THE STRUCTURED COMPACT TAG-SET FOR LUGANDA

Robert Ssali Balagadde[1] and Parvataneni Premchand[2]

[1, 2] Department of Computer Science and Engineering,
University College of Engineering, Osmania University, Hyderabad, India

ABSTRACT

A tag-set for tagging Luganda words has been in absence for quite a long time leading to absence of Luganda Language resources used in Computational Linguistic (CL) and Natural Language Processing (NLP). As a result, this research paper proposes and presents a Structured Compact Tag-set for Luganda (SCTL) in a bid to address this gap, and emphasis has been directed towards presenting the structures. SCTL incorporates a number of new concepts aimed at reducing redundancy in an annotated corpus. In line with this, Tag Length Minimization Strategies (TLMS) have been proposed and implemented in SCTL. The morpho-syntactic properties captured in SCTL were identified through conducting a morphological analysis and word categorization along the various parts of speeches (POS) of Luganda. To demonstrate the suitability of SCTL to tag Luganda text, a sample text extracted from Bukedde, an online Luganda news paper, has been tagged and presented; however, identification and validation of the various tags of SCTL is proposed as a component of continuity of this research work. This paper demonstrates how Concord Number (CN) captured in SCTL can be used to check conventional agreement between words. Storage Efficiencies, namely, η_t (individual) and η_{at} (batch) are novel metrics proposed in this research work, which can be used in evaluating how a particular tag-set is performing in terms of efficient storage usage at tag level and corpus (or batch) level respectively. Finding on the comparison of Storage Efficiencies (η_t and η_{at}) of tags from the four tag-sets of Luganda, Swahili, Russian and Northern Sotho, show that SCTL tags had the highest η_t therefore the highest η_{at} among the tags considered, due to the application of TLMS which maximizes η_t. Finding on the impact of TLMS on these tag-sets using Storage Efficiencies (η_t and η_{at}) as evaluation metrics show that there was a three-folds improvement to Swahili tags, a two-fold to Russian tags, and a 60\% to Northern Sotho tags. SCTL is associated with a number of advantages and have been presented herein. Conclusively, the advent of SCTL has opened the avenue of developing other NLP resources, especially, an annotated Luganda corpus. TLMS is very crucial in highly inflectional languages which have a lot of inherent morpho-syntactic information to capture, in bid to boost their tag storage efficiencies.

KEYWORDS

Structured Tag-set, Structured Compact Tag-set for Luganda (SCTL), Tag Length Minimization Strategies (TLMS), Storage Efficiency, *Concord Number (CN), Mophological Valency (MV).*

1. INTRODUCTION

Lack of a suitable tag-set to capture the morpho-syntactic properties of Luganda words, which properties are very important in the grammatical analysis of the language, has hampered the development of a useful annotated corpus - an NLP (Natural Language Processing) resource - which can be used to develop grammar analysing systems for Luganda, a Bantu language.The universal tag-set, Petrov et al., 2012 [1], is unsuitable for this purpose because it only captures the POS (Part of Speech) of the words which information is inadequate for Luganda grammar analysis. Other Bantu tag-set (SWATWOL - Arvi Hurskainen, 2004 [2], Northern Sotho tag-set -

DOI: 10.5121/ijnlc.2016.5401

Taljard E. et al, 2008[3]; Gertrud Faab et al, 2009 [4], among other) are language specific and consequently, unsuitable for this purpose. The Multilingual Morpho-syntactic Specifications (Erjavec, 2004 [5]; 2009 [6]; 2010 [7]) are not exhaustive leaving out a lot of essential morpho-syntactic properties of Luganda which is important in the grammar analysis. Examples of excluded properties include: the state of noun and adjective; the 14 noun classes as per the ETLC (Extended Traditional Luanda Classification) or 23 classes (proto Bantu noun classification); among others.

Luganda is an agglutinative language, and therefore, highly inflectional. To capture the morphological properties of such languages requires a large tag-set which is naturally manageable using a structured system. In this research work, we propose a compact structure system - namely, Structured Compact Tag-set for Luganda (SCTL) - to capture the morpho-syntactic property of Luganda language, identified in unpublished [8].
In addition, a compact system eliminates the need to encode unnecessary information about a particular POS as is the case with positional tag-set. Table illustrates this point. Note the numerous dashes used to represent 'not applicable' incorporated in the tag. These are eliminated in case of compact tag-set.

Table 1. Sample Tags from a Russian Positional Tag-set

Template	Description	Sample Word	Sample Tag
NNgync - - - - - - - a - -	Noun	Golos (voice)	NNMIS4- - - - - - - A- -
NNgync - - - - - - - a - -	Noun	lodkoj (boatsg:inst)	NNFIS7- - - - - - - A - -
NNgync - - - - - - - a - -	Noun	Kapusta (big bucks)	NNFIS1- - - - - - - A - -
ACg-n - - - - - - - - a - -	short adjective	Krasiv (Beautiful)	ACM - S- - - - - - - A - -

In the development of SCTL, the issue of minimisation of tag length (TL) has been given emphasis because it leads to elimination of annotated corpus storage wastage (storage redundancy). In line with this, a number of strategies have been proposed and implemented in SCTL in a bid to minimize TL as elucidated in section 3. The compactness in SCTL is derived from the implementation of these TL minimization strategies (TLMS), which has added another dimension to the initial meaning of compactness (that is, eliminating dashes in the tag; see Table).Section 4 presents the actual structures making up SCTL capturing the various morpho-syntactic property of the various Luganda POS.

Other than tagging words as demonstrated in Section 5, SCTL can be applied to check the conventional agreement between words as evident in Section 6 and Table 3

SCTL has been evaluated and compared with other tag-sets, first, in order to show how efficient it uses storage resources, and second, to articulate some of its advantages as indicated in Section 7. The overall SCTL advantages have been elucidated in Section 8.

2. RELATED WORKS

2.1 Tag-set Design

There are several criteria to consider when developing a morpho-syntactic tag system for a language. These include: degree of relevant linguistic details, tag-set size; and, uniformity.
Elworthy, 1995 [9] distinguished external and internal criteria for tag-set design. The external criterion dictates that the tag-set must be capable of providing the linguistic (for example,

syntactic or morphological) distinctions required in the output corpora; while the internal criterion on tag-sets ensures the effectiveness of the design criterion.

Elworthy designed an experiment to explore the relationship between tagging accuracy and the nature of the tag-set, using corpora in English, French, and Swedish. The experiment addressed the internal design criterion. The aim of the experiment was to determine, crudely, whether a bigger tag-set is better than a smaller one, or whether external criteria requiring human intervention should be used to choose the best tag-set. General analysis of the results showed that there is no consistent relationship between the size of the tag-set and the tagging accuracy.

Elworthy's general conclusion is that the external criterion should be the main prerequisite in dominating tag-set design; and further suggests that what is important is to choose the tag-set required for the application, rather than to optimize it for the tagger.

Generally, the tag-set size for highly inflected language - like Telegu or Czech or Russian or Luganda - is typically far bigger than for a sparsely inflected language like, English; and it would seem obvious that the size of a tag-set would be negatively correlated with tagging accuracy. The reason being that for a smaller tag-set, there are fewer choices to be made, thus there is less opportunity for an error. However, Elworthy proved to the contrary.

2.2 Types of Tag-sets

There are many ways to classify morphological tag-sets. In this research work, two categories are identified:

1. Atomic: tags are atomic symbols without any formal internal structure, Cloeren Jan, 1993 [10]. Examples include the Penn TreeBank tag-set, Marcus et al., 1993 [11]; Brown Corpus Tag-set, Francis and Kacera, 1979 [12]; 1982 [13]; among others. Notably, even in this tag set, some structure could be found, however, it is rather ad-hoc and very limited.
2. Structured: tags can be decomposed into sub-tags each tagging a particular feature. Any tag-set capturing the morphological features of a richly inflected language is necessarily large. A natural way to make the tags manageable is to use a structured system, in which a tag is a composition of tags each coming from a much smaller and simpler atomic tag-set tagging a particular morpho-syntactic property (e.g., state, class or tense).For large tag-sets, a structured system has many practical benefits, as explained by Jirka and Feldman, 2010 [14] including learnability, systematic description, among others.It is worthwhile noting that it is trivial to view a structured tag-set as an atomic tag-set - for example, by assigning a unique natural number to each tag - while the reverse is an uphill task.

There are two types of structured tags, namely:

a) Positional: Examples include: Czech Positional Tag-set, Hajic, 2004 [15] and MULTEXT-East Tag-set, Ide and Veronis, 1994 [16].Some of its characteristics include: tags are sequences of values encoding individual morphological features; all tags have the same length, encoding all the features identified for the tag-set; and, features not applicable for a particular word have a N/A value or dash.

b) Compact: Examples include: Multext-East, Erjavec, 2004 [5]; 2009 [6]; 2010 [7]; Czech Compact tag-sets, Hajic, 2004 [15]; and CLiC-TALP, Civit, 2000 [17].

Some of its characteristic include: tags are sequences of values encoding individual morphological features; the N/A values or dashes are left out; and positional interpretation vary across different POS.

3. TAG LENGTH MINIMIZATION STRATEGIES (TLMS)

The issue of minimising Tag Length (TL) is given emphasis in the development of SCTL. The importance of these strategies is derived from the fact that tags are used to provide addition information on every word in an annotated corpus, and therefore, the shorter they are without compromising on the information captured, the less storage space they take. In a bid to throw more light on this, an example, which uses ATLS (Average Tag Length per Structure) to estimate storage gain, is explained in the next paragraph.

We know that the TLs for the various POSs in a compact structure tag-set are different and for a particular POS the TL is the same. To easy computation, we take ATLS (μ_L), which is determined using Equation 1 to represent the length of each tag in the tag-set.

$$\mu_L = (\sum L_i) / n \quad (1)$$

where L_i is the length of structure i and n - the number of structures in the tag-set. Suppose we have two tag-sets which capture the same information but with different μ_L of 4 and 15 characters respectively. A comparison between the two tag-sets shows a difference of 11 characters. Assume that the two are used to tag a corpus of 400M words, then there will be a gain or save in storage space of 11 x 400 x 10^6 Bytes (approximately 4.4 GB), which is enough space to save another 4.4 billion characters of words and tags.

In this context, the following TL Minimisation Strategies (TLMS) have been proposed and implemented in SCTL:

- The use of Concord Number (CN) which represents three aspects (ETLC number, plural number, and person) into a single entity (number). CN was created as a result of borrowing a leaf from the concept of redundancy reduction in relational databases by applying Codd's normalisation forms, F. Codd, 1990 [18]. This concept eliminates storage redundancy (or storage space wastage) and
- Table shows CN used in SCTL to capture the corresponding information.

Table 2. Concord Numbers used in the categorisation of Luganda words encoded using either Hexatrigesimal or Duotrigesimal, standard positional numbering systems.

ETLC No	NUMBER	PERSON	CN	ETLC No	NUMBER	PERSON	CN
1	Singular	3rd	1	VIII	Singular	3rd	F (15)
	Plural		2		Plural		G (16)
11	Singular	3rd	3	IX	Singular	3rd	H (17)
	Plural		4		Plural		I (18)
111	Singular	3rd	5	X		3rd	J (19)
	Plural		6	XI		3rd	K (20)
IV	Singular	3rd	7	XII		3rd	L (21)
	Plural		8	XIII		3rd	M (22)
V	Singular	3rd	9	XIV		3rd	N (23)
	Plural		A (10)	I	Singular	1st	O (24)
VI	Singular	3rd	B (11)		Plural		P (25)
	Plural		C (12)		Singular	2nd	Q (26)
VI	Singular	3rd	D (13)		Plural		R (27)
	Plural		E (14)				

The significance of the used of CN can be appreciated by looking at the following example.

suppose we would like to capture information about the three affix pronouns in a trivalent verb form (refer to Section 4.4.4 for more details) in a tag; and for each affix pronoun, we require to capture information about it class, number and person, that mean we require a minimum of nine characters to capture this information in a tag. This translates to TL of 15 characters (9+6). However, with the advent of CN, we require only three characters, which translates to TL of 9; thus, reducing the tag length by 40%.

- Coding numeric and non-numeric information that is captured in the tag-set with a single character. This concept was adapted from the tag-sets of Multext-East (Erjavec, 2004 [5]; 2009 [6]; 2010 [7]). We propose Hexatrigesimal system to capture numeric information exceeding 10 for adaptation purposes for other Bantu languages or other agglutinative languages.

- The use of CN instead of a four character tag to represent affix pronouns in the verb forms. This reduced the TL especially with the trivalent verb forms where the TL was reduced by two folds which is substantiated in the following paragraphs.

  The classification of pronouns into affix and non-affix pronouns has also attributed to the process of minimising TL. The following example shows the importance of this splitting. Suppose that the affix and non affix pronouns were placed in one group; that mean they are represented by a four character tag as by the pronoun structure in
  Table . Suppose we require a tag for trivalent verb form whose structure is shown in Table . In this case we must capture information about the subject prefix (pronoun), primary and secondary object infix (pronoun), among others. The tag length would be 4*3+6=18 characters as opposed to 9 character tag obtained as a result of splitting the pronouns into affix and non affix and representing the affix with a single character, the CN. This results into a reduction in tag length of two folds.

- The use of a compact structured system which ensures that only information relevant to a given POS is encoded, thereby eliminating the use of dashes as explained in the introduction and Table . This eliminates dashes at POS level.

- The use of different structures for a given POS in case grouping the POS subcategories in one structure results into more than one "not applicable situations" (dashes). This strategy eliminates dashes at subcategory level and has been used in monovalent verb forms (refer to Section 4.4.2 for more details).

The compactness in SCTL is derived from the implementation of the TLMS, which adds another dimension to the earlier meaning of compactness which was limited to dropping of N/A values or dashes in positional tag-set in a bid to form compact tag-set. The later meaning is a subset of the former.

4. STRUCTURAL COMPONENTS OF SCTL

In this section, we propose and present the detailed description of the various SCTL structures - grouped alone the POS dimension - used for generation of the tags used for tagging of Luganda words or text.

4.1 Noun Structure

The noun structure produces tag of tag length (TL) of a five characters (that is to say, TL=5) which always begins with character 'N' which stands for 'Noun'. Table shows the values or attributes associated with each position in the noun tag.

Table 3 Attributes for each position of a Luganda noun tag with TL=5

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	N	Noun
2	Type	Specifies Whether noun is common or proper	C	Common Noun
			P	Proper Nouns
3	Concord Number (CN)	Specifies the CN of the noun if applicable.	1-N	Numbers 1 to 23
			X	Not Applicable
4	State	Specifies the state of the noun defined by the presence of or absence of initial vowel (IV).	T	Topic State
			B	Base State
			X	Not Applicable
5	numerals	Specifies the numeric type (except for cardinal number 1-5 which are adjective)	C	Cardinal & Fractional number
			O	Ordinal Number
			X	Not Applicable

The other information encoded or captured by the structure includes information about: whether the word is a common noun or proper noun; CN of the noun; state of the noun; and finally, if the noun is numerical, what numerical type is it.

Table shows examples of words tagged using tag generated by noun and adjective structures. Note that, first, the noun and its qualified adjective have the same CN demonstrating the ease of checking agreement using tag, which feature can be exploited by an NLP application; second, Luganda adjectives come after their qualified nouns which is opposite to English word order.

4.2 Adjective Structure

The adjective structure produces a tag with TL=3 which always begins with character 'A' which stands for 'Adjective'. Table shows the values or attributes associated with each position in the adjective tag. The following information is encoded: the POS of the word; CN of the adjective; and finally, state of the adjective.

Table shows examples of words tagged using tag generated by noun and adjective structure.

Table 4 Attributes for each position of a Luganda adjective tag with TL=3

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	A	Adjective
2	Concord Number (CN)	Specifies the CN of the adjective if applicable	1-N	Numbers 1 to 23
3	State	Specifies the state of the adjective defined by the presence of or absence of initial vowel (IV)	T	Topic State
			B	Base State

Table 5 Examples of Tagged Luganda Noun and Adjective.

Luganda Noun and Adjective and their English Translations	Tagged Noun and Adjective
Ente (cow) ennungi (beautiful)	(ente, NC5TX) (ennungi, A5T)
ente (cows) ennungi (beautiful)	(ente, NC6TX) (ennungi, A6T)
ekitabo (book) ekirungi (beautifu)l	(ekitabo, NC7TX) (ekirungi, A7T)
ebitabo (books) birungi (are beautiful)	(ebitabo, NC8TX) (ebirungi, A8B)
Kizito (Kizito) mulungi (is handsome)	(Kizito, NPXXX) (mulungi, A1B)
amayumba (houses) abiri (two)	(amayumba NCATX) (abiri, AAT)

4.3 Pronoun Structures

Table 6 Attributes for each position of a Luganda tag with TL=4 for a self standing pronouns

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	P	Pronoun
2	Sub-Category	Specifies the sub-categories of the pronoun	E	Emphatic
			D	Demonstrative
			N	'na' Pronouns
			R	Object Relative
			O	Companion SG
			C	Companion PL
			T	Possessive with IV
			B	Possessive without IV
			P	Possessive Personal
			I	Possessive Interclass
			S	Possessive Self Standing
3	Concord Number (CN)	Specifies the CN of the pronoun .	1-R	Numbers 1 to 27
4	Concord Number 2 CN)	Specifies the second CN of the Interclass pronoun.	1-N	Numbers 1 to 23
			X	Not Applicable

Luganda pronouns are sub-divided into affix pronouns and self-standing pronouns (non-affix pronouns). However, only the structure for self standing pronouns is presented. The affix pronouns are not self standing, and therefore, they do not require a tag. What is required from them is the CN. This is in line with TLMS, explained in Section 3.

Table shows the structure for self standing pronoun tag with a length of four characters. The tag always begins with characters 'P'. Information captured by the tag include: the POS of the word and in this case the value is always 'P'; the sub category under the self standing pronoun; CN of the pronoun; and finally, CN of the interclass pronoun.

4.4 Verb Structures

There are five structures considered under verbs and these structures are categorised in accordance with morphological valency (MV), a new concept introduced and proposed in this research work. MV, which has eased the process of classifying Luganda verb forms, is defined as the number of affix pronouns incorporated in the verb morphological form. A tag for a verb form always begins with a character 'V' which stands for verb. More derails on verb classification by MV is presented in unpublished [8].

4.1 Structure for Avalent Verb Form

Table shows attributes for each position of a Luganda tag for verbs of MV=0 and the structure produces a tag with TL=3 which always begins with characters 'V0'. Other information captured by the tag is whether there is a negative morpheme in the verb form.

Table 7 Attributes for each position of a Luganda tag for verbs of MV=0 and TL=3

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	V	Verb
2	Valency	Specifies the MV of the verb form (VF)	0	Zero
3	Polarity	Specifies the polarity of the verb form	+	Affirmative
			-	Negative

4.2 Structures for Monovalent Verb Form

There are two verb form structures consider under Monovalent Verb Form, namely, one for Ordinary Monovalent Verb Form (OMVF) and the other for Luganda Copulae and Actuals (LCA). Since the two tags generated for the verb forms both begin with 'V1', they are disambiguated by their corresponding tag lengths, which are different.

Table shows attributes for each position of a Luganda tag for OMVF and the structure produces a seven character tag which always begins with characters 'V1'. Other than information on POS, MV, presence of a negative morpheme, and modification on verb stem, other information encoded in the tag include: the existence of an initial vowel attached to the subject prefix to form a subject relative clause; CN of the Subject Prefix used in the verb form; and finally, the type of tense incorporated in the verb form.

The second structure, shown in Table is for LCA and produces a tag with TL=4 and the tag always begins with 'V1'. Other than information on POS and MV, other information encoded in the tag include: whether the verb form is a copula or actual; and finally, CN of the copula or actual.

4.3 Structure for Divalent Verb form

Table shows attributes for each position of a Luganda tag for ordinary verbs of MV=2 and the structure produces an eight character tag which always begins with characters 'V2'. In addition to the information captured by the OMVF, this structure captures CN of object prefix used in the verb form.

4.4 Structure for Trivalent Verb form

Table shows attributes for each position of a Luganda tag for verb form of MV=3. The structure produces a nine character tag which always begins with characters 'V3'.

Table 8 Attributes for each position of a Luganda tag for verbs of MV=1

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	v	Verb
2	Valency	Specifies the MV of the verb form (VF)	1	One
3	Polarity	Specifies the polarity of the verb form	+	Affirmative
			-	Negative
4	State	Specifies the state of the VF in terms of forming Subject relative clause (SRC)	T	With IV (SRC)
			B	Without IV
5	Subject Prefix (SP)	Specifies the concord number (CN) of the SP used in the VF	1-R	Numbers 1 to 27
6	Tense	Specifies the tense of the VF	A	Present
			B	Present Perfect
			C	Near past
			D	Far past
			E	Near Future
			F	Far Future
			G	Subjunctive
			H	Conditional
			I	Imperative
			J	Still
			K	So far
			L	Not Yet
			M	Narrative
7	Derivation	Specifies the type of modification (mod) made on the root of the VF	A	Passive
			B	Reflective
			C	Reduplicative
			D	Applicative
			E	Causative
			F	Capable
			G	Neuter
			H	Reversive
			I	Reciprical
			J	Combined mod
			K	Clitic
			N	Normal

Table 9 Attributes for each position of a Luganda tag for copulae and actuals

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	V	Verb
2	Valency	Specifies the MV of the verb form (VF)	1	One
3	Category	Specifies the VF is copula or actual	C	Copula
			A	Actual
4	CN	Specifies the CN of Copula or actual	1-N	Numbers 1-23

In addition to information captured by the divalent verb form structure, the structure for trivalent verb form captures the CN of the secondary object prefix used in the verb form.

4.5 Adverb Structure

The adverb structure, shown in **Error! Reference source not found.** produces a three character tag which always begins with 'D' which stands for adverb. Other information captured include: type of the adverb; and, if applicable, CN of the pronoun attached to the noun dependent adverb.

Table 10 Attributes for each position of a Luganda tag for verbs of MV=2.

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	v	Verb
2	Valency	Specifies the MV of the verb form (VF)	2	Two
3	Polarity	Specifies the polarity of the verb form	+	Affirmative
			-	Negative
4	State	Specifies the state of the VF in terms of forming Subject relative clause (SRC)	T	With IV (SRC)
			B	Without IV
5	Subject Prefix (SP)	Specifies the concord number (CN) of the SP used in the VF	1-R	Numbers 1 to 27
6	Primary Object Infix (POI)	Specifies the CN of POI used in the VF	1-R	Numbers 1 to 27
7	Tense	Specifies the tense of the VF	A	Present
			B	Present Perfect
			C	Near past
			D	Far past
			E	Near Future
			F	Far Future
			G	Subjunctive
			H	Conditional
			I	Imperative
			J	Still
			K	So far
			L	Not Yet
			M	Narrative
8	Derivation	Specifies the type of modification (mod) made on the root of the VF	A	Passive
			B	Reflective
			C	Reduplicative
			D	Applicative
			E	Causative
			F	Capable
			G	Neuter
			H	Reversive
			I	Reciprical
			J	Combined mod
			K	Clitic
			N	Normal

4.6 Particle Structure

Table shows the Luganda particle structure which produces two characters tags which always begins with character 'T' which stands for 'Particle'. Other information captured is the type of the particle.

4.7 Punctuation Structure

The punctuation structure, shown in Table produces a three character tag which always begins with character 'Z' which stands for 'punctuation'. Other information captured by the tag include: type of the punctuation mark; and the sub category under each type of the punctuation mark which correlates with the number of punctuation mark available on a Qwerty keyboard.

One advantage with this structure is that it addresses disambiguation issues associated with the use of firstly, an apostrophe both as a quotation mark and as an inter word mark; and secondly a full stop as an end of sentence mark and as a period for other use (for example, separation of characters in abbreviations). Consequently, the end result is the elimination of the need for developing algorithm to address this disambiguation.

Table 11 Attributes for each position of a Luganda tag for verbs of MV=3

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	v	Verb
2	Valency	Specifies the MV of the verb form (VF)	3	Three
3	Polarity	Specifies the polarity of the verb form	+	Affirmative
			-	Negative
4	State	Specifies the state of the VF in terms of forming Subject relative clause (SRC)	T	With IV (SRC)
			B	Without IV
5	Subject Prefix (SP)	Specifies the concord number (CN) of the SP used in the VF	1-R	Numbers 1 to 27
6	Primary Object Infix (POI)	Specifies the CN of POI used in the VF	1-R	Numbers 1 to 27
7	Secondary Object infix (SOI)	Specifies the CN of SOI used in the VF	1-R	Numbers 1 to 27
8	Tense	Specifies the tense of the VF	A	Present
			B	Present Perfect
			C	Near past
			D	Far past
			E	Near Future
			F	Far Future
			G	Subjunctive
			H	Conditional
			I	Imperative
			J	Still
			K	So far
			L	Not Yet
			M	Narrative
9	Derivation	Specifies the type of modification (mod) made on the root of the VF	A	Passive
			B	Reflective
			C	Reduplicative
			D	Applicative
			E	Causative
			F	Capable
			G	Neuter
			H	Reversive
			I	Reciprical
			J	Combined mod
			K	Clitic
			N	Normal

Table 12 Attributes for each position of a Luganda tag for adverbs

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	D	Adverb
2	Type	Specifies the category of the adverb	T	Time
			L	Location
			M	Manner
			Q	Quality
3	Concord Number (CN)	Specifies the CN of the affix pronoun attached to the noun dependent adverb	1-R	Numbers 1 to 27
			X	Not Applicable

4.8 Unclassified Word Structure

Table show the structure of unclassified words which produces a single character tag. All foreign words, abbreviation, non-Luganda words that are not proper nouns are group under this category

Table 13 Attributes for each position of a Luganda tag for particles

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	T	Particle
2	Type	Specifies the category of the particle	P	Preposition
			C	Conjugation
			I	Interjection
			R	Interrogation
			O	Others

Table 14 Attributes for each position of a Luganda tag for punctuation marks

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	Z	Punctuation
2	Type	Specifies the category of the punctuation mark	S	End of sentence mark
			W	Inter Word Apostrophe
			Q	Quotation mark
			B	Bracket Delimiters
			P	Period
			C	Comma
			O	Colon
			D	Dash
			E	Ellipsis
			X	Others
3	Sub-Category	Specifies the sub-categories under the various categories	S: ., ;, ?, !	The 'S group values
			W: '	The 'W' group values
			Q: ', ",	The 'Q' group values
			B: (,),{,},[,],<,>	The 'B' group values
			P: .	The 'P' group values
			C: ,	The 'C' group values
			O: :	The 'O' group values
			D: -	The 'D' group values
			E: .	The 'E' group values
			X: +,*,%,etc	The 'X' group values

Table 15 Attributes for each position of a Luganda tag for unclassifiable words

POSITION	NAME	DESCRIPTION	VALUE	DESCRIBTION
1	POS	Part of Speech	X	Unclassifiable Word

5. AN EXAMPLE TO DEMONSTRATE THE TAGGING OF LUGANDA WORDS WITH SCTL

This is an example to demonstrate the use of SCTL to tag Luganda text. In this context, sample Luganda text was extracted from the agriculture section of an online local news paper, Bukedde, and presented in three parts:

- Part1: Luganda text with word to word translation in English.
- Part II: Equivalent English translation
- Part III: Luganda text tagged with SCTL

5.1 Part I: Luganda Text With Word to Word Translation in English

Mu (in) kaweefube (efforts) w'okutumbula (of boosting) ebyobulimi (crop farming) n'obulunzi (and animal farming) kkampuni (company) ya (of) Vision Group efulumya (that publish) ne (also) Bukedde ng'ekolagana (in collaboration) ne (with) bbanka (bank) ya (of) DFCU, ekitebe (embassy) kya (of) Budaaki (Denmark) mu (in) Uganda, kkampuni (company) y'ennyonyi (of aeroplain) eya (of) KLM bajja (they will) kusunsula (choose) mu (from) balimi ('plant'

farmers)n'abalunzi ('animal' farmer) abeetabye (who participarted) mu (in) mpaka (compitition) z'omulimi (for farmer) asinga (he / she win) omwaka (year) guno (this).

Enkoko (chickens) zino (these) zaakola (they did) bulungi (good / well) ddala (very) era (and) mu (in / at) kiseera (time / moment) kino (this) nnina (I have) enkoko (chicken) 6,000 ate (and) nga (after) mmaze (I have acomplised) okubuukinga (booking) obukoko (chicks) 4,000. Zino (these) zimpa (they give me) ttule (tray) z'amagi (of eggs) 120 - 125 olunaku (a day) kye (which) ndaba (I see / realize) ng'ate (that) tezikoze (not they worked / performed) bulungi (good / well) kubanga (because) natawaanyizibwa (I was disturbed) nnyo (very much) ekirwadde (disease) kya (of) Newcastle nga (when) zikyali (they were) nto (young) ne (and) zitakula (they not grow) bulungi (good) kuba (otherwise) zandibadde (they would) zimpa (they me give) nga (approximately) ttule (trays) 160.

5.2 Part II Equivalent English Translation

In a bid to boost Agriculture, Vision Group Company which also publishes Bukedde, in collaboration with DFCU bank, Embassy of Denmark in Uganda, and KLM Airline Company, will choose from amongst those who have participated in the farmer's competition the winner of this year.

These chickens did very well and at this moment I have 6000 chickens and I have accomplished booking 4000 chicks. These give me 120 - 125 trays of eggs daily, which performance, I realized, is not good because I was disturbed to a great extent by the Newcastle disease when they were young, and therefore, they did not grow very well, otherwise, they would have given me approximately160 trays.

5.3 Part III Luganda Text Tagged With SCTL

Mu_TP kaweefube_NC1BX w_PB1X '_ZW' okutumbula_NCHTX ebyobulimi_NC8TX n_TC '_ZW' obulunzi_NCCTX kkampuni_NC5XX ya_PB5X Vision_NPXXX Group_NPXXX efulumya_V1+T5AN ne_TC Bukedde_NPXXX ng_TP '_ZW' ekolagana_V1+T5AI ne_TC bbanka_NC5XX\L ya_PB5X DFCU_NPXXXL ,_ZC, ekitebe_NC7TX kya_PB7X Budaaki_NPXXX mu_TP Uganda_NPXXX ,_ZC, kkampuni_NC5XX y_PB5X '_ZW' ennyonyi_NC6TX eya_PT6X KLM_NPXXX bajja_V1+B2AN kusunsula_NCHBX mu_TP balimi_NC2BX n_TC '_ZW' abalunzi_NC2TX abeetabye_V1+T2BN mu_TP mpaka_NC6BX z_PB6X '_ZW' omulimi_NC1TX asinga_V1+B1AN omwaka_NC3TX guno_PD3X ._ZS.

Enkoko_NC6TX zino_PD6X zaakola_V1+B6DN bulungi_DMX ddala_DQX era_TC mu_TP kiseera_NC7BX kino_PD7X nnina_V1+BOAN enkoko_NC6TX 6 ,_ZC, 000_NCBBC ate_TC nga_TP mmaze_V1+BOBN okubuukinga_NCHTX obukoko_NCCTX 4 ,_ZC, 000_NCETC ._ZS. Zino_PD6X zimpa_V2+B6OAN ttule_NC6BX z_PB6X '_ZW' amagi_NCATX 120_NCATC -ZD- 125_A6X olunaku_NCDTX kye_PR7X ndaba_V1+BOAN ng_TC '_PW' ate_TC tezikoze_V1-B6BN bulungi_DMX kubanga_TC natawaanyizibwa_V1+BODJ nnyo_DQX ekirwadde_NC7TX kya_PB7X Newcastle_NPXXX nga_TC zikyali_V1+B6AN nto_AO6B ne_TC zitakula_V1-B6AN bulungi_DMX kuba_TC zandibadde_V1+B6HN zimpa_V2+B6OAN nga_DQX ttule_NC6BX 160_NC6BX ._ZS.

6. CHECKING CONVENTIONAL AGREEMENT USING CONCORD NUMBER

Luganda is noun-centric in the sense that most words in a sentence agree with the noun. Agreement is by noun class, number, and person; and is indicated with prefixes and infixes attached to the beginning of word stems. In this context, CN can be used to check this agreement

since words with same ETLC number, number (plural) and person have the same CN. In the same vein, a tag-set which captures only the POS of the words in the language, like the universal tag-set (Slav Petrov, et al, 2012 [1]), is unsuitable for this purpose.

Table 16 shows tagged Luganda sentences to demonstrate that words which agree with one another (conventional agreement) actually have the same CN. This feature can be taken advantage of in NLP applications checking agreement by simply checking whether their CNs are the same.

Table 16 Tagged Luganda sentences using SCTL. Note that words which agree with one another (conventional agreement) have the same CN.

Luganda sentence and word to word English Translation	**English Equivalent**	**Tagged Sentence**
Enkoko (chicken) zino (these) zaakola (they performed) bulungi (well) ddala_DQX (very)	These Chicken performed Very well	Enkoko (NC6TX) zino (PD6X) zaakola (V1+B6DN) bulungi (DMX) ddala (DQX)
Ogwo (that) gwokka (only) omuti (tree) mulungi.(is beautiful) .	Only that tree is beautiful	Ogwo (PD3X) gwokka (DM3) omuti (NC3TX) mulungi.(A3B)

7. EVALUATION AND COMPARISON OF SCTL WITH OTHER TAG-SETS

7.1 Introduction

Given that SCTL is the first tag-set of its kind developed for Luganda, it has not been possible to conduct meaningful comparison with other tag-sets of the language. However, we have endeavoured to compare SCTL with tag-sets of other inflectional languages, namely, a positional tag-set - Russian positional tag-set, Jirka Hana and Feldman Anna, 2010 [14] - and two atomic tag-sets for Bantu languages which uses a two level tagging process (that is, Swahili Tag-set of SWATWOL - Arvi Hurskainen, 2004 [2] and Northern Sotho tag-set - Taljard E. et al, 2008 [3]; Gertrud Faa et al, 2009 [4]).Table 17 shows a qualitative comparison between these tag-sets articulating the general differences between them.

Table 17 General Comparison of SCTL, Russian tag-set SWATWOL tag-set and Northern Sotho tag-set

Dimension	**SCTL**	**Russian**	**SWATWOL**
Tag-Set Type	Structured compact	Structured Positional	Atomic
POS Identification From Tag	direct	`direct	Not direct rather obscure
Tagging Levels	one	one	two
Modifiability/ Adaptability	Easy	Easy	difficulty

The main difference between SCTL and the other three tag-sets is articulated in the application and implementation of TLMS (Tag Length Minimization Strategies), discussed in Section 3. In this context an experiment has been conducted as demonstrated in the next subsection.

7.2 Experimentation Setup

7.2.1 Comparison of Storage Efficiencies for the Tag-sets

In this experimentation a new metric is proposed for estimating the efficiency of storage usage by the tag-set at individual tag level, expressed in percentage and defined as the ratio of number of different relevant information captured by the tag (I_c) to the tag length (N_t) as shown in Equation 2.

$$\eta_t = 100 * I_c / N_t \quad (2)$$

It is worthwhile noting that the bigger the value of η_t for a given tag, the more efficient is the storage usage by the tag without compromising on the captured information. I_c can be determined by studying the tag-set itself.

Tags for three POS, namely noun, pronoun and punctuation were arbitrarily taken from the four language and their respective storage efficiencies (η_t) were calculated. The results were captured and presented in Table and are visualised in Figure.

Table 18 A comparisons of Storage Efficiencies (ηt) for the three POS tags for the four Languages

POS	Language	Tag Sample	I_c	N_t	η_t (%)
Noun	Luganda(L)	NC2TX	4	5	80
	Russian (R)	NNMIP4 - - - - - - - A - -	7	16	44
	Swahili (S)	N1/2-PL	3	7	43
	Northern Sotho (N)	N02	2	3	67
Possessive Pronoun	Luganda	PB2X	3	4	75
	Russian	PSMXP1- - - R - - - - - -	7	16	44
	Swahili	POSS1/2-PL	3	10	30
	Northern Sotho	PROPOSS02	2	9	22
Punctuation (?)	Luganda	TS?	3	3	100
	Russian	Z#- - - - - - - - - - - - - -	2	16	13
	Swahili	QUESTION-MARK	1	13	8
	Northern Sotho	$?	2	2	100

The average values for η_{at} for the tags for each Language - evaluated using Equation 3 - were computed and captured in Table 19 whose results are visualised in
Figure .

$$\eta_{at} = (\sum \eta_{ti}) / n \quad (3)$$

where η_{ti} is the storage efficiency of tag i, and n - the number of tags under consideration

7.2.2 Assessing the Impact of TLMS on Tag-sets using Storage Efficiencies

TLMS was applied to the tags used in Table and the corresponding storage efficiencies for the new tags were computed using Equations 2 and 3. Also evaluated, using Equation 4, was the improvement in percentage (P_I). The results generated by the experimentation were captured in Table 20.

$$P_I = 100 (\eta^*_{at} - \eta_{at}) / \eta_{at} \quad (4)$$

Where: η^*_{at} is the new storage efficiency and η_{at} - old storage efficiency.

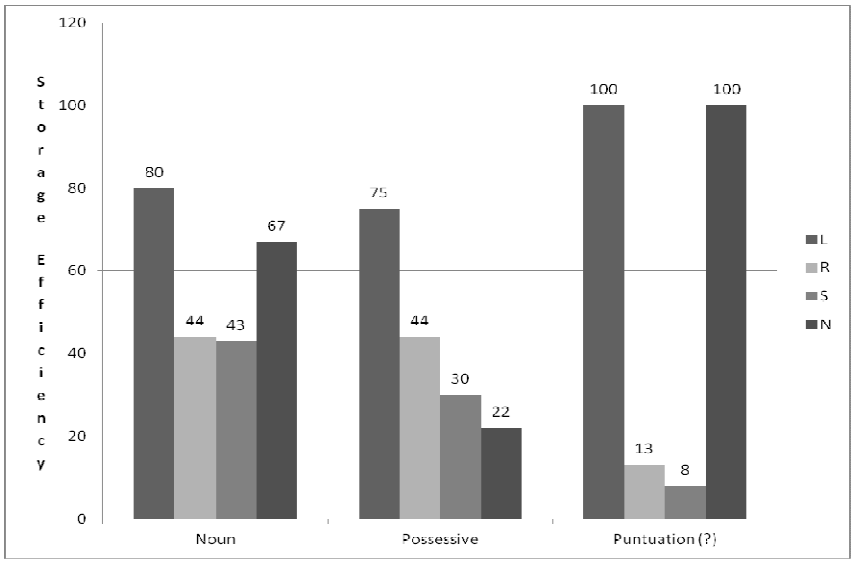

Figure 1 Comparison η_t of the tags in the Languages

Table 19 Comparison of the Average η_{at} of the Languages

Language	**Average η_{at}**	**Rank**
L	85	1
R	34	3
S	27	4
N	63	2

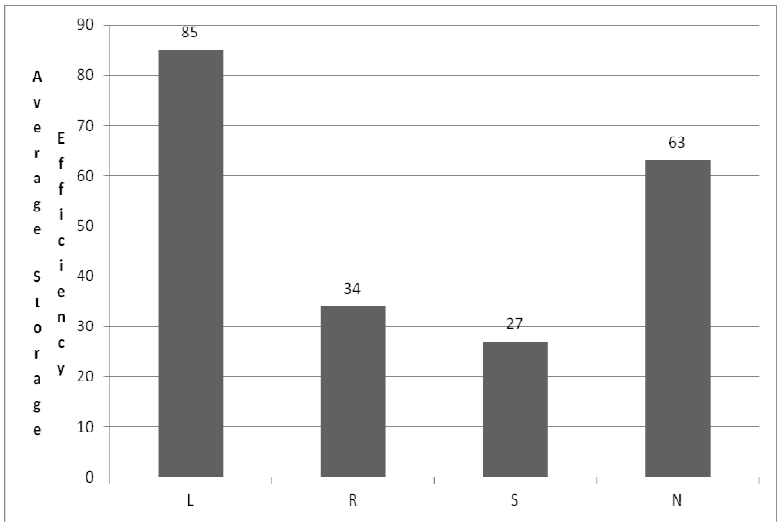

Figure 2.Comparison of the Average η_{at} of the Languages

7.3 Discussion of Results

In reference to Figure , it is evident that SCTL (Luganda) tags had the highest ηt in all the three POS and therefore the highest average ηat - as shown in Figure - due to the application of TLMS which maximizes ηt. Note that the maximum value of ηt is obtained when Ic and Nt are equal.

Table 20 comparison of Storage Efficiencies (η_t) for the three POS Tags for the four Languages

Language	**POS**	**Tag Sample**	$\mathbf{I_c}$	$\mathbf{N_t}$	$\boldsymbol{\eta_t}$	$\boldsymbol{\eta^*_{at}}$	$\boldsymbol{\eta_{at}}$	$\mathbf{P_i}$
Luganda(L)	Noun	NC2TX	4	5	80	85	85	0
	Possessive Pronoun (PP)	PB2X	3	4	75			
	Punctuation(?)	TS?	3	3	100			
Russian(R)	Noun	NNMIP4A	7	7	100	100	34	194
	PP	PSMXP1R	7	7	100			
	?	Z#	2	2	100			
Swahili (S)	Noun	N2	2	2	100	100	27	270
	PP	P2	2	2	100			
	?	Z?	2	2	100			
Northern Sotho (N)	Noun	N2	2	2	100	100	63	59
	PP	P2	2	2	100			
	?	$?	2	2	100			

The low value for η_{at} of less than 35% obtained for both the Russian and Swahili is due to the fact that: firstly, the Russian tag captures a lot of irrelevant information for a given POS by virtue of being a positional tag-set (this is an inherent problem of all positional tag-set). Secondly, the Swahili tag uses a number of characters (more than one) to capture one single form of information in bid to make the tag more mnemonic.

In reference to SCTL, the issue of mnemonic has not been neglected. It has been handled at single character level where the characters are chosen such that they portray the information they capture. The mnemonic issue cannot be over emphasized here because the tags are mainly for computer "consumption".

It is worthwhile noting that the application of TLMS on the other three languages has lead to a tremendously improvement to the storage efficiencies (η_t) of the tags as demonstrated in In reference to Figure , it is evident that SCTL (Luganda) tags had the highest ηt in all the three POS and therefore the highest average ηat - as shown in Figure - due to the application of TLMS which maximizes ηt. Note that the maximum value of ηt is obtained when Ic and Nt are equal.

Table . The improvement impacted to Swahili tags is, approximately, to three-folds, that to Russian tags to two-fold and that to Northern Sotho - 60%. Conclusively, the TLMS is very crucial in highly inflectional languages, which have a lot of inherent morpho-syntactic information to capture, to boost their tag storage efficiencies.

In this experimentation no improvement to the Luganda tags has been reported; however, improvement can be effected to the noun and pronoun tags by splitting the corresponding structures into two. For the Noun - the split is between the common nouns and proper nouns (or named entities) to form two structures; but this has been dedicated to future work when enough information on named entities has been gathers. While, for pronoun the split is between interclass

pronouns and ordinary pronouns but also this has been left to future works when the impact of many structures on performance of the system has been ascertained.

Notably, the applicability of Equation 3 can be extended to corpus level to determine the average value η_{at} for tags used in the entire corpus for a given tag-set. This will go a long way to access how a particular tag-set is firing in terms of efficient storage usage, which is important especially in embedded systems - mobile phones inclusive - where storage resources are limited.

8. ADVANTAGES OF SCTL

1. SCTL is associated with high storage efficiencies due to the application of TLMS

2. CN captured in SCTL can be used to computationally check conventional agreement between words.

3. SCTL eliminates the need for two levels or multilevel processing or tagging which would otherwise demand for more coding and more computational power, especially dealing with Luganda verb forms which have quite some encoded information.

4. The SCTL eliminates the need to develop algorithms for addressing disambiguation issues associated with both the use of apostrophes and full stops in text.

5. The SCTL structures are easily modifiable in that they can be expended to encode more information or contracted to remove unnecessary information. This capability can be exploited to either optimize this tag-set or adapt it to suit other Bantu languages or other agglutinative languages.

6. SCTL being a structured tag-set has many practical benefits including:

    i. Learnability: It is much easier to link traditional linguistic categories to the corresponding structured tag than to an unstructured atomic tag. While it takes some time to learn the positions and the associated values of the Luganda Tag-set, for most people, it is still far easier than learning the corresponding avalanche of tags as atomic symbols.

    ii. Systematic description: The morphological descriptions are more systematic. In each system, the attribute positions are (roughly) determined by either POS or SubPOS. Thus, for example, knowing that a token is a common noun (NN) automatically provides information that the CN, state, and case positions should have values.

    iii. Decomposability: The fact that the tag can be decomposed into individual components has been used in various applications.

    iv. Systematic evaluation: The evaluation of tagging results can be conducted in a more systematic way. Each category can be evaluated separately on each morphological feature. Not only is it easy to detect on which POS the tagger performs the best / worst, but it is also possible to determine which individual morphological features cause the most problems.

9. CONCLUSION AND FUTURE WORK

The advent of SCTL, an NLP (Natural Language Processing) resource, goes a long way to aid the tagging process for the development of an annotated corpus, another NLP resource, and hence open the avenue of developing other NLP applications, like grammar analysers, semantics analysers, to mention but a few. One advantage of languages rich in morphology is that they are easier to process at higher stages of NLP than their counterparts. This means that SCTL, which captures the rich morphology of Luganda, avails itself of this advantage.

Although the number of tags is enormous, it is manageable through the use of tag structures, like those in SCTL. The tag structures are analogous to table structures in a database system while the tags are analogous to the data that these tables store or handle.

Luganda is highly inflected language and therefore, require a large number of tag to capture it morphological properties which play a major role in the analysis of the grammar and semantics of Luganda. The advent of SCTL comes in handy to facilitate these endeavours as well as provides a means to handle this avalanche of tags.

Through the use of CN captured in the SCTL tags, SCTL can be used to check conventional agreement, a vital aspect of Luganda grammar, by simply checking whether CNs of the words in question are the same, and an algorithm is being developed to avail itself of this advantage.
Storage Efficiencies, namely, η_t (individual) and η_{at} (batch) which are novel metrics proposed in this research work, can be used in evaluating how a particular tag-set is performing in terms of efficient storage usage at tag level and corpus (or batch) level respectively. Storage efficiency is important especially in embedded systems - mobile phones inclusive - where storage resources are limited. Finding on the comparison of Storage Efficiencies (η_t and η_{at}) of various tag-sets show that SCTL tags had the highest η_t therefore the highest η_{at} among the tags from Swahili, Russian and Northern Sotho, due to the application of TLMS which maximizes η_t.

TLMS is very crucial in highly inflectional languages which have a lot of inherent morpho-syntactic information to capture, in bid to boost their tag storage efficiencies. Finding on the impact of TLMS on tag-sets using Storage Efficiencies (η_t and η_{at}) as evaluation metrics show that the application of TLMS on the experimental languages has produced tremendously improvement to the storage efficiencies of their tags, precisely, the improvement to Swahili tags is approximately to three-folds, that to Russian tags to two-fold, and that to Northern Sotho - 60%.

SCTL - being a structured tag-set, and therefore, easily modifiable - can be easily adapted for other Bantu languages and other agglutinative languages.

In a bid to improve the storage efficiency of SCTL, the issue of splitting the noun and pronoun structures into two has been dedicated to future work when the impact of many structures on performance of the system has been ascertained. Precisely, for the Noun - the split is between the common nouns and proper nouns (or named entities) to form two structures; while, for pronoun the split is between interclass pronouns and ordinary pronouns.

The issue of identification and validation of the tags of SCTL as well as investigation of the impact of TLMS on the corpora of various languages are dedicated to the continuity of this research work.

REFERENCES

[1] Slav Petrov, Dipanjan Das And Ryan Mcdonald, (2012) "A Universal Part-Of-Speech Tagset", Proceedings Of The Eight International Conference On Language Resources And Evaluation (LREC'12), ISBN-978-2-9517408-7-7.

[2] Hurskainen, A, (2004) "Tagset Of SWATWOL: A Two-Level Morphological Dictionary Of Kiswahili", Institute For Asian And African Studies, University Of Helsinki.

[3] E. Taljard, G. Faab, U. Heid And D.J. Prinsloo, (2008) "On The Development Of A Tag-Set For Northern Sotho With Special Reference To The Issue Of Standardisation", Literator 29(1) April 2008, Pg:111-137, ISSN: 0258-2279.

[4] Gertrud Faaß, Ulrich Heid, Elsab´E Taljard, And Danie Prinsloo, (2009) "Part-Of-Speech Tagging Of Northern Sotho: Disambiguating Polysemous Function Words" Proceedings Of The EACL 2009 Workshop On Language Technologies For African Languages – Aflat 2009, Pages 38–45, Athens, Greece.

[5] Erjavec, Tomaz, (2004) "Mtext-East Version 3: Multilingual Morpho-Syntactic Specifications, Lexicons And Corpora", Proceedings Of The Fourth International Conference On Language Resources And Evaluation, LREC'04, ELRA. Paris, France, Pp. 1535-1538.

[6] Erjavec, Tomaz, (2009) "MULTEXT-East Morphosyntactic Specifications: Towards Version 4", Proceedings Of The MONDILEX Third Open Workshop, Bratislava, Slovakia.

[7] Erjavec, Tomaz, (2010) "MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons And Corpora", Proceedings Of The LREC 2010 Third Open Workshop. Malta.

[8] Balagadde.Ssali Robert, (2016) " Designing A Tag-Set And Models For Spell Checking Intelligent System With A Bias In Luganda Language", Phd Thesis, Osmania University.

[9] Elworthy, D. (1995) "Tag-Set Design And Inflected Languages", 7th Conference Of The European Chapter Of The Association For Computational Linguistics (EACL), From Texts To Tags: Issues In Multilingual Language Analysis SIGDAT Workshop, Dublin, Pp. 1–10.

[10] Cloeren Jan, (1993) "Toward A Cross-Linguistic Tag-Set", Workshop On Very Large Corpora:Academic And Industrial Perspectives.

[11] Marcus, M. P.; Santorini, B.; And Marcinkiewicz, M., (1993) "Building A Large Annotated Corpus Of English: The Penn Treebank", Computational Linguistics, 19(2):313–330.

[12] Francis, W. Nelson; And Kucera Henry, (1979) "A Standard Corpus Of Present-Day Edited American English", Department Of Linguistics, Brown University.

[13] Francis, W. And H. Kucera, (1982) "Frequency Analysis Of English Usage: Lexicon And Grammar", Boston, Houghton Mifflin.

[14] Jirka, Hana And Feldman, Anna, (2010) "A Positional Tagset For Russian", Proceedings Of The 7th International Conference On Language Resources And Evaluation (LREC 2010). Valletta, Malta, European Language Resources Association, 1278–1284, ISBN: 2-9517408-6-7.

[15] Hajic Jan, (2004) "Disambiguation Of Rich Inflection: Computational Morphology Of Czech", Prague, Czech Republic: Karolinum, Charles University Press.

[16] Ide, Nancy And Jean Veronis, (1994) "Multext-East: Multilingual Text Tools And Corpora", Proceedings Of The 15th International Conference On Computational Linguistics (COLING). Vol. I. Kyoto, Japan, Pp. 588-592.

[17] Civit, Montserrat, (2000) "Gua Para La Anotacion Morfologica Del Corpus Clic-TALP (Version 3)", Tech. Rep. WP-00/06. Barcelona, Catalunya: X-Tract Working Paper. Centre, De Llenguatge Computacio (Clic).

[18] Codd, E. F. , (1990) "The Relational Model For Database Management", Addison Wesley Publishing Company, Version 2 Ed., ISBN 0-201-14192-2.

AUTHORS

Robert Ssali Balagadde graduated with Msc (Systems Engineering) from Moscow State University of Mining, Russia in 1995, having specialised in Automated Systems of Data processing and Management; since then, he has been actively involved in teaching various courses at various Ugandan Universities namely, Kyambogo University, Luwero University, and Bugema University. Administratively, Robert has worked as Head of Department of Computer Science as well as Dean Faculty of Social Sciences at Luwero University. Currently, he is a research scholar at Department of Computer Science and Engineering, Osmania University. Robert has published a number of international publications towards developing Natural Language Processing (NLP) resources. His areas of research include: NLP, Artificial Intelligence, Big Data Analytics, Internet of Thing, Data Science, Embedded Programming, and Object Oriented Software Engineering.

Dr. P. Premchand is a Professor at the Department of Computer Science and Engineering (CSE), Osmania University (OU), since 1999. He has successfully supervised over 25 PhD candidates and has a repertoire of more than 100 national and international publications to his name. He is presently serving as the Dean, Faculty of Informatics, OU, and has worked administratively in various capacities in the University including: Dean, Faculty of Engineering; Head of Department, CSE; Chairman, Board of Studies in the Departments of CSE and IT (Information Technology). He is an active member of AICTE-NDA and selection committees of the following Universities: OU, JNTU, ANU, KU, ISRO, NRSA, and ADRIN. His research interests are in the fields of: Image Processing, Software Engineering, and Object Oriented Software Engineering Analysis and Design.