WordNet. A Lexical Database for Swahili Language

Anonymous ACL submission

Abstract

This research aims to create a Swahili language based "WordNet", a lexical database that can group Swahili words into synonym sets (synsets) and provides short definitions and examples. The database should also hold information on the results of related words. The proposed Swahili WordNet will use a modern lexical database using Hybrid approach, a combination of merge and expand approach. The data will be extracted from Wikipedia, Helsinki, 5000 Common Base Synsets (CBCs) and 2446 list of hypernym and hyponym contributed by Okal. will be integrated with a web based/mobile enabled interface that will allow creation and update of synsets and access of information to various users including teachers, students and researchers plus all normal users.

***Keywords*—** Swahili, Wordnet, Lexical database, Natural language processing

1 Introduction

The dictionary has played an integral role in the learning and development of languages around the world. However, with the rapid advancement of technology, it is becoming more time-consuming to rely on traditional dictionaries, which often assumes pre-existing knowledge of the reader (Miller et al., 1990). This limitation led to creation of the Princeton Wordnet (Princeton, 1980) a lexical resource organized around meaning and concepts, enabling users to discover relationships effectively.

Despite global multilingual dictionaries such as BabelNet (Navigli and Ponzetto, 2012) supporting over 500 languages including Swahili, few languages have their own Wordnet, Swahili in particular lacks its own Knowledge base relying online Wikipedia data. This shortage of linguistic resources is further compounded by its classification as low resource-languages (Shikali and Mokhosi, 2020) and assumption of Wordnet availability (Paikens et al., 2022) . This limitation has led to the growth of Cross-lingual Transfer learning (Ruder et al., 2019; Upadhyay et al., 2016) and Multilingual Approaches (Wang et al., 2020), although useful but it compromise language independence and cultural values.

In recent years, Swahili has gained a global interest, becoming the most spoken language in Africa with over 200 million speakers globally. The development of language technology, such as Salama (Hurskainen, 1999), a computational environment for the development of language applications has also contributed to the growth of Swahili. Despite the advancement, it is expected Swahili to spearhead the growth of African Knowledge base, but this effort is currently pushed by South African languages (Madonsela et al., 2016).

We propose Swahili wordnet as a knowledge base. Though there are two approaches to creating a wordnet (Vossen, 1998), most wordnet are created from the original English Wordnet, hence, their structure is based on the English Wordnet. The latter, wordnet is built from grounds up. This has the advantage of maintaining language structure, but it is usually very expensive as compared to the first approach. The choice of the approach to use is influenced by our goals.

One of the goals is to create Swahili wordnet that can be linked to other wordnet through a common base word (Anderson et al., 2010) because, no matter the language disparity there are common words found in almost all the languages in the world and at the same time maintain Swahili language structure. A novel approach will be to combine the two methods to form a Hybrid method. Our second goal is to improve on the existing wordnet tools and make them readily available to others who would want to build wordnet for other African and local languages and also provide a structure that can be used as a baseline in constructing other wordnet more so the Bantu languages that share a lot of similarity with Swahili Language.

Swahili language keeps changing based on various factors such urbanization, technological advancement, social-cultural values and political change and a good example is borrowing. (Ochieng, 2015) reveals, borrowing is not mutual when two languages with different strength are involved. According to Dominance theory (Kachru, 1994) borrowing is often from the high resource language to the low resource language. Ideally Swahili language change will gear towards getting words from English even in cases where a Swahili word equivalent is available. In cases where it is not available, limit the creation of Swahili word to represent what is being addressed. Therefore, keeping tabs on language change without a knowledge base is very difficult, even the current Kamusi is heavily influenced by high resource language. To be able to trace language change and ensure that standard Swahili is maintained and inclusion of other Swahili dialects a knowledge base is of great importance.

Wordnet has various usage from knowledge representation, Machine Translations (MT), lexical relation identification, cross-lingual language reference, dictionary compilation, Ontology development and language learning and teaching (Gopestake et al., 1994; Niles et al., 2004). In language learning it is a good digital resource for learning meaning-related information about words in one or several languages (Dash and Bhattacharyya, 2023) as words are linked using ILI across wordnet. This makes it easier for learners and teachers to quickly learn different languages especially in Kenya where schools offer French, German, Spanish as languages. Swahili wordnet will support 4 out of 7 core competencies in the Competency-Based Curriculum (CBC); ***1***. *Communication and Collaboration*, ***2***. *Citizenship*, ***3***. *Creativity and Imagi-*

nation and ***4****. Digital literacy*. This aligns with government vision 2030 and SDG(education) of providing equitable quality education and opportunity for all.

2 Background

Swahili also known by its local name Kiswahili is a Bantu language spoken in East Africa including Kenya, Tanzania, Uganda and most recently the areas of Congo and Somalia. A first language of most people in the areas of Tanzania and some parts of Kenya especially people born in the cities; a second language to the latter group of people born in rural areas inspired by the need for communication between various communities and ethnicity. Around 20% of Swahili words are loan words mainly from Arabic accumulating 15% followed by English at around 4.6%, followed by Portuguese, Malay and the core Swahili language constituting words mainly from Bantu languages such as the Pokomo and the Mijikenda languages.

Around 15 dialects exist in Swahili, the most important ones are spoken by quite a number of people are KiUnguja which is the modern standard Swahili spoken in Tanzania, Kimvita spoken in Mombasa and other areas of Kenya and KiAmu spoken on the Islands of Lamu (Britannica, 2023) and mostly it is the dialect that was traditionally used in literature and Poetry (Andrey, 2004).

Just like the English language, Swahili has 5 vowels, *a, e, i, o, u* and 23 alphabetical letters which differs from English 26. The letter *q* and *x* are not present in Swahili and *c* is never used alone but in conjunction with the letter *h*, forming a digraph (two letters combined together to form a single sound or phoneme) *ch*. Other digraphs include *dh, gh, kh, ng', ny, sh, th*, and *ng* (Steere, 1882). Out of the 9 digraphs *ng'* is a bit problematic and rarely occurs.

Swahili nouns are separable into classes mostly referred to as a noun class (Marten, 2013), basically it is the categorization of nouns into groups based on their prefix mark. Most prefixes have their singular and plural form, for instance noun class *ki-/vi-*, *ki-/* is a prefix that denotes singular form while *vi-/* denotes plural form. Some other noun classes such as *ya-/ya-* maintain their singularity in both plural form and singular form. Out of the noun class we can further categorize them semantically for instance, the noun class *ki-/vi-* can be categorized into animate object and inanimate object as seen in Table 1

The prefix helps to bring other forms of words such as verbs, adjective into an agreement with a sentence, for instance, *wa-tu wa-li-enda* meaning 'people went'. As seen in the example sentence, the prefix *wa-* helps to create verb-subject agreement.

According to some scholars (Mgullu, 1999; Mohamed, 2001) Swahili is a highly agglutinative language however (Choge, 2017) suggest that this classification is a bit narrow because Swahili has other morphological characteristics that allows it to be further classified in terms of its morphological structure. She adds four other morphological characteristics such as polysynthetic, oligosynthetic, fusional and isolating which are grouped into either synthetic (use affixation in word formation) and non-synthetic category. In addition, Swahili is higly polysemous language which can be attributed to the semantic change of lexemes through social and physical experiences (Gichuru, 2020) which lead to high word level ambiguity hence creating a challenge for MT systems. To show this high level of word ambiguities lets take the word *kaa* which is both a noun and a verb. For a verb it means to dwell, sit and stay while as a noun it means a charcoal or crab. To try and solve this word level ambiguities knowledge of various words under different context they occur is desirable.

3 Related Work

Swahili tools development for text analysis started in 1985 by (Hurskainen, 1999), who built a Swahili language Manager (SALAMA) by combining two major components SWATWOL (a morphological analyzer of Swahili language and SWACGP (grammar parser). The result was a syntactic tool such as part of a morphological analyzer, spelling checker, semantic disambiguator and rule-based system for word level ambiguities, but had shortcomings at the semantic level.

It was around this time when the first Wordnet (Miller, 1985) was developed for the English language, the smallest unit being word/sense pair (Soergel, 1998) identified by a number which are grouped into synsets based on their synonyms. Each synset contains a definition that represents a concept, usage example, a part-of-speech tag comprising of either nouns, verbs, adjectives, and adverbs. All synsets are linked through a semantic network such as hypernyms and hyponyms. The main aim was to create a lexical database that could be used by computers as dictionaries and thesaurus were difficult to be used for NLP related task and thus led to Wordnet.

In a similar study, Eurowordnet was created (Vossen, 1998) in an aim of creating wordnet for other European languages structured as the Princeton Wordnet (PWN). They used two approaches in the development of the multilingual database; 1. expand approach which was used for the Spanish language which let to wordnet that was close to the PWN, 2. the merge model which was used for most of the languages in the multilingual database which resulted in wordnet that was independent of PWN. To effectively create a multilingual database, the equivalence relations between synsets of different language and PWN had to be linked through Inter-Lingual Index (ILI) based on PWN. It led to the development of Domain Ontology and Top concept ontology, which comprised 63 semantic relations distributed across 1024 ILI records with an aim of providing a common framework for the most important concept in various wordnet.

To further improve on Eurowordnet, project Balkanet (Tufis et al., 2004) was created. The aim was to have a better coverage on cross-lingual and better quality than Eurowordnet by developing wordnet for Balkan languages such as Greek, Romanian, Serbian, Turkish and extending Czech wordnet developed by Eurowordnet. They followed the same method proposed by Eurowordnet (EWN), the expand and merge approach. The results were an improvement of the Eurowordnet and 4689 Common Base Concepts (CBCs) extended from the base concepts in EWN.

With the availability of wordnet, (Wanjiku, 2005) build a word sense disambiguation solution using Machine learning and PWN wordnet to solve semantic ambiguity in Swahili. The adopted method used was a corpus-based approach achieved using self-organizing map algorithm; which was used to get semantic categorization of nouns from data using English wordnet and its characteristic modeled as a Bayesian belief network resulting in a semantic tagger.

South Africa later become the first African country to develop its own wordnet (Kotzé, 2006). The first ever African Multilingual Wordnet Codename African Wordnet (Afwn) for its local language, including isiZulu isiXhosa, Setswana, Sesotho sa Leboa, Tshivenda, Sesotho, isiNdebele, Siswati & Xitsonga. They used the expand approach using PWN as the source wordnet because Afrikaans and English language are West Germanic language. The tool used for development was DEBVisDIC, a client server wordnet builder which is freely available. Though the step taken by South Africa towards a Multilingual African Wordnet was a great one, no African countries since have tried developing or interlinking their wordnet to the African Wordnet.

The Global Wordnet Assosiation (GWA), which was built from EWN and PWN, was formed as a non-profit organiza-

Noun Class	Semantic representation	Example
ki-/vi	Animacy	Kitoto
	Inanimacy	Kiti

Table 1: Semantic categorization of noun class

tion because of various groups around the world wanting to build wordnet for their own language. The main aim was to maintain, standardize and interlink wordnet for all languages in the world through the ILI as a universal index of meaning (Background, 2023). Currently, there are over 100 registered Wordnet in the GWA. To further improve on GWA, The Open Multilingual Wordnet (OMW) project was formed with a goal of easing the use of wordnet in multiple languages (Bond and Foster, 2013).

Despite advancements in language technology, a standardized knowledge base for the swahili language remains a critical gap. Efforts to develop Swahili technology have been scattered globally, with a growing emphasis on Artificial Intelligence (AI) and Machine Learning (ML) due to the increased online sharing of information and ideas. However, there has been limited focus on creating Swahili lexicon resources.The development of a Swahili knowledge base is essential not only to enhance NLP capabilities but also to establish a technological foundation for the language,making it more competitive among the world's most widely spoken languages.

4 Methodology

There are two standard methods of building Wordnet (Vossen, 1998), **1. Expand Approach**, a word is extracted from PWN, its equivalent translation in Swahili extracted, and checked if the meaning matches that of the word, a word can have more than one meaning. **2. Merge Approach**, built from a list of Swahili words, compiled either from online resources or a Swahili corpus, their meaning, part of speech tag, usage examples and equivalent relations extracted and then aligned with PWN using equivalence relation such as the ILI. Though not used much, it has the advantage of maintaining the morphological richness of a language and is usually more accurate as compared to the Expand approach.

4.1 Data

The list will be compiled from 5000 CBCs from the proposed EuroWordNet and BalkaNet as of wordnet 2.0 (Concepts, 2023). The second list from the 2446 words (Okal, 2018). The final list will be generated from Helsinki Corpus of Swahili and Wikipedia Swahili data as seen in Table 2.

	Words	Source
First list	5000 CBCs	Globalwordnet
Second list	2446 words (Okal, 2018)	Maseno
Third list	Helsinki Corpus of Swahili + Wikipedia data	Helsinki, sw wikipedia

Table 2: Distribution of Data to be used

4.2 The Hybrid Method

We propose a Hybrid approach (Anderson et al., 2010), a combination of Expand and Merge approach, leveraging the advantages of both methods. The idea is to start with the proposed 5000 CBCs using the expand approach, under each synset, find the equivalent translation from English to Swahili using MT systems such as google translate, chatGPT; *if a synset has one word translation and only one meaning, then we are certain it is the equivalent sense of Swahili, on the other hand if there is more than one meaning, we pick the meaning that matches with our synset and the Swahili word is added to our third list if it's not available. If the synset has multiple Swahili translations, we pick the word that closely matches with the synset in terms of meaning and we check if the rest of translated word/words is in the second list or the third list, if its available in either we ignore the word/words. If it's not available we append the word/words to the third list which will undergo Merge Approach.*

During Translation, when a word has multiple meanings, we will perform a manual lookup in Princeton's synsets to ensure accurate meaning. This process is crucial, but challenging due to difficulties in determining how detailed the meaning should be to match the word sense in PWN (Paikens et al., 2022). To overcome this, we will examine hypernyms, hyponyms, and surrounding words, as well as meanings in Kamusi, to gain insights into word meaning. Words that fail to match will be added to the third list. In Merge Approach, we process the second and third lists. The second list is straightforward, with existing information on part of speech, meaning, and hypernym-hyponym relations. We will add additional relations later. The third list, including words from the Expand approach, will be compiled by a team of lexicographers, researchers, and linguists, with part of speech, meaning, and semantic relations.

In Swahili, a language with complexity in its synonym relationships, limitations arise. The Kitabu cha Visawe and Kamusi lists of synonyms can be misleading due to hypernyms and hyponyms. For example, the word *kundi* (group) has synonyms in Kitabu cha Visawe like *kusanyiko*, *kikosi*, and *umati*. However, Kamusi defines *kundi* as a collective term for objects or people, and *halaiki* as a large number of people. This subtle difference may be hard for non-native speakers to understand. Computers also struggle to grasp language nuances. To avoid confusion, caution is needed when searching for synonyms, and synonyms that are actually hyponyms or hypernyms will be listed separately.

4.3 Semantic relation

Aside from the synonyms, other relations to the synsets will be identified as proposed by (Format, 2023), the study followed the most prominent relations such as:

- **Hypernym and Hyponym** - hypernym is a word whose meaning includes a group of other words. For example, the word mkusanyiko includes words such as kundi, kikosi and with kundi, in its definition, the word mkusanyiko has been used to describe it. Hypernym relations between PWN and Swahili language are mostly the same at the top ontologies of words and even Interlinking them together is straightforward, but as you go down the ontology, the difference is noticeable. A

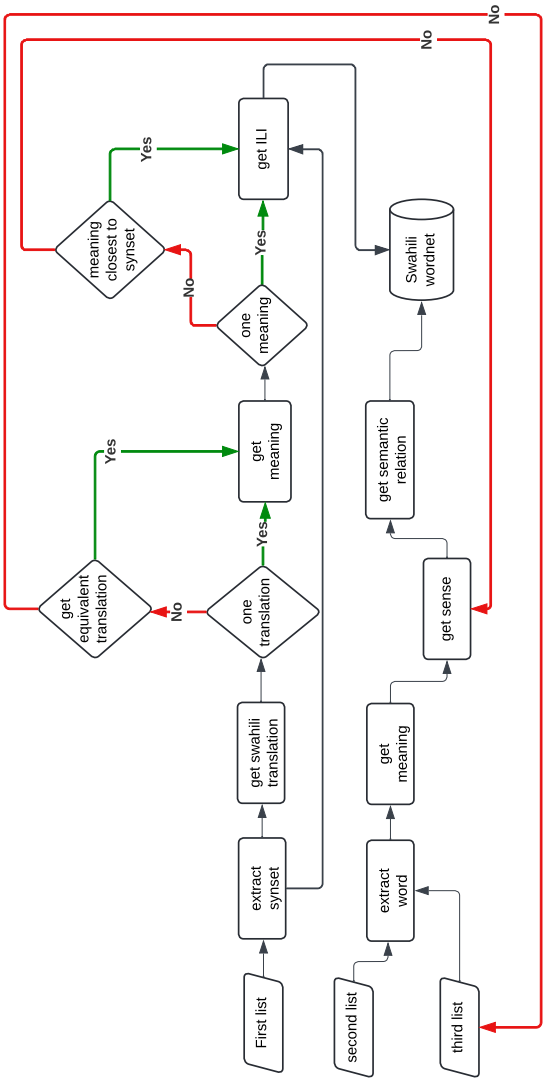

Figure 1: Representation of the proposed Hybrid method

hyponym can be viewed as an inverse relation to hypernym. It is a word whose meaning is included in the meaning of another word. For instance, the word cow is a hyponym for an animal.

- **Be in state and state of** - Link nouns that refer to anything in a particular state expressed by an adjective, for instance the noun rich express the state of being rich as discussed in the GWA format. rich (N) be in state rich (A) rich (A) state of rich (N)

- **Noun class and Plural**- Noun class are found in Swahili language but not in English and so this information is missing from PWN. Nouns (nomino) in Swahili are grouped into noun classes (ngeli) based on the prefix they have. For example, watu people, which is characterized by wa- as a prefix, belong to the group a-/wa-. The wa denotes the plural form of the word mtu, which is a singular (umoja) for the word watu. There are 9 noun classes each having its plural form but some noun classes maintain the word in both singular (umoja) and plural form i.e., the word nguo which belongs to the noun class i-/zi.

- **Antonyms** - a word opposite in meaning to another, for example, good and bad. There are different ways of getting antonyms in Swahili (Ojiambo, 2023).

  I. **Direct antonyms / Utanzu wa pindu**, For instance, hot (moto) and cold (baridi),

  II. **Reverse meaning of a word / Utanzu wa kufanya au kulinyambua neno**, applicable to only verbs. Swahili uses two vowels -U and -O to derive an opposite meaning from verbs that can be reversed in meaning. When the root vowel is a, e, i, u, the reverse marker is -u, when the root vowel is o, the marker is -o, as shown in table 3.

root vowel	marker	words	opposite
a, e, i, u	-u	tega [trap]	tegea [release]
o	-o	choma [poke in]	chomoa [pluck out]

Table 3: Swahili vowel marker and examples

  III. **Gender** - masculine/feminine / Utanzu wa jinsia This is the opposite of animate objects such as human, animals mostly based on gender, for instance, msichana (girl) and mvulana (boy).

- **Meronymy and holonym** - meronym are a form that denote part of something. For instance, a pedal is a meronym of a bicycle while a holonym is the opposite of meronym and it refers to a word that denotes a whole. For instance, in the above example, the word bicycle is a whole of difference parts of things.

A Hybrid approach has one key benefit; It allows us to concentrate more on words that have more than one meaning and at the same time words with one meaning are passed into MT systems, saving time, cost, and translations are more accurate. This approach also preserve Swahili morphological richness. However, it has a drawback; words with single meanings may be mistranslated, particularly those borrowed from English, for instance, the word entity, which is at the top ontology of English wordnet when translated to Swahili means kitu kamili, God can be referred to as an entity to a certain degree in English but in Swahili referring to God as kitu kamili is not societal acceptable. Furthermore, MT systems have higher Precedence to high resource languages compared to low-resource languages, to address such issues all the words will be revised.

5 Design of Swahili Wordnet System

Swahili Wordnet will be accompanied with a wordnet management system drawing inspiration from the Mongolian Wordnet management platform (Hasi and Tang, 2013). The user base of the system will consist of three categories: normal users, contributors, and administrators. Normal users are individuals who wish to interact with the system without requiring advanced technical linguistic knowledge. They will have the ability to search for queries or words within the database and choose how they wish to view the results. Contributors are

a set of knowledgeable users on either linguistics, lexicographers, programmers or NLP practitioners who want to make a contribution to Swahili Wordnet. The administrators will act as an overseer and guide the contributors while doing the revision and updates of Synsets

The system will be designed as a client-server architecture, composed of a web-based system and a python library hosted at GitHub mainly for researchers and NLP practitioners and possibly serve as a collection for other Swahili related tools and resources, such as Stopwords and slang. The main page will comprise of two parts. The first part is the introduction of wordnet, its construction, semantic relations and why wordnet is important. This information will be designed for the public, who want to understand more about words and relations without the underlying knowledge required of lexicography or programming. The second part deals with user management, i.e., logins and registrations, wordnet editor and browser, entailing documentation for contributors and downloads of various resources such as python library and Wordnet in different format such as xml, JSON and DB.

Since the first Wordnet, various tools have been developed to aid in creating wordnet globally such as DebVisdic (Rambousek, 2006), Catalan wordnet (Benitez et al., 1998). While helpful, these tools have a major limitation: they're built around a specific language structure. Swahili, for instance, has its own unique structure. An example is the word *ng'ombe* (meaning cow in english) shows this difference, the apostrophe does not show contractions as opposed to English.

Most systems do not support words in such format, and the solution is to remove the apostrophe and maintain the spacing, for the word *ng'ombe* the word to be registered will be *ng ombe*. This is doable if we are working with wordnet on programming environment; when extracting words using python or any other programming language, a short script can be written that translates the user query into a format the system is able to extract the word correctly. This becomes problematic on non-techies as they would query the word *ng'ombe* as they understand it.

To be able to include Swahili independent structures, we will build on already existing python software codename Hydra (Rizov, 2014) which is a wordnet builder used to build wordnet for the Bulgarian Wordnet and at the same time adhere to Global wordnet (Rambousek and s Horák, 2015) Standards.

Just like other wordnet, designing a system with maintenance in mind is of great importance as it allows the improvement of synsets through revision and addition of other words into the system. Our mission is to create a standard baseline to be used for possible creation of other lexicon resources such as thesaurus and future local language wordnet. Ensuring a standard and up to date synsets will be desirable.

6 Conclusion

Though we are still behind in terms of language resources, the efforts currently being made around the world to improve resource for low resources languages show a good future not only for Swahili Language but also other African languages. Swahili wordnet will not only help the field of NLP but also the development of other lexical tools, such as thesaurus, and also help in learning. When incorporated in learning institution it will provide a platform for learners around East Africa an equal opportunity to learn different languages which otherwise, they will not have access to especially public primary and secondary institutions.

References

Winston Anderson, Laurette Pretorius, and Albert Kotzé. 2010. Base concepts in the african languages compared to upper ontologies and the wordnet top ontology.

Zhukov Andrey. 2004. Old swahili-arabic script and the development of swahili literary language on jstor. *BRILL*, pages 1–15.

GWA Background. 2023. Background document. Accessed: 2023-10-22.

Laura Benitez, Sergi Cervell, Gerard Escudero, Monica Lopez, German Rigau, and Mariona Taule. 1998. Methods and tools for building the catalan wordnet. ArXiv:cmp-lg/9806009 type: article.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet.

Britannica. 2023. Swahili language.

Susan C Choge. 2017. A morphological classification of kiswahili. *Kiswahili*, 80. ; accessed 2023-10-27].

GWA Base Concepts. 2023. Gwa base concepts – global wordnet association. Accessed: 2023-10-22.

Niladri Sekhar Dash and Amrita Bhattacharyya. 2023. Utilizing indowordnet as a digital lexical resource for language learning/teaching purposes.

GWA Format. 2023. Global wordnet formats. Accessed: 2023-10-22.

Tirus Mutwiri Gichuru. 2020. Mabadiliko ya maana za leksia za kiswahili: Mtazamo linganishi wa kikale na kisasa.

Ann Gopestake, Ted Briscoe, Piek Vossen, Alicia Ageno, Irene Castellon, Francesc Ribas, German Rigau, Horacio Rodriguez, and Anna Samiotou. 1994. Acquisition of lexical translation relations from mrds. *Machine Translation*, 9:183–219.

Hasi and En Bo Tang. 2013. Design and implementation of mongolian wordnet management platform. *Advanced Materials Research*, 756:1344–1348.

Arvi Hurskainen. 1999. Salama swahili language manager. *Nordic Journal of African Studies*, 8:139.

Braj B Kachru. 1994. Englishization and contact linguistics. *World Englishes*, 13:135–154. ; accessed 2023-10-25].

Gideon Kotzé. 2006. Building a wordnet for afrikaans: Preliminary research in the form of an inquiry.

Stanley Madonsela, Munzhedzi James Mafela, Mampaka Lydia Mojapelo, and Rose Masubelele. 2016. African wordnet: A viable tool for sense discrimination in the indigenous african languages of south africa.

Lutz Marten. 2013. Structure and interpretation in swahili existential constructions. *Italian Journal of Linguistics*, 25(1):45–73.

Richard S Mgullu. 1999. *Mtalaa wa isimu: fonetiki, fonolojia na mofolojia ya Kiswahili*. Longhorn Publishers.

George A Miller. 1985. Wordnet: Dictionary browser'in information in data.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Mohamed Abdulla Mohamed. 2001. *Modern Swahili Grammar*. East African Publishers. Google-Books-ID: jyNEDFSK9IUC.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

I Niles, A Pease Ike, and undefined 2003. 2004. Linking lixicons and ontologies: Mapping wordnet to the suggested upper merged ontology. *researchgate.netI Niles, A PeaseIke, 2003•researchgate.net*.

Dunlop Ochieng. 2015. Indirect influence of english on kiswahili: The case of multiword duplicates between kiswahili and english.

Ojiambo. 2023. Kiswahili lessons. Accessed: 2023-10-25.

Benard Odoyo Okal. 2018. Uchanganuzi wa kileksikografia wa hiponimia za leksimu nomino na vitenzi vya kiswahili na tafsiri zake katika kiluo na benard odoyo okal tasnifu iliyowasilishwa kwa minajili ya kutimiza baadhi ya mahitaji ya shahada ya uzamifu katika kiswahili idara ya kiswahili na lugha nyingine za kiafrika chuo kikuu cha maseno.

Pēteris Paikens, Mikus Grasmanis, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde, and Laine Strankale. 2022. Towards latvian wordnet.

Princeton. 1980. Wordnet a lexical database for english. Accessed: 2023-10-20.

Adam Rambousek. 2006. Debvisdic–first version of new client-server wordnet browsing and editing tool. *Proceedings of the Third International WordNet Conference (GWC 2006)*. ; accessed 2023-10-27].

Adam Rambousek and Ale s Horák. 2015. *Management and Publishing of Multimedia Dictionary of the Czech Sign Language*, volume 9103, pages 399–403. Springer International Publishing. DOI: 10.1007/978-3-319-19581-037.

Borislav Rizov. 2014. Hydra: A software system for wordnet.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Casper S Shikali and Refuoe Mokhosi. 2020. Enhancing african low-resource languages: Swahili data for language modelling. *Data in Brief*, 31:105951. ; accessed 2023-10-25].

Dagobert Soergel. 1998. Wordnet. an electronic lexical database.

Edward Steere. 1882. A handbook of the swahili language as spoken... | hathitrust digital library.

Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. Balkanet : Aims , methods , results and perspectives . a general overview.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*.

Piek Vossen. 1998. Introduction to eurowordnet. *Computers and the Humanities*, 32:73–89. ; accessed 2023-10-27].

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. *arXiv preprint arXiv:2004.06748*.

Ng' Ang Wanjiku. 2005. Word sense disambiguation of swahili: Extending swahili language technology with machine learning.