Pluralizing Nouns across Agglutinating Bantu Languages

Joan Byamugisha
University of Cape Town
Cape Town
South Africa
jbyamugisha@cs.uct.ac.za

C. Maria Keet
University of Cape Town
Cape Town
South Africa
mkeet@cs.uct.ac.za

Brian DeRenzi
Dimagi
Cape Town
South Africa
bderenzi@dimagi.com

Abstract

Text generation may require the pluralization of nouns, such as in context-sensitive user interfaces and in natural language generation more broadly. While this has been solved for the widely-used languages, this is still a challenge for the languages in the Bantu language family. Pluralization results obtained for isiZulu and Runyankore showed there were similarities in approach, including the need to combine morphology with syntax and semantics, despite belonging to different language zones. This suggests that bootstrapping and generalizability might be feasible. We investigated this systematically for seven languages across three different Guthrie language zones. The first outcome is that Meinhof's 1948 specification of the noun classes are indeed inadequate for computational purposes for all examined languages, due to non-determinism in prefixes, and we thus redefined the characteristic noun class tables of 29 noun classes into 53. The second main result is that the generic pluralizer achieved over 93% accuracy in coverage testing and over 94% on a random sample. This is comparable to the language-specific isiZulu and Runyankore pluralizers.

1 Introduction

The need for generating plurals to ease the use of software tools is well known in several areas. In Natural Language Understanding, plurals are used during parsing to distinguish between atomic elements (simple nouns) and collectives (Schubert, 2015), while in Natural Language Generation, pluralization is essential during quantification (Schubert, 2015) and number agreement. For instance, to inflect properly for the grammatical number in end-user tools, like a calendar properly stating to set off an alarm "1 hour" before and not "1 hours", but also for more advanced tools, such as automatically generating patient discharge notes instructing how many pills the patient has to take. Automated pluralization may use regular expressions of the ending of the nouns, alike for English (Conway, 1998), or extensive linguistic resources to devise pluralization, as for German (Nakisa and Hahn, 1996).

Pluralization for Runyankore and isiZulu, members of the Bantu language family, showed that neither of the above approaches were feasible. It required a combined morphology with syntactic and semantic approach, where the latter was covered by using a combination of the noun and noun class to pluralize, along with a few similar refinements of Meinhof's noun class definition (Byamugisha et al., 2016). A surprising observation was that a similar approach could be used despite the regions where these languages are spoken being thousands of kilometers apart (Uganda and South Africa) and they belong to different Guthrie zones (Guthrie, 1948; Maho, 2009). This suggests there may be sufficient similarities not only for closely related languages, but also for languages in other Guthrie zones.

We thus aim to investigate the generalizability of pluralization for Bantu languages, specifically those with an agglutinating morphology. The following questions guided this investigation:

Q1: Is a combined morphology with syntactic and semantic approach needed to pluralize nouns in other agglutinating Bantu languages as well?

Proceedings of the 27th International Conference on Computational Linguistics, pages 2633–2643
Santa Fe, New Mexico, USA, August 20-26, 2018.

Q2: Are the limitations identified in Meinhof's noun class system for automated pluralization in Runyankore and isiZulu (Byamugisha et al., 2016) present in other agglutinating Bantu languages? If so, how can Meinhof's noun class system be revised systematically?

Q3: Are the same 'exceptions' to pluralization found in other agglutinating Bantu languages, and if so, are the solutions used for isiZulu and Runyankore able to pluralize the same exceptions in these languages?

Q4: In attempting to develop a generic pluralizer, how can the language-specific pluralizers be generalized whilst maintaining roughly the same accuracy?

In order to answer these questions, we investigated the generalizability of the Runyankore and isiZulu approach to agglutinating Bantu languages, both within the same Guthrie zone and across zones. Five agglutinating Bantu languages were selected: two of these, Luganda and isiXhosa, belong to the same zone as Runyankore and isiZulu, respectively, and the rest belong to other zones, being Kinyarwanda (J.10), Kikuyu (E.10), and chiShona (S.10) (Maho, 2009).

We first examined their grammatical structures (Déchaine et al., 2013; Zentz, 2016; Mberi, 2002; Jeon et al., 2015; Baertlein and Ssekitto, 2014; Kimenyi, 2004; Taraldsen, 2010), which revealed that a combined morphology with syntactic and semantic approach, where the noun and its noun class are provided as input, is also required for pluralization. This is principally because there are nouns that have the same prefix but are pluralized differently, due to belonging to different classes. Further, Meinhof's 1948 noun class tables of singular/plural pairs is underspecified as to which option is used when. Therefore, we have revised the tables so as to achieve a deterministic selection during pluralization. Further, the same exceptions to pluralization identified for Runyankore and isiZulu (Byamugisha et al., 2016) were present in each of the five languages, and the same solutions were also applicable. Finally, we developed a pluralizer that can pluralize nouns in agglutinating Bantu languages, with generic pluralization rules but language-specific resources, which achieved similar accuracies as Byamugisha et al., (2016)'s language-specific pluralizers.

This paper is arranged as follows. We briefly introduce relevant aspects of Bantu languages in Section 2. Section 3 justifies why Meinhof's noun class system should be redefined, and describes the revised noun class tables. The generic pluralizer is presented in Section 4 and it is evaluated in Section 5. We discuss in Section 6 and conclude in Section 7.

2 Basics of Bantu Languages

Bantu languages are a group of languages indigenous to Africa, from the south of Nigeria, covering most of central, east, and southern Africa (Nurse and Philippson, 2003). There are Bantu language speaking communities in 27 of the continent's 54 countries, about 240 million speakers, and number of languages ranges from 300 to 680 (depending on the criteria used (Nurse and Philippson, 2003). They are agglutinating, have concordial agreement systems, and all nouns are assigned to a noun class (NC). There are over 20 noun classes, although most languages use fewer (Nurse and Philippson, 2003; Mohlala, 2003). A NC is characterized by a class prefix, a specific singular/plural pairing, and agreement with other constituents (Zentz, 2016; Nurse and Philippson, 2003). The semantic generalizations of the type of nouns in each NC are shown in Table 1 (Keet and Khumalo, 2014; Baertlein and Ssekitto, 2014; Kimenyi, 2004; Jeon et al., 2015; Zentz, 2016; Taraldsen, 2010; Mohlala, 2003).

Bantu languages are conventionally categorized into 16 zones, referred to as Guthrie zones, which are further subdivided into decades; for example, zone J.10 contains languages from J.11 to J.19 (Nurse and Philippson, 2003; Maho, 2009). We selected five languages from five different zones, and their grammars formed the basis of whether a generic pluralizer could be developed. These 5 languages were selected because: (1) they are widely used in their native countries; (2) their linguistics are taught at university level; and (3) they are actively researched, which provides current documentation. Table 2 summarizes the findings into the grammatical features of interest.

The NC classification for each of these languages is based on Meinhof's NC definition, which places the plural form of a noun in a different class from its singular (Nurse and Philippson, 2003; Mohlala, 2003). This is deemed as a standard for defining NCs among linguists, as it facilitates cross-language

Noun Class	Description of nouns typically found in those classes
1 and 2	People and kinship
3 and 4	Plants, nature, and some parts of the body
5 and 6	Fruits, liquids, some parts of the body, loan words, and paired things
7 and 8	Inanimate objects
9 and 10	Loan words, tools, and animals
11	Long thin stringy objects, languages, and inanimate objects
12 and 13	Diminutives
14	Abstract concepts
15	Infinitive nouns
16, 17, and 18	Locative classes
19	Diminutives
20, 21, and 22	Augmentative
23	Locative class

Table 1: Classification of nouns into noun classes.

Language	Guthrie Zone	Number of NCs	Phonological Conditioning
chiShona	S.10	20	Yes, with VC, VE, and VH
isiXhosa	S.40	15	Yes, with VC and VE
isiZulu	S.40	17	Yes, with VC and VE
Kikuyu	E.50	17	Yes
Kinyarwanda	J.60	16	Yes, with VC and VE
Luganda	J.10	21	Yes, with VC and VE
Runyankore	J.10	20	Yes, with VC, VE and VC for VH

Table 2: Relevant features of the selected languages; VC: vowel coalescence; VE: vowel elision; VH: vowel harmony.

comparisons and use (Chavula and Keet, 2014). Further, as noted, the NC determines the agreement markers on the associated lexical categories such as adjectives and verbs. Moreover, there are NCs with the same class prefix but belonging to different NCs (Déchaine et al., 2013; Zentz, 2016; Mberi, 2002; Jeon et al., 2015; Baertlein and Ssekitto, 2014; Kimenyi, 2004; Taraldsen, 2010), which are highlighted in Table 3.

We identified grammatical differences among these languages, which makes the use of the exact same pluralizer impossible: (1) each language has its own lexicon of prefixes, (2) they do not have the same number of NCs; and (3) the rules for phonological conditioning and verbal morphology are language-specific (Mberi, 2002; Jeon et al., 2015; Ferrari-Bridgers, 2009; Muhirwe, 2007; Kimenyi, 2004; Taraldsen, 2010). Regarding the latter, examples include: Runyankore uses vowel coalescence and elision next to a nasal compound (such as '*nk*'), while Luganda uses vowel elision; and in Runyankore, Luganda, and isiZulu, *-u-* + vowel may become *w* (such as *ulu-* + *-azi* = *ulwazi* 'knowledge' in isiZulu), which is not the case in Kikuyu and chiShona. Nonetheless, the noted grammatical similarities among these languages are expected to be sufficient for the generalizability of the pluralization approach.

However, Meinhof's NC definition (Table 3) has several limitations when applied to computational tasks, which impede pluralization and generalizability:

1. Meinhof's NC tables lack the specification for some rules and when to apply them, especially for lesser known cases. In isiZulu, for example, the standard NC1 and NC3 singular prefix is *um*, with its plural prefix *aba* in NC2 and *imi* for NC4 (Byamugisha et al., 2016), yet for NC1 nouns whose stem commences with *a, e, i*, or *o* the plural prefix *ab* is used instead (Byamugisha et al., 2016).
2. The rules for the pluralization of some nouns deviate from those according to Meinhof's NC definition, such as mass nouns (which are neither pluralized no singularized) and prefix exceptions

NC	chiShona	isiXhosa	isiZulu	Kikuyu	Kinyarwanda	Luganda	Runyankore
1	**mu-**	**u-m-**	**u-m-/u-mu-**	**mu-**	**u-mu-**	**o-mu-**	**o-mu-**
2	va-	a-ba-	a-ba-/a-b-	a-	a-ba-	a-ba-	a-ba-
1a	**-**	**u-**	**u-**	**-**	N/A	N/A	N/A
2a	vana-	oo-	o-	**-**	N/A	N/A	N/A
2b	a-	N/A	N/A	N/A	N/A	N/A	N/A
3a	N/A	N/A	**u-**	N/A	N/A	N/A	N/A
2a	N/A	N/A	o-	N/A	N/A	N/A	N/A
3	**mu-**	**u-m-**	**u-m-/u-mu-**	**mu-**	**u-mu-**	**o-mu-**	**o-mu-**
4	mi-	i-mi-	i-mi-	mi-	**i-mi-**	**e-mi-**	**e-mi-**
5	**-**	**i-**/i-li-	**i-**/i-li-	ri-	**i-**/i-ri-	**e-**/e-li-	e-i-/e-ri-
6	ma-	a-ma-	a-ma-	ma-	a-ma-	a-ma-	a-ma-
7	chi-	i-si-	i-si-	ki-/gi-	i-ki-/i-cy-/i-gi-	e-ki-	e-ki-
8	zvi-	i-zi-	i-zi-	ci-/i-	i-bi-	e-bi-	e-bi-
9	**n-**	**i-**/i-n-	**i-**/i-n-	**n-**	**i-/i-n-/i-nz-**	**e-n-/e-m-**	**e-n-/e-m-**
10	**n-**/dzi-	ii-/i-zin-	i-zi-/i-zin-	**n-**	**i-/i-n-/i-nz-**	**e-n-/e-m-**	**e-n-/e-m-**
9a	N/A	N/A	**i-**	**-**	N/A	N/A	N/A
10a	N/A	N/A	N/A	**-**	N/A	N/A	N/A
6	N/A	N/A	a-ma-	N/A	N/A	N/A	N/A
11	ru-	**u-**/u-lu-	**u-**/u-lu-	ru-	u-ru-	o-lu-	o-ru-
10	**n-**/dzi-	ii-/i-zin-	i-zi-/i-zin-	**n-**	**i-/i-n-/i-nz-**	**e-n-/e-m-**	**e-n-/e-m-**
12	ka-	N/A	N/A	ka-/ga-	a-ka-/a-ga-	a-ka-	a-ka-
13	tu-	N/A	N/A	tu-	u-tu-	o-tu-	o-tu-
14	u-	u-bu-	u-bu-	**-**	u-bu-	o-bu-	o-bu-
6	ma-	N/A	N/A	N/A	N/A	N/A	N/A
15	**ku-**	u-ku-	u-ku-	**ku-/gu-**	u-ku-/u-du-	**o-ku-**	**o-ku-**
6	N/A	N/A	N/A	ma-	a-ma-	a-ma-	a-ma-
16	ha-	N/A	N/A	ha-	a-ha-	a-ha-	a-ha-
17	**ku-**	N/A	ku-	**ku-/gu-**	N/A	**o-ku-**	**o-ku-**
18	**mu-**	N/A	N/A	N/A	N/A	**o-mu-**	**o-mu-**
20	N/A	N/A	N/A	N/A	N/A	o-gu-	o-gu-
21	N/A	N/A	N/A	N/A	N/A	a-ga-	a-ga-
21	zi-	N/A	N/A	N/A	N/A	N/A	N/A
6	ma-	N/A	N/A	N/A	N/A	N/A	N/A
23	N/A	N/A	N/A	N/A	N/A	**e-**	N/A

Table 3: Standard NC classification by singular/plural pair (first line and the [2nd/3rd] line, respectively), highlighting the same prefixes across more than one NC for a language. The dashes between the letters in the prefix illustrate separation between the initial vowel (augment) and prefix; '-': empty prefix; 'N/A': the NC is not present in that language (none use NC19 or NC22).

(whose plural NC is different from the standard pairings). Byamugisha et al., (2016) handled this by placing an 'm' on the NC, while prefix exceptions were directly associated with their plurals.

Therefore, we investigated revising Meinhof's NC definition for other agglutinating Bantu languages within and across Guthrie zones.

3 Reclassification of Meinhof's 1948 Noun Class Specification

We investigated the presence of the limitations in Meinhof's NC definition by identifying whether a deterministic output could be achieved based on the standard NCs, and investigating whether the same limitations for mass nouns and prefix exceptions identified for Runyankore and isiZulu exist in the other five selected languages. We found that, there are also NCs that have more than one possible class prefix; e.g., NC7 has three prefixes in Kinyarwanda: *iki* pluralized with *ibi*, *icy* pluralized with *iby*, and *igi* pluralized with *ibi* and NC5 and NC9 both have *i* as a prefix. This makes the rules on pluralization non-deterministic. Additionally, each of these languages classified some mass nouns in singular NCs, which would result in them being incorrectly subjected to pluralization. We thus found it necessary to refine Meinhof's NC definition in order to ensure a deterministic output during pluralization.

There are several alternatives in which this revision can be done: (1) morphologically (and syntactically for compound nouns), where each different class prefix is reclassified as a subdivision of the general class, with the subdivision indicated with roman numerals, such as, NC5, NC5i, etc.; (2) semantically, which would account for special categories of nouns such as mass nouns, singular-only nouns, and prefix

NC	chiShona	isiXhosa	isiZulu	Kikuyu	Kinyarwanda	Luganda	Runyankore
1	mu-	u-m-	u-mu	mu-	u-mu-	o-mu-	o-mu-
2	va-	a-ba-	a-ba-	a-	a-ba-	a-ba-	a-ba-
1a	-	u-	u-	-	N/A	N/A	N/A
2a	vana-	oo-	o-	-	N/A	N/A	N/A
2b	a-	N/A	N/A	N/A	N/A	N/A	N/A
1i	mw-	N/A	u-m-	mw-	u-mw-	o-mw-	o-mw-
2i	v-	N/A	N/A	N/A	a-b-	N/A	N/A
2	N/A	N/A	a-ba-	N/A	N/A	a-ba-	a-ba-
2a	N/A	N/A	N/A	-	N/A	N/A	N/A
1ii	N/A	N/A	u-m-	m-	-	-	-
2i	N/A	N/A	a-b-	N/A	N/A	N/A	N/A
2ii	N/A	N/A	N/A	N/A	ba-	ba-	ba-
2	N/A	N/A	N/A	a-	N/A	N/A	N/A
3a	N/A	N/A	u-	N/A	N/A	N/A	N/A
2a	N/A	N/A	o-	N/A	N/A	N/A	N/A
3	mu-	u-m-	u-mu-	mu-	u-mu-	o-mu-	o-mu-
4	mi-	i-mi-	i-mi-	mi-	i-mi-	e-mi-	e-mi-
3i	mw-	N/A	u-m-	m-	u-mw-	o-mw-	o-mw-
4	mi-	N/A	i-mi-	mi-	N/A	N/A	N/A
4i	N/A	N/A	N/A	N/A	i-my-	e-my-	e-my-
5	-	i-	i-	ri-	i-	e-	e-i-
6	ma-	a-ma-	a-ma-	ma-	a-ma-	a-ma-	a-ma-
5i	N/A	i-li-	i-li-	i-	i-ri-	e-li-	e-ri-
6	N/A	a-ma-	a-ma-	ma-	a-ma-	a-ma-	a-ma-
7	chi-	i-si-	i-si-	ki-	i-ki-	e-ki-	e-ki-
8	zvi-	i-zi-	i-zi-	i-	i-bi-	e-bi-	e-bi-
7i	N/A	i-s-	i-s-	gi-	i-cy-	e-ky-	e-ky-
8i	N/A	i-z-	i-z-	N/A	N/A	e-by-	e-by-
8	N/A	N/A	N/A	i-	i-bi-	N/A	N/A
7ii	N/A	N/A	N/A	N/A	i-gi-	N/A	N/A
8	N/A	N/A	N/A	N/A	i-bi-	N/A	N/A
9a	N/A	N/A	i-	-	N/A	N/A	N/A
10a	N/A	N/A	N/A	-	N/A	N/A	N/A
6	N/A	N/A	a-ma-	N/A	N/A	N/A	N/A
9	n-	i-	i-	n-	i-	e-n-	e-n-
10	dzi-	ii-	i-zi-	n-	i-	e-n-	e-n-
9i	N/A	i-n-	i-n-	N/A	i-n-	e-m-	e-m-
10i	N/A	i-zin-	i-zin-	N/A	i-n-	e-m-	e-m-
9ii	-	N/A	N/A	N/A	-	-	-
10ii	-	N/A	N/A	N/A	-	-	-
9iii	N/A	N/A	N/A	N/A	i-zn-	N/A	N/A
10iii	N/A	N/A	N/A	N/A	i-zn-	N/A	N/A
11	ru-	u-	u-	ru-	u-ru-	o-lu-	o-ru-
10	N/A	ii-	i-zi-	n-	N/A	e-n-	e-n-
10i	N/A	N/A	N/A	N/A	i-n-	N/A	N/A
6	ma-	N/A	N/A	N/A	N/A	N/A	N/A
11i	N/A	u-lu-	u-lu-	N/A	u-rw-	o-lw-	o-rw-
10	N/A	ii-	i-zi-	N/A	i-	N/A	N/A
12	N/A	N/A	N/A	N/A	N/A	a-ka-	a-ka-
12	ka-	N/A	N/A	ka-	a-ka-	a-ka-	a-ka-
13	tu-	N/A	N/A	tu-	u-tu-	N/A	N/A
14	N/A	N/A	N/A	N/A	N/A	o-bu-	o-bu-
12i	N/A	N/A	N/A	ga-	a-ga-	a-k-	a-k-
13	N/A	N/A	N/A	tu-	N/A	N/A	N/A
13ii	N/A	N/A	N/A	N/A	u-du-	N/A	N/A
14i	N/A	N/A	N/A	N/A	N/A	o-bw-	o-bw-
13	N/A	N/A	N/A	N/A	N/A	o-tu-	o-tu-
13i	N/A	N/A	N/A	N/A	u-tw-	o-tw-	o-tw-
14	u-	u-bu-	u-bu-	-	u-bu-	o-bu-	o-bu-
6	ma-	N/A	N/A	ma-	N/A	N/A	N/A
14i	N/A	N/A	N/A	N/A	u-bw-	o-bw-	o-bw-
15	ku-	u-ku-	u-ku-	ku-	u-ku-	o-ku-	o-ku-
6	N/A	N/A	N/A	ma-	a-ma-	a-ma-	a-ma-
15i	N/A	N/A	u-kw-	gu-	u-kw-	o-kw-	o-kw-
6	N/A	N/A	N/A	ma-	a-ma-	a-ma-	a-ma-
15ii	N/A	N/A	N/A	N/A	u-gu-	N/A	N/A
6	N/A	N/A	N/A	N/A	a-ma-	N/A	N/A
16	ha-	N/A	N/A	ha-	a-ha-	a-ha-	a-ha-
17	ku-	N/A	ku-	ku-/gu-	N/A	o-ku-	o-ku-
18	mu-	N/A	N/A	N/A	N/A	o-mu-	o-mu-
20	N/A	N/A	N/A	N/A	N/A	o-gu-	o-gu-
21	N/A	N/A	N/A	N/A	N/A	a-ga-	a-ga-
20i	N/A	N/A	N/A	N/A	N/A	o-gw-	o-gw-
21	N/A	N/A	N/A	N/A	N/A	a-ga-	a-ga-
21	zi-	N/A	N/A	N/A	N/A	N/A	N/A
6	ma-	N/A	N/A	N/A	N/A	N/A	N/A
23	N/A	N/A	N/A	N/A	N/A	e	N/A

Table 4: Reclassified noun classes. The first line in each pairing is the singular and the other line(s) its plural class (if more than one line is paired, the one with the prefix is applicable, or it is N/A); '-': empty prefix; 'N/A': NC is not present in that language.

exceptions; and (3) revising the entire NC classification to account for all fine-grained details on class prefixes, concords, and agreement.

Options (2) and (3) require extensive linguistic analysis of each language, such as that presented by Taraldsen (2010) for Nguni languages. Such analyses are beyond the scope of our work and left to linguists to complete. For immediate computational needs, we thus applied a morphological with syntactic approach to the reclassification. The details on concordial agreement (subject, object, adjective, and possessive concords) are still maintained in our reclassification, enabling its application in other generation processes (such as verb conjugation) beyond pluralization. Our reclassification ensures that there is always a one-to-one singular-plural mapping, hence, ensuring a deterministic outcome. Table 4 presents the reclassified NC definition; the new classes added are those with one or more 'i' after the number.

This resultant reclassified NC definition demonstrates the pluralization rules that Meinhof's NC definition does not. For example, in Luganda, NC 1ii contains nouns that denote kinship, such as mother (*maama*) and father (*taata*); they are pluralized in NC 2ii, ***ba****maama* and ***ba****taata* respectively.

More generally, this reclassification of the NC definition clarifies the rules for pluralization, resulting in a deterministic output. We used similar prefix features among languages to guide our reclassification, based on the following:

(1) Most NCs whose prefix ends in *u* (NCs 1, 3, 11, 13, 14, 17, and 20) have nouns that have *w* instead; this then becomes the new NC, as is the case with NCs 1i, 3i, 11i, 13i, 14i, 17i, and 20i;

(2) Some NCs whose prefix ends in *i* (NCs 7 and 8) also have nouns whose prefix ends in *y*, and this forms the new NC (7i and 8i);

(3) Nouns without prefixes in the singular and/or plural, which were previously not classified in Meinhof's NC definition, are placed in a new NC with '*ii*' such as NCs 1ii, 9ii, and 10ii; NC 2ii is the pairing for 1ii; and

(4) Languages with NCs with multiple prefixes but that do not fit the above criteria, use the standard prefix in the original class, but the alternative prefix in the new class; such as NCs 7 (*ki-*), 7i (*gi-*), 12 (*ka-*), 12i (*ga-*), 15 (*ku-*), and 15i (*gu-*) in Kikuyu.

4 Implementation of Generic Pluralizer

The decision of how to implement the generic pluralizer was based on research into three approaches for multilingual noun pluralization: Grammatical Framework (GF) (Ranta, 2011), SimpleNLG (Gatt and Reiter, 2009), and the pluralizer in Byamugisha et al., (2016). GF and SimpleNLG support multiple languages from different language families, while the pluralizers by Byamugisha et al., (2016) are very accurate for two agglutinating Bantu languages, isiZulu and Runyankore.

Grammatical Framework (GF) is based on type theory and functional programming; it uses distinct functions for singulars and plurals, and linearization rules to handle the language-specific inflectional forms of nouns during pluralization (Ranta, 2011; Ranta et al., 2015). It requires a large resource grammar. An attempt to develop such a resource grammar for Kiswahili (regarded as a Bantu language, with the same NC system that determines noun pluralization, and not as computationally under-resourced) only introduced some parameter types (Ngángá, 2011), but still has no resource grammar. There is thus no existing GF resource grammar for any Bantu language.

SimpleNLG is an NLG realizer with a Java API, where pluralization is achieved through the language-specific lexicon saved as an XML file (Gatt and Reiter, 2009). The lexicon contains 'words' in their base form, as well as attributes for their 'category' (such as noun, verb, adjective, etc.), 'id' for a unique identifier, and 'plural' (Mohamood, 2016). The plural is obtained by creating an inflected word element around the base form, adding appropriate features to it (such as 'plural' or 'past participle'), and then realizing it (Mahamood, 2015).

Byamugisha et al., (2016) developed two separate pluralizers, for isiZulu and Runyankore. They are based on a combined semantic and syntactic approach with morphology, where the semantics of the noun are determined by its NC, and the appropriate plural prefix replaces the singular prefix. This approach requires that the NC, together with the noun, is passed to the pluralizer in order to obtain the correct

plural. The importance of including the NC with the noun during pluralization can be seen in the increase in accuracy for especially the isiZulu nouns, which improved from about 50% correct pluralization for the noun-only pluralizer to about 85% by passing on the NC with the noun, while it increased it from 88% to 92% for Runyankore (Byamugisha et al., 2016). This effect is due to more or less NCs having the same prefix (recall Table 3).

The GF approach requires the definition of a large resource grammar for each language (Ranta, 2011). The attempt to develop one for Kiswahili showed the difficulties with adopting the GF parameter types to fit the grammar of Bantu languages; for example, the NC was represented as the Gender type (Ngángá, 2011). As no resource grammar was derived from these definitions, there is no way to know whether this is sufficient for pluralization and other generation tasks for Bantu languages. For SimpleNLG, the use of a lexicon for pluralization requires that both the singular and plural forms be stated for each word in the lexicon (Gatt and Reiter, 2009; Mohamood, 2016). We found this approach undesirable, as our aim is to generate the plurals, not manually declare them for each noun.

We therefore decided to develop the generic pluralizer based on the approach by Byamugisha et al., (2016), because the NC reclassification in Table 4 provides one singular prefix with one plural prefix for all these languages, and is thus generalizable to other agglutinating Bantu languages, and the need for the NC-based pluralization of compound nouns in (Byamugisha et al., 2016) also holds for other agglutinating Bantu languages.

Additionally, Byamugisha et al., (2016) identified several nouns that did not conform to the NC rules on pluralization, and referred to them as 'exceptions'. These exceptions were handled in three ways: (1) mass nouns (whose NC was marked with an '**m**'); (2) prefix exceptions, which were placed in a separate look-up file with their correct plural; and (3) singular-only nouns that were also placed in a separate look-up file. These same approaches were applied to the five selected languages, and were able to correctly handle these exceptions.

The design of the generic pluralizer thus comprised the following core rules:

(1) Providing a singular noun and its NC, so as to obtain its plural prefix based on its NC;
(2) Ensuring that mass nouns and singular-only nouns are not pluralized;
(3) Obtaining the plural of the main noun, and its plural agreement, during the pluralization of compound nouns (excluding those with adjectives); and
(4) Correctly pluralizing nouns whose pluralization deviates from the NC pairings.

The language-specific NC details are made available to the pluralizer through text files, one for each language, which are read into the program, and availed to the pluralizer as variables, as illustrated in the architecture in Figure 1. The selected language and noun(s) to be pluralized are entered through the **UI**. Through **Languages**, the text files that contain the reclassified NCs, the singular/plural pairings, as well as prefix exceptions are loaded into **Pluralizer** as variables. The following then takes place: 1) If the noun is a mass noun (with an '**m**' on the NC), then it is returned without pluralization; 2) If the noun is found in **Exceptions**, then it is pluralized according to the exceptions prefix instead of the standard one; and 3) If the noun's NC is the same as that in **NCS** (the noun class system), then it is pluralized accordingly.

The generic pluralizer was implemented as a Java application. It, as well as the datasets tested for all languages are available at `https://github.com/runyankorenlg/Generic-Pluralizer`.

5 Evaluation of Generic Pluralizer

We carried out two types of evaluation for the generic pluralizer: verification and validation. The aim of verification is to ensure that all NCs were accounted for correctly, based on what literature states should be in each NC. The validation was aimed at comparing the output of the generic pluralizer to the language-specific ones.

5.1 Materials and Methods

The evaluation was carried out in two phases that used two test-sets, SetI, for verification (internal) testing, and SetC, for coverage testing. SetI was an English wordlist compiled based on what should

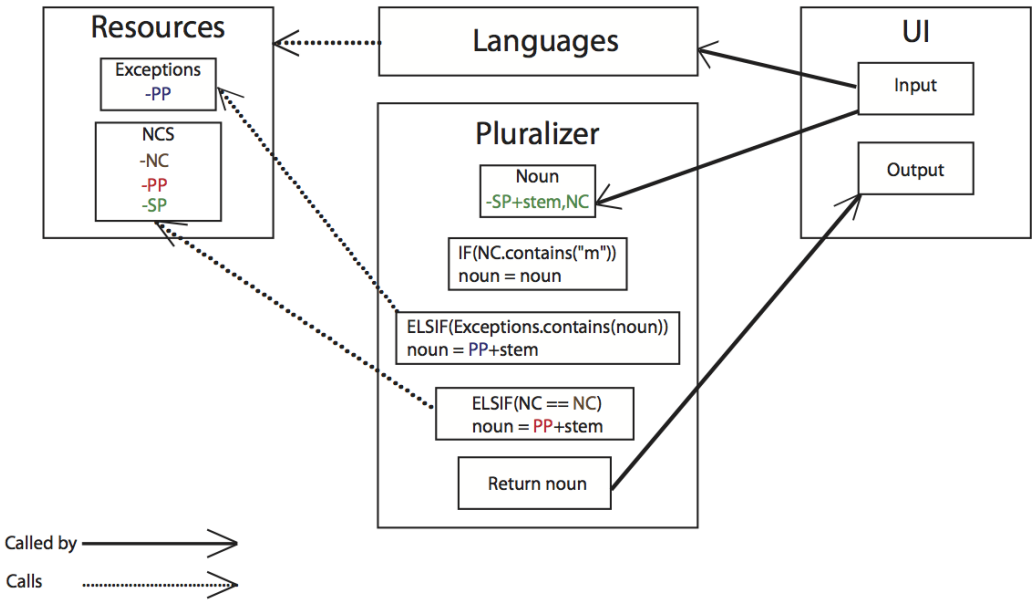

Figure 1: The architecture of the pluralizer; PP: plural prefix; SP': singular prefix; NC: noun class.

generally be in each NC (recall Table 1), and contained 81 nouns such that it includes mass nouns, abstract concepts, and compound nouns. This was done because, for several languages, no computational resources, such as dictionaries, were readily available, from which singular nouns otherwise could have been extracted randomly. The English wordlist was translated, and each noun had its NC; this formed SetI for each language. While not all nouns in the English wordlist were translatable in all languages, all NCs were accounted for in the resulting wordlists.

SetC was the same English wordlist used as Set1R in Byamugisha et al., (2016), and consists of 101 words informed by multiple ontologies. This was also translated by native speakers to chiShona, isiXhosa, Kikuyu, and Luganda, with each noun having its NC. The original Runyankore and isiZulu test-sets from (Byamugisha et al., 2016) were used, but their NCs were updated according to the new NC specification in Table 4.

The metric for evaluation is accuracy, determined as the percentage correct plurals over the test set.

5.2 Results and Analysis

Table 5 shows the accuracy of the generic pluralizer for each language, for SetI and SetC. The accuracy is based on the total number of translatable nouns. The reason for obtaining less than 100% is due to the absence of phonological conditioning. In chiShona for example, ***g**umbo* 'leg' is pluralized as *ma**k**umbo* instead of the generated *ma**g**umbo*. As explained in Section 2, the rules for phonological conditioning are language-specific, and as such could not be generalized. We propose that a separate language-specific phonological conditioning module be developed for each language, and added to that language's resources folder; we leave this for future work. Further, for isiZulu, the lower accuracy in the generic pluralizer is due to it only handling compound nouns with the possessive concord, while the isiZulu pluralizer in Byamugisha et al., (2016) additionally handles compound nouns with adjectives using the adjective concord.

Prefix Exceptions The true exceptions identified in Runyankore and isiZulu are those where pluralization does not correspond to the singular/plural pairings and there are no rules among them. They were thus placed in a separate look-up file with their corresponding plurals. In the generic pluralizer, a separate exceptions look-up file of the same structure was also created for each language (see Figure 1). Prefix exceptions were identified to be quite rare in each language; examples include: the plural of ***i**mbwa* (dog)

	SetI		SetC	
Language	**Translatable Nouns**	**Accuracy (%)**	**Translatable Nouns**	**Accuracy (%)**
chiShona	68	94.12%	74	94.59%
isiXhosa	74	97.30%	83	100%
isiZulu	71	100%	101	97.03%
Kikuyu	77	96.10%	76	94.74%
Kinyarwanda	70	97.14%	-	-
Luganda	75	93.33%	78	97.44%
Runyankore	81	93.83%	88	96.59%

Table 5: Accuracy of the generic pluralizer.

in chiShona as ***i****mbwa* instead of ***dzi****mbwa*; or *dagetari* (doctor) in Kikuyu, pluralized as ***ma****dagetari*, instead of *dagetari*; or ***i****nzu* (house) in Kinyarwanda, whose plural is ***ama****nzu* instead of ***i****nzu*.

Singular-only Nouns Singular-only nouns in isiZulu and Runyankore resulted from nouns for abstract concepts, such as 'greed' and 'thirst', found in several NCs, or infinitive nouns in NCs 15 and 15i. Abstract nouns belonging to plural NCs (for example, ***obu****rungi* 'beauty', Runyankore) were handled in the pluralization algorithm by returning them as they are. Those belonging to singular NCs were placed in a separate look-up file. Similarly, for the five selected languages, abstract concepts can either be found in singular or plural NCs; examples of the singular-only nouns identified in the wordlists are: *moto* 'fire' and *zuva* 'sun' (chiShona), *ubuhle* 'beauty' and *ubuhlobo* 'friendship' in isiXhosa, and *omululu* 'greed' and *enyoonta* 'thirst' in Luganda. The same algorithmic solution was used for those in plural NCs, while a look-up file was used for those in singular NCs. Both solutions were found to be sufficient to handle this class of exceptions.

Mass Nouns Mass nouns that belong to singular NCs are marked with an 'm', so that they can be identified and not be pluralized according to the singular/plural pairings (Byamugisha et al., 2016). This same rule was found to adequately cater for mass nouns that belonged to singular NCs; for example, *doro* (NC5, chiShona) and *umdiliya-omfaxangiweyo* (NC3, isiXhosa) 'wine', and *iria* 'milk' (NC5, Kikuyu). Mass nouns that belong to plural NCs, such as *mae* (NC6, Kikuyu), *mvura* (NC10ii, chiShona), and *amanzi* (NC6, isiXhosa) 'water' are handled algorithmically, as described in the previous section.

Compound Nouns From the translated wordlists, we identified that the main noun of a compound noun is placed before the modifier noun. When pluralizing compound nouns in isiZulu and Runyankore, the main noun was first pluralized, and the plural agreement for the modifier noun was obtained and used to associate with the modifier noun, as in (Byamugisha et al., 2016).

We further found that, similar to isiZulu and Runyankore, phonological conditioning was also required in chiShona, isiXhosa, Kinyarwanda, and Luganda; only Kikuyu does not require phonological conditioning. However, as explained in Section 2, the rules for phonological conditioning are language-specific. We omit its detail here due to space limitations.

6 Discussion

Our investigation into a generic pluralizer for agglutinating Bantu languages has shown that the approach taken by Byamugisha et al., (2016) was applicable to other Bantu languages not only within but also across Guthrie zones. For both the intra-zone and inter-zone investigations, the same approach was found to result in 100% accuracy in some datasets (recall Table 5). Moreover, the investigation into the generalizability offered a systematic way to revise the (outdated) Meinhof NC system into one that is more precise and usable for computational purposes. It clarifies the singular/plural pair rules for pluralization, resulting in a deterministic output with a higher accuracy.

Although the language-specific pluralizers achieved 100% on SetC in Byamugisha et al., (2016), the errors in the generic pluralizer were mainly found to originate from the lack of phonological conditioning. Generally, phonological conditioning is language-specific, but there may be some cases for

generalizability. For example, isiXhosa and isiZulu both have $a + i = e$ and Kinyarwanda, Luganda, and Runyankore, all have $a + vowel = ' '$. However, further research is required into under what circumstances such rules are true, in order to prevent overgeneralization. Further, we intend to extend the generic pluralizer to compound nouns with adjectives using the NC-based adjective concord, as was done in the isiZulu pluralizer (Byamugisha et al., 2016).

The processing of the resources required to add a new language to the pluralizer is quite considerable, with text files required for the noun class system (NC number, class prefixes, and genitive), NC pairings, as well as singular-only nouns and prefix exceptions if known. While the last two, singular-only nouns and prefix exceptions, can be added as errors are identified, the NC system and pairings are required for the pluralization to be done. It may be useful to use the revised NC table to speed up automated annotation of nouns, which then can facilitate resource development for pluralization. This amount of effort is still considerably less than that required by other multilingual noun pluralization implementations, notably, GF (Ranta et al., 2015) and SimpleNLG (Gatt and Reiter, 2009).

The criteria for the revised noun class table, as described in Section 3, can be used by other researchers in computational linguistics for other Bantu languages, who will undoubtably need to revise Meinhof's NC definition, especially when working with multiple languages.

7 Conclusion

We have presented a novel generic pluralizer to pluralize nouns of agglutinating Bantu languages. This pluralizer was based on a combined morphology with syntactic and semantic approach and handling of regular exceptions. The rules for pluralization are based on the classification of nouns into noun classes (NCs). To ensure deterministic outputs, i.e., one plural option for each singular noun, the standard noun class system of Meinhof (1948) had to be redefined from 29 noun classes into 53 noun classes. The generic pluralizer can pluralize simple nouns and exceptions resulting from mass nouns, singular-only nouns, prefix exceptions, and compound nouns. It achieved over 93% for all test-sets, with the remaining errors arising from the need for phonological conditioning and handling compound nouns with adjectives. Future work will center around adding support for adjectives and investigating how to implement phonological conditioning.

Acknowledgements We would like to thank Christine Wangiru Mburu, Goreth Byamugisha, Juliet Naggawa, Naissa Umutoni, Ngoni Choga, and Zola Mahlaza for providing the Kikuyu, Luganda, Kinyarwanda, chiShona, and isiXhosa translations.

This work is based on the research supported by the Hasso Plattner Institute (HPI) Research School in CS4A at UCT and the National Research Foundation of South Africa (Grant Number 93397).

References

Elizabeth Baertlein and Martin Ssekitto. 2014. Luganda nouns inflectional morphology and tests. *Linguistic Portfolios*, 3.

Joan Byamugisha, C. Maria Keet, and Langa Khumalo. 2016. Pluralizing nouns in isizulu and related languages. In *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*, volume 9626, Konya, Turkey. Springer LNCS.

Catherine Chavula and C. Maria Keet. 2014. Is lemon sufficient for building multilingual ontologies for bantu languages? In *11th OWL Experiences and Directions Workshop (OWLED'14)*, volume 1265, pages 61–72, Riva del Garda, Italy.

D. M. Conway. 1998. An algorithmic approach to English pluralization. In C. Salzenberg, editor, *Proceedings of the Second Annual Perl Conference*. O'Reilly. San Jose, USA, 17-20 August, 1998.

Rose-Marie Déchaine, Raphaél Girard, Calisto Mudzingwa, and Martina Wiltschko. 2013. The internal syntax of shona class prefixes. *Language Sciences*, 43:18–46.

Franca Ferrari-Bridgers. 2009. Luganda verb morphology: A new analysis of the suffixes '-ye' and '-a', and their distribution across the indicative, subjunctive, and imperative moods. *Studies in African Linguistics*, 38(1).

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realization engine for practical applications. In *12th European Workshop on Natural Language Generation (ENLG'09)*, pages 90–93, Athens, Greece.

Malcolm Guthrie. 1948. *The Classification of the Bantu Languages*. Oxford University Press, London.

Lisa Jeon, Jessica Li, Samantha Mauney, Anaí Navarro, and Jonas Wittke. 2015. A basic sketch grammar of gíkúyú. *Rice Working Papers in Linguistics*, 6.

C. Maria Keet and Langa Khumalo. 2014. Towards verbalizing ontologies in isizulu. In *4th Workshop on Controlled Natural Languages (CNL'14)*, pages 78–89, Galway, Ireland.

Alex Kimenyi. 2004. Kinyarwanda morphology. In Geert Booij, Christian Lehmann, Joachim Mudgan, and Stavros Skopeteas, editors, *Morphology: An International Handbook for Inflection and Word Formation*, volume 17.2. De Gruyter.

Saad Mahamood. 2015. Simple nlg wiki: Section iv-lexicon. Accessed on 08/02/2018.

Jouni Filip Maho. 2009. Nugl online: The online version of the updated guthrie list, a referential classification of the bantu languages. http://goto.glocalnet.net/mahopapers/nuglonline.pdf.

Nhira Edgar Mberi. 2002. *The Categorical Status and Functions of Auxiliaries in Shona*. Ph.D. thesis, University of Zimbabwe.

Saad Mohamood. 2016. default-lexicon.xml. Accessed on 08/02/2018.

Linkie Mohlala. 2003. The bantu attribute noun class prefixes and their suffixal counterparts, with special reference to zulu. Master's thesis, University of Pretoria, Pretoria, South Africa.

Jackson Muhirwe. 2007. Computational analysis of kinyarwanda morphology: The morphological alternations. *International Journal of Computing and ICT Research*, 1(1):85–92.

Ramin Charles Nakisa and Ulrike Hahn. 1996. Where defaults don't help: the case of the German plural system. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pages 177–182. San Diego, USA, July 12-15, 1996.

Wanjiku and Ngángá. 2011. Swahili inflectional morphology for the grammatical framework. In *Human Language Technology for Development*, Alexandria, Egypt.

Derek Nurse and Gerard Philippson. 2003. *The Bantu Languages: Routledge Language Family Series 4*. Taylor and Francis Routledge, London.

Aarne Ranta, Yan Tian, and Haiyan Qiao. 2015. Chinese in the grammatical framework: Grammar, translation, and other applications. In *8th SIGHAN Workshop on Chinese Language Processing (SIGHAN'08)*, pages 100–109, Beijing, China.

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. Center for the Study of Language and Information (CSLI) Publications, Stanford.

Lenhart Schubert. 2015. Computational linguistics. In N. Edward Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, California, USA. Available from: https://plato.stanford.edu/archives/spr2015/entries/computational-linguistics/.

Knut Tarald Taraldsen. 2010. The nanosyntax of nguni noun class prefixes and concords. *Lingua*, 120(6):1522–1548.

Jason Zentz. 2016. *Forming Wh-Questions in Shona: A Comparative Bantu Perspective*. Ph.D. thesis, Yale University.