, chief' > *inkos**ana*** n. (dim.) 'small or petty chief'

Although the noun suffix *–kazi* is used to derive feminine as well as augmentative forms from nouns in Zulu, corpus investigations by Gauton et al. (2004: 376) show that "the primary significance of the suffix *–kazi* is the expression of the feminine form, with the augmentative significance as secondary." Differentiation between the two significances would need to be made.

• Locative form information is idiosyncratic and therefore needs to be specified as #PCDATA. There are no rules that determine whether a Zulu noun for instance suffixes the locative morpheme *-ini* simultaneously with the locative prefix or not, e.g.

(8)
ikhaya n. 'home' > ***e**khaya* 'at home' (loc.) 'at/in/towards home'
intaba n. 'mountain' > ***e**ntab**eni*** (loc.) 'at/in/to the mountain'.

In some instances locatives may be formed in two completely different ways, e.g.

(9)
umuntu n. 'human being, person' > ***ku**muntu* (loc.) 'to/with/at the person'
umuntu n. 'human being, person' > ***e**muntw**ini*** (loc.) 'among the human race'.

• The feature structures of the POS element for the Bantu languages should be explicitly included where applicable, specifically for nouns and verbs. Whereas Bell and Bird (2000) defer defining these feature structures, we aim to be overt in describing the crucial issues pertaining to these languages.

• The importance of entering the reflexive form of the verb under the base form is emphasised by the fact that in the Sotho languages the base form may change when the reflexive morpheme *i-* is prefixed, e.g. *rata* v. 'love' becomes *ithata* v.

refl. ‘love oneself’. In example (3d) it is illustrated that *ithata* appears as a separate entry and not as sub-entry of the base form *rata*. We see this as a further justification for sub-entries of derived verb forms. Furthermore, the occurrence of an entry *ithata* implies affix information (i.e. reflexive *i*-) in the head, an aspect we want to avoid (cf. Bell and Bird, 2000).

• Transitivity may be influenced by verbal extensions, in other words transitivity may change as extensions are added. Therefore transitivity information should be included for the base form *as well as* for each subsequent form. The following Zulu examples illustrate this phenomenon by means of the suffixation of the applied verbal extension:

(10)

-*vuka*	(intr.)	‘wake up, awake (from sleep)’
-*vukela*	(tr.)	‘awake for, rise for’
-*pheka*	(single tr.)	‘cook’ (as in -*pheka ukudla* ‘cook food’)
-*phekela*	(double tr.)	‘cook for’ (as in -*phekela abantwana ukudla* ‘cook food for the children’)

• In the development of higher human language technology (HLT) and NLP applications we require a precise structure. In the Bell and Bird (2000) model (DTD) the `Variant` element, which is present in `Head`, contains register and dialect information. The `Aux` element, which is found in `Body`, also contains register and dialect. This information therefore may be accommodated in two places. For our purposes we consider this duplication to be undesirable, particularly from a computational point of view, and therefore include this information in the `Body` element only where applicable.

• One of our main purposes is to mark up lexicon information for logical structure in order to provide essential information for the computational language processing task (Ide, 2000). Our goal is to provide a useful and efficient computational resource.

• For the purposes of this paper we use a pure hierarchical element-based DTD, without attribute lists and attributes, to exhibit the *structure* of our data model. This causes the example entries to be rather verbose. We also do not address the important issue of mapping and cross-referencing in this paper.

• Regarding the DTD, our general approach is that `#PCDATA` is mainly used for “structural units of language” (Wittenburg, Peters and Drude, 2000) as deemed useful in computational applications, not for field linguistics type descriptive purposes.

• Reiterating, there is no recursion at `Entry` level. The `MSI` element is obligatory because the `Head` element contains the stem, which requires morphological completion. We allow for one and only one `POS` element per lexical entry. The elements `Noun-features` and `Verb-features` are obligatory and explicitly included, as mentioned before.

- As extensively discussed in previous sections of this paper, the verbal extensions are accommodated as (recursive) subentries `Ext` under the entry `Verb-exts`. Moreover, the `Reflexive-baseform` may exhibit the full complexity of the recursive subentry structure of `Ext` under the `Verbal-exts*` subentry via the `Verb-features` entry, in accordance with the exposition in example (4b). Also, the sense information of each derivational form is presented in the subentry that represents the specific derivational form.

A fragment of our DTD is as follows:

```
<!ELEMENT Entry (Head,Body)>
<!ELEMENT Head (Stem)>
<!ELEMENT Stem (#PCDATA)>
<!ELEMENT Body (Phon-transc*,Tone*,MSI+,Sense+,
Dialects*,Etymology*)>
...
<!ELEMENT MSI (POS)>
<!ELEMENT POS (Noun | Verb | Adverb | Adjective | Relative | Interjective |
Conjunctive | Ideophone | Enumerative | Pronoun | Aux-verb)>

<!ELEMENT Noun-features (Class-pf-s?,Class-pf-p?,Class-
no,Label,Aug*,Fem*,Dim*,Loc*)>
<!ELEMENT Class-pf-s (#PCDATA)>
<!ELEMENT Class-pf-p (#PCDATA)>
<!ELEMENT Class-no (#PCDATA)>
<!ELEMENT Label (#PCDATA)>
<!ELEMENT Aug (Form,Sense)>
<!ELEMENT Fem (Form,Sense)>
<!ELEMENT Dim (Form,Sense)>
<!ELEMENT Loc (Form,Sense)>
<!ELEMENT Form (#PCDATA)>

<!ELEMENT Verb (Root,Verb-features,Refl?,
Perfect-tense?,Redupl?)>
<!ELEMENT Root (#PCDATA)>
<!ELEMENT Verb-features (Transitivity,Label,
Verbal-exts*)>
<!ELEMENT Transitivity (#PCDATA)>
<!ELEMENT Verbal-exts (Ext+)>
<!ELEMENT Ext ((Appl | Caus | Compl | Intens | Neut | Pass |
Recip),Transitivity,Sense+),Ext*)>
<!ELEMENT Appl (#PCDATA)>
...
<!ELEMENT Refl (Reflexive-baseform,
Verb-features?)>
<!ELEMENT Reflexive-baseform(#PCDATA)>
<!ELEMENT Perfect-tense (Long,Short)>
<!ELEMENT Redupl (#PCDATA)>
<!ELEMENT Long (#PCDATA)>
<!ELEMENT Short (#PCDATA)>
...
<!ELEMENT Sense (Gloss+)>
<!ELEMENT Gloss (Eng,Example*)>
<!ELEMENT Eng (#PCDATA)>
<!ELEMENT Example (Usage,Transl)>
<!ELEMENT Usage (#PCDATA)>
<!ELEMENT Transl (#PCDATA)>

<!ELEMENT Dialects (Socio*,Region*)>
<!ELEMENT Socio (Hlon*,Khweth*,Slang*)>
<!ELEMENT Region (#PCDATA)>
```

```
<!ELEMENT Etymology (Proto-form*,Loan-word*)>
<!ELEMENT Proto-form (Ur-B?,CB?)>
<!ELEMENT Hlon (#PCDATA)>
<!ELEMENT Khweth (#PCDATA)>
<!ELEMENT Loan-word (Word,Origin)>
<!ELEMENT Word (#PCDATA)>
<!ELEMENT Origin (#PCDATA)>
<!ELEMENT Slang (#PCDATA)>
<!ELEMENT Ur-B (#PCDATA)>
<!ELEMENT CB (#PCDATA)>
```

In the following two examples the mark up of the “structural units of language” should be noted. The XML entry for the noun -*ntu*, class 1-2, in example (2a) is as follows:

```
<Entry>
  <Head>
    <Stem>ntu</Stem>
  </Head>
  <Body>
    <Tone>3.2.9</Tone>
    <MSI>
      <POS>
        <Noun>
          <Noun-features>
            <Class-pf-s>umu</Class-pf-s>
            <Class-pf-p>aba</Class-pf-p>
            <Class-no>1-2</Class-no>
            <Label>n</Label>
           <Dim>
              <Form>umntwana</Form>
                   <Sense>baby, small child</Sense>
            </Dim>
            <Loc>
              <Form>kumuntu</Form>
              <Sense>to the person</Sense>
            </Loc>
            <Loc>
              <Form>emuntwini</Form>
              <Sense>among the human race</Sense>
            </Loc>
               </Noun-features>
        </Noun>
      </POS>
    </MSI>
    <Sense>
      <Gloss>
        <Eng>human being, person, man</Eng>
      </Gloss>
    </Sense>
    <Etymology>
      <ProtoForm>
        <UrB>muntu</UrB>
      </ProtoForm>
    </Etymology>
  </Body>
</Entry>
```

A fragment of the XML entry for the verb *rata* in example (4b) is as follows:

```
<Verb>
  <Root>rat</Root>
  <Verb-features>
    <Transitivity>t</Transitivity>
    <Label>v</Label>
    <Verbal-exts>
      ...
      <Ext> <!-- Orth. form: ratisa -->
        <Caus>is</Caus>
             <Transitivity>t</Transitivity>
             <Sense>
          <Gloss>
                  <Eng>to cause to love</Eng>
               </Gloss>
             </Sense>
             <Ext> <!-- Orth. form: ratisana -->
               <Recip>an</Recip>
               <Transitivity>t</Transitivity>
               <Sense>
            <Gloss>
                   <Eng>
                  to cause or teach one another to love
                   </Eng>
            </Gloss>
               </Sense>
             </Ext>
           </Ext>
       ...
```

A few closing remarks on recursion seem in order. Much has been written on whether or not recursion is necessary in machine-readable lexicons. We appreciate the view that in terms of archiving (field) linguistics information, recursive descriptions by a multitude of contributors (field linguists, mother-tongue speakers etc.) can become arbitrarily complex and unwieldy. However, we are dealing with modelling the structure of the lexical entities of certain languages for computational purposes, which we consider a somewhat different activity. We argue that well-defined recursion is the correct and intuitive way to capture certain linguistic information such as verbal extensions for the languages of interest.

We therefore concur with Bell and Bird (2000) in that we do not require recursion at their `Lexeme` level (our `Entry` level) for the purposes of the modelling of verbal extensions as they occur in the South African Bantu languages. We do, however, make appropriate use of recursion inside our `Entry` level under `MSI` for the accurate and intuitive modelling of the mentioned constructs.

4. CONCLUSION

In this paper we show that previously applied data models (Bell and Bird, 2000) require modification for machine-readable lexicons for the South African Bantu languages. In particular we question the intuition expressed by Bell and Bird (2000) that "a complete model of dictionaries and lexicons should not need to include recursion of entries". We instead concur in principle with the notion expressed by Weber (2002: 8) that "lexical databases should accommodate (1) derived forms having multiple senses and (2) derived forms ... of the bases from which they are derived". Indeed, we include our recursion in the MSI where from a computational point of view it is available as basic linguistic building blocks for use in, for example, morphological analysis and syntactic analysis. We include the various senses together with their associated derivational forms in the element `Ext` where they are readily available for use in applications that may require them. We therefore propose an alternative model for machine-readable lexicons, which differs in significant ways to ensure maximum inclusiveness of all linguistic information. The model provides flexibility and handles the various representations specifically applicable to Bantu languages, thereby making it applicable to diverse uses of machine-readable lexicons.

The collection of data as well as the model we have developed and proposed, is intended to contribute to 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. We conclude by emphasising that our purpose is not only descriptive in nature, but is aimed at developing machine-readable lexicons as language resources for use in large-scale HLT/NLP applications and the technological development of the South African Bantu languages.

REFERENCES

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

De Schryver, G-M. 2003.
Lexicographers' dreams in the electronic-dictionary age. **International Journal of Lexicography** 16(2): 143–199.

Doke, C.M., Malcolm, D.M., Sikakana, J.M.A. and Vilakazi, B.W. 2005.
English-Zulu Zulu-English Dictionary. Johannesburg: Witwatersrand University Press.

Farrar, S. 2002.
New ways of thinking about lexical resources: a proposal for the Semantic Web. In: *EMELD Workshop on Digitizing Lexical*

Information. [O] Available. http://www.u.arizona.edu/~farrar/publications.html. Accessed on 3 October 2005.

Farrar, S and Lewis, W.D. 2005.
The GOLD community of practice: An infrastructure for linguistic data on the web. [O] Available. http://www.u.arizona.edu/~farrar/papers/FarLew-rev.pdf. Accessed on 4 September 2006.

Gauton, R., De Schryver, G-M. and Mohlala, L. 2004.
A Corpus-based Investigation of the Zulu Nominal Suffix –kazi: A Preliminary Study. In: *Proceedings of the 4th World Congress of African Linguistics New Brunswick 2003,* A. Akinlabi and O. Adesola (eds), pp. 373–380. Köln: Rüdiger Köppe Verlag.

Ide, N. 2000.
The XML Framework and Its Implications for the Developmen of Natural Language Processing Tools. [O] Available. http://www.cs.vassar.edu/~ide/papers/coling00-ws-final.pdf. Accessed on 4 October 2005.

Ide, N., Kilgarriff, A. and Romary, L. 2000.
A Formal Model of Dictionary Structure and Content. In: *Proceedings of EURALEX'00*. [O] Available. http://citeseer.ist.psu.edu/ide00formal.html. Accessed on 4 October 2005.

Itai, A., Wintner, S. and Yona, S. 2006.
A Computational Lexicon of Contemporary Hebrew. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, pp.19–22.

Kriel, T.J. 1976.
The New English-Northern Sotho Dictionary. King William's Town: Educum Publishers.

Mabille, A. 1976.
Sesuto-English Dictionary. Morija: Morija Printing Works.

Mabille, A. and Dieterlin, H. 1988.
Southern Sotho-English Dictionary. Morija: Morija Sesuto Book Depot.

McLaren, J. 1984.
A New Concise Xhosa-English Dictionary. Cape Town: Maskew Miller Longman.

Simons, G.F., Lewis, W.D., Farrar, O.S., Langendoen, D.T., Fitzsimons, B. and Gonzalez, H. 2004.
The semantics of markup: Mapping legacy markup schemas to a common semantics. In: *Proceedings of the 4th workshop on NLP and XML (NLPXML-2004)*, Barcelona, Spain. 2004. pp. 25–32. Also [O] Available. http://www.u.arizona.edu/~farrar/publications.html Accessed on 9 February 2005.

Van Eynde, F. and Gibbon, D. 2002.
Lexicon Development for Speech and Language Processing. Dordrecht, Boston, London: Kluwer Academic Publishers.

Van Wyk, E.B. 1995.
Linguistic Assumptions and Lexicographical Traditions in the African Languages. **Lexikos** 5: 82–96.

Weber, D.J. 2002.
Reflections on the Huallaga Quechua dictionary: derived forms as subentries. [O] Available. http://emeld.org/workshop/2002/presentations/weber/emeld.pdf Accessed on 3 October 2005.

Wittenburg, P., Peters, W. and Drude, S. 2000.
Analysis of Structures from Field Linguistics and Language Engineering. [O] Available. http://www.mpi.nl/lrec/2002/papers/lrec-pap-08-lexical-structures-talk-final.pdf Accessed on 9 February 2006.

Ziervogel, D. and Mokgokong, P.C. 1975.
Comprehensive Northern Sotho Dictionary. Pretoria: J.L. van Schaik.

About the authors: *Sonja E. Bosch* is a professor in the Department of African Languages, University of South Africa. *Laurette Pretorius* is a professor in the School of Computing, University of South Africa. *Jackie Jones* is a senior lecturer in the Department of African Languages, University of South Africa.