On the verbalization patterns of part-whole relations in isiZulu

C. Maria Keet[1] and **Langa Khumalo**[2]
[1] Department of Computer Science, University of Cape Town,
South Africa, mkeet@cs.uct.ac.za
[2] Linguistics Program, University of KwaZulu-Natal,
South Africa, khumalol@ukzn.ac.za

Abstract

In the highly multilingual setting in South Africa, developing computational tools to support the 11 official languages will facilitate effective communication. The exigency to develop these tools for healthcare applications and doctor-patient interaction is there. An important component in this set-up is generating sentences in the language isiZulu, which involves part-whole relations to communicate, for instance, which part of one's body hurts. From a NLG viewpoint, the main challenge is the fluid use of terminology and the consequent complex agreement system inherent in the language, which is further complicated by phonological conditioning in the linguistic realisation stage. Through using a combined approach of examples and various literature, we devised verbalisation patterns for both meronymic and mereological relations, being structural/general parthood, involvement, containment, membership, subquantities, participation, and constitution. All patterns were then converted into algorithms and have been implemented as a proof-of-concept.

1 Introduction

Hitherto text-based human language technologies in South Africa have been developed by CTexT through the Autshumato project, whereas speech technologies have been developed by the Meraka Institute, which include Automatic Speech Recognition (ASR), pronunciation dictionaries and text-to-speech (TTS) technologies under the auspices of the Lwazi project. However, there is no computational technology in all indigenous official languages (including isiZulu), and the HLT audit (Sharma Grover et al., 2011) indicated a huge gap in information and knowledge processing in particular. This is important to address for application areas such as doctor-patient interactions, for which now only a small app with canned bilingual text exist[1]. The app was well-received for being a very small step toward meeting a well-known need of personalised health communication (Mettler and Kemper, 2003; Wilcox et al., 2011). However, due to the entirely manual efforts, the mobilezulu app with its canned text is obviously not scalable to cover all areas of medicine, like captured in standards such as SNOMED CT[2] and for which terminology in isiZulu is being developed (Engelbrecht et al., 2010) and standardised following PANSALB terminology development processes (Khumalo, 2016). SNOMED CT has a logic-based foundation by having the terms, relations, and the constraints that hold among them represented in the Description Logics-based OWL 2 EL ontology language (Motik et al., 2009a). OWL is also becoming popular as structured input for NLG (Bouayad-Agha et al., 2014) and CNLs (Safwat and Davis, 2016). Some results have been obtained in generating grammatically correct natural language sentences in isiZulu for the OWL 2 EL constructors (Keet and Khumalo, 2016), which makes it look promising to use. Exploratory experiments revealed several issues with verbalising axioms involving the pervasive part-whole relations (OWL object properties), however. The part-whole relation is compli-

[1] mobilezulu.org.za and mobilexhosa.org.za
[2] http://www.ihtsdo.org/snomed-ct/

cated by the fluid use in speech and terminology. For instance, structural parts (e.g., the jawbone of the head), involvement (swallowing as part of eating), and membership is generalised as *ingxenye* in isiZulu, yet participation is divided into individual (e.g., the patient) and collective (e.g., the operating team) participation, using different terms. The isiZulu-English dictionary lists 19 translations for 'part' alone (Dent and Nyembezi, 2009). It also introduces the need to process prepositions, which are present only in the deep structure in isiZulu (Mathonsi, 2001), rather than as identifiable isolated words in the better-resourced languages (such as 'of', *von* [DE], *van* [NL], *de* [SP]) that generally do have to be considered in NLG (Baldwin et al., 2009).

Linguistic and cognitive analyses of part-whole relations have resulted in part-whole relation taxonomies, notably the seminal first one by (Winston et al., 1987) and the most recent update in (Keet and Artale, 2008), which have been used successfully in NLP (e.g., (Tandon et al., 2016)). Such analyses start from the underspecified 'part' in natural language to examine what it really is ontologically. For NLG in isiZulu, we face a 'double direction' of analyses for non-English languages: *which parts are there, which terms are used for that, and how?* The general task at hand, thus, is to figure out how the lexicalisation and linguistic realisation of part-whole relations work in isiZulu.

We solve this problem by starting from an established taxonomy of part-whole relations and adjust where needed to cater for differences in conceptualisation as expressed in grammatically correct natural language. Unlike in English, where the same string—like 'has part', 'is part of', and 'contains'—can be plugged in a template unaltered[3], the lexicalisation and linguistic realisation in isiZulu depend on other constituents in the sentence. These include the noun class of the noun that plays the part or whole role in the sentence, the agreement system between a noun and a verb, phonological conditioning, and processing a preposition. In total, there are 13 such constituents for the part-whole relations covered. Instead of templates, this demands for *verbalisation patterns* such that a complete sentence can be generated during runtime. The results presented here thus also provide a first account of how to construct a full—albeit still highly structured—sentence in isiZulu that has more dependent components (so-called 'concordial agreement') than just verb conjugation with the subject concord and quantification with the quantitative concord. These patterns have been converted into algorithms and have been implemented as a proof-of-concept, substantially extending algorithms for verbalising OWL 2 EL axioms with 'simple' relations (verbs) and for pluralising nouns (Keet and Khumalo, 2016; Byamugisha et al., 2016), notably regarding locatives, concords, a preposition, and more comprehensive phonological conditioning.

The remainder of the paper is structured as follows. In Section 2 we outline the preliminaries on part-whole relations and CNLs for isiZulu. We spell out the patterns for the parts and wholes in isiZulu in Section 3. We describe the tool design considerations and implementation in Section 4. We discuss in Section 5 and conclude in Section 6.

2 Preliminaries

Part-whole relations in the context of natural language commenced seriously with (Winston et al., 1987), with various modifications to its latest instalment by (Keet and Artale, 2008) as to which part-whole relations there are. These part-whole relations are also used in NLP (e.g., (Tandon et al., 2016)), and in ontologies and controlled vocabularies in medicine, such as openGalen and SNOMED CT. There is a principal distinction between mereology (parthood) and meronymy (parts in natural language), where the latter includes the former. They are summarised with an example in Table 1.

CNLs are gaining popularity as a version of NLG in the scope of data(base/RDF)-to-text and knowledge(/logic/OWL)-to-text. It has been shown that straightforward templates do not suffice for Bantu languages such as isiZulu, because (almost) *all words* in *any* sentence *need* some processing (Keet and Khumalo, 2016), cf. an occasional rule for flexibility or beautification that one may still rather classify as a template-based approach (van Deemter et al., 2005). This is due mainly to the system of noun classes, the agreement system among the various constituents in a sentence, and the agglutinative characteristics (Keet and Khumalo, 2016). The noun

[3] check, e.g., SWAT NL (Third et al., 2011) or ACE online (Fuchs et al., 2010).

Table 1: Main part-whole relations.

Relation	Example
structural parthood	wall is part of a house, human has part a heart (physical objects)
involvement	eating involves swallowing (processes)
location	city is located in a country (2D region with occupant)
containment	nucleus contained in cell, bolus of food is contained in the stomach (3D region with occupant)
membership	player is member of a team (role & collective)
participation	enzyme participates in a catalytic reaction (object & process)
subquantities	sugar is a subquantity of lemonade, blood sample is a sub quantity of blood (stuffs/masses)
constitution	a vase is constituted of clay (object & stuff)

classes for isiZulu with relevant concords affecting other words in a sentence is shown in Table 2. The noun class system is one of the salient features of the isiZulu language. Every noun belongs to a noun class (NC). The noun is made up of two formatives, the prefix and the stem (e.g., for NC2: *aba-* + *fana* = *abafana* 'boys'). Crucially, the NC governs the agreement of all words that modify the noun. Most NCs are set off into pairs in isiZulu such that most nouns have a singular form in one class and a plural form in another as summarised in Table 2. It must also be pointed out that for the most part the semantics of a noun determines its class (cf. (Twala, 1992)).

So-called 'verbalisation patterns' and algorithms have been developed by (Byamugisha et al., 2016; Keet and Khumalo, 2016), which cover knowledge representation language features from the Description Logic (DL) $\mathcal{ALC}$ (Baader et al., 2008)—hence, OWL 2 EL (Motik et al., 2009a)—such as existential and universal quantification, subsumption, and negation, which have been implemented by the authors in the meantime. The relevant aspects are summarised here to keep the paper self-contained:

- Conjunction 'and' ($\sqcap$ in DL notation), enumerative: *na-* is added to the second noun, using phonological conditioning (see below).
- Subsumption: The copulative is either *y-* or *ng-*, depending on the first letter of the name of the superclass, and added to the name of the superclass; e.g., *inja yisilwane* 'dog is an animal'.
- Quantification, restricted to usage in simple inclusions of the form $C \sqsubseteq \exists R.D$, i.e., 'all Cs R at least one D'. The $\forall$ 'all' is determined by the noun class of the plural of C's name, R is a present tense verb conjugated in concordance with the head noun (C, in plural), and the 'at least one' is made up of the relative concord and quantitative concord of the noun class of D's name and ends with *-dwa*. For instance, uSolwazi $\sqsubseteq$ -fundisa.isifundo becomes *bonke oSolwazi bafundisa isifundo esisodwa* 'all professors teach at least one course': First, *uSolwazi*, in NC3, is pluralised to *oSolwazi*, in NC4. Second, the word for $\forall$ for NC4 is *bonke* and, third, the subject concord for it is *ba-*, making *bafundisa*. Fourth, the noun class of *isifundo* is 7, so the relative concord is *esi-* and quantitative concord is *-so-*, forming *esisodwa* for the verbalisation of $\exists$.

Phonological conditioning occurs in multiple occasions (Miti, 2006), but the one relevant here concerns adding a concord to the noun, because isiZulu does not have two successive vowels in a word. This is known as *vowel coalescence*, and the basic rules are: *-a* + *a-* = *-a-*, *-a* + *e-* = *-e-*, *-a* + *i-* = *-e-*, *-a* + *o-* = *-o-*, and *-a* + *u-* = *-o-*. For instance, *ubisi na+ibhotela* becomes *ubisi nebhotela* ('milk and butter'), *ibhotela na+ubisi* becomes *ibhotela nobisi*, and *nga+ubumba* becomes *ngobumba* 'of clay'. Further, the locative suffix *-ini* is phonologically conditioned by the final vowel: *-a+-ini=-eni*, *-e+-ini=-eni*, *-o+ini=-weni*, *u+ini=-wini*, *-phu* + *-wini* = *-shini*, and the few loanwords that end in *-phu* become *-phini*.

3 Patterns for Parts and Wholes

To describe the patterns, we systematically take the axioms for 'has part' (wp), $W \sqsubseteq \exists hasPart.P$, and 'is part of' (pw), $P \sqsubseteq \exists isPartOf.W$ to demonstrate what is going on linguistically. Ontologically, in a majority of cases, only one of the two reading directions is applicable despite the pervasive informal use of the inappropriate one[4]; such ontological

[4]e.g., it is true that all humans have some heart ($Human \sqsubseteq \exists hasPart.Heart$), but not that all hearts are part of some hu-

Table 2: Zulu noun classes with examples and a selection of concords. NC: Noun class; PRE: prefix; QC: quantitative concord; RC: relative concord; PC: possessive concord.

NC	Full PRE	QC (∀)	RC	QC (∃)	SC	PC
1	um(u)-	wonke	o-	ye-	u-	wa-
2	aba-	bonke	aba-	bo-	ba-	ba-
1a	u-	wonke	o-	ye-	u-	wa-
2a	o-	bonke	aba-	bo-	ba-	ba-
3a	u-	wonke	o-	ye-	u-	wa-
2a	o-	bonke	aba-	bo-	ba-	ba-
3	um(u)-	wonke	o-	wo-	u-	wa-
4	imi-	yonke	e-	yo-	i-	ya-
5	i(li)-	lonke	eli-	lo-	li-	la-
6	ama-	onke	a-	wo-	a-	a-
7	isi-	sonke	esi-	so-	si-	sa-
8	izi-	zonke	ezi	zo-	zi-	za-
9a	i-	yonke	e-	yo-	i-	ya-
6	ama-	onke	a-	wo-	a-	a-
9	i(n)-	yonke	e-	yo-	i-	ya-
10	izi(n)-	zonke	ezi-	zo-	zi-	za-
11	u(lu)-	lonke	olu-	lo-	lu-	lwa-
10	izi(n)-	zonke	ezi-	zo-	zi-	za-
14	ubu-	bonke	obu-	bo-	bu-	ba-
15	uku-	konke	oku-	ko-	ku-	kwa-
17	ku-	lonke	olu-	lo-	lu-	kwa-

Table 3: Abbreviations (Var.) used in the verbalisation patterns.

Var.	Full name	Comment
W	entity playing whole	our abbreviation
P	entity that plays the part	our abbreviation
CONJ	Conjunction	enumerative-and (not a connective-and); *na-*
COP	Copulative	*y-* or *ng-*
LOC	Locative	locative prefix; *ku-* for NC 1a, 2a, 3a, and 17, *e-* otherwise
LOC-SUF	Locative	here used for the locative suffix; *-ini*
PRE	Preposition	only *nga-* is used here
EP	Epenthetic	*-s-*
PASS	Passive tense	*-iw-*
FV	Final Vowel	in this case just *-e* to go with PASS
SC	Subject Concord	$\sim$ conjugation; depends on NC: see Table 2
PC	Possessive Concord	depends on NC: see Table 2
RC	Relative Concord	depends on NC: see Table 2
QCall	quantitative concord	universal quantification; depends on NC: see Table 2
QC	quantitative concord	existential quantification; depends on NC: see Table 2

aspects are beyond the scope of this paper.

Regarding notation, ultimately what is needed is a detailed grammar for the verbalisation patterns. At this stage, however, there is insufficient linguistic knowledge to pursue this. Therefore, we use variables in the patterns, as listed in Table 3, where each variable is to be substituted with the appropriate string (terminal, if it were a CFG), and subscripts, omitting the orthogonal phonological conditioning that is included in the explanation instead. A dash between variables indicates they are part of one word. Subscripts indicate 'agreement' of the various elements. So, for instance, a "$W_{nc_{x,pl}}$" is the entity (its name assumed to be given in the singular) that plays the role of the whole, which is of noun class ("$_{nc}$") x that is to be pluralised, and its preceding "$QCall_{nc_{x,pl}}$" is the term for the universal quantification for the noun class that is the plural of noun class x; e.g., if W is inja, in NC9, then $W_{nc_{x,pl}}$ is izinja in NC10, and its $QCall_{nc_{x,pl}}$ is zonke.

structural/general parts and wholes Let us commence with a parthood relation between objects. The verbalisation patterns in isiZulu (for any noun class) in the 'has part' (*wp*) and 'part of' (*pw*) reading directions are as follows:

wp: $QCall_{nc_{x,pl}}$ $W_{nc_{x,pl}}$ $SC_{nc_{x,pl}}$-CONJ-P_{nc_y} RC_{nc_y}-QC_{nc_y}-*dwa*

pw: $QCall_{nc_{x,pl}}$ $P_{nc_{x,pl}}$ $SC_{nc_{x,pl}}$-COP-*ingxenye* $PC_{ingxenye}$-W_{nc_y} RC_{nc_y}-QC_{nc_y}-*dwa*

Note that the whole-part relation does not have one single string like a 'has part', but it is composed of SC+CONJ, and is thus dependent on both the noun class of the whole (as the SC is) and on the first letter of the name of the part (as the string for CONJ, *na-*, depends on that). The 'is part of' reading direction is made up of the 'part' *ingxenye*, which is a noun that is preceded with the COP *y-* and together amounts to 'is part'. The 'of' is accounted for by the possessive concord (PC) of *ingxenye* (NC9), be-

man ($^*Heart \sqsubseteq \exists isPartOf.Human$), as there are hearts that are part of another, non-human, animal.

ing *ya-*, taking into account vowel coalescence. The SC for concordance with the P has been included because, while in multiple examples, either SC-COP-*ingxenye* or COP-*ingxenye* suffices, in some cases it really does not. The patterns are illustrated in the following two examples for heart (*inhliziyo*, NC9) standing in a part-whole relation to human (*umuntu*, NC1), with the 'has part' and 'is part of' underlined:

wp-ex: bonke abantu banenhliziyo eyodwa

pw-ex: zonke izinhliziyo ziyingxenye yomuntu oyedwa

involved in is the same as for general parts. The salient difference is that both P and W belong to nominals that are in NC15. An example is that eating (*ukudla*) involves swallowing (*ukugwinya*):

wp-ex: konke ukudla kunokugwinya okukodwa

pw-ex: konke ukugwinya kuyingxenye yokudla okukodwa

Observe that "bane-" in the previous example is different from the "kuno-" here, due to the different SCs (*abantu* is in NC2 (ba-) and *ukudla* in NC15 (ku-), and vowel coalescence: *na+i* = -ne- in the former example and *na+u* = -no- here, yet the pattern is exactly the same.

containment has a spatial component to it, which is indicated with the locative affixes (LOC) in the pw direction of verbalisation. Because isiZulu proscribes vowel sequencing, the epenthetic *-s-* is required between the SC and the LOC *e-*. Patterns, for any noun class:

wp: $\text{QCall}_{nc_{x,pl}}$ $\text{W}_{nc_{x,pl}}$ $\text{SC}_{nc_{x,pl}}$-CONJ-P_{nc_y} RC_{nc_y}-QC_{nc_y}-*dwa*

pw: $\text{QCall}_{nc_{x,pl}}$ $\text{P}_{nc_{x,pl}}$ $\text{SC}_{nc_{x,pl}}$-EP-LOC-W_{nc_y}-LOCSUF RC_{nc_y}-QC_{nc_y}-*dwa*

This is illustrated for the usual example (Donnelly et al., 2006) of a bolus of food (*indilinga yokudla*, NC9) that is contained in the stomach (*isisu*, NC7):

wp-ex: Zonke izisu zinendilinga yokudla eyodwa

pw-ex: Zonke izindilinga zokudla zisesiswini esisodwa

The zine- comes from the SC of NC10 of *izisu* 'stomachs', which is followed by the *na+i*=-ne- for CONJ. The zise- is the result of NC10's SC, zi- (see Table 2), the EP -s-, and LOC e-, and then *-u+-ini*=-wini as LOCSUF.

membership The patterns are as for general parthood; e.g., a doctor (*udokotela*, NC1a) is a member of an operating team (*iqembu labahlinzi*, NC5):

wp-ex: onke amaqembu abahlinzi anodokotela oyedwa

pw-ex: bonke odokotela bayingxenye yeqembu labahlinzi elilodwa

subquantities Ontology has so far recognised two core different usages of subquantities. First, as parts, like alcohol is a subquantity of wine, flour of bread and so on. While many of the mass nouns are in NC5 or NC6 in isiZulu, this is not always the case and if in the singular it stays singular and in some cases, the term can be both a count noun and a mass noun, as is the case in English (e.g., 'stone'). Therefore, we change the pattern for part-subquantities so that it omits the pluralisation. Also, one does not count stuffs, so the 'at least one' is omitted as well.

wp: QCall_{nc_x} W_{nc_x} SC_{nc_x}-CONJ-P_{nc_y}

pw: QCall_{nc_x} P_{nc_x} SC_{nc_x}-COP-*ingxenye* $\text{PC}_{ingxenye}$-W_{nc_y}

For instance, water (*amanzi*, NC6) as a subquantity of urine (*umshobingo*, NC3):

wp-ex: wonke umshobingo unamanzi

pw-ex: onke amanzi ayingxenye yomshobingo

The second reading of subquantities is portions, i.e., parts of the whole amount of stuff that are made of the same stuff, be this a tissue sample under the microscope glass that came from a patient's tissue, or the left-half of someone's brain. In isiZulu, there are two types: *umunxa* (NC3) as a kind of 'spatial' portion as in 'the portion of the kitchen where the kitchen utensils are', and *isiqephu* (NC7) as a portion for solid objects, like the tissue. For the 'spatial' portion, we obtain:

wp: $\text{QCall}_{nc_{x,pl}}$ $\text{W}_{nc_{x,pl}}$ $\text{SC}_{nc_{x,pl}}$-CONJ-P_{nc_y}

pw: $\text{QCall}_{nc_{x,pl}}$ $\text{P}_{nc_{x,pl}}$ $\text{SC}_{nc_{x,pl}}$-COP-*umunxa* PC*umunxa*-W_{nc_y}

Observe that the COP is *ng-*, not *y-*, because of the *u*-commencing *umunxa*; e.g., a hospital (*isibhedlela*, NC7) has a portion that is an operating theatre (*ithiyetha yokuhlinzela*, NC9a):

wp-ex: zonke izibhedlela zinethiyetha yokuhlinzela

pw-ex: onke amathiyetha okuhlinzela angumunxa wesibhedlela

For the solid objects type of portion, the whole is an amount of matter (mass noun), thus remains in the noun class it is rather than being pluralised:

wp: QCall_{nc_x} W_{nc_x} SC_{nc_x}-CONJ-P_{nc_y} RC_{nc_y}-QC_{nc_y}-*dwa*

pw: QCall$_{nc_{x,pl}}$ P$_{nc_{x,pl}}$ SC$_{nc_{x,pl}}$-COP-*isiqephu* PC$_{isiqephu}$-W$_{nc_y}$ RC$_{nc_y}$-QC$_{nc_y}$-*dwa*

with as example a blood sample as a portion of blood

wp-ex: Lonke igazi linesampula legazi elilodwa

pw-ex: Onke amasampula egazi ayisiqephu segazi elilodwa

For the W in the *pw*, there is again vowel coalescence: *sa-*+*igazi* = segazi, with *sa-* the PC for *isiqephu*'s NC7. The part P is computationally complicated. It may be a noun phrase, like 'slice of bread', where the 'of' is again catered for by a PC, being the one for the noun class of the noun that is the quantity (slice, piece, bowl, etc). So, e.g., *ucezu* (NC11) has PC *lwa-*, resulting in *lwa-*+*isinkwa* = lwesinkwa 'of bread'. Yet, a 'sample of blood', *isampula legazi*, is considered a compound noun, not a noun phrase.

participation can be divided into two typologies in isiZulu. There is individual type of participation and a group type of participation, like a citizen vs the electorate participating (taking part) in an election. For individual objects, one can include an optional ASP between the SC and COP, restricted to *-be-* in this case. This is not used here so as to match with the rest, assuming that it will suffice. As example, a doctor (*udokotela*, NC1a) participates in an operation (*ukuhlinza*, NC15):

wp-ex: Konke ukuhlinza kunodokotela oyedwa

pw-ex: bonke odokotela bayingxenye yokuhlinza okukodwa

For the collective/group participation, a different 'part' is used, *-hlanganyele*, which is part in the sense of participating by combining to do something, acting in unison (perfect tense). This is verbalised in the singular only:

wp: QCall$_{nc_x}$ W$_{nc_x}$ SC$_{nc_x}$-CONJ-P$_{nc_y}$ RC$_{nc_y}$-QC$_{nc_y}$-*dwa*

pw: QCall$_{nc_x}$ P$_{nc_x}$ SC$_{nc_x}$-*hlanganyele* LOC-W$_{nc_y}$-LOCSUF RC$_{nc_y}$-QC$_{nc_y}$-*dwa*.

Either a LOC as prefix only is allowed, or a locative circumfix can be used, i.e., LOC-W-LOCSUF with vowel elision for the W on both sides. Here, the latter is chosen. For instance, the operating team, (*iqembu labahlinzi*, NC5) participating in an operation (*ukuhlinza*, NC15):

wp-ex: Konke ukuhlinza kuneqembu labahlinzi elilodwa

pw-ex: Lonke iqembu labahlinzi lihlanganyele okuhlinzeni okukodwa

Decomposing the locative aspects that result in okuhlinzeni: the o- is the outcome of the vowel coalescence of LOC *e-*+*u-* and -weni is the outcome of the phonological conditioning *-o*+*-ini*'s LOCSUF.

constitution Also in this case of meronymic part-whole relation, it partially diverges in that there is no variation of 'part' as a noun, but a verb is used, as in the previous case: it is either *-akha* 'build' for objects that are made/constituted of some matter in some structural sense or *-enza* otherwise. As this is verbalised only as wholes being constituted of something, only that one is included:

wp: QCall$_{nc_{x,pl}}$ W$_{nc_{x,pl}}$ SC$_{nc_{x,pl}}$-*akh*-PASS-FV PRE-P$_{nc_y}$.

wp: QCall$_{nc_{x,pl}}$ W$_{nc_{x,pl}}$ SC$_{nc_{x,pl}}$-*enz*-PASS-FV PRE-P$_{nc_y}$.

The PRE here is restricted to *nga-*, with phonological conditioning. Relatively, this construction is similar to the notion of preposition contraction in Romance languages (de Oliveira and Sripada, 2014). For instance, in 'all houses (*izindlu* 'house') are constituted of stone (*itshe*, NC5)', the passive and final vowel causes the -iwe end, and likewise for 'all pills (*amaphilisi*, NC6) are made of starch (*isitashi*, NC7)':

wp-ex: zonke izindlu zakhiwe ngetshe

wp-ex: onke amaphilisi enziwe ngesitashi

The SC is modified because the stem starts with a vowel: if the vowel of the SC is a high vowel (i-; u-) and precedes the vowel of the stem which is low (a-), there is hiatus resolution (Mudzingwa and Kadenge, 2011). The pattern is as follows: *i-* + *a-* = *y* and *u-* + *a-* = *w*. Hiatus resolution is followed by the elision of the initial vowel with the semi-vowel attaching to the initial vowel of the stem (*u-* + *akhiwe* = *yakhiwe*).

This concludes the list of patterns.

4 Design and Implementation

We describe the transformation from the patterns to the algorithms, some tool design considerations, and the architecture of the implementation.

4.1 From verbalisation patterns to algorithms

The variables used in the verbalisation patterns belie what needs to be done in the background, which differs by variable in three principal ways. First,

there are the variables that algorithmically amount to straight-forward *look-up functions* to retrieve something using the noun class, such as the SC, RC, and QC as listed in Table 2. Second, there are *functions that change a word*, notably the pluraliser, which is not simply a case of list look-up (Byamugisha et al., 2016). Third, there are the *functions for phonological conditioning* that are needed for CONJ, LOC, LOCSUF, PC, and PRE. Most of the algorithms to verbalise part-whole relations need all three groups of functions. For instance, Algorithm 1 for the verbalisation of the basic whole-part has the straightforward look-up ones ("get..."), the call to another algorithm for pluralisation (line 7), and one call to the rules for vowel coalescence (phonological conditioning) in line 11. The algorithm for the 'is part of' direction is similar except that instead of line 11, the phonological conditioning is *phonoCondition(*`'ya'`, c_2*)* and sc_1 stringed together with `yingxenye`.

Algorithm 1: Determine the verbalisation of basic whole-part in an axiom

1: $\mathcal{C}$ set of classes, language $\mathcal{L}$, $\sqsubseteq$ for subsumption, $\exists$ 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;
Require: axiom of the form $W \sqsubseteq \exists wp.P$ has been retrieved for verbalisation
2: $c_1 \leftarrow getFirstClass(A)$ {get whole}
3: $c_2 \leftarrow getSecondClass(A)$ {get part}
4: $wp \leftarrow getObjProp(A)$ {get wp type ('default' parthood here)}
5: $NC_1 \leftarrow getNC(c_1)$ {obtain noun class whole}
6: $NC_2 \leftarrow getNC(c_2)$ {obtain noun class part}
7: $c_{pl} \leftarrow$ *pluralise*(c_1, NC_1) {generate plural, using the pluraliser algorithm}
8: $NC_1' \leftarrow getPlNC(NC_1)$ {obtain plural NC, from known list}
9: $a_1 \leftarrow getQCAll(NC_1')$ {obtain quantitative concord (QC(all))}
10: $sc_1 \leftarrow getSC(NC_1')$ {obtain subject concord}
11: $conjp \leftarrow$ *phonoCondition(*`'na'`,c_2*)* {prefix P with the CONJ, phonologically conditioned}
12: $r_2 \leftarrow getRC(NC_2)$ {obtain relative conc. for c_2}
13: $q_2 \leftarrow getQC(NC_2)$ {obtain quant. concord for c_2 from the QC (exists)-list}
14: RESULT $\leftarrow$ ' a_1 c_{pl} $sc_1 conjp$ $r_2 q_2$dwa. ' {verbalise the simple axiom}
15: **return** RESULT

4.2 Design considerations

As the patterns demonstrate, the actual string for 'has part' depends on the noun of the entity that plays the role of the whole and noun of the entity that plays the role of the part, which means that it is not feasible to store all possible strings, but this has to be computed on-the-fly. Yet, OWL requires a single, fixed, string of text for its 'object property' (relationship), i.e., a single IRI (Motik et al., 2009b). Integrating this with OWL means handling object properties differently and full integration with a linguistic model, yet the *lemon* model (McCrae et al., 2012) already needs an extension to deal with the noun classes (Chavula and Keet, 2014), or: that structured representation does not suffice for isiZulu at present. As solving that diverts away from a proof-of-concept implementation of the algorithms for part-whole relations to evaluate whether they and the patterns they implement are correct, we chose an incremental approach with Python instead. Also, the patterns and algorithms presented in (Keet and Khumalo, 2016; Byamugisha et al., 2016) have been implemented in Python, so we extended that with the algorithms for the novel part-whole patterns.

The architecture of the components of the verbaliser are straightforward (see Fig. 1): nouns are stored with their noun class, whereas verb stems are stored to facilitate processing of tense, for automatically determining this has only partial solutions thus far (Pretorius and Bosch, 2003; Spiegler et al., 2010). Each axiom type and each type of part-whole relation relates to a Python function (which calls others). The script is yet to be connected to the SNOMED CT's owl file to fetch the data, so the code emulates that output such that the user adds the terms in the input (see Fig. 2, "->" lines). The code and other examples can be downloaded from `http://www.meteck.org/files/geni/` and a few examples are shown in Fig. 2. It worked for 38 of the 42 test cases (90.5%). The four errors were mainly due to the incomplete pluraliser of (Byamugisha et al., 2016) (e.g., *ucezi* $\mapsto$ *izincezi*, not *izicezi*) and one due to ambiguity of *-akh* vs. *-enz* for constitution.

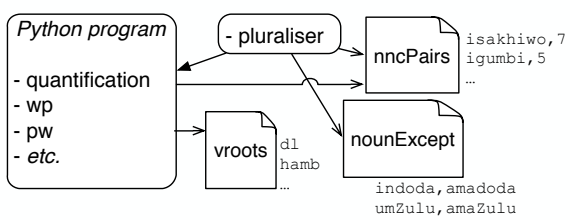

Figure 1: Components of the proof-of-concept implementation of the isiZulu verbaliser. The three txt files were created manually (examples of their contents are shown in `courier` font).

```
> wp('umuntu','inhliziyo')
  'Bonke abantu banenhliziyo eyodwa'
> wp_cp('ukhetho','umphakathi')
  'Lonke ukhetho lunomphakathi owodwa'
> wp_s('umshobingo','amanzi')
  'Wonke umshobingo unamanzi'
> pw('ukugwinya','ukudla')
  'Konke ukugwinya kuyingxenye yokudla okukodwa'
> pw_ci('indilinga yokudla','isisu')
  'Zonke izindilinga zokudla zisesiswini esisodwa'
> constitution('ivazi','ubumba')
  'Onke amavazi akhiwe ngobumba'
```

Figure 2: Screenshot of working code; wp/pw: general wholes/parts; wp_cp: collective participation; wp_s: subquantity; pw_ci: containment.

5 Discussion

The patterns showed that, like in English, isiZulu has several more specific terms for 'part'—*ingxenye*, *indawo*, *isiqephu*, *umunxa*, and *hlanganyele*—although they do not match 1:1 with the established part-whole relation categorisations as in Table 1. Such ontological analyses are left for future work. It does illustrate that in this case sentence planning was a major hurdle compared to just linguistic realisation.

The patterns reconfirm results by (Keet and Khumalo, 2016) that the template-based approach is not feasible for isiZulu, and, by extension, Bantu languages that all share the features of noun classes and concordance. This, however, also makes it an imperative to develop a grammar. While this exercise broadened the scope on understanding what linguistic elements are needed for an NLG, and a quasi pattern language was still sufficient to specify the patterns, with the increased number of elements to keep track of compared to (Keet and Khumalo, 2016), soon this limit will be reached. In addition, rules need to be found so as to process *groups* of tokens so as to know which one is a compound noun and which one is a noun phrase, in order to process them correctly. Hopefully then also sufficient insight is gained to construct a set of requirements for the grammar and either practical ones might be extended, such as the CFG of Ukwabelana (Spiegler et al., 2010), explorations of (Zeller, 2005) worked out in detail, or a natural language-independent approach like in (Kuhn, 2013) may be adjusted, or a new one devised to handle the syntactic elements to generate sentences with the intended semantics.

Finally, although the patterns have been specified for isiZulu only, bootstrapping resources for related Bantu languages—Xhosa, Swati, and Ndebele—based on isiZulu resources have yielded good results (Bosch et al., 2008), and thus solving it for isiZulu will open up HLT prospects for even lesser resourced languages.

6 Conclusions

We devised verbalisation patterns for both meronymic and mereological relations. New constituents in the patterns with respect to related works are, notably, the possessive concord, locative affixes, and a basic treatment of prepositions and the passive tense. The verbalisation patterns were implemented successfully using a proof-of-concept implementation of the algorithms, and tested with 42 examples, resulting in a 90.5% success rate. The patterns reaffirm the infeasibility of the template-based approach for isiZulu and Bantu languages because of the complex morphosyntax.

The patterns also indicated that it is becoming a pressing matter to commence with formally defining a generative grammar for isiZulu. Another avenue will be to take the latest medical terminology terms in isiZulu and create a fully functional medical app.

Acknowledgments

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

References

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

Timothy Baldwin, Valia Kordoni, and Aline Villavicencio. 2009. Prepositions in applications: A survey and introduction to the special issue. *Computational Linguistics*, 35(2):119–149.

Sonja Bosch, Laurette Pretorius, and Axel Fleisch. 2008. Experimental bootstrapping of morphological analysers for nguni languages. *Nordic Journal of African Studies*, 17(2):66–88.

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

Joan Byamugisha, C. Maria Keet, and Langa Khumalo. 2016. Pluralising nouns in isiZulu and similar languages. In A. Gelbkuh, editor, *Proceedings of CICLing'16*, page in print. Springer.

Catherine Chavula and C. Maria Keet. 2014. Is lemon sufficient for building multilingual ontologies for Bantu languages? In C. Maria Keet and Valentina 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.

Rodrigo de Oliveira and Somayajulu Sripada. 2014. Adapting simplenlg for brazilian portuguese realisation. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 93–94, Philadelphia, Pennsylvania, U.S.A., June. Association for Computational Linguistics.

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

M. Donnelly, T. Bittner, and C. Rosse. 2006. A formal theory for spatial representation and reasoning in biomedical ontologies. *Artif Intell Med*, 36(1):1–27.

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

Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. 2010. Discourse Representation Structures for ACE 6.6. Technical Report ifi-2010.0010, Dept of Informatics, University of Zurich, Switzerland.

C. Maria Keet and Alessandro Artale. 2008. Representing and reasoning over a taxonomy of part-whole relations. *Applied Ontology*, 3(1-2):91–110.

C. M. Keet and L. Khumalo. 2016. Toward a knowledge-to-text controlled natural language of isiZulu. *Language Resources and Evaluation*, in print:DOI: 10.1007/s10579–016–9340–0.

L. Khumalo. 2016. Disrupting language hegemony: Intellectualizing African languages. In M. Samuel, R. Dunpath, and N. Amin, editors, *Towards a post-humanist higher education curriculum: Undoing cognitive damage*, page (accepted). SENSE Publishers, Rotterdam.

Tobias Kuhn. 2013. A principled approach to grammars for controlled natural languages and predictive editors. *Journal of Logic, Language and Information*, 22(1):33–70.

N. N. Mathonsi. 2001. Prepositional and adverb phrases in Zulu: a linguistic and lexicographic problem. *South African Journal of African Languages*, 2:163–175.

John McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012. The Lemon cookbook. Technical report, Monnet Project.

M. Mettler and D.W. Kemper. 2003. Information therapy: Health education one person at a time. *Health Prom. Prac.*, 4(3):214–217.

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

Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, Zhe Wu, Achille Fokoue, and Carsten Lutz. 2009a. OWL 2 Web Ontology Language Profiles. W3C recommendation, W3C, 27 Oct.

Boris Motik, Peter F. Patel-Schneider, and Bijan Parsia. 2009b. OWL 2 web ontology language structural specification and functional-style syntax. W3c recommendation, W3C, 27 Oct. http://www.w3.org/TR/owl2-syntax/.

C. Mudzingwa and M. Kadenge. 2011. Comparing hiatus resolution in karanga and nambya: An optimality theory account. *Nordic Journal of African Studies*, 20(3):203–240.

L. Pretorius and E. S. Bosch. 2003. Finite-state computational morphology: An analyzer prototype for Zulu. *Machine Translation*, 18(3):195–216.

Hazem Safwat and Brian Davis. 2016. CNLs for the semantic web: a state of the art. *Language Resources & Evaluation*, in print:DOI: 10.1007/s10579–016–9351–x.

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

Sebastian Spiegler, Andrew van der Spuy, and Peter A. Flach. 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.

Niket Tandon, Charles Hariman, Jacopo Urbani, Anna Rohrbach, Marcus Rohrbach, and Gerhard Weikum. 2016. Commonsense in parts: Mining part-whole relations from the web and image tags. In *Proceedings*

of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16), pages 243–250. AAAI Press.

Allan Third, Sandra Williams, and Richard Power. 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.

E. K. Twala. 1992. The noun class system of isizulu. M.A. dissertation. University of Johannesburg.

Kees van Deemter, Emiel Krahmer, and Mariët Theune. 2005. Real versus tempalte-based natural language generation: a false opposition? *Computational Linguistics*, 31(1):15–23.

L. Wilcox, D. Morris, D. Tan, J. Gatewood, and E. Horvitz. 2011. Characterising patient-friendly micro-explanations of medical events. In *SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*, pages 29–32. ACM.

M.E. Winston, R. Chaffin, and D. Herrmann. 1987. A taxonomy of partwhole relations. *Cognitive Science*, 11(4):417–444.

Jochen Zeller. 2005. Universal principles and parametric variation: remarks on formal linguistics and the grammar of zulu. *Ingede Journal of African Scholarship*, 1(3):20p.