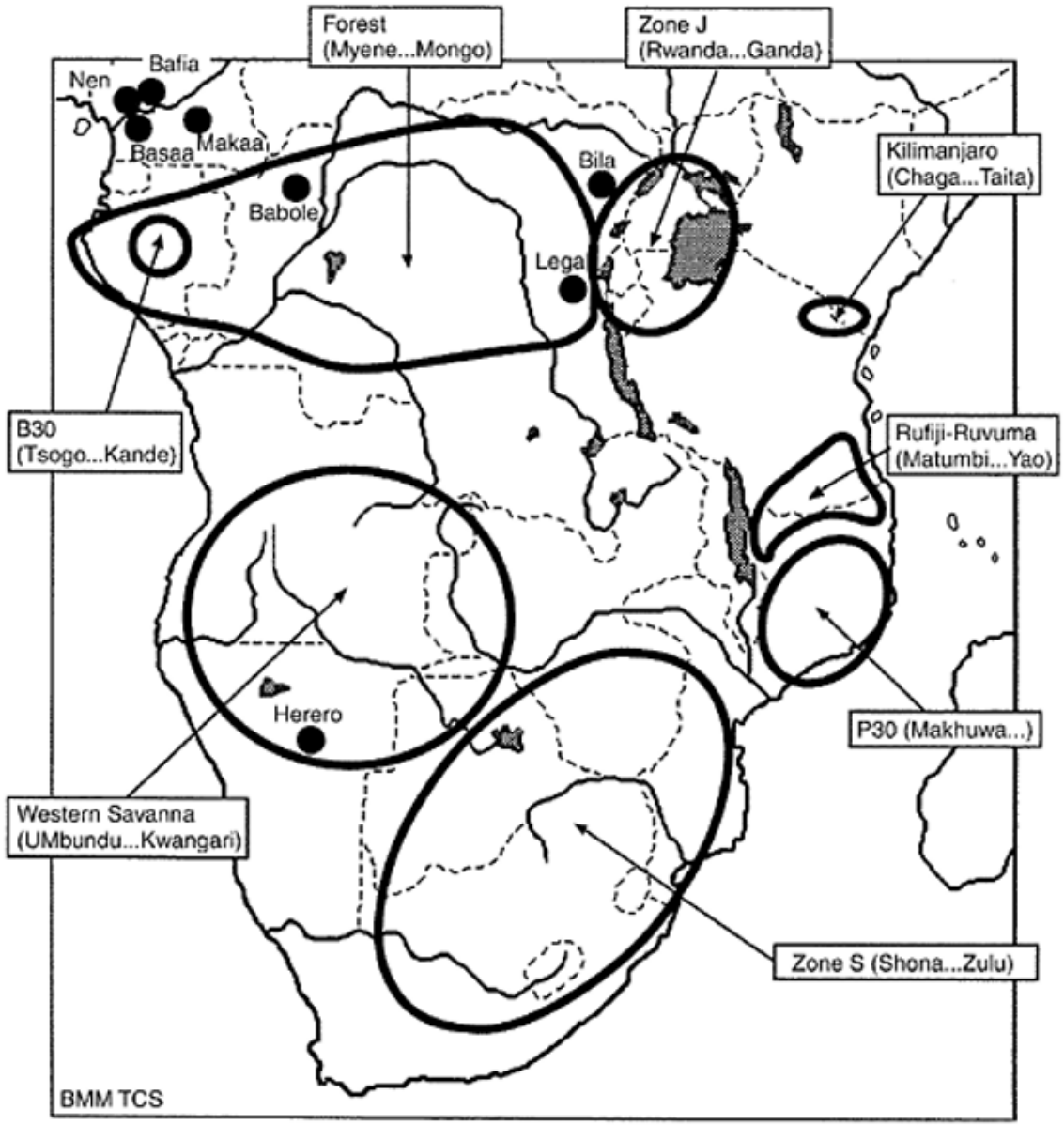
Computational Resources and Tools Developed for Bantu Languages

Joan Byamugisha

August 25, 2021

Overview

Spread of Bantu Languages



Semantic Categorization of Nouns

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

Noun Semantics

- Stem: *-ntu*
  - Person: ***omu****ntu*
  - Thing: ***eki****ntu*
  - Place: ***aha****ntu*
- Stem: *-nyankore*
  - Person from Ankore: ***omu****nyankore*
  - Language of Ankore: ***oru****nyankore*
- The noun prefix is based on a class prefix

Runyankore's Noun Class System

Noun Class (NC), Subject Concord (SC), Possessive Concord (PC), Adjective Concord (AC)

NC	**SC**	**PC**	**AC**
1. o-mu-	-a-	o-wa	o-mu-
2. a-ba-	-ba-	a-ba	a-ba-
3. o-mu-	-gu-	o-gwa	o-mu-
4. e-mi-	-gi-	e-ya	e-mi-
5. ei-/e-ri-	-ri-	e-rya	e-ri-
6. a-ma-	-ga-	a-ga	a-ma-

Generic Verb Morphology

- Comprises bound morphemes, a verb-root, and extensions
- Morphemes preceding verb-root specify person, noun class, aspect, time, negation
- Extensions specify valency-changing categories
- <Initial> <Subject> <Negative> <Tense and/or Aspect> <Object> <Root> <Extension> <Final>
- For tense and aspect:
  - Tense reference occurs before that of the aspect
  - Most languages have one, two, three, or four past tenses
  - Most languages have one, two, or three discrete future tenses

Runyankore's Agglutinative Morphology

- Runyankore: *Titukakimureeterahoganu*.
- Morphemes: Ti-tu-ka-ki-mu-reet-er-a-ho-ga-nu
- English: We have never ever brought it to him/her.

Runyankore Corpora

- Generated a one million sentence general-purpose domain independent corpus
- Created 18,816 different ways of varying the sentence structure
- Generated corpora labeled for morphology and sentiment
- Generated text in seven tenses:

Runyankore Treebank

- $S \rightarrow IG\ FM\ LM\ IF\ VR\ EX\ FN$
- $IG \rightarrow PN\ IT\ SN$
- $PN \rightarrow$ ti | ni
- $IT \rightarrow$ a | o | n | tu | mu | ...
- $SN \rightarrow$ ta
- $FM \rightarrow$ za | ka | riku | rikuza
- $LM \rightarrow$ ki
- $IF \rightarrow$ mu | ba | ...
- $VR \rightarrow\ verbRoot$
- $EX \rightarrow$ w | er | erer | ...
- $FN \rightarrow$ a | e | ire

Pre-trained Runyankore Word Embeddings and Classifiers

- Pre-trained word embeddings from distributional context from one million sentences
- Pre-trained model on morphology
- Pre-trained sentence-level sentiment classifier

Disambiguating Noun Semantics

- Deals with determining the noun class when the class prefix is the same
- Runyankore
  - ***Omu****ti*
  - ***Omu****ntu*
  - ***Omw****aka*
  - ***Omw****ana*
- Kinyarwanda
  - ***Umu****gore*
  - ***Umu****tima*
  - ***Umw****embe*
  - ***Umw****alimu*

From Distributional Context

Query Word	Results
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)
embwa (dog)	*embeba* (rat), *enkyende* (monkey), *empungu* (bird of prey), *enumi* (bull), *enyawaawa* (green ibis)

Semantic Categorization of Nouns

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

Noun Class Determination

Approach	Runyankore	Luganda	Kinyarwanda
Morphological only	69.23	57.53	43.94
Semantic only	66.67	47.95	40.91
Morphology, Syntax, Semantics	87.18	73.97	63.64

From Pre-trained Models and Treebank

- Word segmentation from pre-trained model on morphology
- Sentiment analysis from pre-trained sentiment classifier
- From Treebank
  - Performing machine translation
  - Automated evaluation of generated versus human authored text
  - Computing linguistic diversity