KenCorpus: Kenyan Languages Corpus

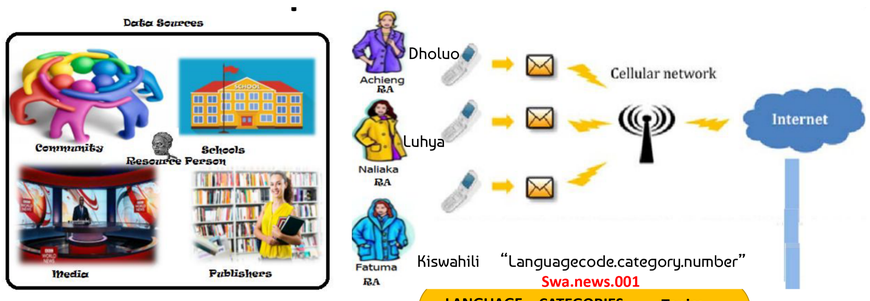
Lilian Wanzare, Edward Ombui, Lawrence Muchemi, Barack Wanjawa, Owen McOnyango and Florence Indede

Maseno University, University of Nairobi, Africa Nazarene University

Kenyan Languages Corpus for Natural Language Processing and Machine Learning

1. Project for the creation, labeling, augmentation and maintenance of datasets for machine learning.
2. Main Objective: To build Datasets that enable specific tasks in Natural Language Processing (NLP) and broader research in Machine Learning with the ultimate goal of supporting social impact.
3. Objectives include to develop:
   - labeled and unlabeled text corpora to support NLP and ML
   - parallel corpora for machine translation
   - corpora to support fundamental NLP tasks: i.e. Part of Speech tagging
   - corpora to support downstream NLP tasks: i.e. question answering
   - speech corpora to support Speech synthesis and Text to Speech conversion

Data Collection



METADATA
- Name of Speaker
- Demographics: gender,
- Date
- Place

LANGUAGE	CATEGORIES	Topics
• **Swahili**	• Letters	Agriculture
• **Dholuo**	• News	Education
• **Luhya**	• Folks	Culture
	• Essays	Commercial
	• Story	Political

Question Answering

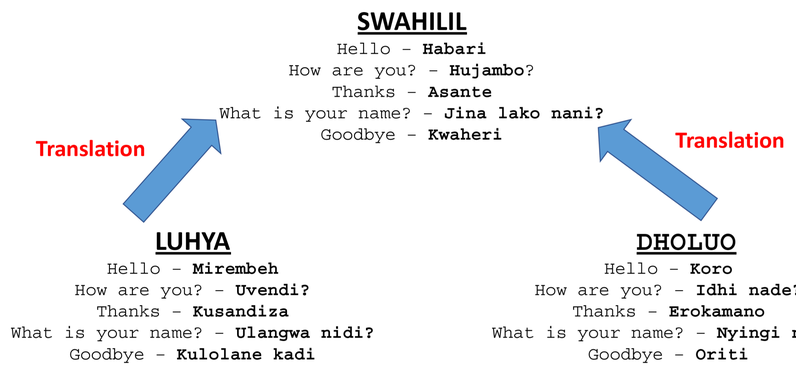
- **Have Questions about the Swahili texts and Answers to those questions**

Asubuhi hiyo ilitupata pambajioni pa hospitali ya Uguzwa. Kando yangu ndugu yangu Pendo alikuwa akijipindapinda kwa uchungu mwingi. Aidha, alikuwa akitetemeka sana. Nilimshikilia asije akaanguka chini. Nilijizuia nisimtazame maadamu kila nilipomtaza machozi yalinidondoka. Nakwambia Pendo alikuwa amedhoofika sana. Maradhi hayo ya kifo yalikuwa yamemdhilisha akabaki gofu la mtu. Alikuwa kakonda ghaya ya kukonda akabaki fremu. Nilitokwa na milizamu ya machozi kwa huruma. Nilihisi kama kwamba mimi ndimi nilikuwa nikiugua. Ama kwa hakika damu ni nzito kuliko maji.
Siku hiyo mimi na ndugu yangu tuliratibiwa kufanyiwa upasuaji. Maadami figo zake zote zilikuwa zimeharibika, ningempa figo yangu moja kulingana na ripoti ya madaktari. Hiyo ndiyo ingekuwa njia ya pekee ya kusaidia la sivyo angetuacha. Mimi na yeye tulikuwa ndugu wa toka nitoke. Tulipendana mithili ya pete na kidole. Mintaarafu na hayo, nilikuwa nimejitolea mhanga kumpa ndugu yangu figo yangu.
…
kencorpus.co.ke

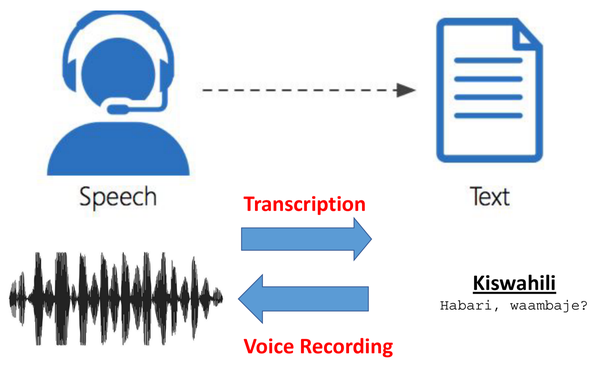
Questions? Answers?

1. Asubuhi ilitupata pambajioi pa hospitali gani - Answer: Uguzwa
2. Mimi na ndugu yangu tuliratibiwa kufanyiwa nini - Answer: Upasuaj
3. …

Translation

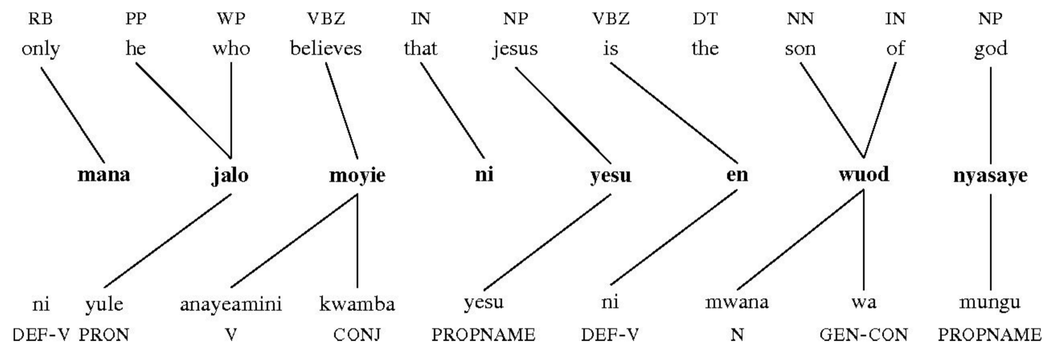


Transcription



Part of Speech Tagging

For Dholuo and Luhya



KenCorpus Datasets

Lang.	Texts	Audio Hrs
Swahili	2585	19
Dholuo	546	99
Luhya	977	57

Table 1. Text and speech data collected.

Data	No.	No.
MT	12,400(sents)	-
POS	143,000(words)	-
QA	1445(texts)	7526(QA)
Trans.	27(Hrs)	30,000(words)

Table 2. Datasets.

Link to Datasets

https://kencorpus.co.ke/

https://dataverse.harvard.edu/dataverse/Kencorpus

Acknolegement

- LACUNA Funds
- Language research assistants and resource persons
- Maseno University
- University of Nairobi
- Africa Nazarene University