JOURNAL OF INFORMATION SCIENCE AND ENGINEERING XX, XXXX-XXXX (2024)
DOI: XX.XXX.XXXX.XX

Developing A Dynamic WordNet for Under-Resourced Languages

STEPHEN OBARE, ABEJIDE ADE-IBIJOLA AND KENNEDY OGADA +
Department of Computer Science
Jomo Kenyatta University of Agriculture & Technology
E-mail: {smobareo@gmail.com}

The development of WordNets has contributed to a number of tasks in Natural Language Processing (NLP) and other research based on human language processing. While there is growing interest in building WordNets for popular languages, there are no major efforts for African languages which are evolving and commonly used by younger generation in social media platforms. Even where there are minimal efforts, no work exist that has comprehensively addressed the challenge of creating and updating such WordNets as new words are coined and meaning of words change. This paper presents a novel technique implemented in a software tool called ”Sense-Mapper” that maps Princeton WordNet synsets to concepts extracted from a lexical resource, extracts additional words from social media platforms, assigns senses to the new words and identify optimal location in the WordNet to insert the new words to cater for evolving vocabulary. We assess the performance and effectiveness of Sense-Mapper using lexical resources and data generated from social media platforms in Kenya and show that the proposed tool achieved an accuracy of 87.34% in mapping senses between lexical resources and 88.75% in updating our WordNet. Sense-Mapper is expected to find application in a number of NLP tasks including those that require assigning senses to out-of-vocabulary (OOV) words and inducing concepts from lexical resources which is important when constructing WordNets for under resourced languages.

Keywords: WordNet, natural language processing, under resourced languages, social media platforms, out-of-vocabulary words.

1. INTRODUCTION

There has been an increase in the use of microblogging sites such as Twitter to relay important information [1] which include occurrences of events such as crime [2], sentiment analysis and predicting election results [3]. A study prepared and published by [4] shows that Kenya, which is a multicultural and multilingual country with over 42 languages, is the second most active in Twitter usage in Africa[1] and microbloggers participate from all parts of the country using a mixture of English, Swahili, Sheng, slang and local dialects in a complex and unstable manner [5], giving rise to OOV words which are evolving in nature and are not included in the current WordNets [6], [7]. The existence

Received May 14, 2024; revised August 13, 2024; accepted September 28, 2014.

+Corresponding author

[1] https://businessdailyafrica.com/corporate/tech

of OOV words, which are defined as, newly coined words or words whose meanings have changed and are not known by existing WordNets, in online and social media platforms and the lack of WordNets that capture the dynamism in their use make processing extremely challenging for current state-of-the-art NLP algorithms [8]. To understand the meaning of OOV words, word sense disambiguation (WSD) tools are implemented to examine contextual information and provide evidence for determining the intended word sense [9]. WSD is a challenging task in developing countries such as Kenya where there is inadequate computer readable text corpora and WordNets adapted to the dynamic local language in use [10].

WordNets have been used for NLP tasks such as WSD and semantic relatedness (SR) since the popularisation of social media platforms [11], [12]. A WordNet is a semantic resource that groups parts of speech (nouns, adjectives, verbs, and adverbs) into linked synonym sets commonly referred to as synsets [13]. The synsets are linked through conceptual semantic and lexical relations. WordNets also include hypernyms and hyponyms in their structure and they have largely been developed in Western countries where the population is large, funding is available, technology is advanced and there is adequate computer-readable text corpora [14]. Most African languages do not have WordNets because of a number of reasons including limited funding, inadequate computer-readable text corpora, dynamic nature of the languages, lack of definate rules of grammar or known syntax [5]. Even with the above challenges, the fundemental question still is not *whether*, but *how* to construct a WordNet that is dynamic enough to capture every day usage of words.

One of the major challenges in constructing WordNets is how to accurately and effectively assess their performance. There are generally two main approaches using in evaluating WordNets: comparing the newly constructed WordNet against an established one or manual evaluation which requires human annotators. Comparing newly created WordNet with a reference WordNet is a challenging task since most newly created WordNets are not only smaller but also difficult to link with or compare to reference WordNets such as Princeton WordNet [15]. The variation in size complicates the process of determining whether the synsets created in a new WordNet are correct when compared to reference WordNets in different languages [16]. Manual evaluation involves human reviewers assessing the accuracy of WordNets either on their own or with automated evaluations [17]. Xu et al., [18] observed that without consistent guidelines, it's hard to gauge how accurate these manual evaluations are. Some studies simply asked manual annotators to decide whether WordNet is semantically similar to a reference one [19]. Others take a more nuanced approach based on weighting accuracy techniques such as Likert scale to rate the degree of correctness [20].

As at the time of writing this paper, there is no evidence in literature on creation of a WordNet that is dynamic enough to address everyday usage of words thereby ensuring that a WordNet does not become obsolete [5]. We address this challenge in this paper by proposing a comprehensive approach for languages used in Kenyan online and social media platforms. We make the following contributions:

1. We present an approach that automatically maps concepts from Helsinki Corpus of Swahili to Princeton WordNet synsets using shared features. We refer to the resulting resource as ”extended WordNet”,

2. We extract OOV words from online and social media platforms, assign senses, find their optimal location and insert them in the extended WordNet,

3. We present an approach that ”listens” continuously to OOV words from online and social media platforms and update our WordNet using contribution 2 above, and

4. We package the above contributions in a software tool called ”Sense-Mapper” and present a WordNet for the dynamic Kenyan language.

The rest of the paper is organised as follows. We present the problem statement, motivation and related work in Section 2. Section 3 outlines the approach we used to construct and update our WordNet. Section 4 describes the experimental design, implementation, evaluation and results. Finally, conclusions and future work is presented in Section 5.

2. PROBLEM STATEMENT, MOTIVATION AND RELATED WORK

This section defines the problem addressed in this paper, presents our motivation and discusses related work on constructing, updating and evaluating WordNets.

2.1 Problem Statement

The problem we address in this paper can be summarized by the following question:

> *given a dynamic language that has no defined grammatical rules or known syntax features and whose vocabulary is not captured in a WordNet, how do we leverage on available lexical resources, online and social media platforms to construct a dynamic WordNet that can take care of changes in word usage?*

We answer the question in Section 3 and present the results of our implementation in Section 4.

2.2 Motivation

The internet is a platform that facilitates a vast number of users to interact and post opinions and views from different regions using a number of languages[2] [21], [22]. This vast user base is rich in a variety of languages that allows internet language to morph[3] [23], [24]. In Kenya, slang, which is newly coined words or expressions which are not found in a WordNet, play a significant role in language use and change [25]. For example, Sheng, which is the language of masses in Kenya has morphed over time: Sheng that was spoken at the beginning of 2023[4] is different from the Sheng that is spoken at the end of 2023 [5], [26].

Example 1, which is based on words extracted from Kenyan online and social media platforms illustrates the dynamic and colloquial usage of a sample slang.

[2] http://www.worldbank.org/en/news/feature/2014/07/03/

[3] http://labs.theguardian.com/digital-language-divide/

[4] https://www.sde.co.ke/pulse/article

Example 1. *a). **wenger** – a miser: example in a sentence: "mpoa wa mine alinipeleka klabu the other day but msee ni Wenger mbaya".*

While the word *Wenger* means *a football manager* in Europe, in Kenya's Urban dictionary's[5] gloss, it means *a person who is unwilling to spend money.* Colloqualism is rampant in Kenya as online and social media facilitates discussion on topics of interest. Constructing a WordNet that can address the challenges posed by such language use in online and social media platforms for the significant number of social media users in Kenya is the main motivation behind this work.

2.3 Related Work

Unlike other regions with WordNets, Kenyan languages consisting of Kiswahili, Sheng and other local languages are not listed in Global WordNet Association[6] making them under resourced [27]. We describe in this subsection previous work on constructing, updating and evaluating WordNets and how they have informed our work.

2.3.1 Existing approaches for constructing WordNets:

WordNets have been constructed using a number of techniques [28], [29], [30] with the most common being merge or expansion approaches [31]:

> a). **Merge approach** – this technique compiles word senses and creates synsets containing all applicable words for a given sense [32] which is suitable for constructing WordNets in well resourced languages.
>
> b). **Expansion approach** – this approach creates a WordNet from existing synsets. Research has shown that this technique is best suited for developing WordNets for under-resourced languages [33], [34], [27], [35], [36]. WordNets are expanded either automatically or manually [37].

We adopt in this paper expansion approach which is based on translation theories on interlingual links and word sense equivalence[38], [39] as the theoretical base for creating our WordNet. A summary of WordNets developed based on expansion approach that are relevant to this work is provided below.

1. **EuroWordNet:** The developers of EuroWordNet project linked several European languages through English WordNet [40]. The project identified a list of common synonyms in English and associated it with its equivalent in other European languages [41]. The English synonyms provided a starting point for the development of EuroWordNet. We adopt a similar approach to extract synsets from Princeton WordNet that formed the starting point for developing the dynamic WordNet.
2. **Persian WordNet:** For Persian WordNet, the authors constructed a core WordNet based on common concepts, expanding the core WordNet and finally enriching the WordNet by adding semantic links [41]. For enriching the WordNet, the authors used a bilingual dictionary, a large amount of Persian and English corpora

[5] https://www.urbandictionary.com/define.php?term=Wenger

[6] http://compling.hss.ntu.edu.sg/omw/

to map Princeton WordNet synsets with Persian words [42]. We expanded our dynamic WordNet using concepts from Helsinki Corpus of Swahili and OOV words extracted from social media media platforms.

3. **Finnish WordNet:** In Finnish WordNet creation, the developers aligned the WordNet with the Princeton WordNet by assuming that most of Princeton Wordnet's synsets represented language-independent real word concepts. [41], [43]. After creating the base WordNet, the development team used a bilingual resource to find new synonyms candidates for enriching Finnish WordNet. The work on Finnish WordNet is relevant to our work since we used different lexical resources to enrich our WordNet.
4. **Polish WordNet:** This WordNet provides an accurate and comprehensive description of Polish lexical semantics and relationships between lexical meanings. It was derived from Polish language data and excluded translations from other languages to increase its accuracy [44].
5. **African WordNet (AWN):** In Africa[7], there have been efforts to construct a WordNet for South African languages [27] with very little effort in several countries such as Kenya. Griesel et al., [27] created a multilingual WordNet based on aligning several South African spoken languages. Based on the alignment approach, the authors further proposed to link the WordNet with global WordNets to make cross-linguistic research and development possible. The DEBVisDic[8] used for editing WordNets, was used to develop the WordNet. Ng'ang'a et al., [45] presented a method to extract meaning of Kiswahili words from corpora based on machine translation (MT). The authors proposed using the extracted semantics in augmenting a lexicon to improve the performance of NLP tasks. The proposed method however lacked the ability to assign meaning to words based on context of use. Hurskainen., [46] developed a Swahili Language Manager (SALAMA[9]), a computational environment used to develop different language applications. The author developed an annotated corpus of Swahili[10] containing about 25 million words to facilitate raw translation from Swahili to English. We propose to map this resource to synsets extracted from Princeton WordNet as a first step towards constructing our dynamic WordNet.
6. **CoreNet:** Kang et al., [47] mapped most of CoreNet's semantic categories, not the word senses. McCrae et al., [48] proposed using human annotators to manually label WordNet synsets to CoreNet's word senses with appropriate lexical relations, however, this required human labor. Kang et al., [49] mapped CoreNet-KorLex-Princeton WordNet-SUMO in order to apply it in broader fields and enhance its international status as a multilingual lexical semantic network.
7. **BabelNet:** Navigli et al., [50] mapped WordNet to Wikipedia using a word-sense disambiguation algorithm that created contexts using surrounding synsets of WordNet entities and article Wikipedia articles. A second step then selected the highest scoring mapping based on structuring the Wikipedia page content using WordNet relations. The authors reported a maximum F-Measure of 82.7% with a precision

[7] https://africanWordNet.wordpress.com/

[8] https://deb.fi.muni.cz/proj_debvisdic.php

[9] https://researchportal.helsinki.fi/en/publications/salama-swahili-language-manager

[10] https://korp.csc.fi/download

of 81.2%, showing that while BabelNet is a high-quality resource, it cannot be considered a gold standard.

8. **Database of Cross-Linguistic Colexifications:** Rzymski et al., [51] proposed a database which was constructed by integrating word lists representing thousands of languages. Each concept is linked to a gloss, a set of lexicalizations, and a category, not part of speech. A unique English word or phrase is assigned to each concept to describe its meaning. However, there are no listings of semantic relations between concepts in BabelNet.

Unlike languages with WordNets whose development have been aided by adequate amounts of quality online resources, indigenous Kenyan languages do not have adequate machine-readable resources. We focus on the methods proposed by [42], [41], [44], [50] and [51] where the task is to extract synsets from Princeton WordNet as a starting point to developing a dynamic WordNet and expand the WordNet with corpus developed in the work of [46].

2.3.2 Updating WordNets:

We have presented relevant techniques for constructing WordNets in Subsection 2.3.1. While natural language processing affords researchers an opportunity to automatically scan millions of social media posts, there is growing concern that automated computational tools lack the ability to understand context and nuance in human communication and language. We present in this subsection existing techniques for detecting OOV words and assigning senses [52].

2.3.2.1 Detecting OOV Words:

Detecting OOV words has received widespread attention despite the significant weaknesses of its implemented approaches. According to Senapati et al., [53], initial efforts at detecting OOV word relied on a manual analysis of texts from sources such as newspapers with the drawback of being time-consuming and tedious. The advent of new data collection methods necessitated the introduction of automated tools that can scan and identify new words [54]. Falk et al., [55] proposed semi-automated detection by extracting relevant features and classifying them using Support Vector Machine (SVM). Breen et al., [56] further proposed SVM that is reliant on language-specific features. Pyo [57] used supervised machine learning and training data to improve the accuracy of OOV collection procedure.

2.3.2.2 Sense Assignment

A number of techniques have been proposed for discerning senses of OOV words and finding an ideal lemma's location based on its gloss(es) [58]. We present some of the key techniques.

1. **Lesk's Algorithm:** Also known as Gloss Overlaps, this algorithm compares a word's sense definitions to identify overlapping definitions from different words within similar contextual proximity [59].
2. **Extended Gloss Overlaps (EGO):** While Lesk's algorithm primarily targeted WSD, EGO [60] approach is geared toward assessing semantic relatedness. EGO extends Lesk's algorithm by incorporating Wornet into the calculation of gloss overlaps by adding hypernyms and hyponyms of lemmas and glosses. EGO compares not only the definition of two terms but also of their respective hypernyms and hyponyms.

3. **Gloss Vectors:** Patwardhan et al., [61] made a significant contribution to this field by exploring the integration of context vectors with WordNet to measure Semantic Relatedness. Traditional context vectors determine the relationship between two words by analyzing the words or context surrounding them. They weigh the closeness of two terms based on the frequency of certain words occurring together in various contexts. Gloss vectors address the challenge faced by EGO and the Lesk algorithm in reducing their reliance on glosses and gloss overlaps to achieve success.
4. **Community enRiched Open WordNet (CROWN):** While EGO and gloss vectors primarily relied on WordNet for determining semantic relatedness, Jurgens et al., [62] introduced Wiktionary[11] into the calculation of SR to improve the identification of the optimal OOV lemma location in a WordNet. CROWN's objective was to augment the WordNet synsets by incorporating OOV terms.
5. **Large Language Models (LLM)** Lietard et al., [63] and Zhou et al., [64] proposed to extract vector representations of word usage using neural Contextualized Language Models (CLM) and feed the representation to a classification for WSD. Though powerful, LLMs are faced with a number of challenges which include: the size and complexity of the datasets on which they are trained is one of the most significant challenges. These models are typically trained on enormous corpora of Internet-sourced text data which are not readily available in third world countries [65]. The training of LLMs is a computationally intensive procedure that requires substantial hardware and energy resources [66]. LLMs can only evaluate a limited number of preceding tokens when generating text due to their limited context window [67]. These limitation present difficulties when working with lengthy text messages from online and social media platforms.

To assign senses to OOV words, we proposes an extension to the above algorithms by integrating additional lexical semantic resources such as Sheng dictionary, Kamusi, Urban dictionary, Shembeteng etc into overlap calculations that incorporates, compares, and evaluates the definitions of lemmas and glosses of hypernyms and hyponyms [68], [69].

2.3.3 Evaluating WordNets:

Manual mapping is used to prepare gold standard data for testing and evaluating the performance and effectiveness of Synset-mapper. We evaluate the ability of our tool to:

a). Correctly map concepts from Helsinki Corpus of Swahili onto synsets in Princeton WordNet;

b). Correctly assign senses to the identified OOV words, and

c). Find optimal location to insert the OOV word in a WordNet.

We then computed the differences or similarities between the action of Synset-Mapper and human evaluators.

[11] https://www.wiktionary.org/

3. WORDNET CONSTRUCTION

We propose the following steps to construct a WordNet while providing freedom and flexibility for refinement as new words are introduced in the vocabulary:

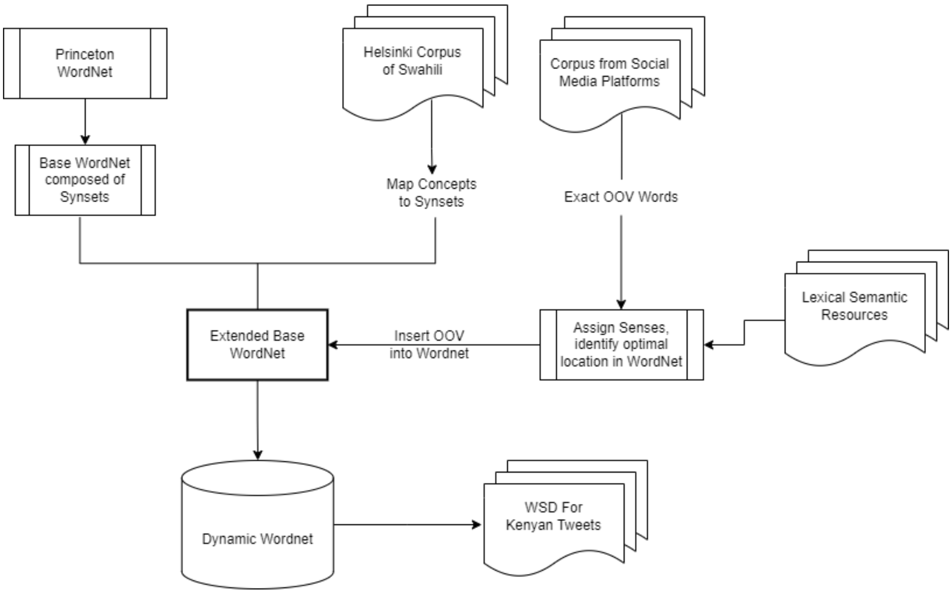

Princeton WordNet
Helsinki Corpus of Swahili
Corpus from Social Media Platforms
Base WordNet composed of Synsets
Map Concepts to Synsets
Exact OOV Words
Lexical Semantic Resources
Extended Base WordNet
Insert OOV into Wordnet
Assign Senses, identify optimal location in WordNet
Dynamic Wordnet
WSD For Kenyan Tweets

Fig. 1. A block diagram for constructing a dynamic WordNet

Constructing base WordNet.

1. Define lemmas of interest from Princeton WordNet.
2. Automatically map concepts from Helsinki Corpus of Swahili to extracted lemmas from Princeton WordNet.

Assigning senses to OOV words.

1. Iterate over online and social media platforms to extract OOV words using Algorithm 3.
2. Using gloss(es) inferred from lexical semantic resources described in Section 2.3.2, determine sense of lemmas from Step 1 using Algorithm 4.

Updating WordNet.

1. Find the optimal location for the lemma assigned sense in Step 2 and insert them in the WorNet using Algorithm 5.
2. Continuously update WordNet by crawling social media platforms for new words, assigning senses and finding optimal location to either attach or merge the new words.

The proposed steps are summarized in Figure 1 and detailed in the subsequent Sections.

3.1 Constructing base WordNet

We start by utilizing part-of-speech (pos) information which can acquire new words and carry semantic load from Princeton WordNet to extract synsets of interest that is mapped to concepts from Helsinki Corpus of Swahili. This approach is motivated by the need to establish a strong way to ensure that the two lexical resources can be optimally linked based on equivalent meanings of terms.

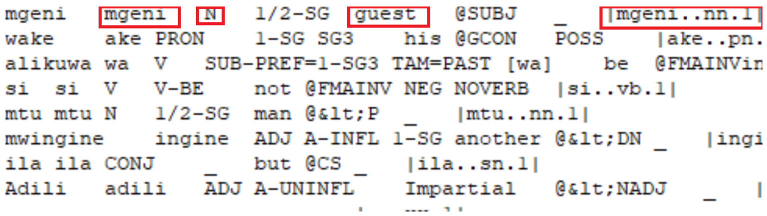
3.1.1 Princeton WordNet

WordNet[12] is comprising 155,287 words and their hypernyms and hyponyms, grouped into 177,659 synsets. Princeton University developed it. The large amount of words and relationships make WordNet an invaluable resource in NLP systems. Each record in WordNet has a lemma with definition, meaning that the acquisition's iteration can be done for every lemma and definition in the dictionary. The acquisition's objective is to find nouns and verbs consisting of a lemma, part-of-speech, and definition.

WordNet's data are stored in low-level representations, with each specific sense's location represented with hard-coded byte-off addresses [70]. The source files are divided into four pos (nouns, verbs, adjectives, adverbs). The lexicographer files each store word and its gloss and pointer to other words in WordNet. These pointers mark relationships such as hypernyms.

3.1.2 Helsinki Corpus of Swahili

Helsinki Corpus of Swahili 2.0 Annotated dataset Version[13] is a valuable resource which contains about 25 million words and has under gone revisions. An annotated word in the corpus has a lemma, part-of-speech, morphological description, gloss, syntactic tag, and verb description. Lemmas from Helsinki Corpus of Swahili are already translated into English words with the translation having an exact matching part-of-speech as shown in Figure 2.



```
mgeni   mgeni  N   1/2-SG guest   @SUBJ    _   |mgeni..nn.1|
wake    ake PRON   1-SG SG3    his @GCON   POSS    |ake..pn.
alikuwa wa  V   SUB-PREF=1-SG3 TAM=PAST [wa]    be  @FMAINVin
si  si  V   V-BE    not @FMAINV NEG NOVERB  |si..vb.1|
mtu mtu N   1/2-SG  man @&lt;P  _   |mtu..nn.1|
mwingine    ingine  ADJ A-INFL 1-SG another @&lt;DN _   |ingi
ila ila CONJ    _   but @CS _   |ila..sn.1|
Adili   adili   ADJ A-UNINFL    Impartial   @&lt;NADJ   _   |
```

Fig. 2. Extract from Helsinki Corpus of Swahili.

We choose Helsinki Corpus of Swahili because of the credibility of the producer, which is the Language Center of Helsinki University. Even though the corpus contains a number of information, we elected to extract only lemma, part-of-speech, and definition because only those three elements are related to WordNet structure.

[12] https://WordNet.princeton.edu/

[13] https://korp.csc.fi/download

3.1.3 Defining Synsets of Interest

For conceptual modeling, we extract nouns since they present part whole, meronymy or holonymy relations. The class also captures the relationship between interlinked synsets and their derivational origin, which indicates their formation. For instance, the derivational origin of the verb to “*develop*” is the root for the noun “*development*”.

3.1.4 Extracting Synsets of Interest to construct base WordNet

We modified the algorithm published on ”*Corpus extraction of noun using nltk”*[14] to nouns from Princeton WordNet.

Algorithm 1: extracting lemmas of interest from Princeton WordNet

```
Princeton WordNet = import( );
Function create base WordNet
    forall words in Princeton WordNet do
        if word is of interest i.e, pos = ”noun” then
            add lemmas to database;
        end
        retrieve only lemmas of similar lemma form (noun check noun) or
          adjectives;
        get lemmas that are derivatives of other lemmas;
    end
    filter further with above criteria to get lemmas of similar forms;
end
```

Algorithm 1 iterated through the Princeton WordNet data files extracting nouns with their glosses which formed our base WordNet. The process can be repeated for the other POS.

3.1.5 Extending base WordNet with concepts from Helsinki Corpus of Swahili

Conceptual similarity based on English definitions from both resources was used to map terms from Helsinki Corpus of Swahili to synsets extracted from Princeton WordNet. To achieve the mapping, we calculate the conceptual terms coverage which measures the number of words in the names of the semantic categories of the lexical resources i.e., a term in Helsinki Corpus of Swahili for a given Princeton WordNet noun sense.

Using the example in Figure 3, Algorithm 2 maps ”*miser*”, which is source concept to ”wenger” thereby maximizing the lexical intersection computed as follows:

$$s_W\left(c_s, c_t\right) = \left|\left\{\mathrm{lex}\left(c_s, \mathscr{L}\right) \cap \mathrm{lex}\left(c_t, \mathscr{L}\right)\right\}\right| \tag{1}$$

where $\mathrm{lex}(c, \mathscr{L})$ is a function that returns the set of lexicalizations of the concept c in a set of languages $\mathscr{L}$.

The gloss of ”*miser*” indicates simply *”a stingy hoarder of money and possessions”*, that would be used for mapping to the Helsinki Corpus of Swahili, ”*wenger*”.

[14] https://nlpforhackers.io/

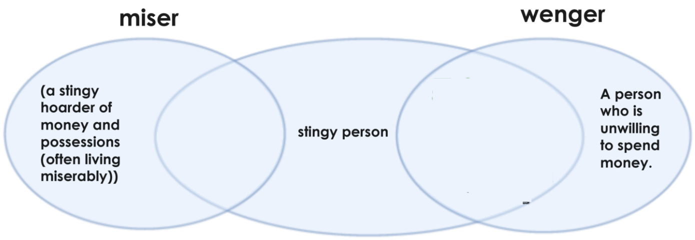

miser
wenger
(a stingy hoarder of money and possessions (often living miserably))
stingy person
A person who is unwilling to spend money.

Fig. 3. Maximizing lexicalization overlaps

```
Algorithm 2: mapping for "miser" and other synsets of interest
  Data: Synsets of interest from Princeton WordNet
  Result: Extended base WordNet
  begin;
  if hasPrinceton WordNet(s, Miser) then
      return wenger;
      attach "wenger" to "miser";
  else
      simultenously map similar lemmas to "miser", "wenger", their
        holonym/meronym to another noun and adjectives;
      end;
      extended WordNet;
  end
```

The result of this process are nouns where each lemma is associated with the appropriate part of speech and corresponding definitions.

3.2 Updating extended base WordNet with OOV

We hypothesise that words which appear in social media platforms and are not in extended WordNet created in Section 3.1.5 have a high probability of being Sheng, slang, slang or any of the local languages, herein referred to as, an OOV word. We extract such words using Algorithm 3, assign senses using Algorithm 4, find their optimal location in extended WordNet and insert the OOV using Algorithm 5.

3.2.1 Extracting OOV

Algorithm 3 takes social media texts messages as input and examines all words that consists of alphanumeric characters and categorizes them into either in vocabulary or OOVs words relative to the extended base WordNet created in Section 3.1.5.

The OOVs are taken as candidates for parsing thereby returning among other words, nouns and verbs as depicted in Figure 4.

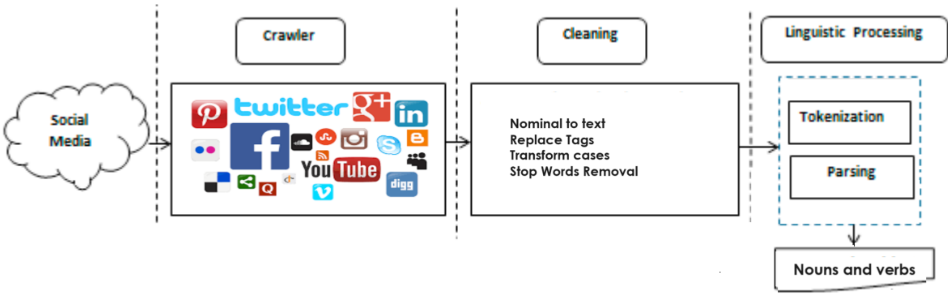


Fig. 4. Framework for information extraction from social media platform

3.2.2 Assigning Senses to OOV

For each OOV, we acquire all information necessary for assigning senses which include lemma definitions, hypernyms and hyponyms from a number of lexical semantic resources such as Sheng dictionary, urban dictionary, Shembeteng, social media corpus generated in Section 3.2.1 etc. Algorithm 4 implements the process of assigning senses to OOV words.

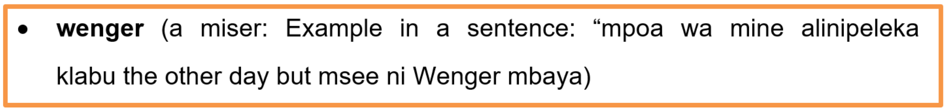
3.2.3 Enriching extended base WordNet

We propose a two step approach to enrich extended WordNet with lemmas created in Section 3.2.2 as follows:

a. **Finding location:** We use a hybrid algorithm by combining overlap, overlap with stemming, Word2Vec to find the optimal location for an OOV word.

b. **Inserting a word into WordNet**: The algorithm then decides if the OOV should be attached or merged into a synset of the selected sense.

Algorithm 5 finds the target synset in extended WordNet for which the OOV lemma should be attached or merged to, by overlapping words in the OOV lemma's gloss with words in each target gloss as well as hypernym and hyponym glosses. We demonstrate the process of finding optimal location and attaching or merging new words to improve the constructed WordNet using Example 1 described in Section 2.2 as follows:



- **wenger** (a miser: Example in a sentence: “mpoa wa mine alinipeleka klabu the other day but msee ni Wenger mbaya)

Fig. 5. Shembeteng gloss for the word ”*wenger.*”

1). Retrieve gloss for the term ***wenger*** from any of the lexical dictionaries mentioned in Section 3.2.2.

2). Create a possible set of hypernym/hyponym candidates by looking at their gloss from the corpus constructed in Section 3.2.1.

Algorithm 3: extract OOV words

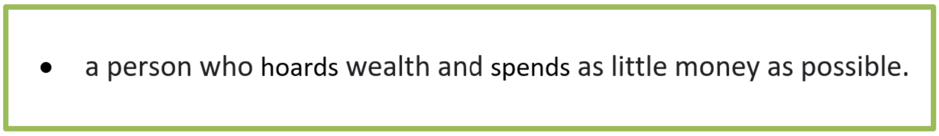
```
input: Corpus stored in CSV FILE;
output: database of OOV tokens;
Function extract non-English words
    forall words in social media corpus do
        preprocess tweets by removing punctiation, tags, urls, stopwords etc;
        tokenize individual words and store in array;
        load preprocessed corpus and pass through base WordNet using linear
          search algorithm;
        lemma w is read from corpus;
        compare w with lemmas in base WordNet;
        if w = base WordNet;
         then
            remove w;
        end
        return percentage of OOV lemmas;
        else extract to a text file;
    end
    Return database of OOV lemmas;
end
```

Algorithm 4: assigning senses to lemmas

```
input: A list of non-English lemma $E$, online resources including Kiswahili
  Dictionary, Urban dictionary, Shembeteng etc;
output: sense assigned lemma;
Function sense assignment
    foreach lemma $e \in E$ do
        Preprocess
        Iterate through each lemma and retrieve each sense of each lemma, store
          array;
        Iterate through array, one by one to obtain a score for each sense;
        Iterate through each sense and obtain the senses gloss;
    end
end
```

Algorithm 5: Inserting a word into WordNet

```
input: A list of sense tagged OOV;
output: optimal location of the input;
Function sense assignment
    Location Algorithms
    if a sense is obtained from WordNet;
    then
        Obtain sense's gloss from the hash initialized by Algorithm 4;
        Retrieve the sense's immediate hypernyms and their glosses and add to
          the expanded sense;
        Retrieve sense's immediate hyponyms and their glosses and add to the
          expanded sense;
        Retrieve and add sense's corresponding synset and glosses and add to
          the expanded sense;
        Cleanup new lemma's gloss;
    end
    Decide whether or not the new lemma should be attached to the synset of
      the chosen sense, or merged into it
end
```



- a person who hoards wealth and spends as little money as possible.

Fig. 6. Gloss for ”*wenger.*”

3). *Person* would be placed in the set of possible candidates since it matches the first round of prepossessing, *miser* is a person and since person is the first word extracted that exists in *wenger's* gloss.

4). Gloss based attachment is then used as each term in *wenger's* gloss is analyzed and the highest scoring related term is selected as the hypernym. In this case *wenger's* gloss makes it the ideal hypernym candidate since person overlaps several times between the glosses, therefore *wenger* is attached to person in WordNet.

4. DESIGN, IMPLEMENTATION AND RESULTS

With over 57,746 noun synsets in Princeton WordNet and 25 million words in Helsinki Corpus of Swahili, it is possible to extract and match all the synsets of interest as described in Section 3 but impractical to evaluate all the synsets. We present the design of a representative sample of the synsets in Section 4.1. We show the implementation of the idea in Section 4.2 and present the results of our evaluation in Section 4.3.

4.1 Data acquisition

The data that was used in this work was from Helsinki Corpus of Swahili dataset, Princeton WordNet and crawled data from social media platforms in Kenya.

4.1.1 Princeton WordNet and Helsinki Corpus of Swahili

Python was used to convert the files into a Pandas data frame. Pandas returned a random sample of items given the sample size. To ensure that we are able to analyse the data, the number of samples was finally set at 8,614, DataFrame.sample (n = 8,614) to return 8,614 random synsets of interest from Princeton WordNet dataset. The file was converted to a text format and then read into a panda's dataframe using the read_csv() function (pandas.read_csv('file.txt', sep = ' ', header=0)). This option allowed the resulting sample to be saved into a CSV, text, or file format like xml and json. The synsets extracted from Princeton WordNet are mapped to equivalent terms in Helsinki Corpus of Swahili.

4.1.2 Social Media Corpus

A corpus was extracted from two social media websites: Twitter, where text was gathered using the Twitter REST API[15] in JSON format for a period of 3 months between May 2023 and August 2023; and Reddit, where we extracted data from the top 15 most popular forums (‘subreddits’) using a webpage crawler[16]. In total, we collected 10,210 Reddit posts (36,940 tokens) and 23,018 Twitter posts (124,324 tokens). We discarded all terms with a raw frequency of less than 5 in both Reddit and Twitter corpus and preprocessed the rest of the data by removing simple non-terms (such as phrases starting with ’*na*’, ‘*a*’ or ‘*the*’).

We then filtered the tokens to only those that did not occur in our extended WordNet. This gave us the ability to find words that would be relevant with high precision, and the annotators agreed that 93.4% of words were worthy of inclusion in the extended WordNet.

4.1.3 Data for Evaluating Synset-Mapper

We crawled about 3,264 tweets from five Twitter handles in Kenya as shown in Table 1 to evaluate the performance of Synset-Mapper.

Twitter account	No. of tweets scraped	OOV
@*KenyanTraffic*	434	26
@*NPSOfficial_KE*	617	23
@*suemc_phee*	1,175	173
@*ntsa_kenya*	206	13
@*DCI_Kenya*	832	76
Total	**3,264**	311

Table 1. Statistics of tweets per account

As a first step, annotators manually identified 311 OOV words, assigned meanings to the words, found their optimal location and inserted the OOV words in the extended

[15] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview.html

[16] https://www.octoparse.com/

WordNet. Synset-Mapper was given the same set of input and its effectiveness on identification of OOV words, assigning senses, finding optimal location and inserting the OOV words was compared with the results from the annotators.

4.2 Implementation

We implemented the approach described in Section 3 as a Java based software tool called ”Sense-Mapper”. Sense-Mapper automatically determined the sample size to be 8,465. The sample size was calculated assuming a confidence level of 94% and a margin of error of 5%.

4.2.1 Mapping Rule

The 8,465 random samples from Princeton WordNet is fed into Sense-Mapper, which automatically maps synsets to their equivalents in Helsinki Corpus of Swahili. Our goal is to match the lemmas of WordNet entries to terms in Helsinki Corpus of Swahili based on gloss definitions so lemma “miser” is matched to “wenger” as shown in Figure 5. This method captured most of the mappings as only 137 Princeton WordNet synsets have no candidates in Helsinki Corpus of Swahili as shown in Table **2**.

Table 2. Mapping WordNet Instances to Helsinki Corpus of Swahili

Extended WordNet	**Size**			**Size**
Entries	16,930			
Princeton WordNet	8,645		Exact matches	7,784
Helsinki Corpus of Swahili	8,645		Broad	479
Unmapped	137		Narrow	65
			Unmapped	137
Sense relations	5,649			

If the Princeton WordNet synset and Helsinki Corpus of Swahili exactly describe the same entity, then we marked it as *Exact*, *Broad* if Helsinki Corpus of Swahili describes several things, of which the entity described by the Princeton WordNet synset is only one off. *Narrow*, if Princeton WordNet synset describes multiple Helsinki Corpus of Swahili. *Unmapped* if Helsinki Corpus of Swahili does not describe the Princeton WordNet synset.

4.2.2 Sense Assignment

Sense-mapper iterated through the corpus that was extracted from the two social media websites as described in Section 4.1.2 comparing the filtered tokens against all synsets in Princeton WordNet and concepts in Helsinki Corpus of Swahili. After filtering, out of about 161,264 tokens, we were left with 456 tokens that were assigned senses. The consolidated statistics for the created resource is presented in Table **3** with a description of the number of extracted entries from Princeton WordNet, entries mapped from Helsinki Corpus of Swahili and new entries added from social media platforms.

Since Princeton WordNet has fewer synsets than Helsinki Corpus of Swahili concepts, this means that we attempt to align each Princeton WordNet synset with a single Helsinki Corpus of Swahili concept. However, in some cases, no alignment is found, due either to not sharing any lexicalizations with a concept in the other resource, or to the

Table 3. Consolidated resource

	Size			Size
Entries	17,386			
Princeton WordNet	8,645		Exact matches	8017
Helsinki Corpus of Swahili	8,645		Broad	612
Unmapped	137		Narrow	105
Social media platforms	456		Unmapped	137
Sense relations	5,649			

one-to-one constraint. 40 out of 456 concepts extracted from social media platforms are not mapped.

4.3 Evaluation and Results

Our evaluation is limited to the 8,645 Princeton WordNet synsets and the corresponding Helsinki Corpus of Swahili concepts which comprise our test set. The first evaluation involves all the synsets and concepts in Princeton WordNet and Helsinki Corpus of Swahili, that is, we map all synsets/concepts between the resources. The second evaluation involves checking the effectiveness of the tool in identifying OOV and assigning senses.

4.3.1 Evaluating Mapping Rule

A group of 10 evaluators from Andela Kenya[17] manually evaluated the mapping set that was used to create the resource described in Table **2**. All 10 evaluators were fluent in English, Kiswahili and Sheng as spoken and written languages. All 10 evaluators were informed about the research purpose and examined the same instruction sets for the experiment. Table 4 presents the number of mappings each evaluator marked as correct or wrong and the accuracy percentage in each test set.

Table 4. Synset-Mapper's score on mapping instances.

Evaluator	**Correct**	**Wrong**	**Total**	**Accuracy**
1	765	163	928	82.44%
2	674	117	791	85.21%
3	876	57	933	93.89%
4	516	83	599	86.14%
5	987	91	1,078	91.56%
6	805	95	900	89.44%
7	735	106	841	87.40%
8	612	138	750	81.60%
9	874	96	970	90.10%
10	732	123	855	85.61%
Total	7,576	1,069	8,645	**87.34**%

The proposed mapping rule achieved an average accuracy of 87.34% which shows

[17] https://andela.com/

its good performance and effectiveness in mapping words. We plotted Synset-Mapper's scores against the number of test sets as shown in Figure 7. Accuracy the tool does not increase as the number of instances reduces. In the test set, it can be observed that the tool performs consistently.

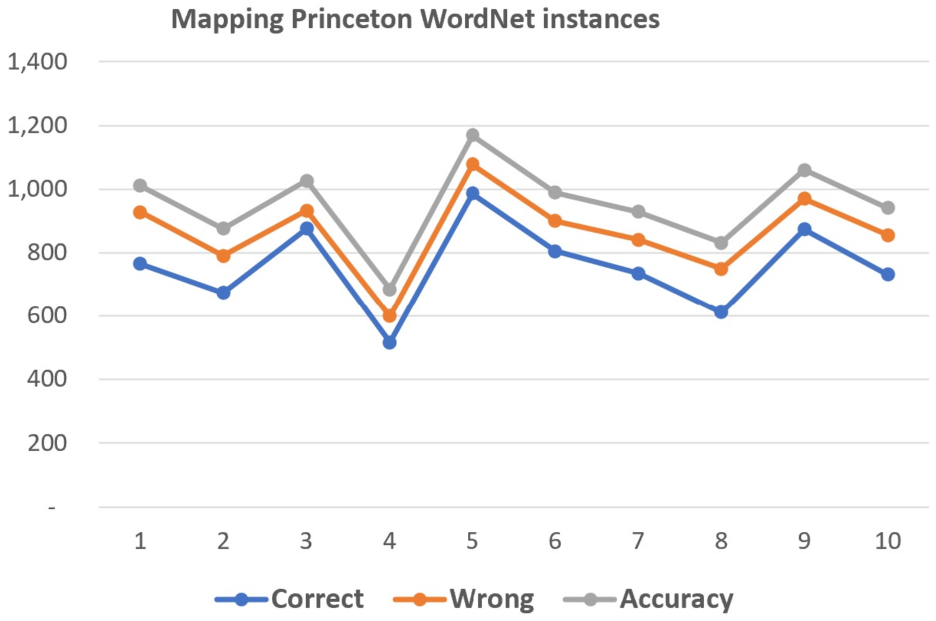


Fig. 7. A graph shows performance based on different sizes of data set

4.3.2 Evaluating Effectiveness

For testing Synset-Mapper's effectiveness in assigning senses, finding location and inserting OOV, annotators identified the 311 OOVs as described in Table 1, assigned senses, found location and inserted the OOVs into extended WordNet. To assign senses, annotators either selected an existing gloss from online lexical resources referring to the word or came up with new definitions. The same OOV words were subjected to Synset Mapper and the results of assigning senses, finding optimal location and inserting the OOV word was compared to the gold standard set as per the annotators.

Out of 311 OOVs, Synset-Mapper correctly identified, assigned senses to and inserted 264/311, 269/311 and 276/311 terms respectively giving an average score of over 85% as shown in Table 5. For evaluating the effectiveness of Synset-Mapper to update our WordNet using OOV words, the tool correctly inserted 276/311 or 88.75% words as shown in Table 5. We attribute the good performance of Synset-Mapper to the notion that OOV lemma's gloss often contained the hypernym to which it should be attached or merged which can be seen in many common lemmas' definitions and not the characteristics specific to the platforms.

Table 5. Synset-Mapper's performance on assigning senses, finding optimal location and inserting OOV words

Social Media platform		**Sense assignment**			**Finding Optimal location**			**Inserting OOV words**		
Hashtag	Total	Correct	Wrong	% Accuracy	Correct	Wrong	%Accuracy	Correct	Wrong	%Accuracy
@KenyaTraffic	26	23	3	87.4	24	2	91.4	24	2	92.3
@NPSOfficial	23	20	3	84.8	20	3	86.1	20	3	87.5
@suemec_phee	173	146	27	88.4	148	25	85.4	152	21	87.9
@ntsa_kenya	13	11	2	88.4	12	1	89.4	11	2	88.3
@DCI_kenya	76	64	12	83.9	66	10	87.4	68	8	89.4

4.3.3 Application of Sense-Mapper

Sense-Mapper will find application in the construction of knowledge graphs that map out the relationships between concepts within specific scientific domains, which can be used for advanced data analytics and reasoning. Additionally, the ability of the tool to disambiguate OOV words based on context can help in retrieving more relevant and precise information which is valuable in research.

In text analysis and data mining, Sense-mapper can be used to enable more sophisticated text analysis by allowing tools to understand the meaning and relationships of words in scientific texts, improving the quality of data mining and information extraction. Sense-mapper can also be used to support NLP tools in tasks like summarization, sentiment analysis, and keyword extraction in scientific documents.

5. CONCLUSION AND FUTURE WORK

We have presented in this paper an approach to construct a dynamic WordNet for common NLP tasks such as WSD in under resourced languages. To achieve this goal, we proposed a comprehensive technique that maps synsets of interest extracted from Princeton WordNet to concepts in Helsinki Corpus of Swahili which is further expanded using terms extracted from social media platforms in Kenya. The technique was implemented in a software tool called ”Sense-Mapper”. Sense-Mapper was able to map Princeton WordNet synset instances to concepts in Helsinki Corpus if Swahili with an average accuracy of 87.34%, identify, assign senses and update our WordNet with new words extracted from social media platforms with an accuracy of 88.75%. To ensure our tool is not deprecated, we propose to continuously monitor usage of new words in social media platforms. A new word would be of interest to this work and would be analysed, assigned sense and attached or merged with existing synsets in our WordNet to keep it up to date. For future work, we will explore the design of a hybrid system incorporating LLMs and deep learning method that infers embedding based on the context for assigning sense and finding optimal location in a WordNet. We will use Sense-Mapper to train LLM models.

REFERENCES

1. G. Golovchinsky and M. Efron, “Making sense of twitter search,” in *Proceedings of CHI 2010 workshop on microblogging: What and how can we learn from it*, 2010, p. 33.
2. N. S. Sasindrakumar, S. S. Ajit, and S. Sureshbabu, “Assist crime prevention using machine learning,” *PCE JCE_*, Vol. 7, no. 1, 2018, p. 75.
3. S. Obare, A. Ade-Ibijola, G. Okeyo, and K. Ogada, “Estimating crime rates using jumping finite automata on tweets,” *IAENG International Journal of Computer Science*, Vol. 50, no. 4, 2023.
4. J. Mwaura, “Digital dissidents or whistle-blowers? a critical analysis of microbloggers in kenya,” *Digital Dissidence and Social Media Censorship in Africa*. Routledge, 2022, pp. 175–194.

5. M. H. Abdulaziz and K. Osinde, "Sheng and engsh: Development of mixed codes among the urban youth in kenya," *International Journal of the Sociology of Language https://doi.org/10.1515/ijsl.1997.125.43*, 1997.
6. A. V. Dwivedi, "Linguistic realities in kenya: A preliminary survey," *Ghana Journal of Linguistics*, Vol. 3, no. 2, 2014, pp. 27–34.
7. H. ElSahar and S. R. El-Beltagy, "A fully automated approach for arabic slang lexicon extraction from microblogs," in *International conference on intelligent text processing and computational linguistics*. Springer, 2014, pp. 79–91.
8. S. Kusal, S. Patil, J. Choudrie, K. Kotecha, D. Vora, and I. Pappas, "A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection," *Artificial Intelligence Review*, 2023, pp. 1–87.
9. D. A. B. Loureiro, "Learning word sense representations from neural language models," *International conference on machine learning*, 2023.
10. J. Dong, "Transfer learning-based neural machine translation for low-resource languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.
11. C. Sindhu, R. Vinoth Kumar, and K. C. T Charandeep Reddy, "Context based sentiment analysis using twitter tweets," *Journal of Survey in Fisheries Sciences*, Vol. 10, no. 2S, 2023, pp. 493–503.
12. L. Abouenour, K. Bouzoubaa, and P. Rosso, "Erratum to: On the evaluation and improvement of arabic wordnet coverage and usability," *Language Resources and Evaluation*, Vol. 47, 2013, pp. 1343–1343.
13. P. Nanjundan and E. Z. Mathews, "An analysis of word sense disambiguation (wsd)," in *Proceedings of the International Health Informatics Conference: IHIC 2022*. Springer, 2023, pp. 251–259.
14. O. Majewska and A. Korhonen, "Verb classification across languages," *Annual Review of Linguistics*, Vol. 9, 2023, pp. 313–333.
15. M. Khodak, A. Risteski, C. Fellbaum, and S. Arora, "Automated wordnet construction using word embeddings," in *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, 2017, pp. 12–23.
16. M. Schulder, S. Bigeard, M. Kopf, T. Hanke, A. Kuder, J. Wójcicka, J. Mesch, T. Björkstrand, A. Vacalopoulou, K. Vasilaki *et al.*, "Signs and synonymity: Continuing development of the multilingual sign language wordnet," in *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, 2024, pp. 343–353.
17. Y. Chasseray, A.-M. Barthe-Delanoë, J. Volkman, S. Négny, and J. M. Le Lann, "A generic hybrid method combining rules and machine learning to automate domain independent ontology population," *Engineering Applications of Artificial Intelligence*, Vol. 133, 2024, p. 108571.
18. H. Xu, J. Lin, S. Pradhan, M. Marcus, and M. Liu, "Annotating chinese word senses with english wordnet: A practice on ontonotes chinese sense inventories," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 1187–1196.
19. M. Pouly *et al.*, "Estimating text similarity based on semantic concept embeddings," *arXiv preprint arXiv:2401.04422*, 2024.

20. S. Wang, G. Zhang, H. Wu, T. Loakman, W. Huang, and C. Lin, “Mmte: Corpus and metrics for evaluating machine translation quality of metaphorical language,” *arXiv preprint arXiv:2406.13698*, 2024.
21. A. Nowak and R. R. Vallacher, “Nonlinear societal change: The perspective of dynamical systems,” *British Journal of Social Psychology*, Vol. 58, no. 1, 2019, pp. 105–128.
22. K. Massoudi, M. Tsagkias, M. De Rijke, and W. Weerkamp, “Incorporating query expansion and quality indicators in searching microblog posts,” in *Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings 33.* Springer, 2011, pp. 362–367.
23. F. Cirillo, F.-J. Wu, G. Solmaz, and E. Kovacs, “Embracing the future internet of things,” *Sensors*, Vol. 19, no. 2, 2019, p. 351.
24. M. Gaikar, J. Chavan, K. Indore, and R. Shedge, “Depression detection and prevention system by analysing tweets,” in *Proceedings 2019: conference on technologies for future cities (CTFC)*, 2019, pp. 1–5.
25. F. Hardini, E. Setia, and U. Mono, “Translation norms of neologism in social media interface,” *LINGUA: Jurnal Bahasa, Sastra, Dan Pengajarannya*, Vol. 16, no. 1, 2019, pp. 15–24.
26. K. Matsumoto, F. Ren, M. Matsuoka, M. Yoshida, and K. Kita, “Slang feature extraction by analysing topic change on social media,” *CAAI Transactions on Intelligence Technology*, Vol. 4, no. 1, 2019, pp. 64–71.
27. S. E. Bosch and M. Griesel, “Strategies for building wordnets for under-resourced languages: The case of african languages,” *Literator (Potchefstroom. Online)*, Vol. 38, no. 1, 2017, pp. 1–12.
28. N. Taghizadeh and H. Faili, “Automatic wordnet development for low-resource languages using cross-lingual wsd,” *Journal of Artificial Intelligence Research*, Vol. 56, 2016, pp. 61–87.
29. D. Fišer and B. Sagot, “Constructing a poor man’s wordnet in a resource-rich world,” *Language Resources and Evaluation*, Vol. 49, no. 3, 2015, pp. 601–635.
30. S. Jimenez and G. Dueñas, “Lar-wordnet: A machine-translated, pan-hispanic and regional wordnet for spanish,” in *Advances in Artificial Intelligence-IBERAMIA 2018: 16th Ibero-American Conference on AI, Trujillo, Peru, November 13-16, 2018, Proceedings 16.* Springer, 2018, pp. 392–403.
31. P. Vossen, “Eurowordnet general document version 3,” *University of Amsterdam*, 2002.
32. B. Broda, R. Kurc, M. Piasecki, and R. Ramocki, “Evaluation method for automated wordnet expansion,” in *Security and Intelligent Information Systems: International Joint Conferences, SIIS 2011, Warsaw, Poland, June 13-14, 2011, Revised Selected Papers.* Springer, 2012, pp. 293–306.
33. M. A. Helou, M. Palmonari, and M. Jarrar, “Effectiveness of automatic translations for cross-lingual ontology mapping,” *Journal of Artificial Intelligence Research*, Vol. 55, 2016, pp. 165–208.
34. J. McCrae, E. Montiel-Ponsoda, and P. Cimiano, “Integrating wordnet and wiktionary with lemon,” *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, 2012, pp. 25–34.

35. D. Kumar, A. Kumar, M. Singh, A. Patel, and S. Jain, “Modern wordnet: An affective extension of wordnet,” *New Trends in Computational Vision and Bio-inspired Computing: Selected works presented at the ICCVBIC 2018, Coimbatore, India*, 2020, pp. 527–536.
36. M. Piasecki, B. Broda, M. Marcińczuk, and S. Szpakowicz, “The wordnet weaver: Multi-criteria voting for semi-automatic extension of a wordnet,” in *Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009 Kelowna, Canada, May 25-27, 2009 Proceedings 22*. Springer, 2009, pp. 237–240.
37. A. F. Khan, J. P. McCrae, F. J. M. Gómez, R. C. González, and J. E. Díaz-Vera, “Some considerations in the construction of a historical language wordnet,” *Global Wordnet Conference 2023*, 2023, p. 101–105.
38. J. Boase-Beier, “Stylistics and translation,” *The Routledge handbook of stylistics*. Routledge, 2023, Vol. 9, pp. 420–435.
39. A. Chesterman, “Translation ethics,” *A history of modern translation knowledge*, 2018, pp. 443–448.
40. C.-R. Huang, *Ontology and the lexicon: A natural language processing perspective*. Cambridge University Press, 2010.
41. M. Griesel and S. Bosch, “Taking stock of the african wordnet project: 5 years of development,” in *Proceedings of the Seventh Global Wordnet Conference*, 2014, pp. 148–153.
42. K. Marszałek-Kowalewska, *Persian Computational Linguistics and NLP*. Walter de Gruyter GmbH & Co KG, 2023, Vol. 2.
43. T. Declerck and S. Olsen, “Linked open data compliant representation of the interlinking of nordic wordnets and sign language data,” in *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, 2023, pp. 62–69.
44. E. Rudnicka, M. Maziarz, M. Piasecki, and S. Szpakowicz, “A strategy of mapping polish wordnet onto princeton wordnet,” in *Proceedings of COLING 2012: Posters*, 2012, pp. 1039–1048.
45. W. Ng’ang’a, “Semantic analysis of kiswahili words using the self organizing map,” *Nordic Journal of African Studies*, Vol. 12, no. 3, 2003, pp. 407–425.
46. A. Hurskainen, “Sustainable language technology for african languages,” *The Routledge Handbook of African Linguistics*. Routledge, 2018, pp. 359–375.
47. J. Kim, Y. Hahm, S. Kwon, and K.-S. Choi, “Automatic wordnet mapping: from corenet to princeton wordnet,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 1–5.
48. J. P. McCrae, I. Wood, and A. Hicks, “The colloquial wordnet: Extending princeton wordnet with neologisms,” in *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*. Springer, 2017, pp. 194–202.
49. S.-J. Kang, I.-S. Kang, S.-J. Nam, and K.-S. Choi, “Mapping between corenet and sumo through wordnet,” *Journal of the Korean Institute of Intelligent Systems*, Vol. 21, no. 2, 2011, pp. 276–282.
50. R. Navigli, M. Bevilacqua, S. Conia, D. Montagnini, and F. Cecconi, “Ten years of babelnet: A survey.” in *IJCAI*, 2021, pp. 4559–4567.

51. C. Rzymski, T. Tresoldi, S. J. Greenhill, M.-S. Wu, N. E. Schweikhard, M. Koptjevskaja-Tamm, V. Gast, T. A. Bodt, A. Hantgan, G. A. Kaiping *et al.*, "The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies," *Scientific data*, Vol. 7, no. 1, 2020, p. 13.
52. D. U. Patton, W. R. Frey, K. A. McGregor, F.-T. Lee, K. McKeown, and E. Moss, "Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 337–342.
53. A. Senapati, "A semi-automated approach for bengali neologism," *SN Computer Science*, Vol. 4, no. 5, 2023, p. 428.
54. D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia tools and applications*, Vol. 82, no. 3, 2023, pp. 3713–3744.
55. I. Falk, D. Bernhard, and C. Gérard, "From non word to new word: Automatically identifying neologisms in french newspapers," in *LREC-The 9th edition of the Language Resources and Evaluation Conference*, 2014, pp. 2–9.
56. J. Breen, "Identification of neologisms in japanese by corpus analysis," *" eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*, 2010, p. 13.
57. J. Pyo, "Detection and replacement of neologisms for translation," Ph.D. dissertation, The Cooper Union for the Advancement of Science and Art, 2023.
58. V. Sheinman, C. Fellbaum, I. Julien, P. Schulam, and T. Tokunaga, "Large, huge or gigantic? identifying and encoding intensity relations among adjectives in wordnet," *Language resources and evaluation*, Vol. 47, 2013, pp. 797–816.
59. R. Saidi and F. Jarray, "Stacking of bert and cnn models for arabic word sense disambiguation," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.
60. B. K. Mishra and S. Jain, "An innovative method for hindi word sense disambiguation," *SN Computer Science*, Vol. 4, no. 6, 2023, p. 704.
61. S. Patwardhan and T. Pedersen, "Using wordnet-based context vectors to estimate the semantic relatedness of concepts," in *Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, 2006, pp. 1–8.
62. H. Haouassi, A. Bekhouche, H. Rahab, R. Mahdaoui, and O. Chouhal, "Discrete student psychology optimization algorithm for the word sense disambiguation problem," *Arabian Journal for Science and Engineering*, 2023, pp. 1–16.
63. B. Liétard, P. Denis, and M. Keller, "To word senses and beyond: Inducing concepts with contextualized language models," *arXiv preprint arXiv:2406.20054*, 2024.
64. X. Zhou, H. Huang, Z. Chi, M. Ren, and Y. Gao, "Rs-bert: Pre-training radical enhanced sense embedding for chinese word sense disambiguation," *Information Processing & Management*, Vol. 61, no. 4, 2024, p. 103740.
65. N. M. Fahad, S. Sakib, M. A. K. Raiaan, and M. S. H. Mukta, "Skinnet-8: An efficient cnn architecture for classifying skin cancer on an imbalanced dataset," in *2023 International conference on electrical, computer and communication engineering (ECCE).* IEEE, 2023, pp. 1–6.

66. X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, "A survey on model compression for large language models," *arXiv preprint arXiv:2308.07633*, 2023.
67. N. Ratner, Y. Levine, Y. Belinkov, O. Ram, I. Magar, O. Abend, E. Karpas, A. Shashua, K. Leyton-Brown, and Y. Shoham, "Parallel context windows for large language models," *arXiv preprint arXiv:2212.10947*, 2022.
68. H. K. Azad and A. Deepak, "A new approach for query expansion using wikipedia and wordnet," *Information sciences*, Vol. 492, 2019, pp. 147–163.
69. M. J. Hussain, H. Bai, S. H. Wasti, G. Huang, and Y. Jiang, "Evaluating semantic similarity and relatedness between concepts by combining taxonomic and non-taxonomic semantic features of wordnet and wikipedia," *Information Sciences*, Vol. 625, 2023, pp. 673–699.
70. M. Apidianaki, "From word types to tokens and back: A survey of approaches to word meaning representation and interpretation," *Computational Linguistics*, Vol. 49, no. 2, 2023, pp. 465–523.