Title: Bantu PARSEME Corpora: Annotation for Swahili and Bapuku MWEs

Name: Matt Malone

Age: 28

Affiliation: City University of New York (CUNY) Graduate Center

Keywords: swahili, bantu, multiword expressions, natural language processing, low-resource languages

My goal for the training school would be to create at least one addition to a PARSEME Corpus. Given that version 1.3 of the corpus contains no Bantu languages, and only one language native to Africa (Arabic), the addition of a MWE-tagged Swahili corpus would significantly diversify the dataset. Depending on the corpus size requirements and time capabilities, I would also like to add a MWE-tagged set for Bapuku (bnm), an endangered language native to Cameroon and Equatorial Guinea.

My research in Vulić et al. (2020) demonstrated how important MWEs are for building Swahili NLP. My contribution to the project was Swahili SimLex-999, a gold standard resource used to evaluate the quality of Swahili word embeddings. The resource itself was simply a list of word pairs, alongside 10 human-annotated judgments of the similarity of the words on a scale 0-10. In translating the original English list of word pairs to Swahili, 12% of the words were translated as MWEs, despite the request to use single words where available. This led me to research the nature of Swahili MWEs, which are extremely common due to compound words (e.g. “airport” in Swahili is *uwanja wa ndege*, “field of planes”), innovated words (e.g. “password”, *neno la siri*, “word of secrecy”), gendered terms (e.g. “groom” is *bwana arusi*, “wedding man”), and other common lexical entries. Additionally, Swahili NPs with adjectives are very often translated with prepositional constructions (e.g. “sad” is *mwenye huzuni*, “owning sadness”). For these reasons, MWEs are ubiquitous in the language, and in order to conduct

proper linguistic analysis using computational methods, these MWEs must be properly identified and tagged.

If selected to attend the training school, I would learn how to properly identify and annotate MWEs according to the PARSEME standards. While I have a significant background in Bantu languages and computational linguistics generally, MWEs are a new field for me, and one I would really enjoy learning more about. This task will be very useful given how often MWEs pose challenges to resource building in NLP and how common MWEs are in Swahili.

Because Bantu languages are structurally very similar, it may be useful during the brainstorming hackathon to discuss the possibility of translating a PARSEME corpus between closely related languages, or variants on a language. Prior to the summer, I will be digitizing resources for Chimwiini, a dialect of Swahili spoken in Somalia. I would be very interested in seeing how we can take advantage of the similarity of these variations to produce multiple PARSEME corpora.

Currently the project is in the planning phase. I use the annotated Swahili Helsinki Corpus to conduct analyses on the language, and if possible, would use this as the basis for the corpus, adding a layer of MWE-identification on top of the current annotation layers; however, I would also be willing to construct a new corpus if necessary. If timing permits, I would create a Bapuku corpus from scratch as well.

References

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, Anna Korhonen; Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. Computational Linguistics 2020; 46 (4): 847–897. doi: https://doi.org/10.1162/coli_a_00391