Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

Toward a knowledge-to-text controlled natural language of isiZulu

C. Maria Keet · Langa Khumalo

Received: date / Accepted: date

Abstract The language isiZulu belongs to the Nguni group of languages, which also include isiXhosa, isiNdebele and siSwati. Of the four Nguni languages, isiZulu is the most dominant language in South Africa, which is spoken by 22.7% of the country's 51.8 million population. However, isiZulu (and even more so the other Nguni languages) still remains an under-resourced language for software applications. In this article we focus on controlled natural languages for structured knowledge-to-text viewed from a potential utility for verbalising business rules and OWL ontologies. IsiZulu grammar—and by extension, all Bantu languages—shows that a template-based approach is infeasible. This is due to, mainly, the noun class system, the agglutination and verb conjugation with concords for each noun class. We present verbalisation patterns for existential and universal quantification, taxonomic subsumption, axioms with simple properties, and basic cases of negation. Based on the preliminary user assessment of the patterns, selected ones are refined into algorithms for verbalisation to generate correct isiZulu sentences, which have been evaluated.

Keywords Bantu languages · isiZulu · Controlled Natural Language · OWL

1 Introduction

While South Africa has been celebrated as having the most enabling constitution in the protection and advancement of African languages and has had a stable democracy for two decades, investment in computational linguistics and

C.M. Keet
Department of Computer Science, University of Cape Town, South Africa. Tel.: +27 (0)21 650 2667, Fax: +27 (0)21 650 2007. E-mail: mkeet@cs.uct.ac.za

L. Khumalo
Linguistics Program, School of Arts, University of KwaZulu-Natal, South Africa. Tel.: +27 (0)31260 3589, Fax: +27 (0)31260 3360. E-mail: khumalol@ukzn.ac.za

human language technologies (HLT) has been limited, especially concerning information and knowledge processing (Sharma Grover *et al.*, 2011). While the paucity of HLTs has been noted, the glaring need to develop them has been motivated. For instance, the "National Recordal System" project[1] by the National Indigenous Knowledge Systems Office (NIKSO) of the South African Department of Science and Technology. A first basic version was officially launched in 2013, but it requires a semantics-driven interface that can interact with natural language interfaces and handle multilingualism in, among others, document search and annotation, and in multimodal model development of the knowledge that is to be stored in the NRS (Alberts *et al.*, 2012), in a similar way as is already possible for multiple Indo-European languages (e.g., (Androutsopoulos *et al.*, 2013; Bosca *et al.*, 2014; Jarrar *et al.*, 2006; Franconi *et al.*, 2010; Ghidini *et al.*, 2009; Kaljurand *et al.*, 2014) and other applications in the Semantic Web context (Bouayad-Agha *et al.*, 2014)). Further, the University of KwaZulu-Natal recently introduced a mandatory isiZulu course for all its students and is driving the development of scientific terminology in isiZulu, and also some other South African universities call for the intellectualisation of the indigenous languages (Msila, 2014). Several larger companies, including Facebook, Google, and Microsoft, have made advances in software localisation, including user interfaces of their software in isiZulu, among other Bantu languages, and there is a Google Translate English-isiZulu (that generates mostly incorrect isiZulu sentences) but their technology is proprietary and inaccessible. These and related endeavours require controlled natural languages and natural language generation systems, machine translation, and multilingualism in knowledge representation (e.g., (Fogwill *et al.*, 2011; Alberts *et al.*, 2012; Chavula and Keet, 2014)) to develop semantics-driven end-user and domain expert interfaces, which do not exist yet.

Some results have been obtained in natural language understanding and corpus building for several languages in the Nguni language group (Pretorius and Bosch, 2003, 2009; Spiegler *et al.*, 2010), of which isiZulu is a member. Multilingual systems are being developed elsewhere (among many, (Bosca *et al.*, 2014; Kaljurand *et al.*, 2014)), and there are large EU projects, such as Monnet[2] that concerned foundations of multilingual ontologies, Molto[3] for machine translation, and, e.g., Organic.Lingua[4] as applied project in organic agriculture. On closer inspection, these advances are not deployable as-is with Nguni languages (discussed later in this article). Starting from scratch and defining a grammar alike described in (Kuhn, 2013; Ranta, 2011), is a resource-intensive challenging effort, for the references for linguistic work for isiZulu and Southern Bantu languages are old and outdated (Doke, 1927, 1935), yet still important, and it will take many years to update them to the current isiZulu. Meanwhile, systems have to be built, hence it is prudent to commence with the basics of a

[1] `https://nrs.dst.gov.za/nikmas/`; last accessed: 20-11-2014.

[2] `http://www.monnet-project.eu`; last accessed: June 2014; offline on 24-12-2014.

[3] `http://www.molto-project.eu`; last accessed 24-12-2014.

[4] `http://www.organic-lingua.eu`; last accessed 24-12-2014.

controlled natural language (CNL) for natural language generation (NLG) in such a way that serves linguists, computer scientists, and domain experts to show relevance. Therefore, we approach it in an incremental fashion by taking common formal language constructs (quantification, implication, etc.) of a practical logic language, such as the OWL 2 EL profile (Motik *et al.*, 2009) that is also used for the SNOMED CT medical terminology and the $\mathcal{ALC}$ Description Logic (DL) language (Baader *et al.*, 2008), as a starting point to focus on CNL and verbalisations of logical theories, i.e., scoping this work within knowledge-to-text. OWL and Semantic Web technologies in general are emerging as a syntax of choice for NLG systems thanks to their standardisation and tool infrastructure (Bouayad-Agha *et al.*, 2014). CNLs for OWL ontologies has received ample attention for several years; good English↔OWL systems exist, notably ACE (Fuchs *et al.*, 2010), as well as pretty-printing verbalisation tools, such as SWAT Natural Language Tools (Third *et al.*, 2011). Related are the results for logic-based conceptual models, notably Object-Role Modeling, which are predominantly for monolingual English (Curland and Halpin, 2007), but limitations of the template-based approach have been noted for other languages (Jarrar *et al.*, 2006). Overall, this raises several questions:

1. What are the verbalisation patterns for isiZulu for the basic logic constructs?
2. Can they be realised with a pure template-based approach, mostly template-based with some rules, or will it require a full-fledged grammar engine?
3. What do the answers to question 1 and 2 entail for an implementation of a controlled natural language?

To answer these questions, we devised the high-level patterns and algorithms for the verbalisation of subsumption, negation, existential and universal quantification, and conjugation. The isiZulu grammar rules are complex so that a template-based approach is not feasible for either of the constructs investigated. This is largely due to the semantics of the noun (name of the OWL class) that affects several other components in a sentence—including the quantifiers, the negation, and the verb (name of the OWL object property)—and the highly agglutinative nature of isiZulu. This infeasibility of templates due to a concordance system is briefly illustrated in the following example.

Example 1 Consider a template for English for a simple axiom with quantification, which can be, e.g.,

All [noun1 pl.] [verb 3rd pers. pl.] at least one [noun2]

The words for 'all' and the 'at least one' in isiZulu, however, depend on the noun class of [noun1] (or [noun1 pl.]) and [noun2], respectively. For instance:

bonke oSolwazi bafundisa isifundo esisodwa
'all professors teach at least one course'

compared to

konke ukusebenza kuyawufeza umsebenzi onqunyiwe owodwa
'all operations achieve at least one task'

To break this down in its components: [noun1 pl.] *oSolwazi* is in noun class 3 and *ukusebenza* is in class 15, therewith determining *bonke* for the former

and *konke* for the latter for the verbalisation of the universal quantification. Then, [noun2] *isifundo* is in noun class 7 and *umsebenzi onqunyiwe* in class 3, determining the *esiso-* for the former and *owo-* for the latter for existential quantification. The same issue exists if one were to have chosen 'each' and 'some'. Further, the conjugation of the verb depends on the noun class of the first noun; e.g., when an operation (noun class 15) 'achieves' something, it is *kuyawufeza*, but when a human (noun class 1) 'achieves', it is *uyawufeza*, and a monkey (noun class 9) 'achieves' *iyawufeza*. That is, there is no single word for a 3rd pers. sg. or 3rd pers. pl. as in, among others, English, but it is dependent on which noun—hence, noun class—participates in the verbalisation of the axiom.

A consequence is that existing multilingual models and verbalisation tools cannot be simply transposed and implemented for the Nguni languages. This paper extends results reported in (Keet and Khumalo, 2014a,b) with, mainly, a novel treatment of Bantu verbs, followed by the treatment of object properties in axioms with complex morphology (when verbalised) that resulted in a novel algorithm, which we also validated experimentally.

In the remainder of this paper, some basic nominal and verbal aspects of isiZulu are described in Section 2, which is followed by the verbalisation patterns in Section 3, and the algorithms in Section 4. We discuss in Section 5 and conclude in Section 6.

2 Basics of isiZulu

IsiZulu is the predominantly spoken language in South Africa with 11.5 million speakers in a population of 51.8 million people. It has a healthy body of literature that has been developed over a hundred and seventy years with its first book *Incwadi Yokuqala Yabafundayo* having been first published in 1837. IsiZulu is a Bantu language that belongs to the Nguni sub-group of languages, which are largely spoken in South Africa. Bantu languages archetypically have a morphological typology that is agglutinating, which makes their structure rich and complex. Other agglutinating languages with extremely complex morphology are Turkish, Hungarian, and Finnish (Durrant, 2013). One of the salient features of isiZulu is the system of noun classes. Every noun has a noun class (NC) and it is the noun class that controls the system of concordance of all words in a sentence whose structure is typically subject, verb and object (SVO). Table 1 shows the isiZulu NC prefixes and is based on Meinhoff (1948). Most of the 17 isiZulu NCs are set in pairs such that they have a singular form in one class and a plural form in another and yet others are latent.

Table 1 Zulu noun classes (NC), with an example for each noun class, based on Meinhoff's classification, and the kind of entities typically classified into that noun class.

NC	**Augment**	**Prefix**	**Stem** (example)	**Word** (example)	**Meaning**
1 2	u- a-	m(u)- ba-	-fana -fana	umfana abafana	humans and other animates
1a 2a	u- o-	- -	-baba -baba	ubaba obaba	kinship terms and proper names
3a (2a)	u- o-	- -	-shizi -shizi	ushizi oshizi	nonhuman
3 4	u- i-	m(u)- mi-	-fula -fula	umfula imifula	trees, plants, non-paired body parts
5 6	i- a-	(li)- ma-	-gama -gama	igama amagama	fruits, paired body parts, and natural phenomena
7 8	i- i-	si- zi-	-hlalo -hlalo	isihlalo izihlalo	inanimates and manner/ style
9a (6)	i- a-	- ma-	-rabha -rabha	irabha amarabha	nonhuman
9 10	i(n)- i-	- zi(n)-	-ja -ja	inja izinja	animals
11 (10)	u- i-	(lu)- zi(n)-	-thi -thi	uthi izinthi	inanimates and long thin objects
14 15	u- u-	bu- ku-	-hle -cula	ubuhle ukucula	abstract nouns infinitives
17		ku-			locatives, remote/ general

2.1 The noun class system

The noun in isiZulu consists of two formatives, the prefix and the stem. Prefixes express number and indicate the class to which a particular noun belongs. A prefix can be characterized as full or incomplete. The full prefix has the augment (sometimes called the pre-prefix) followed by a prefix proper while an incomplete prefix only has the augment, respectively, illustrated in Figure 1. The agglutinating nature of isiZulu compels a number of prefixes to be phonologically conditioned and yet others appear as homographs. The concordance is determined by the morphology of the head noun in the subject position, which then influences the agreement pattern, as shown in Example N1 below, where the abbreviations by convention refer to SUBJ = subject, REL = relative, and 2./9. = noun class 2 and 9 respectively.

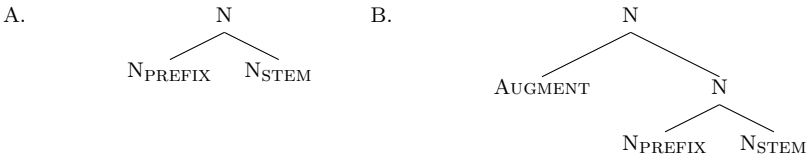

Fig. 1 The structure of isiZulu nouns, where the augment denotes the pre-prefix or initial vowel. A: basic structure for incomplete prefixes; B: for the cases with a prefix proper.

(N1) Abafana abadala bagijimisa inja emnyama
aba-fana **aba**-dala **ba**-gijim-is-a **i-**nja **e**-mnyama
2.-boys **2.**big **2.SUBJ**-chase **9.**-dog **REL-9**.-black
'The big boys are chasing a black dog'

The complex agreement system presents challenges in the development of computational technologies in isiZulu, where the understanding of the basic morphological structure of isiZulu is crucial in the formulation of the technologies.

The NC prefixes of classes 1 and 3 are homographs (*u-m(u)-*) but are crucially conditioned by the morphology of the noun stem to which they attach: *-mu-* before monosyllabic stems and *-m-* for other stems. The *n* of the noun prefixes of classes 9 and 10 coalesces with the following consonant forming prenasalized consonants.

It is imperative to note that, for the most part, semantics of the noun in isiZulu plays a key role in determining which NC the noun falls in within the nominal class system (see column 6 in Table 1). The deeper meaning of the classes as well as the shift and colloquial uses are a subject of deeper investigation (e.g., (Ngcobo, 2010)). Most noun stems belong to only one NC pair (i.e. singular and plural pair), but exceptions exist (e.g., *-ntu*). NC prefixes can also be used to form new noun forms from other noun stems and other stems, like NC:15 that creates infinitives out of verbal stems. The vast majority of the nouns in NC:14 is derived as well: the prefix *-bu-* forms abstract nouns from other noun stems and adjective stems. NC:17 is a non-productive locative class with the noun prefix *ku-*. IsiZulu lacks classes 12 and 13, which are found in other Bantu languages; e.g., Chichewa has them (Bentley and Kulemeka, 2001), which is in Guthrie zone N (unit N31) cf. isiZulu, which is in zone S (unit S42) (Guthrie, 1971). (Guthrie's zones are a referential classification in which the Bantu languages are grouped into fifteen geographical zones labelled with letters and digits signifying a linguistic grouping as well as individual languages (e.g., Zone S). It is one of the two main coding systems for identifying Bantu languages.)

Nominal morphology triggers agreement, as is shown in Example N2:

(N2) Amapoyisa amancane azozihlasela izigebengu eziningi
ama-poyisa **ama**-ncane **a**-zo-**zi**-hlasela **izi**-gebengu e-**zi**-ningi
5.police **5**.small **5.SUBJ**-FUT-**10.OBJ**-attack **10**.robbers REL-**10**.many
'A few police will attack many robbers'

The fact that the subject *amapoyisa* ('police' [-men/-women]) is of NC:5 is reflected both in the agreement prefix on the adjectival *amancane* ('small') and in the subject agreement on the verb. The NC:10 feature of the object *izigebengu* ('robbers') is reflected in the class 10 agreement on the adjective *eziningi* ('many') and in the object concord on the verb. A selection of such agreements is included in Table 2. Although the word order is SVO, variation is evinced to exist in isiZulu since post verbal subject do occur.

Table 2 Zulu noun classes with a selection of concords for the nouns. NC: Noun class; QC: quantitative concord; NEG SC: negative subject concord, PRON: pronominal; RC: relative concord; EC: enumerative concord; oral: oral prefix (see also AU and PRE in Table 1). (Source $QC_{oral+onke}$ list: (Goldsmith and Buthelezi, 2005))

NC	QC (all)		QC (some)			Conjug. and neg.		
	$QC_{oral+onke}$	QC_{nke}	RC	QC_{dwa}	EC	SC	NEG SC	PRON
1	u-onke → wonke	wo-	o-	ye-	mu-	u-	aka-	yena
2	ba-onke → bonke	bo-	aba-	bo-	ba-	ba-	aba-	bona
1a	u-onke → wonke	wo-	o-	ye-	mu-	u-	aka-	yena
2a	ba-onke → bonke	bo-	aba-	bo-	ba-	ba-	aba-	bona
3a	u-onke → wonke	wo-	o-	wo-	mu-	u-	aka-	wona
(2a)	ba-onke → bonke	bo-	aba-	bo-	ba-	ba-	aba-	bona
3	u-onke → wonke	wo-	o-	wo-	mu-	u-	awu-	wona
4	i-onke → yonke	yo-	e-	yo-	mi-	i-	ayi-	yona
5	li-onke → lonke	lo-	eli-	lo-	li-	li-	ali-	lona
6	a-onke → onke	o-	a-	wo-	ma-	a-	awa-	wona
7	si-onke → sonke	so-	esi-	so-	si-	si-	asi-	sona
8	zi-onke → zonke	zo-	ezi	zo-	zi-	zi-	azi-	zona
9a	i-onke → yonke	yo-	e-	yo-	yi-	i-	ayi-	yona
(6)	a-onke → onke	o-	a-	wo-	ma-	a-	awa-	wona
9	i-onke → yonke	yo-	e-	yo-	yi-	i-	ayi-	yona
10	zi-onke → zonke	zo-	ezi-	zo-	zi-	zi-	azi-	zona
11	lu-onke → lonke	lo-	olu-	lo-	lu-	lu-	alu-	lona
(10)	zi-onke → zonke	zo-	ezi-	zo-	zi-	zi-	azi-	zona
14	ba-onke → bonke	bo-	obu-	bo-	bu-	bu-	abu-	bona
15	ku-onke → konke	zo-	oku-	zo-	ku-	ku-	aku-	khona

2.2 Verbs

The verbs in Bantu languages form the most linguistic complex category. They can be conjugated in five different tenses (remote past, recent past, present, immediate future and remote future) and be modified for various aspects and moods. The verb usually agrees with the subject and sometimes also with the object in person and number (as in example N2, above) and in 3rd person for NC as well. To account for these aspects, a verb form can consist of many morphemes. Such complex morphology is characteristic of most Bantu languages, which presents a lot of challenges in the attempts to develop computational technologies in isiZulu.

2.2.1 The Bantu verb morphology

The morphology of the verbal constructions in Bantu is very complex. Wald (Wald, 1987) (p291) observes that the verb shows "[...] the fullest extent of the agglutinative nature of the Bantu language family". The verbal structure consists entirely of bound morphemes. These are the verb root (VR)[5] and a

[5] The following abbreviations are used: A=aspect; ADV=adverb; APPL=applicative; Ext=extension; FV=final vowel; M=mood; NEG=negative tense; OC=object concord; Rad=radical; SG=singular; SC=subject concord; T=tense; VR=verb root; VS=verb stem.

number of affixes. The affixes include subject concord (SC), the object concord (OC), tense, aspect, and mood (TAM), and various derivational extensions. The verb's structure is then terminated by a final vowel (FV). The basic verbal form is summarized as follows (after (Miti, 2006), p299):

(V1) SC - TM - Root - FV

The final vowel (FV) is generally the vowel /-a/. It indicates that the verb radical with which it occurs is used in the indicative mood; e.g.:

(V2) *ndi - cha - end - a* *Chishona: ndichaenda*
1.SC - 1.TM - Root - FV
'I' 'will' go *'I will go'*

The VR in Bantu refers to the base of the verb minus all the concordial and conjugational affixes. Figure 2 shows the complex structure of the verb indicating the verb root, the verb radical and the verb stem.

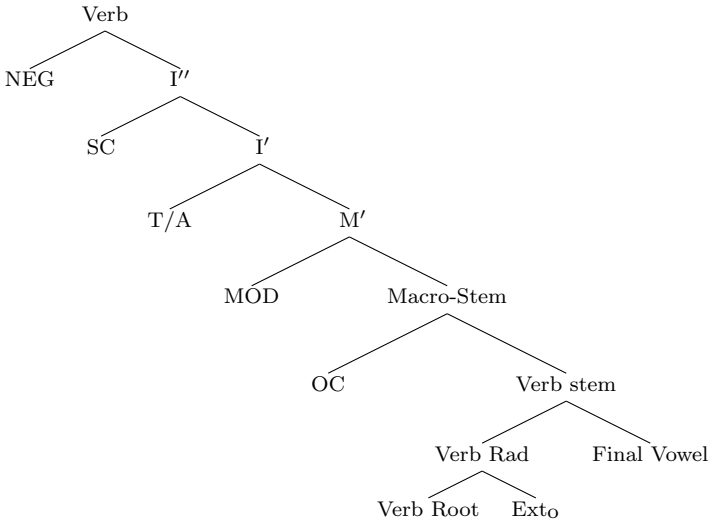

Fig. 2 The structure of a complex verb in Bantu.

2.2.2 The isiZulu verb

Verbal prefixes in isiZulu encode information pertaining to agreement with the subject and object of the verb. Verbal prefixes, like in most Bantu languages, also encode tense/aspect, negation and modality. These prefixes have a somewhat rigid order in which they occur before the VR. This hypothesis has led to the formulation of a Bantu verb slot system resulting in interesting paradigms (Maho, 1999; Mberi, 2002; Khumalo, 2007). Verbal suffixes on the other hand are referred to as extensions which include the causative, applicative, reciprocal, passive, stative, etc. They are viewed as the most interesting

phenomenon in the Bantu verbal complex because of their involvement in the argument structure.

The elements prefixed to the verb stem in isiZulu are usually referred to as clitics. Clitics are independent syntactic elements which appear as part of the host word. This independent element is involved in a morphological merger to appear phonologically as part of a derived word. Example V3 is illustrative:

(V3) *ngi - m - bon - a kusasa isiZulu: Ngimbona kusasa.*
1.SC - 1.OC - Root - FV
'I' 'him' see tomorrow *'I (will) see him tomorrow.'*

The independent elements *ngi-* and *m-* merge to form a derived word *ngimbona*. The clitics are thus syntactic elements, which lack phonological independence. They cannot stand or appear on their own. It is clear that syntactically they are words but phonologically they are not. They are not viewed as phonological words because they fail to satisfy the minimality condition for being a word in Bantu. The Bantu condition is that a word has to minimally consist of two syllables. In the example above, *ngi-* is a single syllable and *m-* is also a single syllable known as "syllabic *m*". The notion of clitics and their grammatical status in Bantu is still a very interesting one.

The isiZulu verbal suffixes are also bound morphemes without any independent status, hence they are also clitics. They are involved in the determination of expressible NP arguments within the sentence. As stated earlier, these include the morphology for encoding the causative, applicative, reciprocal, passive, stative, etc. These suffixes, together with the VR, are terminated by the FV /-a/ and together make up the verb stem (VS) as shown in Figure 2. The following example shows the VR plus the verbal extensions.

(V4) *bon - a isiZulu: bona* 'see' un-extended verb
VR - FV
*bon - **is** - a isiZulu: bonisa* 'make see' extended verb
*bon - **el** - a isiZulu: bonela* 'see for' extended verb
*bon - **an** - a isiZulu: bonana* 'see each other' extended verb

While the suffixes (or verb extensions) clearly introduce a new syntactic element, they however are themselves not independent. They cannot stand as phonological words on their own, hence, they are clitics. The clitics in Bantu can co-occur with the verbal extensions. However, when this happens, they are attached outside the final vowel. The Chishona example in V5 is illustrative.

(V5) *mukomana a-ri-ku-gur-ir-a-zve chisikana*
1.boy 1.SC-T-M-break-APPL-FV-too 7.girl

'The boy is breaking (something) for the girl too'

The order of the extensions and clitics in the example above is worth noting; the clitic *-zve* comes after the FV *-a*. The extensions appear to be more intimately connected to the host VR. The VS is the domain of a number of linguistic processes and it is assumed that the VS has lexical integrity, which makes it an important subdomain in the morphological structure of the verb.

3 Verbalisation patterns

There are two main reasons for the choice for the language constructors that we consider here for the verbalisation patterns that have as scope the generation of grammatically understandable and correct isiZulu sentences. First, prospective application scenarios, and, second, the popular languages in the area of knowledge-to-text with the Description Logics (DL) foundations of most OWL ontology language species. Scientific isiZulu is being developed in nursing (Engelbrecht *et al.*, 2010) and healthcare, which makes it within reach to localise SNOMED CT and related healthcare applications, hence, also its NLG for interaction with isiZulu speakers. The topics of the examples in the following sections are from the general domain, however, so as to use them as explanatory device.

The default logic language one typically starts with in DL is $\mathcal{ALC}$ (Baader *et al.*, 2008), the $\mathcal{A}$ttributive $\mathcal{L}$anguage with $\mathcal{C}$oncept negation, which contains the following elements: *Concepts* denoting entity types/classes/unary predicates/universals, including top $\top$ and bottom $\bot$; *Roles* denoting relationships/associations/binary predicates/properties; *Constructors* 'and' $\sqcap$, 'or' $\sqcup$, and 'not' $\neg$, quantifiers 'for all' $\forall$ and 'exists' $\exists$; *Complex concepts* using constructors: let C and D be concept names, R a role name, then $\neg C$, $C \sqcap D$, and $C \sqcup D$ are concepts, and $\forall R.C$ and $\exists R.C$ are concepts; and *Individuals* (i.e., objects/tokens/named entities can be declared). The meaning is defined by the *semantics* of $\mathcal{ALC}$, using a *domain of interpretation*, and an *interpretation* (as in first order predicate logic with model-theoretic semantics), where Domain Δ is a non-empty set of objects and an interpretation $\cdot^{\mathcal{I}}$ is the *interpretation function*, domain $\Delta^{\mathcal{I}}$. Then, $\cdot^{\mathcal{I}}$ maps every concept name A to a subset $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$; $\cdot^{\mathcal{I}}$ maps every role name R to a subset $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$; $\cdot^{\mathcal{I}}$ maps every individual name a to elements of $\Delta^{\mathcal{I}}$: $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$ (Note: $\top^{\mathcal{I}} = \Delta^{\mathcal{I}}$ and $\bot^{\mathcal{I}} = \emptyset$) Take C and D to denote concepts, R a role, and a and b are individuals, then they have the following meaning, with on the left-hand side of the "=" the syntax of $\mathcal{ALC}$ under an interpretation and on the right-hand side its semantics: $(\neg C)^{\mathcal{I}} = \Delta^{\mathcal{I}} \backslash C^{\mathcal{I}}$; $(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$; $(C \sqcup D)^{\mathcal{I}} = C^{\mathcal{I}} \cup D^{\mathcal{I}}$; $(\forall R.C)^{\mathcal{I}} = \{x \mid \forall y.R^{\mathcal{I}}(x,y) \rightarrow C^{\mathcal{I}}(y)\}$; $(\exists R.C)^{\mathcal{I}} = \{x \mid \exists y.R^{\mathcal{I}}(x,y) \wedge C^{\mathcal{I}}(y)\}$. One can also specify the notion of *satisfaction*, but this is not relevant here and therefore omitted.

We focus here on universal and existential quantification ($\forall$ and $\exists$, respectively), subsumption ($\sqsubseteq$), negation ($\neg$), and object properties (DL roles) whose verbs require conjugation; conjunction ($\sqcap$) has been dealt with in (Keet and Khumalo, 2014a) and disjunction ($\sqcup$) is straightforward and therefore omitted here. We will use the DL notation for conciseness and, where applicable, assume a suitable multilingual encoding has been implemented.

The principal grammatical features that affect verbalisation patterns in isiZulu for the cases we consider are the NC of the name of the OWL class, whether the OWL class is an atomic class or a class expression, the quantifier used in the axiom, and the position of the OWL class in the axiom.

3.1 Quantifiers

3.1.1 Universal Quantification

To keep these first steps feasible, only universal quantification at the start of the concept inclusion axiom is considered. Universal quantification is widely used for verbalising taxonomic subsumption for atomic classes and in the typical 'forall-some' construction (in linguistic terms: the nominal head). The essence of 'all' or 'each' is captured with *-onke* in isiZulu. This is then prefixed with the oral prefix (see also AU and PRE in Table 1) of the NC of that first noun—i.e., a named OWL class/DL concept on the left-hand side of $\sqsubseteq$ in the ontology—and is then modified based on what the prefix was; e.g.:

(U1) `unkosikazi` $\sqsubseteq$ ...	(`wife` $\sqsubseteq$ ...)
wonke unkosikazi ...	('each wife...'; *u-* + *-onke*)
bonke onkosikazi ...	('all wives...'; *ba-* + *-onke*)
(U2) `ucingo` $\sqsubseteq$ ...	(`telephone` $\sqsubseteq$ ...)
lonke ucingo ...	('each telephone...'; *lu-* + *-onke*)
zonke izincingo ...	('all telephones...'; *zi-* + *-onke*)

The oral prefixes are stable for each NC, so then instead of designing an algorithm that first determines the NC, then looks up the oral prefix, and then have a case statement for each, one can pre-compute the complete list of nominal heads and carry out a simple look-up of the word when generating the verbalisation. The composition and list of nominal heads are included in column 2 in Table 2. This also holds for the second option to generate the verbalisation: take *-nke*, and prefix it with the quantitative concord (QC) (see also Table 2). In effect, the real choice that has to be made is between singular or plural, which will be dealt with below. Thus, the patterns, with N=noun taken from the name of the OWL class, are:

a. <QC(all) for NC_1>onke <N_1>;
b. <QC(all) for NC_x>onke <plural of N_1, being of NC_x>.

Either may be preferred, depending on the 'form' of the axiom, like whether it is used in verbalising plain taxonomic subsumption or describing properties of a DL concept (Keet and Khumalo, 2014b).

3.1.2 Existential Quantification

At this stage, we consider only a basic use of existential quantification as in OWL 2 EL, for which there are several verbalisation options:

(E1) `indlovu` $\sqsubseteq \exists$`idla.ihlamvana`	(`elephant` $\sqsubseteq \exists$`eats.twig`)
izindlovu zindla izihlamvana	('elephants eat twigs')
yonke indlovu idla ihlamvana elilodwa	('each elephant eats at least one twig')
zonke izindlovu zidla ihlamvana elilodwa	('all elephants eat at least one twig')
yonke indlovu idla noma yiliphi ihlamvana	('each elephant eats some twig')
zonke izindlovu zidla noma yiliphi ihlamvana	('all elephants eat some twig')
yonke indlovu idla ihlamvanathize	('each elephant eats some twig')

As in English, one can choose between 'at least' and 'some', and again singular vs. plural and with/without verbalising the universal quantification (verbs are dealt with in the next section). The 'at least one' is constructed as follows: the relative concord (RC), determined by the noun class system, is attached to the quantitative concord (QC) and suffixed with the quantitative suffix *-dwa*. This is illustrated in the first three rows in Table 3, and the complete lookup lists for the RC and QC for each NC are included in Table 2.

Table 3 Examples showing the composition of the main component of the two principal patterns for existential quantification.

noun	NC	RC	QC	QSuffix	copulative	EC	ESuffix
ushizi ('cheese')	class 3a	*o-*	*-wo-*	*-dwa*			
ihlamvana ('twig')	class 5	*eli-*	*-lo-*	*-dwa*			
isihlalo ('chair')	class 7	*esi-*	*-so-*	*-dwa*			
ushizi ('cheese')	class 3a				*ngu-*	*-mu-*	*-phi*
ihlamvana ('twig')	class 5				*yi-*	*-li-*	*-phi*
isihlalo ('chair')	class 7				*yi-*	*-si-*	*-phi*

The 'some' is constructed as follows: copulative + enumerative concord (EC) + enumerative suffix *-phi*, and the conjunction *noma* collocates with the enumerative to complete the meaning 'some among many'; see also the last three rows in Table 3. The concord EC is fixed for each NC (see Table 2). The copulative is different from the normal way (see below)—an *-i* or *-u* is added—because the copulative cannot be followed by a consonant that the EC begins with. The clitic *-thize*, and variant form *-thile*, attaches to the noun, which is often the object of the sentence, to express the sense that it is some among many of those objects; *inhlamvanathize* from Example E1 then means 'any one of the twigs'. This is stretching the notion of existential quantification, but it is the only candidate for being an easy template rather than the need for encoding a set of rules. In sum, we obtain the following three core patterns:

a. <QC(all) for NC_x>onke <pl. N_1, is in NC_x> <conjugated verb> <N_2 of NC_y> <RC for NC_y><QC for NC_y>dwa;
b. <QC(all) for NC_x>onke <pl. N_1, is in NC_x> <conjugated verb> noma <copulative *ng/y* adjusted to first letter of N_2><EP of NC_y>phi <N_2>.
c. <QC(all) for NC_x>onke <N_1 in NC_x> <conjugated verb> <N_2>thize;

The first option is less cumbersome to encode, because it does not have to take into account the variation in the copulative (*yi* and *ng*; see subsumption in Section 3.2). Fortunately, the respondents to the survey had an overwhelming preference for this option (Keet and Khumalo, 2014b).

3.2 Subsumption

One can divide the nouns by living vs. non-living things and purely by syntactic means, but for the verbalisations, the main questions are whether singular or

plural would be preferred (S1), whether the universal quantification should be included, and whether one should use the generic or determinate (S2 vs. S3).

(S1) `Herb` $\sqsubseteq$ `Plant`
ihebhu ngumuthi ('herb is a plant')
amahebhu yimithi ('herbs are plants')
wonke amahebhu ngumuthi ('all herbs are a plant')
(S2) `Elephant` $\sqsubseteq$ `Animal`
indlovu yisilwane ('elephant is a animal'; generic)
(S3) `Cellphone` $\sqsubseteq$ `Phone`
umakhalekhukhwini uyifoni ('cellphone is a phone'; determinate)

The syntactic approach entails that to obtain the right copulative, one has to look up the first letter of the noun of the superclass, resulting in *ng* for nouns starting with a-, o-, or u-, else it is *y* (Turner, 19xx). Further, the generic is preferred over the determinate for neutrality of the verbalisation, resulting in the following possible patterns for subsumption:

a. N_1 <copulative ng/y depending on first letter of N_2>N_2.
b. <plural of N_1> <copulative ng/y depending on first letter of plural of N_2><plural of N_2>.
c. <QC(all) for NC_x>onke <plural of N_1, being of NC_x> <copulative ng/y depending on first letter of N_2>N_2.

The above holds for when the subsumption is not followed by negation. If it followed by negation, then the verbalisation for subsumption and negation are combined into one term and the auxiliary verb omitted:

(SN1) `Cup` $\sqsubseteq \neg$`Glass`
indebe akuyona ingilazi ('cup not a glass')
zonke izindebe aziyona ingilazi ('all cups not a glass')

That is, the negative subject concord (NEG SC) of the NC of the first noun (*aku-*) is combined with the pronomial (PRON) of the NC of second noun (*-yona*), where each NC has its version (see Table 2). More precisely as verbalisation patterns, we obtain:

a. <N_1 of NC_x> <NEG SC of NC_x><PRON of NC_y> <N_2 of NC_y>.
b. <QC(all) for NC_x>onke <plural N_1, being of NC_x> <NEG SC of NC_x> <PRON of NC_y> <N_2 with NC_y>.

Negation with object properties is addressed in Section 3.3.2, whereas subsumption in the general case, such as in an axiom like $\forall R.C \sqsubseteq \exists S.(D \sqcap E)$, is left for future work.

3.3 Object properties in axioms

Verbalisation of object properties in English and several other languages relies on the common practice in model development—be they ontologies, structured vocabularies, conceptual models, or business rules—of naming the object property with the verb in 3rd person singular, such as eats and teaches or taught

by, which remains the same regardless of the noun that comes before or after it. Conjugation of verbs in isiZulu, and in all Bantu languages, is rather more complicated, as alluded to in Section 2.2. We delimit the case first to present tense active voice, and defer the reverse verbalisation direction to future work.

3.3.1 Plain object properties and present tense

The basic, but substantial, fact is that the conjugation of the verb depends first on the subject, which is the name of the class on the left from the object property in the axiom. For instance, for monkeys, which are in NC:9/10, the verb is conjugated as in (OP1), whereas the same 'eating' by humans takes a conjugation for NC:1/2 (OP2).

(OP1) `Monkey` $\sqsubseteq \exists$`eats.Fruit`
yonke inkawu idla isithelo esisodwa ('each monkey eats at least one fruit')
zonke inkawu zidla isithelo esisodwa ('all monkeys eat at least one fruit')
(OP2) `Human` $\sqsubseteq \exists$`eats.Fruit`
wonke umuntu udla isithelo esisodwa ('each human eats at least one fruit')
bonke abantu badla isithelo esisodwa ('all humans eat at least one fruit')

That is, the verb stem (*-dla* in the example) is prefixed with the subject concord (SC) of the subject's NC on the noun (noun of the named OWL class). The subject concords are fixed, and included in Table 2. This makes for a straightforward pattern with respect to conjugation:

a. <QC(all) for NC_x>onke <pl. N_1, is in NC_x> <SC of NC_x><verb stem>a <N_2 of NC_y> <RC for NC_y><QC for NC_y>dwa;

Realising this pattern is non-trivial, however, aside from an algorithm with look-up table, which will be touched upon in Section 4.

3.3.2 Negation with object properties

We have seen negation for disjointness with a subsumption axiom, but not yet in conjunction with object properties. Here, also a modelling aspect comes to the fore from ontology development (Rector *et al.*, 2004), illustrated in (OP3) and (OP4):

(OP3) `Leopard` $\sqsubseteq \exists$`eats.`$\neg$`Fruit`
meaning: 'each leopard eats something that is not a fruit' (which can be *anything* that is not fruit; e.g., marshmallows, rabbits, pencils, ...)
(OP4) `Leopard` $\sqsubseteq \neg\exists$`eats.Fruit`
meaning: 'each leopard does not eat something that is a fruit'

In this case, we are interested in the second case, of the type (OP4). Whereas in English one uses the 'does not' (alike the 'is not' from the previous section on disjointness), in isiZulu the verb is modified. There are different ways to arrive at a pattern. First, there are cases where the deconstruction of "a + SC^- + VS + i" works, as in *ngithanda* ('I like to') and *angithandi* ('I do not like to'). Second, we can reuse the NEG SC from before to explicitly negate

the verb (still with the final vowel *-i* as the negative suffix), where the NEG SC is determined by the subject in the noun phrase (the noun of the OWL class that comes before the negation). For instance, in (OP5), we have *aka + dl + i* for the singular and *aba + dl + i* for the plural, with *ugogo* 'grandmother' (NC:1/2), and a different one for *ingwe* ('leopard') (NC:9/10) in (OP6):

(OP5) `Grandmother` $\sqsubseteq \neg\exists$`eats.Apple`
wonke ugogo akadli iapula elilodwa
('each grandmother does not eat some apple')
bonke ogogo abadli iapula elilodwa
('all grandmothers do not eat some apple')

(OP6) `Leopard` $\sqsubseteq \neg\exists$`eats.Apple`
yonke ingwe ayidli iapula elilodwa ('each leopard does not eat some apple')
zonke izingwe azidli iapula elilodwa ('all leopards do not eat some apple')

Thus, also here a template is not feasible, but a rule pattern readily emerges, which we choose with plural:

a. <QC(all) for NC_x>onke <pl. N_1, is in NC_x> <NEG SC of NC_x><verb stem>i <N_2 of NC_y> <RC for NC_y><QC for NC_y>dwa.

Unlike verbalising an axiom with an object property without negation, this certainly does not justify adding a new object property to the vocabulary of the theory (i.e., will not obtain a new URI): while the conjugation does generate a new word in a lexicon with respect to the base case of the verb, logically, one should not squeeze a negation in the name of an object property.

4 Algorithms for the verbalisation patterns

We first describe some considerations of the design of the algorithms together with the algorithms, then illustrate their workings stepping through them, and finally present the exploratory user evaluation.

4.1 Algorithm design

Several constructs discussed in the previous section have more than one option to verbalise it. We have conducted a brief survey in an attempt to uncover the preferences of both linguist and non-linguists isiZulu first-language speakers, and its outcome has been reported elsewhere (Keet and Khumalo, 2014b). Here we focus only on those preferences and our design decisions to devise the algorithms to generate the verbalisation. The reason to go into this detail, is that the patterns hide some complexity (e.g., when it says 'depending on...') and it is also about *how* to obtain the verbalisations rather than on the *what*-emphasis of the patterns.

The algorithm for simple taxonomic subsumption in the sense of named classes in the TBox of an OWL ontology is included as Algorithm 1. It is evident that this is more elaborate than the 'is a' in English verbalisation

Algorithm 1 Verbalisation of simple taxonomic subsumption

1: $\mathcal{C}$ set of classes, language $\mathcal{L}$ with $\sqsubseteq$ for subsumption and $\neg$ for negation; variables: A axiom, NC_i nounclass, $c_1, c_2 \in \mathcal{C}$, a_1 term, a_2 letter; functions: $getFirstClass(A)$, $getSecondClass(A)$, $getNC(C)$, $checkNegation(A)$, $getFirstChar(C)$.
Require: axiom A with a $\sqsubseteq$ has been retrieved **and** named classes on the lhs and rhs
2: $c_1 \leftarrow getFirstClass(A)$ {get subclass}
3: $c_2 \leftarrow getSecondClass(A)$ {get superclass}
4: $NC_1 \leftarrow getNC(c_1)$ {determine noun class by augment and prefix or dictionary}
5: $NC_2 \leftarrow getNC(c_2)$ {determine noun class by augment and prefix or dictionary}
6: **if** $checkNegation(A) == true$ **then**
7: {use *Algorithm 3*}
8: **else**
9: $a_2 \leftarrow getFirstChar(c_2)$ {retrieve first letter of c_2}
10: **select case**
11: a_2 = 'i' **then**
12: RESULT ← 'c_1 yc_2.' {verbalise as taxonomic subsumption with y}
13: a_2 = {'a', 'o', 'u'} **then**
14: RESULT ← 'c_1 ngc_2.' {verbalise as taxonomic subsumption with ng}
15: $a_2 \notin$ {'a', 'i', 'o', 'u',} **then**
16: RESULT ← '*this is not a well-formed isiZulu noun*.'
17: **end select case**
18: **end if**
19: **return** RESULT

templates, which is due to the two variants, *y* and *ng*, for the copulative. Algorithm 1 is essentially the same as introduced in (Keet and Khumalo, 2014a), but the negation component has been moved to a separate algorithm to eliminate duplication of the negation aspect and our first target language for implementation, OWL 2 EL (Motik *et al.*, 2009), does not have negation.

Algorithm 2 presents simple existential quantification, using the *-dwa* option. This is an extended version cf. the one presented in (Keet and Khumalo, 2014a) in that its "*AlgoConjugate*" referral to the need to have that algorithm has been replaced with the procedure to do that conjugation (lines 16-22). It does abstract away from the intricacies of representing a 'zulufied' object property in an ontology and the underlying logics of matching the multiple strings to one vocabulary element in the ontology. The latter could possibly be handled by a rule-based extension of the morphology module of the *lemon* model (McCrae *et al.*, 2012) or of ontolex-lemon[6], as indicated by (Chavula and Keet, 2014), or through a more rigorous treatment of relations by linking up the positionalist with the standard view representation of relationships (Leo, 2008) in the context of OWL ontologies so as to obtain a well-founded separation of the relationship from the lexicon. We leave that implementation component to future work. The second change in the algorithm cf. the one presented in (Keet and Khumalo, 2014a) is that the $pluralizeNoun()$ operation has been changed into a referral to a separate algorithm, *AlgoPluralize*, to pluralise a noun that is yet to be developed. The reason for this is that while indeed there are stable prefixes for plurals (recall Table 1), there are ex-

[6] `http://www.w3.org/community/ontolex/wiki/Final_Model_Specification`; last accessed: 24-12-2014.

Algorithm 2 Determine the verbalisation of existential quantification with object property (basic version, with conjugation)

1: $\mathcal{C}$ set of classes, language $\mathcal{L}$ with $\sqsubseteq$ for subsumption and $\exists$ for existential quantification; variables: A axiom, NC_i noun class, $c_1, c_2 \in \mathcal{C}$, $o \in \mathcal{R}$, a_1 a term; r_2, q_2 concords; functions: $getFirstClass(A)$, $getSecondClass(A)$, $getNC(C)$, $getRC(NC_i)$, $getQC(NC_i)$, $getVSofOP(o)$.
Require: axiom A with a $\sqsubseteq$ has been retrieved **and** an $\exists$ on the rhs of the inclusion
2: $c_1 \leftarrow getFirstClass(A)$ {get subclass}
3: $c_2 \leftarrow getSecondClass(A)$ {get superclass}
4: $o \leftarrow getObjProp(A)$ {get object property}
5: $v \leftarrow getVSofOP(o)$ {get verb stem of object property}
6: $NC_1 \leftarrow getNC(c_1)$ {determine noun class by augment and prefix or dictionary}
7: $NC_2 \leftarrow getNC(c_2)$ {determine noun class by augment and prefix or dictionary}
8: NC_1' ← lookup plural nounclass of NC_1 {from known list}
9: $c_1' \leftarrow AlgoPluralize(c_1, NC_1')$ {call algorithm *AlgoPluralize* to generate a plural from o}
10: a_1 ← lookup quantitative concord for NC_1' {from quantitative concord (QC(all)) list}
11: $r_2 \leftarrow getRC(NC_2)$ {get relative concord for c_2 from the QC_{dwa}-list}
12: $q_2 \leftarrow getQC(NC_2)$ {get quantitative concord for c_2 from the QC_{dwa}-list}
13: **if** $checkNegation(A) == true$ **then**
14: {use *Algorithm 3*}
15: **else**
16: **if** o annotated with present tense **then**
17: $conj_{nc1}$ ← lookup SC of NC_1' {from known SC list}
18: $o' \leftarrow conj_{nc1} v$ {generate conjugated verb}
19: RESULT ← 'a_1 c_1' o'a c_2 $r_2 q_2$dwa.' {verbalise the axiom}
20: **else**
21: RESULT ← '*passive voice and inverses are not supported yet*.'
22: **end if**
23: **end if**
24: **return** RESULT

ceptions as some plural prefixes are phonologically conditioned, as mentioned in Section 2.1. For instance, *-cindezi* does take *in-/izin-* as prefix, but the 'c' of the stem conditions it to have a *-g-* between the prefix and stem, making *ingcindezi/izingcindezi*, whereas other conditions result in *izim-* or *iziny-* (changes underlined). There are several such cases and not all sources seem to agree on the conditions—e.g., *izincingo* in Example U2 by one of the authors (LK) versus *izingcingo* in the isiZulu-English dictionary (Dent and Nyembezi, 2009)—therefore, systematising this for computation is left for future work.

The third algorithm (Algorithm 3) combines the verbalisation patterns of the two cases of negation described in Section 3, i.e., with subsumption and negation in front of an object property, for there are recurring components, and OWL languages differ on their inclusion. The algorithm is called from within Algorithm 1 and 2, which means the basic data about the participating classes and object property are already known and thus do not have to be repeated. Note that also here we have the pluralisation step that has to use an *AlgoPluralize* algorithm. This algorithm possibly can be optimised further once more insight is obtained for more complex cases, but in any case, none of the steps is computationally hard.

Algorithm 3 Verbalisation of negation in an axiom (base cases: taxonomic subsumption and object property)

1: $\mathcal{C}$ set of classes, language $\mathcal{L}$ with $\sqsubseteq$ for subsumption and $\neg$ for negation; variables: A axiom, NC_i noun class, $c_1, c_2 \in \mathcal{C}$, a_1 term, a_2 letter and n, p are concords, v verb stem; functions: $checkNegation(A)$, $getNSC(NC_i)$, $getPNC(NC_i)$.
Require: $checkNegation(A) == true$
2: **select case**
3: negation directly preceded by $\sqsubseteq$ **and** directly followed by c_2 **then**
4: $NC'_1 \leftarrow$ lookup plural nounclass of NC_1 {from known list}
5: $c'_1 \leftarrow AlgoPluralize(c_1, NC'_1)$ {call algorithm *AlgoPluralize* to generate a plural from o}
6: $a_1 \leftarrow$ lookup quantitative concord for NC'_1 {from quantitative concord (QC(all)) list}
7: $n \leftarrow getNSC(NC'_1)$ {get negative subject concord for c'_1}
8: $p \leftarrow getPNC(NC_2)$ {get pronomial for c_2}
9: RESULT $\leftarrow$ 'a_1 c'_1 np c_2.' {verbalise the disjointness (a_1 is QC(all))}
10: negation in front of OP **then**
11: $n \leftarrow getNSC(NC'_1)$ {get negative subject concord for c'_1}
12: RESULT $\leftarrow$ 'a_1 c'_1 nvi c_2 $r_2q_2$dwa.' {verbalise the axiom}
13: negation in front of c_2 **and** A contains an OP **then**
14: RESULT $\leftarrow$ '*verbalisation of this class negation is not supported yet*.'
15: **end select case**
16: **return** RESULT

Overall, it can be seen that isiZulu is quite amenable to computation, and there are no exceptions to the rules alike the numerous case for, e.g., German or alike the irregular verbs in Italian and Spanish. Notwithstanding, the verbalisations are not trivially generated with a template-based approach.

Concerning the computational complexity of the controlled natural language and algorithms, we observe the following. While we have yet to convert it into a formal grammar and investigate rigorously, the algorithms indicate merely several list lookup iterations, which can be done in linear time. Any verbalisation software for isiZulu—and, in fact, any Bantu language—surely will take more time to compute the verbalisation than the template-based approach, but algorithmically, it still looks promising to compute it on-the-fly.

4.2 Applying the algorithms

To illustrate the working of the algorithm, we take two axioms covering different aspects.

(ExZu1)
input: `usolwazi` $\sqsubseteq \exists$`ufundisa.isifundo`
(`professor` $\sqsubseteq \exists$`teaches.course`)
output: Bonke osolwazi bafundisa isifundo esisodwa
('all professors teach at least one course')

To generate the isiZulu verbalisation, the following will happen in essentially four steps: fetching the data, sorting out the first noun and its quantification, then the second noun with its quantification, and lastly the conjugation of the verb. Given that there is an $\sqsubseteq$ and a $\exists$ on its right-hand-side, it will enter

Algorithm 2. After obtaining the vocabulary of the axiom, it will lookup the NC of the first noun, usolwazi (line 6), and of the second one, isifundo (line 7). The NC of the first noun is 1a, so that it can be pluralised (NC:2a) into osolwazi (lines 8-9), and matched with the concord for NC:2a, being bonke (line 10). The quantification on the second noun is done by lines 11-12. Then the conjugation for the first noun: looking up the subject concord for osolwazi's NC (line 17), ba-, which is then attached to the verb stem -fund (line 18), and the overall result put together, obtaining Bonke osolwazi bafundisa isifundo esisodwa, which is returned (line 19).

The second example uses Algorithms 1 and 3.

(ExZu2)
input: `itafula` $\sqsubseteq \neg$`isihlalo` (`table` $\sqsubseteq \neg$`chair`)
output: onke amatafula awasona isihlalo ('All tables are not a chair')

Also here first the basic elements are retrieved (lines 2-3) and their respective NCs looked up (lines 4-5), being NC:5 for itafula and NC:7 for isihlalo. It then checks for negation in the axiom (line 6), which is true, so that it continues with Algorithm 3. The first case in Algorithm 3 holds, so it pluralises itafula to amatafula (line 5, NC:6) and finds the correct for-all (line 6), being onke, and negative subject concord (line 7), being awa-. Taking the NC of isihlalo (NC:7), we obtain the pronominal (line 8), being -sona, and finally string the sentence together (line 9), obtaining onke amatafula awasona isihlalo, which is returned to Algorithm 1 (line 16 in Algorithm 3), which completes the if-else and it also returns that sentence (line 19), completing the verbalisation.

4.3 Exploratory evaluation on the correctness of the algorithms

The algorithms are designed such that they produce grammatically correct sentences, provided the input adheres to what it claims to support. The questions that arises, is: *do the current algorithms indeed generate grammatically correct sentences?* To answer this question, rather than taking a highly qualitative approach with in-depth interviews with 2-3 linguists, we chose to draw from a pool of isiZulu speakers so that the results may also feed into linguistics research.

Survey set-up. The exploratory survey has the following set-up.

1. Extract a set of axioms from extant ontologies in the general knowledge domain (business, pets, etc.), 10 for each algorithm, making sure that in the examples there are nouns of different noun classes;
2. Translate the terms (class names and object property names) into isiZulu;
3. Go through the algorithm to generate their respective CNL sentences;
4. Present the sentences in an online survey (open for a week) where participants can choose between (a) 'grammatical + acceptable', (b) 'grammatical + ambiguous', (c) 'ungrammatical + understandable', or (d) 'ungrammatical + unacceptable'. Participants (students, employees, and linguists) are

recruited from UKZN, who may or may not be an isiZulu linguist (to be indicated on the survey);
5. Evaluate responses.

One author created the list and the other translated, and we went through the algorithms together. In creating the list for simple existential, two axioms were included that should fail on Algorithm 2, being object properties of the pattern 'hasX' whereas the algorithm accepts plain verbs only. For instance, Human $\sqsubseteq \exists$hasPart.Heart would be *bonke abantu banenhliziyo eyodwa* where the underlined word represents the hasPart.Heart-part of the axiom, and Pizza $\sqsubseteq \exists$hasTopping.Cheese would be *wonke amaPhiza anesinongo soshizi esisodwa*, with the underlined part capturing the 'has topping'; hence, this indeed requires further investigation, and they were replaced with other axioms.

In the simple taxonomic subsumption, there were two cases of preference for phonological conditioning, but they were forced into adhering to the algorithm. They were Soybean $\sqsubseteq$ Bean with preference *Ubhontshisi isoya wubhontshisi* and algorithmically *Ubhontshisi isoya ngubhontshisi* was generated and Arm $\sqsubseteq$ Limb as *Ingalo lilungu lomzimba* vs *Ingalo yilungu lomzimba*.

The survey was created in a local installation of Limesurvey using the isiZulu localisation; i.e., not only the questions but also the canned text with instructions, named buttons, and error messages are in isiZulu. It is accessible at `http://limesurvey.cs.ukzn.ac.za/index.php?sid=79641&lang=zu`.

Results and discussion Thirty-two people were invited to participate in the survey. Among them were 16 academics with considerable years in isiZulu language research and language practice, 12 post graduate students, and 4 non-linguists such as administrators. This survey resulted in 11 completed responses, 7 of whom identified themselves as linguists and only 4 as non-linguists, and 9 incomplete responses. As such, this has to be categorised as an exploratory evaluation of the algorithms. Overall, the aggregates are somewhat encouraging toward correctness rather than only understandability, as shown in Table 4. Responses to some of the individual sentences reveal some challenging aspects, which will be described in the remainder of this paragraph.

Questions 1-10 are simple taxonomic subsumption. It is instructive that grammatical and acceptable (option (a)) is an overwhelming choice followed closely by grammatical but ambiguous (option (b)). This is because of the simple structure of these constructions, which is devoid of any complex agreement morphology. There are, however, exceptions. Manager $\sqsubseteq$ Employee, verbalised as *UMphathi ngumsebenzi*, presents a complex response. Whilst most of the respondents (30%) answered with option (a), 20% of the respondents chose option (b), 10% and 5% respectively selected ungrammatical understandable (option (c)) and unacceptable (option (d)), respectively (the remainder did not answer). This may be because in the Zulu culture the perception of a manager is that of a person who does not work but directs others to work. Following that logic *uMphathi* cannot be a worker, and thus while the sentence may be syntactically correct, it would not be semantically. Another construction that presented a complex response was *Ubhontshini isoya ngubhontshisi*

Table 4 Mean, median, and standard deviation (rounded) aggregated by type of verbalisation pattern, based on percentages of the answers given (omitting 'no answer'). 'gr+a': grammatical and acceptable, 'gr+amb': grammatical and ambiguous, 'ugr+u': ungrammatical yet understandable/acceptable, 'ugr+unacc': ungrammatical and unacceptable.

Type	Answer	Mean	Median	St. dev.
Simple subsumption	gr+a	74	79	18
	gr+amb	15	17	12
	ugr+u	3	0	4
	ugr+unacc	8	8	8
Simple disjointness	gr+a	44	42	25
	gr+amb	38	38	19
	ugr+u	3	0	6
	ugr+unacc	15	8	16
Existential quantification	gr+a	74	77	17
	gr+amb	19	14	18
	ugr+u	5	0	7
	ugr+unacc	2	0	4
Total	gr+a	63	70	25
	gr+amb	24	20	19
	ugr+u	4	0	5
	ugr+unacc	9	8	12

($\mathsf{Soybean} \sqsubseteq \mathsf{Bean}$), though not for reasons mentioned above. A total of 8 respondents thought it was grammatical (option (a) or (b)), one selected option (c), and 3 chose option (d). This is due to *Ubhontshi isoya* being a compound word: both lexemes that make up the compound have prefixes, which presents a morphological problem. A way around this is to make it one word *Ubhontshisoya* or translate it as *isoyibhini* or otherwise *ubhontshisi oyisoya*.

Once the sentence takes a complex structure, agreement in isiZulu becomes difficult to follow, hence, there were certain mixed reactions to the simple negation constructions (numbers 11-20). Otherwise there was a general pattern, which indicates that respondents found all the sentences grammatical with others thinking they are acceptable and others thinking they are ambiguous. While not all respondents who answered agree, as with the previous survey (Keet and Khumalo, 2014b), the majority of answers indicated that the automatically generated sentences were grammatical: in 21 of the 30 sentences option (a) was selected $\geq$50% and in 3 sentences option (b) was selected $\geq$50% of the responses; only 3 sentences received an option (d) verdict in $\geq$30% (but still $\leq$42%) of the responses; option (c) never exceeded 18%.

The difficulty with one construction was predictable linguistically: *zonke inyama akuzona imifino* ($\mathsf{Meat} \sqsubseteq \neg\mathsf{Vegetable}$). This is because in isiZulu it is not clear whether *inyama* can take a singular and plural and it belongs to class 9 as a singular (with a plural in class 10 *izinyama*) or in class 4 in the plural form (without a singular), making *yonke inyama akuyona imifino*. It is also likely that the difficulty was caused by the regrettable typo of mixing both (*zonke inyama*) in the survey. This is why this particular construction has 25% option (d) unacceptable, 25% option (b), and only 10% grammatical and acceptable, with the remainder (40%) unanswered.

Simple existential constructions were generally acceptable with a few exceptions pertaining to ambiguity. Comments by respondents mention that some of the constructions are not straight-forward (ambiguous), hence the meaning is difficult to discern. It is also important to state that isiZulu has dialects, depending on which region in which it is spoken, hence some divergences in sentence judgements are a result of dialect variation. We hypothesise that the difference between grammatical and acceptable and grammatical and ambiguous is informed by these salient differences in the dialects of isiZulu. This issue also came afore in the survey reported in (Keet and Khumalo, 2014b), and deserves further investigation by linguists.

5 Discussion

We answer the research questions first and subsequently discuss several other aspects that can be taken into consideration for an isiZulu controlled natural language and prospective applications.

5.1 Answering the research questions

In the introduction of this paper, we posed three research questions we aimed to answer. The first one, what the verbalisation patterns for isiZulu for the basic logic constructs are, has been addressed as follows. The DL $\mathcal{ALC}$ was selected as first target language, bearing in mind the computationally better behaved OWL 2 EL, and verbalisation patterns have been presented in Section 3, covering possible patterns for universal and existential quantification ($\forall$ and $\exists$, respectively), subsumption ($\sqsubseteq$), negation ($\neg$), and object properties (DL roles), whereas conjunction ($\sqcap$) has been dealt with in (Keet and Khumalo, 2014a) and disjunction ($\sqcup$) is similarly straightforward and therefore was omitted.

These patterns enable us to answer the second question, on whether they can be realised with a pure template-based approach, mostly template-based with some rules, or demand for a full-fledged grammar engine. No pattern can be realised with a template-based approach, i.e., for all basic constructs that were investigated, a grammar engine is a necessity. This is principally due to the complexities of isiZulu, with its agglutinative features (compared to the isolating morphology of, e.g., English), the 17 NCs, complex verb conjugation, the modifiers for negation, and context with the position of the symbol in the axiom. These features are emblematic for all Bantu languages, and, hence, the results presented in the previous sections are, in its generality, equally applicable to the other Bantu languages. Such transferability at least among two Nguni languages also has been observed in natural language understanding (Pretorius and Bosch, 2009), and it may be of interest to check that against

the grammar of the Swahili language manager SALAMA[7], with Swahili being in another Guthrie zone.

Concerning the third question—What do the answers to question 1 and 2 entail for an implementation of a controlled natural language?—several observations can be made following from the algorithms. The algorithms themselves each have several steps that can be executed in linear time, hence, look promising for efficient implementation of a grammar engine. However, it also did reveal two tricky parts that are yet to be fully resolved before an implementation is attempted. The first issue has to do with linguistic annotations of the vocabulary of the ontology, notably having to annotate the nouns of the OWL classes (DL concepts) with the NC it belongs to. Preliminary work to solve this (Chavula and Keet, 2014) showed that a simple reuse of the popular *lemon* model (McCrae *et al.*, 2012) or ontolex-lemon that is under standardisation does not adequately address the morphology requirements, and the linguistic model they use, LexInfo (Cimiano *et al.*, 2011), required an extension to also cover Bantu NCs. The second issue concerns the OWL object properties (DL roles). Either each permutation of the conjugation is added as a new object property in the ontology (hence, obtains its own URI) and made equivalent to the others, or only the verb root is added as an element in the ontology, and not only any verbalisation is computed on-the-fly, but also each axiom representation is computed on-the-fly. The former puts a greater strain on automated reasoning but less on the verbaliser, the latter is likely to slow down the ontology development environment, but will be easier for the verbaliser.

5.2 Other considerations

The novel results obtained with the verbs, which are essential in any ontology (as well as conceptual models, linked data, etc.), also revealed that much more will have to be done. Besides the passive voice and other verb modifiers (recall Example V4) and, possibly, also past tense, the complexity regarding the verb is aggravated when generated sentences have to be coherently put together in longer sentences. This can introduce the requirement of including the Object Concord (OC) on the verb morphology, which already encodes the SC together with the NEG and the NC, as was illustrated in example V3 in Section 2.2.2.

A pattern design decision was made to use syntax-based criteria notably for subsumption and universal quantification, which seems to hold always, but it is not clear yet whether that is indeed 100% true or only for most nouns. A more semantics-based approach in these cases—distinguishing between living and non-living things (Keet and Khumalo, 2014b)—may be feasible from a linguistic viewpoint. The reasons for a syntax-based approach were its amenability to computation compared to the semantics-based approach and because of the fluidity of the semantic-based approach, which is susceptible to a variety of sociolinguistic factors. The former would require some way of annotating

[7] `http://www.njas.helsinki.fi/salama/index.html`; last accessed: 8-12-2014.

the names of the OWL classes not only with NC, but also whether it is a living/nonliving thing, which requires more up-front investment in term analysis and a longer algorithm to handle the extra steps. The latter argument—a variety of sociolinguistic factors—is of a more general nature and even more challenging for natural language understanding, which also has to do with the stability of categorisation of nouns in the NCs (as mentioned earlier and also in (Ngcobo, 2010)), with the process of allocating new nouns into NCs (e.g., (Ngcobo, 2013)), and with adding new NCs. For instance, classes 3a and 9a present a complexity because they are new classes that accommodate modern terminology in isiZulu[8]. An example of the second issue—and of that of conflicting resources—is that of the loanword *udokotela* 'doctor', which is of NC:3a according to Wiktionary[9], but in NC:1a according to the `isizulu.net` online dictionary, and stems may be assigned to different classes; e.g., *izilungiselelo* means 'settings' specifically in software, which is of NC:8, which has been modified from the semantically similar *amalungiselelo* 'arrangements, preparations' in NC:6, having both *-lungiselelo* as stem (Keet and Barbour, 2015). This will complicate the annotation process for sentence generation, but is not unsurmountable.

Finally, the generation of a grammar is not helped by the antiquated literature. Doke's seminal work from 1927 and 1935 is outdated, there is no recent available isiZulu grammar book, and the textbooks and lecture notes are quite limited in their coverage and presentation of the grammar. This meant that even the content from Table 2 had to be pieced together from various sources by the authors, and the patterns had to be devised from scratch. The insufficiently structured grammar rules in the outdated documentation made it also clear, however, that committing to a comprehensive specification of the isiZulu grammar in such a way as to be computationally useful and correct, notably by availing of the comprehensive and popular Grammatical Framework (Ranta, 2011), will take a substantial amount of resources. Such resources are not available at present. The need for software in isiZulu and other South African languages has been voiced, as mentioned in Section 1, hence, something has to be done for multilingual knowledge repositories in (South(ern)) Africa now. This paper has presented foundational steps covering the core in that direction, which is benefiting both isiZulu linguistics and ICT in general, and it introduced some interesting new challenges for the verbalisation of logical theories in grammatically rich languages.

6 Conclusions

Due to the complexity of the morpho-syntax of isiZulu, a pure template-based approach to controlled natural language is infeasible. Thus verbalising

[8] The prefixes in these classes are identical to canonical prefixes (class 1a and 5) and are disambiguated semantically, however, the complexity is that their corresponding plurals are found in canonical classes (2a and 6 respectively).

[9] `http://en.wiktionary.org/wiki/Appendix:Zulu_nouns`; last accessed: 17-12-2014.

ontologies—and, more generally, knowledge-to-text—in isiZulu requires more than a template-based approach for each construct we have investigated, being basic use of universal and existential quantification, subsumption, negation, and simple axioms with object properties. The main features of isiZulu complicating verbalisation are the noun class system, the agglutinative nature of isiZulu, and contextual knowledge about the position of the symbol in the axiom. We identified verbalisation patters for selected $\mathcal{ALC}$ language constructs—quantification, subsumption, conjugation, and negation—and devised algorithms to generate grammatically correct isiZulu sentences, which were validated in an exploratory survey.

Avenues for further research into the verbalisation rules include a broader coverage of axioms and conjugation that can handle passive voice as well, and the sociolinguistic component of preferences for one pattern or another, which surfaced in the preliminary experimental evaluation (Keet and Khumalo, 2014b), among other things. There are also questions concerning how to make the ontology multilingual so that it covers the aspects that need to be recorded to facilitate verbalisation. An important next step is also the implementation.

Acknowledgements This work is based on the research supported in part by the National Research Foundation of South Africa (CMK: Grant Number 93397).

References

Alberts, R., Fogwill, T., and Keet, C. M. (2012). Several required OWL features for indigenous knowledge management systems. In P. Klinov and M. Horridge, editors, *7th Workshop on OWL: Experiences and Directions (OWLED 2012)*, volume 849 of *CEUR-WS*, page 12p. 27-28 May, Heraklion, Crete, Greece.

Androutsopoulos, I., Lampouras, G., and Galanis, D. (2013). Generating natural language descriptions from owl ontologies: the naturalowl system. *Journal of Artificial Intelligence Research*, **48**, 671–715.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F., editors (2008). *The Description Logics Handbook – Theory and Applications*. Cambridge University Press, 2 edition.

Bentley, M. and Kulemeka, A. (2001). *Chichewa*. Lincom Europa, Müchen.

Bosca, A., Dragoni, M., Francescomarino, C. D., and Ghidini, C. (2014). Collaborative management of multilingual ontologies. In P. Buitelaar and P. Cimiano, editors, *Towards the Multilingual Semantic Web*, page (in press). Springer.

Bouayad-Agha, N., Casamayor, G., and Wanner, L. (2014). Natural language generation in the context of the semantic web. *Semantic Web Journal*, **5**(6), 493–513.

Chavula, C. and Keet, C. M. (2014). Is lemon sufficient for building multilingual ontologies for Bantu languages? In C. M. Keet and V. Tamma, editors, *Proceedings of the 11th OWL: Experiences and Directions Workshop*

(OWLED'14), volume 1265 of *CEUR-WS*, pages 61–72. Riva del Garda, Italy, Oct 17-18, 2014.

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011). Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, **9**(1), 29–51.

Curland, M. and Halpin, T. (2007). Model driven development with NORMA. In *Proceedings of the 40th International Conference on System Sciences (HICSS-40)*, pages 286a–286a. IEEE Computer Society. Los Alamitos, Hawaii.

Dent, G. R. and Nyembezi, C. L. S. (2009). *Scholar's Zulu Dictionary*. Shuter & Shooter Publishers, 4 edition.

Doke, C. (1927). *Text Book of Zulu Grammar*. Witwatersrand University Press.

Doke, C. (1935). *Bantu Linguistic Terminology*. London: Longman, Green and Co.

Durrant, P. (2013). Formulaicity in an agglutinating language: the case of Turkish. *Corpus Linguistics and Linguistic Theory*, **9**(1), 1–38.

Engelbrecht, C., Shangase, N., Majeke, S., Mthembu, S., and Zondi, Z. (2010). Isizulu terminology development in nursing and midwifery. *Alternation*, **17**(1), 249–272.

Fogwill, T., Viviers, I., Engelbrecht, L., Krause, C., and Alberts, R. (2011). A software architecture for an indigenous knowledge management system. In *Indigenous Knowledge Technology Conference 2011*. Windhoek, Namibia, 2-4 November 2011.

Franconi, E., Guagliardo, P., and Trevisan, M. (2010). An intelligent query interface based on ontology navigation. In *Workshop on Visual Interfaces to the Social and Semantic Web (VISSW'10)*. Hong Kong, February 2010.

Fuchs, N. E., Kaljurand, K., and Kuhn, T. (2010). Discourse Representation Structures for ACE 6.6. Technical Report ifi-2010.0010, Dept of Informatics, University of Zurich, Switzerland.

Ghidini, C., Kump, B., Lindstaedt, S., Mabhub, N., Pammer, V., Rospocher, M., and Serafini, L. (2009). Moki: The enterprise modelling wiki. In *Proceedings of the 6th Annual European Semantic Web Conference (ESWC2009)*. Heraklion, Greece, 2009 (demo).

Goldsmith, J. and Buthelezi, G. (2005). The Zulu Language – Fall 2005. Online course material. University of Chicago, `http://hum.uchicago.edu/jagoldsm/ZuluLanguage/`; last accessed: 24-12-2014.

Guthrie, M. (1971). *Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages*. Number v. 1-2. Gregg.

Jarrar, M., Keet, C. M., and Dongilli, P. (2006). Multilingual verbalization of ORM conceptual models and axiomatized ontologies. Starlab technical report, Vrije Universiteit Brussel, Belgium.

Kaljurand, K., Kuhn, T., and Canedo, L. (in press (2014)). Collaborative multilingual knowledge management based on controlled natural language. *Semantic Web Journal*.

Keet, C. M. and Khumalo, L. (2014a). Basics for a grammar engine to verbalize logical theories in isiZulu. In A. Bikakis *et al.*, editors, *Proceedings of the 8th International Web Rule Symposium (RuleML'14)*, volume 8620 of *LNCS*, pages 216–225. Springer. August 18-20, 2014, Prague, Czech Republic.

Keet, C. M. and Khumalo, L. (2014b). Toward verbalizing logical theories in isiZulu. In B. Davis, T. Kuhn, and K. Kaljurand, editors, *Proceedings of the 4th Workshop on Controlled Natural Language (CNL'14)*, volume 8625 of *LNAI*, pages 78–89. Springer. 20-22 August 2014, Galway, Ireland.

Keet, C. M. and Barbour, G. (2015). Limitations of regular terminology development practices: the case of the isiZulu computing terminology. *Alternation*, **22**(1), 33–70.

Khumalo, L. (2007). *An analysis of the Ndebele Passive Construction*. Ph.D. thesis, University of Oslo, Norway.

Kuhn, T. (2013). A principled approach to grammars for controlled natural languages and predictive editors. *Journal of Logic, Language and Information*, **12**, 13–48.

Leo, J. (2008). Modeling relations. *Journal of Philosophical Logic*, **37**, 353–385.

Maho, J. (1999). A (tentative) verb slot system for Shona. Unpublished report for the ALLEX (African Languages Lexical) Project. Department of Oriental and African Languages, Göteborg University, Sweden.

Mberi (2002). *The categorical status and functions of auxiliaries in Shona*. Ph.D. thesis, University of Oslo, Norway.

McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). The Lemon cookbook. Technical report, Monnet Project.

Miti, L. (2006). *Comaprative Bantu phonology and Morphology*. Cape Town: The Center for Advanced Studies of African Societies (CASAS).

Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A., and Lutz, C. (2009). OWL 2 Web Ontology Language Profiles. W3C recommendation, W3C.

Msila, V. (2014). Africa must take pride of place in higher education. *Mail & Guardian*, pages Nov 14, 2014. http://mg.co.za/article/2014-11-13-africa-must-take-pride-of-place-in-higher-education.

Ngcobo, M. N. (2010). Zulu noun classes revisited: A spoken corpus-based approach. *South African Journal of African Languages*, **1**, 11–21.

Ngcobo, M. N. (2013). Loan words classification in isiZulu: The need for a sociolinguistic approach. *Language Matters: Studies in the Languages of Africa*, **44**(1), 21–38.

Pretorius, L. and Bosch, S. (2009). Exploiting cross-linguistic similarities in zulu and xhosa computational morphology: Facing the challenge of a disjunctive orthography. In *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages - AfLaT 2009*, pages 96–103.

Pretorius, L. and Bosch, S. E. (2003). Enabling computer interaction in the indigenous languages of South Africa: The central role of computational morphology. *ACM Interactions*, **March + April**, 56.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

Rector, A., Drummond, N., Horridge, M., Rogers, L., Knublauch, H., Stevens, R., Wang, H., and Wroe, Csallner, C. (2004). OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In *Proceedings of the 14th International Conference Knowledge Acquisition, Modeling and Management (EKAW'04)*, volume 3257 of *LNCS*, pages 63–81. Springer. Whittlebury Hall, UK.

Sharma Grover, A., Van Huyssteen, G., and Pretorius, M. (2011). The South African human language technology audit. *Language Resources & Evaluation*, **45**, 271–288.

Spiegler, S., van der Spuy, A., and Flach, P. A. (2010). Ukwabelana – an open-source morphological zulu corpus. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 1020–1028. Association for Computational Linguistics. Beijing.

Third, A., Williams, S., and Power, R. (2011). OWL to English: a tool for generating organised easily-navigated hypertexts from ontologies. poster/demo paper, Open Unversity UK. 10th International Semantic Web Conference (ISWC'11), 23-27 Oct 2011, Bonn, Germany.

Turner, N. S. (19xx). IsiZulu sokuzwana (zulu for mutual understanding). Course notes.

Wald, B. (1987). Swahili and the Bantu languages. In B. Comrie, editor, *The World's Major Languages*, pages 991–1014. Oxford: Oxford University Press.