The Development of lexicographical databases, tools, and resources for storing Multilingual Data in Support of the Abstract Wikipedia Project: A Literature Review

Tadiwa Magwenzi
The University of Cape Town Computer Science Department

Abstract

This literature review explores past and current approaches involved in creating a lexicographic database, tools, and resources. With an intent to use these methods to aid Abstract Wikipedia. An additional focus is on the creation of these tools for low-resourced languages. I will draw on a range of sources, including academic papers, technical reports, and online resources, to provide a comprehensive overview of the current state of lexicographic resource development and their applicability to Abstract Wikipedia. The review discusses the importance of including lexicographic data in Wikidata and examines the different approaches and technologies used in developing such a database. Building a high-quality lexicographic database is essential to support the development of multilingual, machine-generated articles in the context of the Abstract Wikipedia project. Overall, this literature review highlights the significance of building high-quality lexicographic databases and resources to support the development of multilingual, machine-generated articles in the context of the Abstract Wikipedia project. This review concludes with a comparison of the different characteristics of lexicographical resources. Key findings include the importance of including lexicographic data in Wikidata, the potential of using Wikibase to develop a reliable lexicographic database, and the challenges of data acquisition, validation, and management.

Keywords Abstract Wikipedia, lexicographic data, multilingual knowledge base, Wikidata, wikifunctions, Wikibase, database development, low-resourced languages, machine-generated articles, data acquisition, data validation, data management, data accessibility.

1 Introduction

Abstract Wikipedia is a new project by the Wikimedia Foundation with its focus on generating articles in any natural language. using Wikidata as an abstract data content source and Wikifunctions as an algorithm store. Developing high-quality lexicographic databases and relevant resources is necessary, and required to achieve this goal. The following literature review explores the challenges and approaches involved in building multilingual lexicographic tools and resources for Abstract Wikipedia. This review will have a particular focus on the needs of low-resourced languages. The review draws on a range of sources, including academic papers, technical reports, and online resources, to provide a comprehensive overview of the current state of the art in lexicographic database development for Abstract Wikipedia. It discusses the importance of including lexicographic data in Wikidata and examines the different approaches and technologies used in developing such a database. The findings suggest that using Wikibase, the same platform used for Wikidata, can be a promising approach to developing a robust and reliable lexicographic database. However, further research is needed to address the challenges of data acquisition, validation, and management, as well as to improve the interoperability and accessibility of the lexicographic database. Overall, this literature review highlights the significance of building high-quality lexicographic resources and tools to support the development of multilingual, machine-generated articles in the context of the Abstract Wikipedia project. In the remainder of this review, we will provide a background on the Abstract Wikipedia project, lexicography, and lexicographic databases. We then will explore the principles of lexicographical databases and the challenges associated with their development. We will also discuss the different approaches to developing multilingual lexicographic databases, such as the use of wordnets and other semantic resources(and their different forms). Finally, we will conclude by highlighting the importance of building high-quality lexicographic resources to support the development of multilingual, machine-generated articles in the context of the Abstract Wikipedia project. This will include a comparison of the different characteristics of lexicographical resources.

2 Background

First, it is important to provide a fundamental background on the Abstract Wikipedia project, its goals, and its structure. The project was proposed by Denny Vrandecic as a solution to the unequal distribution of knowledge across languages in Wikipedia. According to Vrandečić [1], "The knowledge in Wikipedia is very unevenly distributed over the languages: some languages have more than a million articles, but more than 50 languages have only a few hundred articles or less. More importantly, also the number of contributors is very unevenly distributed." He went on to say that "in order to close these knowledge gaps, we are building a multilingual Wikipedia where content is created only once but made available in all languages." The overall goal of the project is to

allow more people to read more content in the language of their choice. To create this content, the aim is to use natural language generation(NLG) and a Renderer function available in and stored in Wikifunctions to generate articles. The goal is to have abstract data in addition to language-specific lexicographic data stored in Wikidata(This will be in the form of Language specific lexemes). The abstract data will then be rendered into natural language using renderer functions stored in Wikifunctions. The process of natural language will also make use of the language-specific Lexicographic data that is also stored in Wikidata.

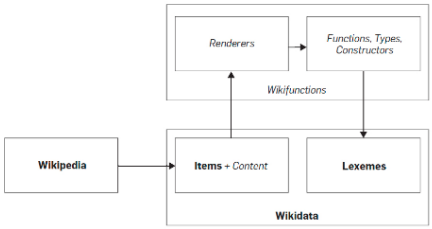

Figure 1. Initial idea of the overall architecture of the project and its relation to traditional Wikipedia[1]

It is important to provide a more in-depth background on the specific aim of this literature review: lexicography and lexicographic databases. Lexicography is the practice of compiling, writing, and editing dictionaries, glossaries, and other types of word lists. It involves the study of the structure, meaning, history, and usage of words, as well as the organization and presentation of this information in a way that is useful to users. According to Bergenholtz [2], "the simplest explanation of lexicography is that it is a scholarly discipline that involves compiling, writing, or editing dictionaries. Lexicography is widely considered an independent scholarly discipline, though it is a subfield within linguistics." He also broke down the field into two related disciplines [2]: practical lexicography, which is the art or craft of compiling, writing, and editing dictionaries, and theoretical lexicography, which is the scholarly discipline of analyzing and describing the semantic, syntagmatic, and paradigmatic relationships within the lexicon (vocabulary) of a language, developing theories of dictionary components and structures linking the data in dictionaries. Both disciplines will be explored in this literature review(although not explicitly) and the ways they can be used to develop lexicographic databases. With this understanding of lexicography, we can now provide background on lexicographic data, databases, and their importance to the Abstract Wikipedia project.

The term 'Lexicographic data' refers to the information used to describe and define words. It is often organized in dictionaries and other lexical resources. This includes information about the word's spelling, pronunciation, part of speech, definition, usage, and other relevant details to the word. Lexemes, on the other hand, are the basic units of meaning in a language, which can be represented by one or more words. Lexemes [4] are the abstract units of language that are associated with a set of inflected forms r word senses. Lexicographic data is used to describe and define lexemes. This functionality enables users to access and understand their meanings, usage, and relationships with other lexemes. To relate this to Abstract Wikipedia, Nielsen [3] notes that "since 2018, Wikidata has included special pages for lexicographic data distinguished from the usual Wikidata 'Q-items' with a new namespace for lexemes. Each page represents one lexeme, its sense(s), and its lexical form(s) together with annotation about them and links between them, both within and between lexemes as well as to the Q-items."

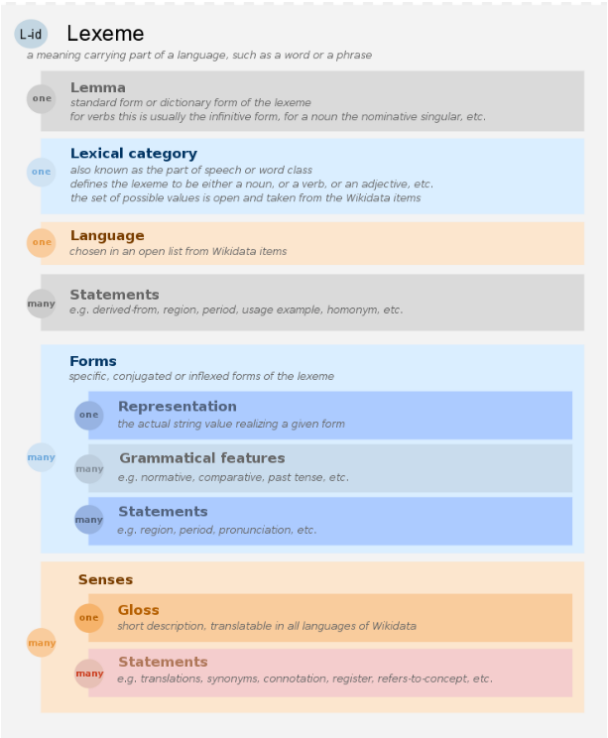

Figure 2. The structure of a Lexeme on Wikidata. [4]

According to the Wikidata documentation [4], a Lexeme is a lexical element of a language, such as a word, a phrase, or a prefix (see Lexeme on Wikipedia). Lexemes are entities in the sense of the Wikibase data model. All of this highlights the project's reliance on lexicographic data and begs the question of how this data would be stored. The solution is a lexicographic database. According to Bergenholtz and Nielsen [5], a lexicographical database is a database constructed to contain lexicographical data. Bergenholtz also illustrated the essential components of such a database, such as a lemma, sense, and grammatical information.

The goal of this literature review is to explore the challenges and approaches involved in developing such high-quality lexicographical databases and other resources for storing multilingual data to support the Abstract Wikipedia project.

3 Principles of Lexicographical Databases

According to Galieva et al. [15], lexicography is the branch of linguistics that deals with the theory and practice of compiling dictionaries. Theoretical lexicography develops typologies of dictionaries, which can vary in format, information, and intended use. Dictionaries serve different functions, such as recording factual information about the world (encyclopedic and defining dictionaries), organizing language content (thesauruses), and standardizing language to aid communication.

Galieva describes a lexicographical system as an information environment in which lexicographical models are implemented. Lexicographical systems and databases are information systems that require specialized software tools and linguistic formalization to effectively integrate database technology. However, linguists must still perform the main work of isolating and describing lexical units.

The first body of work we willl consider is by [5] Bergenholtz et al. In it, they explore the principles of lexicographical databases. They define a lexicographical database as a computerized collection of lexical information that follows established lexicographic principles. The authors propose that a lexicographical database should have four essential components: a headword or lemma, sense or meaning, grammatical information, and illustrative examples. Bergenholtz et al. [5] proposed four essential components of a lexicographical database: a headword or lemma, sense or meaning, grammatical information, and illustrative examples. Bergenholtz et al.also provides a plan and structure for a lexicographical database for Spanish monolingual dictionaries. They proposed 23 fields to be included in the database, such as lemma, sublemma, meaning, synonym, antonym, and internet link. The authors suggest three types of buttons for the user interface: buttons to aid lexicographers in their work, buttons for database operations, and navigation buttons for moving between different UI pages. The authors describe a data format in which each lemma is given a unique ID, and for every polyseme, the lemma ID is added to link it to the lemma to which it belongs. A foreign key in the polyseme table is used to relate the lemma and the polyseme, creating a one-to-many relationship. To link lemmata to each other, a link table is created with fields for the ID of the lemma we want to link from and the ID of the lemma we want to link to. This design provides a clear and structured way of organizing and accessing information about words and their meanings. The work presented in this paper has limitations, such as the omission of newer approaches to lexical data representation such as EKILEX [21] by the Institute of the Estonian Language . Additionally, the paper is written from the perspective of lexicography and may not be as relevant to other fields that use lexical data, such as natural language processing or computational linguistics.

Another recent study on lexicographical databases is Fuertes-Olivera's [6] "The Internet, Digital Initiatives and Lexicography". This study discussed how lexicographical databases have become essential tools in modern lexicography. The study also highlights the advantages of electronic databases over traditional print dictionaries. These advantages include the ability to store and retrieve large amounts of data, create dynamic links between lexical entries, and incorporate multimedia elements. Fuertes' study covered a large variety of of different types lexicographical databases, including monolingual, bilingual, and multilingual databases, along with their respective advantages and disadvantages. In addition, it provides examples of successful and practical lexicographical databases. It also discusses the challenges of creating and maintaining them. Their study also involved corpus-based lexicography. This involves digital databases based on large collections of language samples.Fuertes also states in [6] that, "Corpus-based lexicography is an important development in lexicography, allowing lexicographers to compile dictionaries that reflect the actual use of language". By analyzing a large corpus, lexicographers can identify patterns of language usage that might not be apparent from smaller data sets. Additionally, corpus-based lexicography can provide more up-to-date information than traditional dictionaries, as it is constantly updated as new language samples become available. However, the study did present some limitations to this approach of the lexicographic database, mainly that "corpus-based lexicography is only as good as the corpus it is based on". This highlights and reinforces the need for high-quality corpora of this type of database to be pursued.

In addition to the importance of multilingual databases, as discussed in Abstract Wikipedia, Feurtes [6] noted that "Multilingual databases are becoming increasingly important in a globalized world, where there is a growing need for translation and cross-linguistic communication". The architecture of such a database has the capability for easy cross-language comparisons. For this to be achieved, multilingual databases often use a central indexing system to link equivalent terms in different languages.

They explain that such Indexing systems must be designed to allow for easy access to equivalent terms in different languages. This enables users to search for a term in one language and find equivalent terms in other languages. The Fuertes states that [6] "The architecture of a multilingual database must take into account the unique features of each language, such as grammar, syntax, and vocabulary". This definitely would require a deep understanding of each language and its cultural context. This is because each and

every language has its own unique, language-specific features such as grammar, syntax, and vocabulary. With all this in mind, a multilingual database needs to be designed in a way that takes into account these linguistic and cultural differences to allow for easy access to equivalent terms in different languages. This would require an in-depth understanding of each language and its cultural context to ensure that the translations are accurate and consistent across different languages. In addition, some languages have complex grammatical structures or different systems for expressing tense or aspect. This makes it all the more important to consider the linguistic nuances when creating a multilingual lexicographical database.

Fuentes et all [6]also highlighted important nuances and factors to consider when creating a multilingual lexicographical database. These factors and nuances included translation quality, linguistic nuances, and cultural nuances. Particularly linguistic nuances were stated to be very important to remember and take note of because some languages have complex grammatical structures or different systems for expressing tense or aspect. These differences can make it challenging to develop accurate and consistent translations across different languages.

4 WordNets

One popular form of lexicographical resource are wordnets. A Wordnet is a type of lexical database that organizes words and their meanings based on their semantic relationships. According to McCrae et al. (2020) [18], wordnets have turned out to be one of the most popular types of dictionaries used in natural language processing (NLP) and other areas of language technologies. This can be ascribed mainly to their structure as a graph of words, which is much easier for computers to understand than the traditional form of a dictionary. The Princeton Word Net, [20] served as a template for many similar projects, where the hierarchical structure and semantic relations are kept largely unchanged, and only the content of each synset, i.e. the lemmas, usage examples, and definitions, is translated into the target language.

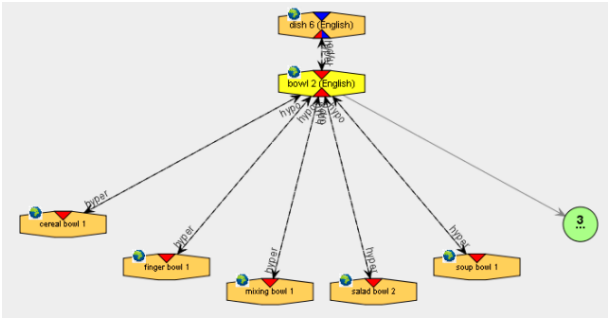

Figure 3. Representation of the noun synset for the word "bowl" in an English WordNet. [18]

4.1 EuroWordNet

This particular paper by Vossen[16] "EuroWordNet: General Documentation" is highly influential in the field of computational linguistics. It describes the creation of a multilingual lexical database for several European languages, including English, Dutch, Spanish, Italian, French, German, and Czech. The database, called EuroWordNet (EWN), serves as a common framework for natural language processing and knowledge representation. It is based on the theory of lexical semantics, which emphasizes the importance of understanding the meaning of individual words and their relationships with other words in a language. The database includes information about word senses, synonyms, antonyms, hyponyms, hypernyms, meronyms, holonyms, and other semantic relations between words. According to Vossen [16], "EuroWordNet: General Documentation 5 1. Introduction EuroWordNet is a multilingual lexical database with wordnets for several European languages, which are structured along the same lines as the Princeton WordNet". It is organized around the notion of a synset, which is a set of words with the same part of speech that can be interchanged in a certain context.

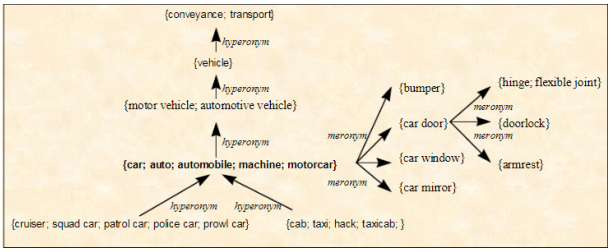

Figure 4. The synsets related to the word 'car'. [16]

The main idea is to have language-specific modules that allow for the independent development of resources while maintaining language-specific structures. To achieve this, the developers made a distinction between language-specific modules and a separate language-independent module. The language modules represent distinct, unique, and autonomous language-specific systems of language-internal relations(or links) between synsets. In sharp contrast, an inference-based ontology is used to represent language-independent relations. The developers also made sure to integrate existing lexical databases and semantic resources, such as WordNet, into EuroWordNet. This allows for a more comprehensive and robust database. The Inter-Lingual-Index (ILI) serves as a mediator between the synsets of the language-specific wordnets, and each synset in the monolingual wordnets has at least one equivalence relation with a record in this ILI. Language-specific synsets linked to the same ILI-record should be equivalent across languages. To give a definition of the ILI, Peters described it as [10]it" an unstructured list

of meaning, and a reference to its source." While EuroWordNet aims to unify the vocabulary of several European languages, it has some limitations. These include limited coverage of only a few European languages (English, Dutch, Italian, and Spanish), lack of consistency in data quality and structure across the languages, incomplete lexical coverage of idiomatic expressions, technical terms, and slang, limited semantic information that may not support advanced computational tasks, and outdated information (last updated in 2006). Researchers should keep these limitations in mind when using EuroWordNet for cross-linguistic lexical analysis.

Overall, Vossen's work is significant because it presents a novel approach to building multilingual lexical databases using the WordNet model. EuroWordNet has since become an important resource for natural language processing research and has influenced the development of similar projects in other languages and regions.

4.2 the African Wordnet (AfWN)

Bosch et al. [17] presented a novel approach to expanding documentation and preservation of African Indigenous Knowledge (AIK) using a digital lexical database. The article states that "the collection of AIK in paper archives and more recently in digital databases is imperative in preserving not only the language but also traditional customs for posterity." Their solution involved the novel application and subsequent expansion of an existing lexical resource for isiZulu, the African Wordnet (AfWN). They demonstrated the conversion of AIK into semantic relations in a wordnet structure. The authors also focused on filling lexical gaps between isiZulu and English, as found in the Princeton WordNet, with culturally relevant synsets.

In the proposed lexical database, words are grouped into sets of synonyms called synsets. The EuroWordNet and BalkaNet projects created the so-called "core base concepts" (CBC) list — a list of seed terms extracted from corpora for various European languages involved in the two projects with which to kickstart wordnet development. However, since the CBC incorporates many concepts that are not lexicalized in African languages, the AfWN resorted to incorporating synsets from a more localized seed list — the SIL Comparative African Wordlist (SILCAWL).

The inclusion of seed terms from this list has already resulted in the inclusion of numerous lexicalized concepts such as the elaborate kinship terms in isiZulu and Sesotho.

The scarcity of resources has been a significant challenge in the development of African wordnets; thus, the expand approach (which involves using an existing wordnet, such as the PWN, as a template and translating its content into the target language) was used. Bosch et al. [17] also state that "African languages and cultures include many unique word senses that are not easily matched to the core set of meanings in the PWN or, for that matter, in other wordnets."

In the study, the authors took the first steps in creating an isiZulu lexical database that addresses lexicalization differences (instances where the source and target languages lexicalize the same concept with a different kind of lexical unit, be it a word, compound, or collocation OR instances where one of the two languages has no lexicalization for a concept at all and results in a lexical unit in either the source or target language being translated with a description of the concept as a phrase. In the latter case, we, therefore, have a so-called lexical gap). Bentivogli [19] defines a lexical gap by stating "a lexical gap occurs whenever a language expresses a concept with a lexical unit whereas the other language expresses the same concept with a free combination of words." To fill such gaps, the SIL Comparative African Wordlist (SILCAWL) is used as a benchmark.

The database makes use of a hierarchical classification of terms, an example of which is in Figure 9.

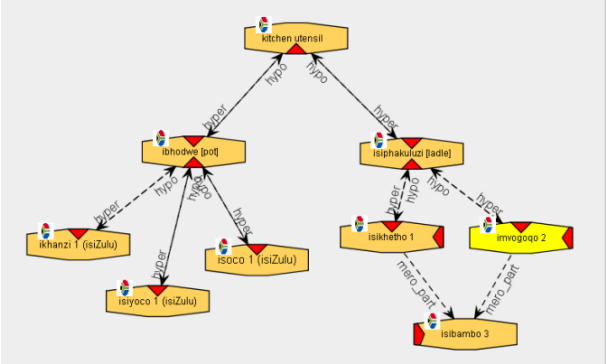

Figure 5. Hierarchical classification of terms related to the term "Kitchen Utensils" [17]

However, the study has a few limitations. Firstly, it only focuses on isiZulu and does not explore other Bantu languages, limiting its generalizability. Secondly, the study only includes a small sample of words, which may not be representative of the entire language, and lacks empirical evaluation of its effectiveness. Thirdly, the authors mainly rely on existing dictionaries and interdisciplinary sources to create the lexical resource, overlooking potential input sources such as native speakers or written texts. Lastly, the study primarily focuses on traditional knowledge, which may not capture the diversity of contemporary usage in isiZulu, thus limiting the resource's ability to fully capture the language as it is spoken and used today.

5 Multilingual Lexical Resources

5.1 BabelNet

BabelNet, is a multilingual resource that integrates information from a wide range of lexical and semantic resources, including WordNet. BabelNet extends the WordNet model by incorporating information from many other resources,

including Wikipedia, Wiktionary, OmegaWiki, and others. This allows BabelNet to provide a much broader and more comprehensive view of word meanings and relationships across multiple languages. Navigili [22] says that "BabelNet is a very large multilingual ontology and semantic network, obtained as a result of a novel integration and enrichment methodology". They explain that the resource is created by linking the largest multilingual Web encyclopedia, Wikipedia to the most popular computational lexicon – i.e., WordNet. TNavigli went on to explain that The integration is performed via automatic mapping and by filling in lexical gaps in resource-poor languages with the aid of Machine Translation (MT). The result of all this is an “encyclopedic dictionary” that provides babel synsets, i.e., concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations. Each babel synset is linked to a unique concept or named entity in the BabelNet ontology. Babel synsets provide a way to represent and link concepts across different languages and enable cross-lingual natural language processing tasks.

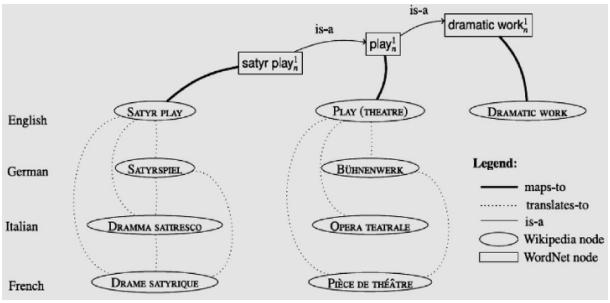

Figure 6. The multilingual structure of a babel synset linking data from Wikipedia and wordnet. [22]

BabelNet focuses on word senses and Named Entities in many languages and aims to provide full lexicographic and encyclopedic coverage. It can be viewed as a multilingual ontology, a large machine-readable encyclopedic dictionary, and a multilingual semantic network. In another paper,[23] , Navilig et all broke down and explained the data model of BabelNet. According to Navigli, BabelNet follows a Graph Model, where "synsets are nodes and edges are semantic relations between them. The relations in BabelNet stem from the underlying resources which provide them. " Navigli’s studies on BabelNet have contributed significantly to the field of Natural Language Processing (NLP). However, these studies are not without limitations and drawbacks. Navigli’s studies on BabelNet have made a significant contribution to the field of Natural Language Processing (NLP); however, they are not without limitations and drawbacks. One major limitation is their heavy reliance on the quality and completeness of the BabelNet knowledge base, which despite efforts to provide accurate information, still contains gaps and errors leading to incomplete or inaccurate results. Additionally, these studies tend to focus on English language data, making the results less generalizable to other languages and cultures. Navigli’s approach also has some technical limitations, such as the use of graph-based algorithms that may not be suitable for large datasets or complex networks. Furthermore, the subjective nature of language and culture can introduce bias into the studies, affecting the accuracy and generalizability of the results. Lastly, these studies may overlook the importance of context in language processing, leading to incomplete or inaccurate results. Despite these limitations, acknowledging them and working to improve upon the approach can advance the field of NLP and enhance our understanding of language and culture.

6 Multilingual Lexicographical Databases

6.1 Kamusi

The Kamusi project is an example of a lexicographical database that utilizes multi-word expressions (MWEs) is the Kamusi Project. This online dictionary website aims to document the lexicons of endangered and less-resourced languages (LRLs) using modern computational tools and techniques. According to Benjamin et al.[13] (2017), the project provides a unified platform and repository for linguistic data that is both easy to use and free to researchers and the public.

The Kamusi Project stores various types of lexicographical data, including definitions, translations, examples, synonyms, antonyms, and usage notes for individual lexemes across multiple languages. It also includes information about the morphology, phonetics, and syntax of the languages in its database. The project uses lexicographical databases to store and organize linguistic data contributed by individual researchers and crowdsourcing. This data is then utilized to produce bilingual and multilingual dictionaries between each language in the system, as well as bedrock linguistic data that can be used in advanced machine applications.

In addition to the above, Benjamin et al. (2017) notes that the Kamusi Project requires offline input systems, privacy systems, and gamification features to encourage data collection and validation by the crowd. These features aim to improve data collection and ensure privacy for sensitive information. Overall, the Kamusi Project is an important initiative that aims to preserve and document the lexicons of endangered and less-resourced languages using modern computational tools and techniques.

6.2 (ANNA) First amalgamated Lexicographical database for Afrikaans and Dutch

Prinsloo et all [24]. discussed the lemmatization and treatment of kinship terms in a proposed English-Sotho, Sotho-English dictionary with an amalgamated lemma list. According to the article, the first step in building such a lexical database is to create a list of kinship terminology for the Sotho languages, followed by determining the most frequently used

forms to be lemmatized in the dictionary due to space restrictions. An amalgamated lemma list is a specific approach to compiling a dictionary where instead of creating separate lists of lemmas (base words) for each language, a combined or unified list of lemmas is created. In this approach, the lemma list includes all the lemmas (base words) for each language, and any duplicates are removed, resulting in a single list of unique lemmas for all the languages in the dictionary [24]. However, the article notes that "the languages to be treated should be closely related, i.e., that they should have a substantial number of words in common," in order to create an amalgamated lemma list. Thus, the study focused on closely related African languages such as Afrikaans/Dutch, Sepedi/Setswana/Sesotho, isiZulu/isiXhosa/Siswati/isiNdebel. To ensure that the approach of an amalgamated lemmalist was correct, Prinsloo et al. compared the 10,000 most frequently used words in Sesotho, Setswana, and Sepedi corpora and found that the vocabulary of these languages overlaps to a large extent. This overlap means that using an amalgamated lemmalist of 22,537 words is more efficient than creating separate dictionaries for each language. The main benefit of this approach is its efficiency and the amount of space it conserves. However, Prinsloo et al. did put forward some limitations that the program has, mainly that ideally, a single term for "uncle" in all three of the Sotho languages would have resulted in additional space saving. However, "ramogolo," "rangwane," and "malome" have the same meanings respectively in all three Sotho languages but refer to different relations in terms of the age of the related person and his position in the family tree

6.3 Russian-Tatar Lexicographical Database

The Russian-Tatar lexicographical database is complex and detailed, with specific markup and models used to describe the internal relationships between the lexical systems of both languages. According to Gantar's paper, [15]It consists of interlinked components, one for each language, with its own internal structure and corresponding grammar and semantic models. Essentially, it is a massive dictionary that contains information about each language's lexicon.

The information about word forms is divided into two parts: the dictionary of base morphemes and the dictionary of inflectional suffixes. The base morphemes dictionary stores information about lemmas and morphological and morphonological types, while the inflectional suffixes dictionary contains possible chains of affixes linked with the base morphemes. These two dictionaries are linked and form the T-component of the database, which includes the T-Base table, the T-Okon table, and the M-Posled table.

Each lexical token in the database has its own set of morphological attributes, which makes it difficult to store all the information in a standard database. As a solution, characteristic vectors consisting of 0s and 1s are used to encode information, with each affixal chain having its own characteristic vector.

7 Lexicographic data storage in Wikidata

According to Vrandečić. [1] , Wikidata is the structured data sister of Wikipedia where users can collaboratively edit a knowledge graph. Wikidata has included special pages for lexicographic data distinguished from the usual Wikidata "Q-items" with a new namespace for lexemes. M. Morshed [25] shed further light on this and explained that "Each page represents one lexeme, its sense(s) and its lexical form(s) together with annotation about them and links between them, both within and between lexemes as well as to the Q-items." This lexicographic data is converted to a semantic Web representation and available in WDQS and for RDFication of the lexeme data, Wikidata uses a combination of classical Wikidata URIs and URIs from (Linguistic) Linked Open Data ontologies(which arere standardized vocabularies or conceptual frameworks that are used to describe and organize linguistic data in a machine-readable way.)

Currently Wikidata stores lexemes from over 668 languages according to Cimiano, [14].

Furthermore According to the Nielsen [3], Russian has the most lexemes (101,137) and forms (1,236,456) of any language in Wikidata, followed by English, Hebrew, Swedish, Basque, French, and Danish. Basque has the most senses (20,272), followed by English, Hebrew, Russian, and Danish.

This highlights one of the main issues in Wikidata, a lack of coverage of low-resourced languages. This lack of representation greatly affects the applicability of Wikidata (and by extension Abstract Wikipedia) to a wider range of users and use cases.

Additionally, M.Morshed et all [25], proposed a method for representing the syntactic dependencies within Wikidata lexicographical data. Their idea was to use a compact format that is applicable to different types of dependency grammars but specifically demonstrated and exhibited with Universal Dependencies (UD). They argued that this method is effective and useful for modeling structures of multi-part elements that may be considered words in some languages. Also, they stated that this method is useful for modeling structures of multi-part elements that may be considered words in some languages. In addition, it will also allow for lexemes with syntax represented in this way to form parts of other lexemes which, as single units, take part in other dependency relations.

However, Morsehd [25] did acknowledge that there would be special cases where modifications would be needed to handle certain syntactic structures. These specific structures would not be immediately accepted by those annotating relevant lexemes. In spite of that Morshed argued that using this representation for multi-part lexemes will make them usable in the syntactic parsing of other texts.

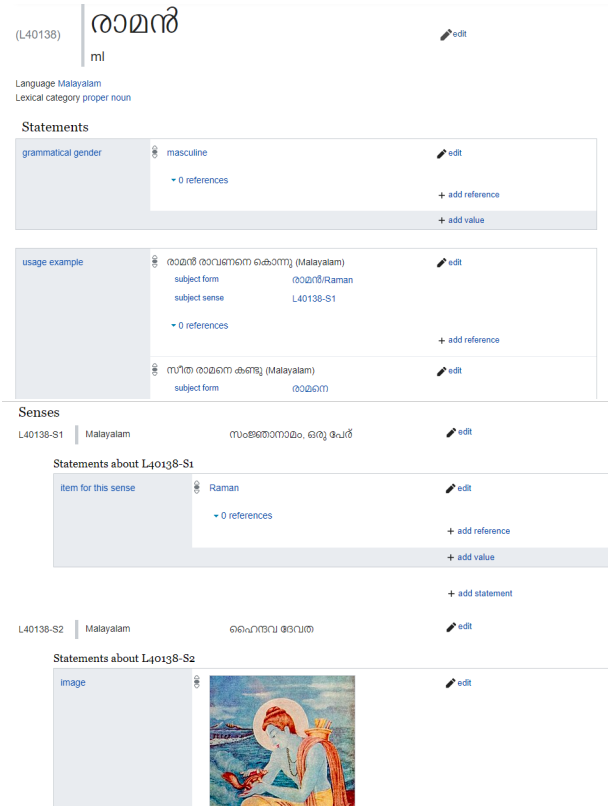

Figure 7. A Representation of a Lexeme for the Malayalam word 'ram'. [11]

8 Multiword Expressions in Lexicographical Databases

Gantar et al.[9] thoroughly discussed the synergy between the lexicographic and natural language processing (NLP) communities in regard to multiword expressions (MWEs). MWEs, according to Gantar [9] are phrases or constructions that consist of more than one word and have a specific meaning that cannot be easily inferred from the meanings of the individual words. They are an extremely important and recent phenomenon for research in linguistics, including its practical applications, as they represent an extensive part of the mental lexicon of native speakers in languages. These MWEs's a challenging proposition for natural language processing (NLP) systems to handle correctly and accurately. This stems from the fact that their meaning cannot always be predicted from the meanings of their component words. In addition, the grammar and syntax of MWEs can be complex, unconventional, and unusual further complicating NLP analysis. Lexicographic databases can address this issue by providing information about MWEs,. This information would include their meanings, typical usage patterns, and any special grammatical/syntactic features they may have. By including MWEs in their databases, lexicographers can help improve the accuracy and effectiveness of NLP systems. Their paper went on to present practical implementations of MWEs in lexicographic databases, each with a different extent of multilingualism. such as the Kamusi Project and the The Algemeen Nederlands Woordenboek (ANW).

9 Monolingual Lexicograpical Database

9.1 Algemeen Nederlands Woordenboek

According to Gantar et all , [9] , The Algemeen Nederlands Woordenboek (ANW) is a freely-available online dictionary and lexicographic database of contemporary Dutch that implements four types of multi-word expressions. The ANW uses a corpus to document the usage of words and phrases in context, providing information on their frequency, distribution, meaning, historical development, and cultural connotations. The ANW data is stored in a MySQL database together with metadata and uses a dedicated Dictionary Writing System consisting of two parts: the lexicographic workstation and the ANW editor. The ANW database is organized around the concept of lexemes, and the definition table of the ANW database contains all the definitions associated with each lexeme, while the usage examples table contains examples of how the lexeme is used in context. In another study, Tiberius et all, [12], explained that the ANW has means of automatic data capture, and data on spelling, inflection, and hyphenation is automatically inserted from the official Word list of the Dutch Language.

10 Summary Of Lexical Resources Covered

Overview of Lexical resources Discussed

Name	Multilingual	Bilingual	Low Resourced	Multi-Word Expressions(MWE)
the African Wordnet (AfWN)	✔	✔	✔	✔
Russian-Tatar lexicographical database data mode	✘	✔	✔	✘
BabelNet	✔	✔	✘	✘
Kamusi Porject	✔	✔	✘	✔
Algemeen Nederlands Woordenboek (ANW)	✘	✘	✘	✔
EuroWordNet	✔	✔	✘	✔

Figure 8. Table summary and critical comparison of the database implementations and tools that were reviewed.

11 Conclusion

In conclusion, developing a lexicographical database and other lexicographical resources for storing multilingual data is a complex but necessary task for supporting projects like the Abstract Wikipedia project. Through our literature review, we have explored various approaches and techniques used in the development of lexicographical resources, such

as fine levels of specification of linguistic markup and models for internal relationships between lexical systems, as well as the use of characteristic vectors for encoding information. We have also seen that lexicographical databases can vary in their format, information, and intended use, serving different functions such as organizing language content, standardizing language for communication, aiding language learning, and providing translation tools. In regards to the Abstract Wikipedia project, which aims to provide multilingual access to encyclopedic content through a centralized database, the development of lexicographical databases and lexicographical resources becomes even more critical.

Overall, the development of lexicographical databases for storing multilingual data is a necessary step toward achieving the goal of providing access to knowledge in various languages. While exists challenges, as with everything, the potential that such databases have cannot be overstated. Ongoing research and development in this area are crucial for advancing multilingual communication and understanding.

References

[1] Denny Vrandečić. 2021. Building a multilingual Wikipedia. *Communications of the ACM* 64, 4 (2021), 38–41. DOI: https://doi.org/10.1145/3442337.

[2] Bergenholtz, H. and Gouws, R.H. 2012. What is Lexicography?. *Lexikos*. 22, 1 (Nov. 2012). DOI:https://doi.org/10.5788/22-1-996.

[3] Nielsen, F. (2020). Lexemes in Wikidata: 2020 status. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)* (pp. 82-86). European Language Resources Association.

[4] Wikidata. (n.d.). Wikidata: Lexicographical data documentation. Retrieved March 14, 2023, from https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation.

[5] Bergenholtz, H. and Nielsen, J.S. 2013. What is a lexicographical database? *Lexikos* 23 (2013), 77–87.

[6] Paola Anna Fuertes-Olivera and Henning Bergenholtz (Eds.). e-Lexicography: the internet, digital initiatives and lexicography. A&C Black, October 20, 2011.

[7] Bonati, Isabella. "Digital Papyrological Editions and the Experience of a Lexicographical Database." *Digital Papyrology II* (2018): 149.

[8] Björn Pétur Svavarsson and Jörgen Fin. 1999. Database systems for lexicographic work. In *Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics*, 126-133. Association for Computational Linguistics.

[9] P. Gantar, L. Colman, C. Parra Escartín, and H. Martínez Alonso, "Multiword Expressions: Between Lexicography and NLP," *International Journal of Lexicography*, vol. 32, no. 2, pp. 138–162, Jun. 2019.

[10] Wim Peters, Piek Vossen, Patricia Diez-Orzas, and Geert Adriaens. Cross-linguistic alignment of wordnets with an inter-lingual-index. In *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, pages 149–179, 1998.

[11] Wikidata contributors. (2023). Lexeme:L40138. Retrieved March 24, 2023, from https://www.wikidata.org/wiki/Lexeme:L40138

[12] C. Tiberius and T. Schoonheim. The Algemeen Nederlands Woordenboek (ANW) and its lexicographical process. *Online publizierte Arbeiten zur Linguistik*, 2016, p.20.

[13] Martin Benjamin and Paula Radetzky. Small Languages, Big Data: Multilingual Computational Tools and Techniques for the Lexicography of Endangered Languages. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW '17 Companion)*, pages 355–360. Association for Computing Machinery, New York, NY, USA, 2017.

[14] Philipp Cimiano, John P. McCrae, and Paul Buitelaar. Lexicon model for ontologies: Community report, 10 May 2016. May 2016.

[15] Galieva, A. M., Nevzorova, O. A., Gatiatullin, A. R. (2014). Towards building Wordnet for the Tatar language: A semantic model of the verb system. In *Knowledge Engineering and the Semantic Web: 5th International Conference, KESW 2014, Kazan, Russia, September 29–October 1, 2014. Proceedings 5*. Springer International Publishing.

[16] P. Vossen and V. Piek. 1198. EuroWordNet: General document. ACM.

[17] S.E. Bosch and M. Griesel. 2020. Exploring the documentation and preservation of African indigenous knowledge in a digital lexical database. *Lexikos* 30, 1-28.

[18] McCrae, J.P., E. Rudnicka and F. Bond. 2020. English WordNet: A New Open-source Wordnet for English. *K Lexical News* 28: 37-44. Available: https://kln.lexicala.com/kln28/mccrae-rudnicka-bond-english-wordnet/

[19] L. Bentivogli and E. Pianta. Looking for Lexical Gaps. In U. Heid et al. (Eds.), *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*, Stuttgart, Germany, 8–12 August 2000, pages 8–12. Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, 2000.

[20] Fellbaum, C. 2010. WordNet. In *Theory and applications of ontology: computer applications*, Dordrecht, Springer Netherlands, 231-243.

[21] Tavast, A., Langemets, M., Kallas, J., Koppel, K. Unified data modelling for presenting lexical data: The case of ekilex. In Ed. J. Čibej, V. Gorjanc, I. Kosem Simon Krek, Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana (pp. 749-761). July 2018.

[22] Roberto Navigli. 2013. A Quick Tour of BabelNet 1.1. In: Alexander Gelbukh (ed.) *Computational Linguistics and Intelligent Text Processing. CICLing 2013*. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, vol. 7816. DOI: https://doi.org/10.1007/978-3-642-37247-6_3.

[23] Roberto Navigli, Marco Bevilacqua, Stefano Conia, Daniele Montagnini, and Francesco Cecconi. 2021. Ten Years of BabelNet: A Survey. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence* (IJCAI '21), August 19–27, 2021, Montreal, Canada. ACM, New York, NY, USA, 4559–4567. DOI: https://doi.org/10.24963/ijcai.2021/632.

[24] Danie J. Prinsloo. (2014). Lexicographic treatment of kinship terms in an English/Sepedi–Setswana–Sesotho dictionary with an amalgamated lemmalist. *Lexikos*, 24, 272-290.

[25] Morshed, M. (2021). Modeling Syntactic Dependency Relationships in Wikidata Lexicographical Data. In *Proceedings of the Wikidata@ISWC Workshop*.