IsiZulu noun classification based on replicating the ensemble approach for Runyankore

Zola Mahlaza, Imaan Sayed, Alexander van der Leek, C. Maria Keet
Department of Computer Science
University of Cape Town
South Africa
{zmahlaza,mkeet}@cs.uct.ac.za, {SYDIMA002,VLKALE003}@myuct.ac.za

Abstract

A noun's class is a crucial component in NLP, because it governs agreement across the sentence in Niger Congo B (NCB) languages, among others. There is a lack of computational models for determining a noun's class owing to ill-documentation in most NCB languages. A promising approach by Byamugisha (2022) used a data-driven approach for Runyankore that combined syntax and semantics. The code and data are inaccessible however, and it remains to be seen whether it is suitable for other NCB languages. We solve the problem by reproducing Byamugisha's experiment, but then for isiZulu. We conducted this as two independent experiments, so that we also could subject it to a meta-analysis. Results showed that it was reproducible only in part, mainly due to imprecision in the original description, and the current impossibility to generate the same kind of source data set generated from an existing grammar. The different choices made in attempting to reproduce the pipeline as well as differences in choice of training and test data had a large effect on the eventual accuracy of noun class disambiguation but could produce an accuracy of 83%, in the same range as Runyankore.

1 Introduction

There has been a resurgence of research focusing on dialogue systems and interfaces in recent years, through efforts focusing on building virtual assistants (Dale, 2016), Large Language Model (LLM) powered systems that are used in industry (Padró and Saurí, 2024), and social robots (e.g., (Pu et al., 2018; van den Berghe et al., 2019)). This trend can also be seen in African Natural Language Processing (NLP) research; where there have been efforts to build ontology verbalisers (Keet et al., 2017; Byamugisha, 2019; Mahlaza and Keet, 2020), determine the requirements for building social robots (Keet, 2021), and data collection aimed at building speech recognition systems (Badenhorst et al., 2011). Within this context, noun class identification/classification has not received a lot of attention in several African languages despite its importance to these existing efforts. We demonstrate the importance of computational methods for noun class/category identification using an example of a hypothetical kitchen robot that communicates in isiZulu and French. Since French nouns are categorised by gender, one needs the capability to determine the gender of each noun; hence, the robot will need to be able to select or generate the right option from *la casserole* 'the$_{feminine}$ pot' vs. *le casserole* 'the$_{masculine}$ pot'. For a kitchen robot to respond in isiZulu, it may need to generate *amazambane a<u></u>thambile* 'the potatoes are soft' by choosing the appropriate prefix (underlined) using automatically detected class of the subject noun, instead of *amazambane li<u></u>thambile* 'the potatoes is soft' or any other prefix. For isiZulu, however, there is no computational noun class identification model that such a robot has to rely on.

Broadening the scope of languages, the work by (Byamugisha, 2022) produced the most promising results on this particular task, obtaining best results using a noun's morphological and semantic information to identify its class, and outperforming other approaches by a large margin. However, it focused only on three languages belonging to Guthrie Zone J—thousands of kilometres apart from Zone S that isiZulu is part of—and produced no open source and re-usable tool for use with other languages. Thus, there is a need to investigate the extent to which its findings hold for languages outside Guthrie Zone J.

The first, and key, task is to determine whether the procedure of (Byamugisha, 2022) is reproducible, serving as a candidate for a method also useful for other Niger-Congo B (NCB) languages, and if it is not, what the (in)surmountable impediments are. Second, if the approach is reproducible, then to ascertain whether the performance of noun

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 469–478
January 20, 2025.

class detection is in the same range of success as for Runyankore that Byamugisha (2022) focussed on. Third, the aim is to make as much data and software available as legally possible.

To investigate the former as a reproducibility study, two co-authors gave the task to two other authors without their knowledge, one of whom had the sole focus of teasing out reproducibility issues and the other as a 'quickly reproduce it' for it to serve as a baseline for designing a better algorithm to outperform said baseline. All used the most well-resourced among the low-resourced languages of South Africa, namely, isiZulu.

The best performing models, emanating from the 'quickly reproduce it' group, affirm Byamugisha (2022), since the model that combines syntax, morphology, and semantics has an accuracy of 83% while the prefix-only model has accuracy of 59%. However, the differences in the choice of training and test data, as well as the design of the module that combines the different linguistic aspects have a significant impact on performance.

In the remainder of the paper, we first describe pertinent details about the NCB noun class system (Section 2) and related work (Section 3). This is followed by the main sections on reproducibility experiment design (Section 4) and results and discussion (Section 5). We close with conclusions in Section 6.

2 Noun complexity in NCB languages

Prior to discussing computational approaches for noun class identification, we first present the complexity of NCB nouns. We do not limit our discussion of noun complexity to only languages in Guthrie Zone J or S since there are features shared across all NCB languages. In NCB languages, nouns are categorised into one of 23 classes. For instance, in isiZulu the noun *umuthi* 'tree' belongs to noun class 3 and its plural *imithi* 'trees' belongs to noun class 4. For each NCB language, nouns are not necessarily categorised into all the 23 noun classes since some of the classes may be obsolete, or the word is not classified as a noun, such as *ekhaya* 'at home' in isiZulu as the locative of *ikhaya* 'home' versus *eka* 'at home' in noun class 23 in Luganda.

We now turn to our running example of a kitchen/social robot to show the impact of the noun class on the text: unlike the French case where the robot can only generate two forms of the text, the noun class of the subject can lead to many more variations of the text; e.g., *zithambile*, *ihambile*, *bathambile*, *lithambile*, *athambile*, *sithambile*, *luthambile*, *buthambile*, and *kuthambile* (all meaning 'soft'), where the underlined parts denote the subject concord values that are dependent on the noun class of the subject. The task of retrieving a concord can be solved easily; however, the question of how to computationally identify the class of a noun still needs to be resolved.

While early linguistic literature argued that the noun classes are not a semantically motivated categorisation (Katamba, 2014), a number of authors have attempted to disprove that for Proto-Bantu[1] (Denny and Creider, 1986) or individual NCB languages such as Kikuyu (Burton and Kirk, 1976), Swahili (Contini-Morava, 1997), Shona (Palmer and Woodman, 2000), Sesotho, Setswana, isiZulu (Ngcobo, 2010, 2013), and Siswati (Demuth, 2000). Despite their proposals, however, there is no consensus on the matter and most authors rely on the rough guide as summarised in Table 1.

The semantic features shown in the Table 1 give the impression that to obtain an effective noun class identifier, one would need an ensemble of different classifiers tuned for the various semantic features or have word representations that capture, even partially, these semantic features. However, since several NCB languages are low-resourced and the above feature list is not exhaustive, there is still a need to investigate the extent to which semantics are required for computational models.

3 Existing computational models and their usability

To the best of our knowledge, there is limited work focusing on noun class identification for NCB languages. As such, we cast the net wider to also include the computational modelling of nouns.

Most existing computational models created for NCB nouns can be classified into: noun pluralization (e.g., (Byamugisha et al., 2016, 2018)), lemmatization, stemming and segmentation (e.g., (Nogwina, 2016; Mzamo et al., 2019; Moeng et al., 2021)), morphological analysis (e.g., (Pretorius and Bosch, 2009)), and POS tagging (e.g., (Eiselen and Puttkammer, 2014; Schlünz et al., 2016)). When considering work for languages from outside the African continent, we see that there are semantic and gender classifiers (Gagliardi et al., 2012; Falk

[1]The hypothetical ancestor of all NCB languages

Table 1: Generalisation of the semantics of the kinds of entities typically found in that noun class (NC). Examples are taken from isiZulu (classes 1-11, 14, 15), Chichewa (12,13,16-18), Hunde (19), Runyankore (20,21), and Luganda (22,23). (Source: adapted from (Byamugisha et al., 2018).)

NCs	Semantics (generalised)	Examples
1 2	People and kinship	*umfana* (nc1) 'boy' *abafana* (nc2) 'boys'
3 4	Plants, nature, some parts of the body	*umuthi* (nc3) 'tree' *imithi* (nc4) 'trees'
5 6	Fruits, liquids, parts of the body, loan words, paired things	*ijikijolo* 'raspberry' *amajikijolo* 'raspberries'
7 8	Inanimate objects	*isihlalo* 'chair' *izihlalo* 'chairs'
9 10	Loan words, tools, and animals	*indlovu* 'elephant' *izindlovu* 'elephants'
11 (10)	Long thin stringy objects, languages, inanimate objects	*ucingo* 'wire' *izingcingo* 'wires'
12 13	Diminutives	*kagalimoto* 'small car' *timagalimoto* 'small cars'
14	Abstract concepts	*ubuhle* 'beauty'
15	Infinitive nouns	*ukucula* 'to sing'
16 17 18	Locative classes	*pamsika* 'round the market' *kumsika* 'at the market' *mumsika* 'in the market'
19	Diminutives	*hyùndù* 'a little bit of porridge'
20 21 22	Augmentative and pejorative	*ogusajja* 'big ugly man' *agasajja* 'big ugly men' *gubwa* 'mutt' (pejorative of dog)
23	Locative class	*eka* 'at home'

et al., 2021) that were not created for the languages in question. As such, they do not consider a similar number of classes thus do not investigate the utility of combining syntax, morphology, and semantics. To the best of our knowledge, only Byamugisha (2022) has investigated how to build effective noun class identifiers for NCB languages.

Byamugisha (2022) created a three-module classifier, whose architecture is provided in Figure 1, using three datasets made up of 2803 Runyankore, 153 Luganda, and 70 Kinyarwanda nouns and split them 70% for training, 20% for validation, and 10% for testing. Following that, they then created three main model variants for identifying the class when provided with a noun (i.e., morphological, semantic, or morphological + syntax + semantic). The morphological model uses "morphological rules" to match a prefix to one of the noun classes, should the class's prefix be unique. The semantic model uses a pretrained FastText embedding model to embed the input noun and then relies on two alternative algorithms to identify between 10-1000 semantically related words. The properties of the neighbouring words are then used to predict noun class of the input noun. The third type is ensemble that starts by using the morphological model and in cases where it fails, it then feeds the noun to the semantic model. While the work concluded that the combination of morphology, syntax, and semantics yields the best performance, the generalisability of this finding across different Guthrie zones is unclear since the author only focused on Zone J. For instance, a key notion with the semantic model is the reliance on the uniqueness of grammatical information that is used to label each input's semantic neighbours. However, it is unclear whether such an assumption holds for NCB languages in other Guthrie zones as well. There is also no open access source code version of their tool that can be used to investigate generalisability across the different Guthrie zones for languages that have the appropriate datasets.

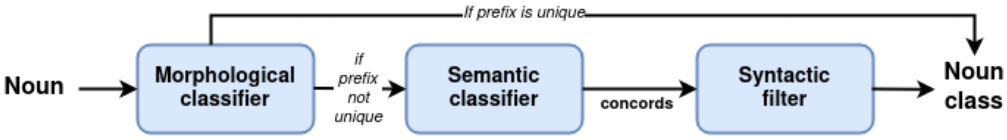

Figure 1: Architecture of Byamugisha's noun classifier that relies on a prefix, semantics, and syntax (Adapted from (Byamugisha, 2022)).

4 Experiment Design

While the work done by Byamugisha (2022) shows some promising results when combining syntax and semantics, there are some details that were not clear in the associated publications (i.e., (Byamugisha, 2020, 2022)). For instance, they use a dataset of sentences to create static word embeddings; however, the nature of the model used to obtain the embeddings is not specified (i.e., skipgram or continuous bag-of-words (CBOW)). Similarly, they state that a dataset is enriched with syntax and morphological information and then used to train a classifier to predict parts-of-speech and morphological information. However, it is not specified whether the dataset is labelled manually or whether each label is assigned to a sentence or the individual words. Owing to this, the work was independently replicated by two groups for isiZulu using different datasets and we confirmed some details with the author.

4.1 Replication: group 1

We understood Byamugisha's classifier as either relying on each noun's prefix or semantics (and syntax) to determine the noun class.

The replicated prefix-only classifier determines the noun class by relying on unique class prefixes—a unique contracted, or full, prefix form. The module does not classify all nouns using the entire prefix since the use of the full prefix alone is unsound for isiZulu. This because some of the prefixes are unique when considered as-is, but are ambiguous when the final consonant or vowel is removed, because of the last vowel being coalesced in the noun. For instance, *izinkwa* 'bread' may be classified either as being in class 10 via *izin-* or class 8 via *izi-*. Based on these observations, this module relies on Table 2 to classify nouns. All other prefixes are treated as ambiguous, and their processing is passed to the semantics-based module as they are outside the scope of prefix-only module.

We understood the second classifier, which uses both the morphology and distributional semantics to predict a noun's class, as combining two main components: a morphology and semantics-based classification module, and a noun class candidate filtering module, illustrated in Figure 2.

The classifier takes a noun and retrieves the top N most similar words by applying a nearest neighbors' algorithm on word representations obtained from a word embedding model. The quantity of N was chosen from the range 10-200 based on producing the highest accuracy. To obtain the embedding model for isiZulu, we decided to rely on a skipgram model that was pretrained on a similarly sized corpus (i.e., 1 million sentences) sourced from (Dlamini et al., 2020). We chose to rely on an existing skipgram model over a new CBOW model because we infer, since Byamugisha is not explicit, that they also used a skipgram model as it is better at capturing sub-word information than CBOW and improves word embedding generation for out-of-vocabulary words (Dlamini et al., 2020). In the second phase of the semantics-based module, each retrieved neighbouring word is annotated with a part-of-speech, concord or noun class prediction. In Byamugisha's work, these annotations are obtained via a model that is trained from labelled training that is generated via a context-free grammar (CFG), as confirmed with the original author via email.

Since we did not have access to such a grammar for isiZulu, we chose to rely on the 2016 web-crawled isiZulu 'mixed' dataset from the Leipzig Corpora Collection (Leipzig University, 2024) (the final dataset had 180 000 sentences) and enriched it with part of speech tags using an automated POS tagger (du Toit and Puttkammer, 2021) and a rules to annotate the subject concord. The dataset was used to train a new multinomial logistic regression model, via FastText (which includes subword information in training), to predict the set `{subject concord, noun class}` for each neighbouring word. We use the trained model to automatically annotate words.

Of the automatically annotated words, the filtering module's goal is to predict the noun class

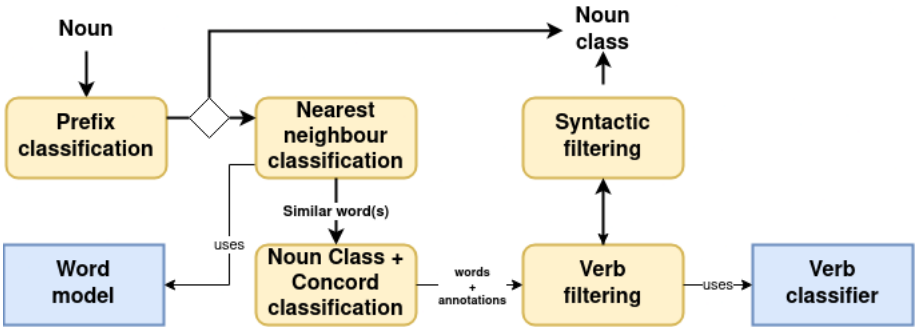

Figure 2: Architecture of group 1's replicated noun classification module.

Table 2: List of classes whose prefixes unique identify the class in isiZulu.

Prefix	aba	abe	ba	be	o	bo	imi	mi	ili	il	li	ama	am	ma
Class	2	2	2	2	2a	2a	4	4	5	5	5	6	6	6
Prefix	isi	si	zi	n	m	zin	zim	lu	ulu	bu	uku	ku	pha	ph
Class	7	7	8	9	9	10	10	11	11	14	15	15	16	16

by relying on the syntax information associated with the neighbouring predictions. Specifically, our replicated filter takes a list of annotated words and predicts a noun class. Byamugisha's module does this by excluding values where the predicted label "is not consistent with the same noun class" (Byamugisha, 2022). Their description of the module's functions has two possible interpretations, i.e., the input noun's neighbouring word contains any of the prefixes associated with its predicted concord or noun class label or the neighbouring word retrieved via the nearest neighbour algorithm has a label that is contained in the set of predictions for its sub-words, hence we chose to pursue two possible versions for the replication. The two versions of the filter remove a neighbour using one of the rules: (1) if the morpheme for the predicted label is not contained in the word or (2) if its predicted label is not in the labels predicted for its sub-words.

We demonstrate the application of these rules using the word *igijima* 'it runs' with the set of bigrams {ig, gi, ij, ji, im, ma}, the pair is predicted to contain subject concord (SC) 9, the first rule checks if the word *igijima* 'it runs' contains the morpheme for SC9, *i*. The second rule predicts an additional label for each subword, and checks if SC9 is in that set.

This syntactic filtering module can be understood as performing syntax-based error checking, rather than extracting additional syntactic information to further disambiguation as with the semantic module.

We tested our replicated classifier on a set of 800 nouns that were curated manually, where the correctness of the noun class was verified using isiZulu.net[2] and the Oxford isiZulu Bilingual Dictionary (de Schryver, Gilles-Maurice, 2015).

4.2 Replication: group 2

We identified and replicated two modules that form the core of Byamugisha's classifier (Byamugisha, 2020): a prefix-based classifier and the classifier that relies on both semantics and syntax. In this section, we will describe our replication process for isiZulu.

The prefix-based classifier model is listed in Algorithm 1. The algorithm relies on Table 3, compiled using the Oxford dictionary (de Schryver, Gilles-Maurice, 2015), to determine if a prefix is sufficient to identify a class. When given a noun as input, the algorithm uses the table of prefixes to establish whether the noun's prefix is associated with a class that has a unique prefix, returning the associated class if one is found (line 2). This is achieved by simply checking whether a noun begins with any of the prefixes found in the table. When the uniqueness check fails (line 1) and the prefix-based model is used without relying on the semantic-syntax module, it randomly selects a class from the list of all the eligible noun classes that have the same prefix value (line 4). For evaluation, when quantifying the performance of the algorithm when it uses random selection on the test set, the algorithm is ran and averaged over 1000 runs. When this module is used in combination with the semantics-syntax module for nouns whose

[2] https://isizulu.net/

Algorithm 1: Specification of the prefix-based noun classifier.

input : A isiZulu noun (*noun*), a prefix table with annotations to determine the sufficiency of determining the noun class by prefix only (*table*), and a boolean to control whether to use the semantics module (*useSemantics*).

output : The noun class of the input noun (*class*).

```
1  if HasUniquePrefix(noun, table, useSemantics) then
2      class ← GetClassFromUniquePrefixes(noun, table);
3  else if useSemantics is false then
4      class ← GetRandClassFromSubsetPrefixes(noun, table);
5  else
6      class ← GetClassUsingSemantics(noun);
7  end
8  return class;
```

prefix fails the uniqueness check, they are passed to the second classifier (line 6).

The second classifier, whose function is documented in Algorithm 2, relies on semantics and syntax. Its functionality extends the prefix-based model, and it is created to handle classes whose prefixes are not unique; instead of randomly selecting the noun class, it selects the class by relying on information found in 'similar' words as the basis. Specifically, we trained a skipgram model, via FastText[3], using a monolingual dataset obtained by aggregating the isiZulu versions of the NCHLT Text Corpus[4], Autshumato Corpora[5], Leipzig Corpus Collection[6], and Common Crawl corpus[7]. When given the input noun, the model is used to identify Nsemantically similar words via the Approximate Nearest Neighbors[8] algorithm. The number of selected neighbouring word representations (i.e., N value) that was optimal in the original work was 110 based on the final accuracy achieved, however, we have found that an approx. value of $N = 60$ is more optimal for isiZulu.

Algorithm 2: List of classes whose prefixes unique identify the class in isiZulu

```
1  vectors ← LoadVectors();
2  classifier ← LoadClassifer();
3  neighbours ← GetAnnoyNeighbours(noun,vectors, n);
4  labels ← {};
5  foreach neighbour in neighbours do
6      label ← classifier.Predict(neighbour);
7      labels.add(label);
8  end
9  classes ← FilterPredictions(labels);
10 class ← MostCommon(classes);
11 return class;
```

The input noun's predicted neighbours are then labelled with automatically predicted labels. This is achieved via a new multinomial logistic regression model, built using FastText and trained on labelled corpora where the label is the noun class chosen based on one or more of the six prefixes that are variants of the subject concord. The corpus was constructed using a rule-based approach since, unlike Byamugisha, there is no context-free grammar to generate and label a dataset. Specifically, we created *ad hoc* rules that make use of the different subject concord-based prefixes associated with each noun class, as described in the Oxford dictionary (de Schryver, Gilles-Maurice, 2015, pS35).

The first phase when annotating data is to identify the relevant nouns and verbs in a sentence and we now turn to describe the process:

1. Identify the last noun in a sentence. A word is considered a noun if it exists in a dataset of nouns extracted from a dictionary (de Schryver, Gilles-Maurice, 2015), used for training and testing the models, as described below.
2. Identify verbs in the sentence. A word is considered a verb if the following criteria is met:
   (a) The word must be the longest word that uses a root from the `vroots.txt` file sourced from (Keet and Khumalo, 2017).
   (b) If a root from `vroots.txt` is found in a word, then it cannot be the leading substring.

[3] https://pypi.org/project/fasttext/
[4] https://hdl.handle.net/20.500.12185/321
[5] https://hdl.handle.net/20.500.12185/575
[6] Obtained by combining the Community '17, Mixed '16, Web '19, and Wiki '21 Leipzig corpora
[7] https://data.statmt.org/cc-100/
[8] https://pypi.org/project/annoy/

Table 3: List of possible isiZulu prefixes found in nouns for various classes. Abbreviations: Am. = Ambiguous, Uni = Unique, P = Prefix, C = Class

Uni	**P**	aba	abe	o	imi	ili	ama	ame	is	isi	in	izin	izim	ulu	ubu	ub	uku	ukw
	C	2	2	2a	4	5	6	6	7	7	9	10	10	11	14	14	15	15
Am	**P**	um	umu	u	umu	um	i	iz	izi	im	u							
	C	1	1	1a	3	3	5	8	8	9	11							

(c) The word cannot be a homograph of the noun identified in Step 1.

3. Create labels for the verb by first extracting a possible prefix from the verb via removing the root that was identified using the `vroots.txt` file described in Step 2. If the prefix matches any of the prefix variants for the various noun classes then each class will be collected as a possible label.

We then expand the sentences, when given the noun and verb positions and possible labels, by annotating the sentence with the collected labels. We demonstrate how this process works via a single example: when given the short sentence *intsha ingagcini* 'the youth must not stop', the dictionary dataset is used to identify that the the word *intsha* 'youth' is a noun that belongs to class 9. The second word is identified as a verb since it contains the verb root *-gcin-* according to `vroots.txt`. The rules then focus on the verb's prefix *inga-* after removing the root. Specifically, they check if the verb leads with the negative marker *a-*, if it does then they check whether the marker is followed by the prefix variant found in negative verbs *-yi-* from Table 4. In this example, it is not but the prefix starts with the subject concord value *i-* and followed by a consonant hence *inga-* is chosen as the label as a candidate for noun class 9.

After obtaining the annotated dataset, we trained a classifier and used it to annotate neighbouring words and remove words whose predicted concord label does not match the true noun class label, as specified in Table 3. In this sense we filter out the neighbours that are not syntactically consistent with the semantic categorisation.

The replicated classifiers are evaluated on a manually collected a dataset of 2279 isiZulu nouns (+ noun class annotations) from the Oxford dictionary (de Schryver, Gilles-Maurice, 2015). This dataset was split into train (80%) and test (20%) sets in a manner that ensures that the noun classes are fairly distributed between the two sets.

5 Results and discussion

The accuracy of the resulting models are reported in Table 5. Group 2's models demonstrated better performance than group 1. The following text compares the differences between the new models and Byamugisha's.

The models that classify nouns based on only their prefix differ by 23% between the two groups. Since group 1's model does not make any predictions for prefixes that are ambiguous and group 2 does so by averaging over multiple runs, this suggests that there is utility in randomly choosing a class amongst possible classes, on average over multiple uses. Nonetheless, Byamugisha's prefix-only model outperforms groups 2's model by 10%. It is unclear whether the two models are directly comparable because Byamugisha's work lacks clarify regarding how this model processes nouns where usage of the prefix, by itself, to determine the noun class is insufficient. It is likely that it ignores such nouns similar to group 1 or employs some other strategy.

There was also a large gap in performance in the classifiers that rely on the prefix, syntax, and semantics. The largest gap in accuracy between group 1 and 2's models was 17%. There are numerous reasons why there could be differences in performance between the models, especially since there were multiple differences taken by the two groups due to interpretation and approach to replication. Of note, while group 1 relied on a similar sized dataset to (Byamugisha, 2022) for training the semantic classifier while the second group did not, the group 1's model performs far worse than Byamugisha's model. This may indicate that ones' possible access to the same kind of resources, and not just similarly sized ones, has an impact on the accuracy—in addition to their interpretation of the original work.

More generally, the two groups identified the following areas for which the original work differs:

1. There is a large difference in the number of cases where the prefix is insufficient to determine the noun class. For Runyankore, only 4

Table 4: List of verbal prefixes, obtained from the subject concord, for noun class 9 taken from (de Schryver, Gilles-Maurice, 2015). Abbreviations: SC = subject concord, C = consonant. The symbol + is used to denote that the regular subject concord is used.

Prefix	Description	Value
SC + C	Subject concord when followed by consonant	i
SC + a/e	Subject concord when followed by the letters -a- or -e-	y
SC + o	Subject concord when followed by the letters -o-	y
SC (negative)	Subject concord in negated verb	yi
SC (situative; continuous)	Subject concord in situative verb	+
SC (subjunctive)	Subject concord in subjunctive verb	+

out of 21 classes have the same prefix hence are considered ambiguous—only two pairs of classes 1/2 (prefix *omu-*) and 9/10 (prefix *em-*) share the same prefixes. By contrast, in group 1's case, 38% of the prefix values are deemed insufficient, by themselves, to predict the noun class. Similarly, 37% of the prefixes considered by Group 2 had the same characteristic. Hence, in the case of our replication groups, we see that there are many more nouns whose classification is the responsibility of the semantic-syntax module, thus overall performance is impacted by the quality of the the module to a larger degree.

2. Byamugisha's work relies on a context-free grammar to generate a dataset with approximately 1 million sentences. This context free grammar is used to create two versions of the dataset, labelled and unlabelled, and these are used to trained models that form the foundation of the semantic-syntax module. There is no equivalent CFG for isiZulu; hence, the groups used classifiers that are trained on similarly sized datasets and crafted labelled datasets. Notably, group 1's careful design of the labelled dataset vs. group 2's use of ad hoc rules did not yield better performance overall.
3. Both groups were unable to identify the precise filtering strategy employed by Byamugisha hence considered multiple variations.

These replication challenges and differences in interpretation highlight that there is still a need to (1) conduct a comprehensive exploration of key functionalities, such as candidate prediction filtering and quantifying the accuracy of the prefix-only model while ensuring transparency regarding its handling of nouns that cannot classified using the prefix only, and (2) investigate more diverse models that integrate semantics, syntax, and morphology, while ensuring that the data is auditable to ascertain whether such models are better across NCB languages, irrespective of the Guthrie zone in which they can be classified.

Table 5: Accuracy results of the replicated models ([†] No predictions are made for nouns whose prefixes are not unique.)

Group	Model	Accuracy
1	Prefix-only[†]	0.36
1	Prefix-semantics-synt. (FilteringRule1)	0.66
1	Prefix-semantics-synt. (FilteringRule2)	0.46
2	Prefix-only	**0.59**
2	Prefix-semantics-synt.	**0.83**

6 Conclusions

Aiming to replicate Byamugisha (2022)'s work on a combined syntactic and semantic approach to classifying nouns in their noun class for isiZulu, our results showed that our best performing models combined syntax, morphology, and semantics yield better performance (83%) compared to relying only on the prefix (59%), which is in line with the findings for Runyankore. However, the differences in the choice of training and test data, as well the design of the models that make up the module that combines syntax, morphology, and semantics, had a substantial effect on the eventual accuracy of noun class disambiguation. Specifically, while group 1's model with prefix-semantics-syntax outperformed their prefix-only models, their performance was not in the same range as the Runyankore models. Possible causes for the difference in accuracy is the lack of a sufficient CFG to create training data in isiZulu and the imprecision with which candidate filtering was implemented in the

original work.

Future work includes investigating the impact of differences in the number nouns that cannot be classified by prefix alone per language and how that effects transferability of the use of semantics, syntax, and morphology and investigating more diverse models (e.g., include pretrained language models) that integrate such knowledge while ensuring that the data is auditable.

7 Limitations and ethical considerations

We replicated and quantified the utility of modular techniques towards noun disambiguation. However, our methodology disregards a number of extant methods such as pre-trained language models since they were not included in (Byamugisha, 2022). Additional work is required to investigate the utility of combining semantics, syntax, and morphology, in the context of such models.

One of our datasets was extracted from a dictionary (de Schryver, Gilles-Maurice, 2015), under the fair use right as codified by the South African Copyright Act[9]; therefore, the dataset is and will not be distributed publicly.

8 Resource availability

The code and shareable datasets of group 1 and group 2 are available at https://github.com/KEEN-Research/NCBnounClassification.

Acknowledgements

This work was financially supported in part by the National Research Foundation (NRF) of South Africa (Grant Number and CPRR23040389063).

References

Jaco Badenhorst, Charl van Heerden, Marelie Davel, and Etienne Barnard. 2011. Collecting and evaluating speech recognition corpora for 11 South African languages. *Language Resources and Evaluation*, 45(3):289–309.

Michael L Burton and Lorraine Kirk. 1976. Semantic reality of Bantu noun classes: the Kikuyu case. *Studies in African Linguistics*, 7(2):157–174.

Joan Byamugisha. 2019. *Ontology verbalization in agglutinating Bantu languages: a study of Runyankore and its generalizability*. Ph.D. thesis, Department of Computer Science, University of Cape Town, South Africa.

Joan Byamugisha. 2020. Generating varied training corpora in runyankore using a combined semantic and syntactic, pattern-grammar-based approach. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 273–282. Association for Computational Linguistics.

Joan Byamugisha. 2022. Noun class disambiguation in Runyankore and related languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4350–4359, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2018. Pluralizing nouns across agglutinating Bantu languages. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2633–2643. Association for Computational Linguistics.

Joan Byamugisha, C. Maria Keet, and Langa Khumalo. 2016. Pluralising nouns in isiZulu and related languages. In *Computational Linguistics and Intelligent Text Processing - 17th International Conference, CICLing 2016, Konya, Turkey, April 3-9, 2016, Revised Selected Papers, Part I*, volume 9623 of *Lecture Notes in Computer Science*, pages 271–283. Springer.

Ellen Contini-Morava. 1997. Noun classification in Swahili. In Robert K. Herbert, editor, *African Linguistics at the Crossroads: Papers from Kwaluseni*, chapter 6, pages 18–22. Rüdiger Köppe Verlag.

Robert Dale. 2016. The return of the chatbots. *Natural Language Engineering*, 22(5):811–817.

de Schryver, Gilles-Maurice. 2015. *Oxford Bilingual School Dictionary: isiZulu and English / Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgisi, Esishicilelwe abakwa-Oxford. Second Edition*. Oxford University Press Southern Africa.

Katherine Demuth. 2000. Bantu noun class systems: loanword and acquisition evidence of semantic productivity. In Gunter Senft, editor, *Systems of Nominal Classification*, chapter 8, pages 270–292. Cambridge University Press.

J Peter Denny and Chet A Creider. 1986. The semantics of noun classes in Proto-Bantu. In Colette G Craig, editor, *Noun classes and categorization*, chapter 3, pages 217–239. John Benjamins Publishing Company.

Sibonelo Dlamini, Edgar Jembere, and Anban Pillay. 2020. Evaluation of word and sub-word embeddings for isizulu on semantic relatedness and word sense disambiguation tasks. In *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pages 1–6.

[9] https://www.saflii.org/za/legis/consol_act/ca1978133/

Jakobus S. du Toit and Martin J. Puttkammer. 2021. Developing core technologies for resource-scarce nguni languages. *Information*, 12(12).

Roald Eiselen and Martin J. Puttkammer. 2014. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3698–3703. European Language Resources Association (ELRA).

Neele Falk, Yana Strakatova, Eva Huber, and Erhard Hinrichs. 2021. Automatic classification of attributes in German adjective-noun phrases. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 239–249, Groningen, The Netherlands (online). Association for Computational Linguistics.

Annie Gagliardi, Naomi H. Feldman, and Jeffrey Lidz. 2012. When suboptimal behavior is optimal and why: Modeling the acquisition of noun classes in tsez. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society, CogSci 2012, Sapporo, Japan, August 1-4, 2012*. cognitivesciencesociety.org.

Francis Katamba. 2014. Bantu Nominal Morphology. In Derek Nurse and Gérard Philippson, editors, *The Bantu Languages*, chapter 7, pages 103–120. Routledge.

C Maria Keet. 2021. Natural language generation requirements for social robots in Sub-Saharan Africa. In *Proceedings of the 2021 IST-Africa Conference (IST-Africa), 10-14 May 2021, South Africa (Virtual)*, pages 1–8. IEEE.

C Maria Keet and Langa Khumalo. 2017. Grammar rules for the isizulu complex verb. *Southern African linguistics and applied language studies*, 35(2):183–200.

C. Maria Keet, Musa Xakaza, and Langa Khumalo. 2017. Verbalising OWL ontologies in isiZulu with Python. In *The Semantic Web: ESWC 2017 Satellite Events*, volume 10577 of *LNCS*, pages 59–64. Springer. 30 May - 1 June 2017, Portoroz, Slovenia.

Leipzig University. 2024. Leipzig Corpora Collection: Zulu mixed corpus based on material from 2016.

Zola Mahlaza and C. Maria Keet. 2020. OWLSIZ: An isiZulu CNL for structured knowledge validation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 15–25, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Tumi Moeng, Sheldon Reay, Aaron Daniels, and Jan Buys. 2021. Canonical and surface morphological segmentation for Nguni languages. *CoRR*, abs/2104.00767.

Lulamile Mzamo, Albert Helberg, and Sonja Bosch. 2019. Evaluation of combined bi-directional branching entropy language models for morphological segmentation of isiXhosa. In *Proceedings of the South African Forum for Artificial Intelligence Research, Cape Town, South Africa, 4-6 December, 2019*, volume 2540 of *CEUR Workshop Proceedings*, pages 77–89. CEUR-WS.org.

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

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

Mnoneleli Nogwina. 2016. Development of a stemmer for the IsiXhosa language. Master's thesis, University of Fort Hare.

Lluís Padró and Roser Saurí. 2024. Fine-tuning open access LLMs for high-precision NLU in goal-driven dialog systems. In *Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability @ LREC-COLING 2024*, pages 33–42, Torino, Italia. ELRA and ICCL.

Gary B. Palmer and Claudia Woodman. 2000. Ontological classifiers as polycentric categories, as seen in Shona class 3 nouns. In Martin Pütz and Marjolyn Verspoor, editors, *Explorations in Linguistic Relativity*, chapter 12, pages 225–249. John Benjamins Publishing company.

Laurette Pretorius and Sonja E. Bosch. 2009. Finite state morphology of the Nguni language cluster: Modelling and implementation issues. In *Finite-State Methods and Natural Language Processing, 8th International Workshop, FSMNLP 2009, Pretoria, South Africa, July 21-24, 2009, Revised Selected Papers*, volume 6062 of *Lecture Notes in Computer Science*, pages 123–130. Springer.

Lihui Pu, Wendy Moyle, Cindy Jones, and Michael Todorovic. 2018. The Effectiveness of Social Robots for Older Adults: A Systematic Review and Meta-Analysis of Randomized Controlled Studies. *The Gerontologist*, 59(1):e37–e51.

Georg I. Schlünz, Nkosikhona Dlamini, and Rynhardt P. Kruger. 2016. Part-of-speech tagging and chunking in text-to-speech synthesis for South African languages. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 3554–3558. ISCA.

Rianne van den Berghe, Josje Verhagen, Ora Oudgenoeg-Paz, Sanne Van der Ven, and Paul Leseman. 2019. Social robots for language learning: A review. *Review of Educational Research*, 89(2):259–295.