Knowledge-based Word Sense Disambiguation for Setswana-English Machine Translation

DECLARATION

I, Tebatso Gorgina Moape hereby declare that this thesis is my work and has not been previously submitted in any form to any other university or institution of higher learning by other persons or by me. I further declare that all the sources of information used in this thesis have been acknowledged.

[redacted]

20 August 2024

TG Moape

Date

Approval for final submission

[redacted]

Supervisor

21/08/2024

Date

ACKNOWLEDGMENTS

Firstly, I would like to thank God for His divine light, wisdom, and unwavering guidance throughout my PhD journey. Without His blessings and grace, this achievement would not have been possible. I am forever grateful for the strength and perseverance He has bestowed upon me during this challenging yet rewarding journey.

I extend my deepest gratitude to my supervisor, Professor Sunday O. Ojo, for his invaluable support, guidance, and mentorship. Your expertise, insights, and constructive feedback have been instrumental in shaping this research and helping me grow as a scholar. Thank you for your patience and understanding and for always believing in me, even during the most challenging times.

I am also very thankful to my co-supervisor, Professor Oludayo O. Olugbara, for his constant encouragement, valuable suggestions, and thorough reviews of my work. Your scientific knowledge, enthusiasm for research and guidance have been crucial in refining my ideas and improving the quality of this thesis.

To my beloved mother, Patricia Thina Moape, words cannot express how grateful I am for your unconditional love, support, and the countless sacrifices you have made to help me reach this milestone. Your strength, resilience, and faith in me have been my guiding light throughout this journey. Thank you for always being there for me and for instilling in me the values of hard work, perseverance, and the pursuit of knowledge.

In loving memory of my late father, Moses Masalesa Moape, I dedicate this thesis to you. Although you are no longer with us, your love, wisdom, and the lessons you taught me continue to inspire me every day. I hope that this achievement makes you proud.

To my siblings and niece, Oratilwe Moape. Lizzy Moape, and Cattleya Moape, thank you for your constant support, understanding, and for being my pillars of strength. Your love, laughter, and encouragement have kept me going through the ups and downs of this thesis journey.

Finally, I would like to acknowledge the University of South Africa for their financial support, through its Academic Qualification Improvement Programme (AQIP) which made this research possible.

Thank you all for being a part of this incredible journey and for making this achievement possible. Above all, I am forever grateful to God for His divine providence and for guiding me every step of the way.

PUBLICATIONS FROM THIS THESIS

Published Journal Publications

Tebatso Moape, Oludayo O. Olugbara and Sunday Olusegun Ojo (2024). Integrating Lesk Algorithm and Cosine Semantic Similarity to Resolve Polysemy for Setswana Language. International Journal of Advanced Computer Science and Applications (IJACSA), April 2024.

Published Conference Publications

Tebatso Moape, Sunday Olusegun Ojo and Oludayo O. Olugbara (2024). Developing Bilingual English-Setswana Datasets for Space Domain. LREC-COLING, May 2024.

Presented Conference Publications

Tebatso Moape, Oludayo O. Olugbara and Sunday Olusegun Ojo (2024). A Meta-Analysis of Word Sense Disambiguation Methods. IEEE 15th International Conference on Mechanical and Intelligent Manufacturing Technologies (ICMIMT 2024), May 2024

Tebatso Moape, Sunday Olusegun Ojo and Oludayo O. Olugbara (2024). Constructing a Context-Driven Setswana-English Lexical Resource and Machine Translation Model. 2024 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems, August 2024.

Publications Currently in Writing

Tebatso Moape, Sunday Olusegun Ojo, Oludayo O. Olugbara (2024). Evaluating trends, patterns, and evolution of Word Sense Disambiguation methods using Bibliometric Analysis.

Tebatso Moape, Sunday Olusegun Ojo, Oludayo O. Olugbara (2024). Constructing a Comprehensive Culturally-Aware Setswana Universal Knowledge Core.

ABSTRACT

There are several challenges that hinder the development of Setswana-to-English machine-translation systems. A key obstacle is the absence of machine-readable knowledge resources. This has prompted the use of the only accessible data, which originates from the government domain. While training machine-translation systems using government-domain data can offer specialized language knowledge, such training introduces obstacles such as limited vocabulary, style variation, bias, and domain specificity. Furthermore, it is noted in the literature that the ongoing problem of polysemy in a machine-translation system reduces the overall accuracy. Polysemy is a linguistic phenomenon in which a single word or phrase has multiple senses, resulting in ambiguity. The task of resolving ambiguity in natural language processing (NLP) is known as word sense disambiguation (WSD). The concept of WSD serves as an intermediate task for enhancing text understanding in NLP applications, including machine translation, information retrieval, and text summarization. Its cardinal role is to enhance the effectiveness and efficiency of these applications by ensuring the accurate selection of the appropriate sense for polysemous words in diverse contexts. This study addresses these challenges by proposing three essential components: a diversity-aware machine-readable knowledge resource for Setswana-English, or the Setswana universal knowledge core (SUKC), a WSD approach to resolving lexical ambiguity; and a corresponding machine-translation model embedded with a WSD capability. Setswana-English data was collected from the existing paper-based bilingual dictionaries to achieve this purpose. Secondly, the study employed professional translators to translate space domain concepts from English to Setswana. The collected lexicon was integrated into the universal knowledge core (UKC). The Lesk algorithm which has seen various adaptations by researchers for different languages over the years was employed to address the inherent polysemy challenges. This study used a simplified, Lesk-based algorithm to resolve polysemy for Setswana; and used the bidirectional encoder representations from transformers (BERT) model for Setswana, and cosine similarity measure to embed Setswana glosses and measure semantic similarity, thus determining the accurate sense. The study employed a rule-based method embedded with the WSD algorithm for machine translation. The translation accuracy of the machine-readable dictionary was assessed by employing the developed machine-translation model; and evaluated using the BLEU score. The proposed model was tested on a combination of sentences containing both ambiguous words and those without ambiguity; and a higher BLEU score of 34.89 was achieved.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

HLT	Human Language Technologies
WSD	Word Sense Disambiguation
MT	Machine Translation
AWN	African WordNet
PWN	Princeton WordNet
ML	Machine Learning
NLP	Natural Language Processing
MRD	Machine Readable Dictionaries
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional Long Short-Term Memory
USR	Unified Sense Representation
Av-ACNN	Average asymmetric convolutional neural networks
CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
SVM	Support Vector Machines
MK-SVN	Multiple Kernel Support Vector Machines
DT	Decision Trees
NBC	Naïve Bayesian Classification
RF	Random Forest
KNN	K-nearest Neighbors
CD	Conceptual Density
RW	Random Walks

GCN	Graph Convolutional Network
UKC	Universal Knowledge Core
SUKC	Setswana Universal Knowledge Core
OOV	Out of Vocabulary

CHAPTER ONE: INTRODUCTION

1.1 Background and Motivation

1.1.1 Background

The linguistic landscape of South Africa is remarkably diverse, with a multitude of languages coexisting within multiple cultures. Despite this linguistic richness, there is a noticeable gap in the availability of advanced natural language processing tools, specifically word sense disambiguation models, for South African languages. This observation, as highlighted by the audit conducted by (Moors *et al.* 2018), proves the urgency and significance of conducting research to address this gap.

Most existing human language technologies (HLT) are biased towards human languages of developed economies, ten of which dominate over 80% of the internet content, which leads to a language digital divide (Duong 2017). The language digital divide remains a significant hurdle in the development and accessibility of technologies across diverse linguistic landscapes. This divide is particularly pronounced in regions with linguistic diversity, such as South Africa, in which official languages such as Setswana face a shortage of linguistic tools and lexical resources.

As a consequence, minority languages such as Setswana suffer from data-sparsity adversity. Data sparsity arises due to the limited availability of linguistic datasets, including both annotated and unannotated corpora, as well as lexical databases required to develop natural language processing (NLP) applications (Garcia-Martinez *et al.* 2020). This poses noteworthy obstacles in advancing effective HLT solutions for these minority languages, which in turn hinders their ability to establish a meaningful digital presence on the HLT landscape and on digital platforms.

In addition, languages such as Setswana are morphologically rich. The Setswana language employs disjunctive orthography characterized by intricate word formations, extensive use of prefixes and suffixes, and a system of noun class concords that plays an important role in conveying meaning and grammatical relationships within the language (Otlogetswe 2008). These language complexities cannot be accommodated by language models trained on datasets for major global languages such as English. This results in having to strike an appropriate balance between NLP tools robustness, and parsimony.

This research addresses the South African languages being under-resourced and the scarcity of linguistic tools, by specifically focusing on the development of a word sense disambiguation (WSD) model within the context of Setswana-English machine translation (MT).

WSD is a crucial component of NLP tasks and applications. The WSD is an open-research problem that is considered artificial-intelligence (AI)-hard (or AI-complete) (Mallery 1988). The solution requires a deep understanding of natural language and the application of common-sense reasoning. AI-complete problems are the most difficult to solve; and are the long-term goals of AI. WSD being AI-hard is due to the contextual nature of ambiguities in human languages and the computational complexity it attracts in NLP. WSD is a challenging task because it requires understanding the context in which a polysemous word is used; and the various senses in which the word can be used. WSD is defined as the process of computationally determining the appropriate sense of an ambiguous word in a specific context of usage. The complexity of WSD becomes more evident in NLP applications such as MT. Consider the following Setswana example:

"*Ba rile tuuu ba di 15, ka lepanta la go draya ga raro fo nokeng nxla*"

The Meta translation system incorrectly translates the Setswana (depicted on Figure 1.1) sentence as:

"They said tuuu the 15th, with a belt that draws three times on the river nxla"

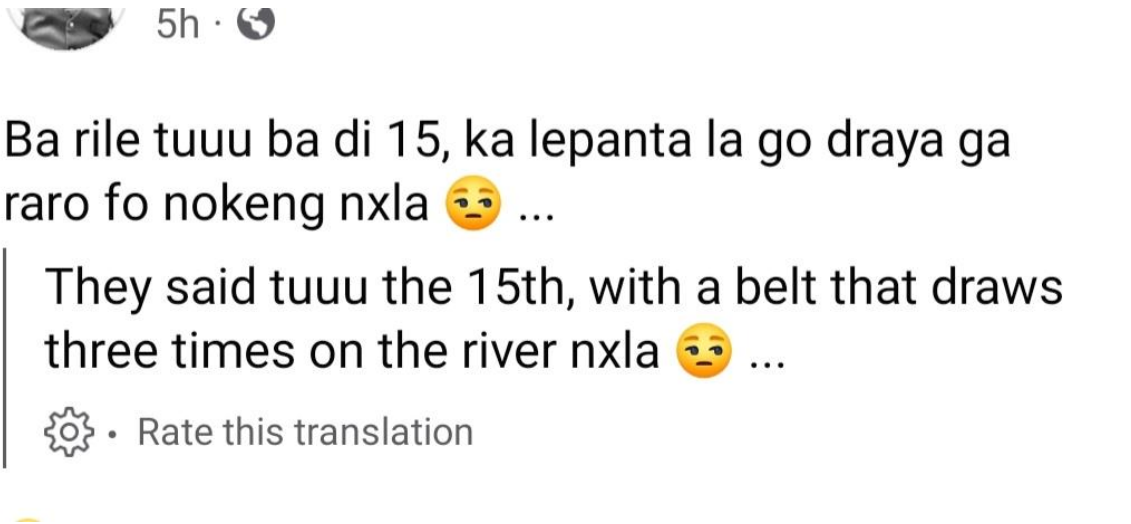


Figure 1.1 Meta's incorrect translation for Setswana polysemous word "noka" (screenshot taken on the 15th of August 2023 15:50)

The incorrect translation is due to the ambiguity of the word '***nokeng***' which can mean 'season', 'waist', or 'river', depending on the context of its use. The Meta translation system was unable to distinguish among these three senses. The inability to correctly disambiguate **"nokeng"** is a result of one of three challenges: the Meta database did not link the word to its multiple senses, the training corpora used did not represent the word in its multiple usage contexts, or data used to train the model was of a small scale. An important aspect to consider when addressing WSD is the need for comprehensive and contextually diverse training data to enable machine-learning models accurately to differentiate between polysemous words. This example highlights the necessity of WSD models for enhancing machine-translation (MT) system performance.

1.1.2 Motivation

The absence of WSD models for Setswana and other South African languages presents a critical gap in the application of NLP methodologies to diverse linguistic contexts. This absence is due to the lack of datasets to train WSD models and benchmark datasets to evaluate them. This absence presents a unique opportunity to contribute to the theoretical foundations of NLP. The development of WSD models for morphologically rich and less-resourced languages poses distinct challenges compared with less morphologically rich and more-resourced languages such as English. Furthermore, the morphological complexity of Setswana, characterized by intricate inflections and derivations presents a vital opportunity for investigating the difficulties and nuances of WSD in morphologically complex languages.

Addressing this situation contributes to the advancement of WSD methodologies, not just for Setswana, but as a reference for other less-resourced languages with similar concerns in Africa. The academic motivation for conducting this research lies in its potential to contribute theoretical understanding and methodological advancements in HLT and NLP applications. Theoretical understanding of the intricacies of how meaning is disambiguated in a language with such morphological richness can pave the way for more robust and accurate WSD models across various linguistic landscapes. The development of WSD models specifically designed for the linguistic characteristics of Setswana, sets a direction for addressing linguistic diversity in HLT and NLP research contributing to existing methodologies.

The practical motivation of WSD extends into NLP applications, including machine translation, text summarization, information retrievals, question-answering systems, inter alia (Navigli 2009). The inability to disambiguate word senses in natural languages impacts the

accuracy and performance of these NLP application systems. The practical motivation for this research is the improved performance of these systems by integrating WSD models in various systems and applications. This further aligns with the broader goal of leveraging technology to bridge linguistic digital divides, ensuring that the benefits of advanced NLP are equally accessible for all languages.

In addition, Setswana, being a low-resource language, faces gaps in the availability of comprehensive linguistic datasets and tools (Marivate *et al.* 2020). This research presents an opportunity to contribute to the development of necessary knowledge resources for Setswana, laying a foundation for future advancements in HLT and NLP. By offering a solution to low-resource languages, this research contributes to the enrichment of linguistic datasets and resources for Setswana.

This research not only addresses the theoretical, methodological, and immediate practical needs for WSD models, but also contributes to the development of knowledge resources for low-resourced languages such as Setswana, by adopting a framework that ensures that the linguistic and cultural specificities of Setswana are catered for, further contributing to a language resource called as Setswana Universal Knowledge Core (SUKC).

1.2 Problem Statement

Word sense ambiguity is a common feature of every human language. In natural language processing (NLP), language ambiguities are addressed through word sense disambiguation (WSD). The WSD problem can be formally defined as follows (Navigli (2009)):

"given a text T as a sequence of words ($w_1, w_2, \ldots, w_n$), WSD is the task of assigning appropriate sense(s) to all or some of the words in T, that is, to identify a mapping A from words to senses, such that $A(i) \subseteq SensesD(w_i)$, where $SensesD(w_i)$ is the set of senses encoded in a dictionary D for word w_i and $A(i)$ is the subset of the senses of w_i which are appropriate in the context T. The mapping A can assign more than one sense to each word $w_i \in T$; however, only the most appropriate sense is selected, that is, $| A(i) | = 1$."

The WSD problem becomes all the more challenging in NLP. The effectiveness of many WSD models depends on the availability and quality of a language corpus. Also, the more morphologically rich a language, the more computationally difficult it is to process, especially in the absence of a commensurately robust language model (Boruah 2022). Therefore, the

combination of a language being morphologically rich and low-resourced becomes problematic in the NLP of the language.

Numerous scholars have presented diverse WSD models across various languages, employing differing approaches and methods to solve the WSD problem. However, the benchmark evaluation datasets and language model used in the development of most of the existing WSD models are based on high-resourced and less-morphologically rich languages such as English (Kumar *et al.* 2019; Patel *et al.* 2021; Song *et al.* 2021). While the WSD models are high-performing for those languages, they are low-performing for more morphologically rich, low-resourced languages (Wiemerslage *et al.* 2022). Hence, they are limited in their direct applicability to NLP tasks and applications development for the latter. The need for WSD for South African languages has long been recognized, and the African WordNet posited as a viable tool for this (Madonsela et al. 2016). However, evidence from the literature reveals that hitherto, there have been no existing WSD models for SA languages in general, and for Setswana in particular (Moors *et al.* 2018). This gap in the literature has implication for the performance of NLP tasks and applications, such as machine translation systems, being developed for these languages.

1.3 Research Questions

In the absence of any existing reference WSD model for Setswana, and lack of direct applicability of existing language models, a research question arising is:

1.3.1 How can a Setswana WSD model that captures the linguistic properties of Setswana based on the adaptation of existing WSD language models be developed for Setswana-English MT?

Derived from this research question is another question, viz:

1.3.2 What is the nature and extent of semantic ambiguities in Setswana in the context of Setswana-English MT?

Given these Setswana ambiguity challenges, another question arising is:

1.3.3 Which existing language model can be adapted for Setswana WSD using knowledge-based approaches?

Given the data sparsity challenge in Setswana NLP, and the language corpus dependency of WSD models, another research question becomes:

1.3.4 How can a lexical resource knowledge base suitable for the development and evaluation of a Setswana WSD model be developed?

Given that WSD is commonly defined in the context of its NLP application (Navigli 2009), a further research question posed is:

1.3.5 How can this WSD-MT model be experimentally evaluated?

1.4 Research Objectives

Considering the research questions outlined above, this study aims to achieve the following research objectives:

1.4.1 Perform a systematic analysis of the nature and extent of semantic ambiguities in Setswana.

1.4.2 Identify and adapt an existing language model for Setswana WSD using a knowledge-based approach.

1.4.3 Develop a Setswana WSD model that captures the linguistic characteristics of Setswana through the adaptation of existing WSD language models for Setswana-English MT.

1.4.4 Develop a lexical resource knowledge base appropriate to the development and evaluation of a Setswana WSD model.

1.4.5 Experimentally evaluate the proposed WSD-MT model.

1.5 Thesis Statement

Based on the foregoing, the thesis statement this study sets out to validate is:

A knowledge-based WSD model for a morphologically rich low-resourced language such as Setswana, can enhance the performance of rule-based machine translation involving the language and a less morphologically rich high-resourced language, such as English.

1.6 Research Questions, Objectives, Methods, and Deliverables Alignment

The research questions, objectives, methods, and deliverables are summarized in Table 1.1

Table 1.1: Research questions, objectives, methods, and deliverables

Research Questions	Research Objectives	Research Methods	Research Deliverables
What is the nature and extent of semantic ambiguities in Setswana in the context of Setswana-English MT?	Perform a systematic analysis of the nature and extent of semantic ambiguities in Setswana.	Literature review and systematic analysis.	A taxonomy of Setswana language ambiguities.
How can a Setswana WSD model that captures the linguistic properties of Setswana based on the adaptation of some existing WSD language models be developed for Setswana-English MT?	Develop a Setswana WSD model that captures the linguistic characteristics of Setswana through the adaptation of existing WSD language models for Setswana-English MT.	Literature review comparative analysis adaptation and modelling.	A Setswana WSD model that effectively captures the linguistic characteristics of the language and demonstrates improved performance compared with the Setswana-English MT.
What existing language model can be adapted for Setswana WSD using knowledge-based approaches?	Identify and adapt an existing language model for Setswana WSD using a knowledge-based approach.	Literature review identification, adaptation, and modelling.	An adapted knowledge-based WSD model for Setswana.
How can a lexical resource knowledge base suitable for the development and evaluation of a	Develop a lexical resource knowledge base appropriate to the development and evaluation of a	Collection and curation mapping and integration.	A lexical-resource knowledge base for Setswana.

Setswana WSD model be developed?	Setswana WSD model.		
How can this WSD-MT model be experimentally evaluated?	Experimentally evaluate the proposed WSD-MT model.	Experimentation.	A Setswana WSD model integrated into MT.

1.7 Thesis Contribution

This research work contributes to knowledge in the field of NLP, in theory, methodology, and in practice, in the following respects:

Modelling:

1. Development ofPuoBERTa Enabled Embedding-based Lesk WSD Model

The primary contribution of this research is the proposal of a novel WSD algorithm specifically designed for the Setswana language. The algorithm is based on the simplified Lesk approach by Kilgarriff and Rosenzweig (2000) but incorporates several key modifications to address the unique linguistic characteristics of Setswana and to enhance disambiguation performance.

The proposed algorithm leverages sentence embeddings generated using the PuoBERTa language model, which is a Setswana-specific version of the BERT model. By employing PuoBERTa by (Marivate et al. 2023), the algorithm captures more effectively the contextual information and compositional nature of the Setswana language. This approach allows for a more accurate representation of word senses and their relationships within the context of Setswana sentences.

Furthermore, the algorithm utilizes the cosine similarity measure to determine the most appropriate sense for a polysemous word. The cosine similarity measure is used to calculate the semantic similarity between the context sentence embedding and the candidate gloss embeddings. By selecting the sense with the highest similarity score, the algorithm effectively disambiguates the intended meaning of the polysemous word in the given context.

The algorithm further incorporates the reduction in computational complexity feature. This feature addresses computational complexity in traditional Lesk algorithms by using a single-operation encoding process for both the context sentence and candidate glosses. This is

achieved by encoding the context sentence and glosses using PuoBERTa in a single step; the algorithm reduces the number of comparisons needed. This approach results in a linear growth rate of comparisons, as opposed to the exponential growth observed in traditional Lesk algorithms.

2. Development of a novel context-aware Setswana-English machine-translation model

Another contribution of this research is the development of a context-aware machine-translation model specifically designed for the Setswana-English language pair. The context-aware nature of the proposed model is a significant advancement, in that it enables the machine-translation system to consider the surrounding context when translating words, phrases, and sentences. The proposed model leverages contextual information to generate more accurate translations. 1200 This model addresses the unique challenges posed by the Setswana language and provides a framework for improving translation quality in resource-scarce settings.

3. A taxonomy of Setswana linguistic ambiguities

This study made a significant contribution to the understanding of ambiguities in the Setswana language by conducting an in-depth analysis of the various types of ambiguity present within the language. Through this study, a novel taxonomy of Setswana ambiguities was developed, which serves as one of the contributions of this thesis.

The developed taxonomy provides a structured and systematic classification of the various ambiguity types encountered in Setswana. By categorizing these ambiguities, the taxonomy provides a clear and concise framework that can be used by both linguists and researchers in the NLP domain. This contribution is particularly valuable for researchers in the development of language models and NLP applications specific to Setswana. The framework provides a comprehensive overview of the linguistic complexities and challenges that need to be addressed when processing and computationally analyzing Setswana text. Using the taxonomy, researchers can better understand the various ambiguity types and their characteristics, which will enable them to design and implement more effective and accurate NLP models and algorithms that cater to these specific ambiguity linguistic characteristics.

Resource:

4. Creation of a Setswana WSD evaluation dataset

A significant obstacle in the development and evaluation of WSD models for low-resource languages such as Setswana is the lack of benchmark datasets. To address this issue, this study developed a WSD evaluation dataset specifically for Setswana. The development of this evaluation dataset is a significant contribution to the research community, because it provides a standardized benchmark for assessing the performance of Setswana WSD models. The availability of this dataset enables researchers to compare and evaluate various WSD approaches, enabling further advancements in this field.

5. Creation of a bilingual Setswana-English lexicon with disambiguation contextual information

Another significant contribution of this research is the creation of a bilingual Setswana-English dictionary that is enriched with disambiguation contextual information. This dictionary serves as a valuable resource for the machine translation model and addresses the limitations of existing resources. In the study, Setswana-English data was collected from various paper-based bilingual dictionaries, including Oxford, Oxford Kiddies, Pharos, and Shuter's. This data was remodelled and combined with existing datasets. Incorporating data from multiple sources, the study developed a more diverse and comprehensive coverage of the Setswana-English language pair.

The collected remodelled lexicons were then transformed into a machine-readable format using the extensible markup language (XML) for facilitating the integration of the dictionary data into the machine-translation model, enabling efficient processing and retrieval of lexical information. A key aspect of the contributed dataset is the inclusion of disambiguation contextual information. The study modelled ambiguities within the dictionary design, associating each lemma with specific context features. This approach allows for the accurate disambiguation of polysemous words based on the surrounding context.

6. Creation of professionally translated bilingual Setswana-English space domain dataset

An additional resource contribution is a new English-Setswana bilingual dataset for the space domain. The dataset was constructed using the expansion method, which involves translating into Setswana a subset of English synsets from the Princeton WordNet (PWN). The translations were conducted by professional translators to ensure high-quality translations. This comprehensive coverage of space-related concepts in Setswana provides a valuable resource for researchers and developers working on NLP applications.

7. Setswana UKC

Another valuable resource is the Setswana Universal Knowledge Core (UKC), a diversity-aware, high-quality, human-centred multilingual dataset that encompasses lexicons from multiple languages interconnected through interlingual concepts. The developed datasets were integrated and incorporated into the UKC, where they can be accessed and queried online as a web service. Researchers and developers who are interested in utilizing this resource can download the XML version and employ it in various downstream NLP applications.

8. Meta-analysis and bibliometric-analysis of word sense disambiguation

WSD is a universal challenge across all spoken languages. This study conducted a meta-analysis to investigate the performance of various WSD approaches and then to measure the heterogeneity among existing studies. Additionally, a bibliometric analysis was conducted to examine the landscape of published research from an African perspective, with a particular focus on South Africa.

The findings of these analyses contribute significantly to the language technology landscape for African languages by demonstrating a substantial gap in the presence and representation of indigenous languages within the NLP domain. The meta-analysis results reveal that while WSD approaches have been extensively studied, there is a notable disparity in the application and evaluation of these methods for African languages. The bibliometric analysis further highlights the underrepresentation of African languages in NLP research, emphasizing the need for increased research and resources dedicated to the development of language technologies for these languages.

Researchers can use the analysis results from this study to identify areas of opportunity and to prioritize future research efforts. The findings highlight the necessity for a concerted effort to bridge the gap between well-resourced languages and underrepresented African languages in the NLP field.

1.8 Thesis Structure

This thesis is structured as follows. Chapter 1 presents the introduction and background, research aim, questions, and objectives. Chapter 2 provides an analysis of the nature and extent of semantic ambiguities in Setswana. Chapter 3 provides a literature review, meta-analysis, and bibliometric analysis of WSD's state-of-the-art approaches and methods. Chapters 4 and 5 respectively present the theoretical foundation and the methods and materials used for the

development of knowledge resources, WSD, and MT algorithms. The experimental evaluation and the discussion of the proposed models are presented in Chapter 6. Chapter 7 presents the summary of the thesis, concluding remarks, and future works. The chapters of this thesis are organized as follows:

Chapter One, Introduction: This chapter presents a background study of WSD as a stand-alone task, an intermittent task in MT; and outlines ML applications that use WSD. The chapter also presents a succinct description of the research motivation, problem statement, aim, questions, objectives, contribution, and thesis structure.

Chapter Two, Setswana: This chapter introduces the Setswana language origin, orthography, and the intricacies of its semantic and syntactic structure. The exploration extends to various types of Setswana words; and a detailed examination of morphologies and sentence structures employed within the language. Furthermore, the chapter presents a taxonomy of the diverse forms of ambiguity inherent in Setswana.

Chapter Three, Analysis of Related Literature: This chapter provides a detailed overview of WSD, its application to NLP, approaches, methods, and knowledge resources. The chapter further presents the latest related work in WSD and its application to MT. In addition, a meta-analysis and bibliometric analysis on WSD methods and research is presented. The preferred reporting items for systematic reviews, meta-analyses, and bibliometric analyses are employed in this chapter to combine the findings of various primary studies related to WSD.

Chapter Four, Theoretical Foundation: This chapter explains the key theoretical concepts and constructs that will be used in the study; and explains how the theoretical concepts will be operationalized and measured in the study.

Chapter Five, Methods and Materials: This chapter highlights the methods and materials used to achieve the outlined research objectives. A series of proposed algorithms is presented in this chapter. In addition, the chapter includes a presentation of resources used and the consolidation process conducted to construct the required knowledge resource.

Chapter Six, Experimentation, Evaluation, and Results: In this chapter, the extensive experiments conducted to validate the proposed algorithms are showcased. Experimental comparisons of the WSD algorithm were carried out using various encoders and combinations of different lexical information, with the results being presented.

Chapter Seven, Summary, Conclusion and Future Work: This chapter provides a summary of the achieved objectives. It concludes the thesis, and suggests future directions for related research studies.

CHAPTER TWO: THE SETSWANA LINGUISTIC CONTEXT

2.1 Introduction

This chapter presents linguistic context of the study for Setswana. The objective of this chapter is to provide a comprehensive overview of the Setswana language and its linguistic properties. The chapter is organized as follows: Section 2.2 introduces the Setswana language, its origin, and provides an overview of its diverse dialects; Section 2.3 presents the semantic and syntactic nature of the Setswana language, revealing a detailed examination of the orthography, morphology, and grammar; Section 2.4 explores the various forms of ambiguities that exist within the Setswana language, crucial for developing a robust and effective WSD model. The summary of the chapter, highlighting the key points and insights gained from the linguistic analysis of Setswana is presented in Section 2.5. This chapter reinforces the importance of the linguistic context in shaping the development and evaluation of the Setswana WSD model.

2.2 Setswana Language

Setswana, also known as Tswana, is genealogically classified as S.31 in Guthrie's nomenclature of Bantu languages. This places Setswana as a Southern Bantu language of Niger-Congo language family classification. It is spoken by approximately 8.2 million people, spreading across a number of Southern African countries. In Botswana, the ethnic Setswana language speakers make up 80% of the population, which was estimated at 2.3 million in 2022. In South Africa, Setswana is predominantly spoken in the Northwest Province and certain districts of the Free State and Gauteng province, with an estimated 4 million speakers. Setswana is also spoken in Namibia and Zimbabwe. It is therefore designated by The African Academy of Languages as a cross-border language.

Setswana belongs to the Sotho language sub-group; its closest relatives are Sesotho sa Leboa, also known as Northern Sotho and Southern Sotho. Figure 2.1 is the graphical representation of the south-eastern Bantu zone, indicating the Sotho group and Setswana in the green blocks.

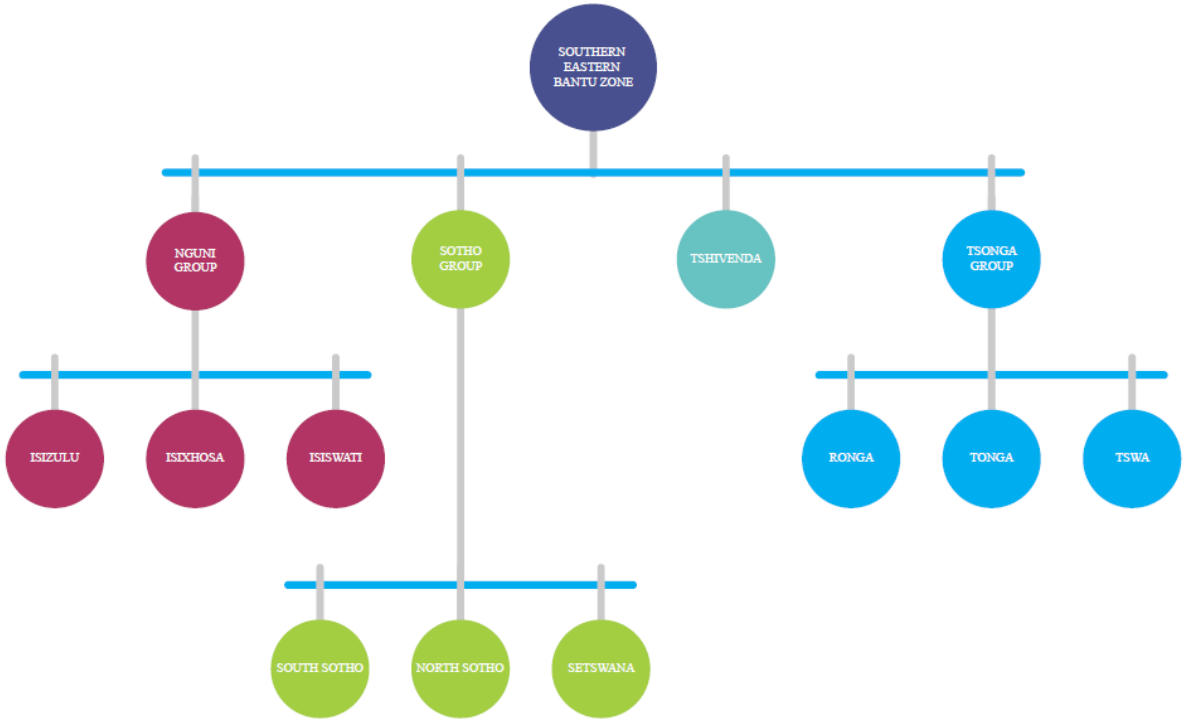


Figure 2.1: Graphical representation of the south-eastern Bantu zone (Otlogetswe 2008)

The three Sotho languages, collectively spoken by at least 16 million people, are closely related; from a linguistic perspective, they can be considered three varieties of a single language. In addition to the three Sotho languages, there are dialects found in every country that speaks Setswana. Figure 2.2 below illustrates Setswana dialects spoken in South Africa, Namibia, Zimbabwe, and Botswana.

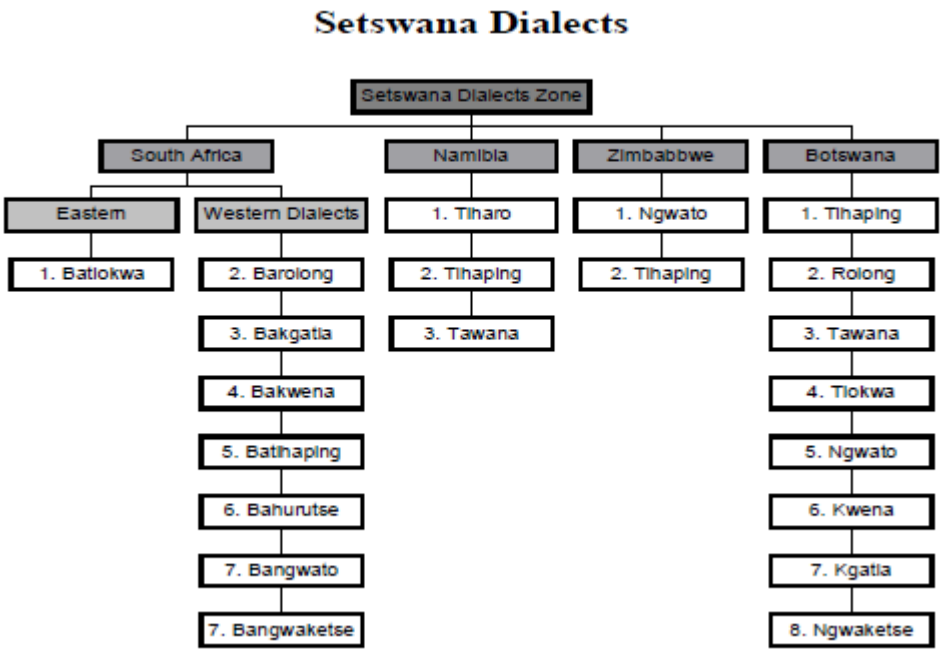


Figure 2.2: Setswana dialects (Otlogetswe 2008)

Figure 2.2 reflects the linguistic diversity and cultural richness of Setswana. By recognizing, valuing, and preserving dialects, inclusivity, cultural understanding, and the appreciation of local traditions are preserved.

There is evidence in the literature that an NLP tool developed for a language can be reused for another language in the same language family group, with changes only in the language corpus. For example, in their study, Dibitso, Owolawi and Ojo (2022), a Setswana part-of-speech tagger was found reusable for Sesotho sa Leboa, another language in the same Tswana-Sotho language family. This is significant in advancing NLP research for low-resourced South African languages.

2.3 Semantic and Syntactic Nature of Setswana

Linguistic typology is a branch of linguistics that studies and aims to categorize languages based on language structure. This includes phonological inventories, grammatical constructions, word order, etc. There are numerous language typologies; however, this section will only cover language typologies which aim to solve WSD in the context of machine translation providing a description of these typologies. These typologies include word order, orthographic and morphological typologies. In a morphologically rich language such as Setswana, word order is important, and the role of a word in a sentence may be indicated

through its morphological markings; therefore, the syntactic structure and the order of morphemes is vital to WSD. In terms of orthography, determining how grammatical relationships are expressed between words helps discern the syntactic and semantic structure of sentences, and assists in the WSD process. The significant morphological changes in Setswana lead to different word forms for the same root word; and understanding these morphological variations is crucial for accurate WSD.

2.3.1 Setswana words

Prior to exploring the characteristics of word structure for Setswana language, it is essential to establish a clear definition of what we mean by a 'word'. According to (Van Rooy and Pretorius 2003), a 'word' is a fundamental unit of language that carries meaning and can either stand alone or be combined with other words to form meaningful utterances. Words are typically a sequence of sounds or written symbols that represent a concept, object, action, or idea. Whether written or spoken, words serve as building blocks of sentences; and play a crucial role in communication. Words can be categorized into various parts of speech, such as nouns, verbs, adjectives, conjunctions, adverbs, interjections, etc., based on their grammatical function and syntactic behaviour within a particular language. For Bantu languages, there are open and closed word categories (Ciaccio, Kgolo and Clahsen 2020). Nouns and verbs are morphologically productive and regarded as open-word categories; while pronouns, particles, conjunctions, adverbs, and interjections constitute closed-word categories as they are morphologically unproductive. 'Morphologically productive' refers to word categories that exhibit a high degree of morphological productivity, allowing for new words or word forms by applying regular and productive morphological rules (Pretorius *et al.* 2009). These rules involve the addition of affixes, prefixes, suffixes, infixes, etc., changes in word stem or root, or other morphological processes to create new words or modify those already existing. 'Morphologically unproductive' refers to word classes that have limited or restricted morphological processes, resulting in a lower degree of productivity and less new word formation from root words. Every language has roots words. This research defines root words as morphemes that carry the basic lexical meaning or principal semantic load (Bennett, 2016). The following sections outline the various types of Setswana words. The characteristics of word structures and word categories in Setswana have several implications for solving WSD. The morphological productivity of certain word categories necessitates nuanced approaches to disambiguation, considering the various derivations of words. The distinctions between the structures and categories, along with the specific morphological processes involved, guide the

development of an effective WSD model tailored to the specific linguistic characteristics of Setswana.

2.3.1.1 Nouns

Setswana has twenty noun classes. Each noun belongs to a specific noun class. Nouns belong to a specific class based on inherent characteristics; and these noun classes determine the agreement patterns with other elements in the sentence (Letsholo and Matlhaku 2014). The noun classes are divided into non-infinitive and infinitive noun classes. Classes 1 to 14 are categorized as non-infinitive; class 15 is categorized as infinitive; while classes 16-20 are locative class nouns. The non-infinitive noun classes' morphological structure consists of a class prefix and a root. The words below provide examples of Setswana non-infinitive words:

molala (neck) -> melala (necks)

molao (law) -> melao (laws)

The noun **molala** (neck) contains a class prefix **mo-** and a root **-lala** and the noun **molao** (law) contains a class prefix **mo-** and a root -**lao**. In the group of these classes, the odd-class numbers are the singular nouns, and the corresponding even-class numbers are plural nouns. In the examples above, the class prefix **mo-** of class 3 denotes the singular, and the class prefix **me-** of class 4 is the plural.

The infinitive noun class 15 is "**go**" which represents the infinitive form of verbs, and is used to express actions or processes, such as:

go tsamaya -> to go

go ja -> to eat

Locative class nouns indicate the location or place where an action or object is situated. These nouns have their own unique prefixes and concords and are used in locative expressions. Locative class nouns are derived from other noun classes, and can vary, depending on the noun they are derived from. Examples include:

mo -> in/at/on

ko -> to/into

fa -> from

wa -> for/of

2.3.1.2 Pronouns

There are various types of pronouns for all the noun classes, namely, absolute, demonstrative, quantitative, and possessive pronouns (Demuth 1989). Pronouns in Setswana agree with noun classes and serve the purpose of modifying or substituting for nouns. Pronouns are used when referring to a noun that has previously been mentioned, allowing for more concise and efficient communication (Berg 2018). An absolute pronoun is used to emphasize or modify a noun. It is also used to replace a noun that has previously been mentioned. Examples of absolute pronouns are:

ke -> me

o -> he/she/him/her

re -> we/us

le -> you

ba -> they/them

A Setswana sentence with an absolute pronoun is as follows:

Batho bana -> specific people

bana emphasizes the noun and has the meaning of specific or particular. In the sentence above the absolute pronoun **bana** emphasizes the preceding Class 2 noun.

A demonstrative pronoun is a type of pronoun used to point out or identify specific nouns or objects in a sentence. It provides information about the location or proximity of the noun in relation to the speaker or the listener. Setswana pronouns indicate one of three possible distances, namely:

Distal position -> “e” or “ele” -> “this” or “these”

Proximal position -> “ona” or “ola” -> “that” or “those”

Post-distal position -> “kwa” -> “over”

The demonstrative pronouns are an essential part of Setswana grammar; and play a crucial role in indicating the location or proximity of objects in relation to the speaker or listener (Van Rooy and Pretorius 2003). Demonstrative pronouns agree with the noun they refer to in terms of class and number.

Setswana distinguishes between inclusive, exclusive, selective, separative, and interrogative pronouns in terms of quantitative usage. Inclusive pronouns are used when referring to a group that includes the speaker and the listener(s) together; while exclusive pronouns are used when referring to a group that includes only the speaker and others, excluding the listener(s). The separative pronouns, also known as disjunctive pronouns, are used to emphasize or separate the subject or object from the verb; and interrogative pronouns are used to ask questions and replace nouns in question sentences. Setswana quantitative pronouns express quantities or numerical values. They provide information about the quantity or number of elements in a sentence.

In the sentence below:

Lona le rata dijo -> You all love food

The pronoun **"lona"** addresses a group that includes both the speaker and the listener(s). It is equivalent to the English "you(plural)".

In the sentence:

Bona ba rata dijo -> They all love food

the pronoun **"bona"** refers to a group that includes the speaker and others, excluding the listener(s). It is equivalent to the English pronouns "they" or "them".

In the sentence:

Nna ke rate dijo -> I love food

the separative pronoun **"nna"** is used for the first-person singular; and translates to "I" or "me" in English. An example of a sentence with an interrogative pronoun is:

Ke dijo tse feng? -> Which food?

The pronoun "feng" asks an interrogative question, while the sentence:

Ke diapolo tse kae? -> How many apples?

quantifies the number of apples, using the quantitative pronoun "**kae**".

In Setswana, possessive pronouns are used to indicate ownership or possession. Possessive pronouns replace nouns, to show that something belongs to someone, and agree with the noun

they replace in terms of class and number. The five different types of Setswana possessive pronouns are:

ya ma/ya ka -> mine/my

ya gago -> your/yours

gagwe -> his/hers

ya rona -> our/ours

tsa bona -> their/theirs

2.3.1.3 Verbs

According to the lexical framework of Setswana, verbs are grouped into three categories, namely, main (principal), auxiliary, and copulative verbs (Creissels 1996). Setswana is an agglutinative and morphology-rich language. The main verb is the verb with the most complex word category. The core of the main verb is the root. Each morpheme occupies a specific slot; and the root is a bound morpheme that carries the basic lexical meaning of the word. The root requires one or more affixes to modify its meaning in order to form a complete word (Creissels 1996). The main verbs include the following elements:

Prefix -> indicates the subject and tense of the verb -> “ke” -> “i”

Tense -> indicates present, past, future -> “o bone” -> “he/she saw”

-> “o bona” -> “he/she sees”

Stem changes -> stem changes in certain tenses or conjugations -> “bone/bona” -> “saw/see”

Negation -> verbs can be negated by adding the prefix -> “watla” (coming) negated to “ga satla” (not coming).

Setswana main verbs undergo conjugation to match the subject and tense; and may undergo stem changes or negation to convey precise meaning (Pretorius *et al.* 2010). Setswana main verbs play a vital role in constructing meaningful sentences and expressing actions, states, and events. The auxiliary verbs work in conjunction with the main verbs to convey tense, aspect, mood, or negation (Pretorius *et al.* 2009). These verbs are used to link the subject of a sentence to additional information, or to express certain grammatical features. The auxiliary verb “**ntse”** (have been) is the past tense of the copulative verb **“nna”.** The main verbs include the following elements:

aspect and tense -> indicate the tense and aspect of a sentence -> “o tla bona” -> “he/she will see”

negation -> forms negative sentences -> “ga sa bona” -> “he/she can no longer see”

verb phrases -> verb phrases are formed by combining the auxiliary verb with the main verb -> “o tla go bona” - > “he/she will see you”

An auxiliary verb is dependent on a complement that can be a main verb, one of the copulative verb phrases, or another auxiliary verb phrase. These copulative verbs are used to identify, describe, and associate lexemes. The identifying copulative verb forms are:

ke, **-se**, **-le**, **-nna**, **-nne** or **-nnile**

The descriptive copulative verb forms are the same as the identifying copulative verb forms.

2.3.1.4 Adverbs

Adverbs provide additional information about the manner, time, place, frequency, degree, or other aspects of an action or state. Setswana adverbs are an important part of speech that modify verbs, adjectives, or other adverbs (Okgetheng, Malema and Tebalo 2022). In addition, Setswana adverbs can be either primitive or derived. Examples of Setswana adverbs are given below:

Primitive adverbs-

jaanong – now

Re fa batho dijo **jaanong** -> we are serving people food **now**

Derived adverbs -

Derived adverbs are formed from nouns, noun roots or verbs by means of an adverbial prefix **ga-.**

Gantsi -> many times

Ba itlhagisitse **gantsi** -> they appeared **many times**

gantsi (often, many times) is derived from the noun root **-ntsi**

Place adverbs –

Kwa ntle – outside

Bana ba tshemekela **kwa ntle.** -> The children are playing **outside.**

Gotlhe/tsotlhe – everywhere

Menang e **gotlhe** ka selemo. -> The mosquitoes are everywhere in summer.

Mo – in/at

Ke e tlgotse **mo** tafoleng. - > I left it **at** the table.

Degree adverbs-

Thata – very/too much

Go tonya **thata** gompieno. - > It is **very** cold today.

Time Adverbs-

Ka – at

Dikgang di tshameka **ka** ura ya boraro. - > The news comes on **at** 03:00.

Morago – after

Go tla tsamaiwa **morago** ga dijo tsa bosigo. -> We will leave **after** dinner.

Pele- before

Tlhapang matsogo **pele** le ja -> Wash your hands **before** you eat.

Setswana adverbs can be placed before or after the verb they modify; and they can also occur at the beginning or end of a sentence. Adverbs add specificity, detail, and precision to the meaning of a sentence, helping to convey the desired message effectively.

2.3.1.5 Conjunctions

A conjunction is used to link words in a sentence. These words are used in expressing relationships, coordinating ideas, indicating cause and effect, and showing contrast or comparison. Examples of Setswana conjunctions are:

Coordinating conjunctions-

le – and

ausi **le** abuti -> sister **and** brother

kgotsa – or

ausi **kgotsa** abuti -> sister **or** brother

jaaka - like, as such

rre **jaaka** mosimane – like father **like** son

mme - and, but

Tsatsi le tswile, **mme** gwa tonya -> The sun is out, **but** it's cold.

Subordinating conjunctions-

gore – that

O robetse ka pele **gore** a kgone go tsoga mo mosong. -> He/she went to sleep early so **that** he/she could wake up early in the morning.

Fa – if

Re tlilo go tsa **fa** ba sa re amogele sentle. -> We will leave **if** they don't welcome us properly.

Ka gonne – because

Re ne tsamaya, ka **gonne** ga ba re amogele sentle. -> We left **because** they did not welcome us properly.

Jalo – thus

O ja thata, ka **jalo** o nonne. -> He/she gained eats a lot, **therefore** he/she gained weight.

Comparative conjunctions-

kanne – than

Mariga a ngwaga o, a tonya **kanne** a ngwaga o fitileng. -> This year's winter is colder **than** last year's.

ka – like

Re fitletse go tswetwe jaaka **ka** maabane. -> They were closed when we arrived yesterday.

Temporal conjunctions-

pele – before

A re apeyeng **pele** baeti ba goroha. -> Let's cook **before** the guests arrive.

fa – when, while

Re tla dira tee, **fa** ba santse ba le mo motlhakanong. -> We will make tea **while** they are in a meeting.

Setswana conjunctions connect ideas, establish relationships, and convey various types of information.

2.3.1.6 Interjections

An interjection is an interposing remark that is used to express reactions and emotions such as joy, approval, disapproval, sorrow, distress, fear, disgust, agreement, disagreement, surprise, disbelief; and to draw the attention of, or call, or urge on, animals (Van Rooy and Pretorius 2003). Setswana interjection examples are as follows:

“bathong” - expresses surprise, disbelief, or shock

Bathong Buyi o tlhokofetse? -> Is it true that Buyi has died?

“Hee!” - expresses pain, distress, or sadness

Letsogo la me le botlhoko, **Hee!** -> My hand is painful!

“Yo!” - expresses excitement or enthusiasm

Yo! ke kereile madi a mantis. -> I received a great deal of money.

“Aowa!” - expresses disagreement or refusal

Aowa! Ga ke battle -> I don’t want to.

“Hao!” - expresses encouragement or agreement.

Hao! o tla kgona go pasa jaaka ngwaga o fitileng. - > You will be able to pass as you did last year.

2.3.1.7 Particles

In Setswana, prepositions are not explicitly distinguished. However, the language utilizes particles to convey the various meanings typically associated with prepositions (Pretorius *et al.* 2008). These particles can either exhibit class agreement or not, depending on the specific case. The possessive and qualificative particles are bound to specific noun classes; while the instrumental, temporal, locative, associative, and comparative particles are not bound to any particular class (Pretorius *et al.* 2008). Examples of these:

possessive particle **“tsa”** indicates possession

Ditokelo **tsa** rona -> our rights

The qualificative particle “**ba**” indicates qualification or characteristic:

batho **ba** bantsi -> many people

The instrumental particle **“ka”** indicates the instrument or means used:

O tsamaile **ka** kololi. -> He/she travelled by car.

The temporal particle **“e” , “ka”** indicates time or duration:

O ngwadile **ka** nako **e** tshesane. -> He/she wrote within a short period of time.

The locative particle **“mo”** indicates a location:

Dijo di **mo** godimo ga tafolo. -> The food is on the table.

The associative particle **“le”** indicates association or connection:

Ke dira **le** bana. -> I work with children.

The comparative particle **“jaaka”** indicates comparison or degree:

O ja thata **jaaka** rragwe. -> He/she eats a lot, just like his/her father.

Setswana particles serve a crucial role in sentence construction; and convey specific relationships between words and phrases. Each particle serves a specific function and contributes to the overall meaning and structure of a sentence.

2.3.1.8 Ideophone

The ideophone in Setswana functions as an exclamation or as a complement. It is a highly expressive word that is often used for sound symbolism regarding colour, sound, smell, taste, and feeling (Van Rooy and Pretorius 2003). Examples of an ideophone:

“tu” -> enhances the meaning -> “ba didimala” (quite)

The sentence below contains the above ideophone:

Ba didimala **tu.**

The **tu** symbolizes the expression of quietness.

Different word categories present different disambiguation characteristics. Setswana nouns and verbs exhibit ambiguity, in that single nouns and verbs have multiple meanings. Pronouns introduce ambiguity when referring to entities in the text; particles have context-dependent meanings. In the case of conjunctions, the interpretations that connect words or phrases vary, based on the relationship they establish with other words. Additionally, meanings of adverbs differ, based on the type of word being modified, similar to interjections. An expression of an emotion written in the same way can be interpreted differently, based on the context in which it is used.

The diverse disambiguation characteristics of Setswana words highlight the complexity of the WSD problem. Effective disambiguation requires consideration of syntactic structures, contextual relationships, and the copious ambiguities introduced by various word categories. Contextual cues and syntactic relationships for nouns and verbs must be leveraged to accurately identify the intended sense of a word. In addition, the WSD model must be able to identify the correct antecedent of pronouns, the syntactic role of adverbs and their intended modifications, the linguistic elements of conjunctions and interjections, as well as the relationship between particles and adjacent words that influence meaning of words in a specific context. To this end, it is important to leverage linguistic knowledge, thus developing a context-aware WSD model presenting accurate interpretations of word senses.

2.3.2 Setswana orthography

Orthography refers to the standardized system of writing of a language (Pretorius *et al.* 2009). Setswana is based on a disjunctive orthography; and is an agglutinative language with rich morphology written based on the Latin alphabet, additional diacritical marks and digraphs, to represent specific sounds. The orthography consists of 25 letters:

a, b, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, š, t, u, v, w, x, y, z,

and includes the following letter combinations representing a single sound:

“ts” represents the sound /ts/ as in “tsamaya” (go)

“tl” represents the sound /tl/ as in “tlala” (full)

“tlh” represents the sound /tl/ as in “tlhoka” (need).

“ng” represents the sound /ng/ as in “mangwana” (auntie)

Because Setswana employs a disjunctive orthography, the prefixes of the verb are generally written disjunctively. This requires a distinction between orthographic and the linguistic word. An orthographic word is a unit that is separated by spaces from other units in the sentence; while a linguistic word denotes units that function as members of a word category; and have their own particular meaning (Pretorius *et al.* 2009). For example, the sentence below contains four orthographic words, but three linguistic words:

Batho ba ja nama. -> The people are eating meat.

"**Batho ba**" (The people) is the first linguistic word, "**ja**" (eat) is the second, and "**nama**" (meat) the third; while the sentence has four lexical units separated by a space.

The rules and structures that govern the formation and arrangement of words, phrases, and sentences in the Setswana language are based on the varied use of the different categories of words introduced in the previous section, with modifying prefixes and suffixes.

Setswana uses a system of noun classes that categorize nouns based on their inherent characteristics. Table 2.1 illustrates the various prefixes used for different noun classes.

Table 2.1: Noun Class Prefix

Class	Prefix	Example
1	mo-	motho (person)
2	ba-	batho (people)
3	mo-	motse (village)
4	me-	metse (villages)
5	le-	lefielo (broom)
6	ma-	mafielo (brooms)
7	se-	setlhare (tree)
8	di-	ditlhare (trees)
10	di-	dijo (food)
11	lo-	lonao (foot)
12	di-	dinao (feet)
13	bo-	bogobe (pap)
14	ma-	magobe (paps)

15	go-	go ja (to eat)
16	fa-	fa pelong (in the heart)
17	kwa-	kwa lebenkeleng (at the shop)
18	mo-	mo gare (inside)

Class 1 nouns can be either without a prefix in the case of proper nouns, or they can have the prefix mo-. On the other hand, Class 2 nouns can have either the prefix bo- for proper nouns or the prefix ba-. Classes 6 & 14 and Classes 8, 10, & 12 are grammatically identical classes. Certain nouns may not have distinct singular or plural forms. Class 13 nouns consist of abstract nouns, and exist only in the singular form. Additionally, certain nouns may have alternative singular or plural forms when the prefixes are phonetically similar. Class 11 (lo-) nouns can also adopt the prefix le-, in which case they follow the same grammatical patterns as Class 5 nouns.

Verbs, on the other hand, can be altered in several ways. There are several types of verb that can be distinguished, namely, the infinitive and the perfect forms, active and passive voice, simple, objective, participatory, and auxiliary verbs. A verb in its regular form can be preceded by the infinitive marker or prefix **go.** The infinitive form of a verb usually takes the suffix. Preceding it with the infinitive marker **go** forms the infinitive, for example:

go j**a** -> to eat

The perfect form of a verb always takes the suffix **–e**. It cannot be preceded by the infinitive marker **go**, but must instead take the same infinitive as the infinitive form, for example:

gopot**e** -> to mould

The active infinitive form and active perfect form of the verb can also be made passive. Active suffix **-a** can be passive **-wa,** e.g.

dira -> dirwa

The perfect infinitive can be:

dirile -> dirilwe

Setswana verbs are derivations of a simpler verb. For example, “go dirisa” can mean to cause or help someone to do; or it can simply mean to use. In another example, “go dirisana”, which

could mean to help each other to do something, more accurately translates into “to cooperate”. Table 2.2 illustrates various types of Setswana verb derivation extensions.

Table 2.2: Setswana Verb Derivation Extensions

Extension	**Stem**	**Role**
Simple	a na	establishes a relationship between a subject and an object element
Applicative	ela nela	introduces an indirect object, indicating the recipient of the action
Causative	eha	suggests that the subject assists the object in performing the action
Neural	aka	indicates the completion of the action on the subject or the possibility of completing such
Extensive	olola	implies that the action is performed frequently, energetically, or excessively
Reversal	ana	signifies either the repetition or undoing of the action
Reciprocal	i	used when the members of a plural collective subject perform the action
Reflective	ela	indicates that the subject is performing the action on him/herself

The simple verb extension establishes a relationship between a subject and an object element. The applicative verb extension introduces an indirect object, indicating the recipient of the action. The causative verb extension suggests that the subject assists the object in performing the action, or is the cause of the action being carried out, similar to the verb’s passive form. The neutral verb extension is highly variable in meaning, indicating the completion of the action on the subject, or the possibility of completing it. The neutral verb can also transform the verb into a state of being, resembling an adjective. The features of the neutral verb extension present challenges for WSD due to the semantic variability it introduces. Recognizing whether a verb with the neutral extension is conveying an action or a state of being is crucial for WSD. Contextual cues and syntactic relationships play a significant role in determining the intended sense of a particular word.

The extensive verb extension implies that the action is performed frequently, energetically, or excessively. The reversal verb extension signifies either the repetition or undoing of the action; while the reciprocal verb extension is used when the members of a plural collective subject perform the action on one another, excluding the passive voice. Lastly, the reflexive prefix, I- indicates that the subject is performing the action on themselves, and I- cannot be used with the passive voice.

The auxiliary verbs are helping verbs that modify a main verb (Berg 2018). There are three types of auxiliary verbs, namely: simple, objective, and participial. Simple auxiliary verbs are used to express tense, aspect, mood, or other grammatical features in verb constructions. Simple auxiliary verbs assist in forming various verb tenses, such as past, present, and future; and can indicate aspects such as continuous or habitual actions. For example:

Ke tla a tsamaya. -> I will go.

Ke ka tsamaya. -> I can go.

Objective verbs are used to introduce an indirect object in verb constructions, for example:

Ke kgona go tsamaya. -> I can go. I am able to go.

Ke tshwanetse go tsamaya. -> I must go.

Lastly, there is the participial verb that serves as an auxiliary verb in certain constructions, and conveys ongoing or continuous actions, for example:

Ga ke ise ke tsamaye. -> I have not yet gone.

Ga a ise a tsamaye. -> He has not yet gone.

The agglutinative nature, disjunctive orthography, and diacritical marks in Setswana contribute unique challenges to WSD. The various noun class prefixes have to be taken into consideration when performing WSD, because they directly influence the semantic interpretations. Contextual information, syntactic analysis, and semantic relations have to be incorporated into a WSD model to accurately disambiguate between the various senses associated with different words and their linguistic properties.

2.3.3 Setswana morphology

Languages have a wide variety of morphological processes available for forming words and languages, and can be categorized based on their morphological patterns. A morpheme is

defined as a meaningful part of a word expressed by the form which exists as an integral part of a word meaning (Ciaccio, Kgolo and Clahsen 2020). For example, the words below:

mosadi (woman) -> mo, sa, di

setlhare (tree) -> se, tlhare

The mo, sa, di, se, tlhare are morphemes with meanings that contribute to the meaning of the word of which they are components. Setswana words contain multiple units of information, morphemes, thus requiring word segmentation, because morphemes become the basic units of analysis, rather than words. As a result, morphological analysis of words may be highly ambiguous, and require word-morphological disambiguation, which may turn morphological segmentation into a non-trivial NLP task.

2.3.3.1 Noun morphology

According to Ciaccio, Kgolo and Clahsen (2020), the primary purpose of a morpheme is to enhance the efficiency of a root or stem word. In Setswana, this is evident in the occurrence of nominal morphemes within morphologically complex words and compounds, in which the noun serves as the framework for their appearance. The Setswana nouns may include morphemes such as grammatical morphemes, which can be prefixes or suffixes, roots, and stems. The grammatical morphemes express grammatical semantic values in word class categories. Grammatical morphemes are attached to roots and stem morphemes. In the grammatical morphemes below:

mo- is a grammatical morpheme of **mo**tsetsana (girl), indicating a female.

Grammatical morphemes are classified according to their position in a word. The morpheme can occur as a prefix at the beginning of a word or as a suffix at the end of a word. For example:

dikilana – di, kil, ana

setlharenyana – se, tlhare, nyana

Di and se are prefixes; ana, and nyana are suffixes of the roots kil and tlhare. Table 2.3 contains prefixes used in the noun morphology; and Table 2.3 illustrates the various types of suffixes for the noun morphology.

Table 2.3: Noun Morphology Prefix

Class	Singular/Plural	Noun Prefix	Variant Prefix	Examples
1	singular plural	mo ba	m ngw b	monna (man) ngawana (child) basadi (women)
1a	singular plural	- bo		ausi (sister) boausi (sisters)
2	singular plural	mo me	m ngw	mosadi (woman)singular? nngwedi (moon) melelo (fires)
3	singular plural	le ma	- m	lekwalo (letter) makwalo (letters)
4	singular plural	se di		selepe (axe) delepe(axes)
5	singular plural	ne di	n m din dim	tsela (road) mpa (stomach) ditsela (roads) dimpa (stomachs)
6	singular plural	lo di		lonao (foot) dinao (feet)
7	singular plural	bo ma	b	bojalwa (beer) majalwa (beers)
8		go		jo ja (to eat)
9		fa go mo	-	fatse (on the floor) go dimo (on top) mo dimo (at the top)

Table 2.4: Noun Morphology Suffix

Type	Suffix	Examples

deverbative	i o a	batsidi (parents) molao (policy) morogwa (sender)
augmentative	gadi	kgalagadi (dried up)
feminitive	gadi	magadi (lobola)
diminutive	ana nyana anyana	tselana(small road) mosadinyana (little woman) mosimanyana (little boy)
locative	ing nnye nyeng	letsatsing (day) bojanyeng (green roots) bojannyeng (at the grass)

Both prefix and suffix are added to the roots and stem words. According to (Pretorius and Berg 2005), a root word is the semantic core of a word; it does not include a grammatical morpheme; it has no word correlate; and it is dependent, as are the prefixes and suffixes. A stem, on the other hand, has a word correlate, and may include one or more grammatical morphemes. Table 2.4 illustrates the various types of Setswana stems.

Table 2.5: Noun Morphology Stems

Type	Grammatical Morphemes	Examples
Simple	non	mma (mother) dipodi (goats)
Complex	1 >	motse (village) banneng (men – locality)
Reduplicated	A partially or fully reduplicated version of a word	mosadisadi (many women) ditlharetlhare (many trees)
Compound	Combination of two different words or of the combination of a word and a word group	modulasetulo (chair person) leebarope (rock pigeon)

The simple stem words do not include any grammatical morphemes, while the complex stems include a root or a simple stem, with one or more grammatical morphemes. Reduplicated stem

words, on the other hand, can be either partially or fully reduplicated versions of a word. Lastly, the compound stems can either be formed as a result of the combination of two different words, or of the combination of a word and a word group.

2.3.3.2 Verb Morphology

The basic form of the verb in Setswana consists of:

an infinitive prefix + a root + a verb-final suffix

The following word:

go bona (to see) consists of:

infinitive prefix -> go

root -> bon-

verb-final suffix ->a

As with the noun morphology, prefixes and suffixes may change; however, the root always forms the lexical core of a word. According to (Pretorius *et al.* 2010), the root of a verb can be preceded by prefixes as reflected in Table 2.6, and suffixes as seen in Table 2.7.

Table 2.6: Verb Morphology Prefix

Morpheme	Prefix	Example
Subject agreement	le la	le a ja (to eat) la ja (to eat)
Object agreement	di	di dula (they stay)
The reflexive	i	ipona(sees oneself)
The aspectual	a	a apolola (to undress)
The temporal	tla	tla baruta (will teach)
The negative	ga sa se	ga ba je (they are not eating) ga re sa battle (we don't want to) o se je dilo tseo (do not eat those things)

Table 2.7: Verb Morphology Suffix

Morpheme	Suffix	Example
Verb-final	a e	ithuta (learned) ithute (learn)
The causative	is	rekisa(sell)
The applicative	el	balela (read)
The reciprocal	an	thusana (help)
The perfect	il	utlwile (heard)
The passive	w	romiwa (sent)

Setswana verbs consist of a verb stem, which carries the core meaning of the verb. The verb stem remains constant, while prefixes and suffixes are added to convey additional information.

2.3.3.3 Derivational, reduplication, and infixes morphology

Setswana employs derivational prefixes and suffixes to create new words or alter the meaning or category of existing words. For example, the prefix "**bo**-" can be added to a verb stem to form a noun; for example:

bomme -> women

bodumedi -> believers

Reduplication involves the repetition of a part of a word to express plurality, intensity, or other semantic nuances. For example:

Go tonya thata thata. -> It's extremely cold.

Infixes are employed and inserted within the stem of a word, to indicate grammatical features or changes; for example:

Dira, dirang, dirileng

The rules and patterns of word formation, including the use of prefixes, suffixes, and other morphological elements are crucial for constructing and understanding Setswana words, phrases, and sentences accurately. Setswana morphology is a complex and integral part of the language; and must be considered when developing a language model. Because Setswana nouns and verbs are morphologically productive, as shown in the examples above, this poses the challenge of variability. The high degree of variability in word forms makes it difficult to create comprehensive knowledge resources that cover all word form combinations. The word

forms often result in a large number of inflections and derivations, contributing to data sparsity (Pretorius *et al.* 2008). The accuracy of WSD models trained on limited data for such a language as Setswana is compromised, as the model struggles to cater to the wide range of morphological variations. Another challenge is with the pronouns, particles, conjunctions, adverbs, and interjections. These word categories are morphologically unproductive and provide fewer contextual clues for disambiguation. This lack of clear morphological information increases the reliance on other syntactic features or lexical semantics for accurate WSD. A question arising from this is: which morphological features should be explicitly encoded in a WSD model for accurate and effective disambiguation?

2.3.4 Setswana grammar

Setswana grammar embodies rules of the language governing the sounds, words, sentences, and other elements, as well as their combination and interpretation (Berg, 2018). There are different types of sentences in Setswana. These sentences can be categorized as either simple, complex, or compound sentences. A sentence is classified according to the division based on the composition or grammatical structure of the sentence, a specific clausal structure that can be an independent or dependent clause (Berg 2018). The independent clause is the main clause that does not depend on another clause; and the dependent clause is a subordinate clause that depends on an independent clause for its existence (Berg *et al.* 2013).

2.3.4.1 Setswana simple sentence

The Setswana simple sentence consists of a subject, and constituents such as objects, obliques, and adjuncts. The simple sentence has a single verbal element and can be categorized as an independent clause. The Setswana sentences that follow are examples of Setswana simple sentences:

Mosadi o apeya nama. -> A woman is cooking meat.

The sentence above has only one independent clause in which the main verb **o apeya** (she cooks) is included in the structure. In terms of analysis, there are two main structures that are often considered in Setswana simple sentences, namely, c-structure, and f-structure (Berg *et al.* 2013). The c-structure which stands for constituent structure, describes the constituent organization and word order. It represents the hierarchical constituent structure of a sentence

and focuses on the organization of constituents, such as noun phrases (NPs), verb phrases (VPs), and prepositional phrases (PPs) (Berg 2018). The structure is as follows:

$$S \rightarrow NP\ VP$$

The S represents the simple sentence; NP is the noun phrase, and VP is the verb phrase. The NP is the first continuant, and the VP is the second continuant.

Monna o sega ditlhare. -> A man is cutting trees.

The sentence above has **monna** as the first continuant NP, and **o sega** as the second continuant VP. The f-structure which stands for the functional structure, represents the functional relationships and grammatical features of the sentence. This structure focuses on the assignment of grammatical roles, such as subject, object, and other functional elements. In Setswana, the f-structure captures information about noun class agreement, tense, aspect, mood, and other grammatical features. In the sentence above, the Setswana word **monna** belongs to noun class 1, associated with singular human nouns. The tense of the simple sentence is the present tense, for example:

$$[S\ SUBJ = NP, OBJ = NP, TENSE = Present]$$

This specifies the subject (SUBJ) and object (OBJ) as noun phrases (NP); and indicates the tense as "present".

The c-structure and f-structure are crucial for understanding the syntactic and semantic aspects of Setswana sentences; and play a significant role in sense disambiguation, machine translation, and parsing. The c-structure provides a foundation for identifying the syntactic roles of constituents within a sentence. This syntactic information is important for the WSD model, helping to disambiguate word senses based on their grammatical relationships. The f-structure, with its inclusion of grammatical features, provides a valuable context for a WSD model to differentiate between various word senses based on Setswana-specific linguistic attributes.

2.3.4.2 Setswana complex sentence

A complex sentence in Setswana consists of one independent clause, and one or more dependent clauses (Letsholo-Tafila 2018). The dependent clauses rely on the independent clause for their meaning. The sentences can be formed using conjunctions such as **fela** (that), **fa** (when), or **kagonne** (because). The sentences below are examples of Setswana complex sentences:

Ke tshepa gore o tla tla gosasa. -> I believe that you will come tomorrow.

The independent clause is **Ke itumelela**, and the dependent clause is **gore o tla tla gosasa**. The dependent clause begins with the conjunction **gore**, which introduces the subordinate clause. Within the dependent clause, **o tla tla** is the verb phrase, which consists of the subject pronoun **o,** the future tense marker **tla**, and the verb **tla**. The symbolic representation of the sentence would be:

$$S \rightarrow NP\ VP$$

$$NP \rightarrow Pronoun: Ke$$

$$VP \rightarrow Verb: tshepa$$

$$Conj: gore$$

$$Clause: o\ tla\ tla\ gosasa$$

Complex sentences in Setswana can have multiple dependent clauses, each serving a specific purpose or providing additional information. The structure and arrangement of these clauses depend on the specific context, and the speaker's intended meaning.

2.3.4.3 Setswana compound sentence

Setswana compound sentences consist of two or more independent clauses that are connected, using coordinating conjunctions such as **le** (and), **ka** (or), **fa** (but), and others. Each independent clause in a compound sentence can stand alone as a complete sentence. The sentences below are examples of Setswana compound sentences:

O tsamaya le dijo tsa maitsibowa. -> She/he is leaving and taking his/her dinner with her/him.

The first independent clause is **o tsamaya,** and the second independent clause is **tsa maitsibowa**. The two clauses are connected by the coordinating conjunction **le**. The symbolic representation of the sentence would be:

$$S \rightarrow S1\ Conj\ S2\ S$$

$$1 \rightarrow NP\ VP$$

$$S2 \rightarrow NP\ VP$$

$$NP \rightarrow Pronoun: O$$

$$VP \rightarrow Verb: tsamaya$$

$$NP \rightarrow Pronoun: dijo$$

$$VP -> Verb: tsa$$

The coordinating conjunction **le** indicates an addition with “and” between the two independent clauses forming a complete sentence.

Disambiguation complexities vary between different sentence categories. Simple sentences do not involve subordinate clauses, conjunctions, and various morphological elements; and therefore they are less complex to disambiguate, compared with complex and compound sentences. The relationships between clauses and the roles of words within complex and compound sentences add an additional layer of complexity to the disambiguation process. These challenges stem from the intricate syntactic structures, diverse semantic relationships, and the roles played by various elements of words within sentences. In complex sentences, the presence of subordinate clauses requires identification of causal and conditional relationships. Compound sentences, with coordinating conjunctions introducing contrasting or coordinating ideas, require an analysis of the relationship between clauses. These complexities highlight the need for the WSD model to go beyond simple sentence structures, and consider the broader context, syntactic hierarchies, and semantic dependencies, to achieve accurate sense disambiguation in complex and compound sentences. The ability of the WSD model to handle these complexities is crucial for effective and contextually aware WSD.

2.4 Setswana Language Ambiguities

Linguistic ambiguity refers to the phenomenon in which words can have multiple meanings, leading to uncertainty for the reader in determining the intended meaning. Ambiguity can occur not only at the level of individual words but also within sentences. Figure 2.3 illustrates a taxonomy of linguistic ambiguities in Setswana.

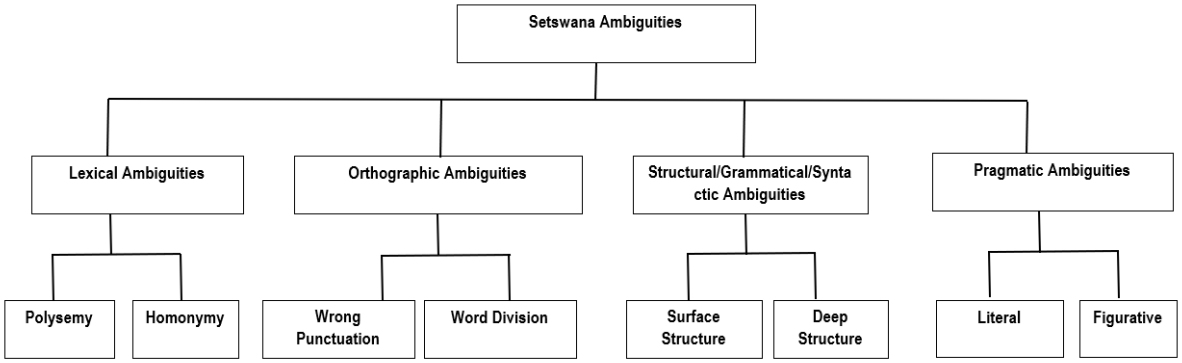


Figure 2.3: A taxonomy of Setswana language ambiguities

As shown in Figure 2.3, Setswana ambiguities can be categorized into lexical, orthographic, structural/grammatical/syntactic, and pragmatic ambiguities. Lexical ambiguity arises when a word is responsible for the ambiguity of a given sentence; that is, ambiguity which is due to lexical factors or lexemes (Rodd 2018). Lexical ambiguities can be classified into two categories, namely, polysemy and homonymy. Polysemy occurs when a word or phrase has multiple meanings. Consider the following sentence in Setswana:

- *Mme wa tshela.*

This sentence is polysemous due to the lexeme '*tshela*' and thus can be interpreted as:

- A mother is well.
- A mother is pouring.
- A mother is crossing.

Homonymy occurs when two or more words are either identical, pronounced or written the same, but they convey unrelated meaning. Consider the word:

- '*madi*'

The word '*madi*' can mean '*blood'* or '*money'.* These are two possible unrelated meanings that can be deduced from this word.

In other instances, ambiguity occurs as a result of orthography permitted by the rules of syntax in a sentence, rather than a specific ambiguous word. This ambiguity is referred to as orthographic ambiguity. Orthographic ambiguity is the type of ambiguity in which the written form of a word or phrase can be interpreted in multiple ways. This type of ambiguity can be

identified from two circumstances resulting in two types of orthographic ambiguity, namely, incorrect punctuation and word division in Setswana. Incorrect punctuation ambiguity results from careless punctuation. If one, for example, writes a paragraph or a sentence without punctuation, it will be impossible for the reader to comprehend what is meant. For example, the following sentence can be interpreted in multiple ways because of punctuation:

- *Mosadi o boletse go re a tsamaye.*
- The woman spoke and said that he/she should go.

Or

- Mosadi o boletse, gore a tsamaye.
- A woman spoke in order to leave.

Or

- Mosadi o boletse a re: “Tsamaya!”
- The woman spoke and said: “Go!”

The comma in the above pair of sentences does not only affect the clause, but also the meaning of the whole sentence.

As stated, Setswana is written disjunctively. This means that certain words are written as separate words or morphemes. Such can result in a sentence having multiple meanings due to word division. Consider the following sentence:

- Ditshaba di maketse.
- The nations are surprised.

and

- Di tshaba di maketse.
- They are running away whilst being surprised.

The sentence “Ditshaba di maketse” which means “the nations are surprised”, could also be taken to mean “they are running away whilst being surprised”, if one overlooks the division found in the word di and tshaba.

In addition to orthographic ambiguity, there is structural ambiguity which can be referred to as grammatical or syntactic ambiguity. Structural ambiguity refers to the type of ambiguity in which the structure of the sentence is solely responsible for multiple meanings, and not because of the presence of an ambiguous lexical item (Hindle and Rooth 1993). Surface structure ambiguity and deep structure ambiguity are two types of structural ambiguity. Surface structure ambiguity is ambiguity which is due to the structure of the sentence at the surface level; it corresponds with the overt, written form of sentences. Consider this sentence:

- Mapogo a mathamaga a tshwere mosimane wa go utswa maabane.

The above sentence can be interpreted as:

- The community guards arrested a boy who stole yesterday.
- The boy who stole was arrested yesterday by the community guards.

It is the positioning of the adverb maabane in the sentence which ambiguates the sentence. One wonders whether the arrest took place yesterday or whether the stealing occurred yesterday. Deep structure ambiguity involves neither a change in meaning of individual words or phrases, nor a change in the groupings of words or phrases, but a change in the logical relations between words and phrases. In this sentence:

- Reabaebetswe o timetse.

the question: Who is lost? can be asked: Is it Reabaebetswe, who does not know where he is, or is it the interlocutors who do not know Reabaebetswe's whereabouts? In the first interpretation, namely, that Reabaebetswe does not know where he is, Reabaebetswe is the agent; whereas in the second interpretation, namely, that some people do not know where Reabaebetswe is, Reabaebetswe is the patient. The intransitive verb o timetse, is associated with two senses in this context: firstly, that he cannot be seen, that is, we do not know his whereabouts; and secondly, that he himself does not know his whereabouts. The ambiguity here involves the subject-verb and the verb-object relations.

Lastly, in Setswana, ambiguity typology is pragmatic ambiguity. In this type of ambiguity, a phrase or a word may have a literal meaning, a dictionary-defined meaning, as well as a figurative meaning and an imaginative meaning, which involves the illogical comparison of unlike things (Ferrari *et al.* 2014). To this end, pragmatic ambiguity can be classified as either literal or figurative. Consider the sentence:

- O ile a mo loma tsebe.

The above sentence can be interpreted literally as:

- He/she bit his/her ear.

or

figuratively, as:

- He/she told him a secret.

This type of ambiguity can also be referred to as metaphorical ambiguity in natural languages. This study focuses on addressing the issue of lexical ambiguity within the Setswana language. The primary aim is to develop a model that can effectively disambiguate multiple meanings associated with lexical words in Setswana. By exploring various techniques and considering the linguistic characteristics of Setswana covered in this chapter, this research aims to contribute to the broader field of the HLT landscape, and enhance the accuracy of NLP applications for Setswana as indicated by (Moors *et al.* 2018), that, presently, there is no word sense disambiguator for Setswana. The absence of an existing word sense disambiguator for Setswana serves as a compelling motivation for this study. This research seeks to address this gap by developing a WSD model tailored to the linguistic specificities of Setswana in the context of MT.

2.5 Summary

This chapter provides a comprehensive overview of the semantic and syntactic nature of the Setswana language, focusing on its linguistic properties and their implications for developing a WSD model. The chapter begins by discussing the importance of linguistic typology in understanding the structure of Setswana, particularly its word order, orthography, and morphology. The chapter highlights the various word categories, such as nouns, pronouns, verbs, adverbs, conjunctions, interjections, particles, and ideophones, along with their specific characteristics and disambiguation challenges. The chapter further presents the Setswana orthography, explaining the disjunctive writing system, noun class prefixes, and verb derivations, which are essential for capturing the semantic nuances and grammatical relationships in Setswana. The section on Setswana morphology further emphasizes the agglutinative nature of the language and the importance of considering morphological variations, such as prefixes, suffixes, and infixes, when developing a WSD model. The chapter

also discusses Setswana grammar, including simple, complex, and compound sentences, and the disambiguation complexities associated with each type. This highlights the need for a WSD model to analyse syntactic structures, semantic dependencies, and the broader context, to achieve accurate sense disambiguation. Finally, the chapter presents a taxonomy of Setswana language ambiguities, categorizing them into lexical, orthographic, structural, and pragmatic ambiguities, scoping the type of ambiguity addressed in this study. Overall, this chapter emphasizes the importance of considering the linguistic properties of Setswana when developing a WSD model, such directly influencing the disambiguation process and the model's ability to accurately interpret word senses in various contexts.

CHAPTER THREE: STATE OF THE ART

3.1 Introduction

This chapter provides a detailed review of the body of literature related to this study. The primary objective of this chapter is to provide a comprehensive review of the research work that has been published in the literature using qualitative and quantitative analysis. The chapter is structured as follows: for qualitative analysis, the first section introduces and describes the task of Word WSD. Subsequently, Sections 3.2 to 3.7 provide an overview of various WSD approaches, knowledge resources, evaluation methods, and the application of WSD in NLP. The aim of this overview is to offer a comprehensive understanding of WSD and its significance in NLP research and applications. Section 3.8 presents a comprehensive review of the research conducted on WSD as a stand-alone task, and its application within MT. This section includes a diverse range of studies conducted in different contexts, employing various approaches, methodologies, techniques, and knowledge resources. The primary objective of this section is to provide a thorough synthesis of the studies that were conducted, emphasizing the key findings that have emerged from previous and current research across different languages. For quantitative analysis, Section 3.9 presents the bibliometric analysis, providing a comprehensive overview of the WSD research landscape, highlighting the key trends, influential works, and prominent researchers in the field. A meta-analysis is presented in Section 3.10 to provide a systematic synthesis of findings from multiple studies, offering a quantitative approach which summarizes and integrates results and research outcomes within the WSD research landscape. In Section 3.11, the WSD open-research problems are discussed, while Section 3.12 identifies gaps in the literature. Finally, Section 3.13 provides a summary of the chapter.

3.2 Word Sense Disambiguation

WSD is an historical task in NLP and AI which, in essence, dates back to Weaver (1949), who recognized the problem of ambiguous words in the context of machine translation. Even today, word ambiguity remains one of the most challenging and pervasive linguistic phenomena in NLP (Bevilacqua, 2021). As illustrated in the Princeton WordNet (https://wordnet.princeton.edu/), the root word for “disambiguation” is “ambiguity” which comes from the Latin word “ambiguitas” meaning “double meaning” or “uncertainty”. The prefix “dis” means “to undo” or “to remove” indicating the process of removing, undoing, or

resolving ambiguity or uncertainty. The word “sense” has several meanings that vary depending on the context in which it is used. According to the Princeton WordNet 3.1 (https://wordnet.princeton.edu/), the word sense has five different meanings as a noun and four different meanings as a verb. Figure 3.1 illustrates the meaning of the word sense from WordNet.

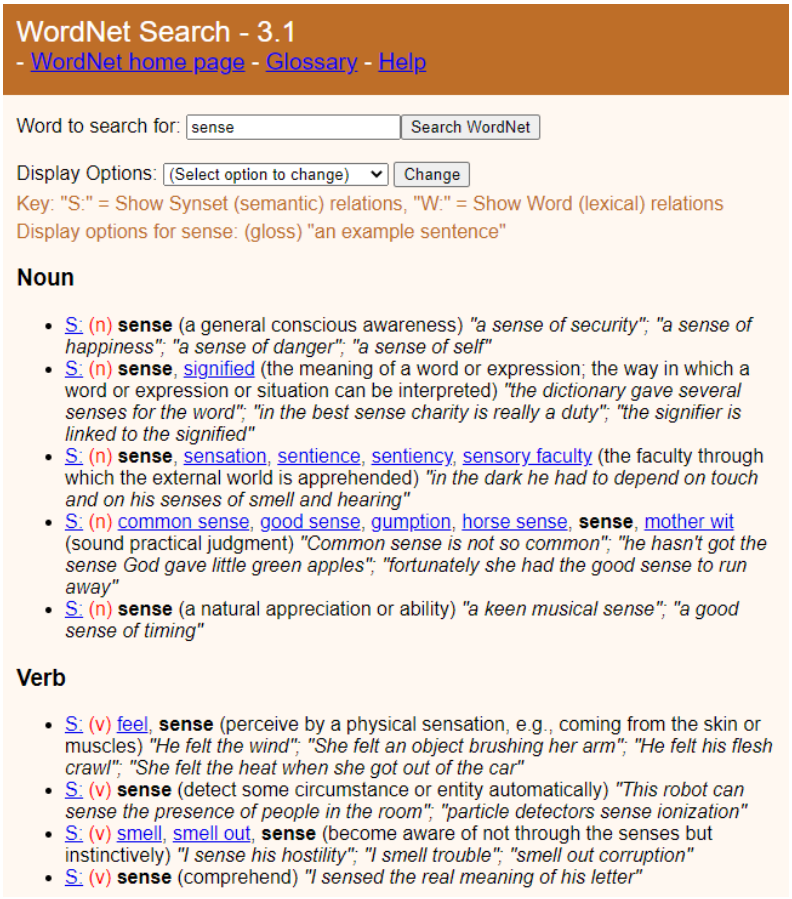


Figure 3.1: The “Sense” synsets from WordNet 3.1 (https://wordnet.princeton.edu/)

The appropriate meaning of sense in this context is the noun sense Number 2 in Figure 3.1 which is “the meaning of a word expression; the way in which a word or expression or situation can be interpreted”. It is not necessarily the case that other senses are incorrect; however, they do not hold relevance within the specific context of “word sense disambiguation”. While human beings unconsciously disambiguate senses, this is a complicated task for machines.

The automation of disambiguation of words with multiple senses has been of interest since the 1950s (Machinery 1950). In the domain of computational linguistics and NLP, WSD is an intermediate task for numerous applications such as machine translation (MT), text summarization, information retrieval, and chatbots systems. Figure 3.2 illustrates the intermediate location of WSD in an NLP pipeline on the SCROLL NLP framework. The name

stands for semantic cross-lingual label parser; and it consists of multilingual syntactic and language-independent semantic modular pipelines. The modularity of the SCROLL framework allows for the seamless integration of WSD alongside other essential NLP components.

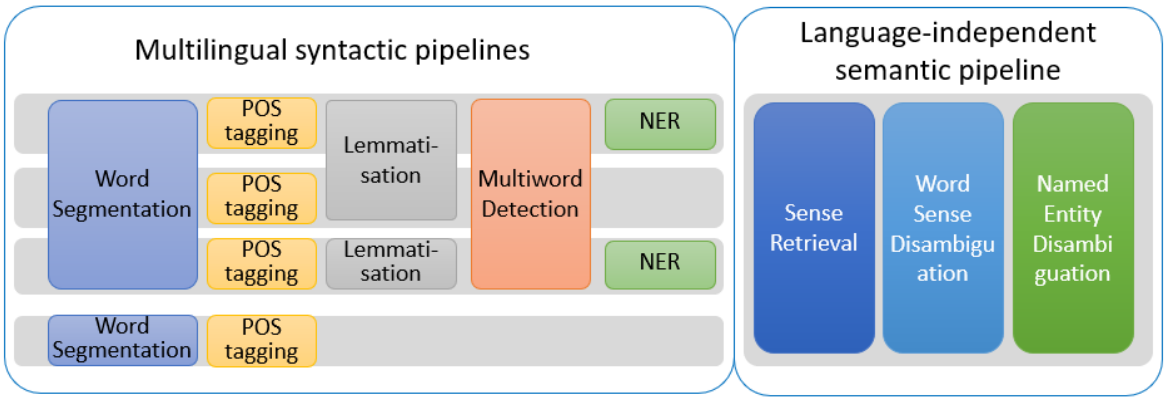


Figure 3.2: Intermediate location of WSD in an NLP pipeline (Gábor Bella, http://knowdive.disi.unitn.it/language-diversity/)

The SCROLL framework allows for the building of NLP pipelines in an extensible and modular way both from language-specific and from language-agnostic components corresponding to common NLP tasks such as tokenization, POS tagging, multiword detection, named entity recognition, WSD etc. As depicted in Figure 3.2, WSD is positioned within the semantic pipeline, focusing on the semantics of a language.

The earliest approach proposed for solving the problem of disambiguation dates to the late 1940s and early 1950s (Booth and Locke 1955). The question of resolving word senses was raised half a century ago; and to this day, this task is still being referred to as an "AI-complete problem", that is, a task whose solution is at least as problematic as the most difficult problems in artificial intelligence (Bevilacqua, 2021).

3.3 Word Sense Disambiguation in NLP

Natural languages are languages used by humans to communicate with one another. NLP can be described as the use of computers to process written and spoken languages for practical applications such as machine translation (MT), information retrieval, and voice-to-text applications. The goal of NLP systems is to facilitate the exchange of information between computers and humans in natural language enabling effective communication between the two. Human beings learn and adapt to environments intuitively, while computers are unable to do so. This is because learning and adaptation is a behavioural aspect that machines do not possess (Powers and Turk 2012). After efforts by the artificial intelligence community to integrate

learning into machines, the field of machine learning (ML) has emerged to educate computer algorithms on assimilating knowledge from data and executing tasks similar to those accomplished by humans. Toward this end, ML offers invaluable input to WSD systems.

The most common challenge encountered by NLP systems is human language ambiguity. Natural languages have words that are ambiguous in the sense that they might have:

a) the same spelling but a different meaning
   e.g., “pelo” meaning either “heart” or “button”

This type of ambiguity is referred to as homonyms.

b) the same spelling but different pronunciation

   e.g., “tshela”, meaning “pour”, “cross”, or “live healthily”

This type of ambiguity is referred to as heteronyms.

c) the same spelling and pronunciation (but not always), but different meaning

   e.g., “noka” meaning either “river”, “waist”, or “season”

This type of ambiguity is referred to as homographs.

Consider the case of communicating with a computer which runs an NLP system. Assuming that the system uses word matching and techniques for interpreting natural language, a Setswana input similar to “*mme o tshela letswai mo nameng*, which means “a mother is pouring salt on the meat”, would probably produce an undesirable output in that finding and analysing commonality among the words is completely different from treating their respective contexts in similar senses. In natural language processing (NLP), WSD is a process of identifying the correct sense of an ambiguous word within a specific context (Kilgarriff 1997). These ambiguous words are sometimes referred to as polysemous words, words that have multiple meanings or senses. In essence, words can be used to refer to various concepts or ideas, depending on the context in which they are used. In formal terms, WSD can be expressed mathematically as a function that takes a word with multiple possible senses and returns the correct sense of that word, depending on the given context:

$$D: W \times C \rightarrow S \qquad (3.1)$$

Let W be a word with multiple possible senses, and *S = {s1, s2, ..., sn}* the set of possible senses of *W*; with *C* being a context in which *W* appears. Then, we define WSD as a function *D* that

maps W and C to the correct sense $s \in S$, as defined in Eq 3.1. The overall framework or structure of a WSD system encompasses various components and processes in disambiguating the intended sense of a word within a given context. Figure 3.3 illustrates the WSD framework with the main components.

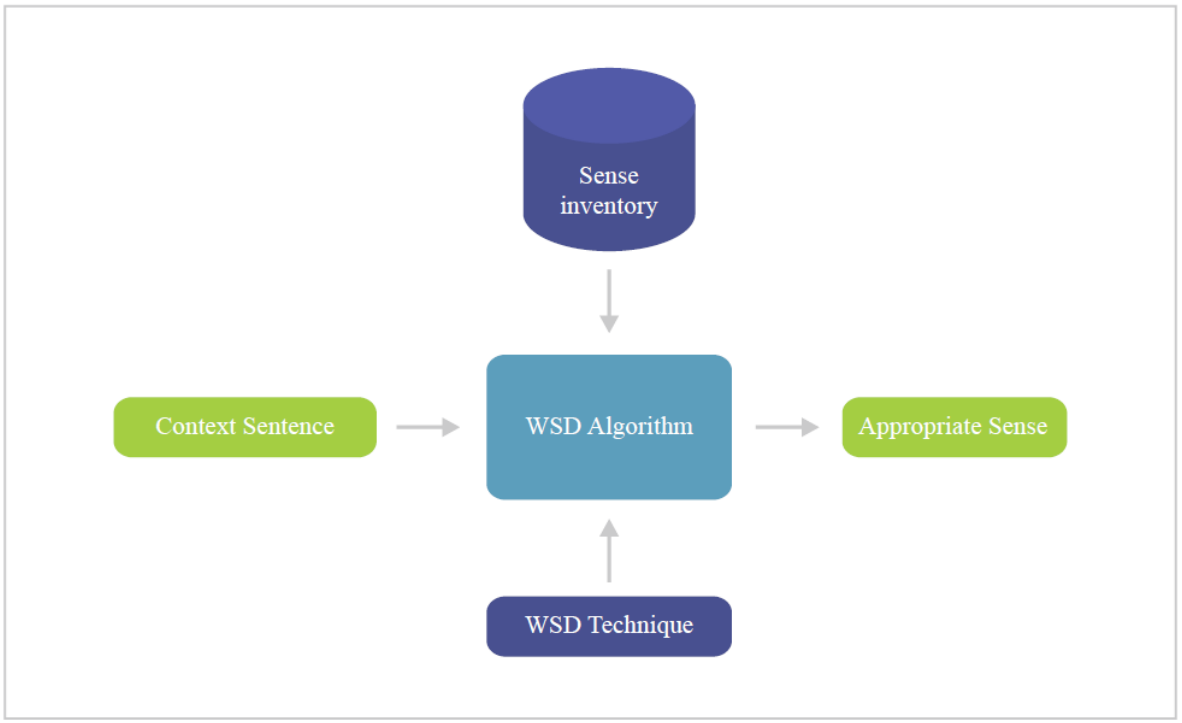


Figure 3.3: WSD framework

The framework comprises three main components: the sense inventory, the WSD technique, and the disambiguation algorithm. The WSD algorithm at the centre of the framework takes the context sentence as input, and outputs the most suitable sense for the ambiguous word within that specific context. To accomplish this task, the WSD algorithm queries the sense inventory to obtain the correct sense of the word. The most commonly used WSD sense inventories in the literature include PWN for English (Orkphol and Yang 2019; AlMousa, Benlamri and Khoury 2022), Indo WordNet for Indian languages (Karuppaiah and Vincent 2021), Arabic WordNet for Arabic (Kaddoura and D. Ahmed 2022), Korean WordNet (Kim and Kwon 2021) for Korean, Polish WordNet (Janz *et al.* 2022) for Polish, among others. The choice of WSD technique varies, depending on the study. These WSD techniques are categorized into different WSD approaches, which are further grouped into WSD methods. A detailed overview of these approaches is provided in the next section. The output of every WSD system is the appropriate sense of an ambiguous word within a specific context, as depicted in Figure 3.3. For further processing in NLP applications, this WSD component is then embedded

into the respective NLP systems. A more in-depth discussion of these applications is presented in the subsequent sections.

3.4 Word Sense Disambiguation Approaches

Over the years, researchers have proposed a diverse range of approaches and methods to address the challenges posed by WSD. These approaches can be broadly categorized into knowledge-based, supervised, unsupervised, and semi-supervised/hybrid methods, each with its own set of methods and techniques. Figure 3.4 provides a diagrammatical overview of the various WSD approaches and their associated methods.

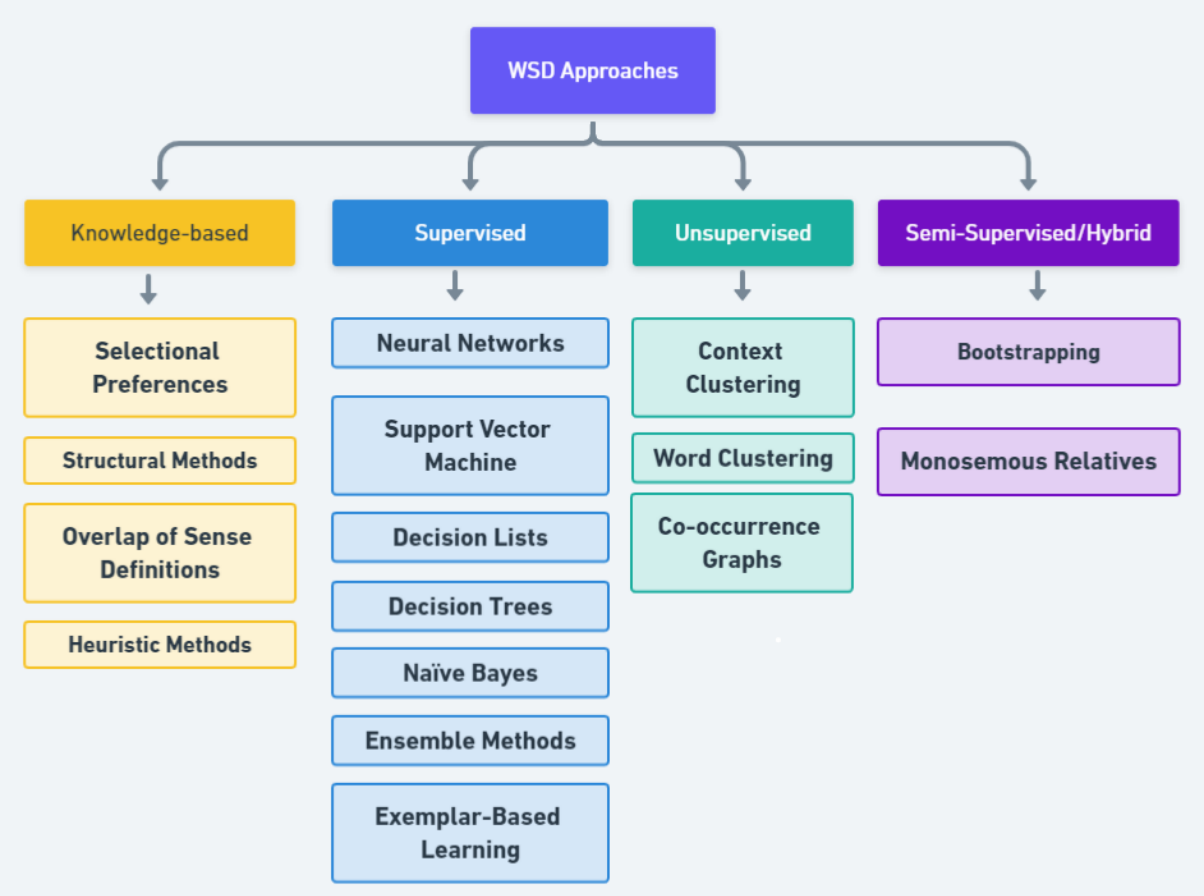


Figure 3.4 A taxonomy of WSD approaches

Knowledge-based methods harness lexical knowledge bases, machine-readable dictionaries, and thesauri. Supervised approaches rely on annotated data to train the model, while unsupervised approaches make use of raw unannotated data. Lastly, semi-supervised approaches combine both annotated and unannotated data, often through a bootstrapping process, to enhance the disambiguation process. These categorizations reflect the various ways in which WSD algorithms leverage available resources and data to improve the accuracy and

effectiveness of sense disambiguation (Wang, Wang and Fujita 2020). The following sub-sections outline these approaches in detail.

3.4.1 Knowledge-based WSD approach

The knowledge-based (KB) WSD approach relies on the use of external knowledge resources to determine the correct sense of an ambiguous word in a given context (Rouhizadeh, Shamsfard and Rouhizadeh 2020). The KB WSD approach applies structured lexical resources which contain information about word senses, word definitions, relationships, and properties. This approach has the advantage of structured knowledge resources that provide rich dependable information about word senses. However, the effectiveness of the approach relies heavily on the quality and coverage of the knowledge resources used (Bevilacqua, 2021). Additionally, the approach struggles with handling ambiguous or rare words that are not well-represented in the resources. Knowledge-based methods include selectional preferences, structural methods, overlap of sense definition calculation methods, and heuristics methods.

3.4.1.1 Selectional preferences method

Early WSD algorithms rely on selectional preferences as a way of constraining the possible meanings of a word in a given context. Selectional preferences process information based on likely relations of word types, assigning to them common sense (Agirre and Martinez 2001). This method of disambiguation requires large collections of selectional preferences: learning accurate semantic constraints requires knowledge of the word senses involved in a candidate relation. Having identified this limitation, several methods were proposed to overcome this, automatically learning selectional preferences based on frequency counts (Montoyo, 2005), information-theory measures (Mihalcea, 2006), or class-to-class relations (Agirre, 2002). In learning word-to-word relations, the frequency count of word-to-word relation is used to measure the semantic fit between words based on the syntactic relation that connects the words. Given two words W_1 and W_2, and the syntactic relation R that associates them, the semantic fit between these words is quantified by counting the number of times the two words occur with the relation R in a corpus which is formalized as a triple $Count$ (W_1, W_2, R). In addition to frequency count, conditional probability determined using Equation (3.2) is used to find the semantic fit between words in a corpus. The probability is calculated based on the selectional preferences relation R learned for two words W_1 and W_2 in which the word W_2 imposes the selectional preferences on W_1 and vice versa, with conditional probabilities in which the roles of the two words are reversed (Carroll and McCarthy 2000).

$$(W_1|W_2, R) = \frac{\text{count}(W_1,W_2,R)}{\text{count}(W_2,R)} \qquad (3.2)$$

In addition to word-to-word relation, selectional preference between a word and a semantic class, or between two semantic classes can be used for WSD. The semantic fit between a word and a semantic class is measured using Eq 3.3, and Eq 3.4, given a word *W* and a semantic class *C* connected by the relation *R*; thus, the selectional association is estimated. The semantic class C in a given relation R is quantified using all the words contained in that class in the selectional association (Agirre and Martinez 2001).

$$P(C|W, R) = \frac{\text{count}(W,C,R)}{\text{count}(W,R)} \qquad (3.3)$$

$$P(W, C, R) = \sum \frac{\text{count}(W,\dot{W},R)}{count(\dot{W})} \qquad (3.4)$$

When disambiguating, the senses of the words are undefined: an equal sense distribution, a word with *N* senses, will have its corpus frequency equally distributed among its possible senses. Compared with word-to-word and word-to-class selectional constraints, class-to-class selectional preferences are more general (Carroll and McCarthy 2000). Eq 3.5 is used to identify the sense of the word W_2 with the maximum probability of co-occurrence of its semantic class with the class of the word it relates to.

$$P\left(W_1^i|W_{2,}R\right) = \frac{count(w_i, w_2,R)}{count\ (w_2,R)} \qquad (3.5)$$

A sense-tagged corpus is mandatory for the disambiguation process because this method assumes the availability of possible senses for words subsumed by classes in the corpus. In the context of developing a Setswana WSD model, the primary limitation lies in its dependency on a sense-tagged corpus, which is challenging to obtain for under-resourced languages such as Setswana. Developing a comprehensive sense-tagged corpus with maximum sense coverage is resource-intensive and time-consuming. Incomplete or biased corpora may lead to suboptimal disambiguation performance. In addition, developing a sense-tagged corpus for Setswana WSD necessitates a comprehensive understanding of the linguistic variances inherent in Setswana, encompassing complexities in grammar, morphology, and sentence structures which may require professional linguists. However, hiring such linguists can be expensive.

3.4.1.2 Structural methods

In the KB WSD approach, structural methods examine the structure of the concept network obtained from a lexical database to determine semantic similarities between words and

concepts. These methods measure semantic similarity computed over a semantic network by calculating the semantic distance between concepts. This method is based on the hypothesis that similar words share a common context; therefore the appropriate sense of a word is located within the shortest semantic distance (Navigli and Velardi 2005). The common structural methods used for WSD are WordNet-based, graph-based, semantic similarity-based, semantic role labelling, and dependency parsing. WordNet is a widely used lexical resource for English. WordNet-based structural methods use WordNet to analyse semantic connections between words in a sentence. Through analysis of the semantic relatedness of words and their respective senses, these methods disambiguate ambiguous words based on the similarity of their surrounding words. The semantic similarity-based methods work similarly to the WordNet-based method. Semantic similarity measure is defined in Eq 3.6:

$$score = \ senses_D \ X \ senses_D \rightarrow [0.1] \qquad (3.6)$$

where $Senses_D$ is the full set of senses listed in a reference lexicon.

The graph-based method explores the structure of a graph to determine the appropriate sense of an ambiguous word (Arab, 2016). Graph-based method techniques include PageRank, Random Walk, Community Detection, extended graph structures and hybrid techniques. Each technique aims to enhance the performance of graph-based methods by refining the way in which structural information is utilized. The choice of a specific technique often depends on the characteristics of the lexical database, the nature of the semantic network, and the linguistic specificities of the target language. This method is based on the lexical chain notion. A lexical chain is sequence of semantically related words in a text. According to this notion, words in a sentence are semantically related by a lexico-semantic relation is-a, has-part, kind-of, part-of relations (Ruas *et al.* 2020). Lexical chains determine context and contribute to the continuity of meaning and the coherence of a discourse (Navigli 2009). Lexical chains can be viewed as a counterpart of the measures of WordNet-based semantic similarity-based methods which, in contrast, are commonly applied in local contexts. Lexical chains are computed as in Eq 3.7:

$$score_{ch}(s_w s_{w'}) = C - d(s_w s_{w'}) - k.turns(s_w s_{w'}) \qquad (3.7)$$

in which C and k are constant variables, d is the shortest distance between the two senses in a lexical resource taxonomy, and *turns* is the number of times the chain "changes direction, i.e., 'turns'." A change of direction is due to the use of an inverse relation. Dependency parsing methods are used to analyse the grammatical relationships between words in a sentence. The method process disambiguation by constructing a dependency tree, which represents the

syntactic dependencies, then identifies the relevant words and their roles in relation to the target word (Ruas *et al.* 2020). To disambiguate, the method uses the dependencies to assign the appropriate sense of the target word based on the context provided by its syntactic neighbours. A dependency constraint rule can be defined as:

If a word w exists in dependency cell r (w, x) or r(x, w), then r (*, x) or r (x, *) are dependency constraints of the word w.

Lastly, semantic role labelling involves identifying and classifying the semantic roles of words or phrases in a sentence. The goal of semantic role labelling is to identify the predicate or the main action of a sentence, assigning specific roles to the words that participate in that action (Navigli and Velardi 2005). The labels represent the semantic relationships between words, providing insights into their roles. The roles can be assigned and represented as in Eq 3.8:

$$V = argmax < sub > x < /sub > P(x \mid S) \quad (3.8)$$

where V is the identified predicate or verb in the sentence S, and P(X|S) is the probability of each role x given the sentence S. The argmax_x denotes selecting the role that maximizes the probability. By leveraging the assigned semantic roles, contextual cues, and role-based features, semantic role labelling assists WSD systems to perform more accurate sense disambiguation decisions, and improves the overall performance of the disambiguation process.

At the core of structural methods is a concept network obtained from a lexical database used to determine semantic similarities between words and concepts for WSD. In the context of Setswana, the concept network for Setswana would need to account for the agglutinative nature of the language. Nodes might represent not only entire words as with other languages such as English, but also the various morphemes that contribute to their meanings. The concept network should consider variations in spelling due to disjunctive orthography, emphasizing the importance of context in interpreting words. Furthermore, edges in the network must reflect the morphological relationships between words, capturing how affixes modify word meanings. The creation of a concept network relies heavily on the availability and diversity of textual data, which is a challenge for resource-scarce languages such as Setswana.

3.4.1.3 Overlap of sense definitions

An overlap of sense definition of a knowledge-based method involves the calculation of word overlap between sense definitions of two or more target words. This method is referred to as

gloss overlap; and it is one of the first algorithms proposed to address the challenge of WSD (Lesk, 1986). This algorithm was developed for sense-disambiguation tasks for all words in a text. The knowledge source required for the algorithm is one or more dictionaries. The algorithm assigns the most appropriate senses of the words in a given context by a measure of dictionary definition overlap. The overlap score for each pair of word senses is computed as follows in Eq 3.9:

$$score_{lesk}(s_1 s_2) = | \, gloss(s_1) \cap \, gloss(s_2) \, | \quad (3.9)$$

Where gloss (S_i) is the bag of words in the textual definition of sense S_i of w_i. Given two words, W_1 and W_2, each with N_{W1} and N_{W2} senses defined in a dictionary, for each possible sense pair W_1^i and W_2^j, $i = 1..N_{W1}$,$j = 1..N_{W2}$, the overlap of the corresponding definitions is determined by counting the number of words they have in common. The sense pair with the highest overlap is selected, and a sense is assigned to each word in the initial word pair.

There are variations of the algorithm, such as simplified Lesk (Kilgarriff and Rosenzweig 2000) and adapted Lesk(Banerjee, 2002), which take cues from the original algorithm. Simplified Lesk is a variation of Lesk that explores the combinatorial explosion of word-sense combinations by performing a separate disambiguation process for each ambiguous word in the input text. The score of this variation is computed as follows in Eq 3.10:

$$score_{leskvar}(s) = context(w) \cap \, gloss(S) \, | \quad (3.10)$$

The context (w) is the bag of all content words in a context window around the target word w. This algorithm finds the correct sense of each word individually by arriving at the sense that leads to the highest overlap between its dictionary definition and the words in the current context. As opposed to the original Lesk that determines all senses of all words in a given text simultaneously, simplified Lesk determines the sense of each word individually, without considering the senses of other words occurring in the same context. The proposed model in this study adopted the Simplified Lesk to develop the PuoBERTa Enabled Embedding-based Lesk WSD Model, making it embedding-based Lesk, which uses vector representations of words to capture semantic relationships between words. The embedding technique uses similarity measure to measure semantic similarity as opposed to calculating overlap.

Another variation of the Lesk algorithm is the adapted Lesk algorithm, in which definitions of related words are used in addition to the definitions of the ambiguous word, to determine the appropriate sense for a word in a given context. A score for each possible combination of senses

definition is calculated to identify the sense with the highest score which is deemed the most appropriate sense. The score of this variation is computed as follows in Eq 3.11:

$$score_{leskext}(s) = \sum s' : s \rightarrow s \mid context(w) \cap gloss(S) \mid \quad (3.11)$$

Contrary to the original and simplified Lesk, adapted Lesk uses hypernyms, hyponyms, holonyms, meronyms, troponyms, attribute relations, and their definitions as context for any given ambiguous word. Presently, Setswana relies solely on the African wordnet(AWN) as machine readable dictionaries(MRD), which offers definitions and examples for specific lemmas and concepts. Due to the absence of these relations, this study was unable to employ variations of Lesk, including the original Lesk, the availability of sentence definitions being limited.

3.4.1.4 Heuristic methods

Heuristic methods evaluate heuristics from various linguistic properties to find a word's sense drawn from large texts. The method consists of rules that assign senses to certain word categories which include the most frequent sense, one sense per discourse, and one sense per collocation (Navigli 2009). For the most frequent sense heuristic, all possible senses that a word may have, one sense occurs more often than the other senses. The most frequent sense method uses word frequency data to assign an appropriate sense to a word. In essence, a word will be assigned its most frequent sense. One major drawback of this method is that sense distributions are not always available, making it impossible to work for languages with small-scale sense-tagged corpora.

Furthermore, different domains have different sense distributions; therefore, a change in domain decreases the performance of this method. The one-sense-per-discourse hypothesis suggests that a word tends to preserve its meaning across all its occurrences in a given discourse. In other words, a word is constantly referred to with the same sense within a specific discourse (Purohit and Yogi 2022). If the meaning is identified in at least one occurrence, it then allows for the automatic disambiguation of all instances of a certain word. This hypothesis has been proven true in coarse-grained sense distinctions; however, with finer-sense distinctions, some words have more than one sense per discourse. The one-sense-per-collocation hypothesis states that a word tends to preserve its meaning when used in the same collocation. This hypothesis asserts that surrounding words strongly and consistently

contribute to determining the sense of a word based on their relative distance, order, and syntactic relationship.

This hypothesis is similar to the one-sense-per-discourse, except for the scope size. Experiments performed with this heuristic on various types of corpora and granularity levels demonstrated a decline in precision; finer levels of ambiguity were encountered, meaning that the heuristic works well for coarse-grained sense distinctions only (Purohit and Yogi 2022). The heuristics of these methods can be used to leverage linguistic properties in determining the sense of words for Setswana. In understanding the syntactic context, contextual information within a given discourse or collocation could be utilized for WSD. However, these senses of words are drawn from large sources of texts which are not available for Setswana.

3.4.2 Supervised WSD approach

The supervised WSD approach tackles the problem of WSD as a classification problem. Classification is a well-established problem in machine learning that involves the categorization of input data into predefined classes based on the labelled training dataset (Conia and Navigli 2021). The classification problem has three aspects, namely, training data, feature selection, and category. The training data is a corpus with a set of examples used to train the algorithm on how to classify new instances. Features are determinants used to determine whether an input belongs to one class or the other; and categories are possible classes in which an input should be assigned. In WSD, classes are possible senses of ambiguous words extracted from a lexical database or dictionary; features are context words, and training data is a sense-tagged corpus.

As shown in Figure 3.4, methods under the supervised WSD approach include neural networks, support vector machine, decision lists, decision trees, naïve Bayes, ensemble methods and exemplar-based learning. Table 3.1 summarizes the characteristic features of each of the knowledge-based methods.

3.4.2.1 Neural networks WSD method

In recent years, neural networks have emerged as powerful tools for WSD. A neural network is an interconnected group of artificial neurons that uses a computational model for processing data based on connectionist approach (Calvo *et al.* 2019). The input of the neural network algorithm is pairs of input features and aimed outcome. The input features are used to partition the training contexts into non-overlapping sets. Neural networks are trained until the desired response is equivalent to the output of the of other unit provided in the training example.

The field of neural networks has witnessed the emergence of multiple types of architectures over time, each with its unique characteristics and capabilities. Attention-based neural network (ANN), recurrent neural network (RNN), convolutional neural network (CNN) and multi-layer perceptron (MLP) are some of the neural network architectures in WSD. The ANN neural network uses attention mechanisms to focus on specific parts of the input sequence. ANN has been applied to WSD by training the network to selectively place attention on the relevant context words for each word in the input sequence, providing sense prediction output based on the processed input sequence (Calvo *et al.* 2019). RNN uses feedback connections to maintain a state memory and process input sequence based on variable length.

LSTM is a type of RNN that utilizes gated recurrent units instead of feedback connections. The main difference between RNN and LSTM lies in the ability to process long-term dependencies (Popov *et al.* 2019). RNN struggles architecturally with capturing long-term dependencies, and suffers from vanishing gradient problem. LSTM addresses these limitations through its ability to capture more complex memory cell structures. In WSD, LSTMs more accurately capture the semantic context and dependencies of words in a sentence, leading to improved sense disambiguation results. While RNN use feedback connections, CNN uses the feedforward neural network on convolutional layers to learn local feature representation from input data. CNNs are used for WSD by applying convolutional filters over the input sequence of words and performing sense prediction (Jadiya, Dondemadahalli Manjunath and Mohan 2022). MLP is a feedforward neural network that takes input as a feature vector representation of the word, and outputs a probability distribution over the possible senses. While neural network methods have shown promising results in WSD for some languages, their application to Setswana in this context faces significant challenges due to the language's morphological complexity, disjunctive orthography, and lack of annotated data for training the neural network.

3.4.2.2 Support vector machine WSD method

Support vector machine (SVM) is a supervised machine-learning algorithm. The SVM algorithms are generally used for classification and regression tasks. The SVM main objective is to find an optimal hyperplane in a high-dimensional feature space that can separate various classes of datapoints within the largest possible margin (Yadav *et al.* 2021). The key components of SVN are feature space transformation, margin optimization, regularization parameters and kernel trick. Feature space transformation enables SVM to map input data from the original feature space to a higher dimensional space using the embedded kernel function; while the margin optimization finds the hyperplane that maximizes the margin between

different classes. The regularization parameter is introduced as a controller that manages the trade-off between achieving a larger margin and minimizing the misclassification of training examples. The kernel trick is used to efficiently compute the dot product in the higher dimensional space without transforming the data. The commonly used kernel functions in SVM are as follows:

Linear kernel:

$K(x, y) = x \cdot y$

where x and y are the input vectors in the original feature space.

Polynomial kernel:

$K(x, y) = (x \cdot y + c)^d$

where c is a constant and d is the degree of the polynomial.

Gaussian (radial basis function) kernel:

$K(x, y) = \exp(-\gamma \|x - y\|^2)$

where γ is a parameter controlling the kernel's width.

In WSD, SVM is used to classify word instances into their appropriate senses. In this method, each word instance is represented by a set of features that train the SVM classifier. The SVM has been shown to achieve the best results in WSD compared with several supervised approaches (Bevilacqua, 2021) however, SVM cannot be used without the presence of an annotated dataset.

3.4.2.3 Decision lists WSD method

Decision lists is a rule-based method that leverages a set of rules to make sense predictions. In WSD, decision lists are used for assigning the appropriate sense to a target word, using an ordered set of rules constructed for categorizing test instances. Decision lists are generally viewed as lists of weighted "if-then-else" rules. This method relies on the concept that certain features and feature combinations strongly indicate specific word senses. The ordering of rules is based on the predictive power, in which the higher score rules are prioritized to enable efficient disambiguation. Given a target word occurrence and its representation as a feature vector, the decision list is examined; and the feature with the highest score that matches the input vector chooses the appropriate word sense to be assigned as in Eq 3.12:

$$S = argmax_{Si \in SensesD(w)} score(S_i) \qquad (3.12)$$

The score of the senses is calculated as the maximum among feature scores. Compared with other supervised approaches, the decision-list method is preferable — it provides interpretability by explicitly showing the decision rules used in the prediction process. This is desirable in the development of medical applications. However, constructing an effective decision list requires careful selection of informative features and crafting of accurate decision rules from the training data. Due to the absence of training data, the decision-list method cannot be applied for the development of Setswana WSD. Without sufficient data, it is impossible to identify and extract meaningful features or to derive reliable decision rules.

3.4.2.4 Decision trees

A decision tree is a tree-structure-based predictive model used for classification. This method recursively partitions the training dataset wherein each internal node of the decision tree represents a test of a feature value, with each branch representing an outcome of the test (Mir *et al.* 2023). Decision trees is a path-based method commonly used to classify word instances into different classes based on their contextual features. While decision lists follow a sequential structure, decision trees follow a hierarchical structure with the root node representing the initial split based on a feature of the target word. Both decision lists and decision trees rely on a set of decision rules to classify ambiguous words into various senses based on the features of the context in which the word appears, while also having distinct differences in terms of their structure, learning algorithms, and decision-making processes, which may lead to variations in their performance and applicability to various WSD tasks. Similar to decision lists, decision trees' effectiveness for Setswana WSD in the absence of an annotated corpus, would be severely limited. Overcoming these challenges would require the collection of labelled training data specific to Setswana.

3.4.2.5 Naïve Bayes

Naïve Bayes (NB) is a probabilistic-based classifier founded on Bayes' theorem. NB calculates conditional probability of each sense of a target word based on the features of the context (Bhattacharjee *et al.* 2020). The NB theorem is expressed in Eq 3.13:

$$\mathrm{P(y \mid x)} = \frac{\mathrm{P(X|Y)P(Y)}}{\mathrm{P(X)}} \qquad (3.13)$$

where:

P(y | x) is the posterior probability of Class y given the features x

P(x | y) is the likelihood probability of features x given the Class y

P(y) is the prior probability of Class y

P(x) is the probability of features x

The naïve Bayes classifier estimates the probabilities P(x | y) and P(y) from the training data and then applies Bayes' theorem to make predictions. The class label with the highest posterior probability is selected as the predicted class for a given set of features.

Naïve Bayes performs well, even with limited training data. NB can provide reasonable results with small training sets, making it suitable for WSD tasks in which annotated data may be either limited or expensive to obtain (Kim and Kwon 2021). However, NB assumes complete feature information for each instance. Handling missing data or out-of-vocabulary words requires additional pre-processing or imputation techniques to avoid compromising performance. Hence, for Setswana, where there may be challenges in obtaining a large, annotated corpus, and the language exhibits morphological richness and agglutinative properties, NB could be a viable option for WSD. However, careful consideration should be given to modelling feature information for each sense instance thoroughly, also coverage of a large language vocabulary which is also a challenge for Setswana.

3.4.2.6 Ensemble Methods

Ensemble methods are a combination of various classifiers integrated into one algorithm to improve the disambiguation accuracy (Navigli 2009). The classifiers commonly combined are of a different nature, with different characteristics. These methods address some of the limitations of supervised techniques. The various types of ensemble methods include stacking, majority voting, probability mixture, rank-based combination, and adaptive boosting (Patel *et al.* 2021). The stacking method combines multiple classifiers by training a meta-classifier that takes the outputs of individual classifiers as input features. Majority voting uses the voting technique in which each component provides a vote for each sense of an ambiguous word. The sense with the majority votes is selected as the appropriate sense. The voting is computed as follows in Eq 3.14:

$$S = argmax_{Si \in SensesD(w)} \left| \{ j : vote(C_j) = S_i \} \right| \qquad (3.14)$$

In probability mixture, classifiers are grouped into different orders. There are first-order classifiers, second-order classifiers, and third-order classifiers. Suppose the first-order classifiers provide a confidence score for each sense of a target word, the confidence scores are normalized:

$$S = argmax_{Si \in SensesD(w)} \sum_{j=1}^{m} -ranksc_j(S_i) \qquad (3.18)$$

The rank is the rank of S_i senses as output by various classifiers.

Adaptive boosting is a process for constructing a best-performing classifier as a linear combination of less-performing classifiers. In this method, weak classifiers are iteratively trained; each subsequent classifier's objective is to correct mistakes made by the previous classifier, thus reducing overall classification error. The classifiers are assigned weights based on their performance; and their predictions are combined for a final output. The classifiers are combined as follows on Eq 3.19:

$$H(x) = sign \sum_{j=1}^{m} \propto_j C_j(x) \qquad (3.19)$$

where x is an example from the training set, (C_1, ... , C_n) are the first-order classifiers that we want to improve, α_j determines the importance of classifier C_j, and H is the resulting strong classifier. H is the sign function of the linear combination of the weak classifiers, which is interpreted as the predicted class. While ensemble methods address some of the limitations of supervised techniques by combining two or more classifiers, this method is not applicable in the context of the development of Setswana WSD model due to the absence of training data.

3.3.2.7 Exemplar-based learning

Exemplar-based learning methods use a set of examples to train a classification model (Navigli 2009). The classification model stores the examples in memory as points on a vector space. The new examples are added to the vector based on the classification outcome. To disambiguate senses, the algorithm compares the features of the target word's context with the stored exemplars; and assigns the most similar sense label (Rais-Ghasem and Corriveau 2020). This approach has several advantages including the ability to handle new and unseen senses efficiently, by comparing the context of the target word with exemplars on the points vector space. However, for Setswana, where the language exhibits morphological richness and agglutinative properties, the method may struggle to generalize well to unseen data, or capture

the full complexity of Setswana's linguistic characteristics, in addition to the lack of annotated datasets which can train the model. Supervised WSD methods rely heavily on the availability of labelled training data. For Setswana, there is a scarcity of annotated corpora, which hinders the application of supervised techniques. Without sufficient labelled examples, training and evaluating supervised models becomes difficult. In addition, Setswana is a morphologically rich language with agglutinative properties and disjunctive orthography. As a result, words are formed by combining multiple morphemes, leading to complex word structures. Supervised methods, particularly those based on fixed-length feature representations, may struggle to capture the fine-grained morphological information present in Setswana words, affecting disambiguation accuracy. This disjunctive orthography also poses a challenge for supervised methods, the separated units possibly carrying important semantic information that influences the sense of the word.

3.4.3 Unsupervised WSD

The task of WSD is at times considered a sense-labelling task for supervised approaches, the process of assigning sense labels to ambiguous words; however, for unsupervised WSD, the task is considered a word discrimination task due to the techniques used (Rahman and Borah 2022). Unsupervised approaches are based on the concept that the words surrounding the target ambiguous word will have similar senses. These methods automatically induce senses from unannotated text by clustering word occurrences and classifying them into clusters, i.e., discriminating senses. Unsupervised methods do not rely on labelled data, and have the capacity to overcome the data-sparsity and knowledge-acquisition bottleneck challenge posed by knowledge-based and supervised approaches (Navigli 2009). The main distinction between knowledge-based, supervised, and unsupervised approaches, is that the unsupervised approach method's aim is to identify sense clusters, while knowledge-based and supervised approach method's aims are to assign sense labels. Unsupervised WSD methods include context clustering, word clustering, and co-occurrence graph.

3.4.3.1 Context clustering WSD method

In context clustering, each occurrence of a target word in a corpus is represented as a context vector. Such vectors are subsequently clustered into groups, each discerning the sense of the target word. Context-clustering methods are based on the hypothesis that the distributional profile of words implicitly demonstrates their semantics (Bhattacharjee *et al.* 2020). The fundamental notion behind this method is that different senses occur in different contexts.

Therefore, clustering contexts enables the identification of a group of senses associated with identical contextual patterns (Jia *et al.* 2021). Context clustering is a data-driven approach and can be applied to various applications and domains. This method leverages large amounts of unstructured text data. The limitations of this method are the challenge of defining and representing context features, the method's sensitivity to noise and variability in the data, as well as the lack of transparency and interpretability of the clustering results (Mennes and van Gulik 2020). To apply this method effectively to Setswana, a substantial unannotated corpus would be necessary to cluster the contexts accurately. However, the available data for Setswana from domain-specific sources, such as government domains, may not be suitable, in that the data is already domain specific.

3.4.3.2 Word clustering WSD method

In contrast to context clustering, there is word clustering. The word-clustering method groups words which carry a specific meaning and are semantically similar. The hypothesis of this method is that words with similar contexts tend to have similar meanings, therefore grouping such words provides additional information for disambiguating the sense of the target word. One common way in which this method is employed for WSD is through the use of distributional semantic models. Here the meaning of the word represented on a vector is based on the distribution of its context words in a large corpus of text. In this context, distributional models can be used to evaluate similarity between words based on their co-occurrence patterns, thus grouping similar words into clusters. Let w be the target word, C;_w the set of contexts of w; s the sense of w; and f_s(C) the frequency of sense s in context C. Let K be the number of clusters, and S_k be the set of senses in Cluster k. We can represent the word clustering approach as follows in Eqs 20, 21 and 22:

1. Collect contexts:

$$C_w = \{C_i \mid w \in C_i \text{ and } \mid C_i \mid = context\_size\} \qquad (20)$$

2. Cluster words:

$$W_k = \{w_i \mid w_i \in C_j \, , j \in [1,|C_w|], w_i \in k\} \qquad (21)$$

3. Disambiguate:

$$s = args\max \frac{1}{C_w} \sum_{C \in C_w} f_s(C) \qquad (22)$$

In this representation, Step 1 collects the contexts of the target word with the specified window size of context size. Step 2 clusters the words in these contexts into K clusters. Finally, Step 3 disambiguates the sense of the target word by selecting the sense with the highest frequency in the cluster that contains the most frequent words in the context. Similar to context clustering, word clustering requires a large amount of unannotated corpus accurately to cluster semantically related words. However, apart from government-domain documents that are publicly available for Setswana, there is insufficient data to effectively utilize this method.

3.4.3.3 Co-occurrence graphs WSD method

Co-occurrence graphs represent the relationships between words based on their co-occurrence patterns in a large corpus of text. In co-occurrence graphs, nodes represent words, and edges represent co-occurrence relationships between words. Co-occurrence graphs provide a way of incorporating knowledge about word associations and semantic relationships into the WSD process. By leveraging the co-occurrence patterns from a large corpus, these graphs can enhance the disambiguation process by considering the broader context and semantic information. However, the effectiveness of co-occurrence graphs depends on the quality and representativeness of the corpus used for their construction, as well as the specific algorithms employed for inference. The use of co-occurrence process can be modelled as:

Create the co-occurrence graph

G = (V, E)

The nodes represent the words in V and the edges represent the co-occurrence relationships between them.

e(i,j)

cooc(w_i,w_j)

The weight of an edge e(i,j) between two words w_i and w_j is given by the co-occurrence frequency count cooc(w_i,w_j) in the context.

G = (V, E) where E = {e(i,j) | i,j in V}, w(i,j) = cooc(w_i,w_j)

For each instance of the target word w in the context C, compute the co-occurrence vector v(w,C) of word w with all other words in C. The co-occurrence vector is defined as:

v(w,C) = [cooc(w,w1), cooc(w,w2), ..., cooc(w,wn)]

Compute the PageRank score of the nodes in N, which are the nodes that co-occur with w in C. The PageRank score of a node i is given by:

PR(i) = (1-d) + d * sum(j in in(i)) {PR(j) * w(j,i) / sum(k in Out(j)) w(j,k)}

where In(i) is the set of nodes that has an edge pointing to node i, Out(j) is the set of nodes that has an edge pointing from node j, d is the damping factor (usually set to 0.85), and w(j, i) is the weight of the edge from node j to node i.

Select the sense with the highest PageRank score among the senses of w. Let the set of senses of w be denoted as S(w); then the sense s* of w is given by:

s* = argmax(s in S(w)) {PR(s)}

The final step is to return the selected senses s*

However, for Setswana WSD, obtaining a sufficiently large and diverse corpus is challenging, limiting the applicability of co-occurrence graphs in this context. Additionally, the agglutinative nature and complex morphology of Setswana may pose challenges for accurately capturing word associations and semantic relationships in co-occurrence graphs. Unsupervised methods present a promising direction for developing a Setswana WSD model, in that they can overcome the limitations of labelled data scarcity. Unlike supervised methods, which require sense-labelled training data, unsupervised methods aim to induce sense clusters from unannotated text by leveraging the distributional properties of words and their contexts. However, as the success of these methods relies heavily on the availability and quality of unannotated Setswana corpora, this data is not available.

3.4.4 Semi-supervised WSD approach

The semi-supervised approach uses both labelled and unlabelled data for training a model. In contrast to supervised learning, in which the model is trained only on labelled data, and unsupervised learning, in which the model is trained only on unlabelled data, semi-supervised learning leverages the benefits of both labelled and unlabelled data to improve model performance (Mennes and van Gulik 2020). These methods are based on an automated bootstrapping process to construct a corpus from human-tagged examples. Semi-supervised WSD methods include bootstrapping and monosemous relatives.

3.4.4.1 Bootstrapping

The bootstrapping technique is a process of training a machine-learning algorithm on a small amount of data, and using the model to label large amounts of unlabelled or unannotated data (Pal *et al.* 2019). The output of the algorithm is then combined with the original data to train a new model literately, until the targeted level of performance results is reached. The objective of this approach is to leverage the availability of large amounts of unannotated data to improve the performance of a model, while minimizing the cost, time, and effort associated with annotating large amounts of data through manual labour (Pal and Saha 2019).

In the context of WSD, the bootstrapping technique involves training a sense classifier on a small amount of manually annotated data before using the trained model to label a large pool of unlabelled or unannotated data. The process is iterative, with the output of the algorithm being combined with the original annotated data to train a new model in each iteration. This cycle continues until the desired level of performance has been reached.

The effectiveness of the bootstrapping technique in WSD depends on several factors, including the quality of the initial annotated data, the size and diversity of the unannotated data, and the choice of classifiers (Navigli 2009). The selection of informative and representative instances from the unannotated data is crucial to the success of the bootstrapping process (Mihalcea 2004). Additionally, the use of confidence thresholds and stopping criteria can help prevent the propagation of errors, ensuring the stability of the iterative learning process (Abney 2004).

While bootstrapping has shown promising results in WSD, it also has some limitations. The performance of the bootstrapping process can be sensitive to the initial annotated data and the choice of classifiers (Navigli 2009). However, for Setswana, a resource-scarce language, bootstrapping may face additional hurdles. Obtaining a sufficient amount of high-quality annotated data to initiate the bootstrapping process can be difficult due to limited linguistic resources for Setswana. Additionally, the iterative nature of bootstrapping may increase errors and inaccuracies in the initial seed data, potentially leading to decreased WSD accuracy.

3.4.4.2 Monosemous relatives WSD method

In the monosemous relatives WSD method, words that are possible synonyms with the target ambiguous words are used to annotate a large corpus of texts and to formulate sense-annotated data for training WSD classifiers. In this method, the process involves identifying unique expressions for a specific sense of a word using different heuristics, searching the web for text snippets containing those expressions, and then tagging each snippet with the relevant sense.

The performance of these methods can be assessed manually, or by comparing the accuracy of WSD systems trained with hand-labelled data and bootstrapped corpus of labelled examples from the web. The monosemous relatives uses data on the web as a corpus. This idea was introduced in 2003 by Kilgarriff and Grefenstette (2003), and has been explored in the field of WSD to automatically build datasets with the goal of addressing the knowledge-acquisition bottleneck and data-sparsity difficulties.

To sum up, developing a WSD solution for Setswana, a resource-scarce language, presents significant problems due to the lack of necessary language resources and annotated data required by most WSD approaches. Knowledge-based approaches, which rely on lexical resources, sense-tagged corpora, and large-text data, are limited by the scarcity of such resources for Setswana. Supervised approaches, including neural networks, support vector machines, decision lists, and others, require annotated training data, which is scarce for Setswana. Additionally, the morphological complexity, agglutinative nature, and disjunctive orthography of Setswana pose additional challenges for these methods.

Unsupervised approaches, such as context clustering, word clustering, and co-occurrence graphs, show promise in that they do not require annotated data. However, their success depends heavily on the availability of large, diverse, and representative unannotated corpora, which is an obstacle for Setswana. The agglutinative nature and complex morphology of Setswana may also pose difficulties for accurately capturing word associations and semantic relationships in these methods. Semi-supervised approaches, such as bootstrapping and monosemous relatives, can leverage both annotated and unannotated data, making them potentially suitable for Setswana WSD. However, obtaining a sufficient quantity of high-quality annotated data to initiate the bootstrapping process can be difficult, due to limited linguistic resources for Setswana. Table 3.1 summarizes the key features of the various types of WSD approaches and methods.

Table 3.1: WSD Approaches Key Features and Methods

Approach	Methods	Key Features
Knowledge-based	Selectional preferences, structural methods, overlap of sense definitions, heuristic methods	Leverages lexical resources, machine-readable dictionaries, and thesauri Relies on structured knowledge resources

Supervised	Neural networks, support vector machines, decision lists, decision trees, naive Bayes, ensemble methods, exemplar-based learning	Treats WSD as a classification problem Requires annotated training data Uses machine learning algorithms
Unsupervised	Context clustering, word clustering, co-occurrence graphs	Induces sense clusters from unannotated text Leverages distributional properties of words and their contexts Requires large unannotated corpora
Semi-supervised	Bootstrapping, monosemous relatives	Utilizes both labelled and unlabelled data Bootstrapping: iteratively trains on labelled data and learns from unlabelled data. Monosemous relatives uses synonyms to annotate data.

Given the resource-scarce nature of Setswana, this study opted for a knowledge-based approach to developing a WSD model for Setswana. The selection of this method required the construction of a novel lexical resource, which was subsequently integrated with existing lexical resources for Setswana. The combination of these resources aimed to mitigate the inherent limitations posed by the resource-scarcity.

3.5 Word Sense Disambiguation Knowledge Resources

WSD relies heavily on the availability of datasets in the form of lexical resources as a key element. Knowledge resources are crucial to providing information on words with their corresponding senses (Navigli 2009). These resources can vary in form and structure from machine readable dictionaries(MRD), thesauri, annotated and unannotated corpora, glossaries, knowledge ontologies, inter alia.

3.5.1 Structured resources

Structured resources play a crucial role in WSD by providing the necessary knowledge and information to determine the correct sense of a word in a given context. These resources are typically organized in a well-defined format, such as a lexical database or ontology; and contain semantic information about words, their senses, and the relationships between them.

3.5.1.1 Machine readable dictionaries (MRD)

An MRD is an electronic version of a traditional dictionary that is specifically designed to be processed and interpreted by a computer (Bevilacqua *et al.* 2021). Contrary to printed dictionaries, an MRD is structured in such a way that allows for efficient storage, retrieval, and manipulation of lexical information by NLP applications. MRDs are typically organized in a structured format, such as a database or an XML-based markup language, enabling automated access and analysis. MRDs include not only the definitions of words but additional linguistic details such as part-of-speech tags, synonyms, meronymy, antonyms, example sentences, pronunciation guides, and sometimes even semantic relations between words, making it a valuable resource for several natural language processing (NLP) applications including WSD (Navigli 2009). The dictionaries, such as the Collins English Dictionary, the Oxford Advanced Learner's Dictionary of Current English, the Oxford Dictionary of English, and the Longman's Dictionary of Contemporary English (LDOCE), were some of the first available in electronic format.

3.5.1.2 WordNets

WordNets are lexical databases that organize words into synsets, sets of synonyms, and provide semantic relationships between these synsets (McCrae, Rudnicka and Bond 2020). The core building blocks of WordNets, synsets, are sets of words that are synonymous or semantically related which is one of the main features that characterize WordNets. Each synset represents a distinct concept or meaning; and words within the same synset are considered interchangeable in certain contexts. WordNets are further characterized by their structured organization of lexical information, which includes semantic relations, and part-of-speech (POS) categorization (Miller and Fellbaum 2007). WordNets capture various semantic relationships between synsets, such as hypernymy (is-a relation), meronymy (part-of relation), and antonymy (opposite relation). These relations create a rich semantic network that enables the exploration of word meanings and their connection (Fellbaum 2010). Furthermore, linguistic data in WordNets is organized based on their POS categories, such as nouns, verbs, adjectives,

and adverbs. This categorization helps in disambiguating word senses by considering the grammatical contexts in which they appear (Beckwith *et al.* 2021).WordNet was originally developed for the English (Princeton WordNet) language; however, it has been adapted to many other languages, as outlined below.

a) English WordNet

The English WordNet is a collaborative, open-source adaptation of the Princeton wordnet (PWN). Its primary goal is to maintain an updated and high-quality English WordNet that can benefit a wide range of users (McCrae, Rudnicka and Bond 2020). The English WordNet was developed to allow easy contributions and enhancements from the community, ensuring continuous improvement. The developers used a straightforward approach, in which a GitHub repository for simple XML documents, which has proven successful with over 18,500 changes and numerous contributions, was established in an attempt to improve the PWN.

b) African WordNet

The African WordNet is a linguistic resource for African languages, which are often under-resourced, linguistically. The project focuses on creating WordNets for languages such as Setswana, isiZulu, isiXhosa, and Sesotho. The construction of the AWN was developed following the structure and principles of the Princeton WordNet, organizing words into synsets, and capturing semantic relations between them. This further included the selection of a core concept set, the translation of synsets from English to the target African languages, and the establishment of semantic relations between synsets.

c) Indo WordNets

Indo WordNet is a multilingual lexical database that covers major Indian languages. Indo WordNet aims to create a linked structure of WordNets for Indian languages, following the design principles of the Princeton WordNet (Bhatt and Bhattacharyya 2011). The development of Indo WordNet involves the creation of synsets for each Indian language, along with the establishment of semantic relations between synsets within each language-specific WordNet (Bhatt and Bhattacharyya 2011). Additionally, Indo WordNet focuses on aligning synsets across various Indian languages, enabling cross-lingual information access and processing.

3.5.1.3 Thesauri

A thesaurus is a specialized reference tool that provides information about word and concepts relationships, such as synonymy and antonymy. The thesaurus offers a valuable collection of

word pairs with similar or opposite meanings. Thesauri assist in resolving word sense ambiguities by providing a broader semantic context and facilitating a more comprehensive understanding of the target word. The rich vocabulary of synonyms and related terms offered by thesauri aids in mapping the appropriate sense to a given word based on the surrounding context. One widely utilized thesaurus in the field of WSD is Roget's International Thesaurus, which includes a vast array of 250,000 word entries organized into six classes and approximately 1000 categories. Researchers often leverage this comprehensive thesaurus to access a rich source of synonyms and related terms.

3.5.1.4 Ontologies

An ontology consists of a taxonomy which represents hierarchical categorizations, and a set of semantic relations that describes the connections between various concepts or nodes. In essence, ontologies are referred to as structured specifications of conceptualizations that capture the knowledge and relationships within specific domains. By providing structured representations of concepts and their relationships, ontologies facilitate various tasks, including natural language understanding and knowledge reasoning, making them relevant for WSD. In WSD, ontologies serve as sense inventories; and provide a structured framework for representing senses or meanings of words. Each sense is defined as a concept within the ontology; and the relationships between senses are captured through semantic relations.

3.5.2 Unstructured resources

Unstructured resources are another important type of knowledge source used in WSD. Unlike structured resources, which have a well-defined format and organized semantic information, unstructured resources are typically raw text corpora or other forms of textual data that do not have a predefined structure. These resources are valuable for WSD because they provide a vast amount of contextual information that can be used to infer the meaning of words in different contexts.

3.5.2.1 Raw corpus

A raw corpus is a collection of unprocessed, unlabelled, and unannotated text documents that serve as the primary knowledge resource of data for training and evaluating WSD systems (Baruah *et al.* 2021). A raw corpus is typically a large-scale collection of text documents gathered from various sources such as books, articles, websites, or other textual resources. Examples of widely used English raw corpus include Brown Corpus (Francis and Kucera

1979), British National Corpus (BNC) (Leech 1992), Wall Street Journal (WSJ) corpus (Paul and Baker 1992), and the American National Corpus (Ide and Suderman 2004).

3.5.2.2 Sense-annotated corpus

A sense-annotated corpus refers to a collection of text documents or data that have been manually annotated with sense labels or sense annotations for target words, in some instances all words (Saeed *et al.* 2019a). The purpose of creating a sense-annotated corpus is to provide labelled data that can be used for training and evaluating supervised WSD systems. Sense-annotated resources serve as valuable resources for training machine-learning models and testing the accuracy of disambiguation algorithms (Navigli 2009). This plays a crucial role in advancing WSD research and developing effective applications in natural language processing; and has enabled the development of the best-performing WSD systems. Examples of widely used English sense-annotated corpus used as standard benchmark WSD evaluation dataset include SemCor (Miller *et al.* 1993).

In the context of developing a WSD model for Setswana, the availability and quality of knowledge resources play crucial roles. Because Setswana is an under-resourced language, there is a scarcity of both structured and unstructured resources in general, and especially those specifically designed for WSD tasks. This poses challenges in terms of developing, training, and evaluating WSD models for Setswana.

Among the structured resources, WordNets have proven to be valuable for WSD in various languages. The African WordNet project, which includes the development of a Setswana WordNet, is a promising initiative in this regard. By organizing Setswana words into synsets and capturing semantic relationships between them, the Setswana WordNet can serve as a foundational resource for disambiguating word senses in Setswana text. However, the coverage and depth of the Setswana WordNet may be limited compared with well-established WordNets such as the Princeton WordNet for English.

Machine-readable dictionaries (MRDs) are another type of structured resource that can be leveraged for Setswana WSD. While there may be limited availability of Setswana-specific MRDs, efforts can be made to digitize and structure existing Setswana dictionaries to create MRDs. These MRDs can provide valuable information about word definitions, part-of-speech tags, and semantic relations, aiding in the disambiguation process.

Thesauri and knowledge ontologies, although not specific to Setswana, can still be useful for WSD tasks. Thesauri offer a collection of synonyms and related terms, which can help in

understanding the broader semantic context of words. Knowledge ontologies, such as domain-specific ontologies, can provide structured representations of concepts and their relationships, facilitating WSD in specific domains.

Unstructured resources, such as raw corpora and sense-annotated corpora, are essential for training and evaluating WSD models. The availability of Setswana raw corpora, such as collections of Setswana text from various sources, is crucial for capturing the linguistic patterns and context in which words appear. However, the limited size and diversity of Setswana corpora compared with well-resourced languages like English may pose challenges in terms of coverage and representativeness.

Sense-annotated corpora, which are manually annotated with sense labels, are particularly valuable for developing supervised WSD models. However, creating sense-annotated corpora for Setswana is a time-consuming and resource-intensive task. Collaborative efforts among researchers, language experts, and the Setswana-speaking community can help in developing such corpora, even if on a smaller scale than English sense-annotated corpora such as SemCor: currently, there is no existing sense-annotated WSD data.

3.6 Word Sense Disambiguation Evaluation Methods

There are two distinct approaches to evaluating the performance and efficacy of WSD systems, intrinsic and extrinsic evaluation. An additional aspect to assessing WSD systems is the evaluation dataset. This section outlines both intrinsic and extrinsic evaluation methods, as well as some of the WSD evaluation benchmark datasets and challenges highlighted in the literature regarding these datasets.

3.6.1 Intrinsic Evaluation

Intrinsic evaluation focuses on evaluating the performance of WSD systems based solely on the disambiguation task without considering any downstream applications (Edmonds and Kilgarriff 2002). This evaluation measures how well the WSD system performs in correctly assigning sense labels to target ambiguous words. Intrinsic evaluation involves using a knowledge-based resource, manually annotated sense-annotated corpora in which the WSD system's performance is evaluated against the annotations in the data. Accuracy (A) on Eq 23 and Eq 24, precision (P) on Eq 25 and Eq 26, recall (R) on Eq 27 and Eq 28, and the F1 score (F1) on Eq 29, are the commonly used evaluation metrics for WSD (Navigli 2009).

Accuracy proportion of the accurately predicted senses over the total senses is annotated in the used dataset. The general formula on (23) and quantified formula (24):

$$A = \frac{correct\ answers}{total\ answers} \qquad (23)$$

$$A = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative} \qquad (24)$$

Precision is the proportion of accurately predicted senses over the proportion of senses returned. The general formula on (25) and quantified formula (26):

$$P = \frac{correct\ answers}{provided\ total\ answers} \qquad (25)$$

$$P = \frac{true\ positive}{true\ positive + false\ positive} \qquad (26)$$

The recall is the proportion of accurately predicted senses over the total identified senses. The general formula on (27) and quantified formula (28):

$$R = \frac{provided\ correct\ answers}{expected\ total\ answers} \qquad (27)$$

$$R = \frac{true\ positive}{true\ positive + false\ negative} \qquad (28)$$

The F1 score is a uniform harmonic mean of precision and recall calculated using the formula on (29):

$$F = \frac{2(precision * recall)}{precision + recall} \qquad (29)$$

In the context of this study, for evaluating the proposed WSD model, accuracy is a straightforward metric that gives an overall assessment of the WSD model's performance by measuring the proportion of correctly predicted senses of the total senses annotated in the evaluation dataset. However, accuracy alone may not provide a complete picture of the model's performance, especially in cases in which the sense distribution is imbalanced, or when the cost of false positives and false negatives differs.

To address the limitations of accuracy, precision and recall were employed. Both these metrics are often used in combination. Precision measures the proportion of correctly predicted senses of the total senses returned by the WSD model, focusing on the model's ability to avoid false positives. On the other hand, recall measures the proportion of correctly predicted senses of the total identified senses, emphasizing the model's ability to identify all the relevant senses.

The F1 score is a harmonic mean of precision and recall, providing a balanced measure of the WSD model's performance. The F1 score is particularly useful when considering both precision and recall equally; and when the sense distribution is imbalanced. The F1 score helps to capture the trade-off between precision and recall, giving a more comprehensive evaluation of the WSD model.

While accuracy, precision, recall, and the F1 score are commonly used metrics for intrinsic evaluation of WSD models, there are other metrics which could be considered, such as receiver operating characteristic (ROC) graphs, and the area under the ROC curve (AUC). ROC graphs plot the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, providing a visual representation of the WSD model's performance. The AUC measures the ability of the WSD model to discriminate between correct and incorrect sense assignments, with a higher AUC indicating better performance.

Based on the conventions and practices in the WSD research community, as well as the specific goals and requirements for Setswana WSD modelling, this study opted to evaluate the proposed model based on accuracy, precision, and recall.

3.6.2 Extrinsic evaluation

Extrinsic evaluation assesses the performance of WSD systems in the context of a specific downstream task or application that relies on WSD (Edmonds and Kilgarriff 2002). With this evaluation, the WSD system is integrated into an application; and measures the impact of WSD on the overall performance of the system. The WSD system measures how effectively WSD

improves the performance of the downstream task, such as machine translation, information retrieval, question answering, or document classification.

3.6.3 WSD evaluation datasets

To facilitate the standardized evaluation and comparison of WSD systems, several benchmark datasets including SemCor (Miller *et al.* 1993), OMSTI, SenseEval-02 (Edmonds and Cotton 2001), SenseEval-03 (Mihalcea, 2004), SemEval-07 (Pradhan *et al.* 2007), SemEval-13 (Navigli, Jurgens and Vannella 2013), SemEval-15 (Moro and Navigli 2015) have been developed and widely adopted by the research community. These datasets serve as gold standards for assessing the performance of WSD models; and provide a common ground for researchers to evaluate and compare their approaches. However, these datasets differ in format, construction guidelines, and underlying sense inventory (Raganato, Camacho-Collados and Navigli 2017).

In the case of the datasets annotated using WordNet, the de facto sense inventory for WSD, there is the barrier of having text annotated with different versions. It is the norm to solve these divergencies individually by using or constructing automatic mappings. However, the quality check of such mapping tends to be impractical, and this leads to mapping errors, which give rise to additional WSD system inconsistencies in the experimental setting (Raganato, Camacho-Collados and Navigli 2017). This issue is directly extensible to the training corpora used by supervised WSD systems.

In fact, there is evidence in the literature that results obtained by either supervised or semi-supervised systems are not completely reliable (Postma *et al.* 2016). A key implication of the foregoing divergency issues is that they prevent us from drawing reliable conclusions on different models. In some cases, ostensible improvements may have been obtained as a consequence of the nature of the training corpus, the pre-processing pipeline, or the version of the underlying sense inventory, rather than of the model itself (Raganato, Camacho-Collados and Navigli 2017).

Moreover, because of these divergencies, the current WSD systems tend to report results on a few datasets only, making it difficult to perform a direct quantitative experimental comparison. Hence, advocacy in the literature for a unified WSD evaluation framework, that takes account of these divergencies (Raganato, Camacho-Collados and Navigli 2017). Furthermore, these datasets are primarily available for well-resourced languages such as English, limiting the accessibility and applicability of standardized evaluation measures for languages with fewer

resources, such as Setswana. Consequently, in this study, an evaluation dataset is crafted from scratch to assess the Setswana WSD model.

To sum up, in every scientific study, an assessment must be carried out to evaluate the effectiveness of the proposed solutions. When evaluating WSD models, researchers choose from a variety of evaluation metrics. Some researchers focus solely on accuracy (Bhatia 2022 Paikens 2022; Durgaprasad 2022), while others exclusively report the F1 score (Maurya, Bahadur and Garg 2022), precision, and recall (Orkphol and Yang 2019). Alternatively, some may report any combination of these metrics (Gutiérrez, Vázquez and Montoyo 2017; Saeed *et al.* 2019b) Al-Saiagh 2018; Calvo 2019). However, it is not common to find studies that report ROC and AUC or incorporate all four metrics simultaneously in the literature.

This study opted to evaluate the performance of the proposed Setswana WSD model using all the four metrics ─ accuracy, precision, recall and F1 score. The choice of evaluation metrics was guided by the conventions and practices in the WSD research community, as well as the specific goals and requirements of the Setswana WSD model. Accuracy, precision, recall, and F1 score were selected as the primary evaluation metrics. Accuracy provides an overall assessment of the model's performance, while precision and recall offer a more nuanced view, considering the model's ability to avoid false positives and identify all relevant senses, respectively. The F1 score, being the harmonic mean of precision and recall, gives a balanced measure of the model's performance, especially when dealing with imbalanced sense distributions. In addition, the study also utilizes the extrinsic evaluation method to assess the performance of the WSD within the Setswana-English machine-translation system.

Although other metrics such as receiver operating characteristic (ROC) graphs and the area under the ROC curve (AUC) could have been considered, they were not employed in this study. The decision to focus on accuracy, precision, recall, and F1 score was based on their widespread use and acceptance in the WSD research community, as well as their ability to provide a comprehensive evaluation of the Setswana WSD model's performance.

Furthermore, another challenge in evaluating the Setswana WSD model was the lack of existing benchmark datasets for Setswana. While there are several well-established datasets such as SemCor, SenseEval, and SemEval for English and other resource-rich languages, these datasets are not directly applicable to Setswana due to differences in language structure, sense inventories, and cultural context. Moreover, the issues of format inconsistencies, annotation guidelines, and sense-inventory versions prevalent in these datasets further complicate their

use for Setswana WSD evaluation. To address this obstacle, a dedicated evaluation dataset was developed specifically for evaluating the Setswana WSD model.

3.7 Word Sense Disambiguation Applications

WSD is an intermediate task for several NLP applications. WSD serves a crucial role in improving the performance and understanding of these applications. The embedding of an ability for an application to be sense- and context-aware leads to improved accuracy of natural language understanding applications (Bhattacharjee *et al.* 2020). The following are some of the NLP applications that have improved through WSD. The Table 3.2 below provides a summary of the key features of various types of WSD applications.

3.7.1 Machine translation

Machine translation (MT) is a process of automatically translating text from one language to another without altering the meaning of the text (Wang *et al.* 2022). An MT system takes source text, i.e., Setswana, as input and converts it to target text, i.e., English, as output. A typical MT has a data source which can be a bilingual dictionary or parallel corpus that aligns the source and target languages. In the context of machine translation, WSD is primarily focused on selecting the most appropriate sense of source words during the translation process (Emelin, Titov and Sennrich 2020). This is a challenging task in that languages exhibit a variety of linguistic characteristics such as orthography and morphological typology. The presence of words possessing multiple meanings further complicates the process. The translation output can be impacted by these words if the disambiguation of senses cannot be achieved by the employed translation technique. WSD algorithms assist MT systems in disambiguating words from one language to another; and within the same language, which results in better performance of MT systems (Premjith *et al.* 2019).

3.7.2 Content and sentiment analysis

With increasing information and content curation on the web, content analysis has become important. Content analysis is the process of analysing data to draw valuable conclusions from the data (Drus and Khalid 2019). In the context of sentiment analysis, WSD improves contextual sentiment analysis to determine the sentiment or opinion expressed in text (Birjali, Kasri and Beni-Hssane 2021). Ambiguous words may convey differing sentiments depending on their meaning in the context. WSD helps to identify the correct sense of such words, ensuring that sentiment analysis models accurately capture the intended sentiment. An

exemplary use case is a hotel company seeking to assess feedback provided by its customers in the form of hotel reviews. By utilizing content analysis tools, the corporation can quantify customer satisfaction through analysis of the data derived from the reviews. WSD algorithms enable these tools to analyse data accurately, particularly in the presence of multiple-sense words in the data (Baiju 2022).

3.7.3 Information retrieval (IR)

Information retrieval is the process of retrieving relevant information or documents from a knowledge base (Luan *et al.* 2019). IR systems are one of the predominant applications of WSD algorithms. The WSD is crucial in IR for ensuring the accuracy and relevance of retrieved documents or information. The WSD contributes to the improved query understanding and enhanced retrieval precision. In IR, users submit queries to retrieve relevant documents or information. These queries may contain ambiguous terms with multiple meanings. WSD helps in accurately interpreting the user's query by disambiguating these terms, ensuring that the system retrieves documents relevant to the intended meaning (Lin 2022). Google is the primary example of an IR system used globally. Integrating WSD into this system improves the accuracy of documents or information retrieved for a particular search input. In circumstances in which ambiguity exists in a search query, WSD can assist in resolving it.

3.7.4 Text summarization

Text summarization applications shorten provided information of text or document into a brief version while preserving key and important information from the input text (Awasthi *et al.* 2021). The objective of these applications is to facilitate the understanding and extraction of relevant information from long documentations, allowing users to comprehend the main ideas without having to read the entire document. WSD assists in resolving the ambiguity of words; and ensures that the correct sense of a word is identified and used in the output summary (Rahman and Borah 2023). If not resolved, ambiguous words can lead to inaccurate representations in the summary, misconstruing the information on the original text. The application of WSD techniques ensures that the system can accurately select the appropriate word senses based on the context and intended meaning.

3.7.5 Question answering (QA) systems

Question answering (QA) systems automatically process and respond to user queries by extracting relevant information from a given knowledge source or a large corpus of documents

(Soares and Parreiras 2020). These systems aim to provide accurate and concise answers to user questions, making use of various natural language processing techniques. WSD plays a crucial role in QA systems by enhancing query interpretation, improving answer retrieval precision, facilitating contextual understanding, supporting semantic matching, handling lexical variations, and enabling multilingual question answering. User questions may contain ambiguous terms or words with multiple senses. WSD helps to determine the appropriate sense of these words, thus providing accurate answers. By disambiguating the word senses, QA systems can better understand the user's question and retrieve the information related to the intended meaning of the words in the question. Table 3.2 highlights WSD applications and their key features.

Table 3.2: WSD Applications and Features

Application	**Key Features**
Machine Translation	- Disambiguates senses of source words to select appropriate translation equivalent - Improves translation accuracy and quality
Sentiment Analysis	- Identifies the correct sense of ambiguous words to determine sentiment polarity - Enhances contextual sentiment analysis
Information Retrieval	- Disambiguates query terms to improve retrieval precision and relevance Facilitates accurate interpretation of user queries
Text Summarization	- Resolves ambiguities in input text to generate accurate and coherent summaries - Preserves key information and intended meaning
Question Answering	- Disambiguates ambiguous terms in user questions to provide precise answers - Enables contextual understanding and semantic matching.

In summary, WSD is a crucial intermediate task that plays a vital role in enhancing the performance and understanding of various NLP applications. In the context of developing a

WSD solution for Setswana, the MT was chosen as the primary downstream application based on the scarcity of MT systems for Setswana, and the absence of a Setswana WSD model.

3.8 Related Works

This section focuses on the published related work in the field of WSD and WSD in MT. There are research studies that focus solely on the task of WSD and use intrinsic evaluation methods without practical application ((Scarlini, 2020); (Ruas, Grosky and Aizawa 2019); (Kumar 2019)), i.e., MT, and studies that explore WSD in the context of MT ((Premjith 2019); (Tang, Sennrich and Nivre 2019); (Luan, 2020)). This section delves into the two types of studies to gain a deeper analysis and review of the related work conducted in this research area.

3.8.1 A review of relevant studies on WSD

The selection of the previewed studies as related works for developing a knowledge-based word sense disambiguation (WSD) solution for Setswana was informed by several key factors. Firstly, these studies encompass a diverse range of WSD approaches, including supervised, unsupervised, semi-supervised, and knowledge-based methods, providing a comprehensive overview of the current state-of-the-art in WSD research. Secondly, the studies cover various languages, including resource-scarce languages, which offer valuable insights into the challenges and strategies for developing WSD solutions in low-resource settings, particularly relevant to Setswana. Thirdly, the selected studies employ varying techniques, such as attention mechanisms, graph-based methods, and hybrid approaches, which have demonstrated promising results in improving WSD performance and can potentially be adapted to the Setswana language context.

Durgaprasad, Sunitha and Padmajarani (2022) proposed a novel algorithm to solve WSD for the Telugu language using BERT based on IndicBERT word embeddings; and used cosine similarity between the targeted word and context vectors to assign the appropriate sense. The performance of their algorithm depends on the larger context window size. The above researchers' proposed algorithm reported an average of 65% and 71% accuracy for three- and four-window sizes. Similarly, this study adopts a comparable methodology, utilizing PuoBERTa (Marivate et al. 2023) to generate sentence embeddings and cosine similarity, thus determining the correct sense. However, in contrast to their approach, this study opted to embed the entire context sentence instead of using different context windows. This decision was made

to ensure that the agglutinative and disjunctive nature of the language, along with all sentence components, are fully considered during embedding.

Zhang *et al.* (2022b) employed a hybrid approach combining the self-attention network and knowledge-based method for sense extraction. The self-attention technique uses the transformer mechanism to process parts of the input sequence in the prediction process, thus capturing the context of the target word; while the knowledge-enhanced technique uses WordNet and BabelNet for English as external knowledge sources to enhance the semantic representation of the context words. The proposed method used HowNet senses, and was evaluated on Chinese lexical sample benchmark datasets, reporting an F1 score of 84%. The self-attention mechanism has shown improved results on NLP semantic tasks; however, this method requires computing attention weights for each input word in a sentence, which is computationally problematical, particularly for long sentences.

(Su *et al.* 2022a) employed the same attention mechanism as (Zhang *et al.* 2022b); however, in this study, the attention neural network was applied to a bidirectional long short-term memory (Bi-LSTM) network output to encapsulate the relevant parts of the input sentence of the target word. The Bi-LSTM was used to learn the contextual representation of the target in the sentence, based on the context window, while BERT was used to capture contextual relationships between words in the input sentence. The triangulation of attention neural network, Bi-LSTM and BERT exhibited an effective approach to multilingual WSD that leveraged semantic relationships between senses across different languages, resulting in a model known as the unified sense representation(USR). The study reported the state-of-the-art performance on the English, Italian, French, Spanish and German benchmark datasets with an 84% F1 score.

Another study that used the neural attention network mechanism is that by Cheng, Tong and Yan (2021). The study combines a capsule network with multi-head attention. A capsule network is a neural network architecture that was introduced by Sabour, Frosst and Hinton (2017), designed to specifically handle spatial relationships and hierarchical representations that could not be dealt with by traditional neural networks. In their study, the capsule network was used to capture complex relationships between words in a target sentence using Bi-LSTM in the first encoding; while the multi-head attention was used to attend to two different aspects of the sentence at the same time to compute the final word sense probabilities. The authors report that the proposed approach achieved an F1 score of 88% on English adjectives.

Attention neural networks have been used in other domains such as the biomedical domain. Zhang *et al.* (2022a) extended the use of attention neural networks to improve the disambiguation accuracy of biomedical words. The study used average asymmetric convolutional neural networks (Av-ACNN) and Bi-LSTM networks to extract features. To determine the semantic category of the biomedical word, the softmax function of the attention network was applied. The experimental results show that the proposed method achieved a disambiguation accuracy of 91.38%. Saeed, Nawab and Stevenson (2021) performed an investigation into the feasibility of deep learning methods for Urdu language WSD. The study utilized convolutional neural networks (CNN), long short-term memory (LSTM), Bi-LSTM and pre-trained word embeddings to conduct WSD. The current investigation found that LSTM and Bi-LSTM performed significantly better than CNN.

The superior performance of LSTM and Bi-LSTM is attributed to their ability to capture long-term dependencies and context information in the input data. This is enabled by the recurrent neural network (RNN) architecture which processes the sequential information of the text data. The study reports 72% accuracy and a 60% F1 score with recommendations on the importance of developing sufficient annotated data to improve the accuracy of these models.

The studies mentioned above propose the use of various deep-learning techniques such as the attention neural network, LSTM, CNN, RNN, capsule network, Bi-LSTM and BERT embeddings to model the context of the target word, and perform sense disambiguation. These studies report achieving state-of-the-art results on various benchmark datasets.

However, these methods rely heavily on larger amounts of annotated data to learn effective representations of words, which is time-consuming and expensive, creating a data-sparsity challenge, particularly for languages with fewer available resources, such as Setswana. In addition, these techniques are extremely abstract to interpret, making it problematic to determine the model's prediction processes. These processes are vital in applications such as biomedicine, in which the model's prediction decision-making process transparency is essential in ensuring the safety of the patients.

Supervised disambiguation methods and approaches require a large amount of data, and have been proven to deliver greater performance than other word-sense disambiguation methods. One major limitation with these methods is their requirement for large-scale, sense-tagged corpora, which is significantly costly in time, funds, and efforts. Unsupervised methods and approaches have been proposed to eliminate this requirement; however, the results have not

been satisfactory due to the semantic occurrence probability of ambiguous words unavailability in unannotated data.

To address this challenge, Kim and Kwon (2021) proposed an approach that uses prior probability estimation for Korean. The approach employs a graph-based approach in which lexical relationships between words and their corresponding senses are mapped on a graph. The method then calculates prior probability for each sense, given a target word. The approach incorporates contextual information obtained from surrounding words to refine probability estimates. The method accomplished an accuracy of 94% on selected Korean ambiguous words.

Kazemi and Karshenas (2021) presented a novel fuzzy-based method to automatically induce word senses. The selection of this method was in terms of the ability of fuzzy logic to represent uncertainty in semantic applications. The approach involves assigning fuzzy membership values to each sense of a target word based on the context in which it appears. The generated memberships were then used to estimate the degree of ambiguity in each word. The appropriate sense was selected based on the ambiguity score. The method attained a 66% F1 score for English.

Alternatively, Angelina and Loukachevitch (2020) proposed an algorithm for the automatic generation and labelling of training data for Russian. The authors used support vector machines (SVMs) and decision trees (DT). The study used the bag-of-words model as the input representation for training the SVM. The SVM was thereafter trained on input data, and used for the prediction of the correct sense. The authors further incorporated manual linguistic rules based on the context in which the ambiguous words occurred; and constructed a decision tree based on the linguistic rules. The study found that their proposed method achieved an accuracy of 71% on the Russian national corpus. A takeaway from this study, especially for resource-scarce languages is that the use of automatically generated corpora can be effective in improving the performance of WSD for languages with limited resources.

The naive Bayes classifier is one of the prominent classification techniques utilized in the task of WSD. Maurya, Bahadur and Garg (2022) perform WSD for English-Sanskrit using naïve Bayesian classification (NBC). Through data pre-processing and feature extraction, the authors trained the classifier using two different datasets. The classifier achieved an F1 score of 85%. Pal Singh and Kumar (2018) used NBC for the Punjabi language WSD and achieved an accuracy of 88%.

While NBC has proven to be effective for WSD, its presumption of the independence between the features given the class labels may not hold true for all data types. As a result, the context words may be independent, thus affecting the meaning of the target word which can lead to misclassification of certain words. Abid *et al.* (2018) explored the performance of NBC, DT, SV, Random Forest (RF), and k-Nearest Neighbors (KNN) on Urdu WSD. The authors used the multi-criteria decision-making method to identify the best-performing technique and found that SVM performs better than the other four techniques: SVM has the ability to handle high dimensional data; and has good generalization performance, which boosts its effectiveness on WSD tasks. On the contrary, Pal *et al.* (2021) conducted the same study as Abid *et al.* (2018) for Bengali; and found that RF performed better than NBC, DT, SV, and KNN, with 92% accuracy. The difference in results may be attributed to languages differences. Abid et al. (2018) focused on Urdu, while Pal et al. (2021) studied Bengali. Different languages have different syntactic and semantic structures, which can affect the performance of various algorithms. The characteristics of the Urdu language might make SVM particularly effective, while Bengali might present features that favour Random Forest. This can be seen with English and Setswana, where the same algorithms might yield different performance outcomes due to variations in linguistic structures and feature distributions. English, with its extensive and varied vocabulary and syntactic flexibility, might benefit from models that can handle high-dimensional and complex feature spaces well, such as SVM. In contrast, Setswana, with its rich morphology, disjunctive orthography and different syntactic patterns, might work well with models that can effectively manage and aggregate diverse features.

The Lesk algorithm is regarded as one of the pioneering techniques in the field of WSD, and has been widely employed in various WSD systems. (Kwon, Oh and Ko 2021) used context-selection techniques to calculate semantic similarity measure between the target word and its possible senses. The Lesk algorithm was used to identify the most probable sense context for the target word by examining the overlap between the context and the definitions of the word senses. This method assists in ensuring that the selected context is semantically related to the target word. To disambiguate, the sense of the target was chosen based on the highest similarity score for the selected context. The approach was evaluated on SenseEval-02, SenseEval-03, SemEval-07, SemEval-13, and SemEval-15 English benchmark datasets; and recorded an F1 score of 72%.

Graph-based methods have been extensively employed to address the WSD problem. AlMousa, Benlamri and Khoury (2022) developed a WSD algorithm, sequential contextual similarity

matrix multiplication, that combines semantic similarity, heuristic knowledge, and document context for disambiguation. The method extracts semantic features of the target ambiguous word from the WordNet knowledge graph and uses the decision tree (DT) algorithm to classify the target word with its appropriate sense. The proposed algorithm outperformed other algorithms when disambiguating English nouns. The algorithm was evaluated on SenseEval-02, SenseEval-03, SemEval-07, SemEval-13, and SemEval-15 English benchmark datasets; and achieved an F1 score of 70%.

The WSD task is highly dependent on the availability of contextual information pertaining to the target word; and it is widely acknowledged that the successful resolution of WSD is almost unachievable without the incorporation of such information. Abdalgader and Al Shibli (2021) proposed a context expansion approach for graph-based WSD. The proposed method expands contextual information of the target word by adding semantically related words to the target word context. The expanded context information was to construct a graph used to compute similarity between the target word and the candidate senses. The sense with the highest similarity score is selected as the appropriate sense of the target word. The study used the random walks (RW) algorithm to compute distributional similarity measures. The proposed method was evaluated on English benchmark datasets and obtained an accuracy of 84%. This approach requires tuning several hyperparameters, which can be time-consuming, and can require extensive experimentation.

Further exploration and optimization of hyperparameters are essential in enhancing the effectiveness and efficiency of the method. Jain and Lobiyal (2022) used ConceptNet and cooperative game theory on the fuzzy Hindi WordNet for WSD. Using the ConceptNet, the authors constructed a multilingual knowledge graph referred to as fuzzy Hindi WordNet; and applied cooperative game theory to disambiguate Hindi word senses. The various word senses were treated as a coalition of players, in which the winning player was the appropriate sense of the target word approach to addressing the WSD problem. Bhatia, Kumar and Khan (2022) present an approach to the problem Hindi WSD using a genetic algorithm to optimize the performance of a Hindi machine-learning language model. In this study, a genetic algorithm based method was used to optimize a Hindi DT classifier. Specifically, the authors optimized the maximum depth and the minimum number of instances required to split a node parameter set of the DT. The genetic algorithm generates candidate solutions that correspond to the set of parameters for the DT, subsequently improving the DT performance.

However, this approach is computationally expensive due to the creation of a large population of candidate solutions iterating through multiple generations for optimization. The proposed method attained accuracy of 80%, which indicated accuracy improvement of the existing work by 8%. Abderrahim and Abderrahim (2022) conducted an experiment with two graph-based methods, conceptual density (CD) and random walk (RW) to resolve WSD in Arabic; and to enhance the performance of the Arabic information retrieval system. The authors used a medium-sized corpus to evaluate the proposed system; RW improved the performance of the system with a 16% precision-improvement rate.

Zhang *et al.* (2021) proposed an approach to WSD that leveraged the capabilities of convolutional networks and graph-based methods, and presented a graph convolutional network (GCN). The GCN neural network operates on a graph structure. As with conventional graphs, the nodes represent entities, and the edges represent relationships between the nodes. The information propagation of a graph-based method such as the RW and PageRank is based on predefined rules; whereas the GCN learns node representations by propagating information on the graph and aggregating information from neighbouring nodes.

Compared with conventional graph-based approaches, GCNs have been shown to be more effective at capturing complex relationships and patterns within the graph, leading to an improved performance in WSD. In addition, GCNs learn node representation without pre-defined rules, making them more efficient and adaptable. However, GCNs can be prone to overfitting on small datasets, which can lead to poor generalization performance on unseen data. The GCN method performed significantly better on the Chinese WSD compared with other graph-based and neural network methods, achieving an accuracy of 82%.

In contrast to supervised approaches that require annotated training datasets, unsupervised approaches do not require any annotated dataset. Rahman and Borah (2022) proposed an unsupervised clustering algorithm to identify the appropriate sense of an ambiguous word. The clustering technique groups words based on the contexts in which they occur, and applies the disambiguation algorithm to disambiguate senses from the clusters. The algorithm was tested on SenseEval-02, SenseEval-03, SemEval-07, and SemEval-13 English benchmark datasets, achieving an F1 score of 79%. Park, Shin and Lee (2022) used the same clustering technique on their BERT sequence labelling programme; and obtained a 97% F1 score on the Korean dataset. The authors announced that the proposed method is useful for discovering compressed sense vocabulary without the need for utilizing a manually constructed thesaurus.

To leverage both annotated and unannotated data, researchers (Geleta 2024; (Garigliotti 2019); Duarte 2021) started exploring semi-supervised approaches. Semi-supervised methods aim to overcome the challenge of limited labelled datasets by incorporating additional information from unlabelled data. One common approach is the use of multiple classifiers as opposed to one classifier. Duarte *et al.* (2021), on the other hand, used a combination of semi-supervised learning and neural word representations, considering various graph-based semi-supervised algorithms with features generated by word embeddings from Word2Vec, FastText, GloVe, BERT, and ELECTRA models. The authors tested combinations of word-embedding models, similarity measures for graph construction, and semi-supervised classification algorithms to disambiguate classical lexical sample WSD datasets. The experiment results indicated that the semi-supervised technique achieved competitive outcomes compared with supervised methods, with highest F score of 88% obtained by the ELECTRA model.

In addition to the conventional methods used in word sense disambiguation (WSD), researchers are actively exploring novel approaches to addressing the challenges of WSD. These innovative methods aim to enhance the accuracy and effectiveness of WSD beyond the traditional techniques. Su *et al.* (2022b) presented a novel approach to addressing WSD for rare and zero-shot words. Rare words are words that have limited occurrences in a given corpus or dataset: these words occur infrequently. Zero-shot words, on the other hand, are words that have not been observed or encountered during the training phase of a machine-learning model.

These words are non-existent in the model, and lack any pre-learned associations or representations in the model. The authors propose a Z-Reweighting technique that leverages the data from high-frequency words in the corpus to improve the disambiguation of low-frequency words or unseen words. The employed method uses a classification technique which is based on the concept that high-frequency words carry relevant semantic information that can be used to transfer the correct sense for rare words. The Z-Reweighting method assigns weights to the senses of high-frequency words, based on their discriminative power in disambiguating rare words. Thereafter, these weights are then used to reweight the features of the high-frequency words in the classification model, giving more importance to the informative senses. The experimental results demonstrated the effectiveness of Z-Reweighting in improving the disambiguation results for rare and unseen words. The approach outperforms several baseline methods, with a 78.9% F1 score on benchmark datasets, illustrating its potential in handling challenging WSD problems, such as the knowledge-acquisition bottleneck.

Contrary to the traditional practice of using a fixed window size for WSD, Li and Suzuki (2021) proposed a novel technique with an adaptive context window length in document representation. The authors' approach focuses on enhancing the accuracy and granularity of word sense disambiguation by incorporating adaptive and hybrid techniques. This enhancement aims to improve the disambiguation process by considering the contextual information surrounding the target word, adapting the disambiguation strategy accordingly. The adaptation ability is incorporated into the method through the use of dynamic context windows based on the specific characteristics of the ambiguous target word. Bag-of-senses instead of bag-of-words was introduced, and used as a disambiguation technique to generate word senses, as opposed to words. The method was tested on standard English datasets and obtained 82% accuracy. Hybridization or triangulation of various methods is common in research to harness strengths of two or more methods.

Heo, Kang and Seo (2020) proposed a scalable process that can address a great number of senses occurring in various domains using hybrid sense classification for large-scale WSD. In this method, disambiguation is processed in accordance with parts of speech in the context and hybrid sense prediction, that classifies the less frequently used senses separately. The hybrid method first utilizes a knowledge-based approach, in which it exploits the information from lexical resources such as WordNet and BabelNet to create the first set of sense candidates for a target ambiguous word. The sense candidates were then expanded using a corpus-based approach, which considers the context of the target ambiguous word in a large-scale corpus. The method employs various statistical measures, such as mutual information and distributional similarity, to rank and select the most appropriate sense for the target word. The hybrid sense classification method was tested on Oxford datasets, and attained competitive results of 94.33% accuracy on adjectives.

Zhong and Wang (2020) propose the use of multiple kernel support vector machines (MK-SVMs) to perform WSD. While conventional SVMs use a one-kernel function, the proposed method uses word embeddings, context, and semantic features to build multiple kernels which are combined to train and test the SVM models. The authors evaluated their approach on several benchmark datasets, and obtained a micro-averaged F1 score of 89%. The obtained results demonstrate an improved performance of the use of multiple kernels as opposed to using a single kernel.

The related works reviewed in this section provide a comprehensive overview of the various approaches and techniques employed in WSD research for various languages. The studies cover a wide range of methods, including supervised approaches using deep learning architectures such as attention neural networks, LSTM, CNN, and BERT embeddings, unsupervised methods based on clustering and graph-based techniques, semi-supervised approaches leveraging both labelled and unlabelled data, and knowledge-based methods employing lexical resources and semantic networks.

The studies demonstrate the effectiveness of incorporating linguistic knowledge and contextual information in improving WSD performance across different languages. Techniques such as context expansion, adaptive context window length, and leveraging semantic relationships between senses have shown promising results in enhancing disambiguation accuracy. Graph-based methods, particularly those using graph convolutional networks (GCNs), have emerged as ideal tools for capturing complex relationships and patterns within semantic networks, leading to improved WSD performance.

However, the studies also highlight the challenges posed by resource-scarce languages, such as limited availability of annotated corpora and the need for efficient utilization of available resources. In the context of developing a knowledge-based WSD solution for Setswana, these related works offer valuable insights and guidance. The success of knowledge-based approaches in other languages suggests the potential for leveraging lexical resources for resource-scarce languages.

3.8.2 A review of relevant studies on WSD in the context of machine translation

The studies reviewed in this section highlight the importance of WSD in improving the performance MT systems across various languages. The integration of WSD algorithms into MT systems has been shown to enhance translation accuracy by enabling the models to better understand the context and meaning of words, thus reducing ambiguity.

Machine-translation (MT) systems have been developed to automate the translation process and facilitate cross-lingual communication. To improve the performance of MT systems, in some instances, researchers embed WSD algorithms. Cohen, Zhong and Li (2022) used a semantic graph for word sense disambiguation in MT. The authors' translation model is based on the n-gram model considering the neighbouring words in terms of spatial proximity. A

semantic graph was constructed; and the conditional probability of a word given all words in a text was designed to be dependent on the keywords, in contrast to the dependency on the closest neighbours of words for the n-gram model. The model parameters were obtained from bilingual corpora data and human translators' feedback. The model allows for optimal statistical decisions with the best disambiguation of a word to be translated. The model was tested on the medical dataset, and literary fiction dataset, and obtained the highest accuracy of 98%.

Chakrawarti, Bansal and Bansal (2022) proposed a WSD approach with a half-breed technique for translating Hindi poems into English. The authors proposed a trie-based dictionary for storing Hindi words and expressions. Syntactic guidelines of Hindi language rules, and tokenized words for distinguishing accurate words in the dictionary were applied. A half-breed MT procedure based on rule-based MT (RBMT) and SMT for WSD was developed. The proposed method improves the semantic and syntactic accuracy of the machine-translation framework as a result of the WSD feature. The proposed system obtained 96.85% on idioms disambiguation and translation.

Zhang, Hauer and Kondrak (2022) presented a method for improving HowNet-based Chinese WSD by combining monolingual contextual information from a pre-trained neural language model with bilingual information obtained from MT and sense translations. The results of the evaluation experiment demonstrated that this new method achieves a state-of-the-art outcome for unsupervised Chinese WSD. The paper also includes a comparative study to evaluate various translation methods and their combinations. The study reports an overall micro-F1 score of 56%, surpassing state-of-the-art systems.

Chauhan, Saxena and Daniel (2022b) propose a new approach to unsupervised neural machine translation that utilizes cross-lingual sense embedding and filtered back-translation to improve translation accuracy for morphologically rich and endangered Indic languages. Cross-lingual sense embedding is a technique used in this approach to improve translation accuracy; this involves mapping the meaning of words in one language to their corresponding meanings in another language. Such is achieved by creating a shared semantic space in which words from different languages are represented as points that are close together if they have similar meanings. By using cross-lingual sense embedding, the model can better understand the meaning of words in both languages and make more accurate translations. This approach improves machine-translation accuracy by capturing the nuances of language that may be lost in traditional word-to-word translations. While the authors show that their approach

outperforms several baseline models in terms of the BLEU score, they also note that their approach requires a large number of computational resources, which may make it difficult to implement in low-resource settings.

Campolungo *et al.* (2022a) presented a manually curated evaluation benchmark that enables a comprehensive study of semantic biases in MT of nominal and verbal words in English, Chinese, German, Italian, Russian and Spanish. The study is based on measuring and understanding semantic biases in NMT that go beyond simple accuracy, and provide novel metrics that summarize the extent of bias in NMT models. The bias is mostly due to the data that the model was trained on. According to the authors, statistics of the annotations suggest that synsets' lexicalizations cannot be used interchangeably when dealing with translations, as their choice depends heavily on the context. This is the case for WSD models. Their model provides an opportunity to identify and mitigate semantic biases in NMT models, leading to more accurate and contextually appropriate translations.

Campolungo *et al.* (2022b) proposed a novel approach to create high-precision sense-annotated parallel corpora. This approach would leverage a multilingual WSD system to tag parallel sentences and refine its predictions by means of cross-lingual word alignments and information from a multilingual knowledge base. The authors proposed this approach to mitigate NMT struggling to disambiguate polysemous words without lapsing into their most frequently occurring senses in the training corpus, NMT models relying solely on the input context to resolve ambiguity. The model disambiguates words to the sense they most frequently encountered during training, regardless of whether the sentence does not provide sufficient context to identify the correct sense. Baseline NMT models were fine-tuned on the developed sense-tagged corpora using a designed loss function, allowing the injection of word-level semantics into the architecture. This approach enables the exploitation of sense annotations during training without introducing any additional requirement at inference time. The authors report that their approach improves translation accuracy and ameliorates the most frequent sense bias.

Chauhan, Saxena and Daniel (2022a) proposed improved unsupervised statistical machine translation through unsupervised WSD for a low-resource Indic language. The authors incorporated WSD into the decoding phase of unsupervised statistical machine translation which improved the translation quality. The improvement was a result of the WSD enabling the MT system to better understand the context and meaning of words, leading to more accurate

translation, unaligned and improperly sensed words presenting difficulties in MT. The authors note that their approach may not be suitable for highly inflected languages or languages with complex syntax. Setswana shares some linguistic characteristics with Indic languages, although there are also significant differences. Some characteristics that Setswana shares with Indic languages include agglutinative morphology, complex verb conjugation, and their polysynthetic nature. Both Setswana and Indic languages exhibit agglutinative morphology, in which words are formed by adding affixes to a root or stem. This results in complex word structures with multiple morphemes conveying different grammatical or semantic information. In terms of complex verb conjugation, both Setswana and some Indic languages such as Sanskrit, Bengali, Gujarati, and Marathi have complex systems of verb conjugation, in which verbs undergo various changes to indicate tense, aspect, mood, and agreement with the subject and object. These languages, including Setswana, are considered polysynthetic, meaning that they have the ability to form complex words or expressions through the agglutination of multiple morphemes. While the method proposed by Chauhan, Saxena and Daniel (2022a) could potentially be applied to Setswana due to shared linguistic properties, the lack of large unannotated datasets for Setswana makes it problematic for unsupervised methods to achieve optimal performance. Therefore, adaptations and further research are necessary to effectively leverage such techniques for Setswana machine translation, given the limited availability of both unannotated and annotated data.

Chingamtotattil and Gopikakumari (2022) present a neural machine translation (NMT) system for translating Sanskrit to Malayalam using morphology and evolutionary word sense disambiguation. The authors utilize an attention-based mechanism to develop the machine-translation system, which allows the model to selectively attend to various parts of the input sentence when generating each word of the output sentence. The authors also use sequential deep-learning approaches such as a recurrent neural network (RNN), a gated recurrent unit (GRU), a long, short-term memory (LSTM), and a bi-directional LSTM (BLSTM) to analyse the tagged data. The attention-based mechanism improves the accuracy of the machine-translation system by allowing the model to focus on specific parts of the input sentence when generating each word of the output sentence. This is in contrast to traditional machine-translation models that generate each word of the output sentence based solely on a fixed-length vector representation of the entire input sentence. The suggested common character-word embedding-based NMT model, which incorporates these techniques, gives a BLEU score of 38.58.

Chauhan *et al.* (2022) propose a method that combines context-aware word translation with a denoising autoencoder to improve the accuracy of machine translation. The denoising autoencoder is used to learn representations of the source and target languages, while the context-aware word translation model is used to align the representations and generate translations. A denoising autoencoder is a type of neural network that is used to remove noise from data. In the context of unsupervised machine translation, a denoising autoencoder can be used to improve the quality of translations by removing noise from the input data. The denoising autoencoder works by taking a noisy input and generating a clean output. The network is trained on pairs of both noisy and clean data; thus it learns to recognize patterns in the input data, thereafter generating accurate outputs. Once trained, the denoising autoencoder can be used to remove noise from the input data before it is fed into the machine-translation model. By using a denoising autoencoder in conjunction with unsupervised machine translation, the resulting translations are more accurate, having fewer errors. This is because the denoising autoencoder helps to remove noise from the input data, which can improve the quality of translations generated by the machine-translation model. The authors of the paper conducted experiments on several language pairs, including English-Hindi and Hindi-English, and showed that their approach outperformed existing unsupervised machine-translation methods with a 20.35 BLEU score.

Vu *et al.* (2020) propose a solution to the lack of openly available bilingual language resources for low-resource languages such as Korean. The authors built an open extensive parallel corpus known as UPC, consisting of two large parallel open corpora for training Korean-English and Korean-Vietnamese MT models. The data was collected from multilingual magazines and online dictionaries, focusing on issues related to everyday life such as economics, education, and religion. The UPC contains up to 969 000 sentence pairs in Korean-English and more than 412 000 sentence pairs in Korean-Vietnamese. The authors also conducted experiments to evaluate the quality of the MT models trained on the UPC, and obtained a BLEU score of 24.21 for SMT and 27.45 for NMT.

The related works reviewed in this section present a diverse range of approaches for incorporating WSD into MT, including semantic graphs (Cohen, Zhong and Li 2022), hybrid techniques combining rule-based and statistical methods (Chakrawarti, Bansal and Bansal 2022), cross-lingual sense embeddings (Chauhan, Saxena and Daniel 2022), and attention-based mechanisms (Chingamtotattil and Gopikakumari 2022). These studies demonstrate the effectiveness of WSD in improving translation quality across various language pairs, such as

Hindi-English, Sanskrit-Malayalam, and Korean-English. Furthermore, the related works emphasize the importance of considering semantic biases (Campolungo *et al*. 2022a), reflecting the most frequent sense bias (Campolungo et al. 2022b) in MT models. These findings underscore the need for a comprehensive evaluation framework that goes beyond simple accuracy metrics and accounts for the contextual appropriateness of translations.

However, the applicability of these methods to Setswana may be problematic due to the unique linguistic characteristics and resource scarcity of the language. While Setswana shares some features with the languages addressed in the related works, i.e., agglutinative morphology, complex verb conjugation, and polysynthetic nature, the limited availability of both unannotated and annotated datasets for Setswana is equally a hindrance to unsupervised and supervised WSD approaches (Chauhan, Saxena and Daniel 2022a).

3.9 Bibliometric-analysis for Word Sense Disambiguation

Bibliometric analysis is a quantitative analysis conducted to examine the patterns and trends within a specific research field. The analysis aims to uncover the most active research areas, emerging topics, and declining themes within the field over time. Bibliometric analysis provides an examination of citation-based metrics; the bibliometric analysis assesses the impact and influence of individual researchers, research groups, institutions, and countries (Iqbal *et al.* 2021). Furthermore, patterns of collaboration among researchers, institutions, and countries are examined by analysing co-authorship networks, citation networks, and co-citation networks. This helps to identify key players, research communities, and knowledge flows within the research field. The bibliometric analysis further determines the most extensively cited publications and authors within the research field, which are used to identify seminal works and leading experts. In this section, the bibliometric analysis is conducted within the WSD research area.

3.9.1 Bibliometric-analysis procedure

This bibliometric-analysis procedure employs quantitative techniques to analyse the bibliometric data and the intellectual structure of the field of study by scrutinizing relationships between various research components (Donthu *et al.* 2021). The analysis enables an illustration of the contribution of disciplines, identification of connections and silos, and identification of trends and research gaps. For this analysis, a process was followed to determine the search

terms, select an appropriate database, establish selection criteria for the search, select software for analysis, and to analyse the results. The steps followed are presented below.

Step 1: Determining the search terms	⟶	Search Term**: "word sense disambiguation", "word sense disambiguation in/for machine translation", "word sense disambiguation in the context of machine translation"**
Step 2: Selection of database	⟶	Database: Scopus
Step 3: Selection criteria for search	⟶	**Year**: 2004 – 2023 **Document Types**: Articles, journals, proceedings papers, book chapters, review articles
Step 4: Selection of software	⟶	**Microsoft Excel**: Tables and visualizations **VOSViewer**: Bibliometric network analysis and mapping
Step 5: Analysis and results	⟶	**Performance Analysis**: Publication and citation trends by country, contribution by affiliation, top journals, frequently cited documents **Science Mapping**: Co-keyword analysis, co-author analysis, co-country analysis

Following a systematic approach and steps ensures analysis validity and reliability. The systematic procedure enhances the consistency and reproducibility of the analysis while promoting transparency and facilitating potential replication by others. By providing clear guidelines and steps, analysis streamlines the research process, reducing errors.

3.9.2 Co-citation analysis

Co-citation analysis explores the relationships and patterns within a research field by examining the frequency that countries, authors, or keyword entities are cited together. This section provides an analysis for these entities to gain insights into the intellectual structure, key players, and emerging trends within the domain of WSD research.

3.9.2.1 Country co-citation analysis

The country co-citation analysis provides insights into the international landscape of WSD research, highlighting the key players and the collaborative networks that have formed over time. By identifying closely related clusters of countries, researchers can better understand the patterns of scientific collaboration and knowledge exchange within the WSD community. This understanding can promote knowledge sharing, and address the challenges faced by WSD research on a global and context-specific scale (Chinchilla-Rodríguez *et al.* 2018). In addition, the country co-citation analysis can shed light on the potential disparities in research output and collaboration among different regions and nations. This information can be used to identify countries or regions that may benefit from increased support, resources, or collaborative opportunities, thus enhancing their contributions to WSD research.

To analyse the collaborative relationships among countries in the field of WSD research, a country co-citation analysis was performed using the association strength method (Bevilacqua *et al.* 2021). A minimum threshold for the number of research outputs from each country was set to 10. This decision aimed to concentrate our study on countries with a significant research output, allowing for more robust data suitable for analysis and comparison. Therefore, only countries with a minimum of ten documents were included, leading to a total of 61 countries divided into clusters, which can be seen in Figure 3.5. These clusters represent sets of countries that exhibit strong research ties and tend to collaborate more frequently with one another (Van Eck and Waltman 2014). The proximity of countries within the visualization indicates the degree of co-occurrence in the cited references of WSD research articles. Countries that are positioned closer to one another in the visualization have a higher likelihood of being cited

together in the same research papers, suggesting a stronger research connection and potential for collaborative endeavours (Perianes-Rodriguez et al., 2016).

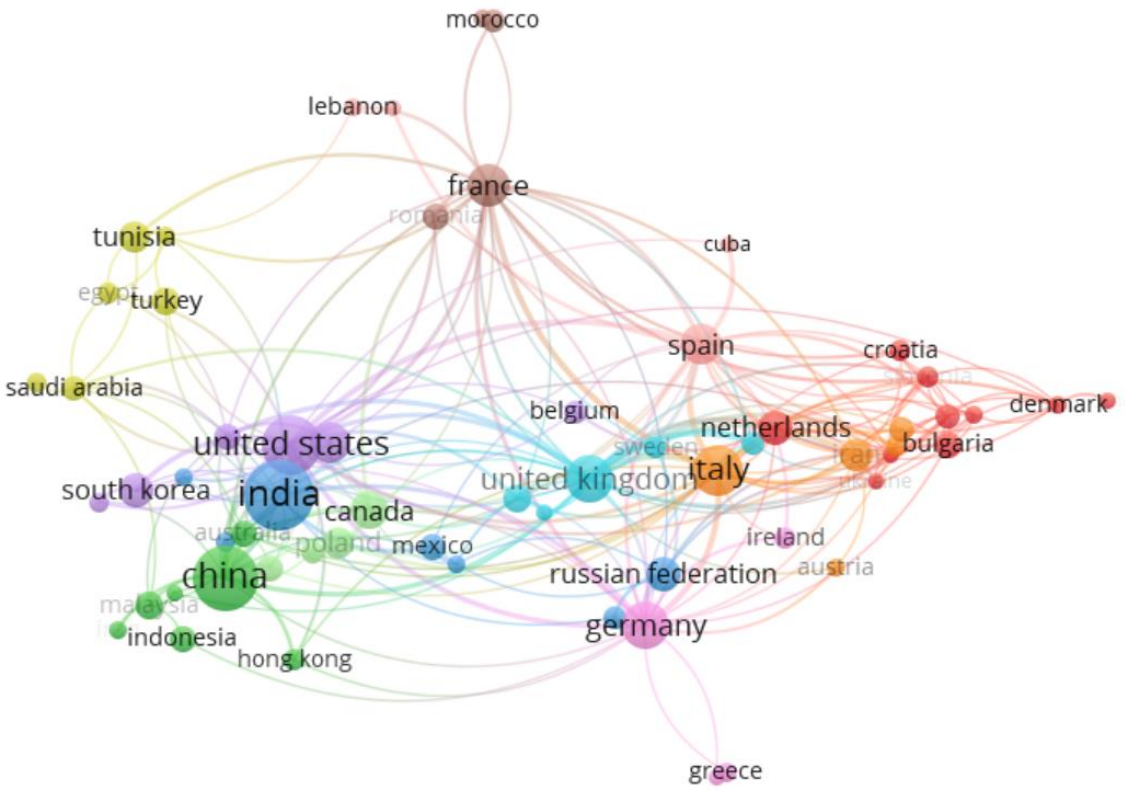


Figure 3.5: Co-country network for "word sense disambiguation"

The size of each node (circle) indicates the number of documents associated with a country (Van Eck and Waltman 2014). The lines represent co-occurrence between two countries, and appear when countries co-occur at least five times. The clusters and positions of the countries suggest that primarily linguistic or regional groupings have formed. For instance, the red cluster predominantly features European countries; the green cluster includes numerous Asian countries; and the top-left corner of the lime-green cluster includes Middle East or North African nations.

From the foregoing co-country analysis, the sub-Saharan African (SSA) countries (including South Africa) cluster is not represented. Thus, the SSA countries cluster suffers WSD knowledge divide, and so lacks contribution to global knowledge in this field of study. Specifically, this is with regard to the lack of contribution to knowledge through WSD solutions provisioning for over 2000 SSA languages. This is a gap that this study contributes to filling.

3.9.2.2 Author co-citation analysis

The co-author analysis offers insights into the role individual researchers play in shaping the direction and progress of WSD research. This analysis helps identify influential authors who have made significant contributions to the field and those who serve as bridges between different research communities by examining the size of the nodes and the density of connections within the network.

A co-author analysis was conducted to investigate the collaborative networks among researchers. The analysis employed the association-strength method to generate clusters of closely related authors (Van Eck and Waltman 2014). To ensure a focus on authors with a substantial contribution to the field, only those with a minimum of five published documents were included in the analysis, resulting in a total of 150 authors being considered.

Figure 3.6 presents the results of the co-author analysis, depicting a visualization of the interconnected clusters of authors. Of the 150 authors included in the analysis, 82 are featured within the central network, which comprises six distinct clusters represented by the colours green, blue, red, yellow, pink, and brown. These clusters represent sets of authors who have strong research ties and who tend to collaborate more frequently with one another.

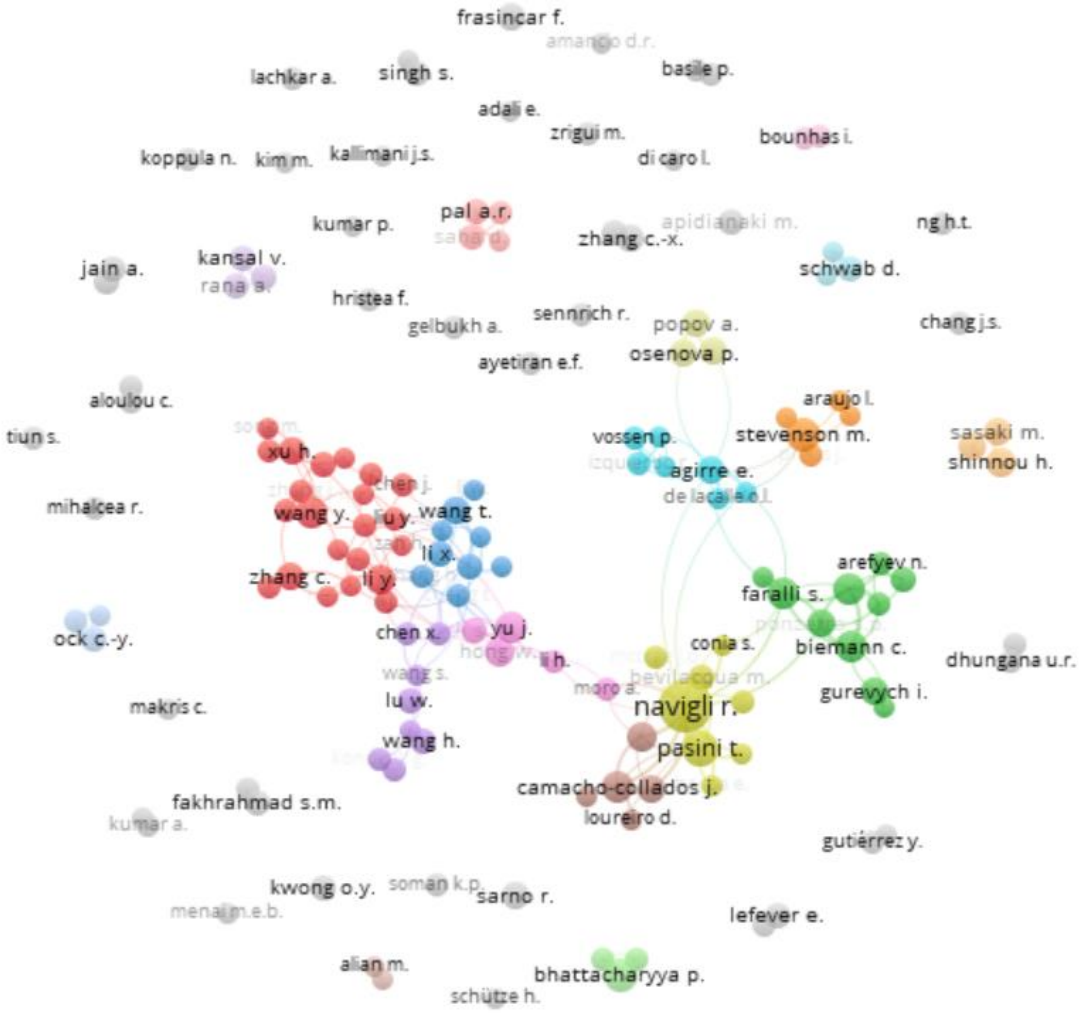


Figure 3.6: Co-author network for “word sense disambiguation”

The proximity of authors within the visualization indicates the degree of co-occurrence in the published WSD literature. Authors positioned closer to one another have a higher likelihood of having co-authored papers together, suggesting a stronger research connection and potential for collaborative work (Van Eck and Waltman 2014). The size of each node or circle in the visualization corresponds to the number of documents associated with a particular author, providing a visual representation of their relative contribution to the field.

The lines connecting the nodes represent co-occurrence between two authors, appearing when they have co-authored at least one paper together. The presence of these lines highlights the collaborative relationships within the WSD research community, allowing for the identification of key players and the formation of research teams.

In this analysis, Roberto Navigli's prominent position with the largest node size and dense connections indicates a high scientific impact and extensive co-citation relationships with other authors or research entities within the WSD research landscape. This suggests that Roberto Navigli's work has been highly influential and extensively cited within the research community, leading to a strong presence and collaboration network in the field of study ((Ajibade, 2019), Sweileh, 2020)).

The absence of cited authors from African and sub-Saharan African (SSA) countries within the network highlights a troubling disparity in representation within the WSD research landscape. This observation suggests a lack of visibility or engagement of researchers from these regions in the broader discourse surrounding WSD. This underrepresentation may reflect various challenges, including limited research infrastructure, and language resources (Nekoto *et al.* 2020). Addressing this gap is crucial not only for promoting diversity and inclusivity in academic research but also for ensuring a comprehensive understanding of WSD across diverse linguistic and cultural contexts.

3.9.2.3 Keyword co-citation analysis

Keyword co-citation analysis provides a more nuanced understanding of the diverse approaches, methodologies, and applications associated with WSD through the identification of keyword and clusters within the network map allowing for the delineation of specific subdomains or research themes. This information can be used to guide future research efforts, facilitate the identification of knowledge gaps, and promote interdisciplinary collaboration among researchers working on related topics (Radhakrishnan *et al.* 2017).

To gain understanding of the key themes and topics within the field of WSD research, an author-keyword analysis was conducted. The analysis identified a total of 2,839 unique keywords used by authors in their published works. Figure 3.7 presents the top 20 keywords sorted by their total occurrences, providing an overview of the most frequently used terms and concepts in WSD literature.

Selected	Keyword	Occurrences	Total link strength
☑	word sense disambiguation	717	1124
☑	natural language processing	179	360
☑	wordnet	141	297
☑	machine learning	49	124
☑	wsd	60	113
☑	sentiment analysis	49	105
☑	information retrieval	39	99
☑	machine translation	47	90
☑	semantic similarity	46	79
☑	text mining	30	69
☑	word sense disambiguation (wsd)	35	68
☑	word embeddings	29	67
☑	ontology	31	66
☑	ambiguity	22	65
☑	clustering	21	62
☑	word embedding	28	62
☑	word sense induction	28	62
☑	lexical semantics	23	58
☑	nlp	24	57
☑	classification	20	55

Figure 3.7: Top 20 author keywords for documents on “word sense disambiguation”

As indicated in Figure 3.7, “word sense disambiguation” and “natural language processing” have the highest occurrences and total link strengths on the keyword co-citation analysis because they represent pivotal concepts within the research landscape being analysed, reflecting their central role and importance within the WSD research landscape. The prominence of these two keywords highlights the fundamental relationship between WSD and the broader field of NLP.

To further investigate the relationships and co-occurrence patterns among the author keywords, a co-occurrence analysis was performed; and the results were visually represented using a network map, as illustrated in Figure 3.8. The network map displays the interconnectedness of the author keywords, with each node or circle representing a specific term. The size of the nodes corresponds to the number of documents that have the corresponding term listed in their author keywords, offering a visual indication of the relative importance and prevalence of each keyword within the WSD research landscape.

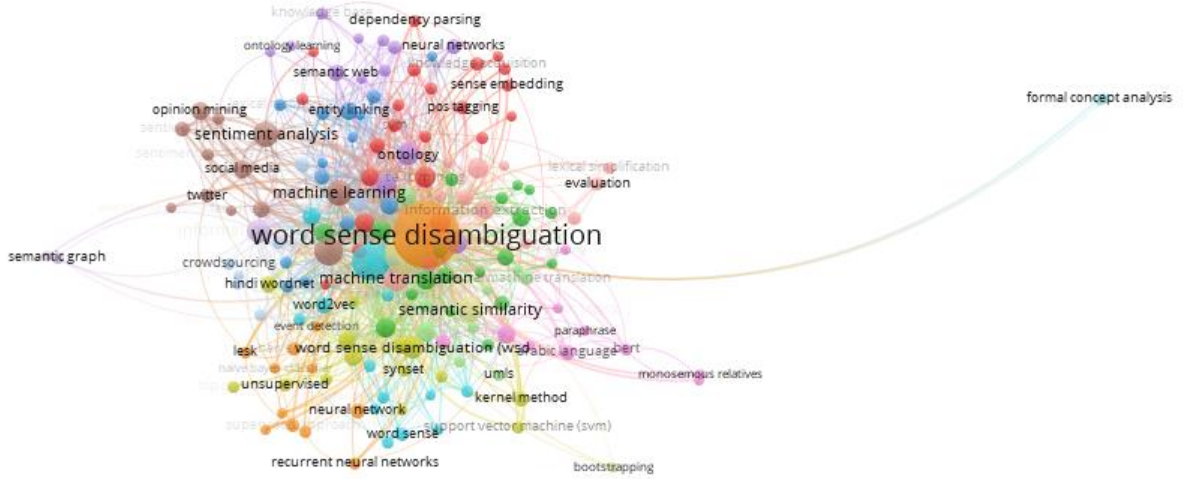


Figure 3.8: Author keyword co-occurrence network for "word sense disambiguation"

Given the centrality of the WSD keyword within the network, accompanied by the largest node size, the observed prominence aligns logically with the theme of this bibliometric analysis on WSD research. The centrality of the "word sense disambiguation" keyword in the co-occurrence network is a clear indication of its significance and relevance to the overall research landscape. This finding aligns with the primary objective of the bibliometric study, which aims to investigate the trends, patterns, and developments within the field of WSD.

3.9.3 Search analysis

A search analysis is crucial in bibliometrics as it forms the foundation for data collection, enabling researchers to identify and retrieve the necessary information for conducting a comprehensive analysis (Radhakrishnan *et al.* 2017).

Figure 3.9 illustrates the distribution of published research records over the past two decades. It is important to note that, while research in the field of WSD dates back to the 1950s (Weaver 1952; Kaplan 1955), the scope of this bibliometric analysis is limited to studies published within the last 20 years. This focus on more recent publications allows for a comprehensive examination of the current state, trends, and developments in WSD research, as well as the identification of emerging themes and future direction.

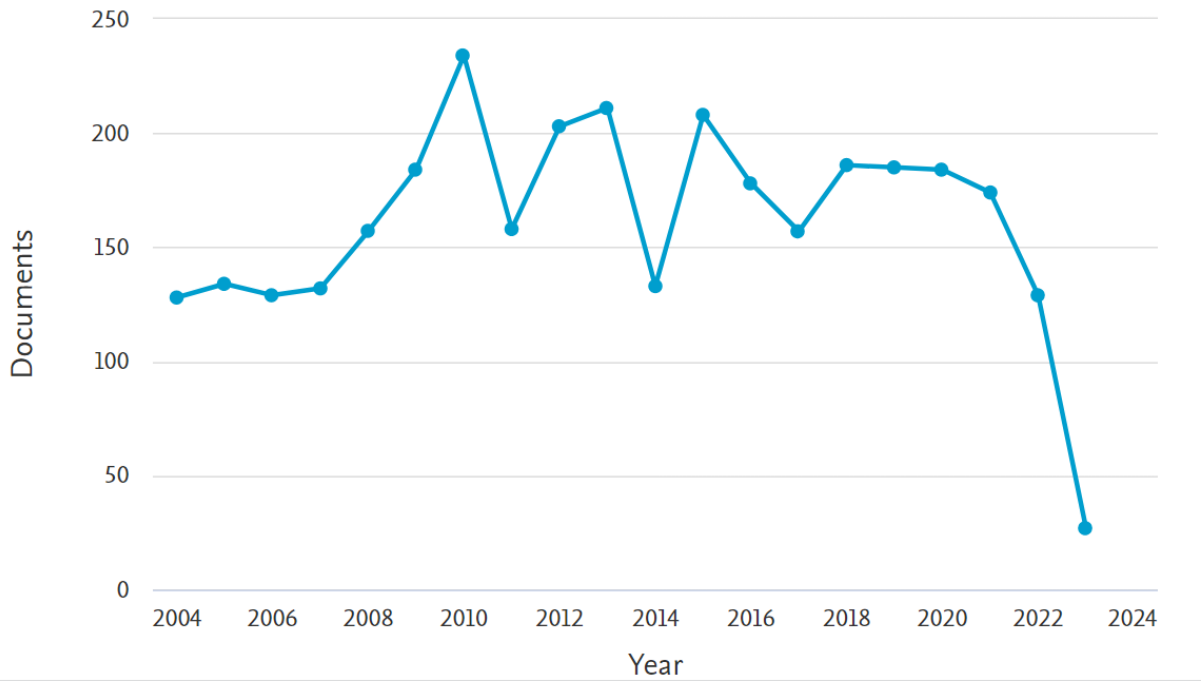


Figure 3.9: Yearly research record on WSD-related research

As illustrated in Figure 3.9, a substantial volume of research was published in the year 2010. In 2010, the field of artificial intelligence (AI) and machine learning (ML) was experiencing significant advancements and notable developments. One of the most prominent developments in 2010 was the emergence of deep learning, a subset of ML that focuses on neural networks with multiple layers. Deep-learning approaches began to attract considerable attention from the research community, because they demonstrated remarkable performance in various NLP tasks, such as language modelling, machine translation, and sentiment analysis (Young *et al.* 2018). In 2010, the emergence of deep learning, the active research efforts in NLP, and the availability of resources created opportunities for scientific discovery and innovation. The increased number of publications in 2010 reflects the momentum within the research community to innovate in language understanding and processing applications.

3.9.4 Top countries and top institutions

The 10 most prolific countries and institutional affiliations regarding WSD-related research are presented in Figures 3.10 and 3.11.

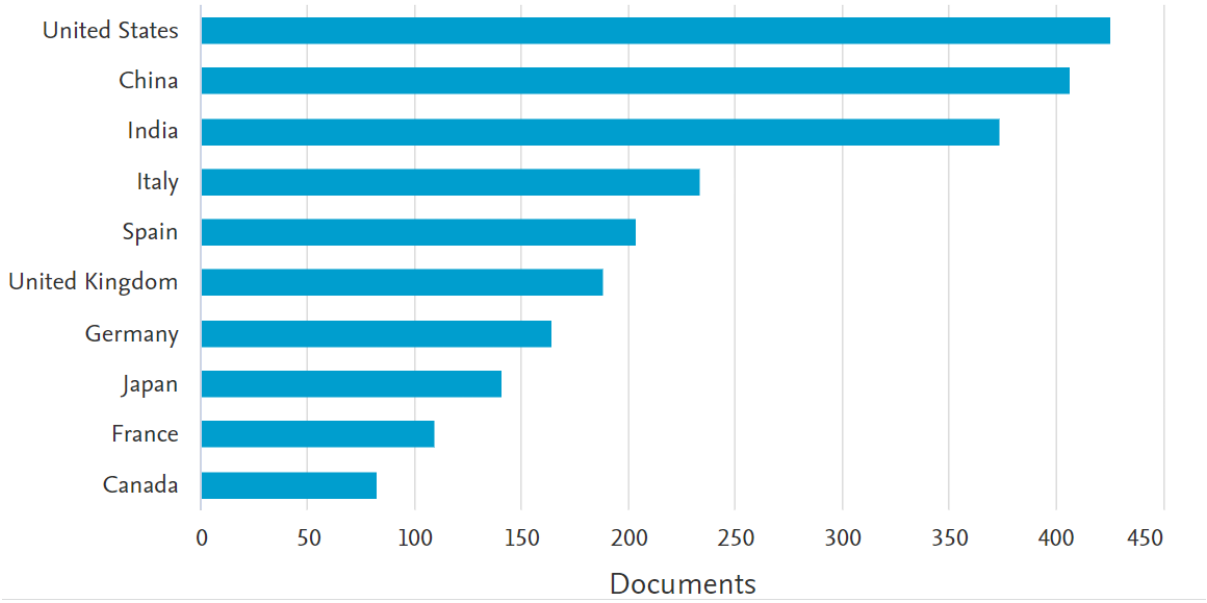


Figure 3.10: The 10 most prolific countries on WSD-related research

At the country level, the United States, as an English-speaking country occupies the first position, with China and India next in line. These three are of the world's largest and most influential countries in terms of population, economy, and geopolitical importance. The United States, China, and India are among the world's most populous countries, with a combined population of over 3 billion people (United Nations 2019). This large population base provides a substantial pool of potential researchers, students, and scholars who can contribute to the growth and advancement of WSD research. The United States, China, and India are also among the world's largest economies, with significant investments in research and development (R&D) (World Bank 2021). Their economic strength allows for greater funding opportunities, infrastructure support, and resources for researchers working in the field of WSD.

The three above-mentioned countries have diverse and rich cultural heritages, long histories, and diverse populations. In terms of language resources, English has a variety of open-source resources such as WordNet, BabelNet, Brown Corpus, SemCor, SemEval, etc., datasets that are openly available for conducting research and developing language tools. China and India have the advantage of having indigenous languages as their official languages, which means that written communication in these countries predominantly occurs in their respective languages. This linguistic characteristic leads to a richer availability of language resources than countries in which English is the official language of communication but is not the first language, such as South Africa. In China and India, various linguistic resources, such as

dictionaries, corpora, and language models, are developed and maintained in their native languages. This abundance of resources in their official languages provides a strong foundation for research and development in natural language processing; and facilitates the advancement of language-related technologies tailored to the specific linguistic characteristics of these countries. Figure 3.11 illustrates the top universities.

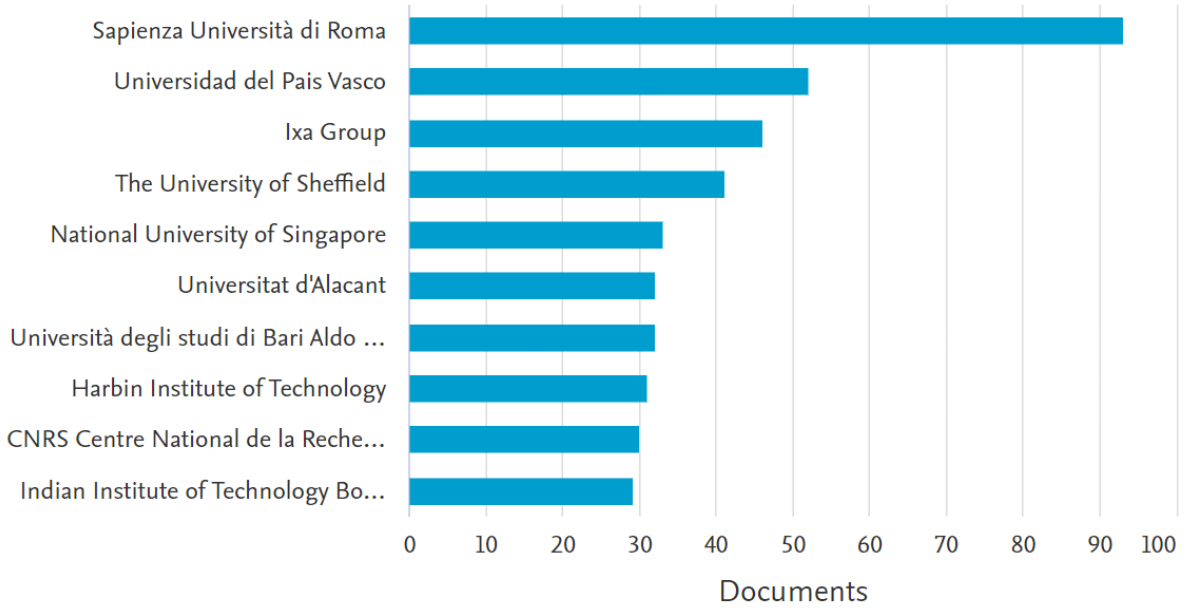


Figure 3.11: The 10 institutional affiliations on WSD-related research

Among the most prolific institutions in the field of WSD are Sapienza Universita di Roma, Universidad del Pais Vasco, and the Ixa group. It is noteworthy that these top-ranking institutions are all based in Europe. This aligns with the fact that Spain, where Universidad del Pais Vasco and the Ixa group are located, holds the fifth position among the top ten countries contributing significantly to WSD-related research. While the United States, China, and India demonstrate their prominence on a national level, with numerous research outputs and significant contributions, the absence of their institutions in the top rankings suggests a diverse landscape of WSD research expertise.

The foregoing analysis, when considered in the context of this study from an African and South African perspective, reveals noteworthy obstacles and disparities in the availability of resources for conducting WSD research. Despite Africa's rich linguistic diversity, with numerous languages spoken across the continent, there is a paucity of resources compared with other regions, such as the United States, China, and India. This lack of resources can be attributed to several factors specific to the African and South African context. One of the

primary reasons for the limited availability of resources is the official language status of indigenous African languages. In South Africa, for example, there are 11 official languages, including indigenous languages such as Setswana. However, English remains the dominant language in official communication, education, and research. This linguistic hierarchy creates a disparity in the development and maintenance of language resources for indigenous languages, the focus and resources more often directed toward English.

In contrast, countries like China and India have the advantage of utilizing their indigenous languages as official languages. This status promotes the development and maintenance of language resources, such as dictionaries, corpora, and language models specifically tailored to their native languages. The availability of these resources provides a solid foundation for conducting WSD research and developing language tools that cater to the unique linguistic characteristics of these languages.

Another notable trial in the African and South African context is the limited collaboration among researchers and institutions. Many researchers work in isolation, often due to a lack of funding, infrastructure, institutional support, or a combination of these shortages. This siloed approach hinders the sharing of knowledge, resources, and best practices, which is essential for advancing WSD research. Additionally, some researchers may not freely share their developed resources, such as corpora, which further limits the accessibility and reproducibility of research findings. The case of (Otlogetswe 2008) Setswana corpus, which contains over a million Setswana words but is not publicly available, illustrates this problem.

3.9.5 Research document types

Figure 3.12 presents the document types that are commonly employed for disseminating WSD research findings and methodologies. These document types may include journal articles, conference papers, book chapters, review articles, and technical reports, among others. The distribution of these document types reflects the mostly used medium for publishing WSD research and the relative significance of each format within the research community.

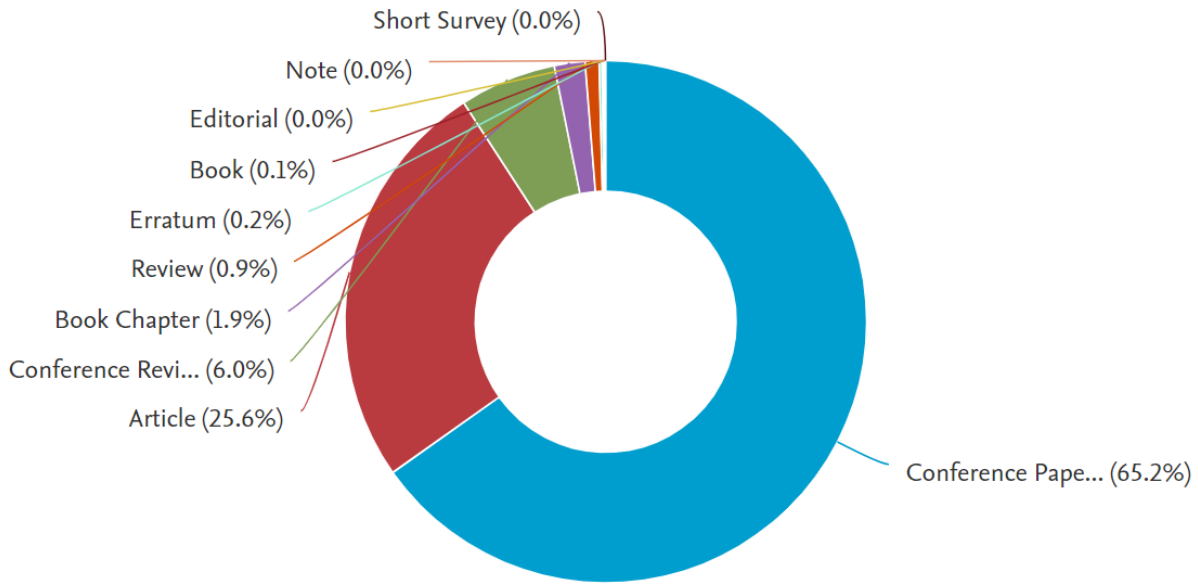


Figure 3.12: Document types on WSD-related research

Conference papers are often published more frequently than other types of academic outputs. This may be because conferences provide opportunities for researchers to connect and collaborate with peers from various other institutions and regions. Conference papers account for 65.2% of the published research in WSD. This may also be because compared with other types of academic outputs, such as journal articles or book chapters, conference papers typically have a faster turnaround time from submission to publication. This process allows researchers to share their work promptly with the scientific community, ensuring that the latest advancements and discoveries in WSD are communicated to a broad audience without delay (Eckmann, Rocha and Wainer 2012).

The analysis and discussion of the proportion of conference papers in WSD research is relevant to this study. In this context, the findings regarding the importance of conferences in WSD research underscore the potential for leveraging these platforms to advance the development of language technologies for Setswana. By actively participating in relevant conferences and presenting our work on the development of the Setswana WSD tool, researchers can engage with the broader WSD community, learn from the experiences and methodologies of other researchers, and establish collaborations that can support the growth and refinement of the proposed language model.

3.9.6 Citation analysis

At the document level, the most cited works cover WSD and the application of WSD in various NLP tasks and applications. Figure 3.13 presents the top twenty most-cited documents based on total citations. The most cited document, from Navigli and Ponzetto (2012)'s study is titled BabelNet: The automatic construction, evaluation, and application of a wide-coverage multilingual semantic network. The authors present an in-depth exploration of BabelNet, a multilingual semantic network, and propose automatic construction method and evaluation processes to build BabelNet, emphasizing its wide coverage across multiple languages and diverse applications. Other most-cited documents peruse semantic matching energy function for learning with multi-relational data in the application to word sense disambiguation (Bordes *et al.* 2014), learning generic context-embedding with bidirectional LSTM (Melamud, Goldberger and Dagan 2016), knowledge-enhanced contextual word representations (Peters *et al.* 2019) and joint learning of words and meaning representations for open-text semantic parsing (Bordes *et al.* 2012).

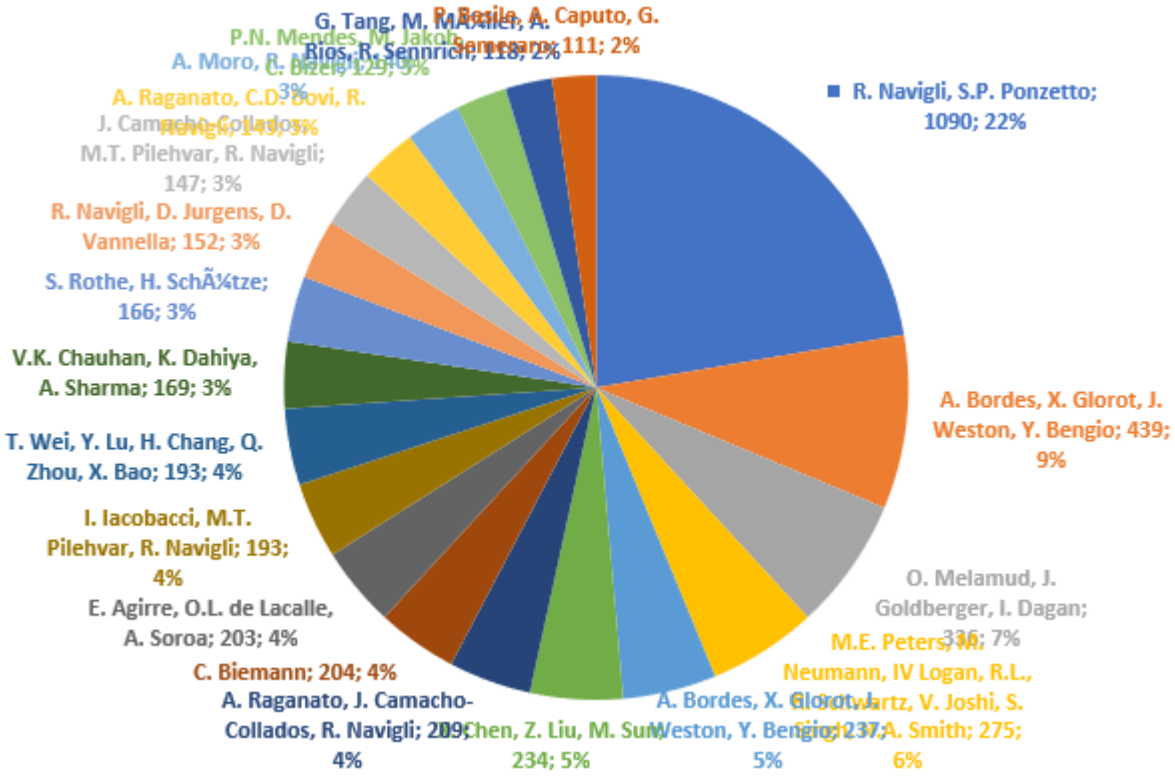


Figure 3.13: Top 20 most cited documents

Figure 3.14 illustrates the top 10 most published authors in terms of documents. The top 10 most cited documents were not produced by the top 10 most published authors. This statistic can be attributed to the fact that the number of publications does not necessarily reflect the impact or quality of the work. The top 10 most published authors may have contributed a large volume of research; however, their papers may not have garnered as much attention or citation as the top-cited documents. This can be seen with Aggirre, E. who is number two in terms of the number of published works, and number 10 in the top-cited category.

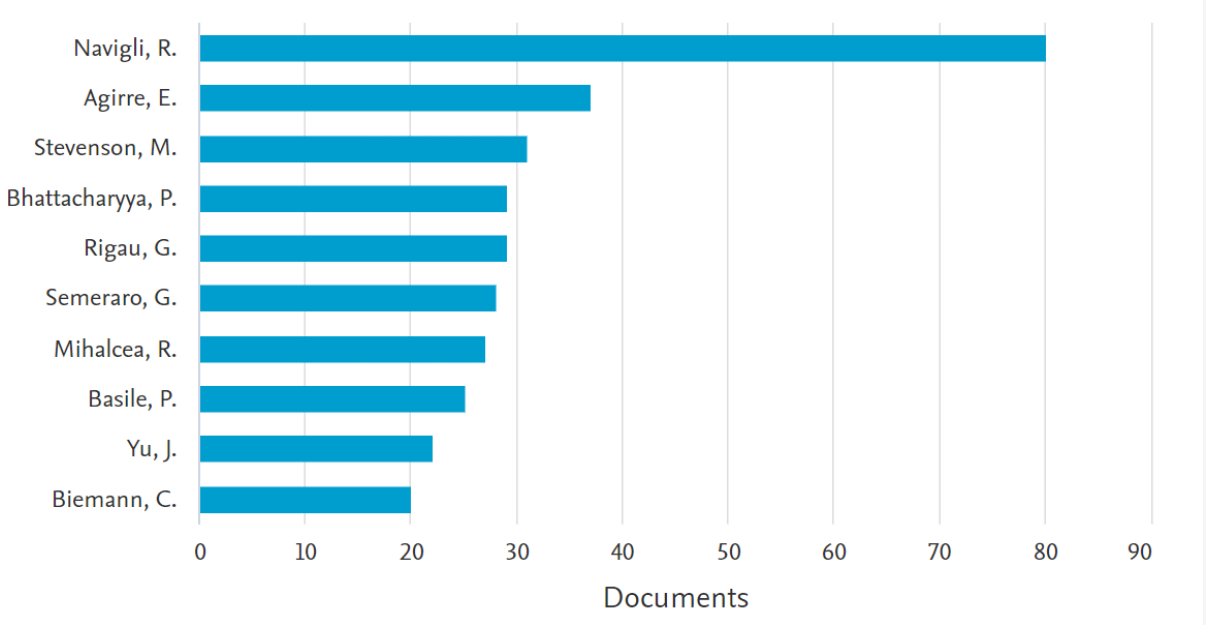


Figure 3.14: Top 10 most published authors

It is relevant that none of these top 10 authors is from sub-Saharan Africa. This lack of representation highlights the persistent disparities and challenges faced by researchers from this region in contributing to global knowledge production in the field of WSD and, more broadly, in NLP. These disparities and obstacles include limited linguistic data, research infrastructure, and funding. This scarcity of resources hinders the ability of African researchers to contribute effectively to the field on both local and global scales.

3.9.7 Bibliometric-analysis results discussion

This section provided a bibliometric analysis that investigates the research contributions and trends in the field of WSD by analysing relationships between various research components. The study follows a systematic procedure to determine search terms, select appropriate databases, establish selection criteria, and analyse the results using quantitative techniques. The analysis reveals that the United States, China, and India are the most prolific countries in WSD

research, attributed to their large populations, economic power, and linguistic diversity. However, the absence of sub-Saharan African countries highlights the disparities in research output and resources for conducting WSD research in this region.

Co-citation analysis at the country, author, and keyword levels provides insights into collaborative networks, influential researchers, and key research themes. The study identifies prominent authors, such as Roberto Navigli, and their extensive co-citation relationships within the WSD research landscape. The keyword co-citation analysis reveals the centrality of "word sense disambiguation" and "natural language processing" in the field.

The study also examines the distribution of research output over time, noting a substantial volume of publications in 2010, coinciding with advancements in artificial intelligence and machine learning. Conference papers are found to be the most common document type, accounting for 65.2% of published WSD research, due to their faster turnaround time and opportunities for collaboration.

The analysis of subject areas and research categories shows that WSD research primarily takes place in the domains of natural language processing and computational linguistics. The most cited documents cover WSD and its applications in various NLP tasks.

The study highlights the hurdles faced by researchers in sub-Saharan Africa, such as limited collaboration, lack of resource sharing, and the dominance of English in official communication and research. These factors contribute to the underrepresentation of African researchers in the global WSD research landscape.

3.10 Meta-analysis of word sense disambiguation approaches

The purpose of this section is to provide a comprehensive and rigorous analysis of the existing research using meta-analysis. Meta-analysis serves as a resourceful research methodology for systematically reviewing and synthesizing findings. The method of selecting the pertinent articles for extraction was created using the preferred reporting items for systematic reviews and meta-analyses (PRISMA). The included research articles were searched from the Scopus and Web of Science databases. All the statistical analyses were performed using the random-effects model implementation in Stata version 17. The overall pooled estimated delay component is presented in forest plots. Overall, thirteen studies were included in the meta-analysis; and the overall pooled estimate was 10% (95% CI: 7%, 30%).

3.10.1 Search strategy

A search of the literature was conducted to identify all published research studies on WSD. The PRISMA flowchart detailing the extraction of relevant studies is presented in Figure 3.15.

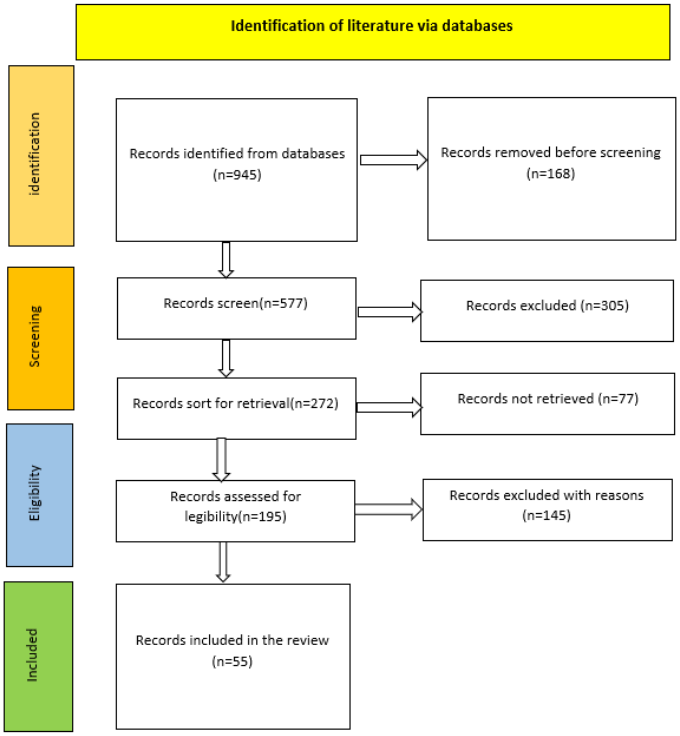


Figure 3.15: Flow diagram of database search using PRISMA framework

The literature search strategy, screening and selection of publications, identification of parameters to be extracted from the studies, quality assessment, data extraction, and reporting results were carried out using the recommendations of preferred reporting items for systematic reviews and meta-analyses (PRISMA) (Moher *et al.* 2009). The academic databases, Scopus and Web of Science were searched to gather published research articles for this meta-analysis. The search terms used during a comprehensive literature search were: “word sense disambiguation” OR “word sense disambiguation for machine translation”; OR “sense disambiguation” OR “disambiguating senses” AND “ambiguity in machine translation”; OR “ambiguities” AND “of senses” OR “in machine translation”. The search resulted with 3987 published articles in total between 2004 and 2023 before applying exclusion criteria relating to publication years, document types, open access, and languages. Upon excluding irrelevant studies, 945 articles were identified for screening and subsequently imported into Microsoft Excel for further analysis. To ensure a comprehensive approach, the reference lists of relevant

papers were also manually examined to identify any citations that may have been overlooked during the electronic database search.

3.10.2 Inclusion and exclusion criteria

The 945 studies identified for screening were subjected to inclusion and exclusion criteria as shown in Table 3.3.

Table 3.3: Meta-analysis Inclusion and Exclusion Criteria

Criterion
Exclusion Criteria
EC1: Research papers not written in English
EC2: Duplicate papers
EC3: Research papers with only abstract
EC4: Research papers not relevant to WSD
EC5: Research papers that do not employ WSD for MT
EC6: Research papers not reporting precision, recall, and F-score
EC7: Review, survey, editorials, commentaries, book chapters, brief studies, thesis, dissertation
Inclusion Criteria
IC1: Research papers written in English
IC2: Full research papers
IC3: Research papers relevant to WSD
IC4: Research papers that employ WSD for MT
IC5: Research papers reporting precision, recall, and F-score

These criteria were carefully designed to ensure that only the most relevant and most high-quality research papers were included in the meta-analysis. Studies were excluded if they were not written in English (EC1), as this is the primary language of scientific communication, and enables a consistent analysis of the literature. Duplicate papers (EC2) were also removed to avoid any potential bias or overrepresentation of specific findings. Additionally, research papers with only an abstract available (EC3) were excluded, because they lacked the necessary detail and information required for a comprehensive meta-analysis.

Furthermore, studies that were not directly relevant to WSD (EC4) or did not employ WSD specifically for machine translation (MT) (EC5) were excluded, because they did not align with the primary focus of this meta-analysis. Research papers that did not report evaluation results precision, recall, and F-score (EC6), were also excluded, these metrics being crucial for calculating meta-analysis variables. Lastly, various types of non-research publications, such as reviews, surveys, editorials, commentaries, book chapters, brief studies, theses, and dissertations (EC7), were equally excluded, to maintain a focus on original, peer-reviewed research contributions.

In contrast, the inclusion criteria were designed to select research papers that were written in English (IC1), available as full-text articles (IC2), directly relevant to WSD (IC3), employing WSD (IC4), and reporting precision, recall, and F-score (IC5). By applying both these inclusion and exclusion criteria, the meta-analysis aimed to synthesize the findings from the most relevant and high-quality studies, providing a reliable and comprehensive assessment of the state-of-the-art WSD methods for MT.

3.10.3 Quality assessment and data extraction

Each article was assessed on merits and relevance, and based on the defined exclusion and inclusion criteria. From the selected studies, studies that met the inclusion criteria were taken for further analysis. A total of 55 studies was included in this systematic review and meta-analysis which had 100% of the information and met all the inclusion criteria. The Excel spreadsheet was populated with article data extracted according to variables listed in Table 3.4. The resulting database consisted of nine variables, which were populated with data that were retrieved through the review of the selected studies.

Table 3.4: Quality Assessment and Data Extraction

Data Extracted No	Contents	Type
1	Title	Research paper title
2	Author	Research paper authors
3	Year	Year of publication
4	Method	Methods used
5	Approach	Approach used
6	Dataset	Dataset used for evaluation
7	Precision	Average precision

8	Recall	Average recall
9	F1 score	Average F1 score

The nine essential variables were extracted from each research paper including the title of the research paper, the authors, and the year of publication, which provided basic bibliographic information, as shown in Table 3.4. Additionally, the methods and approaches employed in each study were documented, allowing for a comparative analysis of the various techniques used in WSD. The dataset utilized for evaluation in each study was also recorded, enabling an assessment of the performance and generalizability of the proposed methods across different data sources. Finally, three crucial evaluation metrics ─ average precision, average recall, and average F1 score were extracted from each study, providing a standardized means of assessing the effectiveness and accuracy of the WSD approaches. By systematically compiling this information into a structured database, the quality assessment and data-extraction process provided the foundation for a comprehensive meta-analysis of the selected studies.

3.10.4 Data synthesis and statistical analysis

To facilitate the statistical analysis, the extracted data were initially compiled and organized on a Microsoft Excel spreadsheet. Subsequently, the data was imported into the STATA statistical analysis software version 17. The effect sizes of each included primary study; and the total pooled effect size of all primary studies was calculated using the extracted data. The random-effects model served as the groundwork for our analysis. Using the Cochrane Q statistic, the study heterogeneity was determined. Consequently, τ^2 and I^2 were employed to measure study heterogeneity (Olugbara *et al.* 2021). I^2 values of 25%, 50%, and 75%, respectively, reflect low, medium, and high heterogeneity. The effect sizes were calculated using the forest plot as a preamble to assessing heterogeneity and biases in the results of the included studies (Shamseer *et al.* 2015). For the purpose of assessing the efficacy of various WSD approaches, a pooled estimate was produced using a DerSimonian and Laird random-effects model. Furthermore, when conducting a moderator analysis in a systematic review with meta-analysis, subgroup analysis and meta-regression are frequently utilized for fine-grained analysis. Subgroup analysis partitions data into smaller groups to compare the sample data. Therefore, to identify the source of study heterogeneity, a subgroup analysis focused on research performance evaluation metrics was performed, i.e., on the accuracy of the included studies. The subgroups were based on the approach used for WSD. In addition, to determine whether

any subsets of the included studies captured the pooled effect size, meta-regression analyses were conducted (Borenstein *et al.* 2010).

3.10.5 Meta-analysis summary

To assess the performance of WSD methods, meta-analyses were conducted using the random-effects model, employing the F1 score and precision as the primary evaluation metrics. The analyses were based on the effect size and standard error of the effect size for each included study. Figure 3.16 presents a visual summary of the meta-analysis results in the form of a forest plot. The figure depicts the pooled effect-size estimates, and the associated confidence intervals, allowing for a comprehensive assessment of the overall performance of WSD methods across the included studies. The forest plot provides a clear and concise overview of the individual study results, as well as the combined effect-size estimate, which allows for the formulation of evidence-based conclusions concerning the effectiveness of WSD approaches.

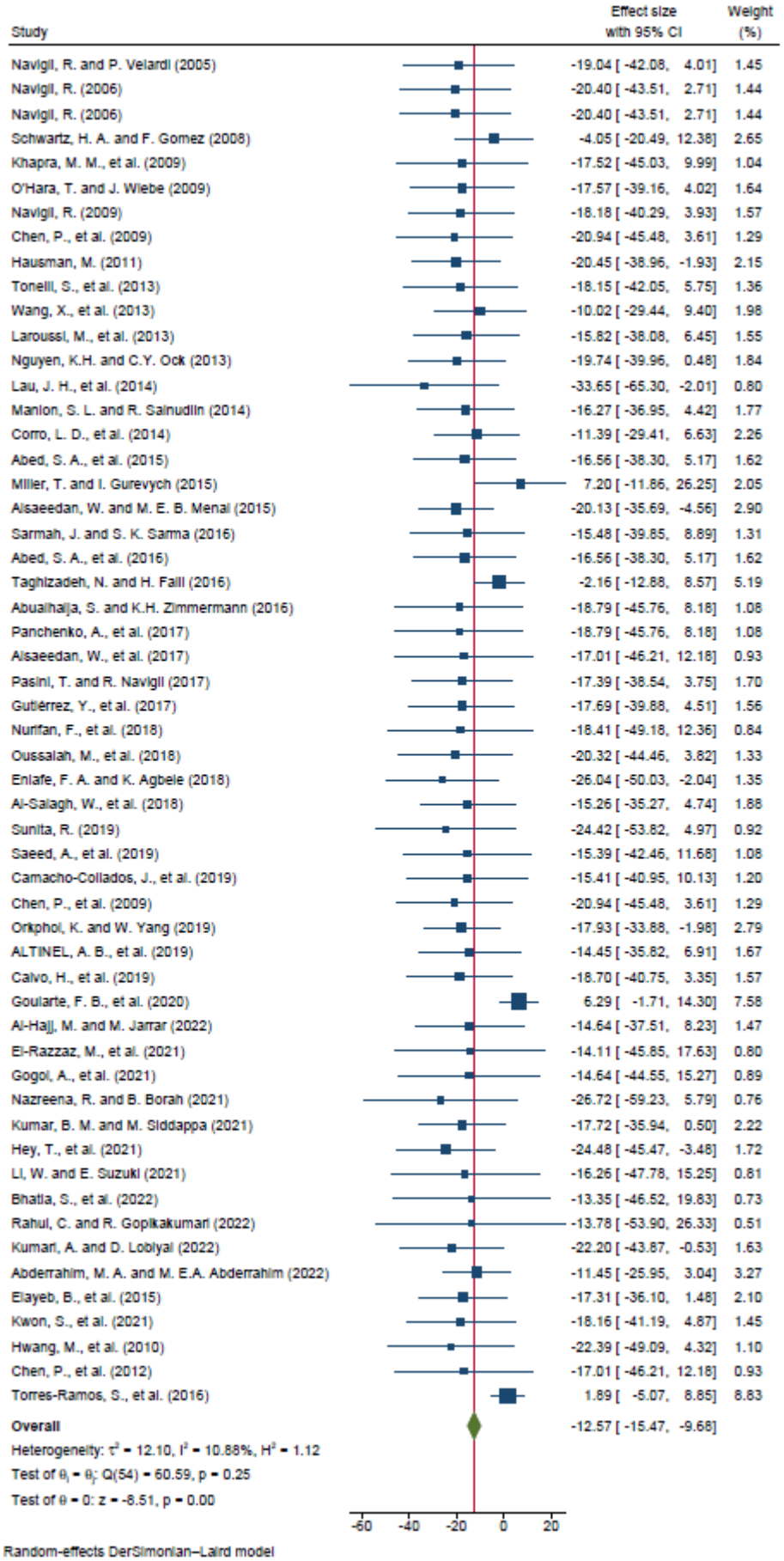

Study
Effect size with 95% CI
Weight (%)
Navigil, R. and P. Velardi (2005) -19.04 [-42.08, 4.01] 1.45
Navigil, R. (2006) -20.40 [-43.51, 2.71] 1.44
Navigil, R. (2006) -20.40 [-43.51, 2.71] 1.44
Schwartz, H. A. and F. Gomez (2008) -4.05 [-20.49, 12.38] 2.65
Khapra, M. M., et al. (2009) -17.52 [-45.03, 9.99] 1.04
O'Hara, T. and J. Wiebe (2009) -17.57 [-39.16, 4.02] 1.64
Navigil, R. (2009) -18.18 [-40.29, 3.93] 1.57
Chen, P., et al. (2009) -20.94 [-45.48, 3.61] 1.29
Hausman, M. (2011) -20.45 [-38.96, -1.93] 2.15
Tonelli, S., et al. (2013) -18.15 [-42.05, 5.75] 1.36
Wang, X., et al. (2013) -10.02 [-29.44, 9.40] 1.98
Laroussi, M., et al. (2013) -15.82 [-38.08, 6.45] 1.55
Nguyen, K.H. and C.Y. Ock (2013) -19.74 [-39.96, 0.48] 1.84
Lau, J. H., et al. (2014) -33.65 [-65.30, -2.01] 0.80
Manlon, S. L. and R. Salnudlln (2014) -16.27 [-36.95, 4.42] 1.77
Corro, L. D., et al. (2014) -11.39 [-29.41, 6.63] 2.26
Abed, S. A., et al. (2015) -16.56 [-38.30, 5.17] 1.62
Miller, T. and I. Gurevych (2015) 7.20 [-11.86, 26.25] 2.05
Alsaeedan, W. and M. E. B. Menai (2015) -20.13 [-35.69, -4.56] 2.90
Sarmah, J. and S. K. Sarma (2016) -15.48 [-39.85, 8.89] 1.31
Abed, S. A., et al. (2016) -16.56 [-38.30, 5.17] 1.62
Taghizadeh, N. and H. Faili (2016) -2.16 [-12.88, 8.57] 5.19
Abualhaija, S. and K.H. Zimmermann (2016) -18.79 [-45.76, 8.18] 1.08
Panchenko, A., et al. (2017) -18.79 [-45.76, 8.18] 1.08
Alsaeedan, W., et al. (2017) -17.01 [-46.21, 12.18] 0.93
Pasini, T. and R. Navigli (2017) -17.39 [-38.54, 3.75] 1.70
Gutiérrez, Y., et al. (2017) -17.69 [-39.88, 4.51] 1.56
Nurifan, F., et al. (2018) -18.41 [-49.18, 12.36] 0.84
Oussalah, M., et al. (2018) -20.32 [-44.46, 3.82] 1.33
Enlafe, F. A. and K. Agbele (2018) -26.04 [-50.03, -2.04] 1.35
Al-Salagh, W., et al. (2018) -15.26 [-35.27, 4.74] 1.88
Sunita, R. (2019) -24.42 [-53.82, 4.97] 0.92
Saeed, A., et al. (2019) -15.39 [-42.46, 11.68] 1.08
Camacho-Collados, J., et al. (2019) -15.41 [-40.95, 10.13] 1.20
Chen, P., et al. (2009) -20.94 [-45.48, 3.61] 1.29
Orikphol, K. and W. Yang (2019) -17.93 [-33.88, -1.98] 2.79
ALTINEL, A. B., et al. (2019) -14.45 [-35.82, 6.91] 1.67
Calvo, H., et al. (2019) -18.70 [-40.75, 3.35] 1.57
Goularte, F. B., et al. (2020) 6.29 [-1.71, 14.30] 7.58
Al-Hajj, M. and M. Jarrar (2022) -14.64 [-37.51, 8.23] 1.47
El-Razzaz, M., et al. (2021) -14.11 [-45.85, 17.63] 0.80
Gogol, A., et al. (2021) -14.64 [-44.55, 15.27] 0.89
Nazreena, R. and B. Borah (2021) -26.72 [-59.23, 5.79] 0.76
Kumar, B. M. and M. Siddappa (2021) -17.72 [-35.94, 0.50] 2.22
Hey, T., et al. (2021) -24.48 [-45.47, -3.48] 1.72
Li, W. and E. Suzuki (2021) -16.26 [-47.78, 15.25] 0.81
Bhatia, S., et al. (2022) -13.35 [-46.52, 19.83] 0.73
Rahul, C. and R. Goplikakumari (2022) -13.78 [-53.90, 26.33] 0.51
Kumari, A. and D. Lobiyal (2022) -22.20 [-43.87, -0.53] 1.63
Abderrahim, M. A. and M. E.A. Abderrahim (2022) -11.45 [-25.95, 3.04] 3.27
Elayeb, B., et al. (2015) -17.31 [-36.10, 1.48] 2.10
Kwon, S., et al. (2021) -18.16 [-41.19, 4.87] 1.45
Hwang, M., et al. (2010) -22.39 [-49.09, 4.32] 1.10
Chen, P., et al. (2012) -17.01 [-46.21, 12.18] 0.93
Torres-Ramos, S., et al. (2016) 1.89 [-5.07, 8.85] 8.83
Overall -12.57 [-15.47, -9.68]
Heterogeneity: τ^2 = 12.10, I^2 = 10.88%, H^2 = 1.12
Test of $\theta_i = \theta_j$: Q(54) = 60.59, p = 0.25
Test of θ = 0: z = -8.51, p = 0.00
-60 -40 -20 0 20
Random-effects DerSimonian–Laird model

Figure 3.16: Forest plot for distribution of effect WSD methods

The meta-analysis summary indicates that the between-study variability is low, as evidenced by the τ2 statistic of 12.10. The I2 statistic, which represents the proportion of true heterogeneity to total observed variation, was found to be 10.88% (p = 0.000). This value is considered significantly low, falling within the 25%, 50%, and 75% interval levels commonly used to categorize heterogeneity. The low I2 value suggests that the observed variability across studies is largely attributable to sampling error rather than true differences in the underlying effects. Consequently, these findings support the notion that the included studies exhibit a high degree of homogeneity, strengthening the validity and generalizability of the meta-analysis results. The low between-study variability and true heterogeneity indicate that the pooled effect-size estimate is likely to be a reliable representation of the overall effectiveness of the investigated WSD approaches.

The Galbraith plot was used in this study to assess the heterogeneity between the included studies. The Galbraith plot, depicted in Figure 3.17, provides a visual representation of the standardized effect sizes (z-scores) of each study against the inverse of their standard errors (1/SE), and the dispersion of the study results with the overall pooled effect-size estimate.

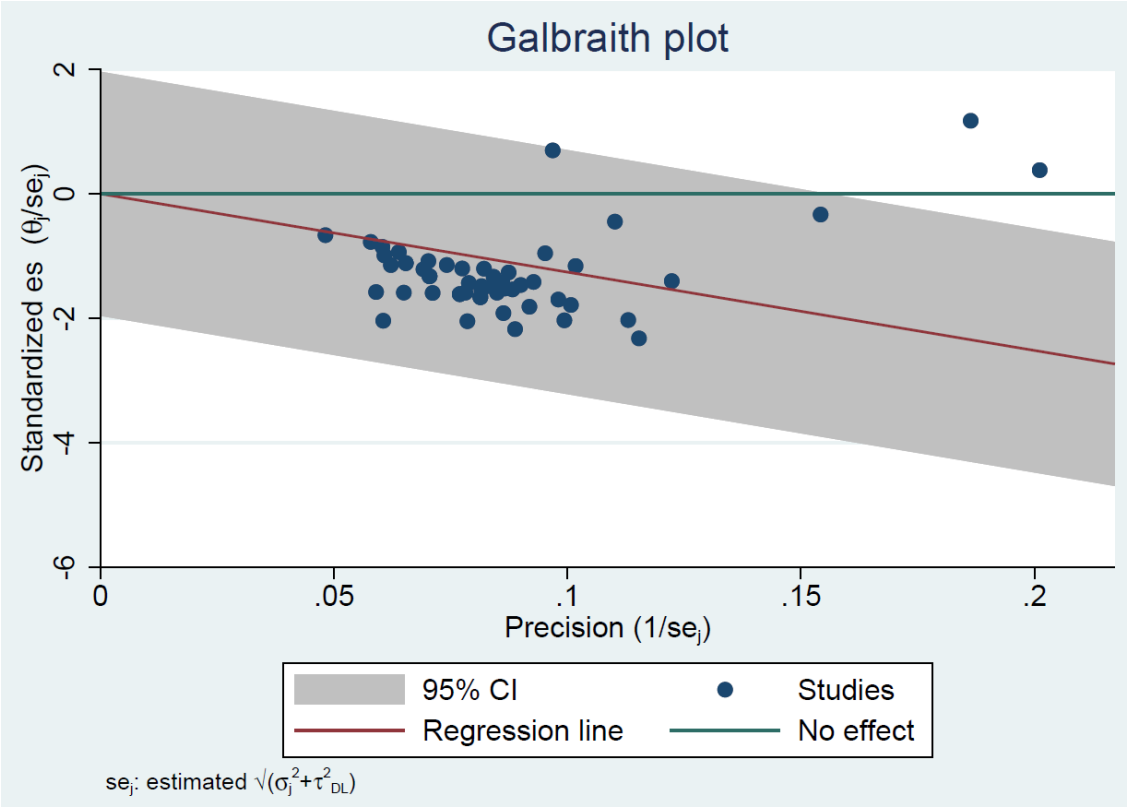


Figure 3.17: Galbraith plot of reviewed studies

As depicted on the Galbraith plot, two studies fall outside the 95% confidence interval bands. This suggests a high level of consistency and low heterogeneity among the majority of the included studies. This finding indicates that the pooled effect-size estimate is likely to be robust and representative of the overall effectiveness of the investigated WSD approaches. The fact that 53 out of 55 studies lie within the confidence interval bands implies that the results are in agreement with the pooled effect-size estimate, reinforcing the reliability of the meta-analysis findings. The low number of outliers suggests that the observed variation in effect sizes across studies is more likely attributable to random sampling error rather than substantial differences in study characteristics or true effect sizes.

3.10.6 Subgroup analysis

To investigate the potential sources of heterogeneity among the included studies and to explore the impact of specific study characteristics on the overall effect-size estimate, subgroup analysis was conducted. Figure 3.18 displays a summary of the subgroup analysis of the different WSD approaches employed in the included studies. The subgroup analysis aimed to identify effect modifiers, such as differences in WSD techniques, that may influence the effectiveness of the methods used in the included studies. By comparing the effect sizes and heterogeneity measures between subgroups, this study was able to assess whether the observed heterogeneity can be explained by specific study characteristics.

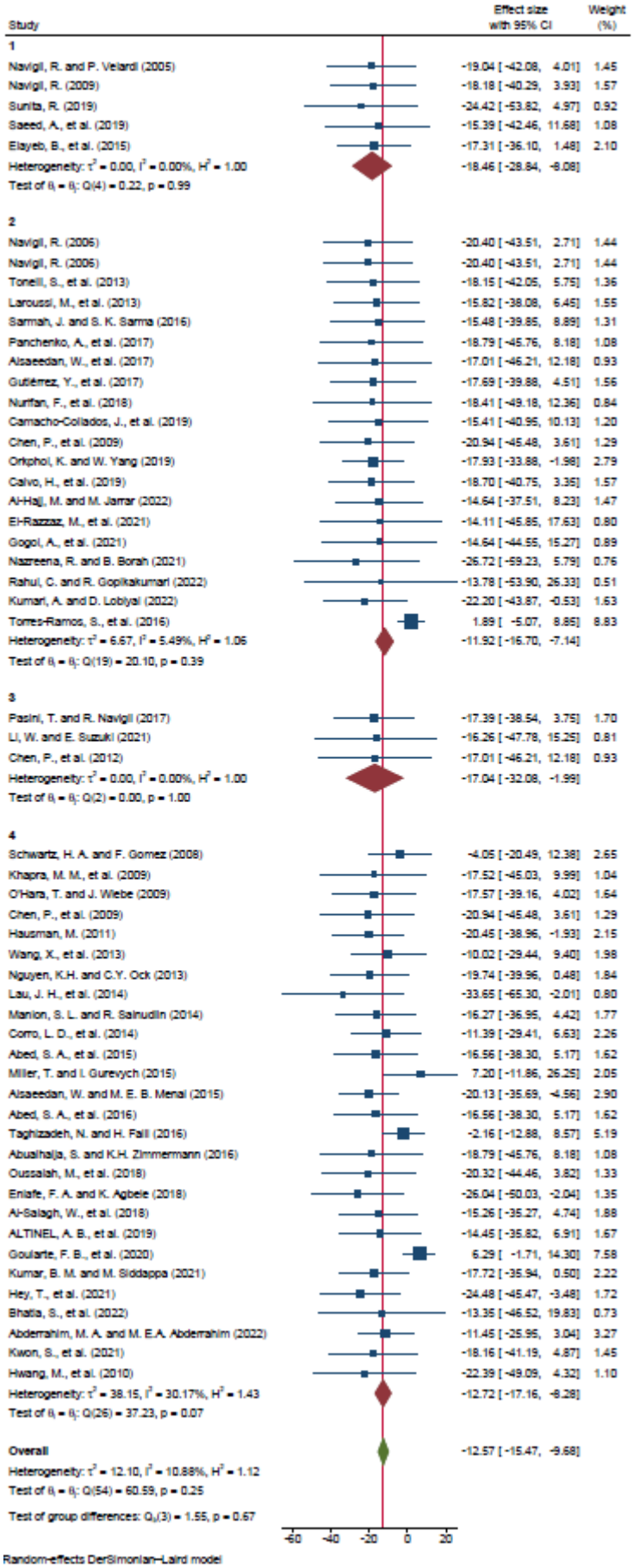

Study
Effect size with 95% CI
Weight (%)
1
Navigil, R. and P. Velardi (2005) -19.04 [-42.08, 4.01] 1.45
Navigil, R. (2009) -18.18 [-40.29, 3.93] 1.57
Sunita, R. (2019) -24.42 [-53.82, 4.97] 0.92
Saeed, A., et al. (2019) -15.39 [-42.46, 11.68] 1.08
Elayeb, B., et al. (2015) -17.31 [-36.10, 1.48] 2.10
Heterogeneity: τ^2 = 0.00, I^2 = 0.00%, H^2 = 1.00 -18.46 [-28.84, -8.08]
Test of $\theta_i = \theta_j$: Q(4) = 0.22, p = 0.99
2
Navigil, R. (2006) -20.40 [-43.51, 2.71] 1.44
Navigil, R. (2006) -20.40 [-43.51, 2.71] 1.44
Tonelli, S., et al. (2013) -18.15 [-42.05, 5.75] 1.36
Larcoussi, M., et al. (2013) -15.82 [-38.08, 6.45] 1.55
Sarmah, J. and S. K. Sarma (2016) -15.48 [-39.85, 8.89] 1.31
Panchenko, A., et al. (2017) -18.79 [-45.76, 8.18] 1.08
Alsaeedan, W., et al. (2017) -17.01 [-46.21, 12.18] 0.93
Gutiérrez, Y., et al. (2017) -17.69 [-39.88, 4.51] 1.56
Nurifan, F., et al. (2018) -18.41 [-49.18, 12.36] 0.84
Camacho-Collados, J., et al. (2019) -15.41 [-40.95, 10.13] 1.20
Chen, P., et al. (2009) -20.94 [-45.48, 3.61] 1.29
Orkphol, K. and W. Yang (2019) -17.93 [-33.88, -1.98] 2.79
Calvo, H., et al. (2019) -18.70 [-40.75, 3.35] 1.57
Al-Hajj, M. and M. Jarrar (2022) -14.64 [-37.51, 8.23] 1.47
El-Razzaz, M., et al. (2021) -14.11 [-45.85, 17.63] 0.80
Gogoi, A., et al. (2021) -14.64 [-44.55, 15.27] 0.89
Nazreena, R. and B. Borah (2021) -26.72 [-59.23, 5.79] 0.76
Rahul, C. and R. Gopikakumari (2022) -13.78 [-53.90, 26.33] 0.51
Kumari, A. and D. Lobiyal (2022) -22.20 [-43.87, -0.53] 1.63
Torres-Ramos, S., et al. (2016) 1.89 [-5.07, 8.85] 8.83
Heterogeneity: τ^2 = 6.67, I^2 = 5.49%, H^2 = 1.06 -11.92 [-16.70, -7.14]
Test of $\theta_i = \theta_j$: Q(19) = 20.10, p = 0.39
3
Pasini, T. and R. Navigil (2017) -17.39 [-38.54, 3.75] 1.70
Li, W. and E. Suzuki (2021) -16.26 [-47.78, 15.25] 0.81
Chen, P., et al. (2012) -17.01 [-46.21, 12.18] 0.93
Heterogeneity: τ^2 = 0.00, I^2 = 0.00%, H^2 = 1.00 -17.04 [-32.08, -1.99]
Test of $\theta_i = \theta_j$: Q(2) = 0.00, p = 1.00
4
Schwartz, H. A. and F. Gomez (2008) -4.05 [-20.49, 12.38] 2.65
Khapra, M. M., et al. (2009) -17.52 [-45.03, 9.99] 1.04
O'Hara, T. and J. Wiebe (2009) -17.57 [-39.16, 4.02] 1.64
Chen, P., et al. (2009) -20.94 [-45.48, 3.61] 1.29
Hausman, M. (2011) -20.45 [-38.96, -1.93] 2.15
Wang, X., et al. (2013) -10.02 [-29.44, 9.40] 1.98
Nguyen, K.H. and C.Y. Ock (2013) -19.74 [-39.96, 0.48] 1.84
Lau, J. H., et al. (2014) -33.65 [-65.30, -2.01] 0.80
Manion, S. L. and R. Salnudin (2014) -16.27 [-36.95, 4.42] 1.77
Corro, L. D., et al. (2014) -11.39 [-29.41, 6.63] 2.26
Abed, S. A., et al. (2015) -16.56 [-38.30, 5.17] 1.62
Miller, T. and I. Gurevych (2015) 7.20 [-11.86, 26.25] 2.05
Alsaeedan, W. and M. E. B. Menai (2015) -20.13 [-35.69, -4.56] 2.90
Abed, S. A., et al. (2016) -16.56 [-38.30, 5.17] 1.62
Taghizadeh, N. and H. Faili (2016) -2.16 [-12.88, 8.57] 5.19
Abualhaija, S. and K.H. Zimmermann (2016) -18.79 [-45.76, 8.18] 1.08
Oussalah, M., et al. (2018) -20.32 [-44.46, 3.82] 1.33
Enlafe, F. A. and K. Agbele (2018) -26.04 [-50.03, -2.04] 1.35
Al-Salagh, W., et al. (2018) -15.26 [-35.27, 4.74] 1.88
ALTINEL, A. B., et al. (2019) -14.45 [-35.82, 6.91] 1.67
Goularte, F. B., et al. (2020) 6.29 [-1.71, 14.30] 7.58
Kumar, B. M. and M. Siddappa (2021) -17.72 [-35.94, 0.50] 2.22
Hey, T., et al. (2021) -24.48 [-45.47, -3.48] 1.72
Bhatia, S., et al. (2022) -13.35 [-46.52, 19.83] 0.73
Abderrahim, M. A. and M. E.A. Abderrahim (2022) -11.45 [-25.95, 3.04] 3.27
Kwon, S., et al. (2021) -18.16 [-41.19, 4.87] 1.45
Hwang, M., et al. (2010) -22.39 [-49.09, 4.32] 1.10
Heterogeneity: τ^2 = 38.15, I^2 = 30.17%, H^2 = 1.43 -12.72 [-17.16, -8.28]
Test of $\theta_i = \theta_j$: Q(26) = 37.23, p = 0.07
Overall -12.57 [-15.47, -9.68]
Heterogeneity: τ^2 = 12.10, I^2 = 10.88%, H^2 = 1.12
Test of $\theta_i = \theta_j$: Q(54) = 60.59, p = 0.25
Test of group differences: $Q_b(3)$ = 1.55, p = 0.67
-60
-40
-20
0
20
Random-effects DerSimonian-Laird model

Figure 3.18: Forest plot for distribution of effect WSD methods sub-group analysis

In this study, the subgroup analysis was conducted by grouping the included studies into four primary categories based on the WSD techniques utilized ─ knowledge-based, supervised, unsupervised, and semi-supervised approaches.

In the subgroup analysis, the included studies were categorized into four distinct groups based on the WSD approaches employed. Group 1 consisted of supervised approaches, with 5 studies and an overall effect size of -18.46. Group 2 comprised unsupervised approaches, with 20 studies and an overall effect size of -11.92. Group 3 included semi-supervised methods, with 3 studies and an overall effect size of -17.04. Finally, Group 4 encompassed knowledge-based approaches, with 27 studies and an overall effect size of -12.72, all reported with a 95% confidence interval.

The effect sizes obtained for each group provide valuable insights into the relative effectiveness of the differing WSD approaches. As per the computed results, the magnitude of the effect sizes varies among the groups, suggesting potential differences in the effectiveness of each approach. The supervised approaches (Group 1) demonstrated the largest effect size of -18.46, indicating that these methods, which rely on labelled training data, may yield the most substantial improvements in WSD performance.

Conversely, the unsupervised approaches (Group 2) exhibited the smallest effect size of -11.92, suggesting that these methods, which do not require labelled data, may have a more limited impact on WSD effectiveness compared with other approaches. The semi-supervised methods (Group 3) and knowledge-based approaches (Group 4) showed intermediate effect sizes of -17.04 and -12.72, respectively, indicating that these techniques may offer a balance between the performance gains of supervised methods and the reduced reliance on labelled data of unsupervised methods.

Furthermore, the subgroup analysis helps to explain potential sources of heterogeneity among the included studies, the effect sizes within each group being more homogeneous compared with the overall meta-analysis. In the subgroup analysis, the $\tau 2$ statistic and its associated P-value were calculated for each group to assess the heterogeneity within the subgroups. The $\tau 2$ statistic represents the between-study variance, with higher values indicating greater heterogeneity among the studies within a subgroup. The P-value associated with $\tau 2$ provides an indication of the statistical significance of the observed heterogeneity.

For Group 1, which consists of supervised approaches, the τ2 statistic is 0.00 with a P-value of 0.99. This suggests that there is no heterogeneity among the studies within this subgroup, the between-study variance being estimated as zero. The high P-value (0.99) indicates that the observed heterogeneity is not statistically significant, providing strong evidence for the consistency and homogeneity of the effect sizes within the supervised approaches subgroup.

Group 2, which comprises unsupervised approaches, has a τ2 statistic of 6.67 with a P-value of 0.39. This indicates the presence of some heterogeneity among the studies within this subgroup, the between-study variance being estimated as 6.67. However, the P-value of 0.39 suggests that the observed heterogeneity is not statistically significant at the conventional alpha level of 0.05. This finding implies that, while there is some variability in the effect sizes within the unsupervised approaches subgroup, the variability is not sufficiently substantial to raise concerns about the consistency of the results.

Group 3, which includes semi-supervised methods, has a τ2 statistic of 0.00 with a P-value of 0.39. Similarly to Group 1, this indicates that there is minimal heterogeneity among the studies within this subgroup, the between-study variance being estimated as zero. The P-value of 0.39 suggests that the observed heterogeneity is not statistically significant, providing evidence for the consistency and homogeneity of the effect sizes within the semi-supervised methods subgroup.

Lastly, Group 4, which encompasses knowledge-based approaches, has a τ2 statistic of 38.15 with a P-value of 0.07. This indicates the presence of substantial heterogeneity among the studies within this subgroup, the between-study variance being estimated as 38.15. The P-value of 0.07, although not statistically significant at the conventional alpha level of 0.05, suggests a trend towards significant heterogeneity. This finding implies that there is considerable variability in the effect sizes within the knowledge-based approaches subgroup, which may warrant further investigation in identifying potential sources of this heterogeneity.

The results of the τ2 statistic and P-values for each subgroup provide valuable insights into the consistency and variability of the effect sizes within each WSD approach. The low τ2 values and non-significant P-values for Groups 1 and 3 indicate that the effect sizes within these subgroups are relatively homogeneous, suggesting that both the supervised and semi-supervised approaches yield consistent results across studies. In contrast, the higher τ2 value and the trend towards significant heterogeneity in Group 4 suggest that the effect sizes within

the knowledge-based approaches subgroup are more variable, which may require further exploration to understand the factors contributing to this heterogeneity.

3.10.7 Meta-regression

To investigate the relationship between the study's moderators and the effect size estimates from individual studies, a meta-regression was conducted. The meta-regression is an extension of subgroup analysis that allows for the examination of multiple moderators and their potential interactions in the meta-analysis. Figure 3.19 illustrates the meta-regression output.

```
  Effect-size label: Effect size
        Effect size: ES
          Std. err.: SE

Random-effects meta-regression                  Number of obs  =        55
Method: DerSimonian-Laird                       Residual heterogeneity:
                                                          tau2 =      10.7
                                                       I2 (%)  =      9.68
                                                          H2   =      1.11
                                                  R-squared (%) =     11.63
                                                Wald chi2(1)    =      0.89
                                                Prob > chi2     =    0.3465
```

_meta_es	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
Year	.3238997	.3440443	0.94	0.346	-.3504147	.9982142
_cons	-665.4267	693.602	-0.96	0.337	-2024.862	694.0083

```
Test of residual homogeneity: Q_res = chi2(53) = 58.68   Prob > Q_res = 0.2751
```

Figure 3.19: Meta-regression model

The tau2 statistic, which represents the between-study variance, was found to be 10.7. This value indicates the presence of residual heterogeneity among the included studies after accounting for the publication year in the meta-regression model. The I2 statistic, which quantifies the proportion of total variation in effect sizes due to heterogeneity, was estimated as 9.68%. This suggests that a relatively small proportion of the variability in effect sizes can be attributed to true differences between studies. The H2 statistic, which is a measure of the extent of heterogeneity, was found to be 1.11. The R-squared value of 11.63% represents the proportion of between-study variance that is explained by the moderator variable included in the meta-regression model. The prob>chi2 value of 0.3465 indicates that the overall meta-regression model is not statistically significant at the conventional alpha level of 0.05. The coefficient of 0.324 represents the change in effect size per one-unit increase in the publication

year. However, the standard error of 0.344 indicates that the coefficient estimate is not precise; and the confidence interval for the coefficient probably includes zero. This suggests that the relationship between publication year and effect sizes is not statistically significant. Figure 3.20 illustrates the relationship between the publication year which is the moderator variable, and the effect size estimates from individual studies.

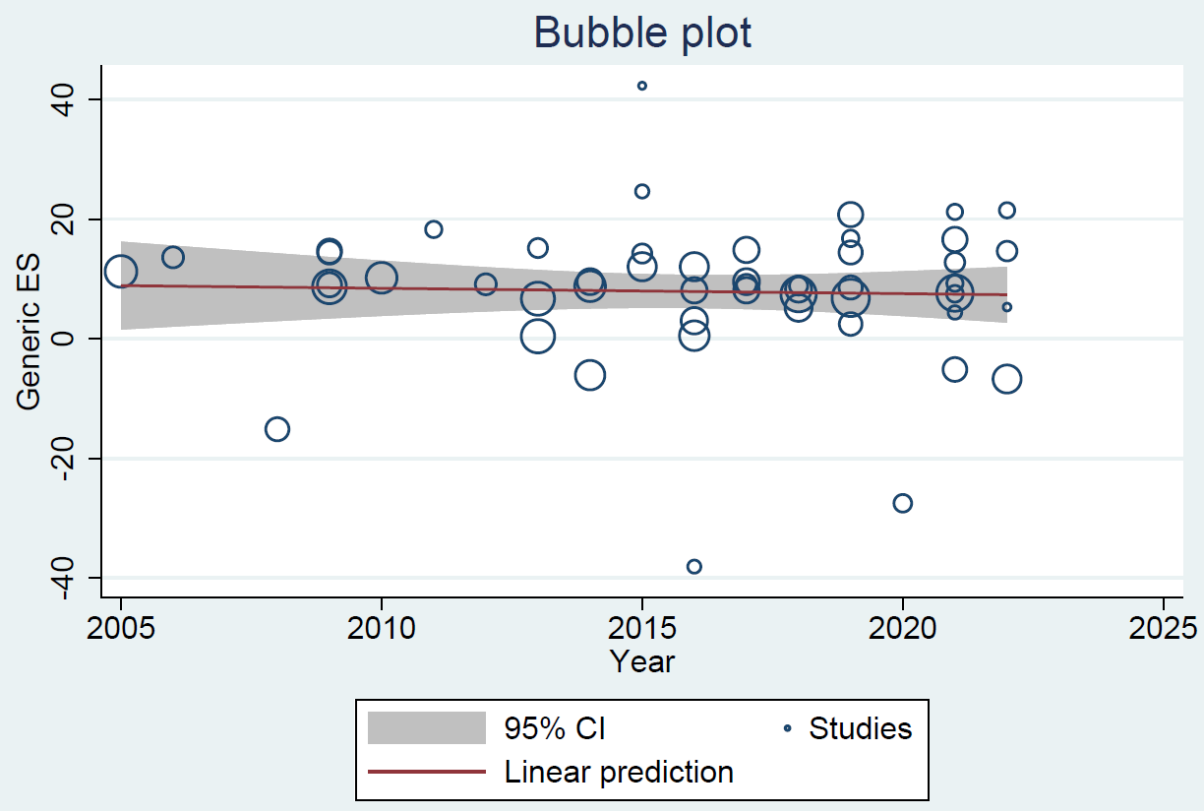


Figure 3.20: Meta-regression based on year

Each bubble in the bubble plot represents an individual study included in the meta-analysis. The size of each bubble is proportional to the precision of the study's effect-size estimate, which is determined by the inverse of the study's standard error or variance. Larger bubbles indicate more precise estimates, while smaller bubbles would indicate less precise estimates. The confidence band on the bubble plot indicates the range within which we would expect 95% of the true effect sizes to fall, given the observed data and the uncertainty in the meta-regression estimates. Studies that fall outside of the 95% CI indicate that the study's effect size estimate is significantly different from what would be expected based on the overall relationship between the moderator and the effect sizes, as estimated by the meta-regression model. Based on the meta-regression model and bubble plot, the publication year accounts for a relatively small proportion of the variability in effect sizes among the included studies.

3.10.8Publication bias

Publication bias refers to the systematic discrepancy between the research published in literature and the entirety of completed studies. Systematic reviews and meta-analyses acknowledge that it is challenging to completely eliminate publication bias (Mathur and VanderWeele 2021). The literature suggests assessing publication bias to ensure reliable conclusions about potential biases and their potential impact on the generalizability of findings. A funnel plot, depicted on Figure 3.21, was used to visually evaluate the publication bias in this study.

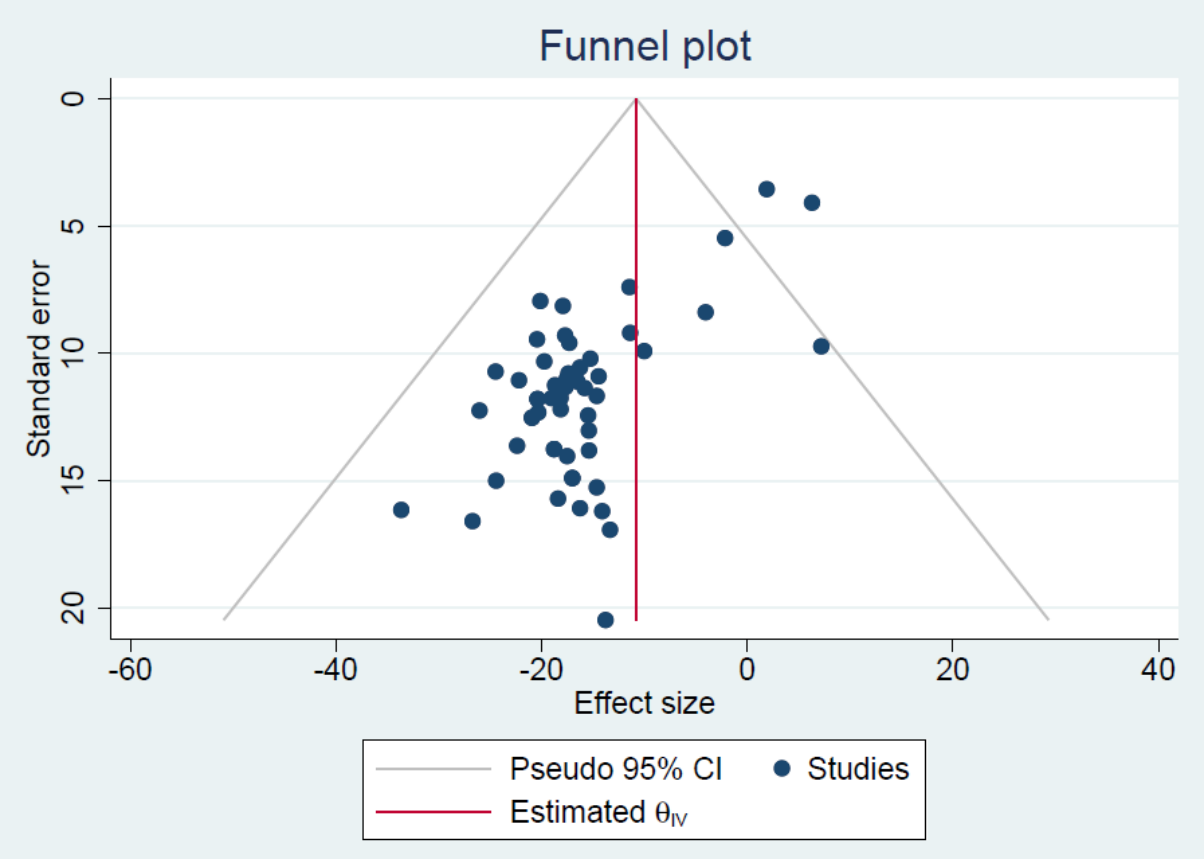


Figure 3.21: Funnel plot with pseudo 95% confidence limits indicating publication bias

According to the funnel plot in Figure 3.21, only two studies fall outside the 95% confidence limits on a funnel plot of a total of 55 studies. This suggests that there is publication bias in the dataset used to conduct the meta-analysis, implying that the distribution of studies is relatively balanced, with minimal evidence of publication bias. Such a symmetrical distribution around the average effect size estimate implies that smaller studies with less precision and potentially more variability are dispersed evenly, while larger studies with greater precision tend to cluster more tightly.

3.10.9 Meta-analysis results and discussion

The conducted meta-analysis followed the PRISMA framework to conduct a comprehensive and rigorous assessment of WSD methods. The objective was to identify the most effective WSD approaches, investigate potential sources of heterogeneity, and provide insights into factors influencing their performance. A total of 55 studies was included in the meta-analysis after a thorough screening process based on predefined inclusion and exclusion criteria.

The low between-study variability and true heterogeneity, as evidenced by the $\tau 2$ and I2 statistics, support the reliability and generalizability of the pooled effect-size estimate. The Galbraith plot in Figure 3.17 further reinforces the consistency of the findings, with only two studies falling outside the 95% confidence interval bands, suggesting low heterogeneity among the majority of the included studies.

The subgroup analysis, which categorized the studies into four groups based on the WSD approaches, employed provided valuable insights into the relative effectiveness of these techniques. Supervised approaches, Group 1, demonstrated the largest effect size, indicating their potential to yield substantial improvements in WSD performance. This finding aligns with the literature, which suggests that supervised methods often outperform other approaches. Conversely, unsupervised approaches, Group 2, exhibited the smallest effect size, suggesting a more limited impact on WSD effectiveness. Semi-supervised Group 3 and knowledge-based Group 4 approaches showed intermediate effect sizes, offering a balance between performance gains and reduced reliance on labelled data.

The subgroup analysis also highlighted potential sources of heterogeneity. The low $\tau 2$ values and non-significant P-values for Groups 1 and 3 indicate homogeneity within the supervised and semi-supervised approaches, suggesting consistent results across studies. The meta-regression analysis, which examined the relationship between publication year and effect sizes, revealed that publication year accounted for a relatively small proportion of the variability in effect sizes among the included studies. The bubble plot in Figure 3.20 further illustrates this finding, with studies dispersed across the plot without a clear trend based on publication year.

The assessment of publication bias using a funnel plot in Figure 3.21 indicated a relatively balanced distribution of studies, with only two studies falling outside the 95% confidence limits. This symmetrical distribution around the average effect-size estimate suggests minimal evidence of publication bias, enhancing the reliability of the meta-analysis findings.

In relation to the present study, whose objective is to develop a WSD model for Setswana in the context of Setswana-English machine translation, the findings of this meta-analysis provided valuable insights into the methodologies for WSD, Setswana being a low-resource language with currently no existing WSD models in literature. The subgroup analysis provided insights into the relative performance of various WSD approaches which guided the selection and adaptation of a suitable method for the Setswana-English WSD machine-translation context. The higher performance of supervised approaches indicates that investing in the creation of sense-annotated datasets for Setswana could yield substantial improvements in WSD accuracy. However, given the resource-constrained nature of Setswana, employing knowledge-based approaches that leverage existing linguistic resources and require less annotated data was a better option.

3.11 WSD Research Issues

In addition to the findings discussed in this chapter, it is crucial to acknowledge the current open problems and hindrances in the field of WSD that require further attention from researchers. Despite the progress made in developing and evaluating various WSD methods, several open problems persist, stalling the widespread adoption and effectiveness of WSD in real-world applications, particularly for native languages. Prevalent key open problems in WSD include:

3.11.1 Data sparsity

WSD is a critical task in natural language processing (NLP) that relies heavily on rich linguistic resources, such as annotated corpora, lexical semantic databases, thesauri, and ontologies (Navigli 2009). However, the availability and coverage of these resources for different languages and domains are limited, with the majority of WSD research focusing on a closed set of languages, primarily English. Also, the WSD problem becomes all the more obtuse in the NLP of a language, due to the dependence of the performance of most WSD models on the corpus available for the language. This limitation hinders the development and evaluation of WSD methods that can generalize effectively to low-resource languages and specialized domains. Additionally, the creation of sense-annotated datasets is a time-consuming and labour-intensive process, often requiring expert human annotators (Pasini 2021). In addition, the performance of most WSD models is highly dependent on the quality and quantity of the corpus available for the target language, making the WSD problem particularly complicated for under-resourced languages (Papini *et al.* 2021). Furthermore, there are divergencies in the

WSD benchmark evaluation datasets, which makes it arduous to perform a direct quantitative experimental comparison of WSD models (Raganato, Camacho-Collados and Navigli 2017). Therefore, developing more comprehensive and high-quality resources, especially for under-resourced languages and specialized domains, is a WSD research issue.

3.11.2 Domain adaptation

WSD models trained on non-specific data may not perform well in specific domains due to domain-specific language and terminology used in the domain (Hassanabadi *et al.* 2021). Adapting trained WSD models to new domains or fine-tuning them to specific domains is an ongoing research quest in improving the applicability and accuracy of WSD systems across various domains without retraining models.

3.11.3 Word sense granularity

Determining the appropriate level of granularity for word senses remains problematic in WSD. The research and distinction between fine-grained senses e.g., differentiating between various shades of meaning, and coarse-grained senses (e.g., grouping similar meanings together), is ongoing (Iacobacci, Pilehvar and Navigli 2016). Developing methods and techniques that can effectively handle both levels of granularity by the same model is a WSD research issue (Conia and Navigli 2021).

3.11.4 Multilingual WSD

Extending WSD methods and techniques to handle multiple languages increases the complexity of the task. Transferring knowledge from resource-rich languages to resource-poor languages, dealing with cross-lingual ambiguity, and addressing language-specific obstacles are problems in multilingual WSD research (Pasini, Scozzafava and Scarlini 2020).

Addressing these WSD research issues requires further advancements in research, algorithms, techniques, and resources for WSD. Researchers are actively exploring innovative approaches, leveraging deep learning, cross-lingual transfer learning, and incorporating rich contextual information to tackle these hurdles, thus improving the state-of-the-art in WSD (Bevilacqua *et al.* 2021). In addition, successfully addressing the quandaries requires collaboration between linguists and computer scientists. Linguists play a crucial role in providing linguistic expertise, insights into the nuances of language, linguistic properties of languages, such as syntactic and semantic structures, collocations, and idiomatic expressions, thus ensuring that WSD models accurately capture the intricacies of languages and word senses (Edmonds and Cotton 2001).

Linguists' contributions ensure that WSD models accurately capture the intricacies of languages and word senses, leading to more reliable and effective WSD systems.

3.12 Gap in the Literature

The absence of a WSD model for Setswana, as indicated in the HLT audit, and the open problems in WSD research mentioned above, highlight a significant gap in the literature. This gap primarily revolves around the lack of resources available for underrepresented languages, particularly native languages, and those from smaller linguistic communities or under-resourced regions.

A review of the existing literature reveals that the majority of WSD models and datasets have been developed and evaluated on major languages. The bibliometric analysis conducted in this study provides compelling evidence of this disparity. The United States, as an English-speaking country, occupies the first position in terms of the number of publications on WSD research, followed by China and India. Strikingly, none of the top 10 countries with the most published research on WSD is located in Africa. This finding underscores a substantial gap in research, particularly concerning African languages in general, and South African languages, specifically.

The scarcity of research focusing on languages with fewer resources is a critical issue that hinders the development of inclusive and diverse language technologies. This lack of attention not only limits the advancement of NLP applications for these languages but also perpetuates the digital divide between well-resourced and under-resourced languages. Consequently, speakers of these underrepresented languages may face barriers to accessing and benefitting from language-based technologies, further widening existing inequalities.

3.13 Summary

In conclusion, this literature review chapter has provided a comprehensive overview of the existing knowledge and research findings in the field of WSD. The current WSD methods, knowledge resources, and evaluation techniques were reviewed; previous published studies were critically analysed; bibliometric and meta-analysis were conducted to identified gaps, limitations, and trends in the literature, laying the groundwork for further research in WSD. The initial sections of the literature review introduced the concept of WSD, its applications in NLP, the various approaches, resources used and evaluation methods in WSD. This foundational knowledge provides a clear understanding of the field and its significance in language-processing tasks. The review of related work encompassed a wide range of studies

conducted in different contexts, employing diverse methodologies and data sources. Through this synthesis, key findings and empirical evidence from previous research were identified, allowing for a comprehensive understanding of the current state of the field. The bibliometric analysis section employed quantitative methods to analyse scientific publications and bibliographic data, revealing patterns, trends, and relationships within WSD research. The findings emphasize the need for increased support and collaboration to bridge the gap in WSD research between sub-Saharan Africa and other regions, ultimately contributing to the development of inclusive and diverse language technologies. A meta-analysis of WSD methods was conducted to examine the publication bias of the WSD methods, and identification of studies that contribute most to the heterogeneity of WSD studies. The findings highlight the effectiveness of supervised approaches, while also acknowledging the potential of semi-supervised and knowledge-based techniques. This corroborates with research in literature; however, due to the absence of annotated datasets for training supervised models, these approaches cannot be used in the context of this study.

CHAPTER FOUR: THEORETICAL FRAMEWORK

4.1 Introduction

This chapter presents the theoretical framework of this thesis. The objective of the theoretical framework chapter is to provide a foundational review of theories, models, and related concepts that form the basis for the methodology used in this study. This chapter is structured as follows: Section 4.2 introduces the distributional semantic hypothesis and explores the distributed representations of words, highlighting essential theories and concepts related to distributed word representation. The Lesk algorithm and its variations are presented in Section 4.3. The bidirectional encoder representations from transformers (BERT) and the transformer architecture are presented in Section 4.4. The lexical knowledge bases and the models used to guide the development of the Setswana resources are presented in Section 4.5. This includes the Princeton WordNet (PWN), Universal Knowledge Core (UKC), African WordNet (AWN), and the bilingual dictionaries used for data extraction. Section 4.6 presents the rule-based machine-translation (RBMT) architecture adopted for Setswana-English machine translation. The similarity measures used in NLP are presented in Section 4.7. Lastly, Section 4.8 details how the various components interconnect to form an overall framework proposed in this research that guides the research methodology. The chapter is summarized in Section 4.9.

4.2 Distributed and Distributional Representations

Natural language is essentially a discrete symbolic representation of human knowledge. Machine learning (ML) and natural language processing (NLP) advances have necessitated the replacement of this discrete symbolic representation with vectors or tensors called distributed and distributional representations, the latter being an approximation of the former (Ferrone and Zanzotto, 2020).

4.2.1 Distributional semantics and hypothesis

Distributional semantics is an approach to semantics that is based on the contexts of words in large corpora. Semantic similarity is the basic notion formalized in distributional semantics. Word embeddings are the modern incarnation of distributional semantics, adapted to work well with deep learning. Distributional semantics aims to describe meaning of words and sentences with vectorial representations. Thus, distributional semantics is an important area of NLP

research, being foundational to embedding models (Ferrone, L. and Zanzotto, F.M., 2020). The survey in (Turney and Pantel, 2010) lends credence to this. Distributional/distributed representations use distributional properties of linguistic entities as the building blocks of semantics, relying fundamentally on a set of assumptions about the nature of language and meaning, referred to as the distributional hypothesis (Harris, 1970). The distributional hypothesis is often stated thus: "words which are similar in meaning occur in similar contexts" (Rubenstein & Goodenough, 1965). In the distributional methodology, distributional facts establish the basic entities of language and the (distributional) relations between them (Harris, 1970). Members of the basic classes of these entities behave distributionally similarly, and therefore can be grouped according to their distributional behaviour, as Harris opined. For example, if two linguistic entities, w_1 and w_2, tend to have similar distributional properties, and they occur with the same other entity w_3, it can be posited that w_1 and w_2 belong to the same linguistic class. Word distributions have a causal relationship with the way meaning is derived from the text. In fact, this theory entails that the statistical distribution of a word causes the semantic representation humans have of the corresponding idea. Distributional semantics models rely on statistical analysis and linear algebraic tools to implement the theory. Latent semantic analysis (LSA) provides the mathematical tool of distributional semantics. LSA is an unsupervised learning method that allows for parsing several texts, analysing similarities between them. The LSA idea is to regard words as points in a so-called distributional space, which is simply a vector space, in which vectors represent words. The vector contains an average of all possible different word usages deduced from the input text corpora (Sahlgren, 2008).

4.2.2 Distributed/distributional representations

Distributed/distributional representations (also known as embeddings) are concepts fundamental in ML and NLP. Embeddings refer to a way of representing data, typically words or phrases, as continuous vectors in a high-dimensional space, capturing a notion of similarity and semantic meaning by allowing an entity to be represented by a pattern of values across many dimensions (Ferrone and Zanzotto 2020). Distributed representations are vectors or tensors of real numbers representing the meaning of words, phrases, and sentences. Vectors for words are obtained observing how these words co-occur with other words in document collections, using corpora (Ferrone and Zanzotto 2020). The fundamental principle of distributional representation of words is based on distributional hypothesis which posits that words with similar distributions have similar meanings (Sahlgren 2008). Distributional

semantics as an area of NLP aims to describe meanings of words and sentences with vectorial representations. Distributional vectors represent words by describing information related to the contexts in which they appear. Thus, distributional representation can be considered a special case of distributed representation (Ferrone, and Zanzotto, 2020).

Through the use of numeric encoding techniques, computers can interpret and perform operations on data efficiently and accurately. Therefore, representing natural language text in numerical form is crucial. This has been a key area of research in NLP. Traditional term-based approaches such as the term frequency-inverse document frequency (TF-IDF) fall short of capturing the semantic meaning of words; and have limitations in the semantic understanding and analysis of natural languages. As a result, methods based on the distributional hypothesis have been introduced.

A number of models such as the Word2Vec (Mikolov, 2013), GloVe (Pennington, 2014) and FastText (Bojanowski, 2017) based on the distributional hypothesis have been proposed and widely used in NLP. Word2Vec is a self-supervised learning model that has been trained on a continuous bag of words (CBOW) and skip-gram methods. The CBOW applies to a word predicted based on its context; while the skip-gram method applies to contexts predicted based on an input word. This model produces distributed word representations through the use of a shallow neural network. GloVe is an unsupervised model that learns word representations by combining global corpus statistics with local window contexts to perform semantic analysis. FastText learns representations for subwords known as character n-grams via CBOW and skip-gram models. FastText uses a bag of character; and n-grams is used to obtain the representation of a word.

While GloVe has been proven to outperform Word2Vec on several benchmarks, one major drawback that cuts across all the models is context independence. Ambiguity and context are inherent to natural languages, therefore it is critical to convey this contextual information to computers and machine-learning algorithms, in the language model (LM). To address the limitation of context independence in Word2Vec, GloVe, and FastText, embeddings from language models (ELMo) was introduced. ELMo is an unsupervised model that uses a bidirectional long short term memory (BiLSTM) to produce contextual text representations (Tenney *et al.* 2019). The model captures contexts in both directions from the internal states of the BiLSTM neural network. As effective as BiLSTM neural networks are in capturing contextual information in natural languages, transformer neural networks have emerged as a

powerful alternative, offering enhanced capabilities in handling the contextual nature of natural languages. The transformer neural network, which is based on distributed/distributional representation, and used in this research to addresses the context-independence limitation of the earlier LMs, is presented in the next section.

4.3 The Lesk Algorithm

The Lesk algorithm, first introduced by Michael E. Lesk in 1986, is a widely used approach WSD (Banerjee and Pedersen 2002). The algorithm is based on the idea that words in a given context tend to share common words, topic, or semantic field. The algorithm disambiguates a target word by selecting the sense whose dictionary definition has the highest overlap with the context words surrounding the target word (Lesk 1986). The original Lesk algorithm laid the foundation for various subsequent variations and extensions aimed at improving its performance and addressing its limitations.

4.3.1 Original Lesk algorithm

The original Lesk algorithm disambiguates a target word by comparing the dictionary definitions of its senses with the definitions of the context words. The Lesk algorithm follows these steps (Lesk 1986):

a. For each sense of the target word, calculate the overlap between its dictionary definition and the definitions of the context words.

b. Select the sense with the highest overlap score as the most appropriate sense for the target word.

The overlap score is typically computed as the number of common words between the sense definition and the context word definitions. The algorithm relies on the assumption that the correct sense of the target word is likely to have a higher overlap with the definitions of the surrounding context words (Lesk 1986).

4.3.2 Simplified Lesk algorithm

The simplified Lesk algorithm, proposed by (Kilgarriff and Rosenzweig 2000), is a variant that reduces the computational complexity of the original algorithm. Instead of comparing the sense definitions with the definitions of the context words, it compares the sense definitions directly with the context sentence (Pal, Maiti and Saha 2013). The simplified algorithm works as follows:

a. For each sense of the target word, calculate the overlap between its dictionary definition and the context sentence.

b. Select the sense with the highest overlap score as the most appropriate sense for the target word.

This simplification reduces the number of comparisons required and makes the algorithm more efficient, especially for larger contexts. Building upon this approach, the proposed model in this study adopted the Simplified Lesk to develop the PuoBERTa Enabled Embedding-based Lesk WSD Model. This enhanced version transforms the algorithm into an embedding-based Lesk, which uses vector representations of words to capture semantic relationships. Instead of calculating overlap, this model employs similarity measures to assess semantic similarity, leveraging the power of word embeddings to improve the accuracy and efficiency of word sense disambiguation.

4.3.3 Adapted Lesk algorithm

The adapted Lesk algorithm, introduced by (Banerjee and Pedersen 2002), extends the original algorithm by considering not only the dictionary definitions, but also the examples and other related information provided in the lexical resources. The algorithm follows these steps:

a. For each sense of the target word, create a sense bag containing its dictionary definition, examples, and other related information (e.g., synonyms, hypernyms).

b. Create a context bag containing the words in the context sentence or paragraph.

c. Calculate the overlap score between each sense bag and the context bag.

d. Select the sense with the highest overlap score as the most appropriate sense for the target word.

The adapted Lesk algorithm aims to leverage additional information from the lexical resources to improve disambiguation accuracy (Banerjee and Pedersen 2002).

One of the major drawbacks of the Lesk algorithm is its computational complexity, which stems from the exponential growth of comparisons needed for numerous candidate senses associated with polysemous words across various lexical resources. The simplified and adapted Lesk's were introduced to overcome this limitation.

4.4. Bidirectional Encoder Representations from Transformers

Bidirectional encoder representations from transformers (BERT) were developed by researchers at Google (Devlin, 2018) as a part of ongoing efforts to advance natural language understanding capabilities. The model was introduced to address challenges posed by unidirectional language models, aiming to capture bidirectional dependencies and improve contextual understanding in natural language processing tasks. By using the transformer neural network (TNN), BERT provides a more comprehensive and nuanced understanding of contextual relationships between words.

4.4.1 Transformer neural network architecture

Figure 4.1 shows the architecture of the transformer neural network (TNN). The TNN architecture relies on self-attention mechanisms.

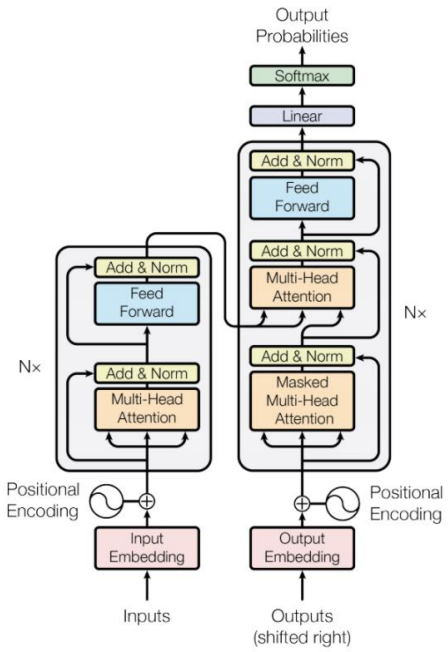


Figure 4.1: The transformer model architecture Vaswani et al. (2017)

As shown in Figure 4.1, the main components of the transformer neural network are the input embedding, positional encoding, encoder and decoder, multi-head attention, masking, and the output. Each of these is explained in the sections that follow.

4.4.1.1 The input embedding block

The fundamental process of any machine-learning process comprises three main stages: input, processing, and output. Preparing input for the transformer architecture starts with initializing an input embedding. In the transformer encoder, the input data is processed as a whole without sequential single passing. All words within a sequence or sentence (input) are simultaneously passed through the model, generating embeddings concurrently. Figure 4.2 depicts the input process encoder of a transformer model architecture.

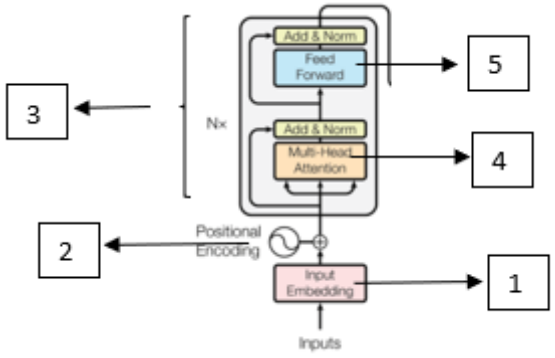


Figure 4.2: Transformer model input embedding eVaswani et al. (2017)

The input embedding, referred to as the "input embedding" in Block 1 in Figure 4.2, is responsible for mapping words to vector representations. However, with natural language, it is essential to consider that the same word can possess distinct meanings in different contexts of usage in sentences. This is where positional encoders, labelled in Block 2 of Figure 4.2, serves the purpose of carrying out context analysis, which results in contextual vectors. By leveraging positional encoders, contextual information is incorporated based on the word's position within a sentence. The generation of these contextual vectors involves using a sine and cosine function, as expressed in the following equations.

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

Where *pos* refers to the position of the word in the sentence, *i* refers to the dimension index of the positional encoding vector, while d_{model} represents the dimensionality of the model (embedding size). The positional encodings are added to the word embeddings to provide contextual information about the word's position in the sentence. By incorporating these positional encodings, BERT distinguishes words with the same token representation but occurring at different positions, thus capturing important positional information during the

model's training and inference processes. The issue of the context-independence limitation of earlier language models raised in Section 4.1 is therefore addressed.

4.4.1.2 Multi-head attention

Following the positional encoding step in which information about the position of words within the sentence or input sequence is captured, the multi-head attention labelled in Block 4 of Figure 4.2 captures the contextual relationships between words. The main purpose of the multi-head attention is to process various types of contextual relationships and dependencies between words in a sentence to allow the model to capture relevant information from different contexts and provide more expressive representations. In this step, the input embeddings are transformed into queries, keys, and values through linear projections in which the projections are performed multiple times with different learned weight matrices, resulting in multiple sets of query, key, and value representations, known as attention heads. From this process, each attention head operates independently and computes attention scores between the queries and keys. The attention scores determine the importance or relevance of each word in the sentence to other words. In addition, the attention scores are used to weight the values, which represent the encoded information of each word. The resulting outputs of all attention heads are concatenated and linearly transformed to obtain the final attention output. This process and output allow BERT to capture the various aspects of the input by attending to several parts of the sentence concurrently.

4.4.1.3 Feed-forward networks

The purpose of feed-forward networks labelled in Block 5 of Figure 4.2 is to introduce non-linear transformations and to enhance the representation of each word independently within the context of the sentence. The feed-forward networks take the output of the multi-head attention mechanism as input, and apply the linear and non-linear transformation and the linear projection as pre-process mechanisms. The linear transformation step occurs when the input representation is projected to a higher-dimensional space using a linear layer, thus allowing the model to capture more complex patterns and relationships in the dataset. The non-linear transformation step is seen when the projected representation is passed through a non-linear activation function to introduce non-linearities, allowing the model to capture complex dependencies and non-linear patterns in the dataset. The last pre-processing step, linear projection, takes the output of the activation function and projects it back to the original dimensionality using another linear layer. This linear projection assists in transforming the representation to a suitable format for the next layers and tasks.

4.4.1.4 Layer normalization

The “Add and Nom” operation labelled in Block 3 of Figure 4.2 plays a significant role in stabilizing and enhancing the flow of information through the encoder layers. This operation refers to the combination of element-wise addition and layer normalization. After passing through the multi-head attention mechanism in Block 4, and the feedforward neural network in Block 5 in each encoder layer, the output must be combined with the residual connection from the input. The “Add” is the addition operation in which the output of the feedforward neural network is added element-wise to the input of the encoder layer. Such allows the model to retain the original information from the input, while incorporating the new information learned during the attention and feedforward neural computations. On the other hand, the “Nom” is the layer-normalization operation in which layer normalization is applied to the combined output. The “Nom” normalizes the values along the feature dimension, ensuring that the output of each encoder layer has a consistent distribution. The “Add” helps mitigate the problem of vanishing gradients and provides a smoother flow of information; the “Nom” helps in stabilizing the training process and improving the model’s generalization. These two operations maintain a strong flow of information while alleviating the potential issues of vanishing or exploding gradients.

4.4.1.5 The output embedding block

The input embedding block in BERT model architecture is responsible for mapping the input tokens vector representations; and the output embedding block is the final layer that maps the contextualized representations of the input tokens to a specific task. Figure 4.2 depicts the output embedding block.

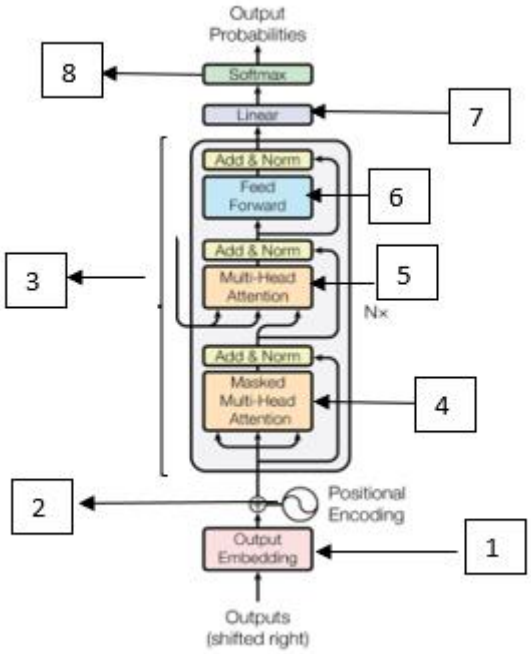


Figure 4.3: Transformer model output embedding Vaswani et al. (2017)

The output embeddings are the contextualized representations of the input tokens obtained from the input embedding block. The positional encoding of the output embedding operates in a similar manner to the positional encoding of the input embedding block providing the unique vector representations for each position in the input sequence.

4.4.1.6 Masked multi-head attention

The masked multi-head attention labelled in Block 4 of Figure 4.3 phase is processed in the pre-training phase in which a small percentage of the input tokens are randomly masked, by replacing them with a special [MASK] token. The masked multi-head attention mechanism is used to predict the original identities of these masked tokens. In this step, the model attends to the unmasked tokens to predict the masked tokens. The attention scores are computed between the masked token and all other tokens, allowing the model to capture contextual relationships and dependencies necessary for accurately predicting the masked tokens. Upon completion of the pre-training phase, the output-embedding block of BERT is used to fine-tune the model to various downstream tasks. The output-embedding block mainly involves task-specific layers and computations based on the requirements of the particular task at hand, which in our research is WSD.

4.4.1.7 Linear and softmax functions

The multi-head attention, feed-forward neural networks, and the addition/normalization phases function in a similar manner as described earlier for the input embedding. The linear layer labelled in Block 7 of Figure 4.3 projects the contextualized representations of the input tokens to a different dimensional space. The linear layer helps learn task-specific patterns and representations by mapping the high-dimensional input representations to a desired output space. This transformation enables the model to extract relevant features and capture task-specific information from the BERT representations. The softmax layer labelled in Block 8 of Figure 4.3 is applied after the linear layer and converts the output of the linear layer into probabilities across different classes. The softmax function normalizes the scores for each class, ensuring that they add up to 1. This allows for the interpretation of the output as a probability distribution over the classes, indicating the model's confidence or likelihood for each class. The class with the highest probability is typically selected as the predicted class during inference.

4.5 Lexical Resources Development Models

Natural language processing requires lexical resources for language model training and testing. The quantity and quality of lexical resources available for a language have a bearing on the performance of NLP tasks and application systems for the language. This section examines some models of lexical resources development that are of relevance to this study.

4.5.1 Bilingual dictionary model

Dictionaries are lexical resources wherein information is organized primarily alphabetically without any semantic relations between words. Bilingual dictionaries translate words and phrases from one language to another. Various dictionaries employ diverse organizational structures. Some dictionaries offer one-to-one mappings of word translations in two or more languages. Others provide word translations and word classes, such as nouns, verbs, and adjectives. Additionally, certain dictionaries offer word translations, word classes, definitions, and examples, while some combine all three elements in their entries. In the context of this study, bilingual dictionaries played a crucial role as valuable resources for extracting data, contributing to the creation of datasets that underpin linguistic analyses and WSD modelling.

4.5.2 WordNet model

A WordNet is a lexical database of semantic relations between words. The WordNet represents a lexical resource model that links words into semantic relations including synonyms, hyponyms, and meronyms. The synonyms are grouped into synsets with short definitions and usage examples. The WordNet can thus be seen as a combination and extension of both a dictionary and a thesaurus. Since the advent of the English Princeton WordNet (Miller, 1995) as a model of lexical resource for NLP, several WordNets of different world languages and language groups, have emerged (Bond, Kuribayashi, and Fellbaum, 2018). These include the African WordNet, Arabic WordNet, Hindi WordNet, Euro-WordNet, BabelNet, and so on. In this subsection, WordNets that are relevant to this study as lexical resources are considered.

4.5.2.1 The Princeton WordNet (PWN) model

Princeton WordNet (PWN) is an on-line lexical reference system whose design is inspired by psycholinguistic theories and principles of human lexical memory (Miller *et al.* 1990). English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Various relations link the synonym sets. The PWN has revolutionized the traditional dictionaries in which lexical information is organized primarily alphabetically without any semantic relations identified between the information contained in the dictionary. Another reason for the development of the PWN was the introduction of the age of computers in which lexical databases that can be read by computers were a necessity. During this time, instead of merely converting traditional dictionaries to online dictionaries, the PWN was proposed as an effective combination of traditional lexicographic information and modern high-speed computation (Miller *et al.* 1990).

The PWN was developed by a group of psychologists and linguists at Princeton University whose objective was to construct a lexical database that would provide an aid in searching dictionaries conceptually, as opposed to alphabetically. WordNet divides the lexicon into five lexical categories, namely, nouns, verbs, adjectives, adverbs, and function words. Nouns are words that represent people, places, things, or concepts. Nouns serve as a building block of WordNet as they form parts of sentences, functioning as either subjects or objects. Figure 4.4 illustrates the sense for the noun “hand” in WordNet.

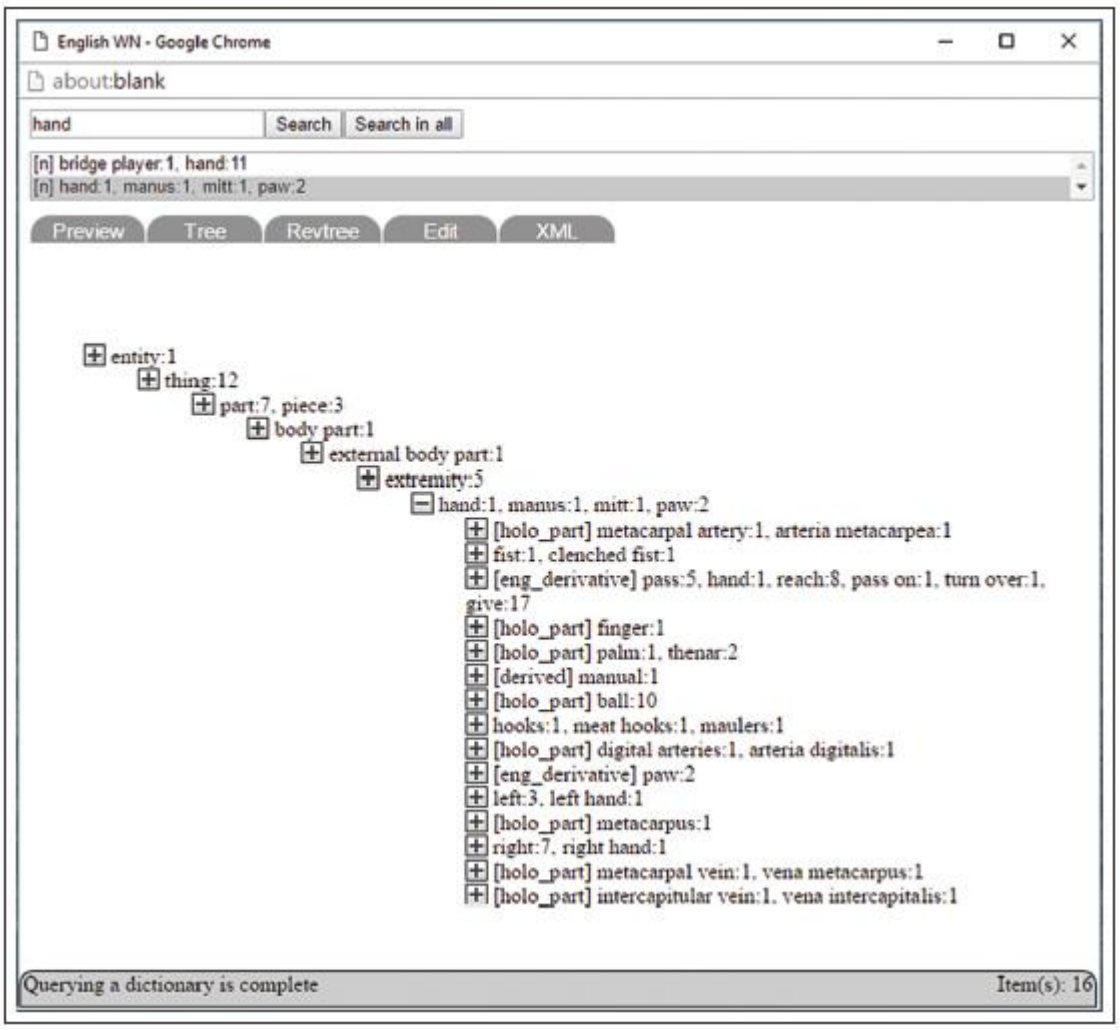


Figure 4.4: Noun “hand ” in WordNet

Figure 4.4 illustrates the semantic nature of WordNet and the hierarchical relationships it captures. The various types of relations are described in detail in the following paragraphs. Verbs express actions and describe in the sentence what subjects either do or experience. Adjectives modify or describe nouns, providing additional information about their qualities, characteristics, or attributes, while adverbs modify verbs, adjectives, and other adverbs. Adverbs provide additional information about how an action is performed, or the circumstances in which it occurs; and lastly, function words, also known as grammatical words or structure words serve a grammatical or syntactic role in a sentence. Function words are stored separately as part of the syntactic component of language. Figure 4.4 represents hierarchical relationships that WordNet captures for nouns; however, the same relationships for verbs, adjectives, and adverbs, are captured in the same way.

Relations play a significant role in WordNet in that they establish connections with distinct concepts. These relations enable the representation of semantic associations, and provide a comprehensive understanding of how different word categories are interlinked within the WordNet framework. Synonymy is the primary relationship among words in WordNet with connecting words such as “shut” and “close”, or “car” and “automobile”. Synonyms represent words with the same meaning. Commonly, the words that belong to the same synset have interchangeable usage in various contexts. Princeton WordNet consists of 117,000 synsets, each connected to other synsets through a small number of conceptual relations. Furthermore, each synset includes a concise definition referred to as gloss, along with one or more example sentences illustrating the word usage. In WordNet, words with multiple meanings are mapped to distinct synsets. This is to ensure that each form-meaning pair within WordNet remains unique.

The most commonly observed relationship among synsets in WordNet is the super-subordinate relationship, also known as hyperonymy or hyponymy, which represents the “ISA” (is-a) relationship. This relationship connects more general synsets, such as [furniture] and [piece_of_furniture], to increasingly specific items, such as [bed] and [bunk bed]. As a result, WordNet establishes that the category of furniture includes beds, which in turn includes bunk beds. Similarly, concepts such as bed and bunk bed fall under the broader category of furniture. In WordNet, all noun hierarchies trace back to the root node [entity]. The hyponymy relation exhibits transitivity, meaning that if an armchair is a type of chair, and a chair is a type of furniture, then an armchair is also considered a type of furniture. WordNet makes a distinction between types (common nouns) and instances (specific individuals, countries, and geographic entities). For example, an armchair is classified as a type of chair; while Barack Obama is an instance of a president. Instances always occupy the leaf or terminal nodes in their respective hierarchies.

Meronymy, the part-whole relation stands between synsets such as [chair] and [back, backrest], [seat] and [leg]. Parts are inherited from their superordinates: if a chair has legs, an armchair also has legs. Parts are not inherited “upward” because parts may be characteristic only of specific kinds of things rather than the class as a whole; hence chairs and kinds of chairs have legs; but not all kinds of furniture have legs. Verb synsets in WordNet are also organized into hierarchies. Within these hierarchies, verbs located towards the lower levels represent increasingly specific ‘manners’ that characterize an event. For example, the hierarchy [communicate]-[talk]-[whisper] demonstrates a progression from a general act of

communication to a more specific manner of speaking with lowered volume. It is important to note that the specific manner expressed by troponyms (verbs towards the bottom of the hierarchy) depends on the semantic field. Volume, as shown in the previous example, is just one dimension along which verbs can be elaborated. Other dimensions include speed (e.g., move-jog-run) or intensity of emotion (e.g., like-love-idolize). WordNet also establishes links between verbs that describe events that necessarily and unidirectionally entail one another. For instance, verbs such as [buy] and [pay], [succeed] and [try], [show] and [see], are interconnected to signify the inherent relationship between these actions. Lastly, adjectives are organized in terms of antonymy. Pairs of "direct" antonyms such as hot-cold and happy-sad reflect the strong semantic contract of their members. Each of these polar adjectives in turn is linked to a number of "semantically similar" adjectives: hot is linked to balmy, boiling, flaming, and baking; and cold to chilly, frosty, etc. Semantically similar adjectives are "indirect antonyms" of the contra member of the opposite pole. Figure 4.5 shows the structure of the XML synset model in a typical WordNet-LMF (Lexical MarkupFramework)

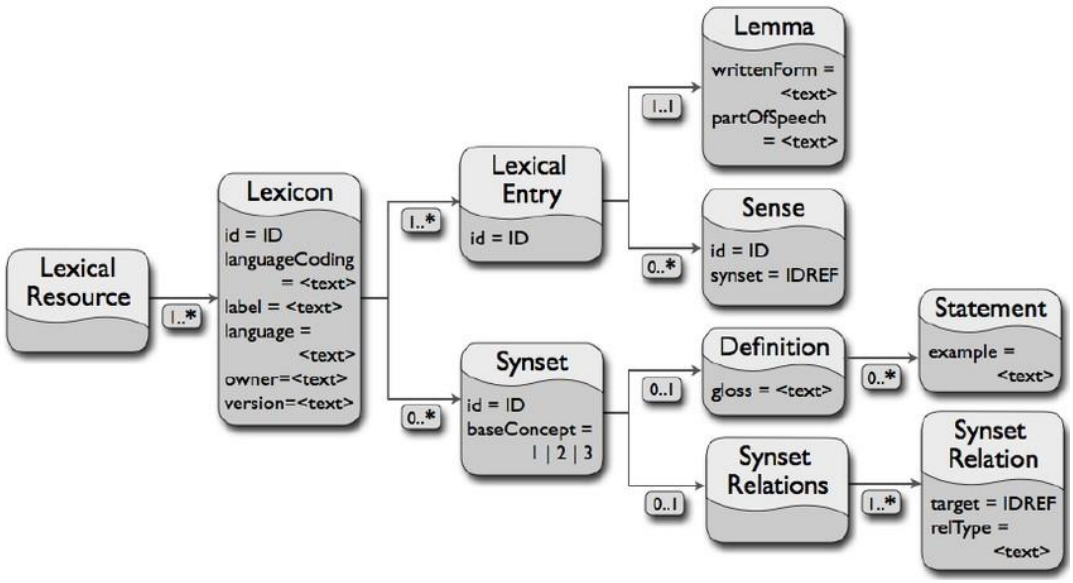


Figure 4.5: Structure of the XML synset model in WordNet-LMF (Henrich, 2010)

WordNet serves as a valuable resource in NLP, providing rich lexical knowledge; and facilitating various NLP tasks and applications such as WSD, machine translation, text summarization, lexical semantics, inter alia. The rich semantic relations captured in WordNets are part of the contextual information exploited in enhancing the performance of NLP tasks and application systems.

4.5.2.2 The African WordNet (AWN) model

The African WordNet (AWN) project was initiated with the aim of promoting multilingualism and facilitating the development of language tools and resources for South African (SA) languages. In 2007, international experts conducted a training workshop for linguists, lexicographers, and computer scientists, introducing WordNets for African languages, with a specific focus on SA languages (Bosch, 2017). Currently, WordNets have been developed for Setswana (tsn), isiXhosa (xho), isiZulu (zul), Sesotho sa Leboa (nso), and Tshivenda (ven), Sesotho (sot), Xitsonga (xho), isiNdebele (nde) and Siswati (ssw)

The AWN was developed using the expanded model that is based on the structure of the PWN and various development strategies. According to Bosch and Griesel (2017), the expanded model provides a tested structure on which to build a new lexical resource; and it is appropriate for less resourced languages. One major drawback of the expanded method is that it assumes that the new language shares an underlying structure with PWN. This is not always true given the difference in morphology and orthography between English and the African indigenous languages.

The first stage of development was linguists manually identifying synsets from the PWN and translating them to the five local languages. The focus in this first stage was noun parts included in PWN, nouns making up the bulk of the lexicon; and this would enable the WordNets to grow at a steady pace. The linguists followed the 3-part structure of the PWN which comprises a localized noun synset, a definition, and a usage example for each of the five languages. Figure 4.6 illustrates an AWN Setswana synset for “seatla” meaning “hand” on the WordNet editor DEBVisDic.

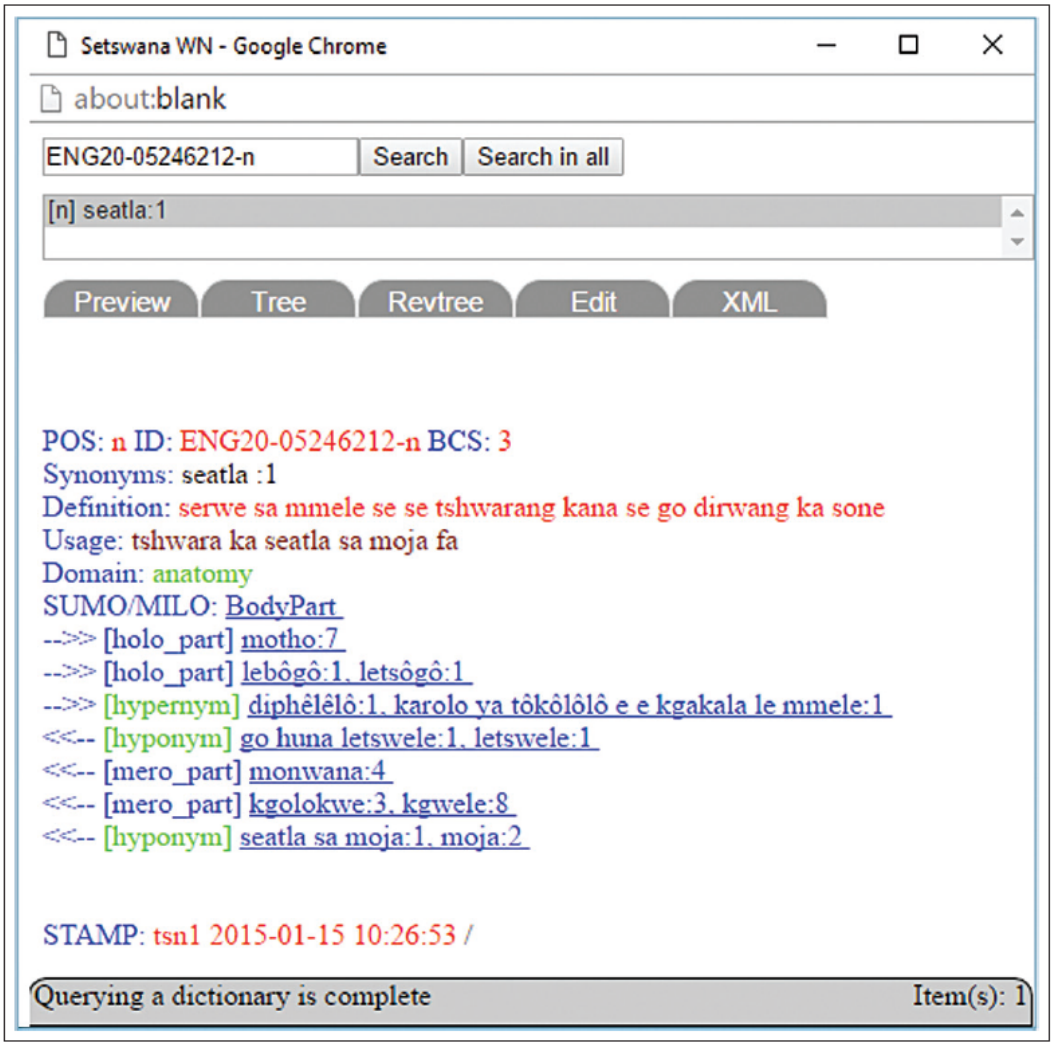


Figure 4.6: Setswana synset for “seatla” (Bosch and Griesel 2017).

The synset consists of the definition, usage example, domain, and semantic relations for *seatla*. While the development of WordNets for African languages is an ongoing project, presently, the AWN has 53 982 synsets, 9279 definitions, and 28 853 usage examples for the five mentioned SA languages. Table 4.1 shows the AWN statistics.

Table 4.1: The status quo of the data contained in the AWN (Bosch and Griesel 2017). The developed and localized synsets are significant given the data scarceness of the SA languages resources, and the time-consuming manual labour dedicated to the constructions of the resources.

The project development was guided by various developmental strategies. The first strategy was identifying concepts and synsets that are used frequently and in a broad spectrum of domains. For this step, a list of Princeton core concepts was used to identify the concepts that

should be included in the WordNet first. In addition to the Princeton core concepts, a list of common base concepts from the EuroWordNet Project and BalkaNet Project was also used.

Table 4.1: The Status Quo of the Data Contained in the AWN, as of 2017 (Bosch and Griesel 2017)

Language	Synsets	Definitions	Usage examples
Northern Sotho (NSO)	8412	1178	5253
Tshivenda (VEN)	4270	209	4270
IsiZulu (ZUL)	10 782	2179	5112
IsiXhosa (XHO)	14 715	2198	7015
Setswana (TSN)	15 803	3515	7203
Total	**53 982**	**9279**	**28 853**

The list of common base concepts was "the fundamental building blocks for establishing the relations in a WordNet and give information about the dominant lexicalization patterns in languages" (Bosch and Griesel 2017). The organic growth strategy was utilized as the second approach, aiming to incorporate language-specific concepts and synsets that reflect the unique language and its culture, rather than solely translating the English common base concepts. With this strategy, linguists began by incorporating the most typical sense of a frequently used word and allowed that sense to guide the expansion to subsequent senses. This organic approach resulted in, for example, the Setswana component of the AWN encompassing numerous figurative meanings and unforeseen relationships. The third strategy was studying corpus frequencies. To test this approach, a multilingual parallel corpus, including all 11 official SA languages was acquired from the Resource Management Agency in which a frequency list for Tshivenda was extracted and the 5000 most frequent terms in the multilingual African wordlist compared with the list of the English base and core concepts acquired from the first strategy. The extracted frequency list included concepts and synsets that represent the unique African language usage.

An additional objective of the AWN project was to include a set of synsets that will be shared across all languages involved. The goal was to facilitate cross-linguistic semantic and syntactic analysis. To achieve this, the fourth strategy, semantic domains, was used. The existing structure of the WordNet was leveraged as a semantic ontology, incorporating the hierarchical structures of the suggested upper merged ontology (SUMO) and the mid-level ontology (MILO). These frameworks provided a foundation for categorization that was both machine-readable and easily understood by human interpreters. The final strategy employed is referred to as linking. To establish multilingual connectivity for all the languages in the AWN, each synset is linked to the PWN using a unique identification code (ENG ID).

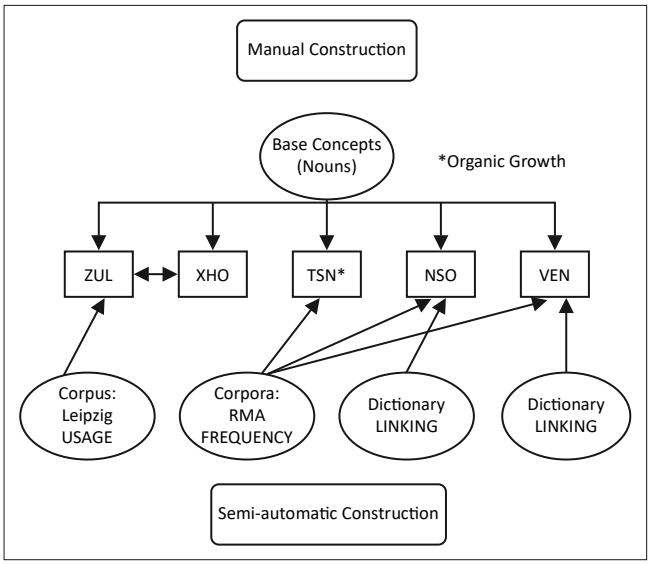


Figure 4.7: Summary of the various components used in the development of AWN

By leveraging available electronic resources, such as bilingual dictionaries, potential synset links in the PWN were identified. For example, the Sesotho sa Leboa dictionary entry for *'almond tree'* was matched with the English lemma in the PWN, extracting the corresponding ENG ID and definition. A spreadsheet was created to list these potential matches, allowing linguists to indicate true matches and provide novel usage examples. Linguists evaluated the matches by indicating whether the definition aligned with the dictionary entry. A "yes" or "no" indication was used, and linguists provided novel usage examples for each matched synset. The provided information from linguists was automatically transferred to an appropriate XML database format, incorporating additional details from the PWN such as SUMO and MILO classifications, domain, and hierarchical structure. Figure 4.7 shows the summary of the various components used in the development of AWN (Bosch and Greisel 2017) .

Table 4.2: PWN, EWN and AWN Statistics

	Princeton WN	English WN	African WN
Synsets	117,791	120,054	53 982
Lemmas	159,015	163,079	53982
Relations	378,203	383,825	38132

Table 4.2 illustrates the disparity in data size between Princeton WordNet and English WordNet in comparison to the African WordNet. The count of synsets and relations in the AWN is notably smaller when contrasted with the two lexical resources. This highlights the data sparsity inherent in African languages compared with English. The AWN lexical resource has been posited as a viable tool for WSD (Madonsela et al. 2016), however, no study has explored this direction. Hence, the present study.

4.5.3 The universal knowledge core (UKC) model

To design our language resource, a multilingual lexical resource, the universal knowledge core (UKC) proposed by Giunchiglia, Batsuren and Freihat (2018) was adopted. The UKC is organized into three layers depicted in Figure 4.8. The UKC adapts the PWN model. However, in the UKC, the synsets which in different languages codify the same meaning, are clustered into language agnostic concepts. Additionally, the UKC semantic relations link concepts, and not synsets, as in the PWN, and create a language-independent semantic network.

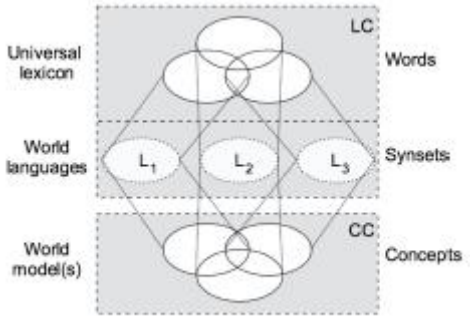


Figure 4.8: The UKC structure (Giunchiglia 2018)

Figure 4.8 shows the three layers in the UKC model. The first layer from bottom-up reflects the world models, the concepts core (CC). The concepts core comprises a semantic network of

language-independent concepts as nodes. Each concept possesses a distinct identifier that sets it apart from other concepts. Within the semantic network, various semantic relations such as hyponymy and meronymy connect the relations of these concepts.

The integrated second and third layer, the language core (LC) which encapsulates the synset layer encompasses words, senses, synsets, glosses, and examples that are backed by the UKC. The LC and PWN fulfil similar functions. In the PWN, each synset in the LC is uniquely associated with a particular language; and, within that language, with at least one word. However, unlike the PWN, synsets in the LC are interconnected with concepts. Additionally, there is a requirement that a concept can only be created if it is lexicalized in at least one language. Considering the multilingual nature of the UKC, there exists a one-to-many relationship between concepts and synsets. Figure 4.9 illustrates the connection between synsets and concepts.

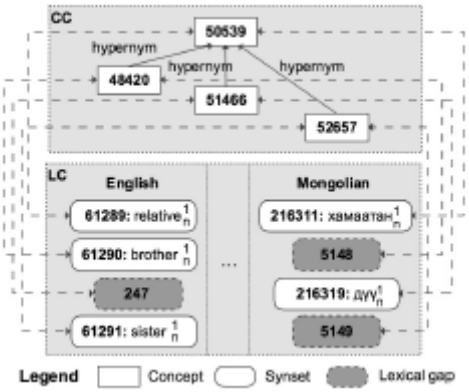


Figure 4.9: A fragment of the semantic network of concepts and their synsets (Giunchiglia 2018)

The LC layer corresponds to the PWN; however, an additional feature of the LC is that within this layer, each synset is uniquely associated with a specific language, and, within that language, it is connected to at least one word. In contrast to the PWN, synsets in the LC are linked to concepts; and there is a strict requirement that each synset be associated with one and only one concept. An additional constraint in place for concept creation is that, to create a concept, it is necessary for there to be at least one language in which the concept is represented by lexicalized terms. Due to the multi-lingual nature of the UKC, there is a one-to-many relationship between concepts and synsets. Figure 4.9 shows how synsets and concepts are related. This research adapted the UKC model to construct the localized lexical resource for Setswana, the Setswana universal knowledge core (SUKC).

4.6. Machine Translation: Rule-Based Approach and Theoretical Framework

Machine translation (MT) is the automatic conversion of text of speech from a source natural language to a target language, using machine-learning algorithms. Given the inherent highly complex nature of natural languages such as word ambiguities, word grammatical relations in one language that might not exist in another language, language syntactic differences, the rich morphology of certain languages and the differing orthography amongst languages, MT becomes a challenging task.

4.6.1 The rule-based approach

Various types of MT approaches have been proposed over the years since the inception of automatic language translation. MT systems can be classified according to their core methodology. Within this categorization, there are three primary approaches ─ rule-based machine translation (RBMT), statistical machine translation (SMT), and neural machine translation (NMT) (Kituku, 2016). Since this study adopts a rule-based MT approach, the following section will focus on rule-based and the theoretical framework of RBMT.

RBMT is an approach to MT that relies on explicit linguistic rules and patterns to translate text from one language to another. In the rule-based approach, human experts specify a set of linguistic rules to outline the translation process. There are three methods in the RBMT approach, which are, direct (or Dictionary/Lexicon-based), transfer-based, and interlingua methods. Figure 4.10 illustrates the three methods. The figure illustrates the progressive level of analysis required as we move towards the apex of the triangle representing the three methods.

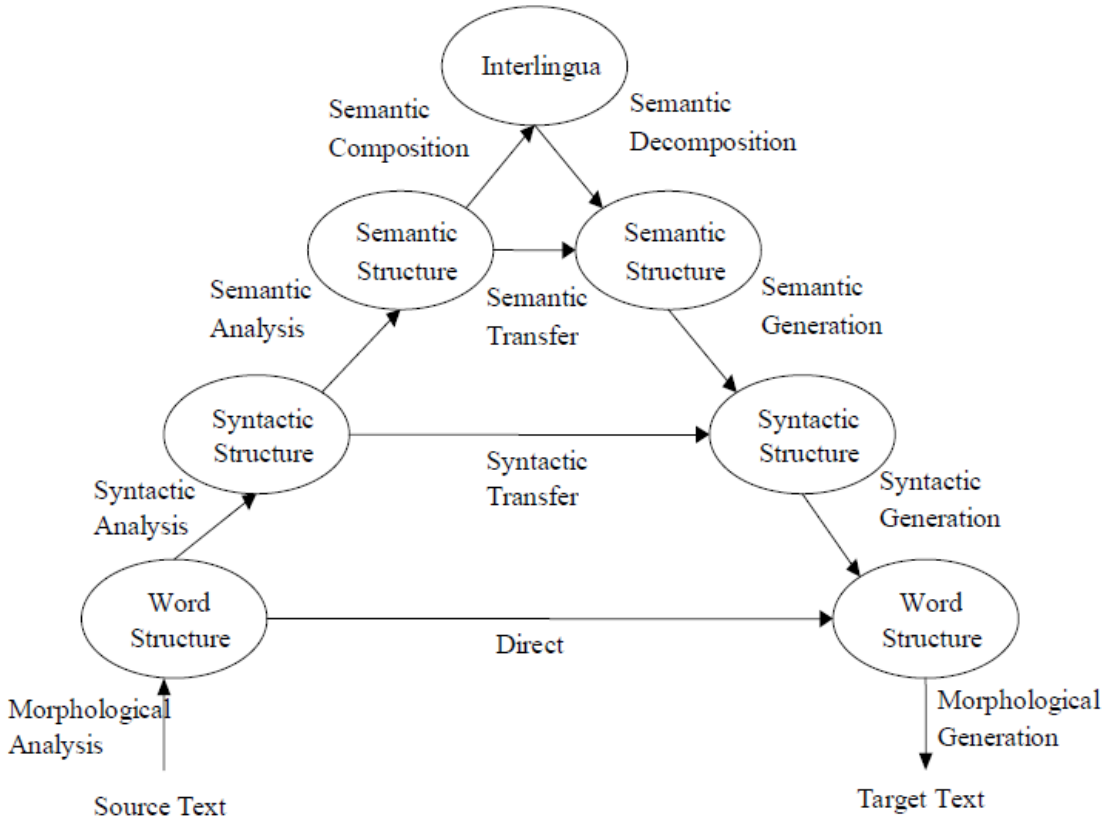


Figure 4.10: The Vauquois triangle for RBMT methods (Vauquois, 1968)

In the direct (also known as the dictionary/lexicon-based) method, translation rules are constructed to directly map source-language segments to target-language segments. The mappings can be word to word or phrase to phrase, from source language to their corresponding target language equivalents without extensive linguistic analysis. The key components of the direct method in RBMT are lexical mapping, limited linguistic analysis, rule precedence, domain-specific rules, and word/phrase translation. Lexical mapping encompasses translation rules; and defines mappings between specific source-language words or phrases and their target language equivalents. These mappings can be based on one-to-one correspondences or include variations, such as considering inflections, word-order changes, or differing lexical choices. In the direct method, there is minimal linguistic analysis compared with other RBMT methods. Instead of focusing on complex syntactic or semantic analysis, the emphasis is on direct lexical and phrase-based translations. This simplifies the rule-creation process and reduces the computational complexity of the translation system. Rule precedence is crucial for the RBMT direct method, in which multiple rules match a particular source language segment; the rule with the highest precedence is applied to generate the translation. Rule prioritization is important to ensure accurate and contextually appropriate translations. Domain adaptation is

fundamental in that systems employing the direct method can incorporate domain-specific rules to handle terminology, jargon, or specific language usage in a particular body of knowledge. These rules help improve translation quality in specific domains by providing more precise and accurate translations. Lastly, in the word/phrase translation, translation rules are designed to handle specific source-language phrases, and provide the corresponding translations in the target language. These rules can cover common idiomatic expressions, collocations, or frequently occurring phrases.

The transfer RBMT method leverages linguistic knowledge and structural correspondences between source and target languages to achieve accurate translations. This method involves analysing the source sentence, extracting its syntactic and semantic information, and applying transformation rules to generate the corresponding translations. The main components of the transfer method are source-language analysis, linguistic rules, structural mapping, semantic transfer, lexical selection, and rule composition. The transfer method begins with certain linguistic analysis such as part-of-speech tagging, syntactic parsing, and semantic analysis of the source language input to identify the grammatical structure, word dependencies, and semantic roles of the source sentence. Linguistic rules are a set of predefined linguistic elements, such as noun phrases, verb phrases, and syntactic dependencies used to transform the analysed source-language structures into their target-language equivalents. The structural mapping component involves identifying equivalent syntactic constructions, word order patterns, and grammatical relationships in the target language guided by the predefined linguistic rules that capture the structural correspondences between the two languages. In addition to the structural mapping component is the semantic transfer that incorporates the preservation of meaning of the source sentence during the translation process. This is achieved by applying semantic rules that capture the semantic relationships between the words and phrases in the source and target languages, together with their roles and constraints. The lexical selection component handles ambiguities that might occur within the translation process. The lexical selection includes disambiguation rules based on the context, or the use of lexical semantic resources in choosing the most suitable translation equivalents. Lastly, the rule composition allows for the integration of various linguistic rules and constraints applied sequentially or hierarchically, based on the analysis of the source sentence, thus ensuring accurate and coherent translations output.

The interlingua RBMT method is the pre-eminent method used in RBMT to generate translations by utilizing an intermediate language representation, the interlingua, that serves as

a neutral, language-independent representation. The interlingua captures the underlying meaning and structure of the source language sentence. The six components of this strategy are interlingua representation, source-language analysis, interlingual transfer, target-language generation, structural mapping, and lexical selection, lastly, disambiguation and coherence. The interlingua representation involves creating a formal representation of the meaning and structure of the source language sentences. The aim of this representation is to store the essential linguistic and semantic feature information, while being independent of specific source or target languages. The source language analysis component's function is similar to that of the transfer strategy, which is mainly to perform part-of-speech tagging, syntactic parsing, and semantic analysis. In the interlingual transfer, the analysed source-language sentence is used to transfer the linguistic information to the interlingua representation. This transfer process involves mapping the analysed source-language structures to their corresponding interlingua representations. The interlingua captures the underlying meaning and structure of the source sentence in a language-independent manner. Following the interlingual transfer is the target-language generation component that applies transformation rules and the mapping of the interlingua structures and semantics to their corresponding structures in the target language for translation generation. The transformation rules specify how the interlingua representations should be transformed into the target language. The fifth component, structural mapping and lexical selection includes identifying equivalent syntactic constructions, word-order patterns, and grammatical relationships in the target language. This basically involves mapping the interlingua source structures to their corresponding target structures. The lexical selection rules are applied to choose appropriate target-language words or phrases based on the interlingua representation. Ultimately, there is the disambiguation and coherence component that addresses ambiguities arising from the source language and ensures coherence in the generated translations. The ambiguities are resolved using the information available in the interlingua representation governed by defined rules and constraints to maintain the coherence and consistency of the translations.

4.6.2 The rule-based approach theoretical framework

The RBMT theoretical framework encompasses the foundational principles and concepts that guide the design and operation of RBMT systems. The framework involves various linguistic and computational components that work together to perform accurate translation from source language to target language. This section outlines the key elements of the theoretical framework for RBMT. The main elements included in the framework include linguistic rules, lexical

resources, parsing and analysis, rule application, disambiguation, and error handling; and lastly, rule development and maintenance. Linguistic rules are the core of the RBMT system determined by linguists and domain experts. These rules define the grammar, syntax, semantics, structural patterns, word order, and other linguistic properties necessary for translation. There cannot be RBMT without lexical resources (Kituku, 2016). These resources include dictionaries, thesauri, and WordNets, that provide information about word meanings, translations, synonyms, and other lexical properties. Linguistic rules are used in conjunction with lexical resources to determine the appropriate translation equivalents for words and phrases. The parsing and analysis element of the framework consists of parsing algorithms and linguistic processors used to identify parts of speech, to parse sentence constituents, and apply grammatical and syntactic rules, enabling the system to generate appropriate target-language structures and translations. Following the analysis is the rule application in which linguistic rules are executed in a predefined order, with each rule such as pattern matching, rule selection, and rule composition used in transforming a specific linguistic structure or pattern into its target-language counterpart. Ambiguity is handled in this step. This is where homonyms, polysemous words, or sentence structure are disambiguated based on contextual information or further rules while performing error handling. Error-handling rules identify and rectify translation errors based on predefined error patterns or heuristics. The final element of the framework is the rule development and maintenance. This is an iterative process of rule expansion and refinement to improve translation quality. Figure 4.11 illustrates the architecture of a RBMT.

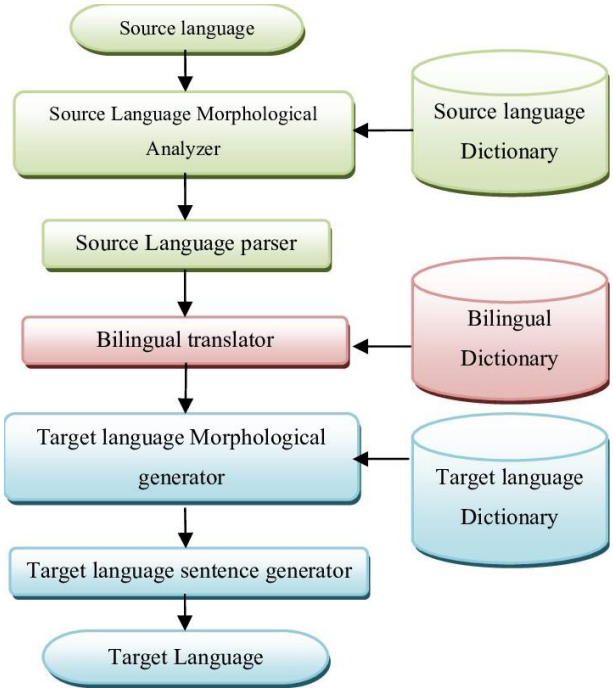


Figure 4.11: Architecture of the RBMT paradigm (Kituku, 2016).

The morphological analyser analyses the source text and provides the morphological information of the words in the source text. The output of this step is the root form of the analysed words. The part of speech tagger tags the output of the morphological analyser with appropriate part of speech tags. Lexical selection for the equivalent target lexemes of the source text is performed, then the lexical and morphological transfer is applied. The target language morphological generator works as a generator of appropriate target language words for the given grammatical information, while the post generator functions as a composer of suitable target-language sentences.

4.7. Similarity Measures

The distributional methodology based on the distributional hypothesis introduced in Section 4.2 is concerned with semantic similarity which is about meaning similarity (Sahlgren, 2008). Text semantic similarity measurement is the core of NLP tasks. According to Lin (1998), semantic similarity is defined as the semantic commonness between two text inputs. The greater the semantic commonness, the higher the similarity, and the lower the semantic commonness, the lower the semantic similarity. Various text semantic similarity techniques have been proposed and employed in NLP. These techniques can be classified into text representation and distance, subdivided into various methods (Wang and Dong 2020). Figure 4.12 illustrates the categorization of the methods.

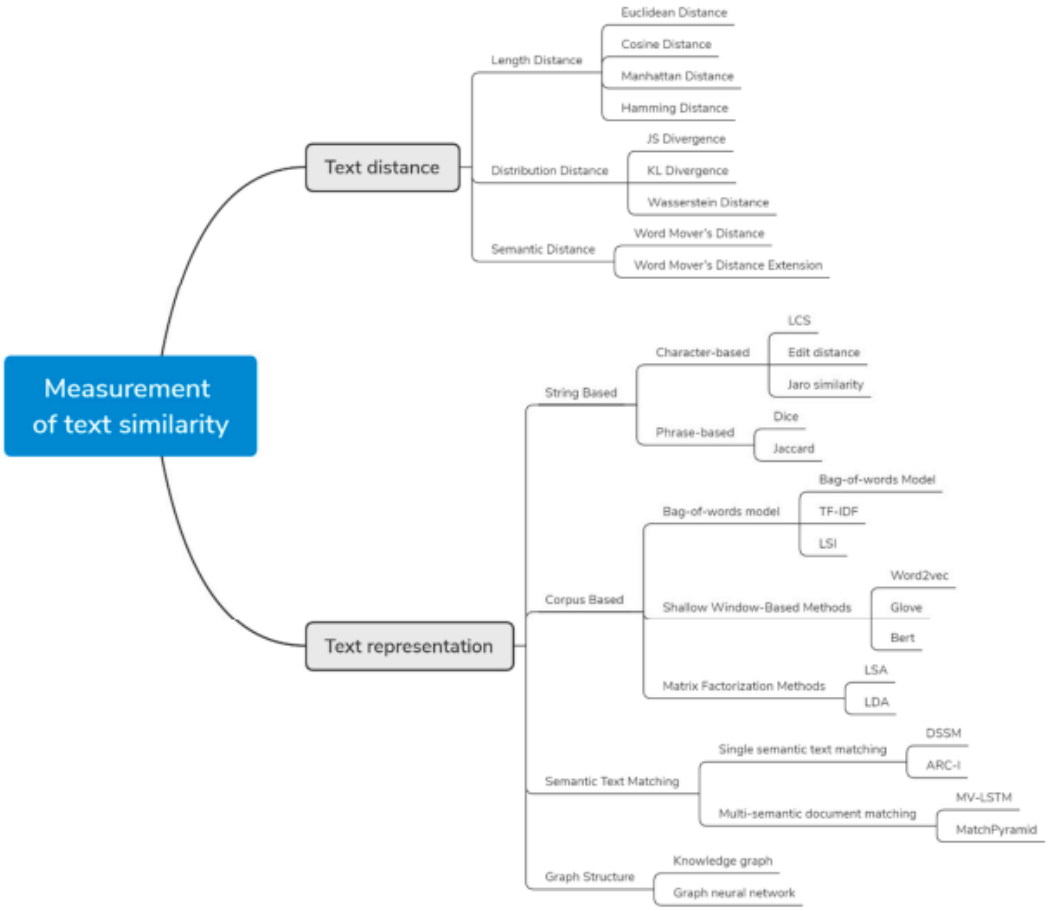


Figure 4.12. Measurement of text similarity categorization (Wang 2020).

4.7.1 Text-distance similarity

The text-distance similarity measure calculates the semantic proximity between two texts units, focusing on their relative distances. This measure enables the quantification of the proximity of the words in terms of the meanings and contextual relationships. Analysing the distances between the words enables an understanding of how closely or distantly related the words are within the given text. This measure consists of three distinct approaches for measuring distance, taking into account the length, distribution, and semantics of the analysed text.

4.7.1.1 Length distance

This method involves measuring length distance between the two texts inputs. The method leverages the numerical characteristics of text to calculate the length-based distance of text vectors. This approach quantifies the differences or similarities in the lengths of textual elements; and allows quantitative comparison. The distance-measures approaches are described below:

4.7.1.2 Euclidean distance

Euclidean distance is a widely used mathematical measure that quantifies the distance between two points in a multi-dimensional space (Deza *et al.* 2009). Mathematically, the Euclidean distance is the straight line distance between two points in Euclidean space. The equation for calculating the Euclidean distance between two points in a multi-dimensional space is as follows:

$$d(S_a, S_b) = \sqrt{\sum_{i=2}^{n} (S_a^{(i)} - S_b^{(i)})^2}$$

Where *a* and *b* represent the coordinates of the two points in the multi-dimensional space. As the equation shows, Euclidean distance is computed by taking the square root of the sum of the squared differences between the corresponding coordinates of the two points. In the context of text-similarity-measure applications, Euclidean distance can be employed to compare the feature vectors of two text inputs. Each text is represented as a vector on vector space, in which each dimension corresponds to a specific feature or characteristic. By calculating the Euclidean distance between the two vectors, the similarity or dissimilarity between the texts based on their feature values is determined. A smaller Euclidean distance indicates a higher degree of similarity, while a larger distance suggests greater dissimilarity.

4.7.1.3 Cosine distance

Cosine distance, also known as cosine similarity, is a measure that quantifies the similarity or dissimilarity between two vectors in a multi-dimensional space. Contrary to measuring the distance between the two points, the cosine similarity measure transforms the measure into the angle problem corresponding to the two points in the vector space .The cosine distance between two vectors, *a* and *b*, is calculated using the cosine of the angle between them. The equation for cosine similarity is:

$$Sim(S_a, S_b) = \frac{S_a . S_b}{||S_a||. ||S_b||}$$

Where "$S_a \cdot S_b$" represents the dot product of vectors *a* and *b*, which involves multiplying the corresponding components of the vectors, thereafter adding them. The notation "$||S_a||$" and "$||S_b||$" denotes the magnitude or Euclidean norm of vectors *a* and *b*, respectively. The cosine-distance similarity measure allows effective comparison of the similarity between vectors,

including text documents represented as vector representations, determining their degree of either relatedness or dissimilarity (Deza *et al.* 2009).

4.7.1.4 Manhattan distance

Manhattan distance is calculated as the sum of the absolute differences between the two vectors in a multi-dimensional space. This measure is named after the way a taxicab moves through city blocks, where it can only travel along orthogonal directions. The Manhattan distance works only if the points are arranged in the form of a grid. With this technique, the problem at hand gives more priority to the distance between the points only along with the grids, not the geometric distance. The formula for Manhattan distance is as follows:

$$Sim(x, y) = |x_1 - x_2| + |y_1 - y_2|$$

The Manhattan distance between two points, *a* and *b*, is calculated by adding the absolute differences between their corresponding coordinates. As opposed to the Euclidean distance, which measures the straight-line distance between two points, Manhattan distance follows a path along the axes of the coordinate system (Deza *et al.* 2009).

4.7.1.5 Hamming distance

Hamming distance is a metric used for comparing two binary data strings inputs. This technique measures the minimum number of substitutions required to change one string into another by replacing individual characters. While comparing two binary strings of equal length, Hamming distance is the number of bit positions in which the two bits are different. The Hamming distance between two strings, *a* and *b*, is calculated by comparing corresponding characters at each position and counting the number of positions in which the characters differ. The formula for Hamming distance is as follows:

$$HM = \sum A[i] \neq B[i]$$

The Hamming distance between two strings, a and b, is denoted by *A[i]*, which represents the character at position *i* in string *a*. *B[i]* represents the character at position *i* in string *b*; and Σ denotes the sum over all positions *i*. The Hamming distance is equal to the total count of positions in which the characters in the two strings are different (Norouzi, Fleet and Salakhutdinov 2012).

4.7.1.6 Distribution distance

Distribution distance refers to a measure of dissimilarity or divergence between two probability distributions. This technique quantifies how different or similar two distributions are, based on certain criteria. According to Wang and Dong (2020), there are two difficulties with using length distance to calculate similarity. The first problem is that length distribution distance is suitable for symmetrical problems, such as Sim *(A, B) = Sim (B, A),* but for question *Q* to retrieve answer *A*, the corresponding similarity is not symmetrical, and not based on the length distance between two points. The authors further state that the second problem is that there is a risk in using length and distance to judge similarity, without knowing the statistical characteristics of the data being used. There are several methods of measuring distribution distance, and the choice of method depends on the specific context and requirements of the analysis. The distribution distance approaches are described below:

a) Jensen-Shannon divergence

Jensen-Shannon divergence, abbreviated as JS divergence, is a symmetrized version of the Kullback-Leibler (KL) divergence. This measurement quantifies the similarity between two distributions by measuring their average divergence from the average of both distributions. The formula for JS divergence is as follows, assuming that two documents belong to two different distributions *P1* and *P2*:

$$JS(P_1||P_2) = \frac{1}{2}KL(P_1||\frac{P_1 + P_2}{2} + \frac{1}{2}KL(P_2||\frac{P_1 + P_2}{2}$$

To calculate the JS divergence, the KL divergence between each distribution and their average must be computed, before taking the average of those two KL divergence values, weighted by 0.5. This approach is commonly used in conjunction with latent Dirichlet allocation (LDA) to compare the topic distributions of new documents with the topic distributions of documents in a corpus. This ranges from 0 for identical distributions to 1 for completely dissimilar distributions (Wang and Dong 2020).

b) Kullback-Leibler divergence

Kullback-Leibler (KL) divergence measures the information lost when one distribution is used to approximate another. KL divergence is sometimes referred to as relative entropy in which the various degrees of a probability distribution and a second reference probability distribution are measured. The formula for KL divergence is as follows:

$$d(p||q) = \sum_{i=1}^{n} p(x) \log \frac{p(x)}{q(x)}$$

where *d(p||q)* represents the KL divergence between distributions *P* and *Q* represent the probabilities of the events *x*. The logarithm used in the KL divergence formula can be taken with Base 2 or Base e, a natural logarithm, depending on the specific application or context.

c) Wasserstein distance

Wasserstein distance is a measure of the distance between two probability distributions (Weng 2019). This technique is also known as earth mover's distance (EMD). The technique calculates the optimal transport cost between the distributions and provides a non-negative value. The equation for EMD can be represented as follows:

$$W(p_r, p_q) = inf_{y \sim \pi(p_r, p_q)} E_{(x,y)\sim y}[||x - y||]$$

This formula finds the optimal transportation plan that minimizes the total cost or work required to transform distribution *Pr* into distribution *Pq*, satisfying the mass conservation constraints; while *(x, y)* states the percentage of dirt that should be transported from Point *x* to Point *y* so as to make *x* follow the same probability distribution of *y* (Vallender 1974). EMD is commonly used in image processing and computer vision.

4.7.1.7 Semantic distance

Semantic distance refers to a measure of either similarity or dissimilarity between the meanings or semantic representations of words, phrases, sentences, or other linguistic units (Kusner *et al.* 2015). This measure is used when there is no common word in the text, therefore the similarity obtained by using the distance measure based on length or distribution may be relatively small. In this instance, the calculation of the distance is performed at the semantic level, quantifying the conceptual or semantic difference between these units based on their inherent meaning or semantic relationships (Wang and Dong 2020). The main method used to determine semantic distance is the word mover's distance (WMD) (Kusner *et al.* 2015) explained below.

a) Word mover's distance (WMD)

WMD quantifies the dissimilarity based on the semantic meaning of the words and their relationships. This technique is derived from the concept of earth mover's distance (EMD) or Wasserstein distance (Kusner *et al.* 2015), which calculates the minimum amount of "work"

required to transform one distribution into another. The equation for word mover's distance can be defined as follows:

$$WMD(d_1, d_2) = min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij(i,j)}$$

The *WMD(D1, D2)* represents the word mover's distance between documents *D1* and *D2* with constraints for mass conservation of the documents to measure the minimum distance required for a word in one text to reach a word in another text in the semantic space, so as to minimize the cost of transporting text 1 to text 2 (Wu *et al.* 2018). WMD provides a non-negative value, where a smaller distance indicates a higher semantic similarity between the documents, while a larger distance indicates greater dissimilarity.

b) Word mover's distance extension

WMD has been extended in several ways to improve its applicability to different contexts. WMD extension uses the improved version of the distance known as the Mahalanobis distance, as opposed to Euclidean distance (De Maesschalck, Jouan-Rimbaud and Massart 2000). The Euclidean distance takes into account every dimension in a space as the same weight, meaning that there are no varying weights for any of the dimensions. However, this measure does not consider the correlation between the dimensions, which the Mahalanobis distance does.

4.7.2 Text representation

Text representation approaches represents the text as numerical features that can be quantified and calculated. Texts or strings can be similar in two ways, namely, lexically, and semantically. Lexical similarity refers to a similar character sequence of words combined to form a text. Lexical similarity focuses on the similarity between words based on their surface form or linguistic features; while semantic similarity focuses on the meaning or semantic content of words, phrases, or sentences. Lexical similarity assesses the relatedness or resemblance of linguistic units based on their semantic representation or conceptual similarity. Lexical similarity is measured through various measurements of text representation. Semantic similarity is measured through the string-based, corpus-based, semantic-text-matching, and graph-structure-based methods described below.

4.7.2.1 String-based similarity

The string-based similarity measures are techniques used to quantify the similarity or dissimilarity between two texts of characters. The techniques used compare the strings based

on their structural or sequence-based characteristics, without considering the semantics or meaning of the strings. As shown in Figure 4.12, there are two distinct methods used to classify character-based methods and phrase-based methods.

4.7.2.2 Character-based similarity

A character-based similarity calculation is based on the similarity between characters in the text, expressing the similarity between texts (Wang and Dong 2020). Longest common substring (LCS), editing distance, and Jaro similarity are the three most commonly used techniques in character-based similarity. The LSC focuses on identifying the longest contiguous sequence of characters that is shared between two strings. Its similarity measure is calculated as follows:

$$LCS = (Length\ of\ the\ longest\ common\ substring) \\ / (Average\ length\ of\ the\ two\ strings)$$

By dividing the length of the longest common substring by the average length, the LCS similarity measure is normalized to a value between 0 and 1, indicating the degree of similarity (Irving and Fraser 1992). The edit distance, on the other hand, represents the minimum number of transformations such as insertions, deletions, or substitutions required to convert the string from S_a to S_b. This technique captures the similarity between two strings, based on the number of operations needed to make them identical. The edit distance similarity measure is calculated as follows:

$$ED = 1 - (Edit\ Distance\ /\ Max\_Length)$$

The edit distance represents the minimum number of edit operations needed to transform one string into another. The max_length is the length of the longer string of the two being compared. The two forms of definition of editing distance are L-distance by Levenshtein (1966); and the D-distance by Damerau (1964). The L-distance transformation operations only include delete, insert, and replace operations; and the operations of D-distance include delete, insert, replace, and adjacent exchange.

The Jaro similarity quantifies the similarity between two strings S_a and S_b based on their character matching and transpositions. The Jaro similarity takes into account the order of characters and the similarity of characters in the compared strings. For two strings S_a and S_b, the Jaro similarity representation is as follows:

$$Jm = (m / S_a + m / S_b + (m - t) / m) / 3$$

The m represents the number of matching characters between the two strings S_a and S_b ; t represents half the number of transpositions between the matching characters. The three key factors of Jaro similarity are character matching, character order, and transpositions. While Jaro provides a way of measuring the similarity between strings, it does not consider the semantics or meaning of the strings; and may not be suitable for all types of string similarity comparisons.

4.7.2.3 Phrase-based similarity

In the phrase-based method, the basic unit of the phrase-based method is a phrase word. The fundamental difference between the phrase-based method and the character-based lies in the level of granularity at which they compare and evaluate the similarity between strings. The main methods of phrase-based similarity measures are the Dice coefficient and the Jaccard. The Dice coefficient similarity measure is a technique used to quantify the similarity between two sets or strings based on their overlap. The Dice coefficient is calculated using the following formula:

$$Dice(S_a, S_b) = \frac{2\ x\ comm(S_a S_b)}{len(S_a) + len(b)}$$

The comm (S_a, S_b) indicates the number of collinear phrases, that is, the number of the same characters in the strings S_a and S_b (Dice 1945). This measure provides a value between 0 and 1, in which 1 indicates a perfect match, and 0 indicates no similarity. The Jaccard similarity is defined as the size of the intersection divided by the size of the union of two sets (Jaccard 1912). The Jaccard similarity measure focuses on the relative overlap or shared elements between the sets. The Jaccard similarity coefficient is calculated using the following formula:

$$S(S_a, S_b) = \frac{S_a\ \cap\ S_b}{S_a\ \cup\ S_b}$$

The S_a and S_b are the sets of strings being compared; the $S_a\ \cap\ S_b$ represents the size or cardinality of the intersection of sets S_a and S_b; and $S_a\ \cup\ S_b$ represents the size or cardinality of the union of sets S_a and S_b. The Jaccard similarity coefficient provides a value between 0 and 1, in which 1 indicates a perfect match or complete similarity, and 0 indicates no similarity.

4.7.2.3 Corpus-based similarity

Corpus-based similarity measure methods involve analysing and differentiating texts based on patterns, statistics, and information derived from a database of documents, or a corpus. As a result of the nature of these methods, they consider the broader context of the language usage

and rely on statistical analysis to identify patterns and relationships. The widely used corpus-based similarity measure approaches include the bag-of-words model, distributed representation, and matrix factorization methods.

a) Bag-of-words (BoW) model

The bag-of-words model simplifies the textual information by considering only the frequency or presence of individual words, disregarding the word order and grammar. The basic idea is to represent the document as a combination of a series of words (Wang and Manning 2012). To process the data, the model creates a vector representation of the text, called a "document vector", in which each dimension of the vector corresponds to a unique word in the vocabulary (Salton and Buckley 1988). The output document vectors serve as an input for various NLP tasks, except that this model does not consider the semantic or syntactic relationships between words, and the word order is ignored. The BoW model has the extensions such as TF-IDF (term frequency-inverse document frequency) and the LSI (latent semantic indexing). The TF-IDF can be used to weigh the importance of words based on their frequency in the document and the entire corpus, providing a more refined representation of the text. LSI, on the contrary, computes the similarity between documents based on their underlying latent semantic structure, while taking into account the overall structure of a document collection; and captures the latent relationships between words and documents.

b) Shallow window-based methods

The shallow window-based methods focus on a small window of adjacent words or tokens in a document known as a context window. The methods consider the local context of words within a fixed-size window to extract features and capture relationships between neighbouring words. Due to the shallow technique, low-dimensional vectors can be trained in unstructured text with no mark, which makes similar words closer in distance (Wang and Dong 2020). Word2Vec, GloVe and Bert are the three main developed and most widely used techniques for text vectorization. Word2Vec and GloVe learn distributed representations of words by predicting the context words within a given window. The resulting word embeddings capture semantic and syntactic relationships between words. Word2Vec has two pre-training models, the continuous word-bag (CBOW) and word-skipping (skip-gram) models. Bidirectional encoder representation from transformers (BERT) covers masked language model and next sentence prediction, which capture expression and sentence-level representation (Devlin *et al.* 2018).

c) Matrix factorization methods

Matrix factorization methods have long been used for latent semantic analysis (LSA). These methods are used for generating low-dimensional word representations, and employ low-rank approximations (Wang and Dong 2020). The goal of matrix factorization is to discover latent features or factors that capture the underlying structure or patterns in the data. Information captured by matrix factorization matrices varies by application. LSA has paved the way for the exploration and analysis of latent Dirichlet allocation (LDA). The main difference between LDA and LSA in the context of matrix factorization lies in their underlying objectives. LSA aims to capture the latent semantic structure of a document collection and obtain low-dimensional representations or latent semantic vectors for words and documents (Landauer and Dumais 1997); while LDA aims to discover the latent topics present in a document collection, estimating the topic distributions for each document, and word distributions for each topic (Mardones-Segovia *et al.* 2021). However, both methods involve matrix factorization, dimensionality reduction; and both utilize probabilistic frameworks.

4.7.2.4 Semantic text matching

Semantic text matching refers to the task of determining the similarity or relatedness between two or more pieces of text based on their semantic meaning (Sahami and Heilman 2006). The text matching process involves examining the underlying semantic representation of a specific text as opposed to syntactic and lexical or surface-level features. Semantic text captures semantic similarity between texts, their context, intent, and shared concepts. The two types of semantic text matching are single text matching and multi-semantic document matching, elaborated below.

- Single text matching

Single semantic text matching focuses on comparing and evaluating the similarity between two individual texts. Semantic text matching mainly includes the deep-structured semantic model (DSSM), the convolutional latent semantic model (CDSSM), the Architecture-I for matching two sentences (ARC-I), and the Architecture-II of convolutional matching model (ARC-II). The DSSM represent texts as fixed-length vectors in a continuous space, in which the vectors of semantically similar texts are closer to one another compared with those of dissimilar texts. ARC-I is a representation learning-based model that obtains multiple combinatorial relationships between adjacent feature maps by the convolution layer. The ARC-II model is an interactive learning model that captures local patterns and interactions between words in the

input texts. The ARC-II focuses on encoding both the local and global information of the texts to learn their semantic similarity.

- Multi-semantic document matching

Multi-semantic document matching is a process of comparing and assessing the similarity between multiple documents or texts as a whole. Instead of comparing pairs of individual texts, it focuses on understanding the overall semantic relationships and similarities among a set of documents. On the basis of measuring semantics between documents, the deep-learning model uses a single-granularity vector to represent document text, and to synthesize the matching degree between texts (Wang and Dong 2020). There are two multi-view bi-LSTM (MV-LSTM) and MatchPyramid. MV-LSTM and MatchPyramid have different architectural designs and approaches from text matching. MV-LSTM focuses on incorporating multiple views of the input texts using separate LSTMs, while MatchPyramid utilizes convolutional operations to capture matching patterns. The choice between these models depends on the specific requirements of the text-matching task and the nature of the input data.

4.6.2 Graph-based representation

Graph-based text representation is an approach in which words or concepts are represented as nodes in a graph; and the relationships between them are represented as edges. Semantic similarity based on graph structure refers to a method of measuring the similarity between two pieces of text or concepts by analysing their representations as graphs (Zhu and Iglesias 2016). The advantage of graph-based representation and calculation of text similarity lies in that the links between nodes are established through the edges of graph structures to determine the degree of similarity between nodes (Wang and Dong 2020). The two mainly used graph-based techniques are the knowledge and neural network graphs, explained below.

4.7.2.5 Knowledge-graph representation

Knowledge-graph representation learning employs machine-learning techniques to map the entities and relationships within a knowledge graph onto a continuous low-dimensional vector space while preserving the fundamental structure and properties of the original knowledge-graph, enabling a more efficient processing and analysis (Chen, Jia and Xiang 2020). Through the knowledge-graph representation process, the entities and relationships between the nodes of the graph are encoded as vectors, allowing for various computational tasks such as similarity measurement, relationship inference, and prediction. Knowledge-graph-representation learning methods have been proven to handle large-scale knowledge graphs efficiently; and are

convenient to use in expanding the calculation of similarity directly in the continuous vector space (Chen, Jia and Xiang 2020).

4.7.2.6 Graph neural network representations

Graph neural networks (GNNs) have emerged as an efficient tool for analysing and understanding graph-structured data. GNNs are useful in scenarios in which there are numerous levels of data, connections, and hierarchical relationships between the data elements. In this case, GNNs serve as a connectionism model, which captures the dependency of the graph through the message transmission between the nodes of the graph (Bhattacharya, Natarajan and Saha Roy 2020). One of the key advantages of using GNNs for semantic similarity measurement is their ability to capture both local and global information. The iterative message-passing technique of GNNs allows the incorporation of information from neighbouring nodes, capturing the context and relational information that contributes to semantic similarity. Contrary to the conventional neural network, GNNs maintains a neural state that has the ability to represent information of any depth from its neighbourhood (Wu *et al.* 2020). Table 4.3 summarizes the strengths and limitations of each of the measures in relation to the development of a WSD model for Setswana.

Table 4.3: Strengths and Limitations of Each of the Measures in Relation to WSD

Similarity Measure	**Strength**	**Limitation**
Text Distance Similarity		

Length Distance	Captures the idea that shorter words might be more similar in meaning	Does not consider the semantic relationships between words
Euclidean Distance	Capturing geometric relationships	Does not perform optimally with sparse or high-dimensional data
Cosine Distance	Captures semantic similarity, especially in high-dimensional spaces	Does not consider the structural relationships between words
Manhattan Distance	Measures the distance between two points as the sum of the absolute differences of their coordinates	Is not effective in capturing complex relationships in high-dimensional spaces
Hamming Distance	Efficient for measuring similarity between fixed-length strings	Does not capture semantic similarities well for variable-length words or phrases
Distribution Distance	Effective for capturing semantic relationships based on word co-occurrence	Requires large amounts of text data for accurate distributional representations
Semantic Distance	Reflects the semantic relatedness between words based on their meanings	The quality of semantic distance relies heavily on the quality and coverage of the underlying word embeddings.
Text Representation		
String-based	Efficient for handling short and common words	Does not capture semantic relationships and nuances
Character-based	Effective for handling morphologically rich languages	Struggles to capture higher-level semantic meanings

Corpus-based	Utilizes statistical patterns and co-occurrence information	Requires a large and diverse corpus for optimal performance
Semantic Text Matching	Effective for capturing context and for identifying similar meanings	Highly dependent on the quality and coverage of pre-trained embeddings or models
Knowledge-graph Representation	Effective for capturing structured information and semantic hierarchies	Requires a well-structured knowledge graph relevant to the domain
Graph Neural Network Representations	Captures complex relationships and dependencies between words or entities	Requires careful tuning and may be more complex to implement

Semantic similarity measures play a crucial role in WSD and MT tasks. The choice of a semantic similarity measure depends on the specific characteristics of the data and the nature of NLP tasks in relation to the linguistic characteristics and properties of the language in question.

4.8 The Theoretical Framework

The foregoing provides the theories, models, and concepts that serve as the building blocks for the study's theoretical framework. The question to answer in a theoretical framework is how these building blocks are integrated together into a coherent whole that informs the methodological framework for the study. This question is answered in this section. Figure 4.13 is a conceptual diagram showing the study's theoretical framework.

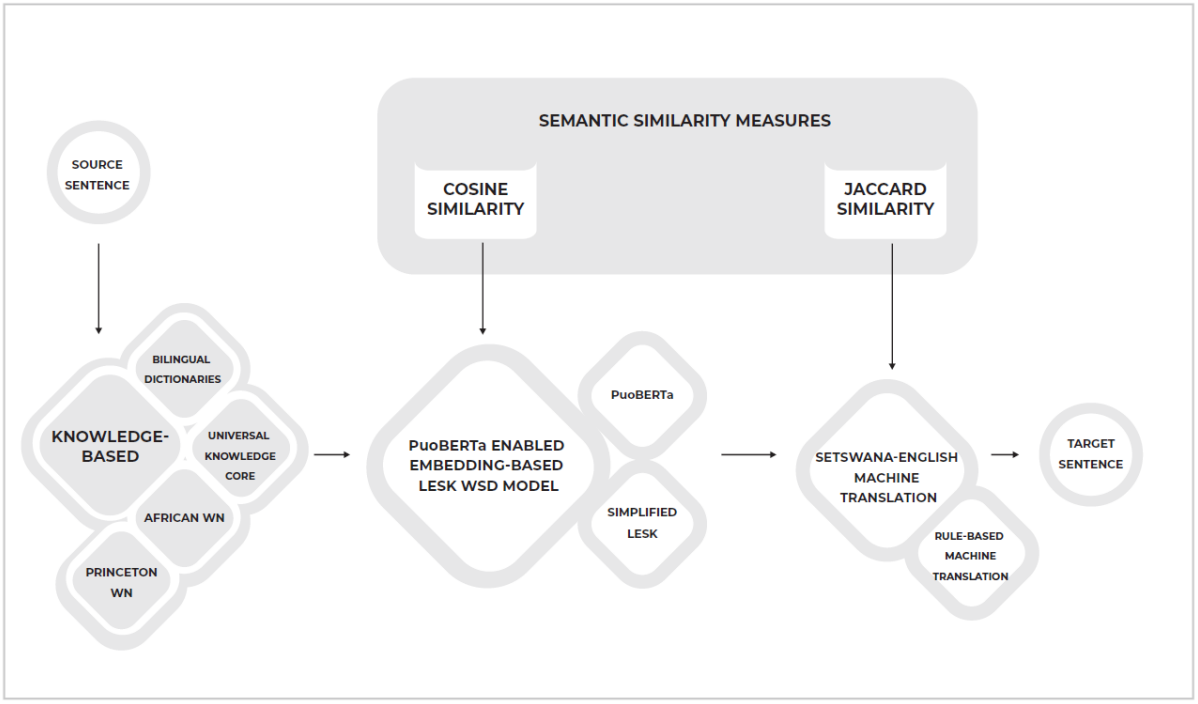


Figure 4.13: Theoretical framework conceptual diagram

This study introduces a knowledge-based word sense disambiguation (WSD) model using BERT for Setswana-to-English machine translation. The theoretical framework guiding the proposed methodology comprises four integral components, as delineated in Figure 4.13. The first component encompasses foundational theories and resources for constructing a knowledge resource for Setswana. These resources, inclusive of bilingual dictionaries, the AWN, PWN, and the UKC, guided the development of a knowledge-based resource. The second component is a PuoBERTa Enabled Embedding-based WSD Lesk Model, grounded in distributional semantic theory, the transformer architecture, and BERT. This model is based on the multilingual BERT pre-trained model, augmented and fine-tuned specifically to accommodate the linguistic nuances and properties of Setswana. The third component focuses on machine translation, based on the RBMT theory and architecture. Lastly, for evaluating the proposed model, the study employs semantic similarity measures, using cosine similarity for WSD and Jaccard similarity for translations. Combined, these elements constitute a comprehensive theoretical framework that underlies the proposed model.

4.9. Chapter Summary

The theoretical framework chapter has provided a comprehensive overview of several key elements that form the foundation of the research methodology. By exploring distributed word representation, bidirectional encoder representations from transformers (BERT), Princeton

WordNet (PWN), universal knowledge core (UKC), African Wordnet (AWN), rule-based machine translation, and semantic similarity measures, this chapter has established the theoretical underpinnings necessary to address the identified research problem. The introduction of these elements sets the stage for the subsequent methodology chapter, in which a detailed explanation of their specific application will be provided. This methodology chapter will delve into the practical implementation of each method, illustrating how it is employed to tackle the research questions and objectives of this study. By combining these diverse elements, the research aims to leverage the power of distributed word representation, advanced language models such as BERT, comprehensive lexical resources such as Princeton WordNet and African WordNet, knowledge bases such as universal knowledge core, and rule-based machine translation, to solve the problem of WSD in the context of Setswana-English machine translation.

CHAPTER FIVE: METHODOLOGY

5.1 Introduction

This chapter presents the methodology employed in developing a knowledge-based WSD model for Setswana in the context of Setswana-English MT. The primary objective of this chapter is to provide a detailed description of the practical application and adaptation of the resource and theoretical frameworks discussed in Chapter Four. The chapter is organized as follows: Section 5.2 presents the graphical representation of the methodological framework; the resource construction process is described in Section 5.3, and the mapping process in Section 5.4. The WSD-MT is presented in Section 5.5, while Section 5.6 summarises the chapter.

5.2 Methodological Framework

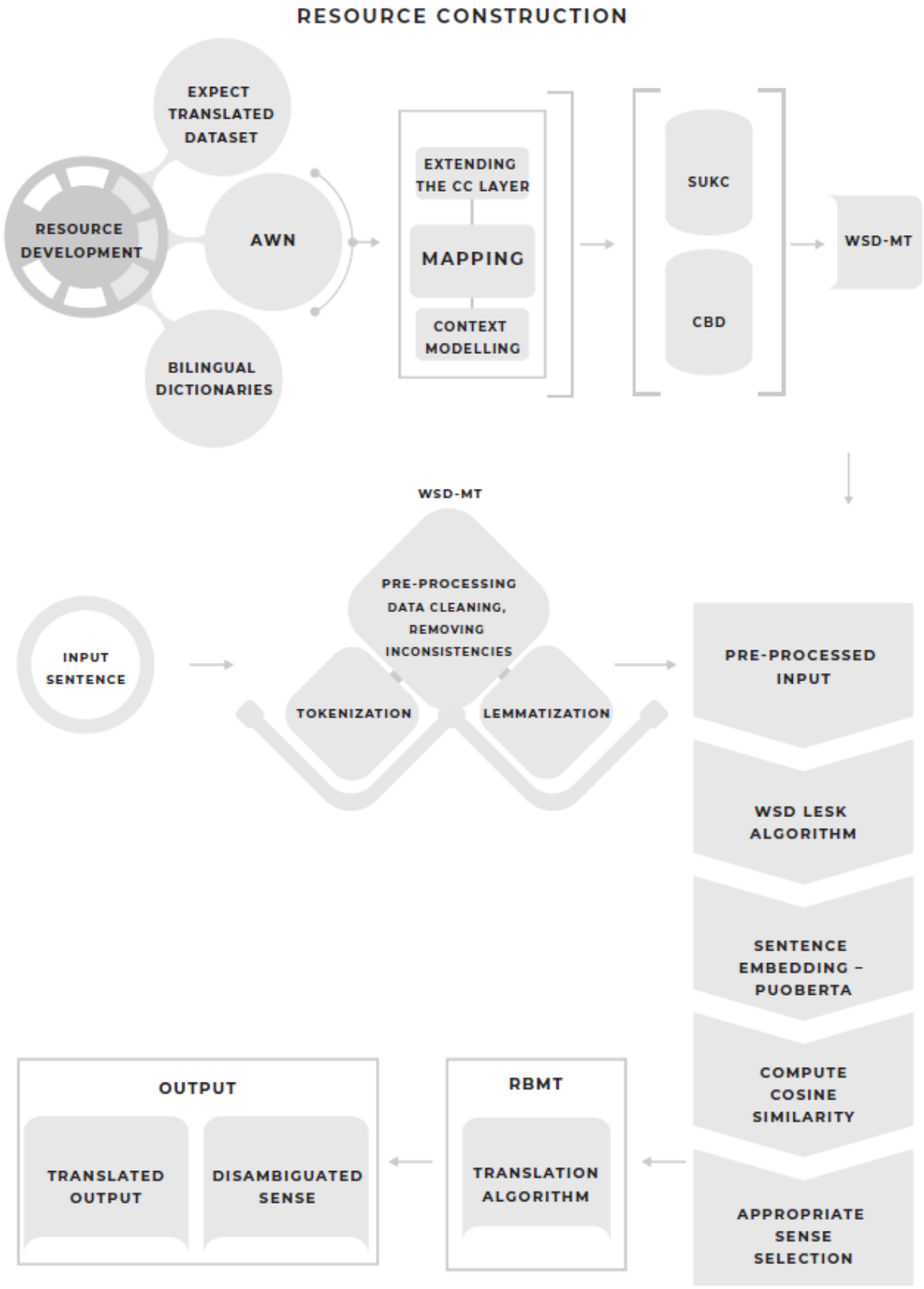


Figure 5.1: Methodological framework (AWN: African Wordnet; CBD: Contextual Dictionary; SUKC: Setswana Universal Knowledge Core; WSD-MT: Word Sense Disambiguation-Machine Translation; RBMT: Rule-Based Machine Translation)

5.3 Lexical Resource Construction

The resource-development process involved three main components: expert translation, Setswana bilingual dictionaries, and the integration of the AWN. Figure 5.2 illustrates the resource-construction component with its elements and processes.

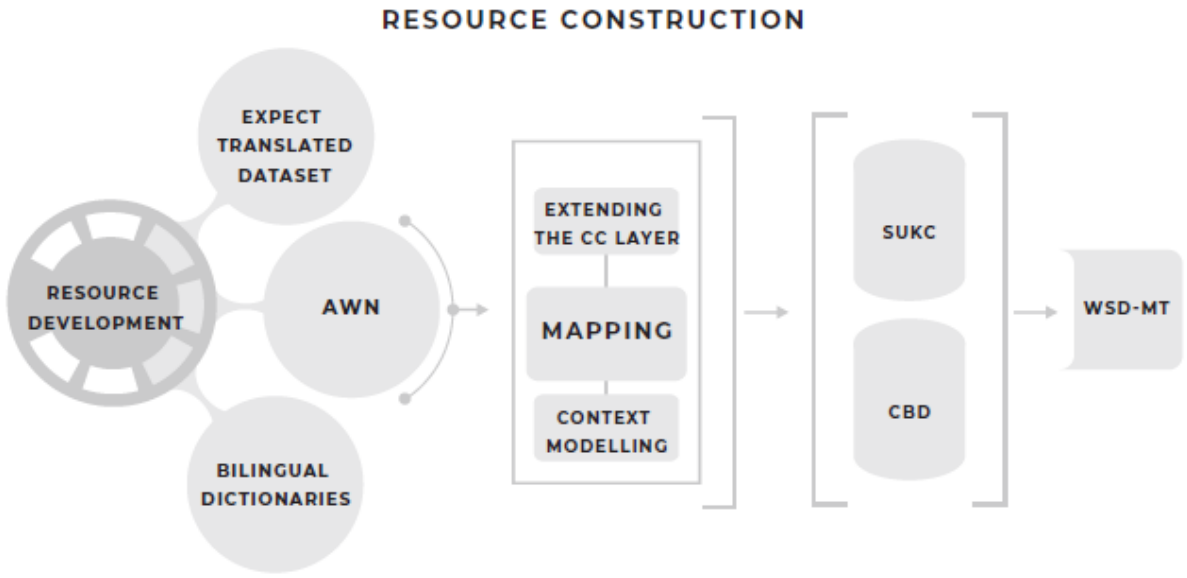


Figure 5.2: Lexical resource construction

The expert-translation component includes professional translators using Microsoft Excel spreadsheets containing source and target lemmas, synsets, glosses, and examples from the space domain. Due to budget constraints, the study further employed the use of paper-based Setswana-English bilingual dictionaries to increase linguistic coverage. Lastly, the AWN served as an integral resource addition to the datasets, complementing the expert translations and bilingual dictionaries. The data from the three sources, expertly translated dataset, AWN, and bilingual dictionaries, were mapped into the universal knowledge core (UKC). This mapping included context modelling for developed context bilingual dictionary (CBD), and extending the concept core layer of the UKC in order to accommodation the Setswana specific concepts that did not exist in UKC, which resulted in Setswana universal knowledge core (SUKC). The two resources were used in the implementation of the Setswana word sense disambiguation and machine translation (WSD-MT) model. The following sections and subsections describe the elements of the resource-construction component.

5.3.1 Expert translation

5.3.1.1 Microsoft Excel entries

For the space domain data translation, a Microsoft Excel spreadsheet containing source and target lemmas, synsets, glosses, and examples, was used. The entries are defined as:

a. Synsets

A synset is a set of synonyms that represents a single concept or idea in linguistics, which consists of lemmas. Each synset represents a unique concept, and words within the same synset are considered synonymous with one another. Synsets provide a way of organizing and understanding the relationships between words and their meanings in a structured format.

b. Gloss

A text or sentence that describes the concept, i.e., a lemma.

c. Example

A text or sentence(s) that clarifies the exact meaning of the described concept. Examples are also used to clarify and demonstrate how the lemma/concept is used in a sentence.

5.1.1.2 Microsoft Excel translation columns

The translation spreadsheet consisted of the following columns:

a. Synset lemmas columns

Column C: contains a comma-separated list of lemmas of the source language.

Column D: The translator provides a comma-separated list of the synset lemmas of the target language.

b. Synset gloss columns

Column F: contains the synset gloss in the source language.

Column G: The translator provides the synset gloss in the target language.

c. Synset examples columns

Column I: contains the synset examples in the source language.

Column J: The translator provides the synset examples in the target language.

5.1.1.3 Validation

The same Excel sheet used for translation was used for validation. The validator fields consist of the following:

a. Synset lemmas validation

The validator validates the lemmas in column E in the target language. The validator can choose between:

- Accepted: If the validator finds that the lemmas are complete and do not contain any errors such as spelling errors, he/she enters A (for accepted).
- Rejected: If the validator finds that the lemmas are not correct, or there are missing lemmas, or lemmas that do not belong to the synset, he/she enters R (for rejected) and provides justification for the decision in the validator notes column.

b. Synset gloss validation

The validator provides a validation on the synset gloss column H. The validator can choose between:

- Accepted: If the validator finds that the synset gloss describes the synset correctly and does not contain errors such as spelling errors, he/she enters A, and provides justification.
- Rejected: If the validator finds that the synset gloss does not describe the synset or it contains errors, he/she enters R, and provides justification.

c. Synset example validation

The validator provides his decision on the synset examples in Column K. The validator can choose between:

- Accepted: If the validator finds that the synset examples are correct and they do not contain errors; and if there are no synset examples that may be necessary to describe how to use the lemmas, she/he enters A. It is possible to accept synsets without examples if the translator did not provide them, nevertheless, the validator accepts the translator's decision.
- Rejected: If the translator did not provide examples, and the validator does not accept the translator's decision, or if he/she finds errors in the examples, he/she enters R and provides justification.

5.3.2 Setswana bilingual dictionaries

Due to budget constraints, this study further employed the use of paper-based Setswana-English bilingual dictionaries to increase the linguistic coverage in addition to the expert-translated data. The four bilingual dictionaries used are Oxford, Oxford Kiddies, Pharos, and Shuter's. The resource development followed the process in Figure 5.3.

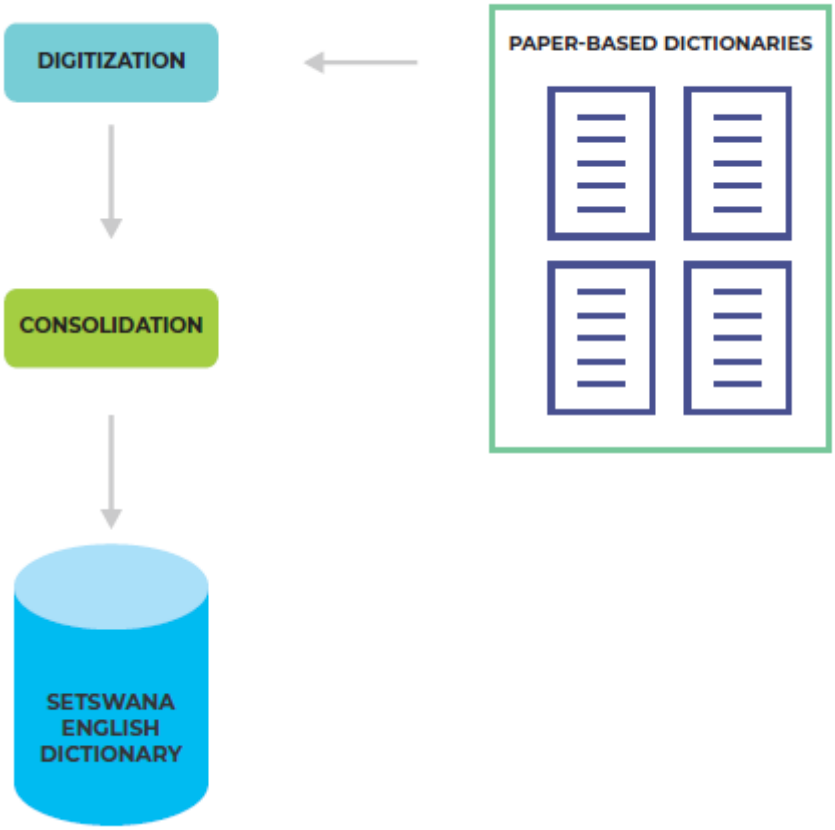


Figure 5.3: Digitized bilingual dictionary construction process

The process was divided into three major phases: data collection, data digitization, and data consolidation. The data-collection process involved researching, purchasing, and compilation of dictionaries. The data-digitization process involves transcribing the data from each dictionary; while data consolidation involves consolidation, mapping, and removing duplicates and overlaps from the four dictionaries. The following subsections describe the phases in much more detail.

5.3.2.1 Data collection

Dictionaries: Paper-based Setswana-English bilingual, Oxford, Oxford Kiddies, Pharos, and Shuter's dictionaries were used.

5.3.2.2 Digitization

The dataset obtained from the four dictionaries was transcribed. The data was transcribed according to the formatting of the dictionaries. The Oxford dictionary had a Setswana lemma, part-of-speech tag, English translation of the lemma, Setswana usage example, and word class; while Oxford Kiddies had Setswana lemmas and their equivalent English lemmas translation only. Pharos dictionary had a Setswana lemma, part-of-speech tag, English translation, English definition, and English synonyms. The Shuter's dictionary had a Setswana lemma, part-of-speech tag, English lemma translation, Setswana synonyms, and English synonyms. Each dictionary was transcribed into its own Microsoft Excel file, which resulted in four different files.

5.3.2.3 Consolidation

Python was utilized to consolidate the four files. A dedicated script was developed to automate this process, aiming to avoid manually inspecting files to identify and remove duplicate entries and overlaps from the dictionaries. Algorithm 5.1, as outlined below, was employed for this purpose.

Algorithm 5.1: Data consolidation

Input: Setswana-English transcribed files (*Files*), Input file path (*Input_path*), Output file path (*Output_path*),

Output: consolidated English-Setswana mappings (*translation_dataset*)

1. EXLlist ← Listdir (*Input_path*)

2. EXReaddataframe ← pd.Dataframe()

3. exlf ← pd.Dataframe()

4. ***For*** exlf in EXLlist

if exlf.endwith(".xlsx")

EXReaddataframe = pd.read_excel(Input_path+ exlf)

exlf = exlf.append(EXReaddataframe)

exlf = exlf.drop_duplicates()

5. end

6. end

7. exlf.to_excel(Output_path+" consolidated_dictionary.xlsx")

return consolidated_dictionary

Algorithm 5.1 consolidates multiple Excel files containing Setswana-English translations into a single data frame, removes any duplicate entries, and then saves the consolidated data to a new Excel file that makes up the Setswana-English bilingual dictionary depicted in Figure 5.3.

5.3.3 African WordNet

The AWN served as an integral valuable addition to the datasets used in this study, augmenting linguistic coverage. Employed as a lexical resource, it complemented the translated data and bilingual dictionaries. The AWN currently houses 53,982 synsets, 9,279 definitions, and 28,853 usage examples across the five South African languages. However, this study primarily focused on the 15,803 Setswana synsets, supported by 3,515 definitions and 7,203 usage instances.

5.4 Lexical Resource Mapping

The data mentioned above, expect translated data, Setswana bilingual dictionaries, and the AWN were integrated and mapped to create a comprehensive and cohesive dataset for the Setswana-English WSD and MT system. The mapping process involves two main components: the UKC mapping and the context-modelling mapping as shown in Figure 5.1. The mapping and integration served a crucial role in creating a unified and comprehensive dataset that leverages the strengths of multiple resources.

5.4.1 UKC mapping

For the purpose of mapping the translations within the universal knowledge core (UKC), the importing and merge features of the UKC framework were utilized. The UKC is already integrated with Princeton WordNet (PWN) synsets (Giunchiglia et al. 2014). Data from the AWN was merged using the PWN synset IDs that were already mapped within the AWN. The professionally translated data of the space domain was merged using the UKC IDs that were already part of the space domain data extracted from the UKC.

The importing of translations into the UKC was executed in a fully monotonic manner: if a newly translated word lexicalized an existing synset that already existed in AWN, it was added to it as a synonym; while if no such synset existed in the UKC, it was not. During this process, the provenance of each synset and word was stored. Whenever the same lexicalization was provided by both sources, they were fused; however, the provenance indicated both sources.

The automated merge operation did not take into account the same word being provided by both resources, the AWN and the professionally translated datasets, but with differing orthographies. This is a relevant issue for native languages in general, words in such languages often not having canonical orthographies prescribed by an authoritative source, or having more than one of them. For this reason, based on local dialects, the same word is often spelled in multiple ways. The duplicates of such words were not eliminated ─ both the different orthographies are correct ─ instead, they were added to the synsets, because the presence of multiple acceptable orthographies contributes to the richness of the resource.

The UKC formally represents phenomena of lexical diversity (Giunchiglia et al. 2017). As such, the lexical gaps were also imported into the UKC. The lexical gaps marked up by the translator and validator were formalized in the UKC as special synsets with glosses. This integration included the incorporation of culture-specific terms into the UKC's concept core layer, subsequently facilitating their inclusion in the language core layer. This feature of the UKC layers underscores its capability of accommodating novel concepts, which stands as a distinct advantage.

This methodological approach yielded the creation of the SUKC, as depicted in the resource construction segment of the methodology architecture shown in Figure 5.2.

5.4.2 Context-modelling mapping

For context modelling, each lemma is associated with a specific keyword. If the lemma has a one-to-one source-to-target mapping, it is linked to its source lemma and equivalent translation. For example:

"ntlo" (Setswana) -> house (English)

In this case, the lemma "ntlo" is mapped to its English equivalent, "house". In instances involving ambiguous lemmas, a one-to-many source-to-target mapping was implemented, as illustrated by:

"Madi: boreledi jo bo khibidu jo botswa mo mmeleng, o tswa madi"(Setswana) -> A red fluid that comes out through a cut in the body, she is bleeding (English).

"Madi: chelate e e berekiswang go reka dilo, o reka borotho ka madi" (Setswana)-> money the most common medium of exchange, she is buying bread with money (English).

The two lemmas of “madi” are mapped in the same row under the same lemma keyword. In this study, the descriptor “substrates” was used to denote various translations. In addition to the dictionary data, a set of conjunction lemma mappings obtained from (Wilken, Griesel and McKellar 2012) were included in the dictionary. The purpose of these mappings was to ensure that the Setswana conjunction words were mapped with the correct English conjunction words. In Setswana, the conjunction of nouns and the conjunction of verbs are different. When nouns are joined, “and” is translated as “le”, but when verbs are joined, “and” is translated as “mme”. Other conjunctions that are translated with the correct Setswana word “is” or, “but” and “because” are mapped with “kgotsa”, “mme” and “ka gore”. The following maps were defined:

[noun] [“le”] [possible: the or an or a] [possible: adjective] [noun] → map “le” with “and”.

[verb] [“le”] [possible: the or an or a] [possible: adverb] [verb] → map “mme” with “but”.

[conjunction “kgotsa”] → map “kgotsa” with “or”.

[conjunction “mme”] → map “mme” with “but”.

[conjunction “ka gore”] → map “ka gore” with “because”.

To effectively use consolidated and mapped dictionary data in an IDE (integrated developed environment), the data was converted to an extensible markup language (XML). This decision was prompted by the observation that the large size of the Microsoft Excel file containing the dictionary data resulted in a significant slowdown of the programme’s execution and processing duration. The performance degradation was attributed to the inherent limitations of the Excel format in handling substantial datasets within an IDE environment.

XML, a widely recognized and standardized markup language, was selected as the target format for the data-conversion process. XML offers several advantages that align with the requirements of efficient data manipulation and processing within an IDE (Chingamtotattil and Gopikakumari 2022). XML provides a structured and hierarchical representation of data, enabling easy parsing and traversal of the dictionary entries. This structured format allows for precise and targeted access to specific elements within the dictionary, thereby optimizing data retrieval and manipulation operations.

A definition of document type was used to convert the data. A document type definition (DTD) outlines the components of XML documents, defining the elements and attributes of the document’s structure. XML elements serve as the fundamental building blocks. XML elements

function as containers that hold texts, elements, and attributes. An XML document encompasses one or more elements identified by start and end tags. Attributes define specific properties of an element, associating a name with a string value. The DTD for the lexical database used in this study is illustrated in Figure 5.4, followed by a detailed description of the element tags within the XML document.

```
<!DOCTYPE Dictionary
[
<!ELEMENT dictionary (Language,Entry)>
<!ELEMENT language (source language, target language)>
<!ELEMENT Entry (word)>
<!ELEMENT word (pos)>
<!ELEMENT word (sense)>
<!ELEMENT word (case, translations)>
<!ELEMENT sourcelanguage (#PCDATA)>
<!ELEMENT targetlanguage (#PCDATA)>
<!ELEMENT case (#PCDATA)>
<!ELEMENT translation (#PCDATA)>
<!ATTLIST ENTRY entryid CDATA #REQUIRED>
<!ATTLIST WORD senses CDATA #REQUIRED>
<!ATTLIST SENSE senseid CDATA #REQUIRED>
<!ATTLIST WORD translations CDATA #REQUIRED>
<!ATTLIST WORD transsubstrateid CDATA #REQUIRED>
]>
```

Figure 5.4: Document type definition for the Setswana-English dictionary

The XML tags in the DTD were used for the following:

- Lexical database tag - The root element of the lexical database is represented by the tag. The root element includes the category attribute, specifying the resource types, i.e., monolingual, bilingual, or multilingual. In this instance, the category attribute is bilingual, as the resource is for Setswana-English. The creation-date attribute indicates the database's creation date; and the encoding attribute specifies the language-encoding method. For this database, the chosen encoding method is Unicode. Unicode is a character-encoding standard that assigns a unique numeric value referred to as a code point to each character in a wide range of scripts and languages (Langemets, Loopmann and Viks 2010). Unicode was selected to handle the diacritics that may be found in Setswana. The attribute specifies the title of the lexical resources, which was assigned to the Setswana-to-English-translation database.
- Language tag - The language tag element specifies the languages in the database, consisting of the element source and target language. The source language is Setswana, and the target language is English.
- Entry tag - The entry tag element denotes each of the lexical entries in the database with the attribute entry-id, which assigns a unique numeric identification number known as the primary key to each of the translated words.

- Word tag - The word tag is a sub-element of the entry tag that contains the Setswana lemma with a sense attribute that indicates the number of possible English translations of the lemma.
- POS Tag - The POS tag is a sub-element of the word tag that contains the Setswana part of speech.
- Sense Tag - The sense tag defines each possible translation of the Setswana word using the translation and case tags. The sense tag has sense-id and POS attributes. The sense-id assigns a unique primary key to each of the possible translations. The POS attribute specifies the part of the speech in that sense of the word, i.e., noun, pronoun, verb, adjective, adverb, and conjunction. The translation tag specifies the English translation of the sense, while the *case* tag provides an example of the sense being used.

The outcome of this process was the context-aware bilingual dictionary, illustrated within the resource construction component in the methodology architecture depicted in Figure 5.1.

5.5 PuoBERTa Enabled Embedding-based Lesk WSD Model

The PuoBERTa enabled embedding-based Lesk WSD model adopted the simplified Lesk model and uses the Setswana Bert model, PuoBERTa, to generate sentence embeddings for the context sentence and polysemous word glosses and uses semantic similarity measure Cosine similarity to determine the correct sense of the polysemous word within that specific context. The reason for this selection is that compared with other variants, simplified Lesk's primary objective is to maintain disambiguation effectiveness while simplifying the computation by reducing computational complexity (Kilgarriff and Rosenzweig 2000). This algorithm reduces computational complexity by using limited context instead of considering the entire context surrounding the polysemous word. Simplified Lesk limits the scope to a predefined window of adjacent words; and has the ability to work well for languages with limited resources, making it suitable for Setswana as a resource-scarce language. Another method that simplified Lesk adopts to decrease computational complexity involves minimizing linguistic features. This is achieved by reducing reliance on extensive linguistic features and syntactic structures within the context, simplifying the feature set. However, this is a crucial capability when disambiguating agglutinative and morphologically rich languages such as Setswana. To address this, the linguistic features are integrated into the encoding process, utilizing PuoBERTa to

encode the complete context sentence and the respective glosses to generate sentence embeddings. Sentence embeddings capture contextual information and the compositional nature of a language. This provides a representation that reflects the combination and interaction of words within a sentence (Scarlini, Pasini and Navigli 2020). Our algorithm is more closely related to the simplified Lesk algorithm (Kilgarriff and Rosenzweig 2000) but leverages the importance of word sense definitions using embeddings and similarity measures. The encoding process using PuoBERTa is as follows:

- The first step of the encoding process is preprocessing and tokenization. The input is passed into the input embedding layer which maps each token ID to a fixed-size vector (embedding) that encodes information about the token in a high-dimensional space. These embeddings are the input to the next layer of PuoBERTa.
- The second step is contextual embedding by the self-attention layers. The self-attention layers are the core of the transformer's ability to understand context. Each token's embedding is updated by aggregating information from all other tokens in the sequence, weighted by their relevance (attention scores). This mechanism allows each token to attend to every other token in the input sentence, enabling the model to capture the context of a word by considering its relationships with other words in the sentence.
- The third step is contextual representation by feed-forward layers. The feed-forward layers allow the model to learn complex representations by applying non-linear transformations. Each token embedding is passed through these layers to generate a more refined representation that captures both its original meaning and the context derived from self-attention which enhances the representation.
- The pooling layer is applied on the fourth step. This layer aggregates information from the entire sequence into a single vector. This vector is used for sentence-level WSD task. To obtain the sentence-level embedding, the model pools information from the token embeddings that represents the entire sentence.

After obtaining sentence and gloss embeddings, the cosine similarity between these vectors is calculated to determine which gloss best matches the sentence context. The metric used to measure semantic similarity is cosine similarity, as it effectively captures the directional similarity between vectors, making it suitable for assessing the relationship between gloss embeddings and context representations (Orkphol and Yang 2019). Research that employs cosine similarity for disambiguation includes (Sarika and Sharma 2016), (Sari, Manurung and Adriani 2010), (Yatabe and Sasaki 2020), (Hari and Kumar 2023). As depicted on the Figure

5.5, the input sentence is first pre-processed through data cleaning, then passed in the Lesk algorithm. The WSD component of the WSD-MT pipeline translation algorithm consists of the following steps:

Input: Ambiguous word (W), Context sentence (C), Sense inventory with glosses for each sense of W.

Disambiguation:

1. Tokenization:

Tokenize the context sentence (C)

2. Encode

Encode context sentence (C) using PuoBERTa

3. Sense Selection:

For each sense of the ambiguous word (W), retrieve the corresponding glosses

4. Encode and Measure Similarity:

Encode corresponding glosses (G) using PuoBERTa

Measure semantic similarity using cosine similarity on (5.1) between C and G

$$\cos(\theta) = \frac{\mathrm{C}\cdot\mathrm{G}}{\|\mathrm{C}\|\|\mathrm{G}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (5.1)$$

5. Sense Ranking:

Rank the senses based on the degree of similarity. The sense with the highest similarity is considered the most likely appropriate sense.

6. Disambiguation:

Assign the sense with the highest similarity score as the disambiguated sense for the ambiguous word (W).

The steps above are formalized as Algorithm 1 below:

Algorithm 5.2: WSD Algorithm

Input: target word (*w*), context sentence (*cs*)

Output: best disambiguate senses (*bestDef*) of the *w*

1. STX ← Pre-process(RemoveStopwords(Tokenized(Text))
2. Syn = GetSynsetFromUKC(w)
3. Est = defEncode(STX)
4. highestSim = 0
5. ***For*** s in Syn
    Sdef = GetDefinition(s)
    Eqs = defEncode(Sdef)
    Similarity = Cosim(Est, Eqs)
    if Similarity > highestSim
        highestSim = Similarity
        bestDef = s
6. end
7. End

return bestDef

Algorithm 5.2 illustrates the pseudocode of the proposed method. The process starts with the input of the target word (*w*) and the context sentence (*cs*) into the system. The provided context sentence is pre-processed. Following this, the system retrieves synsets of the target word from the UKC, extracting various glosses associated with the target word. The context sentence is then encoded using PuoBERTa. Next, each synset gloss of the target word is encoded, and a similarity measure is computed for each gloss similar to the context sentence. The algorithm identifies and returns the gloss with the highest similarity measure as the correct definition of the target word.

5.6 Setswana-English Rule Based Machine Translation

Following the WSD task in sequence is the translation task. To ensure that the meaning is preserved during translation. The methodology translates on word and phrase level and integrates the words and phrases back into the sentence, ensuring grammatical and contextual coherence using the developed Setswana-English language resource. The MT component of the WSD-MT pipeline translation algorithm consists of the following steps:

Input: Setswana sentence

1. Tokenize and lemmatize input

Input: Sentence S.

Output: Lemmatized and tokenized sentence W.

Description: Tokenize and lemmatize the input sentence S, removing special characters in the process.

2. Find keyword lexical entry id and sentence length

Input: Lemmatized sentence W.

Output: Keyword index wi and sentence length Ln.

Description: Retrieve the keyword index (wi) and calculate the length of the sentence (Ln).

3. Find translation substrates from the based on the context in the dictionary

Input: List of words (inwords), values, word position (wpos), and a parameter (d).

Output: Substrate value (substrate).

Description: Extract a substrate value based on a Jaccard coefficient comparison between the initial context and a merged context of substrates. The loop iterates through values, compares contexts, and returns the substrate if a match is found.

4. Calculate the Jaccard coefficient

Input: Two strings, str1 (A) and str2 (B).

Output: Jaccard coefficient sm using Equation 5.2.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (5.2)$$

Description: Calculate the Jaccard coefficient (sm) for two strings (str1 (A) and str2 (B)). This coefficient measures the similarity between sets by dividing the size of the intersection by the size of the union.

5. Merge words depending on the Jaccard coefficient

Input: Sentence S in Setswana context.

Output: Merged term mw.

Description: Merge the words in the input sentence (S) into a single term (mw) using the merge function. The loop iterates over entries in the sentence, combining them into a cohesive term.

The steps above are formalized as the Algorithm 5.3 below:

Algorithm 5.3: MT Algorithm

Input: target word (*w*), context sentence (*cs*)

Output: translated sentence (*ST*) of the context sentence

1. Mw ← merge(words, x0, xn)

2. ***for each*** entryW ***in*** x0, xn

term = term + words(entryW)

3. ***return*** term

4. *sm*←jaccardco(str1, str2)

5. ***return*** sm

6. sb← getsubstrate(inwords, values, wpos, d)

7. ***while*** values[x] != empty and x < d

diwords = wordtokenize(values[x])

ndiwords = wpos + wcount

dicontext = merge(diwords, 0, wcount)

incontext = merge(diwords, wpos, ndiwords)

sm = jaccardco(dicontext, incontext)

8. substrate = values[y + d]

9. ***return*** substrate

10. W ← Lemmatized(RemoveSpecialCharacters((Tokenized(*S*))))

11. w_i← getKeyWordIndex(W)

12. Ln ← getLengthSentence(W)

13. While wpos < Ln

If w_i in dict

output, wpos = getsubstrate(inwords, values, wpos, d)

ST = ST + " " + output

14. ***return*** ST

Algorithm 5.3 illustrates the pseudocode of the proposed translation model. The algorithm involves a combination of linguistic processing, contextual analysis, and translation strategies for translating Setswana sentences into English. The algorithm utilizes Jaccard coefficients for context matching, and substrates for translation. This study acknowledges the scarcity of resources available for the Setswana language, as highlighted throughout the thesis. Consequently, the current database may not contain all Setswana words, leading to the out of vocabulary (OOV) challenge. To address this issue, the algorithm incorporates a semi-automated feature for resource growth. When the algorithm receives words as input that are not found in the database, they are written to a file for subsequent incorporation. This approach enables the continuous expansion and enrichment of the database, enhancing the system's ability to handle a broader range of Setswana vocabulary over time. Addressing the OOV challenge through this semi-automated resource growth mechanism, the proposed translation model demonstrates adaptability and scalability.

5.7 WSD-MT Pipeline

The WSD-MT pipeline integrates the PuoBERTa Enabled Embedding-based Lesk WSD Model, and rule-based machine translation (RBMT). Figure 5.2 illustrates the resource-construction component with its elements and processes.

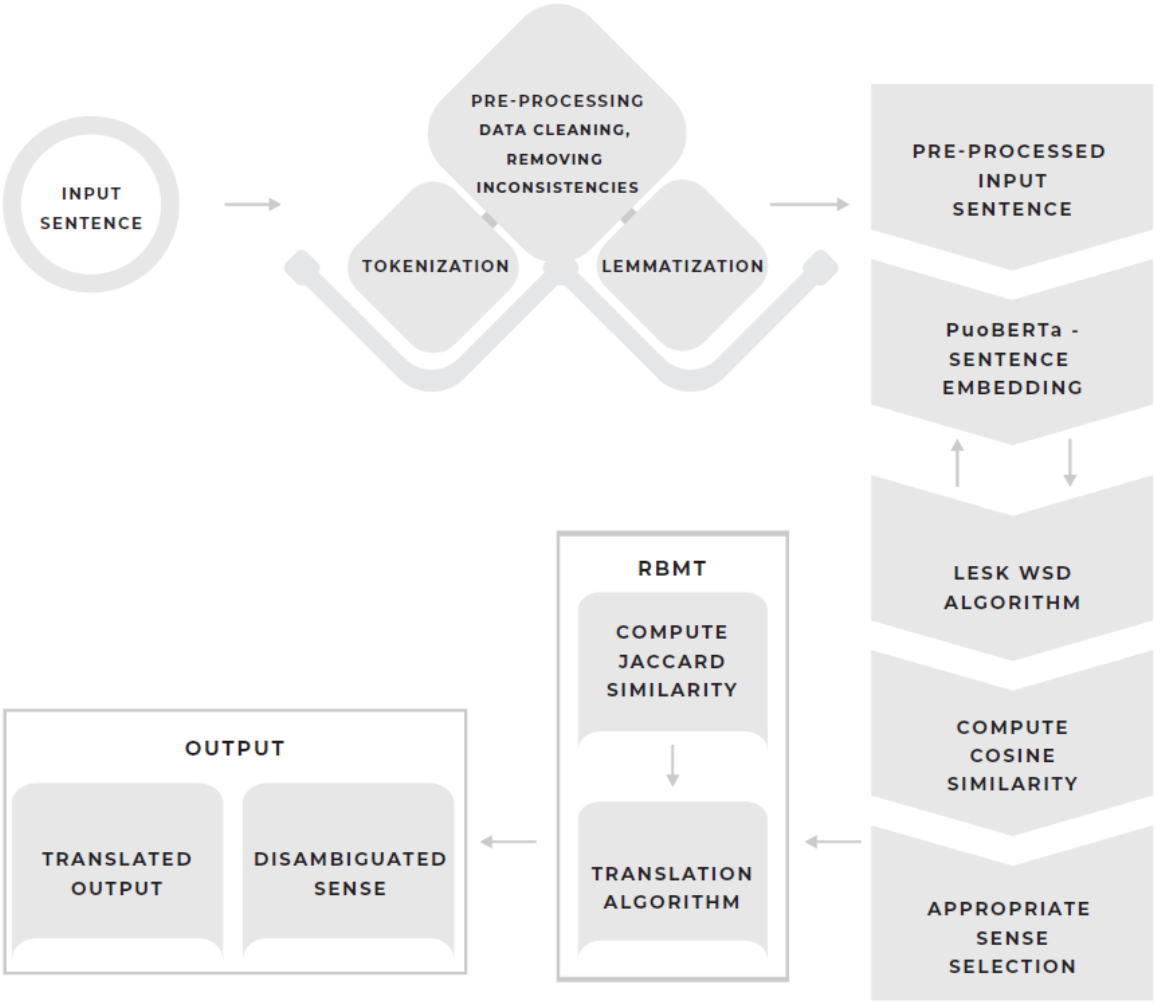


Figure 5.5: WSD-MT

As illustrated on Figure 5.5 the input into the pipeline is the Setswana input sentence which gets preprocessed by removing stopwords and special characters, and then tokenizing the text. Using the modified simplified Lesk, the algorithm retrieves all possible senses of the ambiguous word from the Setswana-English lexical resource. The algorithm then uses PuoBERTa to encode the context sentence and each possible sense into sentence embeddings. The algorithm then calculates the cosine similarity between the context and each sense, and subsequently select the sense with the highest semantic similarity as the most appropriate sense for the polysemous word within that specific context.

Once the best sense of the word is determined, the algorithm proceeds with translating the entire sentence into English. It ensures that the polysemous word is translated according to the selected sense. The translation process uses a rule-based approach, which handles the structure and grammar of the sentence while incorporating the correct translation of the disambiguated

word. To measure the accuracy of target translations and the retrieved candidate translation, the algorithm uses Jaccard semantic similarity. Jaccard semantic similarity is used in the MT to compare segments of text, such as matching phrases or substrings between the source and target languages, to find the best translation equivalents. This method is particularly useful in assessing the degree of overlap between different word groups or phrases, especially when handling translations that involve phrases with multiple possible translations depending on context. The translations with the highest semantic similarity were chosesn as the target translations. The final output is a fully translated English sentence that accurately reflects the intended meaning of the original Setswana sentence, with special attention given to the correct interpretation of any ambiguous words. The WSD-MT pipeline algorithm is as follows on Algorithm 5.4:

Algorithm 5.4: WSD-MT Pipeline Algorithm

Input: target word (*w*), context sentence (*cs*)

Output: translated sentence (*ST*) with the correct sense of the ambiguous word

8. STX ← Pre-process(RemoveStopwords(Tokenized(Text))

9. Syn = GetSynsetFromUKC(w)

10. Est = defEncode(STX)

11. highestSim = 0

12. ***For*** s in Syn

Sdef = GetDefinition(s)

Eqs = defEncode(Sdef)

Similarity = Cosim(Est, Eqs)

if Similarity > highestSim

highestSim = Similarity

bestDef = s

13. end

14. End

15. ***return*** bestDef

16. bestDefWord = getWordFromDefinition(bestDef)

17. translatedWord = TranslateWord(bestDefWord)

18. ST = “ ” # Initialize translated sentence

19. wi ← getKeyWordIndex(W)

20. Ln ← getLengthSentence(W)

21. While wpos < Ln:

if wi in dict:

output, wpos = getsubstrate(inwords, values, wpos, d)

sm←jaccardco(ST, output)

if *sm* > tsimilarity

ST = ST + " " + output

22. ST = Replace(ST, w, translatedWord)

23. ***return*** ST

Integrating WSD with RBMT is was important for accurately translating Setswana sentences containing polysemous words into English. The use of PuoBERTa was particularly effective because of the transformer-based architecture it is based on which efficiently captures contextual information within sentences. By encoding the context and comparing it with potential senses of the ambiguous word, the algorithm can determine the most appropriate meaning based on the sentence's specific usage. Additionally, integrating this with a rule-based translation system allows for precise control over how the disambiguated word is translated, addressing the limitations of purely statistical or neural machine translation models, which may not handle polysemy effectively without disambiguation. This model ensures that the translated sentence not only reflects the correct meaning of the ambiguous word but also maintains grammatical accuracy and coherence in the target language, a necessity for translating resource-scarce languages like Setswana, where large parallel corpora is not be available for training data-driven models leading to the data sparsity challenge.

5.8 Summary

This chapter provides a description of the methodology employed in the development of a knowledge-based WSD model for Setswana in the context of Setswana-English MT. It covers the resource construction process, including the use of expert translations, bilingual dictionaries, and the AWN, as well as the resource mapping to the UKC; ultimately creating the Setswana language resource, SUKC, and the context-aware mapping of the CBD, as well as the conversion of dictionary data into XML format for efficient use. The chapter presents the adapted simplified Lesk algorithm for Setswana disambiguation, incorporating linguistic features through sentence embeddings using PuoBERTa and cosine similarity measure. The

chapter also describes the MT algorithm utilizing Jaccard coefficients and substrates substation for translation based on RBMT. The next evaluation chapter presents the experimentation, evaluation, and results of the algorithms presented in this chapter. It outlines both the intrinsic and extrinsic evaluation methods, along with the dataset used and the comparison analysis with existing research work.

CHAPTER SIX: EXPERIMENTATION, EVALUATION AND RESULTS

6.1 Introduction

This chapter presents the experimentation, evaluation, and results of the PuoBERTa Enabled Embedding-based Lesk WSD Model, the Setswana-English RBMT model and the integrated WSD model into the Setswana-English RBMT. The objective of this chapter is to provide a comprehensive analysis and assessment of the developed algorithms, demonstrating their performance, and applicability in addressing the identified research challenges. This chapter is structured as follows: Section 6.2 presents the evaluation framework used in this study and zooms into the development of evaluation datasets, evaluation metrics, and the experimental setup. The evaluation and results for the WSD, MT, and WSD-MT are presented in Section 6.3. Section 6.4 discusses the results and Section 6.5 summarizes the chapter.

6.2 Evaluation Framework

The evaluation framework for the proposed Setswana WSD-MT model comprises four essential components: evaluation datasets, evaluation metrics, experimental setup and settings, and the results obtained from the evaluation process. Figure 6.1 provides a visual representation of the evaluation framework.

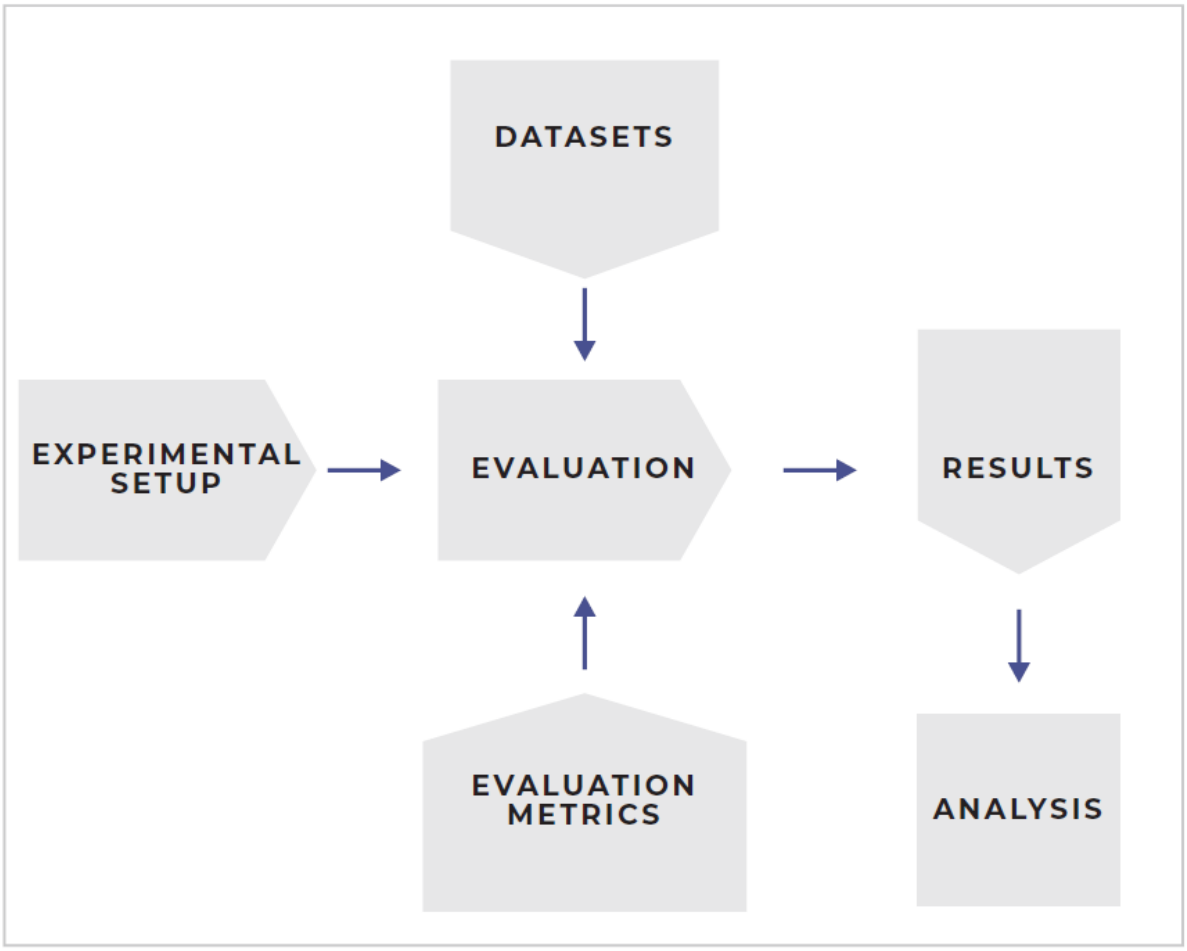


Figure 6.1: Evaluation framework

The evaluation components consist of the datasets, evaluation metrics, and experimental setup. These components are important for evaluating the performance and effectiveness of the Setswana WSD model. For the datasets, due to the lack of benchmark datasets for Setswana, which are available for languages such as English, new benchmark datasets were created and developed for this study. The development process followed the same procedures and structures as the existing English benchmark datasets to ensure comparability and reliability.

The choice of evaluation metrics was based on the metrics commonly used in the literature for comparative analysis. This approach allows for a consistent and standardized assessment of the Setswana WSD model's performance against existing research. The experimental setup and settings were designed to ensure a systematic evaluation of the Setswana WSD model. For the analysis, the evaluation results, were then analysed and compared with existing work in the literature. The analysis involves a comprehensive comparison of the Setswana WSD model with existing WSD research across all available approaches. To gain deeper insights, the analysis is further narrowed down to studies that directly employ the same algorithm as the one

used in this study, namely the Lesk algorithm. Although these studies used different variations and techniques, comparing them provides a more focused analysis and enables closer comparisons.

In addition to the WSD aspect, the evaluation framework also encompasses the analysis of MT and the integration of WSD with MT (WSD-MT). To ensure a fair comparison, the current studies on Setswana-English and English-Setswana MT translation systems are analysed. This analysis takes into account the datasets used to train these systems; and provides comparative results to assess the impact of WSD on the overall translation quality.

The comparative analysis is a critical component of the evaluation framework. By comparing the Setswana WSD model with existing research, both in terms of WSD approaches and MT systems, the study aims to demonstrate the model's effectiveness and its potential to advance and contribute to the state-of-the-art in Setswana language processing.

6.2.1 Datasets

6.2.1.1 WSD datasets

To construct the evaluation data set, the Senseval-3 lexical sample structure by Mihalcea, Chklovski and Kilgarriff (2004) was adopted. Senseval-3 is one of the evaluation benchmark datasets, and a follow-up to Senseval-1 and Senseval-2. The dataset was built using the Open Mind Word Expert system proposed in (Chklovski and Mihalcea 2002). To construct the evaluation dataset, words that were mapped to more than one synset in the AWN were extracted, together with their glosses and example sentences. Figure 6.2 illustrates an example of an ambiguous word in AWN. Algorithm 6.1 was used to extract ambiguous words in the WordNet.

```
    <LexicalEntry id="awn_tsn-mma-n">
      <Lemma writtenForm="mma" partOfSpeech="n"/>
      <Sense id="awn_tsn-ENG20-09350337-n-1-mma" synset="awn_tsn-
ENG20-09350337-n">
      </Sense>
      <Sense id="awn_tsn-ENG20-09582021-n-2-mma" synset="awn_tsn-
ENG20-09582021-n">
      </Sense>
      <Sense id="awn_tsn-ENG20-09663987-n-3-mma" synset="awn_tsn-
ENG20-09663987-n">
      </Sense>
    </LexicalEntry>
```

Figure 6.2: Lexical entry "mma" in AWN

Figure 6.2 depicts a lexical entry "*mma*" associated with three separate noun synsets. This indicates that the word "*mma*" has three senses, depending on the context in which it is used.

Algorithm 6.1 : AWN multi-sense words extraction

Input: AWN file (*ANWfile*)

Output: Setswana sense inventory (*sense_inventory_dataset*)

1. AWNtree ← ETparse (*ANWfile*)
2. AWNroot ← AWNtree.getroot()
3. workbook ← openpyxl.Work()
4. workbook.title ← "sense_inventory_dataset"
5. ***For*** lex_entry in AWNroot.findall('.//lexicalEntry')
   sense_ids = lex_entry.findall('.//sense/senseId')
   if sense_ids > 1
   synset = lex_entry.find('.//synset').text
   written_form = lex_entry.find('.//writtenForm').text
   part_of_speech = lex_entry.find('.//partOfSpeech').text
   definition = lex_entry.find('.//definition').text
   example = lex_entry.find('.//example').text
   worksheet.append([synset, written_form, part_of_speech, definition, example])
   end
6. End
7. Worksheet.save("sense_inventory_dataset.xlsx")
8. ***return*** sense_inventory_dataset

Additional words, glosses, and example sentences were extracted from Oxford and Pharos bilingual dictionaries; the different senses were indicated with number superscripts in the dictionaries. From these resources, an evaluation dataset of 1200 Setswana sentences was created. Currently, there is no existing WSD evaluation dataset for Setswana. This dataset serves as a valuable resource for evaluating and benchmarking models designed to address the complexities of ambiguities in Setswana. Figure 6.3 illustrates the snippet of the Setswana sense inventory.

Figure 6.3: Setswana sense inventory

The target words are organized in alphabetical order in the inventory. Each target word has an instance entry with an instance ID and a source file, which indicates the origin from which the target word was extracted. Sources can include Oxford and Pharos dictionaries, as well as the African WordNet (AWN). The entry instance also includes the context sentence for the target word, its possible senses, the appropriate sense for that target word within the specific context, and the English translations of the senses. Figure 6.4 illustrates the Senseval-2 English sense inventory structure.

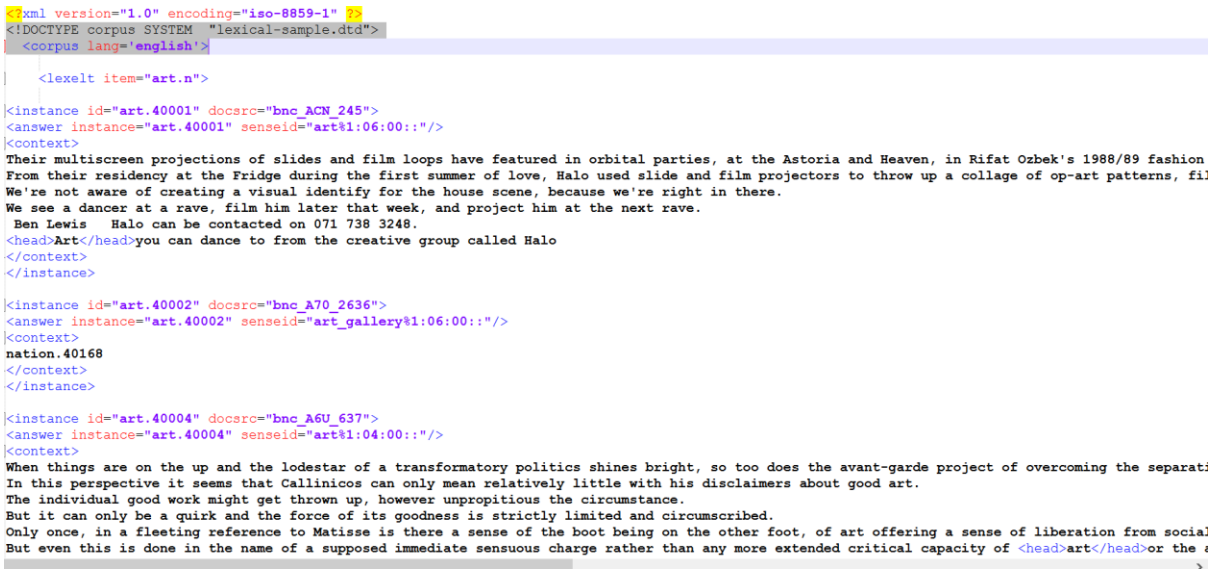


Figure 6.4: Senseval-2 English sense inventory

The Senseval-2 and Setswana sense inventories share similar organizational structures in terms of instances, source files, context sentences, and instance IDs, but they differ in several key

aspects. Senseval-2, designed for English, contains a significantly larger dataset compared to the Setswana inventory. Unlike Senseval-2, the Setswana inventory does not include sense IDs, as these are obtained and linked with WordNet synsets or use a head element to identify polysemous words, instead relying on the target word itself. A notable difference is that the Setswana inventory includes English translations of Setswana glosses derived from dictionaries, while Senseval-2 is monolingual, containing only English data.

The data was initially collected and organized in an Excel spreadsheet. Figure 6.5 illustrates a snippet of the data organized in the spreadsheet file.

Setswana Ambiguous Wor	Setswana Context Sentence	Setswana Possible Glosses	Setswana Context Gloss	English Dictionary Definitions
aba 1	Ke tla aba nako ya go ithuta letsatsi lengwe le lengwe	beya fa thoko;go phatlhalla	beya fa thoko	to set aside something (like a task, time or mo
aba 2	Metsi a ile a aba ka potlako fa fatshe	beya fa thoko;go phatlhalla	go phatlhalla	to spread
baba 1	Letsatsi le baba thata mo marung a thari	go fisa;go tlhtlhonya	go fisa	hot
baba 2	Letlalo la me le a baba	go fisa;go tlhtlhonya	go tlhtlhonya	skin irritation
baka 1	Mme o rata go baka dikuku	go baka ka setofo; go dira sengwe	go baka ka setofo	to bake
baka 2	Dikgwebo di tlile go baka ditiro tse dintsi mo motseng o	go baka ka setofo; go dira sengwe	go dira sengwe	to make something happen
batla 1	Ke batla dipampiri tsa me tse di nlatlhegetseng	go sekasekana le sengwe; go tlhoka seng	go sekasekana le sengwe	to look for
batla 2	Re batla thuso ya gago go rarabolola bothata jono	go sekasekana le sengwe; go tlhoka seng	go tlhoka sengwe mo go masisi	to need
belaela1	Ke belaela gore a go tla nna dipula letsatsi la ka moso	go nagana gore sengew se ka diragala; go	go nagana gore sengew se ka diragala	a feeling when you are unsure about somethin
belaela2	Ba belaela gore motho yo o tshwereng madi ke legodu	go nagana gore sengew se ka diragala; go	go nna le mokgwa ka sengwe se se dirage	to believe that someone is guilty of something
bodirelo 1	Bodirelo jono bo dira diaparo tse dintle	lefelo leo le dirang diaparo; lefelo leo mo	lefelo leo le dirang diaparo	place where goods are manufactured
bodirelo 2	Ditiro tse dintsi mo bodirelong jono di akaretsa go itaya	lefelo leo le dirang diaparo; lefelo leo mo	lefelo leo modiro wa matsogo o direlwang	place where manual work is done
bodulo1	Bodulo jwa gagwe bo kwa motseng wa Mathibestad	lefelo le motho a dulang kwa go lona; sen	lefelo le motho a dulang kwa go lona	home; place where one lives
bodulo2	Ke Ke batla bodulo mo phatlalatsong lapeng	lefelo le motho a dulang kwa go lona; sen	sengwe se motho a ka dulang mo go sone	place to sit
bofelo 1	Bofelo jwa pina e e ne e le botlhoko	bo felelo jwa sengwe; motho kgotsa seng	bo felelo jwa sengwe	1 last point of something 2 after all the others;
bofelo 2	Ke tlaa nna mo mothong wa bofelo yo o yang kwa tlhop	bo felelo jwa sengwe; motho kgotsa seng	motho kgotsa sengwe sa mafelelo	last
bogale1	O ile a bua le nna ka bogale fa a ntlhalosa bothata jwa g	tsela e e tshosang; sengwe se se tlhaban	tsela e e tshosang	in a threatening way
bogale2	Thipa eno e bo bogale	tsela e e tshosang; sengwe se se tlhaban	sengwe se se tlhabang kgotsa se se kotsi	with a point
boima 1	okgoni jwa go rarabolola dipalo tse bo ne bo le boima g	sengwe se se thata; sengew se se bokete	sengwe se se thata	to be difficult

Figure 6.5: Excel Setswana sense inventory

The dataset comprised 1200 Setswana sentences, of which 602 contained target words. An additional 408 sentences with repeated target words, including context sentences and usage examples sourced from the AWN, Oxford dictionaries and the remaining 190 sentences contained monosemous words. The datasets comprised of 43 polysemous words and 14 polysemous concords e.g. (a – of the, a – then, a-did not, a-these, a-can, ga-not, ga-of, ga-at). This diverse collection of sentences provides a comprehensive representation of various word types and their context to enable a thorough evaluation of the PuoBERTa Enabled Embedding-based Lesk WSD Model performance across different linguistic scenarios.

6.2.1.2 Machine-translation datasets

For the evaluation of the MT component, a parallel corpus consisting of source language sentences and their corresponding translations in the target language was constructed from the Oxford and Pharos dictionaries. These dictionaries were specifically selected due to their inclusion of complete sentences in both Setswana and English, rather than solely providing lemmas. Figure 6.6 depicts Oxford's extracted data. Only usage examples and their translations were used as parallel sentences. Figure 6.7 illustrates the data extracted from the Pharos

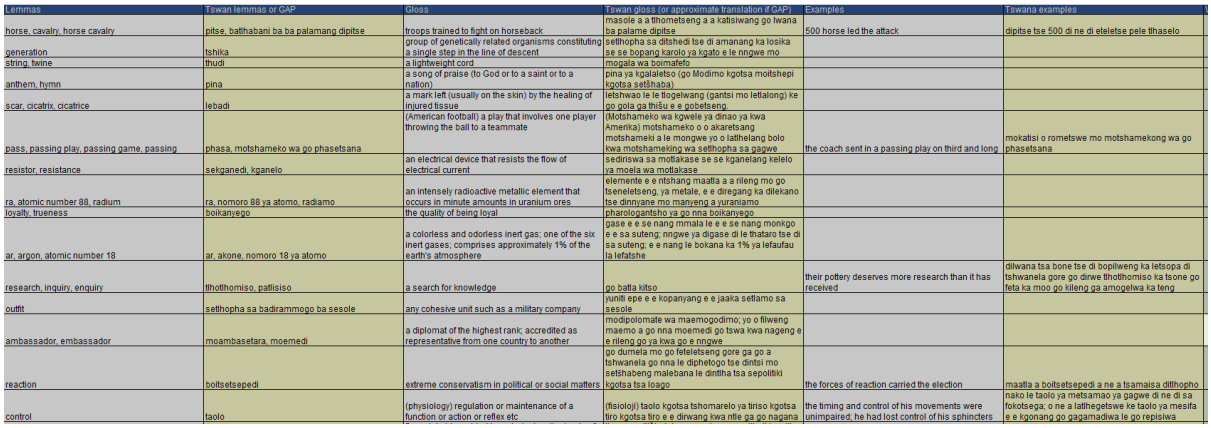
dictionary. This particular dictionary provided only definitions for the words. In the figure, the polysemous words are highlighted, drawing attention to the terms that have multiple meanings or senses. In Figure 6.8, the data from the professionally translated material is presented. Both glosses and usage examples were utilized. However, only entries that included usage examples were taken into consideration. Figure 6.9 illustrates the Setswana-English parallel corpora extracted from the Autshumato corpus. These corpora were originally in text files and were then transferred into an Excel file to maintain consistency with the structure of other data extracted from dictionaries.

Setswana	Part of Speech	Class	Synonym	Dark square[translation]	light Square[usage example]	English usage example
aba	verb			give	re tlaa aba dikobo,re di abela batlhoki	we weill give blankets to the needy
abela	verb			distribute	setheo sa tla abela dikolo tsa mo gae dibuka n	the company will distribute free books to local schools
abelela	verb			guess	a o ka abelela gore ke rata mmala ofe thata	can you guess what my favourite colour is
abola	verb			make bigger	abola lefuti leo gore medi e kgone go tsena se	make the hole bigger so that the roots can fit in easily
abula	verb		gagaba	crawl	thato o santse a ithuta go abula	thato is just learning to crawl
abuti	noun	cl,10-plural cl2abo	aubuti,mogolole	elder bother	thabo ke abuti	thabo is my elder brother
adima	verb			borrow	1.tsaya sengwe mo mongweng ka tumelano y	I borrowed a pen from her
adimia	verb			hire out	ba adimisa dikoloi mo bajanaleng	they hire out cars for taurists
aega	verb			lean	aega lleve mo lebotaneng	lean the ladder against the wall
aena	verb			iron	ke feditse go aena	I have finished ironing
aena	noun	cl9,n-plural c1o din		iron	dirisa aene e e bothitho ntsha masosoba	use a warm iron to remove these creases
aesekirimi	noun	ci 9-plural ci 10		ice cream	ke bokae leswana la aesekirimi	how much ias a scoo[of ice cream
afe	enumerative	c1 mo-singular cl 5-le		which one	maboko a gagwe ke efe	which ones are his poem
aga	verb			build	dira bonno ka go bopela	construct.we will build a house here
aga	adverb			always	se se afa se diragala	this always happens
agelela	verb			put a fence round,enclose	agelela jarata ka terata go kganelela dinkukwa	put a fence round the yard to keep the sheep out
agisanya	verb			reconcile	dira gore bayho ba tshedisane ka kagiso.boo b	the two have been making every effort to reconvile
ahaa	interjection			aha	ahaa jannong ke a bona gore o raya jang	aha, now I see what you mean
ahee	interjection			oh[denotes agreement]	ahee jaanong ke a tlhaloganya	oh yes, now I understand
ahelelee	interjection			got you		[used to show that somebody has been found out]
aids	abbreviation			aids		
aitse	adverb			truly	aitse ba dirile tiro e ntle e le ruri	they have truly done a magnificent job
aiyelele	interjection			yippee		
aka	verb		fosa	lie	ga o tshwanela go aka	you should not tell a lie
akabala	verb			be amazed,be astonished	ke ne ka akavala tota fa ke utlwa gore o tlhots	I was absolutely amazed when I heard he'd failed exams
akabatsa	verb			amaze,astonish	o ne a are akabatsa ka go o dire jang	she amazed us by winning the race
akantsha	verb			help with ideas	ke tlaa go akantshagore o dire jang	I will help you with ideason what to do
akanya	verb		nagana,gopola	think	o akanya gore re dire jang	what do you think we should do

Figure 6.6: Oxford extracted data

baba	hot				
baba	skin irritation	itch			
babalesegile	not in danger	safe			
babogedi	people watching or listening to a show	audience			
badirelwa	someone who plays for a service	client			
baenari	of number system based on 0 and 1	binary			
baeng	guest	company			
baesekele	1 two-wheeled vehicle powered by pedalling 2 short of bicycle, motorbike or motorcycle	bicycle	bike		
baesekopo	cinema picture; movie	film	movie		
baete	unit of electonic data	byte			
baka	to bake				
baka	to make something happen	cause			
baki	1 neat jacket worn by learners of a school or club members 2 garment worn over cloths for warmth	blazer	jacket		
bakilwe	cooked in oven	baked			
bakiphaloso	floating jacket to prevent one from drawning	life jacket			
bakukunedi	member of a small irregular fighting group, usually with a political aim, kind of warfare	guerrilla	guerilla		
balei	formal dance performed with pointed toes	ballet			
balekhoni	platform that a door leads to, built onto an outside wall one storey or more up	balcony			
baluni	rebber sac filled air	ballon			
bana	young boys or girls	children			
bangwe	not all, a few	some			
banka	financial institution where money is held and managed	bank			
banka ya lebota	a machine used for banking transaction such as drawing money	automated teller	ATM		
banna	word or sign on men's toilet door	gents			
bapa	of lines that are always the same distance apart	parallel			
bapisa	to find similarities and differences between things	compare			
bara	place where alcohol is served	bar			
barekisamebileng	person selling goods from non-permanent retail space	hawker			
basadi	sign on toilet for females	ladies			
baswa	young person or people	youth			

Figure 6.7: Pharos extracted data



Lemmas	Tswan lemmas or GAP	Gloss	Tswan gloss (or approximate translation if GAP)	Examples	Tswana examples
horse, cavalry, horse cavalry	pitse, batlhabani ba ba palamang dipitse	troops trained to fight on horseback	masole a a tlhometseng a a katisiwang go lwana ba palame dipitse	500 horse led the attack	dipitse tse 500 di ne di eteletse pele tlhaselo
generation	tshika	group of genetically related organisms constituting a single step in the line of descent	setlhopha sa ditshedi tse di amanang ka losika se se bopang karolo ya kgato e le nngwe mo		
string, twine	thudi	a lightweight cord	mogala wa boimatefo		
anthem, hymn	pina	a song of praise (to God or to a saint or to a nation)	pina ya kgalaletso (go Modimo kgotsa moitshepi kgotsa setshaba)		
scar, cicatrix, cicatrice	lebadi	a mark left (usually on the skin) by the healing of injured tissue	letshwao le le tlogelwang (gantsi mo letlalong) ke go gola ga thišu e e gobetseng		
pass, passing play, passing game, passing	phasa, motshameko wa go phasetsana	(American football) a play that involves one player throwing the ball to a teammate	(Motshameko wa kgwele ya dinao ya kwa Amerika) motshameko o o akaretsang motshameki a le mongwe yo o latlhelang bolo kwa motshameking wa setlhopha sa gagwe	the coach sent in a passing play on third and long	mokatisi o rometswe mo motshamekong wa go phasetsana
resistor, resistance	sekganedi, kganelo	an electrical device that resists the flow of electrical current	sediriswa sa motlakase se se kganelang kelelo ya moela wa motlakase		
ra, atomic number 88, radium	ra, nomoro 88 ya atomo, radiamo	an intensely radioactive metallic element that occurs in minute amounts in uranium ores	elemente e e ntshang maatla a a rileng mo go tseneletseng, ya metale, e e diregang ka dilekano tse dinnyane mo manyeng a yuraniamo		
loyalty, trueness	boikanyego	the quality of being loyal	pharologantsho ya go nna boikanyego		
ar, argon, atomic number 18	ar, akone, nomoro 18 ya atomo	a colorless and odorless inert gas; one of the six inert gases; comprises approximately 1% of the earth's atmosphere	gase e e se nang mmala le e e se nang monkgo e e sa suteng; nngwe ya digase di le thataro tse di sa suteng; e e nang le bokana ka 1% ya lefaufau la lefatshe		
research, inquiry, enquiry	tlhotlhomiso, patlisiso	a search for knowledge	go batla kitso	their pottery deserves more research than it has received	dilwana tsa bone tse di bopilweng ka letsopa di tshwanela gore go dirwe tlhotlhomiso ka tsone go feta ka moo go kileng ga amogelwa ka teng
outfit	setlhopha sa badirammogo ba sesole	any cohesive unit such as a military company	yuniti epe e e kopanyang e e jaaka setlamo sa sesole		
ambassador, embassador	moambasetara, moemedi	a diplomat of the highest rank; accredited as representative from one country to another	modipolomate wa maemogodimo; yo o filweng maemo a go nna moemedi go tswa kwa nageng e e rileng go ya kwa go e nngwe		
reaction	boitsetsepedi	extreme conservatism in political or social matters	go dumela mo go feteletseng gore ga go a tshwanela go nna le diphetogo tse dintsi mo setshabeng malebana le dintlha tsa sepolitiki kgotsa tsa loago	the forces of reaction carried the election	maatla a boitsetsepedi a ne a tsamaisa ditlhopho
control	taolo	(physiology) regulation or maintenance of a function or action or reflex etc	(fisioloji) taolo kgotsa tshomarelo ya tiriso kgotsa tiro kgotsa tiro e e dirwang kwa ntle ga go nagana	the timing and control of his movements were unimpaired; he had lost control of his sphincters	nako le taolo ya metsamao ya gagwe di ne di sa fokotsega; o ne a latlhegetswe ke taolo ya mesifa e e kgonang go gagamadiwa le go repisiwa

Figure 6.8: Professionally translated data

Setswana Sentence	English Sentence
Kgolegelo ya ntlha e solofetswe go konotelwa ka Firikgong 2007	The first prison is expected to be completed by January 2007
Nako ya go aga e gorogile	The time to build has arrived
A o ka nagana	Can you imagine
Ke le eleletsa katlego e e tswelelang thata mo dipuisanong tsotlhe tsa lona	I wish you much continued success in all your deliberations
Fa re filwe matlhakore a le 3 re ka rarabolola dikhutlo dingwe le dingwe tse 3	When we are given 3 sides we can solve for any of the 3 angles
katlego le bohuma	Prosperity and poverty
Lephata la lelapa le tlhoka mekgwatiriso e e latelang	The household sector requires the following measures
Ditekanyo di akaretsa gore	Measures include that
Go tlhanola ga se tiro	Inverse is not a function
Dintlha tse dingwe tse di ka akarediwang go tshegetsa sekai sa moithuti	Other points which may be included to support candidates ' argument
Tse ga di ise di laolwe sentle mme ditsamaiso di tshwanetse go tsengwa tirisong go siamisa :	These have not been well controlled and measures must be implemented to rectify this
Ke raya gore a mme se sa ntse se dira tlhotlhomiso ka ga mathata a re neng re lebane le one	I mean is it still doing research on the problems that were facing
Eno e tla nna tiragalo e kgolo go di feta tsotlhe mo lefatsheng e e kileng ya nna gone	This will be the greatest event on earth ever
Re tshwanetse go tlhotlheledi wa ke boitshwaro jo bo siameng mo kgwebong	In the conduct of business we have to be values-driven
Re tlhoka kelo e e oketsegileng ya kgolo ya lotseno le bothapiwa	We seek a faster rate of growth of incomes and employment
Ba tlhomamisa gore ga re boeletse diphoso tse di dirilweng go sele	They are ensuring that we do not repeat mistakes that have been made elsewhere
Go maleba le mo go Karolo ya puso e e dirang tiro e e kgethegileng	The same is applicable to the Organ
Kgato ya Motheo ke tokomane e e farologaneng	Foundation Phase is a separate document
Ela tlhoko ka go tshwenyega go tlhatloga ga ditlhwatlhwa le dithwatlhwa tse di kwa godim	Note with concern the escalating food prices and high energy prices
Re ikaeletse go fitlhelela palogare ya ditiro di le 200000 ka ngwaga mo dingwageng di le tlha	We are looking at averaging 200000 jobs annually over the next five years

Figure 6.9: Autshumato parallel data

The data was consolidated, in the progress the duplicate entries and unnecessary data variables were removed. Subsequently, the remaining data was mapped and formatted according to the procedure detailed in section 5.4.2. This rigorous data preparation resulted in the refined set of variables illustrated in Figure 5.4.

The selected sentences from the parallel corpus serves as a ground truth benchmark, enabling a reliable assessment of the performance of the machine-translation algorithms employed in this study. To ensure the highest quality and reliability of the evaluation dataset, only sentences from published dictionaries were considered. This approach guarantees that the dataset has undergone rigorous quality assurance processes, minimizing the presence of errors or inconsistencies that could potentially skew the evaluation results. The use of a ground truth dataset from reputable sources is crucial for obtaining accurate and consistent insights into the performance of the MT system. The selected evaluation dataset comprises a total of 5,000

sentence pairs, providing a substantial and diverse sample for assessing the MT system's performance across various linguistic contexts and complexities.

The use of a high-quality, representative, and adequately sized evaluation dataset is essential for evaluating the effectiveness of the MT system. By leveraging the parallel corpus sourced from the Oxford and Pharos dictionaries, this study ensures that the evaluation results are based on a reliable and comprehensive benchmark, enabling a fair and accurate assessment of the MT system's performance in the context of Setswana-English translation.

6.2.2 Evaluation metrics

To evaluate the performance of the proposed model, both intrinsic and extrinsic evaluation measures were used. Intrinsic evaluation metrics accuracy (6.1), precision (6.2), recall (6.3) and F1 score (6.4) were used.

$$A = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative} \quad (6.1)$$

$$P = \frac{true\ positive}{true\ positive + false\ positive} \quad (6.2)$$

Accuracy (A) is the ratio of correctly predicted instances to the total number of instances. Precision (P) is the ratio of true positive predictions, the correctly predicted positive instances to the sum of true positive and false positive predictions.

$$R = \frac{true\ positive}{true\ positive + false\ negative} \quad (6.3)$$

Recall is the ratio of true positive predictions to the sum of true positive and false negative predictions, instances that are positive but incorrectly predicted as negative. It measures the model's ability to identify all positive instances in the dataset.

$$F = \frac{2(precision * recall)}{precision + recall} \quad (6.4)$$

The F1 score (F) is the harmonic mean of precision and recall. The F1 score provides a balance between precision and recall. Figure 6.3 shows the proposed model's performance.

Extrinsic evaluation measure was employed to measure the performance of the downstream Setswana-English MT application. The bilingual evaluation understudy (BLEU) evaluation

metric by Papineni *et al.* (2002) was used. This metric measures how well a machine-generated translation aligns with human-generated reference translations using Equation 6.5.

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \cdot log\big(\, precision_{\,n}\big)\right) \quad (6.5)$$

BLEU is based on precision, which calculates the percentage of n-grams in the machine-generated translation that also appears in the reference translation. The score ranges from 0 to 1, with 1 indicating a perfect match between the machine-generated translation and the reference translation. One of the components of the BLEU score calculation is the brevity penalty (BP). The BP is a metric used to penalize translations that are too short compared with the reference translations. The proposed algorithm was tested on professionally translated Setswana-English pieces of data. Using the NLTK library in Python Anaconda, the output translation of each Setswana sentence was passed to an evaluation function with candidate and reference translation, in which individual precision and brevity penalty was calculated.

Individual precision is a metric used to evaluate the precision of a translation system at the individual sentence level, while brevity penalty addresses the issue of length differences between the candidate and reference translations. The brevity penalty penalizes the system if the generated translation is shorter than the reference translation, to prevent the system from favouring shorter translations to achieve higher precision scores. The overall BLEU score is calculated by combining these two metrics.

6.2.3 Experimental setup

All experiments were conducted using the Python programming language with preinstalled NLTK (Bird, Klein and Loper 2009) and scikit-learn library (Pedregosa *et al.* 2011) run in Windows 10 Pro, 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz 1.69 GHz, 17GB installed ram. For experimentation, each context sentence was passed into the algorithm for pre-processing and encoding. Thereafter, the glosses of the polysemous word in that context were extracted from the UKC, and subsequently encoded iteratively to generate gloss embeddings. With each iteration, a similarity measure was computed between the context and gloss embeddings; and the gloss with the highest similarity was selected as the appropriate sense gloss, which was then compared with the ground truth.

6.3 Experimental Evaluation

6.3.1 WSD evaluation and results

6.3.1.1 Evaluation

To evaluate the WSD model, the study utilized the dataset presented in section 6.2.1.1. The evaluation file includes a column for the ground truth, which contains the correct sense of the target word against which the algorithm's output is compared. As explained in detail in Chapter 5, section 5.5, the model receives an input in the form of a Setswana context sentence, which is pre-processed. The model further inputs candidate senses of the target word from the dataset. Both the sentence and senses are vectorized using PuoBERTa to generate sentence embeddings. The PuoBERTa-Enabled Embedding-based Lesk WSD Model then outputs what it determines to be the correct sense of the target word. This returned sense is compared to the sense in the ground truth column, the Setswana context gloss to assess the model's accuracy. From the output, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are calculated and computed. A true positive occurs when the model correctly identifies the sense of a polysemous word, matching the ground truth. A false positive occurs when the model incorrectly identifies a sense for a polysemous word that doesn't align with the ground truth. A true negative is recorded when the model accurately identifies a word as monosemous, in line with the ground truth. A false negative occurs when the model fails to identify the correct sense of a polysemous word, either by selecting an incorrect sense or mistakenly classifying it as monosemous. These metrics are used to calculate the model's precision, recall, and F1 score, which is then computed using the evaluation metrics in section 6.2.2 above.

6.3.1.2 WSD results

The results presented in this section encompass the evaluation outcomes of the WSD model, to illustrates the metrics that are reported for and not necessarily for a comparative analysis as the reported results are for different languages with different datasets (see Appendix C, Table 6.1), which presents the reported results.

Furthermore, as this study used the Lesk-based WSD algorithm, a dedicated comparative analysis is conducted specifically for Lesk-based methods. It is important to note that most studies in this category solely report accuracy as the primary performance metric. Consequently, the comparative analysis of Lesk-based approaches is only based on the accuracy metric. Figure 6.3 presents a visual representation of the comparative performance based on accuracy.

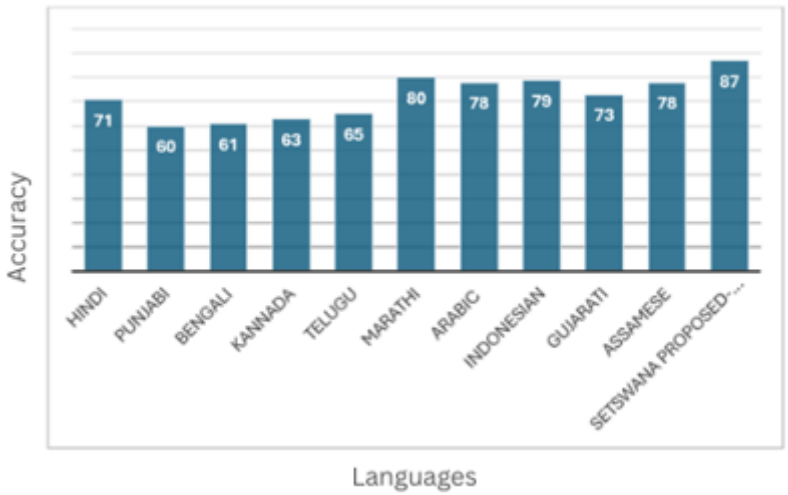


Figure 6.10: Comparative analysis of proposed Lesk with other Lesk-based algorithms

It crucial to note that, currently, there is no existing WSD model specifically for Setswana, nor is there a benchmark dataset available for this language. This absence of a WSD model and benchmark dataset presents a significant challenge for direct evaluation and comparison. Consequently, the comparative analysis included in this study is conducted against other Lesk-based algorithms, which, although evaluated on different datasets, represent the most feasible approach given the current limitations. While acknowledging that comparing models across different datasets may introduce variations, this study aimed to provide the closest possible comparison given the constraints.

6.3.2 MT evaluation and results

6.3.2.1 Evaluation

The evaluation measure used in the experiment, BLEU, was explained in detail in section 6.2.2. The evaluation dataset comprised professionally translated Setswana-English sentence pairs extracted from Oxford and Pharos dictionaries, as well as government domain data. This dataset included sentences with monosemous words. These reference translations served as the gold standard against which the machine-generated translations were compared. The RBMT system was then applied to the prepared Setswana sentences, producing English translations. These output translations were systematically compared against the reference translations using the predefined metrics, with the BLEU score being computed automatically. Discrepancies between the machine output and the reference translations were carefully noted and analyzed to identify specific areas where the RBMT system was underperforming. Finally, the results of

these evaluations were compiled and thoroughly analyzed to assess the overall performance of the translation system and to highlight areas for potential improvement.

6.3.2.2 MT results

Table 6.1 provide overview of the evaluation results, with a comparative analysis with the existing Setswana-English machine-translation systems.

Table 6.1: MT Results and Comparative Analysis

Research	**Language Pairs**	**Dataset**	**BLEU**
Wilken, Griesel and McKellar (2012)	English-Setswana	Autshumato (governmental domain)	28.80
Abbott and Martinus (2018)	Setswana-English	Autshumato dataset (governmental domain)	33.12
Martinus and Abbott (2019)	English-Setswana	Autshumato parallel corpora (governmental domain)	15.60
Lastrucci *et al.* (2023)	Setswana-English	Vuk'uzenzele, ZA-Gov-multilingual, Autshumato (governmental domain)	29.84
Rule-Based Machine Translation	Setswana-English	Autshumato (governmental domain) + Oxford, Macmillan, Oxford Kiddies, Pharos, and Shuter's	30.62

A BLEU score of 30.62 out of a possible 100 indicates that the RBMT model's translations, while not perfect, show a moderate level of similarity to the reference translations. This score suggests that the model is capturing some of the meaning and structure of the original Setswana text, but there is still significant room for improvement.

6.3.3 WSD-MT evaluation and results

6.3.3.1 Evaluation

The evaluation procedure employed for the WSD-MT is the same as the approach outlined in section 6.3.2.1. However, this evaluation incorporated an additional element, the context sentences used in section 6.2.1.1 that contained polysemous words. This inclusion allowed for

a more comprehensive assessment of the system's ability to handle ambiguous terms in translation.

6.3.3.2 WSD-MT results

Table 6.2 provide overview of the WSDT-MT evaluation results, with a comparative analysis with the existing Setswana-English machine-translation systems.

Table 6.2: WSD-MT Results and Comparative Analysis

Research	**Language Pairs**	**Dataset**	**BLEU**
Wilken, Griesel and McKellar (2012)	English-Setswana	Autshumato (governmental domain)	28.80
Abbott and Martinus (2018)	Setswana-English	Autshumato dataset (governmental domain)	33.12
Martinus and Abbott (2019)	English-Setswana	Autshumato parallel corpora (governmental domain)	15.60
Lastrucci *et al.* (2023)	Setswana-English	Vuk'uzenzele, ZA-Gov-multilingual, Autshumato (governmental domain)	29.84
Rule-Based Machine Translation	Setswana-English	Autshumato (governmental domain) + Oxford, Oxford Kiddies, Pharos, and Shuter's	30.62
PuoBERTa Enabled Embedding-based Lesk WSD Model, the Setswana-English RBMT	Setswana-English	Autshumato (governmental domain) + Oxford, Oxford Kiddies, Pharos, and Shuter's	34.89

The PuoBERTa Enabled Embedding-based Lesk WSD Model, Setswana-English RBMT performance increased from 30.62 to 34.89 after incorporating the WSD functionality. This indicates that the indicates that the WSD-enhanced model produces more accurate translations that are closer to the reference translations. The improvement likely reflects the model's enhanced ability to correctly interpret and translate ambiguous words or phrases in context, leading to more appropriate word choices and better capture of the intended meaning of the source text.

6.4 Discussion

The results presented in Appendix C, Table 6.1 encompass the evaluation outcomes of the WSD models of WSD research found in the literature across various languages and approaches. To maintain consistency across the reported results, studies that did not report precision, recall, and F1 score were excluded from the table. Among the 54 studies reviewed, 68.5% focused on WSD for the English language, employing a range of approaches, methods, and techniques. The prevalence of English WSD research can be attributed to the availability of rich lexical resources, such as the Senseval datasets, commonly used as standard benchmark dataset, which offer diverse and substantial data for training, testing, and algorithm development. The availability of these lexical resources has been a key enabler for the extensive exploration of WSD in the English language.

The analysis reveals that semi-supervised approaches yield the highest average score of 0.87 for English WSD, underscoring their effectiveness in accurately disambiguating word senses. This finding aligns with the literature, which emphasizes the advantages of combining both supervised and unsupervised methods to leverage their respective strengths. The meta-analysis study further corroborates this observation, confirming that supervised methods generally achieve superior results; and the integration of multiple approaches enhances overall performance. However, the application of these methods to Setswana WSD is currently hindered by the scarcity of both annotated and unannotated data required for comprehensive model training. As a result, this study opted for a knowledge-based approach as a viable alternative, for a low-resource language, such as Setswana.

Arabic emerged as the second-most studied language, with four WSD research papers included in the analysis. The highest average F1 score of 0.89 is obtained using a supervised approach, further reinforcing the notion that supervised methods excel when sufficient training data is available. Hindi, with three studies, demonstrates a notable average score of 0.88 using knowledge-based techniques. As a morphologically rich language characterized by a complex system of inflectional morphology and word structures, Hindi WSD benefits from the ability of knowledge-based approaches to capture and leverage linguistic information. The combination of multiple morphemes, such as prefixes, suffixes, and infixes, to convey different meanings, tenses, genders, and numbers, poses unique challenges for WSD in Hindi. The high average score achieved by knowledge-based methods in Hindi aligns with the findings of this study, which reports an average score of 0.89 for Setswana, another morphologically rich language.

Assamese, with two studies, exhibits an average score of 0.85 using both unsupervised and supervised approaches, particularly the naïve Bayes algorithm, which is widely employed in WSD tasks. French, Australian, and Bengali, each represented by a single study, show varying performance scores of 0.55, 0.64, and 0.80, respectively, with the highest score obtained using a knowledge-based method.

The success of knowledge-based approaches in morphologically rich languages, such as Hindi and Setswana, highlights the potential of leveraging linguistic knowledge and resources to tackle the complexities of these languages. The scarcity of both annotated and unannotated data for Setswana, similar to other low-resource languages, necessitates the exploration of alternative approaches that can effectively capture and utilize the language's unique features and structures.

In terms of the WSD results, Sharma and Joshi (2019) used an evaluation corpus of 3000 context sentences, of which 2143 were correctly disambiguated, achieving an accuracy of 71% for Hindi. Singh and Singh (2015) tested the modified Lesk on 15 Punjabi polysemous words and achieved an average accuracy of 60% for all the ambiguous words. Pandit *et al.* (2018) used two test sets, the first test set with 10, and the second test set with 12 Bengali polysemous words, achieving an accuracy of 61%. Using a single word with 5 different senses and 2153 context sentences, researchers Kannada, Parameswarappa and Narayana (2011) obtained 63% accuracy; while Eluri and Siddu (2020) obtained 65% testing with 150 context sentences for Telegu. Patil *et al.* (2021)'s algorithm was evaluated on 6 Marathi polysemous words, with a total 14 senses; and achieved an overall accuracy of 80%. Arabic (Zouaghi, Merhbene and Zrigui 2011) and Indonesian (Basuki *et al.* 2019) achieved almost the same accuracy with 0.1% difference, 78% and 79%. The Arabic WSD was tested on 50 polysemous words with 20 context sentences per word. For Indonesia, the algorithm was evaluated on 140 context sentences. Assamese obtained 78% evaluated on an annotated Assamese corpus with 15606 polysemous nouns. This study achieved an accuracy of 87% on 1200 Setswana context sentences, of which 602 contained target words with additional 408 sentences with repeated target words plus 190 sentences that contained monosemous words and 57 polysemous words.

For the machine translation, the initial study conducted by Wilken, Griesel and McKellar (2012) on Autshumato achieved a BLEU score of 28.80; followed by Abbott and Martinus (2018) with a score of 33.12; and Martinus and Abbott (2019) with a score of 15.60. Notably,

all these studies were conducted on the same dataset. Subsequently, in Lastrucci *et al.* (2023), an additional dataset was introduced, resulting in a BLEU score of 29.84. In comparison, the proposed work obtained a BLEU score of 30.62 without the WSD integration.

In Table 6.2, Abbott and Martinus (2018) MT slightly outperformed our model, with a BLEU score of 33.12. Despite some inaccuracies, the model's translations often make sense. For example, it translated "moithuti yo o iseng a aloge wa ngwaga wa bobedi" as "student in second year" instead of the reference "a second-year undergraduate". While not exact, this translation is still meaningful, showing the model's ability to convey the general idea even when specific terms are missing from its vocabulary. Another interesting observation is the model's use of synonyms which are the words in the dictionary. For instance, it translated a phrase as "to place or press foot on" instead of "to place or press leg on". This demonstrates the model's ability to use related terms, as "foot" and "leg" are often interchangeable in Setswana (both referred to as "leoto"). These examples highlight the challenges in translating between languages with different linguistic properties. While the model may not always produce perfect translations, it often captures the essence of the original text.

The BLEU improved from 30.62 to 34.89 after incorporating WSD significantly improving the model. While an increase of 4.27 points might seem small, in BLEU score terms, this represents a substantial improvement, especially considering the complexity of translation between dissimilar languages like Setswana and English. Despite this progress, there's still room for enhancement, as a score of 34.89 out of 100 indicates that the translations, while better, are not yet at a level comparable to human translation. However, this improvement validates the approach of integrating WSD into the machine translation process, demonstrating that addressing word sense ambiguity does indeed lead to better translation outcomes. Table 6.3 illustrates how the WSD improved the translations.

Table 6.3: WSD-MT translations improvements

Source: Mosadi o ne a apere kobo e ntsho MT: The woman was wearing a black blanket WSD-MT: The woman was wearing a black dress *The word "kobo" can mean both "blanket" and "dress" in Setswana. WSD helped the system understand that in the context of clothing, "kobo" likely refers to a dress*
Source: O ne a lela ka lentswe le le kwa godimo MT: He was crying with a voice that is up high

WSD-MT: He was crying loudly *The phrase "lentswe le le kwa godimo" literally means "voice that is up high," but WSD helped the system understand that in this context, it means "loudly."*
Source: O ne a noka nama ka letswai MT: He river the meat with salt WSD-MT: He seasoned the meat with salt *The word "noka" can mean "waist" "river" and "season" in Setswana. WSD helped the system understand that in the context of clothing, "noka" likely refers to a season*

Although a direct comparison with the previously developed Setswana-English MT translation cannot be made due to the use of different datasets, this study demonstrates that expanding the dataset to address data sparsity and incorporating WSD significantly improves MT performance.

In terms of the datasets, the Autshumato dataset is a parallel corpus that was extracted from government resources and web crawls created by Abbott and Martinus (2018). The ZA-Gov-multilingual dataset constructed by Lastrucci *et al.* (2023) consists of scraped data from the South African government websites, in which all the state documents are translated and publicly shared. The Vuk'uzenzele dataset was constructed from editions of the South African government magazine, which mainly covers present political content. The Autshumato dataset, ZA-Gov-multilingual, and Vuk'uzenzele dataset exhibit similarities, and are comparable in terms of content and characteristics.

The limitation of freely available resources has prompted the utilization of the only accessible data, which originates from the government domain. While training machine-translation systems using government domain data can offer specialized language knowledge, this introduces challenges such as limited vocabulary, limited style variation, bias, and domain specificity. The formal and specific language style found in government documents results in restricted vocabulary and less diversity in linguistic expressions, potentially causing difficulties for the machine-translation system in handling colloquial or informal language. In addition, biases inherent in government domains could be unintentionally encoded into the machine-translation model; and the domain-specific nature of the training data may limit the system's performance when faced with content from other domains.

To address these challenges, in this study, Setswana-English bilingual dictionaries from Oxford, Oxford Kiddies, Pharos, and Shuter's were incorporated into existing datasets to expand the data and coverage. This addition introduces variations to the datasets traditionally used for Setswana-English machine translation. The integration of diverse data sources is crucial because it enhances the robustness and adaptability of machine-translation systems, ensuring their effectiveness across a broader range of linguistic contexts, making this a contribution of this study. Another contribution is that the translation algorithm outperformed existing Setswana-English machine-translation systems, obtaining a higher BLEU score by 1.77%. The study attributes this improvement to the integration of WSD context modelling in dictionary design and algorithm development.

Although a direct comparison cannot be conducted as a results of employing different datasets, this study proved that increasing the dataset to address data sparsity and incorporating the WSD does improve the performance of MT.

6.5 Chapter Summary

This chapter provided a comprehensive overview of the experimentation, evaluation, and results for three key components of the research: the PuoBERTa Enabled Embedding-based Lesk WSD Model, the Setswana-English RBMT model, and the integrated WSD-MT model. The chapter started by outlining the evaluation framework, detailing the development of evaluation datasets, metrics, and the experimental setup. The chapter then presented in-depth evaluations and results for each model. The WSD, MT process, and the combined WSD-MT approach, analyzing their individual and integrated performances were discussed. The discussion section interpreted these results, contextualizing them within the broader research objectives and highlighted key findings. Through this systematic evaluation, this chapter demonstrated the applicability of the developed PuoBERTa Enabled Embedding-based Lesk WSD Model for Setswana-English machine translation.

CHAPTER SEVEN: SUMMARY, CONCLUSION AND FUTURE WORKS

7.1 Summary

The primary focus of this study was on answering the following main research question:

How can a Setswana WSD model that captures the linguistic properties of Setswana based on the adaptation of existing WSD language models be developed for Setswana-English MT?

which led to the main research objective:

Develop a Setswana WSD model that captures the linguistic properties of Setswana based on the adaptation of existing WSD language models in the context of Setswana-English MT. Effectively address the ambiguity challenges specific to Setswana.

A comprehensive analysis was conducted to examine ambiguities inherent in Setswana. This constituted Research Question 1:

- What is the nature and extent of semantic ambiguities in Setswana in the context of Setswana and English MT?

This research question led to Research Objective 1, which was:

- Perform a systematic analysis of the nature and extent of semantic ambiguities in Setswana.

Using a literature review and systematic analysis as a methodology, Chapter 2 comprehensively studied and presented the Setswana language's linguistic and ambiguity properties. This chapter first introduced the origin of the Setswana language and the semantic and syntactic nature of Setswana by outlining the different word groups, orthography, phonology, and morphology, as well as various sentences, being simple, complex, and compound sentences. The chapter lastly introduced various language ambiguities in Setswana. These ambiguities are categorized as lexical, orthographic, structural, and pragmatic ambiguities. To this end, a taxonomy of Setswana ambiguities was presented. Researchers can use this taxonomy to better understand the various ambiguity types and their characteristics, which can assist in the development of more effective and accurate NLP models and algorithms that cater to these specific ambiguities and linguistic characteristics. This study's scope was to focus on lexical ambiguities.

Given the Setswana ambiguity challenges, the second research question was:

- What existing language model can be adapted for Setswana WSD using knowledge-based approaches?

This led to the Second Research Objective, which was:

- Identify an existing language model for Setswana WSD using a knowledge-based approach.

Chapter 3 was conducted to determine which existing language model can be adapted for Setswana WSD using knowledge-based approaches, given the ambiguity and linguistic characteristics identified in Chapter 2.

Using various types of literature analysis, this chapter provided a comprehensive review of existing models in WSD; and the research work that has been published in the literature using both qualitative and quantitative analysis. The chapter introduced and described the task of WSD and provided an overview of various WSD approaches, knowledge resources, evaluation methods, and the application of WSD in NLP. The aim of this overview was to provide a comprehensive understanding of WSD and its significance in NLP research and applications.

The chapter further presented a comprehensive review of the research conducted on WSD as a stand-alone task; and its application within machine translation (MT). This included a diverse range of studies conducted in different contexts, employing various approaches, methodologies, techniques, and knowledge resources. The primary objective of this section was to provide a thorough synthesis of the studies that were conducted, emphasizing the key findings that have emerged from previous and current research across different languages.

For quantitative analysis, the chapter conducted the bibliometric analysis, providing a comprehensive overview of the WSD research landscape and highlighting the key trends, influential works, and prominent researchers in the field. A meta-analysis was presented to provide a systematic synthesis of findings from multiple studies, offering a quantitative approach to summarizing and integrating results and research outcomes within the WSD research landscape.

Chapter 3 also identified the WSD open research problems; and discussed gaps in the literature. The findings from this chapter were used to determine which existing language model can be adapted for Setswana WSD using knowledge-based approaches, considering the ambiguity and linguistic characteristics identified in Chapter 2. The chosen and identified existing WSD model was the simplified Lesk model.

Having identified the model, the Third Research Question was:

- How can a Setswana WSD model that captures the linguistic properties of Setswana based on the identified model be developed for Setswana-English MT?

This research question led to the third research objective:

- Develop a Setswana WSD model that captures the linguistic properties of Setswana for Setswana-English MT.

Chapters 4 and 5 answered the research question and addressed the third research question.

Chapter 4 presented the theoretical framework of this thesis, providing a foundational review of theories, models, and related concepts that formed the basis for the methodology used in this study. The chapter introduced the distributional semantic hypothesis and explored the distributed representations of words, highlighting essential theories and concepts related to distributed word representation. It also presented the Lesk algorithm and its variations, BERT, and the transformer architecture.

Chapter 5 presented the methodology employed in developing a PuoBERTa Enabled Embedding-based Lesk WSD Model knowledge-based WSD model for Setswana in the context of Setswana-English MT. The primary objective of this chapter was to provide a detailed description of the practical application and adaptation of the resource and theoretical frameworks discussed in Chapter Four. The chapter presented the graphical representation of the methodological framework and described the resource construction process and the mapping process. The WSD-MT was also presented in this chapter.

Given the data sparsity challenge in Setswana NLP, and the language corpus dependency of WSD models, the fourth question was:

- How can a lexical resource knowledge base suitable for the development and evaluation of a Setswana WSD model be developed?

which resulted in the objective:

- Develop a lexical resource knowledge base appropriate for the development and evaluation of a Setswana WSD model.

The lexical knowledge bases and the models used to guide the development of the Setswana resources, including the Princeton WordNet (PWN), Universal Knowledge Core (UKC), African WordNet (AWN), and the bilingual dictionaries used for data extraction, were discussed in Chapters 4 and 5. The chapters also presented the RBMT architecture adopted for Setswana-English machine translation and the similarity measures used in NLP. The chapters further detailed how the various components interconnect to form the resource-construction framework proposed in this research that guided the development of the lexical resource.

With the availability of a WSD-MT model and the lexical resource, the next task was to evaluate the proposed models, WSD as a stand-alone, and within the MT task. This led to the Fifth Research Question:

- How can this WSD model be experimentally evaluated?

This research question led to Research Objective 4, which was:

- Experimentally evaluate the model.

This research question was answered in Chapter 6. This chapter presented the experimentation, evaluation, and results of the proposed models. The objective of this chapter was to provide a comprehensive analysis and assessment of the developed algorithms, demonstrating their

performance and applicability in addressing the identified research challenges. The chapter described the evaluation framework used in this study and zoomed into the development of evaluation datasets, evaluation metrics, and the experiment settings. The results and analysis were also presented in this chapter.

7.2 Conclusion

The thesis statement that this study was set to validate is:

A knowledge-based WSD model for a morphologically rich low-resourced language such as Setswana, can enhance the performance of rule-based machine translation involving the language and a less morphologically rich high-resourced language, such as English.

This thesis successfully validated the aforementioned statement through the following:

Using the developed lexical resources for Setswana, the research led to the development of a PuoBERTa Enabled Embedding-based Lesk WSD Model knowledge-based WSD model that achieved an impressive accuracy of 87%. Furthermore, the integration of this WSD model into existing machine-translation systems resulted in a significant improvement in Setswana-English translation quality, as evidenced by an increase in the BLEU score from 30.62 to 34.89 for rule-based machine translation.

7.3 Future Works

In future research, the following avenues will be explored:

- Refinement of the algorithm: continuous refinement of the Setswana WSD algorithm to improve its accuracy and adaptability to evolving linguistic nuances.
- Language resource expansion: the Setswana UKC will be continuously expanded and updated to accommodate emerging lexical and semantic variations, ensuring the resource's growth relevance over time.
- Cross-linguistic adaptability: exploring the adaptability of the developed WSD model to other Bantu languages with similar linguistic characteristics to broaden the model's applicability and impact.
- Web service and real-world deployment: further work will focus on the practical deployment of the developed model in real-world translation web service settings, considering user feedback and addressing challenges that may arise in practical applications. This will enable more extensive user-centric evaluations.

References

Abbott, J. Z. and Martinus, L. 2018. Towards neural machine translation for African languages. *arXiv preprint arXiv:1811.05467*, Article ID.

Abdalgader, K. and Al Shibli, A. 2021. Context expansion approach for graph-based word sense disambiguation. *Expert Systems with Applications*, 168: 114313.

Abderrahim, M. A. and Abderrahim, M. E.-A. 2022. Arabic Word Sense Disambiguation for Information Retrieval. *Transactions on Asian and Low-Resource Language Information Processing*, 21 (4): 1-19.

Abid, M., Habib, A., Ashraf, J. and Shahid, A. 2018. Urdu word sense disambiguation using machine learning approach. *Cluster Computing*, 21: 515-522.

Abouenour, L., Bouzoubaa, K. and Rosso, P. 2013. On the evaluation and improvement of Arabic WordNet coverage and usability. *Language resources and evaluation*, 47: 891-917.

Adeliyi, T. T., Ogunsakin, R. E., Adebiyi, M. and Olugbara, O. 2021. A meta-analysis of channel switching approaches for reducing zapping delay in internet protocol television. *Indonesian Journal of Electrical Engineering and Computer Science*, 22 (3).

Agirre, E. and Edmonds, P. 2007. *Word sense disambiguation: Algorithms and applications*. Springer Science & Business Media.

Agirre, E. and Martinez, D. 2001. Knowledge sources for word sense disambiguation. In: Proceedings of *Text, Speech and Dialogue: 4th International Conference, TSD 2001 železná Ruda, Czech Republic, September 11–13, 2001, Proceedings 4*. Springer, 1-10.

Agirre, E., De Lacalle, O. L., Fellbaum, C., Hsieh, S.-K., Tesconi, M., Monachini, M., Vossen, P. and Segers, R. 2010. Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In: Proceedings of *Proceedings of the 5th international workshop on semantic evaluation*. 75-80.

AlMousa, M., Benlamri, R. and Khoury, R. 2022. A novel word sense disambiguation approach using WordNet knowledge graph. *Computer Speech & Language*, 74: 101337.

Angelina, B. and Loukachevitch, N. 2020. All-words word sense disambiguation for Russian using automatically generated text collection. *Cybernetics and Information Technologies*, 20 (4): 90-107.

Awasthi, I., Gupta, K., Bhogal, P. S., Anand, S. S. and Soni, P. K. 2021. Natural language processing (NLP) based text summarization-a survey. In: Proceedings of *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 1310-1317.

Baiju, V. 2022. Word Sense Disambiguation in the domain of Sentiment Analysis through Deep Learning. Article ID.

Banerjee, S. and Pedersen, T. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In: Proceedings of *International conference on intelligent text processing and computational linguistics*. Springer, 136-145.

Baruah, N., Gogoi, A., Sarma, S. K. and Borah, R. 2021. Utilizing corpus statistics for Assamese word sense disambiguation. In: Proceedings of *Advances in Computing and Network Communications: Proceedings of CoCoNet 2020, Volume 2*. Springer, 271-283.

Basuki, S., Kholimi, A. S., Minarno, A. E., Sumadi, F. D. S. and Effendy, M. R. A. 2019. Word sense disambiguation (WSD) for Indonesian homograph word meaning determination by LESK algorithm application. In: Proceedings of *2019 12th International Conference on Information & Communication Technology and System (ICTS)*. IEEE, 8-15.

Beckwith, R., Fellbaum, C., Gross, D. and Miller, G. A. 2021. WordNet: A lexical database organized on psycholinguistic principles. In: *Lexical Acquisition*. Psychology Press, 211-232.

Bennett, W. G., Diemer, M., Kerford, J., Probert, T. and Wesi, T. 2016. Setswana (South African). *Journal of the International Phonetic Association*, 46 (2): 235-246.

Berg, A. S. 2018. Computational syntactic analysis of Setswana. Article IDNorth-West University (South Africa), Potchefstroom Campus.

Berg, A., Pretorius, R. P. L., Butt, M. and King, T. H. 2013. The representation of Setswana double objects in LFG. In: Proceedings of *Proceedings of the LFG13 conference*. 111-130.

Bevilacqua, M., Pasini, T., Raganato, A. and Navigli, R. 2021. Recent trends in word sense disambiguation: A survey. In: Proceedings of *International Joint Conference on Artificial Intelligence*. International Joint Conference on Artificial Intelligence, Inc, 4330-4338.

Bevilacqua, M., Pasini, T., Raganato, A. and Navigli, R. 2021. Recent trends in word sense disambiguation: A survey. In: Proceedings of *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Bhatia, S., Kumar, A. and Khan, M. M. 2022. Role of Genetic Algorithm in Optimization of Hindi Word Sense Disambiguation. *IEEE Access*, 10: 75693-75707.

Bhatt, B. and Bhattacharyya, P. 2011. IndoWordNet and its linking with ontology. In: Proceedings of *Proceedings of the 9th International Conference on Natural Language Processing (ICON-2011)*.

Bhattacharjee, K., ShivaKarthik, S., Mehta, S., Kumar, A., Phatangare, S., Pawar, K., Ukarande, S., Wankhede, D. and Verma, D. 2020. Survey and gap analysis of word sense disambiguation approaches on unstructured texts. In: Proceedings of *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 323-327.

Bhattacharya, A., Natarajan, S. and Saha Roy, R. 2020. Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. In: Proceedings of *ACM India Joint International Conferenceon Data Science and Management of Data*. ACM.

Bird, S., Klein, E. and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Birjali, M., Kasri, M. and Beni-Hssane, A. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226: 107134.

Booth, A. D. and Locke, W. N. 1955. Machine translation of languages. Article ID.

Bordes, A., Glorot, X., Weston, J. and Bengio, Y. 2012. Joint learning of words and meaning representations for open-text semantic parsing. In: Proceedings of *Artificial intelligence and statistics*. PMLR, 127-135.

Bordes, A., Glorot, X., Weston, J. and Bengio, Y. 2014. A semantic matching energy function for learning with multi-relational data: Application to word-sense disambiguation. *Machine Learning*, 94: 233-259.

Borenstein, M., Hedges, L. V., Higgins, J. P. and Rothstein, H. R. 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, 1 (2): 97-111.

Boruah, P. 2022. A novel approach to word sense disambiguation for a low-resource morphologically rich language. In: Proceedings of *2022 IEEE 6th Conference on Information and Communication Technology (CICT)*. IEEE, 1-6.

Bosch, S. E. and Griesel, M. 2017. Strategies for building wordnets for under-resourced languages: The case of African languages. *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 38 (1): 1-12.

Calvo, H., Rocha-Ramirez, A. P., Moreno-Armendáriz, M. A. and Duchanoy, C. A. 2019. Toward universal word sense disambiguation using deep neural networks. *IEEE Access*, 7: 60264-60275.

Campolungo, N., Martelli, F., Saina, F. and Navigli, R. 2022a. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In: Proceedings

of *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4331-4352.

Campolungo, N., Pasini, T., Emelin, D. and Navigli, R. 2022b. Reducing Disambiguation Biases in NMT by Leveraging Explicit Word Sense Information. In: Proceedings of *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4824-4838.

Carroll, J. and McCarthy, D. 2000. Word sense disambiguation using automatically acquired verbal preferences. *Computers and the Humanities*, 34: 109-114.

Chakrawarti, R. K., Bansal, J. and Bansal, P. 2022. Machine translation model for effective translation of Hindi poetries into English. *Journal of Experimental & Theoretical Artificial Intelligence*, 34 (1): 95-109.

Chauhan, S., Daniel, P., Saxena, S. and Sharma, A. 2022. Fully unsupervised machine translation using context-aware word translation and denoising autoencoder. *Applied Artificial Intelligence*, 36 (1): 2031817.

Chauhan, S., Saxena, S. and Daniel, P. 2022a. Enhanced unsupervised neural machine translation by cross lingual sense embedding and filtered back-translation for morphological and endangered Indic languages. *Journal of Experimental & Theoretical Artificial Intelligence*, Article ID: 1-14.

Chauhan, S., Saxena, S. and Daniel, P. 2022b. Improved unsupervised neural machine translation with semantically weighted back translation for morphologically rich and low resource languages. *Neural Processing Letters*, 54 (3): 1707-1726.

Chen, X., Jia, S. and Xiang, Y. 2020. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141: 112948.

Cheng, J., Tong, W. and Yan, W. 2021. Capsule Network Improved Multi-Head Attention for Word Sense Disambiguation. *Applied Sciences*, 11 (6): 2488.

Chinchilla-Rodríguez, Z., Miao, L., Murray, D., Robinson-García, N., Costas, R. and Sugimoto, C. R. 2018. A global comparison of scientific mobility and collaboration according to national scientific capacities. *Frontiers in research metrics and analytics*, 3: 17.

Chingamtotattil, R. and Gopikakumari, R. 2022. Neural machine translation for Sanskrit to Malayalam using morphology and evolutionary word sense disambiguation. *Indonesian Journal of Electrical Engineering and Computer Science*, 28 (3): 1709-1719.

Chklovski, T. and Mihalcea, R. 2002. Building a sense tagged corpus with open mind word expert. In: Proceedings of *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions*. 116-122.

Ciaccio, L. A., Kgolo, N. and Clahsen, H. 2020. Morphological decomposition in Bantu: A masked priming study on Setswana prefixation. *Language, Cognition and Neuroscience*, 35 (10): 1257-1271.

Cohen, F. S., Zhong, Z. and Li, C. 2022. Semantic graph for word disambiguation in machine translation. *Multimedia Tools and Applications*, 81 (30): 43485-43502.

Conia, S. and Navigli, R. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In: Proceedings of *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 3269-3275.

Cotton, S., Edmonds, P., Kilgarriff, A. and Palmer, M. 2001. SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems. *Toulouse, July*, Article ID.

Creissels, D. 1996. Conjunctive and disjunctive verb forms in Setswana. *South African Journal of African Languages*, 16 (4): 109-115.

Damerau, F. J. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7 (3): 171-176.

De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D. L. 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50 (1): 1-18.

Demuth, K. 1989. Discourse functions of independent pronouns in Setswana. *Current issues in African linguistics*, 6: 176-193.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, Article ID.

Deza, E., Deza, M. M., Deza, M. M. and Deza, E. 2009. *Encyclopedia of distances*. Springer.

Dibitso, M., Owolawi, P. A. and Ojo, S. O. 2022. An Hybrid Part of Speech Tagger for Setswana Language using a Voting Method. In: Proceedings of *International Conference on Intelligent and Innovative Computing Applications*. 245-253.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26 (3): 297-302.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N. and Lim, W. M. 2021. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133: 285-296.

Drus, Z. and Khalid, H. 2019. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161: 707-714.

Duarte, J. M., Sousa, S., Milios, E. and Berton, L. 2021. Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations. *Information Sciences*, 570: 278-297.

Duong, L. 2017. Natural language processing for resource-poor languages. *University of Melbourne*, Article ID.

Durgaprasad, P., Sunitha, K. and Padmajarani, B. 2022. Resolving Lexical Level Ambiguity: Word Sense Disambiguation for Telugu Language by Exploiting IndicBERT Embeddings. In: *Communication, Software and Networks: Proceedings of INDIA 2022*. Springer, 357-368.

Eckmann, M., Rocha, A. and Wainer, J. 2012. Relationship between high-quality journals and conferences in computer vision. *Scientometrics*, 90 (2): 617-630.

Edmonds, P. and Cotton, S. 2001. Senseval-2: overview. In: Proceedings of *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*. 1-5.

Edmonds, P. and Kilgarriff, A. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, 8 (4): 279-291.

Eluri, S. and Siddu, V. 2020. A Knowledge Based Word Sense Disambiguation in Telugu Language. *International Journal of Engineering and Advanced Technology (IJEAT) ISSN*, Article ID: 2249-8958.

Emelin, D., Titov, I. and Sennrich, R. 2020. Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. *arXiv preprint arXiv:2011.01846*, Article ID.

Fellbaum, C. 2010. WordNet. In: *Theory and applications of ontology: computer applications*. Springer, 231-243.

Ferrari, A., Lipari, G., Gnesi, S. and Spagnolo, G. O. 2014. Pragmatic ambiguity detection in natural language requirements. In: Proceedings of *2014 IEEE 1st International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*. IEEE, 1-8.

Francis, W. N. and Kucera, H. 1979. Brown corpus manual. *Letters to the Editor*, 5 (2): 7.

Garcia-Martinez, M., Aransa, W., Bougares, F. and Barrault, L. 2020. Addressing data sparsity for neural machine translation between morphologically rich languages. *Machine Translation*, 34: 1-20.

Garigliotti, D. 2019. Semi-supervised learning for word sense disambiguation. *arXiv preprint arXiv:1908.09641*, Article ID.

Giunchiglia, F., Batsuren, K. and Freihat, A. A. 2018. One world–seven thousand languages. In: Proceedings of *Proceedings 19th international conference on computational linguistics and intelligent text processing, CiCling2018*. 18-24.

Gouskova, M., Zsiga, E. and Boyer, O. T. 2011. Grounded constraints and the consonants of Setswana. *Lingua*, 121 (15): 2120-2152.

Gutiérrez, Y., Vázquez, S. and Montoyo, A. 2017. Spreading semantic information by word sense disambiguation. *Knowledge-Based Systems*, 132: 47-61.

Habash, N. 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In: Proceedings of *Proceedings of ACL-08: HLT, Short Papers*. 57-60.

Hari, A. and Kumar, P. 2023. WSD Based Ontology Learning from Unstructured Text Using Transformer. *Procedia Computer Science*, 218: 367-374.

Harris, Z. S. 1954. Distributional Structure. WORD 10, 2--3 (1954), 146--162. *Google Scholar Google Scholar Cross Ref Cross Ref*, Article ID.

Hassanabadi, S., Kříž, J., Chung, W., Lütfüoğlu, B., Maghsoodi, E. and Hassanabadi, H. 2021. Thermodynamics of the Schwarzschild and Reissner–Nordström black holes under higher-order generalized uncertainty principle. *The European Physical Journal Plus*, 136: 1-13.

Heo, Y., Kang, S. and Seo, J. 2020. Hybrid sense classification method for large-scale word sense disambiguation. *IEEE Access*, 8: 27247-27256.

Hindle, D. and Rooth, M. 1993. Structural ambiguity and lexical relations. *Computational linguistics*, 19 (1): 103-120.

Iacobacci, I. J., Pilehvar, M. T. and Navigli, R. 2016. Embeddings for word sense disambiguation: An evaluation study. In: Proceedings of *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016-Long Papers*. Association for Computational Linguistics (ACL), 897-907.

Ide, N. and Suderman, K. 2004. The American National Corpus first release. In: Proceedings of *LREC*.

Iqbal, S., Hassan, S.-U., Aljohani, N. R., Alelyani, S., Nawaz, R. and Bornmann, L. 2021. A decade of in-text citation analysis based on natural language processing and machine learning techniques: An overview of empirical studies. *Scientometrics*, 126 (8): 6551-6599.

Irving, R. W. and Fraser, C. B. 1992. Two algorithms for the longest common subsequence of three (or more) strings. In: Proceedings of *Combinatorial Pattern Matching: Third Annual Symposium Tucson, Arizona, USA, April 29–May 1, 1992 Proceedings 3*. Springer Berlin Heidelberg, 214-229.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11 (2): 37-50.

Jadiya, A., Dondemadahalli Manjunath, T. and Mohan, B. R. 2022. A Comparative Study of Deep Learning Models for Word-Sense Disambiguation. In: *Advanced Machine Intelligence and Signal Processing*. Springer, 245-257.

Jain, G. and Lobiyal, D. 2022. Word Sense Disambiguation using Cooperative Game Theory and Fuzzy Hindi WordNet based on ConceptNet. *Transactions on Asian and Low-Resource Language Information Processing*, 21 (4): 1-25.

Janz, A., Dziob, A., Oleksy, M. and Baran, J. 2022. A unified sense inventory for word sense disambiguation in polish. In: Proceedings of *International Conference on Computational Science*. Springer, 682-689.

Jia, L., Tang, J., Li, M., You, J., Ding, J. and Chen, Y. 2021. TWE-WSD: An effective topical word embedding based word sense disambiguation. *CAAI Transactions on Intelligence Technology*, 6 (1): 72-79.

Kaddoura, S. and D. Ahmed, R. 2022. A comprehensive review on Arabic word sense disambiguation for natural language processing applications. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 12 (4): e1447.

Kaplan, A. 1955. An experiment study of ambiguity and context. *Mechanical Translation*, 2: 39-46.

Karuppaiah, D. and Vincent, P. D. R. 2021. Word sense disambiguation in Tamil using Indo-WordNet and cross-language semantic similarity. *International Journal of Intelligent Enterprise*, 8 (1): 62-73.

Kazemi, P. and Karshenas, H. 2021. Fuzzy Word Sense Induction and Disambiguation. *IEEE Transactions on Fuzzy Systems*, 30 (9): 3918-3927.

Kilgarriff, A. 1997. What is word sense disambiguation good for? *arXiv preprint cmp-lg/9712008*, Article ID.

Kilgarriff, A. and Grefenstette, G. 2003. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29 (3): 333-347.

Kilgarriff, A. and Rosenzweig, J. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34: 15-48.

Kim, M. and Kwon, H.-C. 2021. Word Sense Disambiguation Using Prior Probability Estimation Based on the Korean WordNet. *Electronics*, 10 (23): 2938.

Kumar, S., Jat, S., Saxena, K. and Talukdar, P. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In: Proceedings of *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5670-5681.

Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K. 2015. From word embeddings to document distances. In: Proceedings of *International conference on machine learning*. PMLR, 957-966.

Kwon, S., Oh, D. and Ko, Y. 2021. Word sense disambiguation based on context selection using knowledge-based word similarity. *Information Processing & Management*, 58 (4): 102551.

Landauer, T. K. and Dumais, S. T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104 (2): 211.

Langemets, M., Loopmann, A. and Viks, Ü. 2010. Dictionary management system for bilingual dictionaries. *eLexicography in the 21st Century: New Challenges, New Applications*, Article ID: 425-429.

Lastrucci, R., Dzingirai, I., Rajab, J., Madodonga, A., Shingange, M., Njini, D. and Marivate, V. 2023. Preparing the Vuk'uzenzele and ZA-gov-multilingual South African multilingual corpora. *arXiv preprint arXiv:2303.03750*, Article ID.

Leech, G. N. 1992. 100 million words of English: the British National Corpus (BNC). *어학연구*, Article ID.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of *Proceedings of the 5th annual international conference on Systems documentation*. 24-26.

Letsholo, R. and Matlhaku, K. 2014. The syntax of the Setswana noun phrase. *Marang: Journal of Language and Literature*, 24: 22-42.

Letsholo-Tafila, R. 2018. A characterisation of Setswana complex sentences. *Language Matters*, 49 (2): 80-106.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In: Proceedings of *Soviet physics doklady*. Soviet Union, 707-710.

Li, W. and Suzuki, E. 2021. Adaptive and hybrid context-aware fine-grained word sense disambiguation in topic modeling based document representation. *Information Processing & Management*, 58 (4): 102592.

Libovický, J., Rosa, R. and Fraser, A. 2019. How language-neutral is multilingual BERT? *arXiv preprint arXiv:1911.03310*, Article ID.

Lin, D. 1998. An information-theoretic definition of similarity. In: Proceedings of *Icml*. 296-304.

Lin, J. 2022. A proposed conceptual framework for a representational approach to information retrieval. In: Proceedings of *ACM SIGIR Forum*. ACM New York, NY, USA, 1-29.

Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M. and Hajishirzi, H. 2019. A general framework for information extraction using dynamic span graphs. *arXiv preprint arXiv:1904.03296*, Article ID.

Machinery, C. 1950. Computing machinery and intelligence-AM Turing. *Mind*, 59 (236): 433.

Mallery, J. 1988. *Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers. Master's thesis*: MIT Political Science Department, Cambridge, MA.

Mardones-Segovia, C., Choi, H.-J., Hong, M., Wheeler, J. M. and Cohen, A. S. 2021. Comparison of estimation algorithms for latent dirichlet allocation. In: Proceedings of *The Annual Meeting of the Psychometric Society*. Springer, 27-37.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R. and Modupe, A. 2020. Low resource language dataset creation, curation and classification: Setswana and Sepedi. *arXiv preprint arXiv:2004.13842*, Article ID.

Martinus, L. and Abbott, J. Z. 2019. Benchmarking Neural Machine Translation for Southern African Languages. *arXiv preprint arXiv:1906.10511*, Article ID.

Mathur, M. B. and VanderWeele, T. J. 2021. Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. *Research Synthesis Methods*, 12 (2): 176-191.

Maurya, A. S., Bahadur, P. and Garg, S. 2022. Approach Toward Word Sense Disambiguation for the English-To-Sanskrit Language Using Naïve Bayesian Classification. In: Proceedings

of *Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022*. Springer, 477-491.

McCrae, J. P., Rudnicka, E. and Bond, F. 2020. English WordNet: A new open-source wordnet for English. *K Lexical News*, 28: 37-44.

Melamud, O., Goldberger, J. and Dagan, I. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In: Proceedings of *Proceedings of the 20th SIGNLL conference on computational natural language learning*. 51-61.

Mennes, J. and van Gulik, S. v. d. W. 2020. A critical analysis and explication of word sense disambiguation as approached by natural language processing. *Lingua*, 243: 102896.

Mihalcea, R., Chklovski, T. and Kilgarriff, A. 2004. The Senseval-3 English lexical sample task. In: Proceedings of *Proceedings of SENSEVAL-3, the third international workshop on the evaluation of systems for the semantic analysis of text*. 25-28.

Miller, G. A. and Fellbaum, C. 2007. WordNet then and now. *Language Resources and Evaluation*, 41: 209-214.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3 (4): 235-244.

Miller, G. A., Leacock, C., Tengi, R. and Bunker, R. T. 1993. A semantic concordance. In: Proceedings of *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Mir, T. A., Lawaye, A. A., Rana, P. and Ahmed, G. 2023. Comparative Analysis of Decision Tree and k-NN to Solve WSD Problem in Kashmiri. In: Proceedings of *International Conference On Innovative Computing And Communication*. Springer, 243-254.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. and PRISMA Group, 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, 151 (4): 264-269.

Moors, C., Wilken, I., Calteaux, K. and Gumede, T. 2018. Human language technology audit 2018: Analysing the development trends in resource availability in all South African languages. In: Proceedings of *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*. 296-304.

Moro, A. and Navigli, R. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In: Proceedings of *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. 288-297.

Navigli, R. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41 (2): 1-69.

Navigli, R. and Ponzetto, S. P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193: 217-250.

Navigli, R. and Velardi, P. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, 27 (7): 1075-1086.

Navigli, R., Jurgens, D. and Vannella, D. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In: Proceedings of *Second Joint Conference on Lexical and Computational Semantics (SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 222-231.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S. H., Kabongo, S. and Osei, S. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*, Article ID.

Norouzi, M., Fleet, D. J. and Salakhutdinov, R. R. 2012. Hamming distance metric learning. *Advances in neural information processing systems*, 25.

Okgetheng, B., Malema, G. and Tebalo, B. 2022. TAGGING: Setswana Complex Qualificatives & Adverb. In: Proceedings of *2022 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE)*. IEEE, 13-18.

Olugbara, C. T., Letseka, M., Ogunsakin, R. E. and Olugbara, O. O. 2021. Meta-analysis of factors influencing student acceptance of massive open online courses for open distance learning. *The African Journal of Information Systems*, 13 (3): 5.

Orkphol, K. and Yang, W. 2019. Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. *Future Internet*, 11 (5): 114.

Otlogetswe, T. J. 2008. Corpus design for Setswana lexicography. Article IDUniversity of Pretoria.

Pal Singh, V. and Kumar, P. 2018. Naive Bayes classifier for word sense disambiguation of Punjabi language. *Malaysian Journal of Computer Science*, 31 (3): 188-199.

Pal, A. R. and Saha, D. 2019. Word sense disambiguation in Bengali language using unsupervised methodology with modifications. *Sādhanā*, 44: 1-13.

Pal, A. R., Maiti, P. K. and Saha, D. 2013. An approach to automatic text summarization using simplified lesk algorithm and wordnet. *International Journal of Control Theory and Computer Modeling*, 3 (4): 15-23.

Pal, A. R., Saha, D., Dash, N. S., Naskar, S. K. and Pal, A. 2019. A novel approach to word sense disambiguation in Bengali language using supervised methodology. *Sādhanā*, 44: 1-12.

Pal, A. R., Saha, D., Naskar, S. K. and Dash, N. S. 2021. In search of a suitable method for disambiguation of word senses in Bengali. *International Journal of Speech Technology*, 24: 439-454.

Pandit, R., Sengupta, S., Naskar, S. K. and Sardar, M. M. 2018. Improving Lesk by incorporating priority for word sense disambiguation. In: Proceedings of *2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT)*. IEEE, 1-4.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311-318.

Papini, M., Tirinzoni, A., Restelli, M., Lazaric, A. and Pirotta, M. 2021. Leveraging good representations in linear contextual bandits. In: Proceedings of *International Conference on Machine Learning*. PMLR, 8371-8380.

Parameswarappa, S. and Narayana, V. 2011. Target word sense disambiguation system for Kannada language. In: Proceedings of *3rd International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2011)*. IET, 269-273.

Park, J. Y., Shin, H. J. and Lee, J. S. 2022. Word Sense Disambiguation Using Clustered Sense Labels. *Applied Sciences*, 12 (4): 1857.

Pasini, T. 2021. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In: Proceedings of *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 4936-4942.

Pasini, T., Scozzafava, F. and Scarlini, B. 2020. CluBERT: A cluster-based approach for learning sense distributions in multiple languages. In: Proceedings of *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4008-4018.

Patel, N., Hale, J., Jindal, K., Sharma, A. and Yu, Y. 2021. Building on Huang et al. GlossBERT for Word Sense Disambiguation. *arXiv preprint arXiv:2112.07089*.

Patil, A. P., Ramteke, R., Bhavsar, R. and Darbari, H. 2021. Marathi Language Word Sense Disambiguation using Modifies Lesk Algorithm.

Paul, D. B. and Baker, J. 1992. The design for the Wall Street Journal-based CSR corpus. In: Proceedings of *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12: 2825-2830.

Peters, M. E., Neumann, M., Logan IV, R. L., Schwartz, R., Joshi, V., Singh, S. and Smith, N. A. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.

Popov, A., Koprinkova-Hristova, P., Simov, K. and Osenova, P. 2019. Echo state vs. lstm networks for word sense disambiguation. In: Proceedings of *International Conference on Artificial Neural Networks*. Springer, 94-109.

Postma, M., Ilievski, F., Vossen, P. and Van Erp, M. 2016. Moving away from semantic overfitting in disambiguation datasets. In: Proceedings of *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*. 17-21.

Powers, D. M. and Turk, C. C. 2012. *Machine learning of natural language*. Springer Science & Business Media.

Pradhan, S., Loper, E., Dligach, D. and Palmer, M. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In: Proceedings of *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*. 87-92.

Premjith, B., Soman, K., Anand Kumar, M. and Jyothi Ratnam, D. 2019. Embedding linguistic features in word embedding for preposition sense disambiguation in English—Malayalam machine translation context. *Recent advances in computational intelligence*, Article ID: 341-370.

Pretorius, L., Viljoen, B., Pretorius, R. and Berg, A. 2008. Towards a computational morphological analysis of Setswana compounds. *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 29 (1): 1-20.

Pretorius, L., Viljoen, B., Pretorius, R. and Berg, A. 2010. A finite state approach to Setswana verb morphology. In: Proceedings of *Finite-State Methods and Natural Language*

Processing: 8th International Workshop, FSMNLP 2009, Pretoria, South Africa, July 21-24, 2009, Revised Selected Papers 8. Springer, 131-138.

Pretorius, R. and Berg, A. 2005. The morphological analysis of Setswana nouns. *Journal for Language Teaching= Ijenali Yekufundzisa Lulwimi= Tydskrif vir Taalonderrig*, 39 (2): 274-291.

Pretorius, R., Berg, A., Pretorius, L. and Viljoen, B. 2009. Setswana tokenisation and computational verb morphology: Facing the challenge of a disjunctive orthography. In: Proceedings of. Association for Computational Linguistics.

Purohit, A. and Yogi, K. K. 2022. A Comparative Study of Existing Knowledge Based Techniques for Word Sense Disambiguation. In: Proceedings of *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2021*. Springer, 167-182.

Radhakrishnan, S., Erbis, S., Isaacs, J. A. and Kamarthi, S. 2017. Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. *PloS one*, 12 (3).

Raganato, A., Camacho-Collados, J. and Navigli, R. 2017. Word sense disambiguation: a uinified evaluation framework and empirical comparison. In: Proceedings of *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 99-110.

Rahman, N. and Borah, B. 2022. An unsupervised method for word sense disambiguation. *Journal of King Saud University-Computer and Information Sciences*, 34 (9): 6643-6651.

Rahman, N. and Borah, B. 2023. Query-Based Extractive Text Summarization Using Sense-Oriented Semantic Relatedness Measure. *Arabian Journal for Science and Engineering*, Article ID: 1-42.

Rais-Ghasem, M. and Corriveau, J.-P. 2020. Towards Exemplar-based Polysemy. In: Proceedings of *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*. Psychology Press, 566-571.

Rodd, J. 2018. Lexical ambiguity. *Oxford handbook of psycholinguistics*, Article ID: 120-144.

Rouhizadeh, H., Shamsfard, M. and Rouhizadeh, M. 2020. Knowledge based word sense disambiguation with distributional semantic expansion for the Persian language. In: Proceedings of *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, 329-335.

Ruas, T., Ferreira, C. H. P., Grosky, W., de França, F. O. and de Medeiros, D. M. R. 2020. Enhanced word embeddings using multi-semantic representation through lexical chains. *Information Sciences*, 532: 16-32.

Ruas, T., Grosky, W. and Aizawa, A. 2019. Multi-sense embeddings through a word sense disambiguation process. *Expert Systems with Applications*, 136: 288-303.

Sabour, S., Frosst, N. and Hinton, G. E. 2017. Dynamic routing between capsules. *Advances in neural information processing systems*, 30.

Saeed, A., Nawab, R. M. A. and Stevenson, M. 2021. Investigating the Feasibility of Deep Learning Methods for Urdu Word Sense Disambiguation. *Transactions on Asian and Low-Resource Language Information Processing*, 21 (2): 1-16.

Saeed, A., Nawab, R. M. A., Stevenson, M. and Rayson, P. 2019a. A sense annotated corpus for all-words Urdu word sense disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18 (4): 1-14.

Saeed, A., Nawab, R. M. A., Stevenson, M. and Rayson, P. 2019b. A word sense disambiguation corpus for Urdu. *Language Resources and Evaluation*, 53: 397-418.

Sahami, M. and Heilman, T. D. 2006. A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of *Proceedings of the 15th international conference on World Wide Web*. 377-386.

Sahlgren, M. 2008. The distributional hypothesis. *Italian Journal of Disability Studies*, 20: 33-53.

Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24 (5): 513-523.

Sari, S., Manurung, R. and Adriani, M. 2010. Indonesian WordNet Sense Disambiguation using Cosine Similarity and Singular Value Decomposition. *ICSIIT 2010*, Article ID: 234.

Sarika and Sharma, D. K. 2016. Hindi word sense disambiguation using cosine similarity. In: Proceedings of *Proceedings of International Conference on ICT for Sustainable Development: ICT4SD 2015 Volume 2*. Springer, 801-808.

Scarlini, B., Pasini, T. and Navigli, R. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In: Proceedings of *Proceedings of the AAAI conference on artificial intelligence*. 8758-8765.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P. and Stewart, L. A. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *Bmj*, 349.

Sharma, P. and Joshi, N. 2019. Knowledge-Based Method for Word Sense Disambiguation by Using Hindi WordNet. *Engineering, Technology & Applied Science Research*, 9 (2).
Shiwen, Y. and Xiaojing, B. 2014. Rule-based machine translation. In: *Routledge Encyclopedia of Translation Technology*. Routledge, 224-238.

Singh, J. and Singh, I. 2015. Word sense disambiguation: enhanced Lesk approach in Punjabi language. *International Journal of Computer Applications*, 129 (6): 23-27.
Soares, M. A. C. and Parreiras, F. S. 2020. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, 32 (6): 635-646.
Song, Y., Ong, X. C., Ng, H. T. and Lin, Q. 2021. Improved word sense disambiguation with enhanced sense representations. In: Proceedings of *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4311-4320.
Su, Y., Zhang, H., Song, Y. and Zhang, T. 2022a. Multilingual Word Sense Disambiguation with Unified Sense Representation. *arXiv preprint arXiv:2210.07447*.
Su, Y., Zhang, H., Song, Y. and Zhang, T. 2022b. Rare and Zero-shot Word Sense Disambiguation using Z-Reweighting. In: Proceedings of *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4713-4723.
Sunilkumar, P. and Shaji, A. P. 2019. A survey on semantic similarity. In: Proceedings of *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*. IEEE, 1-8.
Tang, G., Sennrich, R. and Nivre, J. 2019. Encoders help you disambiguate word senses in neural machine translation. *arXiv preprint arXiv:1908.11771*, Article ID.
Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R. and Das, D. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, Article ID.
Vallender, S. 1974. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18 (4): 784-786.
Van Eck, N. J. and Waltman, L. 2014. Visualizing bibliometric networks. *Measuring scholarly impact: Methods and practice*, Article ID: 285-320.

Van Rooy, B. and Pretorius, R. 2003. A word-class tagset for Setswana. *Southern African linguistics and applied language studies*, 21 (4): 203-222.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vossen, P. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10: 978-994.

Vu, V.-H., Nguyen, Q.-P., Shin, J.-C. and Ock, C.-Y. 2020. UPC: An Open Word-Sense Annotated Parallel Corpora for Machine Translation Study. *Applied Sciences*, 10 (11): 3904.

Wang, H., Wu, H., He, Z., Huang, L. and Church, K. W. 2022. Progress in machine translation. *Engineering*, 18: 143-153.

Wang, J. and Dong, Y. 2020. Measurement of text similarity: a survey. *Information*, 11 (9): 421.

Wang, S. I. and Manning, C. D. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 90-94.

Wang, Y., Wang, M. and Fujita, H. 2020. Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge-Based Systems*, 190: 105030.

Weaver, W. 1952. Translation. In: Proceedings of *Proceedings of the Conference on Mechanical Translation*.

Weng, L. 2019. From gan to wgan. *arXiv preprint arXiv:1904.08994*.

Wiemerslage, A., Silfverberg, M., Yang, C., McCarthy, A. D., Nicolai, G., Colunga, E. and Kann, K. 2022. Morphological Processing of Low-Resource Languages: Where We Are and What's Next. *arXiv preprint arXiv:2203.08909*.

Wilken, I., Griesel, M. and McKellar, C. 2012. Developing and improving a statistical machine translation system for English to Setswana: a linguistically-motivated approach. In: Proceedings of *Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa*. 114.

Wu, L., Yen, I. E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.-Y., Ravikumar, P. and Witbrock, M. J. 2018. Word mover's embedding: From word2vec to document embedding. *arXiv preprint arXiv:1811.01713*.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32 (1): 4-24.

Yadav, R. K., Jiao, L., Granmo, O.-C. and Goodwin, M. 2021. Interpretability in Word Sense Disambiguation using Tsetlin Machine. In: Proceedings of *ICAART (2)*. 402-409.

Yatabe, R. and Sasaki, M. 2020. Semi-supervised word sense disambiguation using example similarity graph. In: Proceedings of *Proceedings of the graph-based methods for natural language processing (TextGraphs)*. 51-59.

Young, T., Hazarika, D., Poria, S. and Cambria, E. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13 (3): 55-75.

Zhang, C.-X., Liu, R., Gao, X.-Y. and Yu, B. 2021. Graph Convolutional Network for Word Sense Disambiguation. *Discrete Dynamics in Nature and Society*, 2021: 1-12.

Zhang, C.-X., Pang, S.-Y., Gao, X.-Y., Lu, J.-Q. and Yu, B. 2022a. Attention Neural Network for Biomedical Word Sense Disambiguation. *Discrete Dynamics in Nature and Society*, 2022: 1-14.

Zhang, G., Lu, W., Peng, X., Wang, S., Kan, B. and Yu, R. 2022b. Word Sense Disambiguation with Knowledge-Enhanced and Local Self-Attention-based Extractive Sense Comprehension. In: Proceedings of *Proceedings of the 29th International Conference on Computational Linguistics*. 4061-4070.

Zhang, L. and Zhao, X. 2020. An overview of cross-language information retrieval. In: Proceedings of *Artificial Intelligence and Security: 6th International Conference, ICAIS 2020, Hohhot, China, July 17–20, 2020, Proceedings, Part I 6*. Springer, 26-37.

Zhang, X., Hauer, B. and Kondrak, G. 2022. Improving HowNet-Based Chinese Word Sense Disambiguation with Translations. In: Proceedings of *Findings of the Association for Computational Linguistics: EMNLP 2022*. 4530-4536.

Zhong, L. and Wang, T. 2020. Towards word sense disambiguation using multiple kernel support vector machine. *International Journal of Innovative Computing, Information and Control*, 16 (2): 555-570.

Zhu, G. and Iglesias, C. A. 2016. Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29 (1): 72-85.

Zouaghi, A., Merhbene, L. and Zrigui, M. 2011. Word Sense disambiguation for Arabic language using the variants of the Lesk algorithm. *WORLDCOMP*, 11: 561-567.

Appendix

Table 6.1 : WSD Reported Results

The outcomes of the WSD model experiments are delineated in Table 6.1, detailing precision (P), recall (R), and F1 score (F). The initial entry in the table, highlighted in red, represents the results of the proposed WSD model for Setswana; and the conducted research per language is highlighted in varying colours.

Language	Author	Year	Approach	Method/System	Dataset	P	R	F
Setswana	**Moape, T., et al.**	**2024**	**Knowledge-Based**	**PuoBERTa Enabled Embedding-based Lesk WSD Model**	**SUKC**	**0.89**	**0.88**	**0.89**
Arabic	Abderrahim, M. A. and M. E.A. Abderrahim	2022	Knowledge-Based	PageRank and conceptual density	Arabic WordNet	0.69	0.45	0.51
	Al-Hajj, M. and M. Jarrar	2021	Supervised	Bert	Arabic context-gloss pairs	0.85	0.71	0.76
	El-Razzaz, M., et al.	2021	Supervised	Bert	Arabic WordNet	0.92	0.87	0.89
	Laroussi, M., et al.	2013	Unsupervised	Vote procedure	Arabic corpus	0.79	0.68	0.73
Assamese	Gogoi, A., et al.	2021	Unsupervised	Cuckoo search algorithm	Assamese corpus	0.87	0.84	0.86

	Sarmah, J. and S. K. Sarma	2016	Supervised	Naive Bayes classifier	Assamese corpus	0.81	0.74	0.78
Australian	Navigli, R.	2009	Supervised	CQC algorithm	Macquarie Concise Dictionary, WordNet	0.64	0.64	0.64
Bengali	Khapra, M. M., et al.	2009	Knowledge-based	PageRank algorithm	Bengali datasets	0.80	0.80	0.80
English	Abed, S. A., et al.	2015	Knowledge-based	Harmony search algorithm	SemCor	0.67	0.64	0.65
	Abed, S. A., et al.	2016	Knowledge-based	Evolutionary method	SemCor	0.67	0.64	0.65
	Abualhaija, S. and K.H. Zimmermann	2016	Knowledge-based	D-Bees algorithm	Semeval-07	0.79	0.79	0.79
	Alsaeedan, W. and M. E. B. Menai	2015	Knowledge-Based	Self-adaptive GA	Senseval-03	0.44	0.49	0.46
	Alsaeedan, W., et al.	2017	Unsupervised	Self-adaptive genetic algorithm, Ant colony optimization	S2FGAW, S3FGAW, S07FGAW, and S07CGAW	0.83	0.83	0.83

	Al-Saiagh, W., et al.	2018	Knowledge-based	Hybrid swarm intelligence	Senseval-03	0.65	0.57	0.61
	ALTINEL, A. B., et al.	2019	Knowledge-based	Semantic kernels	Senseval-03	0.73	0.65	0.68
	Calvo, H., et al.	2019	Supervised	Neural networks	Senseval-03	0.63	0.63	0.63
	Camacho-Collados, J., et al.	2019	Unsupervised	Distributional semantic similarity	BabelNet, WordNet, Wikipedia, Wikidata, Nasari	0.83	0.76	0.79
	Chen, P., et al.	2009	Knowledge-based	TreeMatch automatic lexical dependency	Semeval-07	0.74	0.74	0.74
	Chen, P., et al.	2012	Semi-supervised	Dependency parser	Senseval-03, Semeval-07	0.83	0.83	0.83
	Corro, L. D., et al.	2014	Knowledge-based	Syntactic and semantic pruning	Semeval-07	0.74	0.52	0.60
	Eniafe, F. A. and K. Agbele	2018	Knowledge-based	Optimized Lesk-based	Semeval-13	0.66	0.65	0.60
	Goularte, F. B., et al.	2020	Knowledge-based	Co-occurrences and morphosyntactic	MSC patterns tweets	0.80	0.24	0.37
	Gutiérrez, Y., et al.	2017	Unsupervised	Personalized PageRank algorithm	Semeval-13	0.65	0.65	0.65

	Hausman, M.	2011	Knowledge-based	Genetic algorithm GA	Senseval-03	0.52	0.54	0.53
	Hey, T., et al.	2021	Knowledge-based	Page rank	Wikipedia-based Variants, WordNet	0.51	0.59	0.55
	Hwang, M., et al.	2010	Knowledge-based	Semantic network	Senseval-3	0.75	0.79	0.77
	Kwon, S., et al.	2021	Knowledge-based	Graphical semantic word similarity measure	Senseval-2, Senseval-3, SemEval-07, SemEval-13, SemEval-15	0.73	0.69	0.71
	Lau, J. H., et al.	2014	Knowledge-Based	Topic models	UKWAC, Twitter	0.76	0.89	0.77
	Li, W. and E. Suzuki	2021	Semi-supervised	Hybrid context-based topic model	20Newsgroups	0.87	0.87	0.87
	Manion, S. L. and R. Sainudiin	2014	Knowledge-based	Iterative subgraph	BabelNet	0.65	0.59	0.62
	Miller, T. and I. Gurevych	2015	Knowledge-based	DKPro WSD framework	Day website humorists	0.21	0.13	0.16
	Navigli, R.	2006	Unsupervised	SSI algorithm	Senseval-03	0.60	0.60	0.60

	Navigli, R. and P. Velardi	2005	Supervised	SSI Algorithm+baseline	Senseval-03	0.69	0.68	0.67
	Nazreena, R. and B. Borah	2021	Unsupervised	Sense-oriented sentence semantic relatedness	MPSC corpus	0.76	0.90	0.82
	Nguyen, K.H. and C.Y. Ock	2013	Knowledge-based	Ant colony optimization	Senseval-03	0.57	0.57	0.57
	Nurifan, F., et al.	2018	Unsupervised	Word2vec	Wikipedia, Oxford English Dictionary	0.85	0.86	0.85
	O'Hara, T. and J. Wiebe	2009	Knowledge-based	Semantic role modelling	FrameNet	0.63	0.62	0.63
	Orkphol, K. and W. Yang	2019	Unsupervised	Word2vec	Senseval-03	0.51	0.49	0.50
	Oussalah, M., et al.	2018	Knowledge-based	CatVar Sem & Syntactic	Semcor	0.74	0.72	0.73
	Panchenko, A., et al.	2017	Unsupervised	Meta-combination dependency features	TWSI dataset	0.79	0.79	0.79
	Pasini, T. and R. Navigli	2017	Semi-supervised	Train-o-Matic, Denser semantic networks	Senseval,Senseval-01, Senseval-03, Semval-07,Semeval15	0.65	0.60	0.62

	Schwartz, H. A. and F. Gomez	2008	Knowledge-based	Token, type, and selector	Semeval-07	0.84	0.50	0.63
	Tonelli, S., et al.	2013	Unsupervised	Multilingual frame annotation	Wikipedia annotations	0.76	0.73	0.75
	Torres-Ramos, S., et al.	2016	Unsupervised	Alpha-Beta associative memory	Senseval-2	0.58	0.26	0.36
	Wang, X., et al.	2013	Knowledge-based	Topical and semantic association	Semcor	0.77	0.60	0.67
French	Elayeb, B., et al.	2015	Supervised	Possibilistic and probabilistic theory modelling	Romanseval	0.57	0.54	0.56
Hindi	Bhatia, S., et al.	2022	Knowledge-based	Generic algorithm	Hindi WordNet	0.82	0.89	0.90
	Kumari, A. and D. Lobiyal	2022	Unsupervised	Distributed words representation	Hindi Wikipedia	0.58	0.60	0.59
	Sunita, R.	2019	Supervised	Decision tree	Health database	0.70	0.84	0.79