

# Research questions for thesis

- How do Bantu languages, and Swahili in particular, integrate into the Universal Dependencies framework?
  - Being as there are no existing Bantu treebanks at all, what are the relevant syntactic phenomenon to consider when making a treebank?
    - Discuss relevant morphological phenomenon in Swahili (for the morphological features in UD)
    - Discuss some relevant syntactic features (for dependency relations)
    - Provide some suggestions for how other Bantu languages could be handled in UD as part of an effort to encourage future treebanks for Bantu languages.
    - Discuss additional Bantu syntactic phenomenon using resources on Bantu syntactic typology.
- How do these phenomena fit into the universal dependencies framework?
  - How is the specific corpus developed for this work created?
    - Automatic POS tagging, morphological tagging, lemmatization and functional tagging using the Helsinki corpus
    - Annotation procedure (e.g. sampling and sentence selection procedure)
    - Automatic labelling procedure
- What considerations need to be taken to allow for accurate parsing using a relatively small dataset of mixed origin?
  - What strategies can be used to enable parsing on a small dataset?
  - Given that the treebank produced is partially created by rules and partially hand annotated, what is the best way to integrate the hand annotated trees with the trees created using rules?
    - How should the hand annotated data be integrated into train, eval, and test data splits?

---

Revision #2

Created Thu, Feb 6, 2020 4:34 PM by [kenneth](#)

Updated Thu, Feb 6, 2020 8:37 PM by [kenneth](#)