http://lexikos.journals.ac.za

A Computational Approach to Zulu Verb Morphology within the Context of Lexical Semantics

Sonja E. Bosch, *Department of African Languages, University of South Africa, Pretoria, South Africa (boschse@unisa.ac.za),*
and
Laurette Pretorius, *School of Interdisciplinary Research and Graduate Studies, University of South Africa, Pretoria, South Africa (pretol@unisa.ac.za)*

Abstract: The central research question that is addressed in this article is: How can ZulMorph, a finite state morphological analyser for Zulu, be employed to add value to Zulu lexical semantics with specific reference to Zulu verbs? The verb is the most complex word category in Zulu. Due to the agglutinative nature of Zulu morphology, limited information can be computationally extracted from running Zulu text without the support of sufficiently reliable computational morphological analysis by means of which the essential meanings of, amongst others, verbs can be exposed. In this article we describe a corpus-based approach to adding the English meaning to Zulu extended verb roots, thereby enhancing ZulMorph as a lexical knowledge base.

Keywords: ZULU VERB MORPHOLOGY, VERB EXTENSIONS, LEXICAL SEMANTICS, COMPUTATIONAL MORPHOLOGICAL ANALYSIS, ZULMORPH, ZULU LEXICAL KNOWLEDGE BASE, BITEXT

Opsomming: 'n Rekenaarmatige benadering tot Zoeloe werkwoordmorfologie binne die konteks van leksikale semantiek. Die sentrale navorsingsvraag wat in hierdie artikel onder die loep kom, is: Hoe kan ZulMorph, 'n eindige toestand morfologiese ontleder vir Zoeloe, gebruik word om waarde toe te voeg tot Zoeloe leksikale semantiek, met spesifieke verwysing na Zoeloe werkwoorde? Die werkwoord is die mees komplekse woordkategorie in Zoeloe. As gevolg van die agglutinerende aard van Zoeloe morfologie kan net beperkte inligting sonder die ondersteuning van voldoende betroubare rekenaarmatige morfologiese analise uit lopende Zoeloe teks onttrek word. Die morfologiese inligting stel essensiële betekenisse van, onder andere, werkwoorde, bloot. In hierdie artikel beskryf ons 'n korpusgebaseerde benadering om die Engelse betekenis aan uitgebreide werkwoordwortels van Zoeloe toe te ken en sodoende ZulMorph as leksikale kennisbasis uit te brei.

Sleutelwoorde: ZOELOE WERKWOORDMORFOLOGIE, WERKWOORDUITBREIDINGS, LEKSIKALE SEMANTIEK, REKENAARMATIGE MORFOLOGIESE ANALISE, ZULMORPH, ZOELOE LEKSIKALE KENNISBASIS, BITEKS

1. Introduction

The integral role of the Internet and the world-wide web in facilitating the production and consumption of enormous amounts of information in digital space depends on the ability of computers to perform a wide variety of tasks involving human language. This requires, amongst others, computational approaches to representing and understanding world knowledge on the one hand, and knowledge about human language in machine-processable form on the other hand. Central to this endeavour is the notion of meaning or semantics, and more specifically lexical semantics, generally defined as the linguistic study of the meaning of individual words, and the meaning-related connections between words. Moreover, contemporary research in lexical semantics as such also relies on natural language processing (NLP) for a wide range of computational approaches and on large electronic corpora that have "revolutionized the possibilities of investigating usage patterns in real language across genres and cultures and further develop probabilistic usage-based ideas." (Paradis 2012).

Typical computational lexical semantics tasks include word sense disambiguation in context, computing word similarity and word relatedness, as well as other relations between words, and semantic role labelling (Jurafsky and Martin 2009). In turn, NLP applications such as machine translation, question answering, information retrieval, information extraction, text classification and multilingual conversational agents, to name but a few, rely on these basic tasks in realising a digital space in which the users of diverse languages can participate in cross-lingual knowledge production and consumption. Performing computational lexical semantics tasks across languages brings the added complexity of requiring access to NLP support in multiple languages. For under-resourced languages it has become common practice to use a well-resourced language such as English as a type of pivot language for providing word meaning and cross-lingual lexical semantics.

Lexical semantic knowledge has up to now been captured mainly through two approaches. "The first is the knowledge-based approach, in which human linguistic knowledge is encoded directly in a structured form, resulting in various types of lexical knowledge bases. The second is the corpus-based approach, in which lexical semantic knowledge is learnt from corpora and then represented in either explicit or implicit manners." (Gurevych et al. 2016: xiii). Broadly speaking, lexical knowledge bases are knowledge bases that provide lexical information about words of a particular language.

In this article the focus is on Zulu, an official language of South Africa, which is, amongst others, characterised by its rich agglutinative morphology in which the verb is the most complex word category. In spite of its official status, Zulu is considered an under-resourced language. When dealing with under-resourced languages, it is common practice to use as much of the available language data and resources as possible. For this reason, both kinds of approaches to lexical semantic knowledge are employed: hand-crafted expert linguistic/ lexical knowledge in machine-processable form as well as growing volumes of

electronic Zulu text corpora. There is also a deeper linguistic justification for employing these two complementary approaches: The first exploits the regularity of linguistic structure — in our case the basic morphological structure and the so-called predictable meanings associated with morphemes, in this case the verb extensions. The second caters for the irregularities, the idiosyncrasies that occur in all languages, and for the "unpredictable" lexicalised meaning of extended verb roots. More specifically, *we show how ZulMorph, a comprehensive hand-crafted finite state morphological analyser for Zulu, and the South African Constitution (SAC), a small electronically available parallel English–Zulu corpus which is an official document of the highest order, translated into all official languages, can contribute to Zulu lexical semantics with English as pivot language*.

2. Basic approach

A lexical knowledge base (LKB) is a digital knowledge base "that provides lexical information about words" (Gurevych et al. 2016). Conceptually the most basic unit or entry in a lexical knowledge base is the so-called (*lemma*[1], *meaning*) pair[2]. While our ultimate aim is to construct such pairs for all the words of Zulu, *nouns* and *verbs* are specifically important since they play a central role in knowledge representation — nouns usually name concepts about which information is represented and verbs often express relationships between concepts. Moreover, verbs are the morphologically most complex word category in Zulu. The verb in Bantu languages, in general, incorporates a great deal of information, to the extent that it may even stand alone as a sentence. It is for this reason that we focus on the latter word category in this article.

We propose a computational approach based on ZulMorph. As a comprehensive hand-crafted finite state morphological analyser, ZulMorph not only contains lemmas of most Zulu words, based on various paper dictionaries, other language resources and text books for Zulu (Pretorius and Bosch 2003; Bosch and Pretorius 2006), but it is also arguably the most complete model of the morphological structure of Zulu words. So, when presented with a valid Zulu word, it provides the lemma as part of the full morphological analysis of the word. What ZulMorph does not yet provide, is the meaning of the lemma.

Representing the meaning, also often referred to as the sense, of a lemma is well-known to be hard (see, for example, Faruqui 2016) and has been studied extensively for a language such as English, generally considered a well-studied and digitally well-resourced language. Jurafsky and Martin (2009) provide an excellent introduction to and overview of computational approaches to the representation of word meaning and word sense in English. Therefore, since computational word meaning representation approaches and resources (Lazaridou et al. 2013) for Zulu are not readily available, *we propose a cross-lingual approach with English as pivot language for providing the meaning of a Zulu lemma*. More specifically, we enhance ZulMorph to output a lemma, as well as its English translation equivalent as the meaning of the lemma. Endowed with this added

capability, we then propose that ZulMorph, as basic Zulu LKB, would enable the user to rely on the rich computational infrastructure of English word meaning representation in further processing and applications.

The structure of the article is as follows: Section 2 outlines the approach followed to address the stated problem. Section 3 provides a brief overview of Zulu verb morphology with specific reference to verb extensions, their complexity, their predictability of meaning and related lexicalisation issues. We specifically emphasise morphological (lemma) and semantic (meaning) challenges. In Section 4 ZulMorph is presented as an approach to lemmatisation. As before, the focus is on verbs, their roots and their extensions. In Section 5 the hand-crafting of a basic Zulu LKB from existing paper dictionaries and grammar texts as a "snapshot" of Zulu lexical semantic information is presented. In section 6 the focus is on a corpus-based approach to semi-automatically extracting new verb roots, new extensions and new lexicalised meanings[3] for the possible addition to the ZulMorph-based LKB. Section 7 concludes the article and provides suggestions for future work.

3. Zulu verb morphology

The morphological composition of the verb is considerably more complex than that of any other word category in Zulu. A number of slots, preceding and also following the verb root, may contain numerous morphemes with functions such as derivation, inflection for tense-aspect and marking of nominal arguments. Examples are cross-reference of the subject and object by means of class- (or person-/number-) specific markers, locative affixes, morphemes distinguishing verb forms in clause-final and non-final position, negation morphemes and so forth. In this article we concentrate on the so-called verb extension morphemes (Poulos and Msimang 1998: 183-207). As is the case with most Bantu languages, the complex verb morphology of Zulu is characterised by the use of so-called verb extensions to extend or adapt the meaning of a particular verb. By means of a verb extension or a combination of extensions "definite variations of meaning are derived, variations which in English can only be made by the use of auxiliary verbs, adverbs or prepositions." (Doke 1973: 135).

In the inflectional morphology of Zulu the basic meaning of a verb root in Zulu may therefore be modified by suffixing one or more extension morphemes to the verb root[4], e.g.

(1a)	*-phind-a* -verb.root-terminative	'repeat, do again, return, go back, fold, make double, duplicate'
(1b)	*-phind-an-a* -verb.root-reciproc.ext-terminative	'fold one into the other/coil together'
(1c)	*-phind-el-a* -verb.root-appl.ext-terminative	'repeat for, fold for; return'

(1d)	*-phind-el-el-a* -verb.root-appl.ext-appl.ext-termina-tive	'repeat again and again; return again and again'
(1e)	*-phind-is-a* -verb.root-caus.ext-terminative	'cause to repeat, return, fold, send back; retaliate, take vengeance, avenge oneself'
(1f)	*-phind-is-el-a* -verb.root-caus.ext-appl.ext-termina-tive	'send back for, send back to; retaliate against, repay vengeance, revenge oneself upon'

It is significant that the verb root *-phind-* may use 22 different combinations of verb extensions of which 6 feature as headwords in the Zulu–English Dictionary (ZED) (1964: 662-663). In the outer matter (ZED 1964: ix), it is indicated that separate entries have been made for "verbal derivatives" (extended verb stems) that "convey some meaning or idiomatic usage not deducible from the inherent significance of the derivative form", e.g.

(2a)	*-hamb-a*	'travel, move along'
(2b)	*-hamb-el-a*	'visit, be on good terms with'

In other cases, where the "inherent significance of the derivative form" is easily deducible from the basic verb stem, the derivative forms are listed in brackets after the entry of the basic form, e.g.

(3)	*-pikiz-a* (pass. *-pikizwa*; ap. *-pikizela*; caus. *-pikizisa*)	'wriggle about, waggle'

According to Wilkes (1971: 261) there is theoretically no limit to the number of verb extensions that may be suffixed to a verb root. However, the database of over 6000 examples collected for his study (Wilkes *op. cit.*) contained very few examples with more than three verb extensions being used simultaneously.

In summary, verb extensions are a key feature of Zulu verbs and their meanings and have to be accounted for in a LKB for Zulu, both in terms of the easily deducible meanings and also the lexicalised and idiomatic usage.

3.1 Morphological challenges

Within a rule-based approach to morphology, the following are examples of morphological challenges (morphotactics and morphophonological alternation rules) that are encountered with regard to verb extensions:

(a) Some basic verb roots resemble extended verb roots, e.g. the verb root *-hlangan-* 'come together; unite; connect' in which the morpheme *-an-*

resembles the reciprocal extension. In this case it is not an extension but part of the verb root.

(b) Rule-based palatalisation occurs in the formation of passives when the final syllable of a verb root begins with a bilabial consonant, also when such a verb root is separated from the passive extension *-w-* by another extension, e.g.

(4a)	*-boph-a* -verb.root-terminative	'tie, fasten, button up'
	-bosh-w-a -verb.root-pass.ext-terminative	'be tied, fastened, buttoned up'
(4b)	*-boph-el-a* -verb.root-appl.ext-terminative	'tie for, imprison for'
	-bosh-el-w-a -verb.root-appl.ext-pass.ext-terminative	'be tied for, be imprisoned for'

Occasionally however, idiosyncrasies occur when bilabials appearing elsewhere in the verb root are palatalised, e.g.

(5)	*ezisetshenziswa* *-sebenz-is-w-a* -verb.root-caus.ext-pass.ext-terminative	'that are used'
	-setshenz-is-w-a (not *-sebenziswa** as expected)	

(c) The order of extensions is not always fixed. For instance the passive extension usually follows other extensions, e.g.

(6a)	*-akh-el-w-a* -verb.root-appl.ext-pass.ext -terminative	'be built for'

In some cases, the reciprocal may, however, follow the passive extension, e.g.

(6b)	*-akh-el-w-an-a* -verb.root-appl.ext-pass.ext-recip.ext -terminative (cf. Van Eeden 1956: 657)	'be built for each other'

It should be noted that we do not deal separately with verb roots that end in *-k-* and *-l-* and which are subject to varying modifications in the formation of the causative (e.g. *-vuk-is-a* > *vu-s-a; -vel-is-a* > *-ve-z-a*). The reason is that such extended roots are lemmatised as such in most dictionaries, e.g. Dent and

Nyembezi (1969: 506-507) contains the entries *-vuka* (v) 'wake up; rise up' and *-vusa* (v) 'awaken; rouse up; warn against danger; lift up'.

3.2 Semantic challenges

Whereas the basic meaning of verb roots is easily accessible from existing dictionaries, the semantic challenge lies in the extended or lexicalised meanings that come about when the verb root is extended by means of a variety and combination of verb extensions. In most grammatical descriptions of the Bantu languages, verb extensions are considered to be inflectional suffixes since "they do not change the word category to which a word belongs, but add a regular, predictable meaning to the word" (Kosch 2006: 109). The predictable meanings of extended verb roots can be summarised as in Table 1[5]:

Type of extension	Extension	Predictable meaning
passive	*-w-* or *-iw-*	be, being
applicative	*-el-*	for, on behalf of, to
intensive	*-isis-*	expresses intensity
causative	*-is-*	cause to, help
neuter	*-ek-*	cause or assist to perform an action
reciprocal	*-an-*	each other
completive	*-elel-*	action carried out to perfection or completion

Table 1: Predictable meanings of Zulu verb extensions

Not all verb roots may take all extensions arbitrarily since there are restrictions on the combinations of certain meanings (Poulos and Msimang 1998: 183). The following examples are ungrammatical (*) because the neuter extension is incompatible with the meaning of the two verbs and therefore signifies a semantic restriction:

(7a) *-ephuk-a* 'get broken; die suddenly' > *-ephuk-ek-a**

(7b) *-shon-a* 'sink, go down, die etc.' > *-shon-ek-a**

Exceptions occur when the meaning of an extended verb root is lexicalised, and therefore becomes unpredictable to a large extent. Kosch (2006: 106) singles out an extension such as the causative which is prone to lexicalisation in combination with certain verb roots. The result is an unpredictable meaning and a display of derivational properties, e.g.

(8a) *bon-a* 'see'
-verb.root-terminative

	-bon-is-a -verb.root-caus.ext-terminative	'show'
(8b)	*-lum-a* -verb.root-terminative	'bite, suffer sharp pain, itch'
	-lum-is-a -verb.root-caus.ext-terminative	'cause to bite/itch; give a bite of food to/share with'

The applicative extension is also used to indicate "in a direction" when followed by a noun indicating location, e.g.

(9)	*-gijim-el-a ezintabeni* -verb.root-appl.ext-terminative	'seek shelter in the mountains'

An interesting case is found with the meanings of the verbs *-khohla* 'escape from the memory, slip from the memory' and *-khola* 'satisfy, have confidence in', in the sense that they are unexpectedly used as transitive verbs in the passive, e.g.

(10)	*-khohl-w-a*	'forget, overlook'
	-khol-w-a	'be satisfied, believe in'

The predictable versus lexicalised meaning phenomenon has been considered from various perspectives that are important for our computational approach to the lexical semantics of Zulu verbs.

On the one hand, the predictable nature of meaning has been documented and provides justification for us to include such regularity in our computational model of Zulu verbal lexical semantics through the "standard" (rule-based) semantic annotation of verb extensions in ZulMorph. According to Wilkes (1971: 50-51) the adding of a verb extension in Zulu does not imply a radical modification of the lexical-semantic aspect of a verb since this remains basically the same. The modification that takes place is that of the manner in which a process progresses or is executed, while the nature of the process remains unchanged. In cases of combinations of verb extensions being suffixed to a verb, it is only the first suffix after the basic verb root that modifies it. Each of the following extensions in turn modifies the foregoing modification (extended root). This modification process is demonstrated in Figure 1. We return to this sequencing of extensions and their "composite" meanings in Section 5.1.

On the other hand, Chabata (1998: 146) points out that verb extensions in the Bantu language Shona are considered to be derivational morphemes and not inflectional morphemes, one of the reasons being that "they usually change the meanings of the verb roots in question in highly significant ways". This suggests that there is good reason to also make provision in our Zulu LKB for

verb extensions to have "unpredictable" lexicalised meanings. These are not systematic and cannot be captured by means of rules. They have to be found individually mainly through corpus-based approaches and added to the LKB as part of its maintenance and continued enhancement.

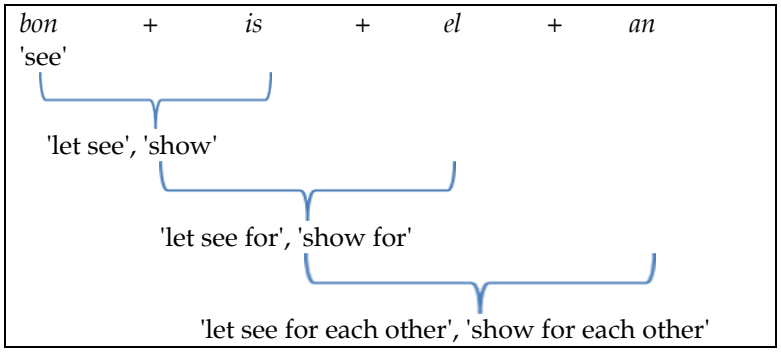


Figure 1: Left-associativity of the compositional meaning of the extended verb root *-boniselan-*, with 'let see' lexicalised as 'show'.

4. Computational Zulu verb morphology and lemmatisation

Before providing the essential details of ZulMorph as the basis for a basic Zulu LKB, we develop the core notion of *Zulu word sense pair*, in this case for verbs.

4.1 What is the lemma and word sense pair of a Zulu verb?

We start by illustrating by means of an example what a *word sense pair* — a (lemma, meaning) pair in English is. We then use this to explicate the notion of *Zulu word sense pair* — a word sense pair in which the lemma is in Zulu and its meaning is the English translation equivalent[6].

Example 1: 'travels' is a word in the English sentence 'He travels to Johannesburg.' The appropriate meaning of 'travels', according to the Princeton WordNet[7], is "undertake a journey or trip". The lemma of 'travels' is 'travel'[8] and therefore the English word sense pair is (travel, undertake a journey or trip).

But what constitutes the lemma and the Zulu word sense pair of a Zulu verb? Four important aspects have to be addressed:

(i) Lemmatisation via morphological analysis: a standard approach to lemmatisation is through computational morphological analysis (Jurafsky and Martin 2009: 645). For Zulu, the complex agglutinative morphologi-

cal structure of a Zulu verb includes, amongst others, the verb root and its verb extensions. For the purposes of Zulu verbal lexical semantics, the verb root together with its extensions, i.e. the extended verb root, constitutes the lemma of the Zulu verb. This decision is based on the insight that the lexical semantics of the Zulu verb is determined by the verb root AND its extensions since, as we have seen in Section 3, the extensions are meaning changing suffixes to the root. This aspect is addressed in Section 4.3;

(ii) Assigning a meaning in the form of its English translation equivalent to the verb root. This aspect is addressed in Section 4.4;

(iii) Assigning English meaning(s) to the verb extensions so that they can be combined (composed) with the meaning of the verb root. This aspect is addressed in Section 4.4;

(iv) Combining the information in (ii) and (iii) to yield a lemma and a word sense pair for any given Zulu verb in which the meaning is provided as the English translation equivalent of the Zulu lemma, English being our pivot language.

But how is this composite meaning of the Zulu lemma, as defined in (i), obtained? In Table 1 of Section 3.2 the predictable meanings of the respective verb extensions are given and the question now arises as to how a sequence of meanings is combined into one meaning for the extended verb root as Zulu lemma. To answer this question we rely on the left-associative compositional nature of the meaning of the verb root and its sequence of extensions, as already documented by Wilkes (1971) (see Section 3.2 and Figure 1). We illustrate this by means of Example 3. Although we primarily base our modelling of the "composite" meaning on the predictable meaning of extensions, we also attend to lexicalised meaning where relevant.

Example 2: Consider the word *uyahamba* 'he travels'. Through morphological analysis (see example (2a)) we obtain the lemma *hamb*[9]. The appropriate word sense pair is (*hamb*, travel). In further applications[10] the lemma *hamb* may then be linked to the original Princeton WordNet sense "undertake a journey or trip" via the English translation equivalent 'travel'. For the words *uyahambisa* and *uyahambela* the word sense pairs are (*hambis*, cause to travel) and (*hambel*, travel on behalf of/travel towards) or (*hambel*, visit) if the lexicalised meaning of (2b) is used. Similarly linking to the Princeton WordNet could further yield (*hambis*, cause to undertake a journey or trip) and (*hambel*, undertake a journey or trip on behalf of) or (*hambel*, go to certain places as for sightseeing).

Example 3: For the word *uyahambelisa* the lemma is *hambelis*. Its meaning is obtained by composing the respective meanings from the left, as shown by means of the bracketed representation of the lemma (((*hamb*)*el*)*is*) : 'cause to'

(meaning of *hambel*) => 'cause to travel on behalf of' , 'travel towards' or 'visit' if the lexicalised meaning is used. This yields the Zulu word sense pair (*hambelis*, cause to travel on behalf of/travel towards) or (*hambelis*, cause to visit). As before, this may then be further expanded via the Princeton WordNet to (*hambelis*, cause to undertake a journey or trip on behalf of) or (*hambelis*, cause to go to certain places as for sightseeing).

For any verb in Zulu, we are now able to conceptually provide its Zulu word sense pair. In subsequent sections we show how this lexical semantic information is computationally obtained and encoded in ZulMorph as basic Zulu LKB.

4.2 ZulMorph

ZulMorph was developed with the Xerox finite state toolkit (Beesley and Karttunen 2003) and has also been successfully compiled with Foma (Hulden 2009), a free, open source finite state toolkit. The two central problems of morphology, viz. *morphotactics* (rules for morpheme sequencing) and *morphophonological alternation rules* (rules for spelling and sound changes) are computationally modelled by and implemented as finite state transducers, which are then composed to form one single transducer, which constitutes the morphological analyser. For modelling the morphotactics the **lexc** programming language with its cascading continuation classes of morpheme lexicons (Beesley and Karttunen 2003: 210) is provided and for the alternations rules, **xfst**, a language for using the extensive Xerox finite state calculus, is used[11].

It is well-known that the coverage of a finite state morphological analyser such as ZulMorph is determined by (i) the accurate and complete modelling of the morphological structure of the language, and (ii) the comprehensiveness of the noun stem and verb root lexicons. Only valid Zulu words, of which the noun stems or verb roots are present in the respective lexicons, can be analysed correctly. For such a morphological analyser to be maximally useful, these stem and root lexicons need to be maintained and extended as new words enter the language. This remains ongoing work.

In principle, the cascading continuation classes of morpheme lexicons model the filling of slots in the morphological structure of the verb. However, the slots that we are interested in here are those for the verb root and its extensions, since together these constitute the lemma. While the order of the verbal prefixes is fixed (cf. Poulos and Msimang 1998: 305), this is not the case for the extensions. There is no fixed order or number since these are semantically determined. Indeed, as discussed in Section 2, the various verb extensions are not compatible with all verb roots, and there are no hard and fast rules that determine the possible combinations, i.e. roots with extensions, as well as extensions with one another. Comprehensive information on these combinations is not available — not even paper dictionaries provide complete information on combinations and sequences for all verb roots. The inclusion of such

"idiosyncratic" information about verb roots and their (semantically) valid extensions in ZulMorph further emphasises its role as one of the most comprehensive computational models yet of Zulu morphology.

4.3 Modelling the Zulu verb lemma

Before explaining the computational modelling of the Zulu verb lemma, we return to the morphological challenges of Section 3.1and how we address them. Challenge (a) concerns the common ambiguity of human language for which no real solution exists except to deal with it through semantic context-based disambiguation at a later stage of processing — at the morphological level such limited over-generation will thus occur. Challenge (b) is non-rule-based and is met by hand-crafting the analyser to accurately model all the individual known cases.

Challenge (c) is closely related to aspect (i) in Section 4.1 and is the core of this section. In modelling verbs and their lemmas in ZulMorph, we make provision for different possibilities: a known basic root with no extensions, a known basic root with its own attested sequence(s) of extensions and a known basic root with an as yet unattested (i.e. new) sequence of extensions. Verbs based on basic roots that are not included in ZulMorph will not be analysed. As we shall see in this section, the distinction between morphology and the root lexicon becomes somewhat fuzzy in the case of the Zulu verb and its extensions in that the attested extension sequences of any specific basic verb root should be marked on the relevant basic root and thereby become part of the "lexicon".

In order to describe the modelling of the Zulu verb lemma and its meaning, we briefly explain the notion of `Lexicon` in **lexc**, as well as the technical use of so-called flag diacritics in both the Xerox toolkit and Foma. We show how they are used to record information about the verb lemma in **lexc.**

4.3.1 The verb root lexicon, extension sequences and flag diacritics

In order to keep explanations short, an example is used instead of trying to explain the technical details in a more general setting. The example **lexc** script for the root *-hamb-* is given in Appendix A. As a code fragment for explanatory purposes, it does not, for example, show how verbal prefixes are modelled. It consists of broadly four sections: the preamble in which certain so-called multi-character symbols are declared, the verb root lexicon that typically contains thousands of roots, but for the example contains only the entry *-hamb-*, the modelling of the verb extensions and finally the morpheme lexicon containing the verb terminative *-a*. Each section is briefly discussed.

In the preamble two tags, `[ATT]` and `[NEW]`, are declared for distinguishing between attested extension sequences for *-hamb-* and possibly newly discov-

ered ones in the output produced by ZulMorph, as well as a number of so-called flag diacritics[12] that are used to mark the attested extension sequences of any particular verb root in the verb root lexicon (`LEXICON VRoot`) in the second section. This lexicon contains various entries for the verb root *-hamb-* each annotated with a `P` flag diacritic that encodes the specific attested extension sequence. It also shows the next continuation class (`Lexicon VExt`) containing the morpheme lexicon from which the next morpheme in the input verb should be matched. The third section shows the morpheme lexicons of next morpheme(s) (extensions) that may follow the basic verb root in accordance with the structure of the verb. As mentioned before, we distinguish between the basic root with no extensions, attested and new extensions. This is modelled by `LEXICON VExt` and its continuation classes `VerbTerm`, `VExtAttested` and `VExtNew`. In lexicon `VExtAttested` the `R` flag diacritic is used to match precisely the attested extension sequence that was marked by the corresponding `P` flag diacritic in the verb root lexicon entry. The lexicon `VExtNew` and the cyclic lexicon `VExtNew2` model any new extension sequence of arbitrary length. The fourth section shows the last continuation class, `LEXICON VerbTerm`, which models the final verb terminative morpheme, here `-a`, followed by `#` to indicate that no further (input) morphemes may follow.

While the attested extension sequences are precise and correct, the cyclic modelling of the recognition of new as yet unattested sequences may cause over-generation in that any arbitrary (finite) sequence of extensions, even sequences that are semantically not plausible, will be recognised. This implementation is specifically useful for the purposes of mining new sequences of extensions from a corpus with the understanding that any new sequence will be subjected to human elicitation before inclusion in ZulMorph as an attested sequence.

4.3.2 Coverage

By way of example, the verb root *-hamb-* in Appendix A has eight different extension sequences. In ZulMorph *-hamb-* has eleven attested sequences: *-w-*, *-ek-*, *-el-*, *-is-*, *-isis-*, *-elw-*, *-elel-*, *-elan-*, *-isw-*, *-isel-* and *-isan-*.

ZulMorph contains 8 031 basic roots and 28 477 (extended) verb roots with attested extension sequences, bringing the number of entries in the verb root lexicon of ZulMorph to approximately 36 000. From the extensive data harvested from available paper dictionaries, grammar textbooks and other paper resources, 113 different extension sequences were identified, with the first 20 most frequent sequences (see Appendix B) representing more than 97% of all attested extensions. Statistics about the number of extensions per basic verb root are provided in Appendix C. We note that 22 of them allow between 20 and 30 combinations of one or more verb extensions. The number of lexicalised headwords, as recorded by Doke and Vilakazi (1964), is given in brackets. For example, the basic verb root in ZulMorph with the largest number of extensions,

viz. 30, is *-fan-* ('resemble'). The basic root *-bon-* ('see') has 28 extension sequences. Moreover, ZulMorph contains 6 153 basic verb roots that have at least one attested extension and 1 878 that have no attested extensions. In Appendix D we list the basic verb roots that have the longest attested extension sequences, as recorded in ZulMorph, for example:

(11) *-ling-an-is-el-a* 'equalise for/make equal for'

The extensive coverage of both Zulu morphology and its verb roots, basic and extended, in ZulMorph provides the basis for the LKB of the next section.

5. Hand-crafting a basic LKB for Zulu

Hand-crafting a basic LKB for Zulu consists of a systematic and comprehensive usage of the expert knowledge that has been published and made available for Zulu. Three kinds of information need to be encoded — firstly the morphology, secondly the Zulu lemmas and thirdly their meanings. Since ZulMorph is an accurate model of Zulu morphology and its comprehensive coverage of Zulu verb lemmas was addressed in the previous section, we now turn our attention to the acquisition and inclusion of their meanings using a so-called expert knowledge-based approach, as already alluded to in Section 4.1 (ii)–(iv). More specifically, a meaning in the form of an English translation equivalent is assigned to each verb root and its extensions. While our main focus is on predictable meaning as a first step, lexicalised meaning is also considered.

5.1 Representing the meaning of the lemma

The first step in adding meaning to each basic verb root in ZulMorph is including the English translation equivalent to each basic verb root in the `VRoot` lexicon and the predictable meaning to each extension in the `Attested`, `VExtNew` and `VExtNew2` lexicons. For example, the code fragments

```
hamb(travel)@P.Basic.ON@: hamb@P.Basic.ON@ VExt;
hamb(travel)@P.ExtEL.ON@: hamb@P.ExtEL.ON@ VExt;
```

and

```
an(each_other)[RecipExt]: an                    VExtNew2;
el(for)[ApplExt]: el                            VExtNew2;
is(cause_to)[CausExt]: is                       VExtNew2;
el(for)[ApplExt]@R.ExtEL.ON@: el@R.ExtEL.ON@    VerbTerm;
```

yield the following analyses[13]:

```
uyahambela:
u[SC][1]ya[LongPres]hamb(travel)[VRoot]el(for)[ApplExt][ATT]a[VT]
u[SC][2ps]ya[LongPres]hamb(travel)[VRoot]el(for)[ApplExt][ATT]a[VT]
u[SC][3]ya[LongPres]hamb(travel)[VRoot]el(for)[ApplExt][ATT]a[VT]
```

```
bayahambelisana:
ba[SC][2]ya[LongPres]hamb(travel)[VRoot]el(for)[ApplExt]is(cause_
to)[CausExt]an(each_other)[RecipExt][NEW]a[VT]
```

The respective word sense pairs are (*hambel*, travel for/travel towards) and (*hambelisan*, cause to travel for/towards each other). Note the composite meaning in the latter pair.

By adding basic meanings to the 8 031 basic verb roots and by including the predictable meanings of the various extensions (7 in total) we are able to provide not only a first approximation of the meaning of each of the ~36 000 entries in the verb root lexicon, but also produce word sense pairs for all the Zulu verbs that are based on these basic roots. Keeping in mind that the extensive Princeton WordNet for English has 11 529 verbs, the ZulMorph coverage of the Zulu extended verb root semantics is significant and can already be used in applications, as alluded to in Section 1.

Adding lexicalised meaning is the most resource intensive part of endowing ZulMorph verb analyses with accurate lexical semantics since it has to be added manually for each verb root individually. For each basic verb root and a particular extension sequence for which a lexicalised meaning is available, the meaning of the *basic* root is replaced by the lexical meaning of the *extended* root while the meaning of the extension that caused the lexicalisation is no longer explicit. The tag `[LEX]` shows that lexicalisation has occurred. As before, the predictable meanings of any subsequent extensions, if present, are still shown. By way of example we consider the extended root *-hambel-*, which also has the lexicalised meaning of 'visit'. Therefore, the verb root lexicon entry is as follows:

```
hamb(visit)[VRoot]el[ApplExt][LEX]@P.Lex.ON@ @P.Basic.ON@: hambel@
P.Lex.ON@ @P.Basic.ON@ VExt;
```

and yields the analyses

```
uyahambela:
u[SC][1]ya[LongPres]hamb(visit)[VRoot]el[ApplExt][LEX]a[VT]
u[SC][2ps]ya[LongPres]hamb(visit)[VRoot]el[ApplExt][LEX]a[VT]
u[SC][3]ya[LongPres]hamb(visit)[VRoot]el[ApplExt][LEX]a[VT]
```

The resulting word sense pair is (*hambel*, visit).

In summary, by annotating each entry in the verb root lexicon with its meaning (either predictable or lexicalised) and by providing the meanings of the 113 extension sequences, the morphological analysis of any Zulu verb will contain sufficient semantic information to support a basic notion of semantic linking or interoperability — a possibility that did not exist before.

6. Enhancing the Zulu LKB through a corpus-based approach

Improving and updating an electronic LKB to keep it current and maximally

useful, specifically for an under-resourced language such as Zulu, is essential for its digital (web) presence, as discussed in Section 1. Having exploited available paper resources such as dictionaries, grammar textbooks, wordlists and terminologies etc., the obvious next step is to "mine" electronically available language corpora for new lexical information to add to ZulMorph. Such lexical information includes new verb roots, new extension sequences, and new (as yet unrecorded) lexicalised meanings of extended roots as they occur in authentic language use. For this purpose we propose in this section a semi-automated corpus-based approach to the extraction of new lexical information about verbs.

By way of example, the SAC (parallel English and Zulu versions) that has been sentence-aligned is used. It was chosen for mainly four reasons: firstly it is publicly available in all the official South African languages, secondly it is assumed to have been professionally quality assured, thirdly it is by its very nature well-structured and lends itself to accurate sentence alignment, and fourthly it uses contemporary formal language. The idea is that this process should be continued as new parallel corpora become available in due course.

The extraction of bilingual lexical information from bitexts[14] has a long tradition. Tiedemann (2011) provides an overview of techniques that may be applied for this purpose. Although he focuses on statistical approaches to word alignment, he also briefly discusses a number of non-statistical techniques for lexicon extraction from bitexts (Tiedemann 2011: 100-102). While automatic word alignment "is just too noisy to be useful for qualitative investigations", these non-statistical techniques "focus on the extraction of reliable translation equivalents", usually emphasising high precision links between words and multi-word units.

The approach that we follow in this article may also be seen as such a non-statistical technique aimed at high precision.

New basic verb roots lead to morphological analysis failures. Through human elicitation and by individually considering these failures, new basic roots are identified and added to ZulMorph, together with their English translation equivalents. Alternatively, we could apply the guesser variant of ZulMorph to the failures and in this way obtain new verb root candidates. These also need to be subjected to human linguistic scrutiny before adding them to ZulMorph. The occurrence of *a new extension sequence* is tagged in the morphological analyses of a verb as `[NEW]`. Such a sequence is then manually checked and added to ZulMorph, as shown in Section 5. For *additional attested sequences for specific basic verb roots* basically the same procedure is followed.

For the extraction of *new (lexicalised) meanings and (extended) roots as they occur in authentic language use* we employ bitexts — it is here that the sentence aligned parallel corpus plays a central role. For each sentence may we proceed as follows:

1. Perform part of speech (POS) tagging of the English sentence. For this purpose we used TreeTagger[15].

2. Perform a morphological analysis of the Zulu sentence, using ZulMorph.
3. Isolate the verbs in the English sentence using the POS tags, and the verb roots and their extensions in the Zulu sentence using the morphological analysis tags for the verb root and its verb extension, and align these (the POS tags and morphological tags). This directly links the English lemma[16], i.e. the new (lexicalised) meaning, which is our translation equivalent for the new Zulu word sense pair, and the Zulu (extended) verb root, the Zulu lemma in our new word sense pair.
4. Add the information to ZulMorph so that it includes the new Zulu word sense pair.

In this semi-automated process steps 1 and 2 are automated while steps 3 and 4 as yet require manual intervention.

Specific examples that have been extracted in this way are shown in Tables 2–6[17].

In Table 2 we demonstrate how a new lexicalised meaning 'impart' has been detected for *-dlulis-* in the verb alignment process. In sentence <s103>[18] the English verb 'impart' links up with the extended Zulu verb root *-dlulis-*, forming a new lexicalised addition to those already listed for *-dlulisa* in the ZED (1964: 162), namely 'cause to pass; carry past, send past …'. Verb alignment between a new lexicalized meaning 'impart' and the Zulu lemma *dlulis* therefore results in the new word sense pair (*dlulis*, impart).

English sentence <s103> fragment	**Zulu sentence <s103> fragment**
… freedom to receive or impart information …	… inkululeko yokuthola noma ukudlulisa imininingwane …
English — Word, POS and lemma	Zulu — morphologically analysed with ZulMorph, disambiguated manually
<s103> freedom NN freedom to TO to receive VB receive or CC or impart VB impart information NN information	<s103> inkululeko: i[NPrePre][9]n[BPre][9]khululeko.9-10[NStem] yokuthola: ya[PC][9]u[NPrePre][15]ku[BPre][15]thol[VRoot]a[VT] noma: noma[Conj] ukudlulisa: u[NPrePre][15]ku[BPre][15]**dlul[VRoot]is[CausExt]**[ATT]a[VT] imininingwane: i[NPrePre][4]mi[BPre][4]niningwane.3-4[NStem]

Table 2: New lexicalisation of Zulu lemma *dlulis*

Table 3 demonstrates verb alignment between the new lexicalized meaning 'limit' and the Zulu lemma *nciphis* to form a new word sense pair (*nciphis*, limit). The English verb 'limit' in sentence <s286> links up with the extended Zulu verb root *-nciphis-* and produces a new lexicalised supplement to those already

listed for *-nciphisa* in the ZED (1964: 532): 'diminish; make small, less; minimize'.

English sentence <s286> fragment	**Zulu sentence <s286> fragment**
… no law may limit any right …	… awukho umthetho onganciphisa noma yiliphi ilungelo …
English — Word, POS and lemma	Zulu — morphologically analysed with ZulMorph, disambiguated manually
<s286> no DT no law NN law may MD may limit VB limit any DT any right NN right	<s286> awukho: a[NegPre]u[SC][3]khona[Adv] umthetho: u[NPrePre][3]mu[BPre][3]thetho.3-4[NStem] onganciphisa: o[RC][3]nga[Pot]**nciph[VRoot]is[CausExt]**[ATT]a[VT] noma: noma[Conj] yiliphi: yi[CopPre]li[EC][5]phi[EnumStem] ilungelo: i[NPrePre][5]li[BPre][5]lungelo.5-6[NStem]

Table 3: New lexicalisation of Zulu lemma *nciphis*

Verb alignment between the novel lexicalised meaning 'register' and the Zulu extended verb root *-bhalis-* is shown in Table 4. A new word sense pair (*bhalis*, register) is created for possible inclusion in dictionaries (e.g. ZED, and isiZulu.net) where *-bhalisa* has not yet been listed as headword. It should be noted however, that the SZD (1969: 309) lists *-bhalisa* as headword with the meaning 'put name on waiting list', while the OZSD (2010: 18) does in fact list *-bhalisa* with the meaning 'register'.

English sentence <s2471> fragment	**Zulu sentence <s2471> fragment**
… to register that immovable property …	… wokubhalisa leyo mpahla engagudluki …
English — Word, POS and lemma	Zulu — morphologically analysed with ZulMorph, disambiguated manually
<s2471> to TO to register VB register that DT that immovable JJ immovable property NN property	<s2471> wokubhalisa: wa[RC][3]u[NPrePre][15]ku[BPre][15]**bhal[VRoot]is[CausExt]** [ATT]a[VT] leyo: leyo[Dem][9][Pos2] mpahla: n[BPre][9]phahla.9-10[NStem] engagudluki: e[RC][9]nga[NegPre]gudluk[VRoot]i[VTNeg]

Table 4: New lexicalisation of Zulu lemma *bhalis*

In Table 5 it becomes clear how a new lexicalised meaning 'affirm' has been identified for the Zulu lemma *qinisekis* in the verb alignment process. In sentence <s51> the English verb 'affirm' links up with the extended Zulu verb root *-qinisekis-*, forming a new lexicalised addition to those already listed for *-qinisekisa* in the OZSD (2010: 198): 'make sure; make certain'. Verb alignment between a new lexicalized meaning 'affirm' and the Zulu lemma *qinisekis* therefore results in the new word sense pair (*qinisekis*, affirm), which could also qualify for inclusion in a dictionary such as ZED, where *-qinisekisa* has not yet been listed as headword. The same procedure applies to the occurrence of the extended Zulu verb root *-qinisekis* as occurs in <s45> and < s157> respectively, resulting in two further new word sense pairs (*qinisekis*, ensure) and (*qinisekis*, secure).

English sentence <s51> fragment	**ZUL SAC text <s51>**
… affirms the democratic values of human dignity …	… uqinisekisa amagugu entando yeningi yokwazisa isithunzi somuntu …
ENG — Word, POS and lemma	Zulu — morphologically analysed with ZulMorph, disambiguated manually
<s51> affirms VBZ affirm the DT the democratic JJ democratic values NNS value of IN of human JJ human dignity NN dignity	<s51> uqinisekisa: u[SC][3]**qin[VRoot]is[CausExt]ek[NeutExt]is[CausExt]**[ATT]a[VT] amagugu: a[NPrePre][6]ma[BPre][6]gugu.5-6[NStem] entando: a[PC][6]i[NPrePre][9]n[BPre][9]thando.9-10[NStem] yeningi: ya[PC][9]i[NPrePre][9]n[BPre][9]ningi.9[NStem] yokwazisa: ya[PC][9]u[NPrePre][15]ku[BPre][15]az[VRoot]is[CausExt][ATT]a[VT] ya[PC][9]u[NPrePre][15]ku[BPre][15]az[VRoot]is[CausExt]a[VT] isithunzi: i[NPrePre][7]si[BPre][7]thunzi.7-8[NStem] somuntu: sa[PC][7]u[NPrePre][1]mu[BPre][1]ntu.1-2[NStem]

Table 5: New lexicalisation of Zulu lemma *qinisekis*

The example in Table 6 reflects verb alignment between the new lexicalised meaning 'refer back' and the Zulu lemma *buyisel* resulting in a new word sense pair (*buyisel*, refer back). The English verb 'refer' together with its RB (adverb) 'back' in sentence <s364> links up with the extended Zulu verb root *-buyisel-* and produces a new lexicalised supplement to those already listed for *-buyisela* in the ZED (1964: 96), namely 'restore to; return to … make amends to; replace for or by … retaliate … fill up again (as river)'. The same concerns the OZSD (2010: 22) where *-buyisela* is listed with the meanings 'return to; bring back to … restore (to)'. Another example that confirms this new word sense pair occurs in <s766> of the SAC, namely:

> **… referring** a Bill **back** to the National Assembly …
> **… ukubuyisela** uMthethosivivinywa emuva ePhalamende …

English sentence <s364> fragment	ZUL SAC text <s364>
… it may refer a dispute back to the organs of state involved …	… ingayibuyisela leyo ngxabano kulezo zingxenye zombuso ezithintekile …
ENG — Word, POS and lemma	Zulu — morphologically analysed with ZulMorph, disambiguated manually
<s364> it PP it may MD may refer VB refer a DT a dispute NN dispute back RB back to TO to the DT the organs NNS organ of IN of state NN state involved VBN involve	<s364> ingayibuyisela: i[SC][9]nga[Pst]yi[OC][9]**buy[VRoot]is[CausExt]el[ApplExt]**[ATT]a[VT] leyo: leyo[Dem][9][Pos2] ngxabano: n[BPre][9]ngxabano.9-10[NStem] kulezo: ku[LocPre]lezo[Dem][10][Pos2] zingxenye: zin[BPre][10]ngxenye.9-10[NStem] zombuso: za[PC][10]u[NPrePre][3]mu[BPre][3]buso.3-4[NStem] ezithintekile: ezi[RC][10]thint[VRoot]ek[NeutExt][ATT]ile[VTPerf]

Table 6: New lexicalisation of Zulu lemma *buyisel*

Finally, the discussion of how the chosen bitext and the semi-automated process were used to uncover new lexical information is concluded by considering further examples in Tables 7 (new root), 8 (new extension sequences) and 9 (new lexicalisations).

Extended verb root and its new English lexicalisation from SAC bitext	*-chibiyelw-* 'amended by'
Verb root	*-chibiyel-*
Extension	*-w-*
Examples from SAC bitext	*si-chibiyel-w-e* 'amended by' <s511>, <s869>, <s1863> *i-chibiyel-w-e* 'amended by' <s1200> *li-chibiyel-w-e* 'amended by' <s2467>
Comments	An example of a new verb root and its extension that does not as yet occur in the ZulMorph embedded verb root lexicon: *-chibiyel-w-* the meaning of which is 'amended by' as harvested from the SAC. The fact that *-chibiyela* does occur in the monolingual ISZ (2006: 143), serves as affirmation of the validity of the verb stem.

Table 7: New root *-chibiyel-* identified from parallel bilingual SAC corpus

Verb and its new English new lexicalisation from SAC bitext	*-xoxisan-* 'negotiate'
Verb root	*-xox-*
Extensions	*-is-an-*
Example from SAC bitext	*uku-xox-is-an-a* 'negotiating' <s1973>

Comments	*-xox-is-an-a* is not listed as headword in the ZED (1964), nor is the extension string *-is-an-* listed under the entry *-xoxa* 'narrate, tell, give account, hold conversation, converse, chat' (ZED 1964: 868). The extension string *-is-an-* for the verb root *–xox-* is also not an attested combination in ZulMorph.

Table 8a: New extension sequence *-is-an-* for *-xox-* identified from parallel bilingual SAC corpus

Verb and its new English new lexicalisation from SAC bitext	*-hlinzekelw-* 'be provided for'
Verb root	*-hlinz-*
Extensions	*-ek-el-w-*
Examples from SAC bitext	*ku-hlinz-ek-el-w-a* 'be provided for' <s174> *zi-hlinz-ek-el-w-e* 'be provided for' <s1434> *ku-nga-hlinz-ek-el-w-a* 'may be provided for' <s1629>
Comments	*-hlinz-ek-el-w-* is not listed as headword in the ZED (1964), nor is the extension string *-ek-el-w-* listed under the entry *-hlinzeka* "get skinned, murdered, operated upon … prepare food for expected visitor" (ZED 1964: 329). The extension sequence *ek-el-w-* in combination with the verb stem *–hlinza* does not occur in the monolingual ISZ (2006: 486), and it is also not an attested combination in ZulMorph.

Table 8b: New extension sequence *-ek-el-w- for -hlinz-* identified from bilingual SAC corpus

Extended verb root and its new English lexicalisation from SAC bitext	*-shicilel-w-* 'published by'
Verb root	*-shicilel-*
Extension	*-w-*
Examples from SAC bitext	*u-shicilel-w-e* 'is published' <s307> *u-shicilel-w-e* 'be published' <s1258>, <s1334>
Comments	An example of a new lexicalisation of a verb root and its extension: *-shicilel-w-* the meaning of which is 'published by' as harvested from the SAC. The verb stem *–shicilela* is not listed in the OZSD (2010), although it does occur in the SZD (1969: 479) with the meaning 'print; make an impression', in isiZulu.net with the meaning 'print, publish' as well as in the monolingual ISZ (2006: 1104), which is a confirmation of the validity of the verb stem. Furthermore, it is interesting to note that isiZulu.net does not recognise the passive extension in combination with *-shicilela*. It should be noted that the verb root *-shicilel-* does not as yet occur in the ZulMorph embedded verb root lexicon.

Table 9a: New lexicalisation of *-shicilel-* identified from parallel bilingual SAC corpus

Extended verb root and its new English lexicalisation from SAC bitext	*-khankasel-* 'campaign, lobby'
Verb root	*-khankas-*
Extension	*-el-*
Examples from SAC bitext	*nelo-ku-khankas-el-a* 'campaign' <s118> *o-ku-khankas-el-a* 'campaign' <s1571> *uku-khankas-el-a* 'lobby' <s1582>
Comments	An example with a "new" meaning that differs from the basic (original) meaning of the verb stem: *-khankasa* according to the ZED (1964: 380) is "move in horseshoe formation, with a view to intercepting or outmanoeuvring" or according to the SZD (1969: 387) the meaning is "press on (as army), be on the war-path". The new lexicalisation of the extended verb root *-khankasel-* as identified from parallel bilingual SAC corpus is 'to campaign for', 'to lobby'. The verb stem *-khankasa* is listed neither in the OZSD (2010) nor in isiZulu.net, although it does occur (with its extension *-el-)* in the monolingual ISZ (2006: 552), which is a confirmation of the validity of the verb stem.

Table 9b: New lexicalisation of *-khankas-* identified from parallel bilingual SAC corpus

7. Conclusion and future work

We have shown how ZulMorph, a comprehensive hand-crafted finite state morphological analyser for Zulu, and a small electronically available parallel English–Zulu corpus, namely the South African Constitution (SAC), which is an official document of the highest order, translated into all official languages, can enrich Zulu lexical semantics with English as pivot language.

While our approach to enhancing ZulMorph to produce Zulu word sense pairs applies to all word categories, our focus was on the verb as the morphologically most complex word category in Zulu. This complexity arises mainly from (sequences of) verb extensions that are suffixed to the basic verb root to produce modified or new verb meanings. We noted that although a morphological analyser may provide accurate morphological analyses of Zulu verb constructions, these analyses do not offer much information in terms of the meaning of the verb. This constitutes a major impediment to a computational understanding of what a Zulu verb means, and therefore also to applications such as, for example, information extraction from Zulu text, question answering in and from Zulu, machine translation between Zulu and any other language and Zulu natural language generation. In this article we presented a Zulu LKB that uses the well-resourced English language as pivot language towards addressing this challenge.

It is important to note that for a language such as Zulu (morphologically complex and under-resourced) statistical and machine learning approaches

have not yet yielded sufficiently accurate results for the applications mentioned above. Recent experience has shown that building the necessary high-quality, sufficiently large electronic corpora for Zulu has proven more difficult and expensive than handcrafting ZulMorph. This is clear from the fact that ZulMorph actually exists while no corpus-driven statistical approach to Zulu computational (verb) morphology has, as yet, yielded results that are comparable to those of ZulMorph. It is our view that the Zulu LKB that we have reported on in this article has the potential to serve as an important and novel component in future hybrid systems (robust combinations of handcrafted, rule-based, statistical and data-driven machine learning approaches) for Zulu lexical semantics.

Our core contribution is twofold:

— the enhancement of ZulMorph to constitute a large basic LKB for Zulu that, for any input verb, produces a word sense pair consisting of the Zulu lemma of the verb (here the extended root) and its meaning (here its English translation equivalent). The meaning is computationally composed from the meaning of the root and the predictable meaning of its verb extensions;

— a proposed semi-automated corpus-based approach in which existing NLP tools, viz. TreeTagger and ZulMorph, and a bitext in the form of the electronically available sentence-aligned English–Zulu parallel corpus, are used to expose new verb roots, new extension sequences and new lexicalisations of existing verbs and their extensions for addition to the Zulu LKB.

Future work may include increasing the automation of the process while also extending the process to other word categories to offer a more comprehensive Zulu LKB. We also envisage using further parallel English–Zulu corpora across a variety of domains as they become available to extend ZulMorph and the Zulu LKB, and eventually experimenting with the use of the Zulu LKB in some of the mentioned applications. In the longer term we may consider developing LKBs for other languages for which finite state morphological analysers are available.

Endnotes

1. The canonical or so-called citation form of a surface word form. For example, *write* is the lemma of the surface forms *writes*, *wrote* and *written* (cf. Section 3).
2. See, for example, Gurevych et al. (2016: 1)
3. Lexicalisation is also discussed in detail in subsequent sections.
4. For the sake of convenience a verb root followed by one or more extensions, is called an extended root in this article.
5. Also cf. De Schryver (2010: 178).

6. A *word* is taken to be a surface word form as found in a sentence or an utterance; a *lemma* is a specific grammatical form of a word, often also referred to as citation form or canonical base form; *lemmatisation* is the process of mapping a word to a lemma; *meaning* is the denotation, referent, or idea associated with a word; and a *translation equivalent* is a corresponding word or expression in another language (see, for example, Jurafsky and Martin 2009: 645; Gurevych et al. 2016: 1).
7. Princeton University "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>
8. In English the canonical base form of the verb (travel, travels, travelling, travelled) is 'travel'.
9. While we consistently use the hyphen (-) to indicate morpheme boundaries, we view the lemma as an entity that can stand on its own in the context of a word sense pair and therefore the notion of morpheme boundary is not important and therefore not indicated.
10. A discussion of such applications falls outside the scope of this article.
11. The detailed explanation of the **lexc** and **xfst** languages falls outside the scope of the article. The interested reader is referred to Beesley and Karttunen (2003).
12. Flag diacritics provide a light-weight approach to feature-setting and feature-unification operations for enhancing modelling accuracy and runtime efficiency. Specific uses are to enforce separated dependencies and mark idiosyncratic morphotactic behaviour (see Beesley and Karttunen 2002) for a comprehensive exposition). In **lexc** and **xfst** flag diacritics are so-called multicharacter symbols with a distinctive spelling: `@operator.feature.value@` and `@operator.feature@` where the operators are `P` (positive (re)setting), `N` (negative (re)setting), `R` (require test), `D` (disallow test), `C` (clear feature) and `U` (unification test). The features and values are specified by the user. In ZulMorph flag diacritics are used extensively to, amongst others, model the Zulu noun class system (Bosch and Pretorius 2002; Pretorius and Bosch 2003), long distance dependencies (Pretorius and Bosch 2008), part of speech information and a wide variety of other morphotactic constraints that apply in Zulu. In this article the focus is on their use for annotating each basic verb root with its valid and attested extension sequences.
13. The morphological tags, enclosed in [and], are listed in Appendix E.
14. The term bitext originally referred to "documents along with their translations into other languages to be used in translation studies. Since then, bitexts have attracted a lot of interest in a larger community with many other applications in mind. Therefore, it is now common to let the term bitext refer to a wider range of parallel resources, not only original documents and their direct translations. … An important characteristic of a bitext is the property that there is some kind of correspondence between the two texts coupled together, for example, translational equivalence." (Tiedemann 2011: 1)
15. http: //www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
16. TreeTagger terminology
17. The English POS tags are at http: //www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf and the Zulu tags are provided in Appendix E.
18. For example, the number <s103> refers to sentence 103 in the bitext.
19. The number in brackets after the root provides the number of lexicalised headwords, as recorded by Doke and Vilakazi (1964).

20. Class numbers are added to the tags as `[c]` where c is the class number. For example, `[BPre][5]` denotes the basic prefix of class 5.

References

Dictionaries and Corpora

ISZ = **Mbatha, M.O. (Ed.).** 2006. *Isichazamazwi SesiZulu*. Pietermaritzburg: New Dawn Publishers.

OZSD = **De Schryver, G.-M. (Ed.).** 2010. *Oxford Bilingual School Dictionary: Zulu and English / Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgisi.* Cape Town: Oxford University Press Southern Africa.

SAC = *South African Constitution*. Available at: http://www.justice.gov.za/legislation/constitution/SAConstitution-web-eng.pdf and http://www.constitutionalcourt.org.za/site/theconstitution/thetext.htm.

SZD = **Dent, G.R. and C.L.S. Nyembezi.** 1969. *Scholar's Zulu Dictionary*. Pietermaritzburg: Shuter & Shooter.

ZED = **Doke, C.M. and B.W. Vilakazi.** 1964. *Zulu–English Dictionary*. Second Edition. Johannesburg: Witwatersrand University Press.

isiZulu.net. http: //isizulu.net

Other literature

Beesley, K.R. and L. Karttunen. 2003. *Finite State Morphology*. Stanford: CSLI Publications.

Bosch, S.E. and L. Pretorius. 2002. The Significance of Computational Morphological Analysis for Zulu Lexicography. *South African Journal of African Languages* 22(1): 11-20.

Bosch, S.E. and L. Pretorius. 2006. A Finite-state Approach to Linguistic Constraints in Zulu Morphological Analysis. *Studia Orientalia* 103: 205-227.

Chabata, E. 1998. Using the Predictability Criterion for Selecting Extended Verbs for Shona Dictionaries. *Lexikos* 8: 140-153.

De Schryver, G.-M. 2010. Revolutionizing Bantu Lexicography — A Zulu Case Study. *Lexikos* 20: 161-201.

Doke, C.M. 1973. *Textbook of Zulu Grammar.* Sixth Edition. Cape Town/Johannesburg: Longman Southern Africa.

Faruqui, M. 2016. *Diverse Context for Learning Word Representations.* Ph.D. Thesis. Pittsburgh, PA: Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

Gurevych, I., J. Eckle-Kohler and M. Matuschek. 2016. *Linked Lexical Knowledge Bases: Foundations and Applications.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Hulden, M. 2009. Foma: A Finite-state Compiler and Library. *Proceedings of the EACL 2009, Demonstrations Session*: 29-32, Athens: Association for Computational Linguistics.

Jurafsky, D. and J.H. Martin. 2009. *Speech and Language Processing*. Second Edition. New Jersey, USA: Pearson Education.

Kosch, I.M. 2006. *Topics in Morphology in the African Language Context.* Pretoria: Unisa Press.

Lazaridou, A., M. Marelli, R. Zamparelli and M. Baroni. 2013. Compositionally Derived Representations of Morphologically Complex Words in Distributional Semantics. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 4–9 2013: 1517-1526. Association for Computational Linguistics.

Paradis, C. 2012. Lexical Semantics. Chapelle, C.A. (Ed.). 2012. *The Encyclopedia of Applied Linguistics*: 3357-3356. Oxford: Wiley-Blackwell.

Poulos, G. and C.T. Msimang. 1998. *A Linguistic Analysis of Zulu*. Cape Town: Via Afrika.

Pretorius, L. and S.E. Bosch. 2003. Finite-State Computational Morphology: An Analyzer Prototype for Zulu. *Machine Translation. Special Issue on Finite-state Language Resources and Language Processing* 18: 195-216.

Pretorius, L. and S. Bosch. 2008. Containing Overgeneration in Zulu Computational Morphology. *Southern African Linguistics and Applied Language Studies* 26(2): 209-216.

Tiedemann, J. 2011. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies, May 2011 (doi: 10.2200/S00367ED1V01Y201106HLT014).

Van Eeden, B.I.C. 1956. *Zoeloe-grammatika*. Stellenbosch/Grahamstown: Die Universiteitsuitgewers en -Boekhandelaars.

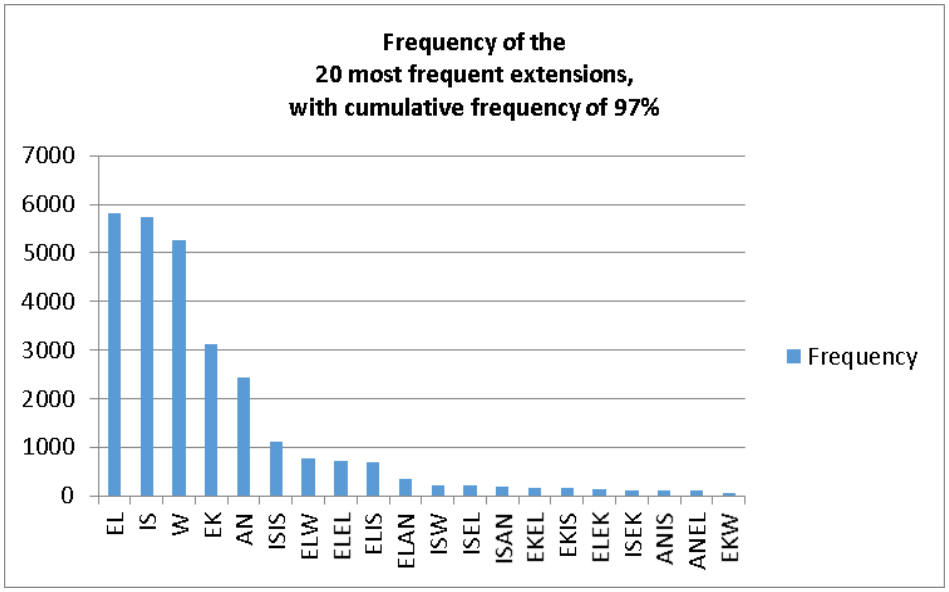
Wilkes, A. 1971. *Agtervoegsels van die werkwoord in Zulu.* Unpublished Doctoral Thesis. Johannesburg: Rand Afrikaans University.

Addendum

Appendix A: Example lexc script for the verb root *-hamb-* 'walk, travel'

```
Multichar_Symbols
@P.Basic.ON@ @R.Basic.ON@ @D.Basic@
@P.ExtEK.ON@ @R.ExtEK.ON@
@P.ExtELAN.ON@ @R.ExtELAN.ON@
@P.ExtELEL.ON@ @R.ExtELEL.ON@
@P.ExtEL.ON@ @R.ExtEL.ON@
@P.ExtISAN.ON@ @R.ExtISAN.ON@
@P.ExtISEL.ON@ @R.ExtISEL.ON@
@P.ExtISIS.ON@ @R.ExtISIS.ON@
@P.ExtIS.ON@ @R.ExtIS.ON@
[ATT] [NEW]

...
LEXICON VRoot
hamb@P.Basic.ON@ VExt;
hamb@P.ExtEK.ON@ VExt;
hamb@P.ExtEL.ON@ VExt;
hamb@P.ExtIS.ON@ VExt;
hamb@P.ExtISIS.ON@ VExt;
hamb@P.ExtELEL.ON@ VExt;
hamb@P.ExtELAN.ON@ VExt;
hamb@P.ExtISEL.ON@ VExt;
hamb@P.ExtISAN.ON@ VExt;

LEXICON VExt
@R.Basic.ON@: 0@R.Basic.ON@        VerbTerm;
[NEW]@R.Basic.ON@: 0@R.Basic.ON@   VExtNew;
[ATT]@D.Basic@: 0@D.Basic@         VExtAttested;

LEXICON VExtNew
an[RecipExt]: an     VExtNew2;
ek[NeutExt]: ek      VExtNew2;
el[ApplExt]: el      VExtNew2;
is[CausExt]: is      VExtNew2;

LEXICON VExtNew2
! Recursion to cater for unknown order and arbitrary number of extensions
an[RecipExt]: an     VExtNew2;
ek[NeutExt]: ek      VExtNew2;
el[ApplExt]: el      VExtNew2;
is[CausExt]: is      VExtNew2;
                     VerbTerm;

LEXICON VExtAttested
ek[NeutExt]@R.ExtEK.ON@: ek@R.ExtEK.ON@  VerbTerm;
el[ApplExt]@R.ExtEL.ON@: el@R.ExtEL.ON@  VerbTerm;
is[CausExt]@R.ExtIS.ON@: is@R.ExtIS.ON@  VerbTerm;
el[ApplExt]an[RecipExt]@R.ExtELAN.ON@: elan@R.ExtELAN.ON@   VerbTerm;
elel[ComplExt]@R.ExtELEL.ON@: elel@R.ExtELEL.ON@           VerbTerm;
is[CausExt]an[RecipExt]@R.ExtISAN.ON@: isan@R.ExtISAN.ON@   VerbTerm;
is[CausExt]el[ApplExt]@R.ExtISEL.ON@: isel@R.ExtISEL.ON@    VerbTerm;
isis[IntensExt]@R.ExtISIS.ON@: isis@R.ExtISIS.ON@          VerbTerm;

LEXICON VerbTerm
a              #;
```

Appendix B: Number and frequency of extension sequences in the 28 477 lexical entries in ZulMorph



Appendix C: Basic verb roots with the most extension sequences

No. of extension sequences	Verb root
30	*fan (6)*[19]
29	*enz (8)*
28	*bon (7), fund (6)*
27	*az (7), buth (6), buy (6)*
26	*hlab (8), muk (5)*
24	*qal (4), sh*
23	*ling (4)*
22	*emuk (5), photh (5)*
21	*futh (6), phik (4), phind (5)*
20	*nbuk, khub, phath, thath, zal*
19	*bek, dl*
18	*al, bhac, bhek, boph, cob, gan, mangal, qin, shay*
17	*akh, beth, chith, fik, gab, phis, song, thol, vimb*
16	*fun, khohl, khol, phamb, qand, thel, yek, zw*
15	*amuk, band, bind, chach, cim, cin, dlul, eq, f, fic, hlom, lung, phish, qond, vul*
14	*bang, bung, chath, ehl, elam, esab, hlal, hlol, hlum, hol, kham, khaph, khul, mel, ngen, nik, nqum, phons, qed, swel, theng, theth, val*
13	*bhumbuth, cash, chaz, ding, dlal, dluny, ehluk, emul, encik, gudl, hlinz, lov*

Appendix D: Basic verb roots with the longest extension sequences

Extension sequence	Verb root
anisanel	*ling*
aniselan	*ahluk, cin, ehluk, futh, ling*
aniselel	*futh, ling, phamb*
elekelis	*phish*
elelanel	*buth, photh*
elelanis	*bek*
elelisan	*hlab*
elelisel	*hlab*
isanisis	*jojoz*
isekelan	*sh*
isekelis	*sh*
aniselw	*ahluk, cin, ehluk, futh, ling, phamb*
aniswan	*ling*
elekelw	*phish*
eleliswa	*hlab*
elelwan	*buth, photh*
eliselw	*balek*
isekelw	*sh*
iselwan	*dlal, gan*

Appendix E: ZulMorph morpheme tag set (only tags that occur in the examples are provided)

Tag	Description
"Class[20], person and/or number dependent tags"	
BPre	Basic prefix
Dem	Demonstrative pronoun
EC	Enumerative concord
NPrePre	Noun preprefix
OC	Object concord
PC	Possessive concord
SC	Subject concord
RC	Relative concord
"Tags independent of class, person and/or number"	
Adv	Adverb
ApplExt	Applied extension
CausExt	Causative extension
Conj	Conjunction
CopPre	Copulative prefix
EnumStem	Enumerative stem
IntensExt	Intensive extension
LocPre	Locative prefix
LongPres	Long present tense
NegPre	Negative prefix
NeutExt	Neuter extension
NStem	Noun stem
Pot	Potential
RecipExt	Reciprocal extension
VT	Verb terminative
VTNeg	Verb terminative negative
VTPerf	Verb terminative perfect
VRoot	Verb root
ATT	Attested verb extension sequence
LEX	Lexicalisation
NEW	New verb extension sequence