# Noun Class Disambiguation in Runyankore and Related Languages

**Joan Byamugisha**
IBM Research Africa
45 Juta Street, Braamfontein
Johannesburg, South Africa
joan.byamugisha@ibm.com

## Abstract

Bantu languages are spoken by communities in more than half of the countries on the African continent by an estimated third of a billion people. Despite this populous and the amount of high quality linguistic research done over the years, Bantu languages are still computationally under-resourced. The biggest limitation to the development of computational methods for processing Bantu language text is their complex grammatical structure, chiefly in the system of noun classes. We investigated the use of a combined syntactic and semantic method to disambiguate among singular nouns with the same class prefix but belonging to different noun classes. This combination uses the semantic generalizations of the types of nouns in each class to overcome the limitations of relying only on the prefixes they take. We used the nearest neighbors of a query word as semantic generalizations, and developed a tool to determine the noun class based on resources in Runyankore, a Bantu language indigenous to Uganda. We also investigated whether, with the same Runyankore resources, our method had utility in other Bantu languages, Luganda, indigenous to Uganda, and Kinyarwanda, indigenous to Rwanda. For all three languages, the combined approach resulted in an improvement in accuracy, as compared to using only the syntactic or the semantic approach.

## 1 Introduction

Over the last three decades, there has been an increase in the development of computational resources for Bantu languages. Much of this work applies the knowledge gained on Bantu language linguistics to digitize textual resources and develop software tools for text processing for a single language or for a group of languages. Among the textual resources created are the SAWA English-Kiswahili parallel corpus for machine learning (De Pauw et al., 2011) and the labelled and unlabelled Runyankore datasets (Byamugisha, 2020). The text processing resources include morphological tools, such as morphological analyzers for isiZulu (Bosch and Pretorius, 2003, 2004), isiXhosa, seSwati, and isiNdebele (Bosch et al., 2008); text generation tools, such as a morphological generator for isiZulu (Bosch and Pretorius, 2003) and surface realizers for isiZulu (Keet et al., 2017) and Runyankore (Byamugisha et al., 2017a,b); part-of-speech taggers for Kiswahili, Ciluba, Northern Sotho, and isiZulu (De Pauw et al., 2012); and noun pluralization tools for isiZulu and Runyankore (Byamugisha et al., 2016), chiShona, isiXhosa, Kikuyu, Kinyarwanda, and Luganda (Byamugisha et al., 2018).

Despite these efforts, Bantu languages are still among the most computationally under-resourced languages in the world. This is due to their complex grammatical structure, mainly the noun class system, verb morphology, and agglutinative morphology. In this paper, we focus on the noun class system only, the hallmark of Bantu nominal morphology (Katamba, 2003). This system places every noun in a language into a class, based on the semantics of a noun first (such as whether the noun is of a human or non-human entity), then the morphology of a noun next (which is based on the prefix of a noun) (Katamba, 2003). The importance of the noun class to computational text processing goes beyond classifying nouns, and also determines the formulation of other parts of speech, such as adjectives, verbs, possessives, determinants, grammatical number, etc. because a noun class is central to an extensive system of concordial agreement that determines morphological composition (Katamba, 2003).

Given that there exists in some noun class systems class prefixes that are not unique among different noun classes, a problem of ambiguity exists when determining a noun class using a class prefix only, as is the case with morphological approaches. We refer to the process of determining the correct



*Proceedings of the 29th International Conference on Computational Linguistics*, pages 4350–4359
October 12–17, 2022.

noun class under these circumstances as noun class disambiguation, which, to the best of our knowledge, has not yet been solved. Katamba (2003) state that using a noun's semantics and extending beyond morphology to syntax by considering concords, can overcome the limitations of morphological approaches, but it requires large resources that capture the context in which a noun is used. We therefore investigated whether it is possible to extend morphological approaches to syntactic approaches by including the syntax of an entire sentence, and further combine this with the semantics of a noun in order to undertake noun class disambiguation. We used the following questions:

1. To what extent can a noun's semantic generalizations, in the form of nearest neighbors, work to disambiguate among noun classes to identify its noun class correctly?

2. Can the presence of sub-word information in word vectors in one language contribute enough semantic information to improve noun class disambiguation in another Bantu language?

We investigated the applicability of a combined syntactic and semantic approach as a means of noun class disambiguation among singular nouns, using word vectors pre-trained using FastText (Bojanowski et al., 2016; Joulin et al., 2016) on two one million sentence datasets in Runyankore, one dataset unlabelled and the other labelled with morphological information including the noun class (Byamugisha, 2020). The syntactic method relies both on the morphology of a noun based on its prefix and on the syntax of other grammatical units in a sentence which are determined by a noun class. The semantic method uses the noun class labels of the nearest neighbors of a query word. Only singular nouns were used in this investigation because there is ground-truth data from the singular wordlists used in the noun pluralization tools with which to evaluate the results, and these same tools apply the knowledge on the singular/plural pairings to solve the problem of plurals computationally. We started with two datasets in Runyankore, the same language as the word vectors used, and obtained accuracies of 80.54% and 85.23% on two test sets. We then investigated whether, using a combined approach with the same Runyankore word vectors, a correct noun class determination can be made for another Bantu language. Using Luganda and Kinyarwanda, we obtained accuracies of 73.97% and 63.64%, respectively. To the best of our knowledge, this is the first computational attempt to use both the syntactic and semantic characteristics of a noun to disambiguate among noun classes with the same class prefix.

The rest of this paper is arranged as follows: Section 2 provides information on Bantu languages, with a focus on the noun class system; Section 3 details the materials, methods, and results from using a combined syntactic and semantic approach to determine a noun class; Section 4 discusses the implications of this work; and we conclude in Section 5.

## 2 Bantu Language Noun Class System

Bantu languages are a group of languages indigenous to Africa (Nurse and Philippson, 2003). They extend from the south, below Nigeria, to most of central, east, and southern Africa, as shown in Figure 1 (Nurse and Philippson, 2003). There are Bantu-speaking communities in 27 of the continent's 54 countries, with about 240 million speakers (Nurse and Philippson, 2003). The exact number of languages classified as Bantu ranges from 300 to 680, based on different criteria by different authors (Nurse and Philippson, 2003).

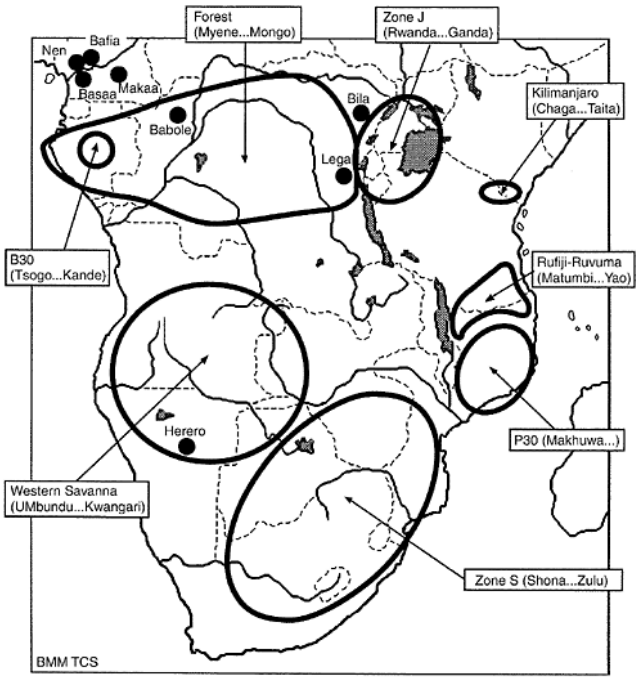


Figure 1: The spread of Bantu languages across Africa (Nurse and Philippson, 2003)

Bantu languages assign all nouns to a class re-

ferred to as a noun class (NC); and there are over 20 noun classes, though some NCs have fallen into disuse in most languages (Nurse and Philippson, 2003; Mohlala, 2003; Maho, 1999). The semantic generalizations of the types of nouns in each class 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; Katamba, 2003; Maho, 1999).

| Noun Class | Description of Associated Nouns |
|---|---|
| 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, and paired things |
| 7 and 8 | Inanimate objects |
| 9 and 10 | Tools and animals |
| 11 | Long thin stringy objects, languages, and inanimate objects |
| 12 and 13 | Diminutives |
| 14 | Abstract concepts |
| 15 | Infinitives and parts of the body |
| 16, 17, and 18 | Locative classes |
| 19 | Diminutives |
| 20, 21, and 22 | Augmentatives |
| 23 | Locative class |

Table 1: Classification of Bantu nouns into noun classes (the 'and' indicates that the two classes are a singular/plural pairing)

The simple noun comprises a prefix and a stem (Katamba, 2003); for example, *omuntu* 'person' in Runyankore, which can be analyzed as the prefix *o-mu-* and stem *-ntu*. Therefore, in addition to the semantic categorization of nouns shown in Table 1, nouns are categorized morphologically also, according to the prefixes they take (Katamba, 2003). In many Bantu languages, the class prefix may be preceded by a formative referred to as the augment, pre-prefix, or initial vowel (Katamba, 2003; Maho, 1999). In the above example of *omuntu*, the class prefix *o-mu-* possesses the augment *o*. The augment is not found in all Bantu languages (Katamba, 2003; Maho, 1999).

Noun classes do not only classify nouns, but are at the heart of an extensive system of concordial agreement (Katamba, 2003), where each class determines the agreement with: concord patterns; nominal prefix in nouns, locatives, and adjectives; numeral prefix; pronominal prefix for substitutives, connectives, possessives, demonstratives, and determinatives; initial prefix in absolutive verb forms; and the verbal infix (Katamba, 2003; Maho, 1999; Zentz, 2016; Nurse and Philippson, 2003; Tayebwa, 2014). Given how central the noun class system is to Bantu language computational linguistics, it is important to identify a means of determining the noun class of a noun.

## 3 Noun Class Disambiguation Using Syntax and Semantics

In this research, we aimed to find out to what extent determining the semantics of a noun is beneficial to disambiguating among noun classes, in order to determine its noun class correctly. In Byamugisha (2020), it was shown that from word vectors trained on one million Runyankore sentences in an unsupervised manner, the nearest neighbors (co-occurrence vectors) obtained for a query word reflect the semantic generalizations of noun classes shown in Table 1. We, therefore, used a noun's nearest neighbors as a representation of its semantic generalizations. We investigated further whether, with the pre-trained word vectors in one Bantu language, the nearest neighbors of a query word in a different Bantu language improve the disambiguation, and, consequently, determination of a noun class in this other Bantu language. Four criteria are relied upon when selecting languages for this research: (1) the availability of linguistic information about its noun class; (2) the presence of an augment in a class prefix of a noun, as is the case with Runyankore; (3) the placement of people and kinship nouns in the same noun class (instead of two different classes), as is the case in Runyankore's noun class system; and (4) the availability of ground-truth data with which to evaluate the output. Based on this, Luganda and Kinyarwanda were selected from the existing computational resources for seven Bantu languages in Byamugisha et al. (2018).

Our approach to disambiguating among noun classes of a noun also extends beyond the morphology of a noun and considers the presence of the concords in a sentence that are indicative of a noun class. We combine this method with a noun's semantics as defined by its nearest neighbors. The details on each of these two methods are explained in the following sections.

### 3.1 Syntactic Approach

As explained by Katamba (2003), the assignment of nouns to a class is partially morphological, using rules based on class prefixes (which may or may not be unique) according to the noun class system of a particular language. A syntactic approach to disambiguating among noun classes involves

analyzing the syntax of an entire sentence to confirm whether other grammatical units such as the possessive, subject, or object concords (which are always unique in a noun class system) are indicative of a particular noun class. Runyankore, Luganda, and Kinyarwanda have different noun class systems, and, therefore, require different morphological rules for each language. The class prefixes for the noun class systems of these languages are shown in Table 2.

| Noun Class | Runyankore | Luganda | Kinyarwanda |
|---|---|---|---|
| 1 | o-mu- | o-mu- | u-mu- |
| 2 | a-ba- | a-ba- | a-ba- |
| 3 | o-mu- | o-mu- | u-mu- |
| 4 | e-mi- | e-mi- | i-mi- |
| 5 | e-ri-/ei- | e-li-/e- | i-ri-/i- |
| 6 | a-ma- | a-ma- | a-ma- |
| 7 | e-ki- | e-ki- | i-ki-/i-cy-/i-gi- |
| 8 | e-bi- | e-bi- | i-bi- |
| 9 | em-/em- | e-n- | i-/i-n-/i-nz- |
| 10 | em-/em- | e-n- | i-/i-n-/i-nz- |
| 11 | o-ru- | o-lu- | u-ru- |
| 12 | a-ka- | a-ka- | a-ka-/a-ga- |
| 13 | o-tu- | o-tu- | u-tu-/u-du- |
| 14 | o-bu- | o-bu- | u-bu- |
| 15 | o-ku- | o-ku- | u-ku-/u-gu- |
| 16 | a-ha- | wa- | a-ha- |
| 17 | o-ku- | ku- | N/A |
| 18 | o-mu- | mu- | N/A |
| 19 | N/A | N/A | N/A |
| 20 | o-gu- | o-gu- | N/A |
| 21 | a-ga- | a-ga- | N/A |
| 22 | N/A | N/A | N/A |
| 23 | N/A | e- | N/A |

Table 2: The noun classes for Runyankore (Asiimwe, 2014), Luganda (Baertlein and Ssekitto, 2014), and Kinyarwanda (Kimenyi, 2004), showing the class prefixes. The dashes between the letters in the prefix illustrate separation between the augment and prefix; and 'N/A', the NC is not present in that language.

Table 2 shows the noun classes and the class prefixes for Runyankore, Luganda, and Kinyarwanda. They have different numbers of noun classes–20 in Runyankore, 21 in Luganda, and 16 in Kinyarwanda–because some noun classes have fallen into disuse in most languages (Nurse and Philippson, 2003; Mohlala, 2003; Maho, 1999). None of these languages has classes 19 and 22.

Morphological rules, based on a class prefix, are used first to attempt to determine a noun's noun class. However, there are some prefixes that are the same for different noun classes. In Table 2, these are classes 1, 3, and 18 with prefix *o-mu-* and classes 15 and 17 with prefix *o-ku-* in Runyankore; classes 1 and 3 with prefix *o-mu-* and classes 5, 9, and 23 with prefix *e-* in Luganda; and classes 1 and 3 with prefix *u-mu-* and classes 5 and 9 with prefix *i-* in Kinyarwanda. This results in ambiguity during noun class determination, which cannot be resolved by a morphological approach only. On the other hand, concords can be used to disambiguate between nouns belonging to different classes but with the same class prefix because their concords differ (Katamba, 2003; Maho, 1999). Table 3 shows the subject concords in Runyankore's noun class system.

| Noun Class | Class Prefix | Subject Concord |
|---|---|---|
| 1 | o-mu- | -a- |
| 2 | a-ba- | -ba- |
| 3 | o-mu- | -gu- |
| 4 | e-mi- | -gi- |
| 5 | ei-/e-ri- | -ri- |
| 6 | a-ma- | -ga- |
| 7 | e-ki- | -ki- |
| 8 | e-bi- | -bi- |
| 9 | e-n-/e-m- | -e- |
| 10 | e-n-/e-m- | -zi |
| 11 | o-ru- | -ru- |
| 12 | a-ka- | -ka- |
| 13 | o-tu- | -tu- |
| 14 | o-bu- | -bu- |
| 15 | o-ku- | -ku |
| 16 | a-ha- | -ha |
| 17 | o-ku- | -ha- |
| 18 | o-mu- | -ha- |
| 20 | o-gu- | -gu- |
| 21 | a-ga- | -ga- |

Table 3: The Subject concords of the Runyankore noun class system, showing that concords are unique across classes with the same prefix

Table 3 shows that for classes 1, 3, and 18 with prefix *o-mu-* and classes 15 and 17 with prefix *o-ku-*, the subject concords are unique among them, and can thus be used to disambiguate among these classes. We, therefore, extended beyond morphology, to syntax, by including an entire sentence in our approach.

### 3.2 Semantic Approach

According to Katamba (2003), the assignment of nouns to a class is also partially based on semantic generalizations of the types of nouns in each class, as shown in Table 1. In Byamugisha (2020), the nearest neighbors obtained from word vectors trained on one million Runyankore sentences were found to have a high level of semantic relatedness. Table 4 from Byamugisha (2020) shows the nearest neighbors for query words with the prefix *o-mu-* but belonging to different noun classes.

The examples in Table 4 present a case of noun class ambiguity using three query words (*omuntu*,

| Query Word | Nearest Neighbors |
|---|---|
| *omuntu* (person) | *omugyesi* (reaper), *omutaahi* (companion), *omukoreesa* (overseer), *omushomesa* (teacher), *omukuru* (elder) |
| *omuti* (tree) | *omutumba* (banana tree), *omwani* (coffee tree), *omuzaabibu* (grape or grapevine), *omucungwa* (orange), *omugusha* (sorghum) |
| *omukono* (arm) | *omunwa* (mouth), *omutwe* (head), *eriino* (tooth), *enkokora* (elbow), *okuguru* (leg) |

Table 4: Nearest Neighbors for query words with the prefix *o-mu-* (Byamugisha, 2020)

*omuti*, and *omukono*) which all have the same noun prefix, *o-mu-*. However, they belong to different noun classes, with *omuntu* in class 1, semantically for people according to Table 1, and *omuti* and *omukono* in class 3. The nearest neighbors of these query words reflect a semantic distinction among them, which cannot be determined syntactically. On the other hand, the nearest neighbors of *omukono* in Table 4 belong to different noun classes morphologically: *omunwa* and *omutwe* in class 3, *eriino* in class 5, *enkokora* in class 9, and *okuguru* in class 15 according to Table 2. Therefore, whilst the nearest neighbors help to exclude class 1 for *omukono*, the disambiguation is made syntactically, by going beyond a noun's prefix and extending to the concords in a sentence. Our approach thus combines syntax and semantics in order to leverage the benefits of both to disambiguate among noun classes.

### 3.3 Materials

The materials used to develop and evaluate a combined syntactic and semantic approach to noun class disambiguation were obtained from existing computational resources. These included: (1) pre-trained word vectors in Runyankore from Byamugisha (2020); (2) a classifier trained on one million sentences, in Runyankore labelled for parts-of-speech and morphology (including the noun class assignment of nouns, subject and object concords, and adjective prefixes) from Byamugisha (2020); (3) a dataset of singular nouns and their correct noun classes in Runyankore to act as ground-truth during development, obtained from Set1r and Set2r in Byamugisha et al. (2016); and (4) a dataset of singular nouns and their correct noun classes in Runyankore, Luganda, and Kinyarwanda, to act as ground-truth during evaluation, obtained from SetI and SetC in Byamugisha et al. (2018). Table 5 shows the language and number of nouns in each of these datasets.

| Dataset | Language | Number of Nouns |
|---|---|---|
| Set1r | Runyankore | 92 |
| Set2r | Runyankore | 2542 |
| SetI | Runyankore | 81 |
| SetC | Runyankore | 88 |
| SetI | Luganda | 75 |
| SetC | Luganda | 78 |
| SetI | Kinyarwanda | 70 |
| SetC | Kinyarwanda | - |

Table 5: Details on datasets used

The datasets[1] in Table 5 contain a noun and its noun class. These ground-truth datasets contain singular nouns only (with the exception of mass nouns), thus, we consider only singular nouns so far. The need for noun class disambiguation is evident in the statistical characteristics of these datasets: 25.0% of nouns in Set1R have the prefix *om-* and 51.09% have the prefix *e-*; 39.89% of nouns in Set2r have the prefix *om-*, while 34.97% have the prefix *e-*. For Runyankore, 17.05% of nouns in SetI have the prefix *om-* and 40.91% have the prefix *e-*; while 28.4% of nouns in SetC have the prefix *om-* while 58.02% have the prefix *e-*. For Luganda, 23.07% of nouns in SetI have prefix *om-* and 38.46% have prefix *e-*; while 16.0% of nouns in SetC have prefix *om-* and 45.33% have prefix *e-*. For Kinyarwanda, with SetI only, 18.57% of nouns have prefix *um-* and 44.29% have prefix *e-*.

### 3.4 Methods

We used an iterative approach to develop a tool that uses a noun's syntax and semantics to disambiguate first, and then determine its noun class. This involved testing different versions of the tool for Runyankore using Set1r and Set2r, adding new functionality, and then testing again. The tool was developed as a Python module using the Gensim library[2] to load the pre-trained word-vectors of a language model for Runyankore from Byamugisha (2020). The FastText library[3] was used to train a classifier on a one million sentence labeled dataset also from Byamugisha (2020). The dataset was

[1] Set1r and Set2r can be obtained from https://github.com/ThesisResources/RunyankorePluralizer, while SetI and SetC can be found at https://github.com/runyankorenlg/Generic-Pluralizer.

[2] For more details on the Gensim Python library, see https://pypi.org/project/gensim/.

[3] For more details on the FastText Python library, see https://pypi.org/project/fasttext/.

split into 70% for training, 20% for validation, and 10% for testing. We used the default values for the training parameters of the FastText library. Figure 2 shows the architecture of the method used for noun class disambiguation.

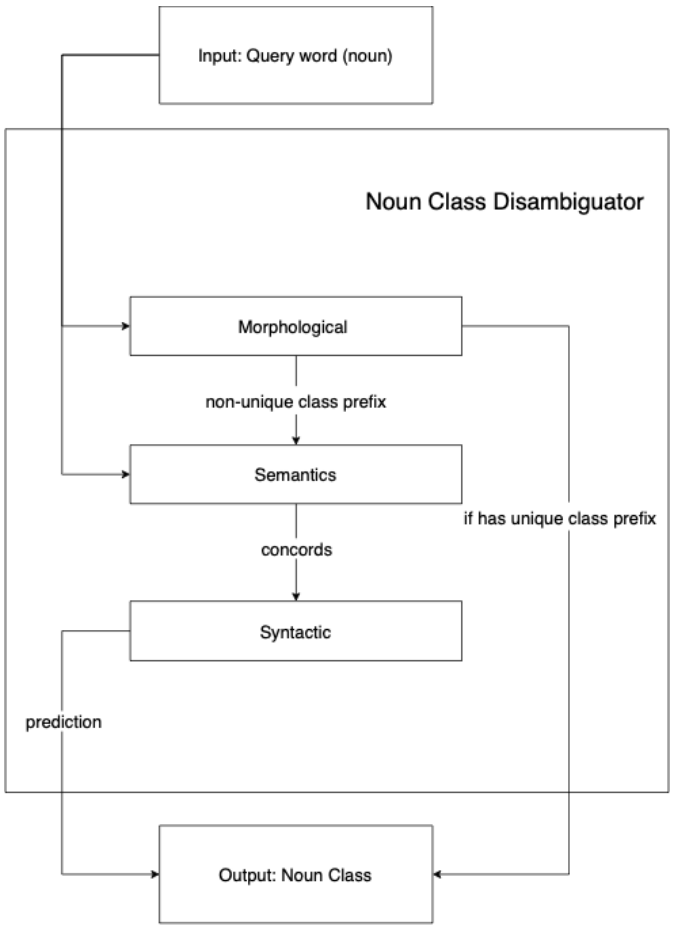


Figure 2: Method taken for noun class disambiguation

Figure 2 shows how a noun class of a noun is determined, as well as where disambiguation is required. Given a noun, a morphological step is performed first, but this results in a noun class only if the class prefix of the noun is unique (see tables 2 and 3 for the class prefixes). On the other hand, if the class prefix is not unique, then the noun is regarded as a query word and undergoes the semantic method, during which the following happens:

- Extracting a vocabulary from the pre-trained language model, and including all the sub-word information in the model as part of the vocabulary;
- Finding the noun's nearest neighbors from the pre-trained language model, which should include the sub-word results in the vocabulary; and
- Predicting, from the top-n nearest neighbors obtained, a label for each nearest neighbor using the classifier.

The results from the semantic method go through the syntactic method, where concords in an entire sentence can be used to determine the final label because concords are unique among noun classes (Katamba, 2003; Maho, 1999) (see Table 3 for subject concords in Runyankore). This involves:

- Excluding items from the list of nearest neighbors if the label predicted for their concord (sub-word) is not consistent with the same noun class; and
- Predicting the final noun class based on the final list of nearest neighbors

Table 6 shows the improvement in accuracy on the two test sets as different versions of the tool were implemented.

| Version | Set1r | Set2r |
|---|---|---|
| Classical nearest neighbors | 52.22 | 41.13 |
| Annoy nearest neighbors | 54.22 | 41.26 |
| Top 10 Annoy nearest neighbors + no verbs | 63.33 | 54.67 |
| Top 110 Annoy nearest neighbors + no verbs | 73.86 | 65.72 |
| Top 110 Annoy nearest neighbors + no verbs + retrofitted | 76.14 | 68.45 |
| Morphological and top 110 Annoy nearest neighbors + no verbs + retrofitted + syntactic | 85.23 | 80.54 |

Table 6: Improvements in accuracy of the different versions developed to disambiguate among noun classes on two Runyankore test sets

The first version of the tool was based only on the semantics, and used the classical method to obtain the nearest neighbors. There was an improvement in performance when Annoy[4] nearest neighbors were used instead of the classical method, as shown in Table 6. Next, we used a set of ten Annoy nearest neighbors, omitting any verbs from this set because the verb root does not contain any noun class information, and this also improved the accuracy. We investigated the optimum number of nearest neighbors that result in the highest accuracy, starting from 10 up to 1000. We found 110 to be

[4]For details on the Annoy algorithm, see https://pypi.org/project/annoy/1.0.3/.

the optimum number that results in the accuracies shown in Table 6. This number is also large enough to include the concords required for the syntactic approach.

The next version of the tool added on to the previous one by using the retrofit algorithm by Faruqui et al. (2015) in an attempt to enable words to be related semantically in a better manner by having similar vector representations. We extracted a lexicon of nouns from a Runyankore dictionary (Taylor, 2009), arranged according to their noun classes, and used it to retrofit the pre-trained word vectors. We then calculated the accuracy at this point and obtained the results shown in Table 6. This represents the current combination of methods associated with the semantic and syntactic approaches. The final version starts with a morphological method to handle nouns with unique class prefixes in the input datasets, followed by the semantic and syntactic methods when disambiguation is required, in order to obtain the final accuracies of 85.23% and 80.54% on Set1r and Set2r, respectively.

Next, we report on the results from evaluating this combined approach on Runyankore, Luganda, and Kinyarwanda.

### 3.5 Results on Evaluation

We carried out an evaluation of the tool developed to disambiguate and determine a noun's class along two lines of enquiry: (1) whether a combined approach achieves better results than a morphological or semantic method alone; and (2) whether, without changing the underlying resources that are the basis of the semantic approach, a combined approach can still achieve better results than the individual approaches when applied to languages other than Runyankore. Table 7 shows the results of the evaluation for Runyankore, Luganda, and Kinyarwanda ("NCS" refers to "Noun Class System"). The metric for evaluation is accuracy, determined as the percentage of correct noun classes determined over the test sets.

The results in Table 7 show clearly that a combined syntactic and semantic approach achieves better results than the individual approaches. Though not shown in Table 7, accuracies of 84.88% and 70.67% were obtain on SectC for Runyankore and Luganda, respectively (there was no dataset for SetC in Kinyarwanda). Additionally, the results show that an improvement in performance is achieved in Luganda and Kinyarwanda despite the underlying semantic resources belonging to Runyankore. This is an important result given the under-resourced state of Bantu languages, where the reuse of resources during text processing will always be advantageous, because developing the same resources in Runyankore for another Bantu language requires significant time and effort (Byamugisha, 2020).

| Approach | Runyankore | Luganda | Kinyarwanda |
|---|---|---|---|
| Morphological only | 69.23 | 57.53 | 43.94 |
| Semantic only | 66.67 | 47.95 | 40.91 |
| Combination, with Runyankore NCS | 87.18 | 67.12 | 31.82 |
| Combination, with language NCS | 87.18 | 73.97 | 63.64 |

Table 7: Results of the evaluation on the morphological only, semantic only, and combined syntactic and semantic approaches on Runyankore, Luganda, and Kinyarwanda datasets

While it is not surprising that the best results are achieved when the syntactic approach is based on the concords of a test language's noun class system, it is nonetheless interesting that there is a 10% improvement in accuracy for Luganda even when the syntactic approach uses Runyankore's noun class system. This can be explained by how similar their noun class systems are (see Table 2), suggesting, possibly, a potential for direct reuse of the tool among closely related languages.

## 4 Discussion

We developed an approach to disambiguate among noun classes in order to determine a noun's class that combines a noun's syntax and semantics in order to achieve the best results. Our approach is in line with literature on the Bantu noun class system that states that the assignment of nouns to a class is based on semantic generalizations of the types of nouns in each class, as well as morphologically, according to their class prefixes, and syntactically, according to the concords they take (Katamba, 2003). While our work is focused on Bantu languages, it is important to note that noun class systems are a strong areal feature in Africa, with an estimated two-thirds of the languages on the continent having noun classes (Katamba, 2003). The theory on which our approach is based might be applicable

to another family of languages in Africa.

That a combined syntactic and semantic approach is necessary during text processing for Bantu languages is not novel, as this is the basis of the tools on noun pluralization for Runyankore and isiZulu (Byamugisha et al., 2016) and chiShona, isiXhosa, Kikuyu, Kinyarwanda, and Luganda (Byamugisha et al., 2018). What is emphasized by our results, however, is that, for the semantics, a little goes a long way. Only 385 singular nouns were used in the generation of the Runyankore datasets (Byamugisha, 2020) on which the word vectors were trained; yet, as seen in the results in Table 6, an accuracy of over 80% was achieved on Set2r that comprises over 2000 nouns.

The main reason for needing a semantic approach is to overcome noun class ambiguity, a situation where nouns belonging to different noun classes have the same class prefix. Therefore, though useful, the reuse of the resources in one language to determine the semantics of another language is limited if the source language does not account for the nouns affected in the target language. For example, in addition to classes 1 and 3 having the same prefix in Luganda and Kinyarwanda, classes 5, 9, and 23 in Luganda also have the same prefix, *e-*, and classes 5 and 9 in Kinyarwanda also have the same prefix, *i-*; yet their semantic distinctions are not all captured in the Runyankore resources. Having language-specific resources is the desired outcome, though the ability to reuse resources provides a benefit.

The increase in accuracy on test sets in languages that are different from the resources on which the semantics are determined is also a notable finding. Though the selection of Luganda and Kinyarwanda was purposively based on the availability of resources, their results cannot be removed entirely from their grouping in Guthrie zones as compared to Runyankore. Guthrie zones are a referential classification of Bantu languages which groups together languages with similar linguistic features (such as phonetic, semantic, and syntactic) that are geographically colocated, without presupposing their historical relatedness (Schadeberg, 2003). Guthrie zones categorize Bantu languages into 16 geographic zones, which are labeled using the letters A, B, C, D, E, F, G, H, J, K, L, M, N, P, R, and S; and these are further subdivided into decades (zone J.10, where Runyankore belongs, contains individual languages labeled from J.11 to J.19, while J.20, J.30, etc. represent different zones) (Nurse and Philippson, 2003; Schadeberg, 2003; Maho, 2009). Figure 3 shows how this classification covers the Bantu languages throughout the African continent.

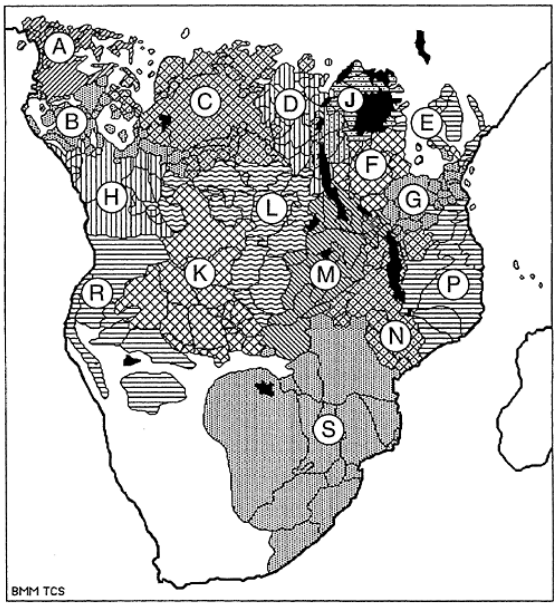


Figure 3: The spread and classification of Bantu languages into Guthrie zones (Maho, 2009)

Runyankore and Luganda are in zone J.10, while Kinyarwanda is in zone J.60 (Maho, 2009). When considered from the perspective of Guthrie zone classification, the results in Table 7 provide evidence, though limited, for the reuse of language resources both within and across Guthrie zones, with the best results obtained for the former. However, this needs to be investigated further before any conclusions can be made.

## 5 Conclusion

In this paper, we have presented the use of a combined syntactic and semantic approach to address noun class ambiguity when determining a singular noun's class. The semantic approach, which is based on the type of the noun, is necessary to perform noun class disambiguation, which addresses the main limitation of the morphological approach, that is, the presence of nouns with the same prefix but belonging to different noun classes. We used a noun's nearest neighbors in word vectors as representations of a noun's semantic generalizations, and a noun's concords, together with its morphology, as representations of a noun's syntactic association. We developed a tool based on our

combined approach using pre-trained word vectors in Runyankore, and evaluated the accuracy of the tool with two Runyankore datasets, achieving up to 87.18% accuracy. We also evaluated the reusability of our approach to other Bantu languages, Luganda and Kinyarwanda, whilst relying on Runyankore resources for the semantic approach. We achieved accuracies of 73.97% in Luganda and 63.64% in Kinyarwanda. Additionally, for all three languages, the combined approach performed better than individual morphological and semantic approaches.

## References


Allen Asiimwe. 2014. *Definiteness and Specificity in Runyankore-Rukiga*. Ph.D. thesis, Stellenbosch University, Western Cape, South Africa.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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

Sonja E. Bosch and Laurette Pretorius. 2003. Building a computational morphological analyser/generator for Zulu using the Xerox finite-state tools. In *Workshop on Finite-State Methods in Natural Language Processing, 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 27–34, Budapest. Association for Computational Linguistics.

Sonja E. Bosch and Laurette Pretorius. 2004. Software tools for morphological tagging of Zulu corpora and lexicon development. In *4th International Language Resources and Evaluation Conference*, pages 1251–1254, Lisbon, Portugal.

Joan Byamugisha. 2020. Generating varied training corpora in Runyankore using a combined semantic and syntactic, pattern-grammar-based approach. In *13th International Conference on Natural Language Generation*, pages 273–282.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2017a. Evaluation of a Runyankore grammar engine for healthcare messages. In *10th International Conference on Natural Language Generation (INLG 2017)*, volume 4, pages 105–113, Santiago de Compostela, Spain. ACL.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2017b. Toward an nlg system for Bantu languages: first steps with Runyankore (demo). In *10th International Conference on Natural Language Generation (INLG 2017)*, volume 4, Santiago de Compostela, Spain. ACL.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2018. Pluralizing nouns in agglutinating Bantu languages. In *27th International Conference on Computational Linguistics (COLING 2018)*, pages 2633–2643, Santa Fe, New Mexico, USA. ACL.

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, pages 271–283, Konya, Turkey. Springer LNCS.

Guy De Pauw, Gilles-Maurice de Schryver, and Janneke van de Loo. 2012. Resource-light Bantu part-of-speech tagging. In *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SaLTMiL 8 - AfLaT 2012)*, pages 85–92, Istanbul, Turkey.

Guy De Pauw, Peter W. Wagacha, and Gilles-Maurice de Schryver. 2011. Exploring the SAWA corpus: Collection and deployment of a parallel corpus English-Swahili. *Language Resources and Evaluation*, 45(3):331–344.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. *arXiv:1411.4166 [cs]*.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Francis Katamba. 2003. Bantu nominal morphology. In *The Bantu Languages: Routledge Language Family Series 4*, chapter 7, pages 103–120. Taylor and Francis Routledge, London.

C. M. Keet, M. Xakaza, and L. Khumalo. 2017. Verbalising owl ontologies in isiZulu with python. In *14th Extended Semantic Web Conference (ESWC'17)*, volume 10577, pages 59–64, Portoroz, Slovenia. Springer LNCS.

C. Maria Keet and Langa Khumalo. 2014. Towards verbalizing logical theories in isiZulu. In *4th Workshop on Controlled Natural Languages (CNL'14)*, volume 8625, pages 78–89, Galway, Ireland. Springer LANI.

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.

Jouni Maho. 1999. *A Comparative Study of Bantu Noun Classes*. Ph.D. thesis, Goteborg University, Goteborg, Sweden.

Jouni Filip Maho. 2009. NUGL online: The online version of the updated Guthrie list, a referential classification of the Bantu languages.

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.

Derek Nurse and Gerard Philippson. 2003. Introduction. In *The Bantu Languages: Routledge Language Family Series 4*, chapter 1, pages 1–9. Taylor and Francis Routledge, London.

C. Thilo Schadeberg. 2003. Historical linguistics. In *The Bantu Languages: Routledge Language Family Series 4*, chapter 9, pages 143–181. Taylor and Francis Routledge, London.

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

Doreen Daphine Tayebwa. 2014. Demonstrative determiners in Runyankore-Rukiga. Master's thesis, Norwegian University of Science and Technology, Norway.

C. Taylor. 2009. *A Simplified Runyankore-Rukiga-English Dictionary*. Fountain Publishers, Kampala, Uganda.

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