UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours Project Final Paper 2023

Title: AfriLex: Building a Lexicographical Database for Niger-Congo B Languages and Enabling Data Upload to Wikidata

Author:Tadiwa Magwenzi

Project Abbreviation: WiNG

Supervisor(s): Maria Keet

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	15
Theoretical Analysis	*0*	*25*	0
Experiment Design and Execution	*0*	*20*	0
System Development and Implementation	*0*	*20*	15
Results, Findings and Conclusions	*10*	*20*	15
Aim Formulation and Background Work	*10*	*15*	15
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	
Total marks		**80**	

AfriLex: Building a Lexicographical Database for Niger-Congo B Languages and Enabling Data Upload to Wikidata

Tadiwa Magwenzi
University Of Cape Town

ABSTRACT

The Niger-Congo B languages, including the popular Bantu languages, are known for their complex grammatical traits and detailed linguistic nuances. Platforms like Wikidata, while abundant in lexicographical data, have a noticeable deficiency in low-resourced languages, especially the Niger-Congo B languages. The absence of comprehensive and precise lexicographical data adversely affects projects that depend on Wikidata as a foundational resource, such as 'Abstract Wikipedia'.To address this, a specialized separate, platform that aligns with Wikidata's infrastructure is essential. Tailored to the unique nuances of the Niger-Congo B language family, this platform streamlines data collection and uploads, thus enhancing the overall linguistic richness of the Wikidata.

Our solution was a lexicographical database application, "AfriLex" which includes a database specifically tailored for the Niger-Congo B language family, and an interface to allow users to upload, select, and determine the quantity of lexicographical data, as well as download data and access other database functions. In order to facilitate efficient collection and batch uploads of this data to Wikidata we crafted the WingUCTBOT, an automation bot approved by Wikidata. Ultimately, the application successfully test-uploaded five hundred isiZulu nouns, twenty chiShona Verbs, and ten isiZulu Adjectives to Wikidata, although we were unable to be granted permission to upload to the main Wikidata site. We also implemented a verb form generator to atomically generate the various forms of the verbs in the database. The application was successful in creating a platform that can increase the quality and quantity of lexicographic data in Wikidata for Niger-Congo B Languages.

KEYWORDS

Wikidata, Niger Congo B Languages, Lexicography, Databases, Abstract Wikipedia

1 INTRODUCTION

The Niger-Congo B languages, which span across sub-Saharan Africa, are celebrated for their vast geographical distribution and profound linguistic diversity [25]. A prime example is isiZulu, a member of the Bantu subgroup, which showcases intricate morphological structures and an agglutinative nature – a feature characterized by the formation of words through the addition of prefixes and suffixes [27]. Bantu languages, distinguished by traits such as nasal consonant clusters [30], can be found from Cameroon to Kenya. Among these, Swahili stands out. While it is natively spoken by five million people, it also serves as a second language for an impressive 30 million individuals [26]. Lexicographic data dives deep into details like spelling, pronunciation, and definition. At its core, lexemes, which are the fundamental units of this data, symbolize abstract meaning units and correlate with word forms [31, 33].

Our focus is Wikidata, which is a free and open knowledge base that anyone can edit [1]. It is a collaborative project by the Wikimedia Foundation, the same organization that runs Wikipedia. Wikidata aims to collect and organize the world's knowledge in a structured way so that it can be easily accessed and reused.[1] One of its standout features is its emphasis on language-independent data. This means that Wikidata is designed to store information in a manner that is not tied to any specific language. As a result, data can be accessed, edited, and utilised by individuals from diverse linguistic backgrounds. To allow this, Wikidata has included special pages for lexicographic data distinguished from the usual Wikidata 'Q-items' [17] with a new namespace for lexemes. Each page represents a 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.

Despite its stature as a prominent knowledge base, Wikidata exhibits a notable deficiency in lexicographical data for numerous African languages, particularly those belonging to the Niger-Congo-B family. This shortfall becomes particularly conspicuous when assessing the lexemes, the fundamental lexical units of a language, available for these languages in comparison to more globally recogniSed languages. For example, as of 2022, the Zulu language, a member of the Niger-Congo B family, accounted for a mere 1,000 lexemes on Wikidata, in stark contrast to English which boasted over 400,000. Such stark imbalances underscore the prevalent underrepresentation of African languages on international platforms like Wikidata[16]. Such a gap in accurate representation adversely affects projects that depend on Wikidata as a foundational resource, such as 'Abstract Wikipedia'.Abstract Wikipedia is an initiative designed to decouple the content of Wikipedia articles from their written language, generating language-neutral abstracts instead [16].

Given the complex linguistic nuances of the Niger-Congo B languages, there's an evident need for a dedicated platform where lexicographical data from these languages can be collected and edited. Such a platform, tailored to align seamlessly with Wikidata, would allow for batch uploading, a feature that promises not just to increase the volume of lexicographical data, but also its quality. By housing this data on a separate platform first, provides an opportunity to apply various database functions to refine and perfect the data before integrating it into Wikidata. Moreover, harnessing the unique features of Niger-Congo B languages could offer insights to reshape Wikipedia, enriching its comprehension and representation of Bantu languages. This endeavor aims to enhance Wikidata's linguistic database, molding it into a more inclusive and sophisticated hub for global languages.

Our solution to these problems was the AfriLex database application. An interactive database application that sits upon a lexicographical database (separate from Wikidata) designed specifically to model Niger-Congo B languages. The database was designed to capture the linguistic complexity of Niger-Congo B languages. AfriLex is designed to effortlessly upload Niger-Congo B lexemes to Wikidata in bulk, these batches can be of specific languages, sizes, and grammatical categories. Additionally, users can seamlessly access a myriad of other database functions, such as SPARQL queries and downloads through the integrated interface, thus making their interaction with the platform both productive and enjoyable. The application also included a Verb Form Generator which was based on similar work done by Hyman et al. l [10]. This feature produced the various verb forms for all the verb stems found in the database through the use of linguistic marker elements such as the Subject, Tense, and Object Markers. In order to facilitate efficient collection and batch uploads of this data to Wikidata we crafted the WingUCTBOT, an automation bot, approved by Wikidata, that handles the data collection and upload processes.

As this report will show, we were able to successfully locally store, and test upload five hundred isiZulu nouns, twenty chiShona Verbs, and ten isiZulu Adjectives to Wikidata. The verb generator was also successfully able to generate 8500 correct Verb Forms, which was an accuracy of 85%. Although, ultimately we were unable to get approval in time to mass upload to the main Wikidata site.

To reiterate, the primary goal of the project was to increase the quality and quantity of lexicographic data in Wikidata for Bantu Languages in hopes of benefiting projects such as Abstract Wikipedia which relies on it. The following report will show that we achieved this goal and will detail how we came to that conclusion. First, we will review related works and provide a background, then we will break down the design process of the application, and how it was implemented. Following that, we will then describe how each aspect of the project was evaluated and then detail the results of said evaluation.

2 BACKGROUND AND RELATED WORK

The Niger-Congo B languages, prominently featuring the Bantu subcategory, stand out in Africa's linguistic panorama due to their intricate grammar and linguistic details. While platforms like Wikidata have made strides in cataloging global linguistic data, there's a conspicuous deficit when it comes to these specific African languages. This lacuna is particularly evident when gauging the number of lexemes available for these languages compared to globally dominant tongues. Initiatives like "AfriLex" aim to bridge this gap, tailoring a database specifically for the Niger-Congo B languages, and ensuring seamless synchronization with Wikidata's vast infrastructure. In our background and related works section, we will delve deeper, exploring the broader context of this linguistic imbalance, the significance of Wikidata, and the inception and impact of the AfriLex initiative.

2.1 Niger-Congo B languages Linguistic Features

The Niger-Congo B language family, as cited in [15], can trace its beginnings to the Proto-Bantu language that emerged approximately 2,500-3,000 years ago in West Africa. Driven by factors such as trade and migration, speakers of these languages began to disperse, leading to a rich diversification of dialects. A key subgroup, the Benue-Congo, encompasses a vast array of Bantu languages and has significantly shaped Africa's linguistic narrative [26]. Through the Bantu expansion from 3000-2000 BCE, these tongues found homes across Central, Eastern, and Southern Africa. Presently, languages like Swahili have gained immense popularity, while numerous others remain restricted to specific communities, risking obsolescence. A notable consistency among these languages is their profound verb-to-verb derivations. John T. Bendor-Samuel delves into the distinctive linguistic features of this family in his 2018 article "Niger-Congo languages" from the Encyclopedia Britannica [4]. A prominent trait is the noun class system where nouns receive categorisation through distinct affixes, often influencing other parts of the sentence to create harmony. The tonal system is another defining characteristic, with the majority of languages possessing two to three pitch levels. Bendor-Samuel's article also sheds light on aspects like vowel harmony, nasalised vowels, and verb serialization, providing a broad insight into the Niger-Congo languages, though possibly not capturing the depth or the socio-cultural contexts of each language.

2.2 Lexicography and the Current State of Lexicographical Data

Lexicography, as defined by Galieva et al. [28], pertains to the creation of dictionaries, combining linguistic theory and practical application. These systems necessitate specialized software tools and linguistic formalization, with linguists isolating and describing lexical units. Bergenholtz et al. [13] emphasize that a lexicographical database should encompass a headword, sense, grammatical details, and examples. Lexibank [34], inspired by GenBank, consolidates lexical datasets from diverse languages, adhering to FAIR principles. Meanwhile, the World Atlas of Languages [35] documents over 8,000 languages, highlighting those still in use. The SeLA project, as detailed in "Ongoing work on e-lexicography in the SeLA project" by Heid, Ulrich [48], is a collaborative initiative involving universities from Germany, South Africa, and Namibia. The project delves into the intricacies of electronic lexicography. Its interdisciplinary nature allows for a comprehensive exploration of data acquisition from corpora, lexicographic data representation, and user orientation in electronic dictionaries. SeLA also emphasizes the development of a user-centric theory of lexicography. Notably, the project has been instrumental in offering training courses and workshops in South Africa, focusing on lexicographic theory, data acquisition, and user-focused dictionary design.

2.3 Lexicographic Databases

Bergenholtz et al.[13] defined a lexicographical database as a computerized collection of lexical information. They identified four critical components: a headword or lemma, sense or meaning, grammatical information, and illustrative examples. Additionally, they proposed a structured lexicographical database for Spanish monolingual dictionaries with 23 fields including lemma, meaning, synonym, antonym, and internet link. In contrast, FuertesOlivera's study[14] emphasized the significance of lexicographical databases in modern lexicography, underscoring their advantages over print dictionaries, like data storage, dynamic linking, and multimedia integration. The study explored various types of databases, highlighting their pros and cons, and cited practical examples. It also delved into corpus-based lexicography, emphasizing its role in understanding actual language usage by analyzing extensive language samples. Fuertes commented that Corpus-based lexicography allows dictionaries to reflect the actual use of language.

2.4 Wikidata and Abstract Wikipedia

Vrandecic et al [1] wrote that Wikidata functions as a centralized storage repository, primarily for the wikis managed by the Wikimedia Foundation. eliminates the need for redundant data maintenance across individual wiki projects by centralizing information like statistics and dates. They went on to say that at it's core, Wikidata uses "items", each with a unique identifier, described further using "statements" that pair properties with values. [3] This structured data can be displayed in any language, ensuring up-to-date information across platforms. Various tools, including the Wikidata Query and Reasonator, facilitate access, and the data can be integrated into other platforms using the Wikidata API. Lexemes, as previously stated [37] are the abstract units of language that are associated with a set of inflected forms or word senses. A lexeme is used to describe and define Lexicographic data. This functionality enables users to access and understand their meanings, usage, and relationships with other lexemes according to the Wikidata documentation [37]. The paper "Lexemes in Wikidata: 2020 status" by Finn Arup Nielsen [17] offers an insightful update on the status of lexemes in Wikidata, emphasizing its multilingual aspects. Using descriptive statistics and the Wikidata Query Service (WDQS), Nielsen delves into the growth and intricacies of lexemes, senses, and lexical forms since 2018. While the paper showcases the advancements, such as the use of tools like the Lexicator for Russian lexemes and the Wikidata Lexeme Forms for aiding editors, it also underscores the limitations. A significant challenge is the limited extent of lexeme data and annotations for etymology and senses, with Russian being a notable exception due to its automated integration from the Russian Wiktionary. The paper also touches upon the complexities of linking non-noun lexemes and the copyright concerns surrounding lexicographical data in Wikidata. Despite these challenges, Nielsen's work paints a picture of continuous growth and evolution in the realm of Wikidata's lexicographic data.

Abstract Wikipedia is a project proposed by Denny Vrandecic as a solution to the unequal distribution of knowledge across languages in Wikipedia. According to Vrandečić [16] 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, the number of contributors is very unevenly distributed. The project aims to enhance content accessibility in desired languages using NLG and a Renderer function in Wikifunctions. It utilizes lexemes and lexicographic data from Wikidata. Lexemes, fundamental language units, are described with data that aids in understanding their use and connections[19].

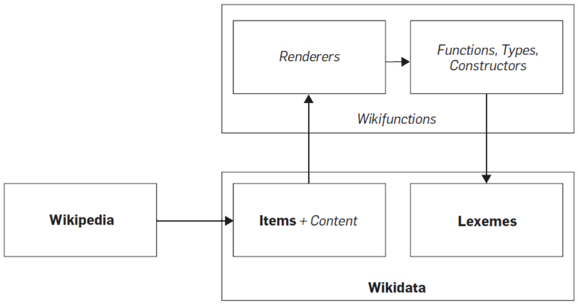


Figure 1: Architecture of the multilingual Wikipedia proposal, Abstract Wikipedia.[16]

2.5 Previous Efforts and Solutions

Bosch and Faaß[7] populated a MySQL database with Zulu nouns and English translations, crafting a model for African languages. Despite benchmarks like ISO 24613:2008 and Spohr's XML/OWL, they emphasized 'sense description' over the traditional lemma, contending other components depend on this description. Their work, from the "Scientific eLexicography for Africa" [48] project, aims for online dictionaries like Northern Sotho and bilingual lexicons for Xhosa–English and Zulu–English. They planned to release their tools publicly by 2015. Their distinctive approach, especially prioritizing MySQL over XML/OWL and downplaying the lemma, offers a fresh perspective, aiming for adaptability to users' evolving needs.

In relation to endeavors focused on building a Bantu language database, insights from Inge Kosch's [8] research from the Department of African Languages at the University of South Africa are of paramount significance. Kosch underscores the intricate relationship between dictionary users and lexicographers, especially within Bantu languages. She postulates that an effective Bantu language database is not merely dependent on its content, but also on the user's proficiency in dictionary skills and understanding of Bantu linguistic structures. Kosch's findings intimate that database developers, akin to lexicographers, should be cognizant of the varying capabilities of their users—spanning from basic word searches to intricate language-specific consultations. Consequently, it becomes essential for those spearheading the Bantu database project to ensure a harmonious blend of user-friendly interfaces, robust user guidelines, and rich linguistic content. Kosch's study ultimately sheds light on the symbiotic relationship between database design and user competency, particularly in the realm of Bantu languages.

3 DESIGN

In the pursuit of a system embodying both robustness and scalability, the architecture of AfriLex was crafted as a modular entity. This design paradigm ensures that each significant function of the system resides within a unique module, facilitating efficient management, future scalability, and streamlined troubleshooting.

3.1 Requirements Analysis

The requirements for the AfriLex platform were gathered using a multi-faceted approach to ensure accuracy and comprehensiveness. Primarily, face-to-face consultations with the project supervisor provided direct insights into the core needs and expectations. Additionally, a thorough review and consultation with similar lexicographical databases and linguistic resources offered a broader perspective on industry standards and best practices.

To further validate and refine these requirements, a triangulation method was employed. This involved cross-referencing the gathered requirements with other Niger-Congo B linguistic databases. By comparing and contrasting the needs identified by the project supervisor with insights from these specialised linguistic sources, a more holistic and informed set of requirements was established. This triangulation ensured that the requirements were not only aligned with the project's goals but were also grounded in the linguistic intricacies of the Niger-Congo B languages.

Once collated, these requirements were translated into clear software objectives that directly informed the architectural design and overall system development of the platform. Utilising the waterfall development methodology, the requirements were methodically categorized and prioritised based on their relevance to the project's goals and the dependencies they introduced. Such a structured approach ensured that the development team could focus on and prioritise the most crucial features, ensuring efficient use of resources and time.

The requirements analysis for this project can be divided into three main parts:

1. The requirements for the development of a separate lexicographical database for Niger-Congo B languages.
2. The requirements for the batch upload of data from this database to Wikidata.
3. The requirements for an interface to the database.

3.1.1 Requirements for the development of a separate lexicographical database. It was essential that the database could store lexicographical data from any Niger-Congo B language. Additionally, it should be capable of modeling language family-specific grammatical features. . Above all, the database needs to be robust, stable, and user-friendly.

3.1.2 *Requirements for the batch upload of data from the database to Wikidata.*

For the batch upload of data from the database to Wikidata, the process needs to be automated, efficient, and accurate. Moreover, the uploaded data should meet the standards set by the Wikidata API, passing its numerous value and format checks.

3.1.3 *Requirements for the interface to the lexicographical database.*

It should be responsive and fast to ensure a smooth user experience. Security measures need to be implemented to safeguard both the data and user information. The design should incorporate user profiles,. Users ought to have the capability to select lexemes based on their grammatical type, and language, or even upload data directly from Wikidata. Additionally, the interface should feature a SPARQL query point, presenting results directly to the user

These requirements are essential for the successful development and implementation of this project. By meeting these requirements, the project will be able to address the challenges faced by the lexicographical component of Abstract Wikipedia and make a significant contribution to the linguistic diversity of Wikidata.

3.2 Design Approach

The modular design of the AfriLex database application was intentionally chosen in anticipation of future growth. As more languages or features are introduced, new modules can be seamlessly developed and integrated, ensuring the existing system would remain undisturbed[40]. Additionally, this modular approach enhances maintainability, allowing individual components to be updated, debugged, or replaced independently. Foremost is the AfriLex Database Module, the very cornerstone of the system, which has been tailored to capture and model linguistic data. It was designed for adaptability to the intricate linguistic features of the Niger-Congo B languages. This diversity of these languages is comprehensively represented. Serving as a conduit between users and the extensive lexicon of AfriLex is the User Interface Module. It allows the execution of SPARQL queries to enable data uploads and downloads, it also presents users with a platform that's both intuitive and seamlessly interactive. The data Upload Module is responsible for pulling data in bulk from the AfriLex database and uploading it to Wikidata. Assisted by the WingUCTBOT, this module efficiently uploads lexemes along with their associated senses and forms to Wikidata. And finally, the Verb Form Generator module is another key component of AfriLex. Rooted in linguistic academia, it's tailored to automate the generation of diverse verb forms by utilizing linguistic marker elements.

3.3 AfriLex Database Design

To ensure a comprehensive and representative database model, we embarked on an iterative design cycle. Initially, a cross-section of prominent Niger-Congo B language resources and analogous database projects were analysed [49]. This preliminary investigation aimed to identify commonly represented and crucial linguistic features of the Niger-Congo B languages that the AfriLex database should encapsulate. Drawing insights from this cross-sectional study, the initial ORM was crafted, serving as a prototype that encapsulated the core linguistic features. This initial ORM model then underwent iterative refinements based on insights from various other resources and database analyses, ensuring the model was comprehensive and accurately represented the linguistic features found in other works. This iterative design cycle also ensured a reduced risk of failure [36] as problems were identified and improvements were made before further development was made with the rest of the application. Using an Object-Relational Mapping model for a Niger-Congo B language database allows for an intuitive object-oriented representation of complex linguistic structures and offers an abstraction layer [6], simplifying development while ensuring data integrity. This approach facilitated focus on linguistic intricacies without getting entangled in database-specific nuances.

3.3.1 *Key Entities and Relationships.*

The database grasps the linguistic depth of Niger-Congo B languages through a mesh of entities and attributes. The LanguageFamily, entity enumerates various Niger-Congo B language families, distinguished by identifiers, denominations, and geographic spreads. The Language entity delves deeper into individual Niger-Congo B languages, shining light on attributes like designations, ISO codes, and dominant territories. Inherently connected to its progenitor LanguageFamily with a many-to-one predicate, it branches further to portray linguistic details, embodied by entities like Morpheme, Phoneme, SoundChangeRule, and InflectionalCategories. The Word entity presents a rich tapestry of lexical entries highlighting lemmas and interpretations. Every word is woven with a LexicalEntry to affiliations with linguistic constituents such as Inflections and Tones. The LexicalEntry is the heart, embracing an array of word data ranging from identifiers to user data and beyond.

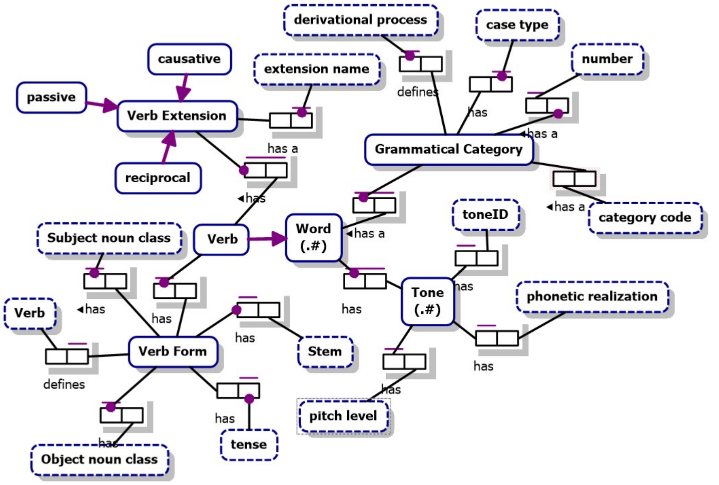


Figure 2: Submodel of the Verb Form, Tone, Grammatical Category, and Verb Extension Entities

A full visualisation of the entire Database Model is attached in the appendix.See Appendix section 4

Continuing from the foundational entities, the database further encapsulates the tonal intricacies of Niger-Congo B languages through the TonalPattern entity. This is inherently linked to the Language entity, ensuring each tonal pattern is language-specific. The Tone entity, with attributes like phoneticRealization and pitch level, is associated with the Word entity, capturing the tonal essence of each word. The User entity, vital for user management and interaction ensures secure and personalized access to the lexicographical database application. The verbform entity is designed to capture relative concords related to generated verb forms. while VerbalMorphology focuses on the grammatical intricacies with attributes such as Tense, Aspect, Mood, Voice, and Valence. The VerbExtension entity provides insights into verb modifications with attributes like, causative, reciprocal, and passive.

3.4 Database Management System (DBMS) Selection and Schema Selection

MySQL was selected for AfriLex due to its reliability, performance, and flexibility—essential for efficient linguistic database management. Its SQL-based structure ensures consistent data handling. The inclusion of MySQL Workbench provides an integrated toolset for easy design, development, and oversight of AfriLex. It efficiently handles large datasets [12], accommodating the linguistic intricacies of Niger-Congo B languages. Its ACID support ensures data reliability, and its scalability caters to AfriLex's growth. Additionally, MySQL Workbench simplifies real-time monitoring and recovery, promoting stability.

The decision to opt for SQL over XML/OWL, despite the latter's prevalence in recent lexicographical data models, was driven by SQL's superior scalability and adaptability. SQL excels in managing large datasets, offering efficient querying, and maintaining data integrity. Its relational structure facilitates versatile data modeling, making it optimal for intricate lexicographical endeavors. A populated MySQL database can be equated to a standoff XML system, where both can represent necessary data items and their interrelations and SQL offers a swift, straightforward implementation without the requisites of DTDs, XML editors, or commercial Dictionary Writing Systems. [7]

3.5 Inclusion of Language-Specific Features and Linguistic Accuracy

The AfriLex ORM diagram, tailored for the Niger-Congo B languages, encapsulates their intricate linguistic features. The word ’ entity ’ encompasses all linguistic units, while entities like NounStem, VerbStem, and ’inflection’ capture the essence of agglutination, a core trait highlighted by Faaß et al. [7]. Grammatical structures are outlined via the GrammaticalCategory, and GrammarRule, ’ entities, with specific constraints for unique Niger-Congo B attributes. The ORM also integrates the Nominal and Verbal Derivation System, emphasizing word derivation and inflection, as stressed by Faaß et al. [7]. Aligning with insights from Byamugisha et al. [20], the ’NounClass’ entity underscores the significance of noun classes in this linguistic domain. Entities like Vowel, VowelCluster, and VerbStem capture the phonological intricacies of the Niger-Congo B languages. The Vowel and VowelCluster entities focus on individual sounds and combinations, respectively, while the VerbStem provides insights into verb roots. Other entities, including ’inflectionalcategories’ and ’phonologicalpattern’, further enrich the understanding of phonetic variations.

In essence, the AfriLex ORM, enriched by these entities and their intricate relationships, offers a profound and comprehensive representation of the Niger-Congo B languages, capturing their lexicographical, grammatical, and phonetic nuances with utmost precision.

3.6 Data Pipeline Design

In regards to the Data Preparation and Upload to Wikidata, to streamline the efficient management of large lexeme batches from Niger-Congo B languages, we developed a custom bot tailored specifically for this task, WingUCTBOT. Recognizing the impracticality of manual uploads due to the vast volume of linguistic data and meeting one of the key requirements which was to allow the batch upload of lexicographical data, the bot serves to accelerate the upload process while ensuring data consistency and integrity. It is also designed to handle potential API rate limits, ensuring Wikidata's servers are not overwhelmed. Should the API generate any errors, the bot logs them for review but continues its operation, preventing any disruption in the upload process. The strategic development and deployment of this bot not only optimize the upload speed but also prioritize data quality and seamless integration, ultimately fulfilling our objective of a comprehensive representation of Niger-Congo B languages on Wikidata.

3.7 Application Infrastructure and Architecture

The AfriLex database is designed for an intuitive user experience, combining a lightweight Python Flask backend with a dynamic JavaScript frontend. Flask, chosen for its modularity [38] and easy MySQL integration, ensures an agile and scalable backend. The front end, powered by JavaScript, provides a responsive UI and real-time data updates, enhanced by specialised libraries for interactive UI elements.

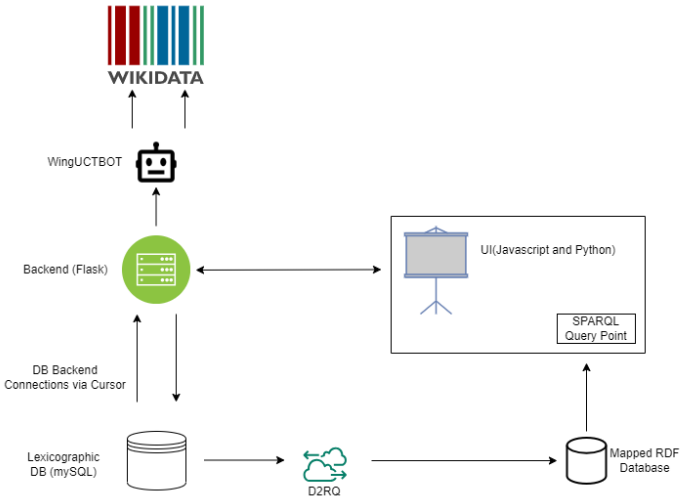


Figure 3: AfriLex Application Architecture Diagram

The integration of Flask and JavaScript is pivotal for seamless data exchange between the front and back end, ensuring data integrity. Both ends have validation checks for data accuracy. JavaScript libraries visually represent relationships from the data model, providing users with an interactive platform to navigate the data.

3.8 SPARQL Endpoint Integration Feature Design

The inclusion of a SPARQL endpoint was a strategic move to enhance data querying capabilities and integration with semantic web technologies. The constraints of traditional querying methods often come to the forefront of traditional linguistic data. The core reason behind this design decision stems from SPARQL's ability to accommodate semantically rich and complex queries, offering users the flexibility to delve profoundly into linguistic datasets. Bergenholtz et all stated that it is of critical importance in a user-driven lexicographic approach is the need to ensure that the target users of a specific dictionary gain unimpeded access to the data they need in order to achieve an optimal retrieval of information[5]. Perez et all [Pérez et al. 2009] stated that SPARQL's graph pattern matching facility enables complex querying which is invaluable for a lexicographical database where you may need to retrieve entries based on intricate relationships or characteristics. This adaptation entailed several design refinements, essentially reshaping and optimizing the database's architecture to be more conducive to semantic querying.

3.9 Verb Form Generation Feature Design

The linguistic community needed a tool for Niger-Congo B languages due to their complex verb structures. The Verb Form Generator, inspired by Larry M. Hyman's work [10] and Keet's insights on grammar rules and verb generation [11], was developed to enhance lexicographic databases and clarify verb morphologies. It utilized the Niger-Congo B noun class to customize verb forms for Shona.

The Verb Form Generator's primary objective is to efficiently produce verb forms for the Shona language, drawing from the Niger-Congo B noun class classification principles. Structurally, the generator seamlessly incorporates four crucial linguistic elements: the Subject Marker (SM) denoting the action's executor; the Tense Marker (TM) highlighting the action's timing (past, present, or future); the Object Marker (OM) representing the action's recipient or target; and the Verb Root, conveying the fundamental nature of the action, with examples like "run" or "fear." To ensure precise verb form generation, the generator is underpinned by a comprehensive list of concords and affixes which were sourced from Students in African 671 [39], and the list of isiZulu concords compiled by Keet [27] served as a reference.

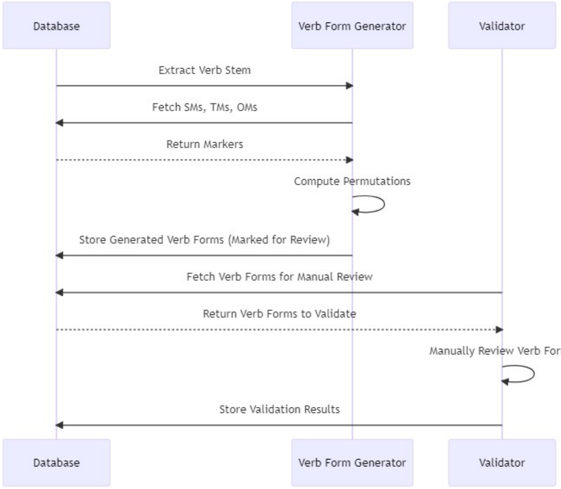


Figure 4: Verb Form Generator Sequence Diagram

3.9.1 Operational Mechanics: Using the Verb Stems currently in the database, the generator generates potential combinations of SM, TM, and OM. This systematic permutation approach ensures the generation of all possible verb forms rooted in the provided verb. It then stores each inflected form of the verb in the database and marks each generated one for further review. The forms are stored as full inflections in the Inflection table. At the same time, the individual components of the form i.e. the marker combination and the verb stem were stored in the verb form table. The design maintains the root's integrity and consistency while modifying the surrounding markers, complying with the syntactic and semantic norms of the Shona language.

4 IMPLEMENTATION

The AfriLex database application consists of a MySQL database, a Flask backend, and a frontend developed in JavaScript and Python. Leveraging Python and Flask, it offers efficient data management and backend operations. Key features include algorithms for data handling with Wikidata, an intuitive user interface enhanced by Materialize CSS, and specialized tools like the Verb Form Generator for Shona linguistic analysis. The system also supports SPARQL queries supported by D2RQ [47].

4.1 Components

The AfriLex database application comprised three, distinct components that each worked together to achieve the project goals. The first is the MySQL DataBase, which was managed by the MySQL workbench. The backend component interacted with the database, and it was realized as a Flask Application. Finally, the front end was written in plain JavaScript and Python.

4.1.1 Environment and Tools. In the AfriLex development environment, Python is the primary language, chosen for its simplicity and extensive web development libraries. We use Flask, a Python microweb framework, for backend operations, routing, and template rendering. Flask was selected for the backend because of its highly level of flexibility [41], allowing us the freedom to configure exactly how the data in our application was to be transformed and transmitted. MySQL manages the database, interfaced via mysql.connector in Python. OS, Requests, and JSON modules handle OS-specific functions, HTTP requests, and JSON data, respectively.

AfriLex's backend comprises the Flask application for HTTP routing, mysql.connector for database interactions, and the JSON Module for API communications. Flask's Request Module manages client data, Rendertemplate displays HTML templates, and Session oversees user details. URL handling utilities include urlfor, redirect, and urllib.parse, with Datetime handling date operations. Error management combines Flask components with Python's mechanisms, and when interfacing with the Wikidata API, stringent checks ensure data accuracy.

4.2 Important Algorithms

4.2.1 Data Extraction, Transformation, and Upload to Wikidata. The process of data extraction, transformation, and upload to Wikidata is initiated by the `cursor.execute` function, which retrieves a specific number of LexicalEntry IDs based on the `batchsize`. These IDs are then probed through successive SQL queries to gather details about the associated word, including its language, grammatical category, inflections, and meanings. Each query is crafted to pull data from different tables in the MySQL database, leveraging table joins and WHERE clauses for precise data collection. Once extracted, the data is locally stored in variables, such as `language_code` for the word's language and `cat` for its grammatical category, ensuring it's readily available in the desired format for any subsequent transformations or validations. An essential step in this process is the data integrity checks. One crucial check involves determining the presence or absence of inflections. This approach guarantees that the data uploaded to Wikidata is both accurate and specifically tailored to its content. A small section of the extraction code is below.

```
query = (
    "SELECT␣gramfeaturesid,␣inflection␣"
    "FROM␣inflection␣AS␣i␣"
    "JOIN␣lexicalentry␣AS␣l␣ON␣i.base␣=␣l.Word␣"
    "WHERE␣l.LexicalEntry_id␣=␣%s"
)
cursor.execute(query, (lexical_entry_id,))

# Fetch the results and remove duplicates
results = cursor.fetchall()
```

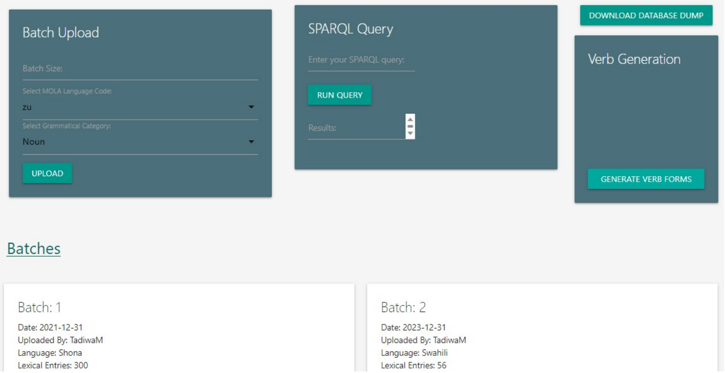
```
unique_forms = list(set(results))

forms = unique_forms
```

Listing 1: "Extracting the different forms of a lexical entry"

4.2.2 Bot for Wikidata Upload. The data conversion phase readies data in a JSON format to match Wikidata's lexeme standards, using values like lemmas and language from the database. The lexeme structure depends on specific data, especially inflections. For API interaction, a LOGINTOKEN is first obtained and used for Wikidata server authentication. After this, a CSRFTOKEN is secured to ensure request security, guarding against web vulnerabilities. Data is then submitted via the wbeditentity API endpoint, with the bot suggesting new lexeme entities based on structured JSON data. Post-submission, the API response is assessed, with successful uploads noted by the 'success' field. A successcount counter tracks all successful operations, culminating in a final report.

4.3 User Interface Implementation

The AfriLex system features an intuitive interface crafted with HTML, enhanced by the Materialize CSS and JavaScript frameworks. Key components include the "Batch Upload Card" for lexical data uploads, the "SPARQL Query Card" for executing queries, and the "Verb Form Generator" for exploring Shona verb morphologies. Users can easily upload data, view query results, and delve into linguistic intricacies. Additionally, a teal button facilitates database downloads. Authentication is streamlined with distinct login and registration panels, both offering real-time feedback. The cohesive design, blending teal and blue-grey tones, ensures a seamless user experience.



Select MOLA Language Code: zu

Select Grammatical Category: Noun

Batch: 1
Date: 2021-12-31
Uploaded By: TadiwaM
Language: Shona
Lexical Entries: 300

Batch: 2
Date: 2023-12-31
Uploaded By: TadiwaM
Language: Swahili
Lexical Entries: 56

Figure 5: AfriLex Application Interface

4.4 SPARQL Query Endpoint Implementation

The AfriLex system has integrated support for SPARQL queries through the utilization of the D2RQ platform. This platform serves as a bridge connecting relational databases to RDF, enabling seamless mapping from SQL to RDF triples without altering the foundational SQL structure. This preservation ensures that the original database design remains intact while still benefiting from the enhanced capabilities of RDF. Upon initiating the AfriLex application, an embedded D2RQ server is activated. This server presents the SQL database in the form of an RDF view, accessible through a localized SPARQL endpoint. This design ensures uninterrupted availability for users to conduct semantic queries via SPARQL. When users input SPARQL queries, the D2RQ engine comes into play. It interprets these queries, translating them to their SQL counterparts. Post-data retrieval, it then converts the results back into RDF format, ready for user consumption.

4.5 Verb Form Generation Feature Implementation

Building upon the design principles outlined in the previous section, the implementation of the Verb Form Generator was carried out in a systematic manner to ensure accuracy, and thoroughness the first step involved setting up a table that stored the Shona noun class classifications, concords, and affixes. The markers and noun classes were sourced through consultation with literature, in particular, the University of Wisconsin-Madison Students in African 671 [39]

The core functionality involves iterating through the extracted data, coupled with predefined object markers, to generate comprehensive verb forms. This process is meticulously structured to cater to multiple tenses (Past, Present, Future) and various noun classes. The algorithm employs dictionaries for mapping noun classes and tenses to their corresponding Wikidata IDs. For instance, the tensetowikidata dictionary associates tenses with their respective Wikidata representations. Similarly, the classtowikidata dictionary maps noun classes to their Wikidata counterparts, streamlining data consistency. Within the main loop, for each row in the sorted results from the database, tense prefixes are combined with object markers and the verb's root form to yield the desired verb inflection. These constructed verbs are then printed for verification and subsequently inserted into the inflection table in the database. The constructed verb forms are persistently stored in the database using the INSERT INTO SQL command. Each verb form's attributes, such as its grammatical tense, subject class, and object concord, are meticulously cataloged to enhance data retrieval efficiency later. The verb forms are inserted into the Inflections table as they are forms however a duplicate entry is made to the verb form table which stores the Subject and Object Noun classes, the markers, and the stems of each generated verb form.

Batch Processing: To optimize the verb construction process for a variety of verbs, the generator is engineered to accept a list of verbs. This batch-processing capability ensures that a multitude of verb forms are generated in a single run, bolstering productivity.

The Verb Form Generator, while a standalone feature, was seamlessly integrated into the main AfriLex system, taking advantage of its modularity, and this ensured that users could easily navigate between different functionalities without any disruptions.

5 EVALUATION

5.1 Evaluation Overview

Evaluating the outcomes of a project is paramount to understanding its success and areas of improvement. Such an assessment not only gauges the effectiveness of the implemented strategies but also provides insights into the real-world impact of the project. In the context of this endeavor, the primary objective was to augment the quality and quantity of lexicographic data in Wikidata for Bantu Languages. By measuring our achievements against this goal, we can ascertain the tangible contributions made towards enriching Wikidata's linguistic database for these significant African languages.

5.2 Data Compatibility with Wikidata

Format Compatibility Assessment: To ensure compatibility with Wikidata, we uploaded our data to the Wikidata test API, which validates the data format. Upload results can be viewed in the "recent changes" section of the TestWikidata site. Using the Wikidata API, we received detailed feedback on each entry. Out of 500 test uploads, we noted the success rate and documented failure points. Errors like invalid grammatical features, incorrect language codes, or lexeme discrepancies were identified, helping pinpoint both our strengths and areas for improvement.

5.3 Linguistic Representation

Through thorough cross-evaluation and consultation of various resources, including other Bantu languages and similar Niger-Congo B database projects, we successfully identified a comprehensive list of 163 distinct Niger-Congo B language features. These features were subsequently categorized into the following representative subgroups: Noun Class System, Nominal Morphology, Verbal Morphology, Syntax, Microstructure, Derivation and Inflection System, Word and Stem Lemmatization, Writing System, Morphosyntactic Challenges, and Sense Elements. By organizing the features into these subgroups, we were able to assess the effectiveness of the database in reflecting the intricate linguistic features in each aspect of Niger Niger-Congo B languages.

5.3.1 Completeness Verification: Once these features and subfeatures were clearly defined, they were directly compared with the features represented in the AfriLex database. The completeness evaluation consisted of a direct numbered enumeration of the amount of Niger-Congo B features that were represented in the database. The evaluation also included a detailed category-by-category analysis and an overall assessment of the representation of Bantu features. In other words per each category of features, the percentage was represented in the Afrilex database.

5.3.2 Coverage Analysis: We conducted our coverage analysis through a *Comparative Assessment*. Our database was systematically compared with seven other renowned Niger-Congo B linguistic resources and similar lexicographic databases. These selected resources are recognised for their comprehensive coverage and authoritative representation of the Niger-Congo B linguistic domain. The comparison was executed using the following methodology: For each feature category identified earlier, every resource, including Afrilex, was evaluated based on the number of features from that category they encompassed or reflected. This comparison was carried out for each Niger-Congo B feature category and for all the linguistic features therein. The sources are a mixture of research papers and Niger-Congo B databases

The sources selected are as follows:S

- Bantu Morphosyntactic Variation (BMV) Database -DB [49]
- African WordNet -DB [32]
- A General Lexicographic Model for a Typological Variety of Dictionaries in African Languages -DB[**?**]
- Towards machine-readable lexicons for South African Bantu languages -Paper [**?**]
- Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes -Paper [42]
- Linguistic description of Shiwa, Bantu language of Gabon: phonology, morphology, syntax, lexicon -DB [42]
- Software Tools for Morphological Tagging of Zulu Corpora and Lexicon Development -Paper[46]

By comparing the number of features in our database with those in other resources, we essentially aimed to gauge the comprehensiveness and distinctiveness of our resource. The comparative assessment would also provide a roadmap for potential refinements, enhancements, and additions in future iterations of our database.

5.3.3 Database Model and Structure Evaluation: The database model underwent evaluation by a distinguished Ph.D. candidate and computer science language expert Toky Hajatiana Raboanary. Their expertise is indispensable in determining the database's faithfulness to the linguistic nuances of the Niger-Congo B languages. Instead of using predetermined test data, the evaluator conducted the assessment based on their knowledge and expertise, utilizing a comprehensive form to provide feedback. This form was specifically designed to capture intricate details about various linguistic and structural facets of the database. The evaluation metrics within this form emphasised elements such as accuracy rate, instances of misrepresentations, and the potential presence of previously unidentified linguistic patterns in the database.

5.4 Evaluation of the Verb Form Generator

The primary goal of this evaluation is to ensure the accuracy and authenticity of the generated verb forms by the Verb Form Generator module, with a focus on their linguistic correctness in the context of the Shona language. The evaluator was a native speaker of Shona, possessing a deep-rooted understanding of the language's morphological nuances, Their inherent linguistic intuition was crucial in gauging the authenticity of generated verb forms. As for the Evaluation Method, for each generated verb form, a sample noun from the subject noun class and another from the object noun class will be selected. Using these nouns, a sentence will be constructed with the generated verb form at its core. The constructed sentences was then presented to the native speaker. They will evaluate the sentences based on their fluency, grammatical correctness, and overall linguistic authenticity within the Shona language. Then they provided feedback on each sentence, indicating whether it sounds natural and is grammatically correct, or if adjustments are needed.

An example would be the following: From the verb stem 'seka', meaning 'laugh' in chiShona. the verb Generator would generate the form 'chaka museka', among many others. This particular form uses the noun class seven in the past tense for the Subject and noun class one for the Object. The sentence would then be created by adding an example noun from a ready-made list to the start of the verb, such as in this case "Chidhoma' meaning clown. Thus the sentence would be 'Chidhoma chaka museka', the clown laughed at him/her. This sentence would then be assessed by a native speaker for correctness and the verb form would be marked for validity.

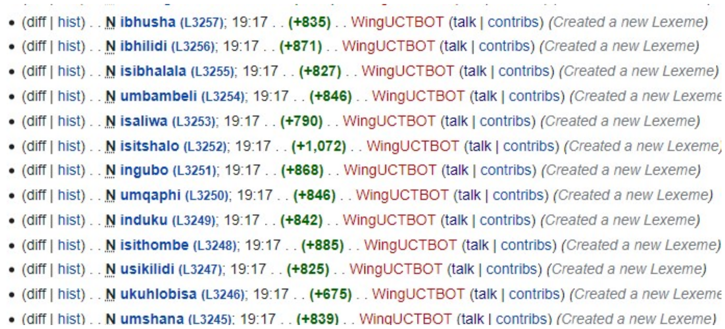
The evaluation hinges on several key metrics. Firstly, the Accuracy Rate represents the percentage of sentences deemed linguistically correct and fluent by the native speaker out of the total sentences constructed. Secondly, the Types of Errors provide a specific categorisation of mistakes, whether they are morphological incompatibility, incorrect tense formation, or noun-verb agreement discrepancies. Lastly, Consistency will be assessed by analysing how uniformly the Verb Form Generator produces correct verb forms across diverse tenses and noun classes.

6 RESULTS AND DISCUSSION

Objective Recap: The principal aim of our project was to enhance Wikidata's repository by elevating both the quality and volume of lexicographic entries pertaining to Bantu Languages. This endeavor was rooted in the recognition of the underrepresentation of these languages on the platform and the subsequent need to bridge this informational gap.

6.1 Data Compatibility with Wikidata Evaluation Results

The evaluation process for our data's compatibility with Wikidata commenced with an upload to the Wikidata test API. Upon completion, we visited TestWikidata's recent changes page and noted the outcomes of the lexeme uploads conducted by our bot. The figure below is from the test Wikidata site, It demonstrates successful upload to the test API site which indicates the data passes the wiki data compatibility and validity checks. [22] A comprehensive sampling of words was considered for this test, representing diverse grammatical categories. Our sample consisted of five hundred Nouns, ten Verbs, ten pronouns, and twenty adjectives. Upon analysis, we recorded the following results:



- (diff | hist) . . N ibhusha (L3257); 19:17 . . (+835) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)
- (diff | hist) . . N ibhilidi (L3256); 19:17 . . (+871) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)
- (diff | hist) . . N isibhalala (L3255); 19:17 . . (+827) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)
- (diff | hist) . . N umbambeli (L3254); 19:17 . . (+846) . . WingUCTBOT (talk | contribs) (Created a new Lexeme
- (diff | hist) . . N isaliwa (L3253); 19:17 . . (+790) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)
- (diff | hist) . . N isitshalo (L3252); 19:17 . . (+1,072) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)
- (diff | hist) . . N ingubo (L3251); 19:17 . . (+868) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)
- (diff | hist) . . N umqaphi (L3250); 19:17 . . (+846) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)
- (diff | hist) . . N induku (L3249); 19:17 . . (+842) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)
- (diff | hist) . . N isithombe (L3248); 19:17 . . (+885) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)
- (diff | hist) . . N usikilidi (L3247); 19:17 . . (+825) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)
- (diff | hist) . . N ukuhlobisa (L3246); 19:17 . . (+675) . . WingUCTBOT (talk | contribs) (Created a new Lexeme
- (diff | hist) . . N umshana (L3245); 19:17 . . (+839) . . WingUCTBOT (talk | contribs) (Created a new Lexeme)

Figure 6: Wikidata Recent Changes Page

Verbs: 10 out of 10 were successfully uploaded (100% percent success rate). **Nouns:** 500 out of 500 were successfully uploaded (100% perfect success rate). **Pronouns:** 8 out of 10 were successfully uploaded (97.75' perfect success rate). **Adjectives:** 18 out of 20 were successfully uploaded (97' perfect success rate). In total, out of 540 words from diverse categories, 536 were successfully uploaded, yielding a general success rate of 99.26%. The discrepancies noted were primarily due to formatting inconsistencies or non-standard linguistic representations. The database was proven to be largely successfully compatible with Wikidata mainly because of the application's specific format and validity checks before any attempt to update to Wikidata is made. However, when a lexical entry would fail these checks they would be marked for review and the attempt to upload would be aborted.

Upload Success Analysis: Despite the test data being approved multiple times, and back and forth for the better part of a month, the Wikidata officials did not approve an upload access to the release version of Wikidata.

6.2 Linguistic Representation

6.2.1 ***Completeness Verification Results:*** The Afrilex DB's evaluation against linguistic features reveals its strengths and areas for improvement. The Noun class system is well-represented at 68%, with 15 of 22 features. Nominal morphology stands out with 89% coverage, highlighting the database's capability in noun morphologies. Verbal morphology and Syntax are covered at 82% and 56%, respectively, indicating room for enhancement. Microstructure and Derivation and Inflection System are at 70% and 55%. Word and Stem Lemmatization is at 57%, suggesting refinement areas. The Writing Systems and Analysis is at 40%, indicating a need for focus. Morpho-syntactic Challenges and Sense Elements are around 67%. Impressively, the Verb Extension System is fully captured at 100%. Overall, while Afrilex DB covers many linguistic aspects, some subsystems warrant deeper exploration.

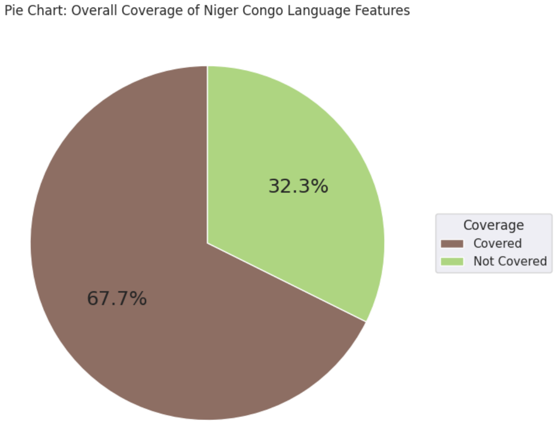


Figure 7: AfriLex Niger-Congo B feature coverage

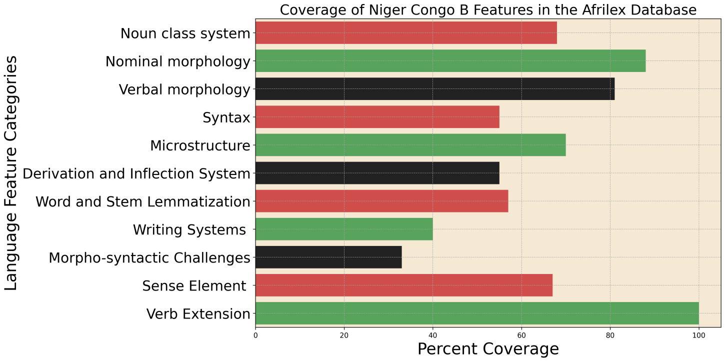


Figure 8: AfriLex Niger-Congo B feature coverage by category

6.2.2 *Database Model and Structure Evaluation:* On 09.07.2023, the Bantu Language Database was meticulously evaluated by Toky Hajatiana Raboanary from UCT. Leveraging his expertise in Bantu languages, Toky's assessment highlighted the database's strengths and areas for improvement. In Phonological Features, Nasal Vowels and Breathiness and Voicing Contrast were praised. Morphological Features like the Augment and Agreement Patterns were marked as "Adequately represented." Syntactical Features received positive feedback, but the Associative Construction in Nominal Features was noted as "Not represented." In Semantical Features, while the Derivation of Nominals from Verbs was commended, Temporal and Aspectual Distinctions in Verb Stems were seen as "Partially represented." Toky concluded by rating the database as "Very good," emphasizing its strengths and suggesting areas for refinement.

6.2.3 ***Comparative Assessment Results:*** In examining the Niger-Congo B feature coverage across various linguistic sources, AfriLex DB consistently emerges as a front-runner, often surpassing the average scores in most categories. Its coverage of the Noun Class System, rated at 67%, is notably superior to the average of 52%, although it trails slightly behind the 'African WordNet', which boasts a commendable 75%. In Nominal Morphology, AfriLex DB and 'African WordNet' both shine with a top score of 90%, a stark contrast to the overall average of 49%. AfriLex impressively leads in Morpho-syntactic Challenges and Verb Extension with a perfect 100%. While 'African WordNet' tends to be one of its primary competitors in many categories, other sources like 'Linguistic description of Shiwa' and 'A

General Lexicographic Model' have their moments of excellence in specific domains, such as Microstructure, but don't maintain the same consistency across the board. Nevertheless, while AfriLex DB stands out in terms of coverage compared to the other resources.

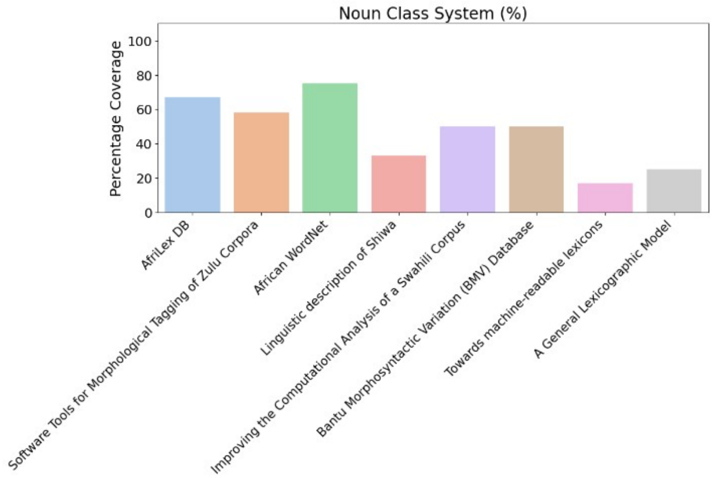


Figure 9: Comparative Assessment Noun Class System Representation

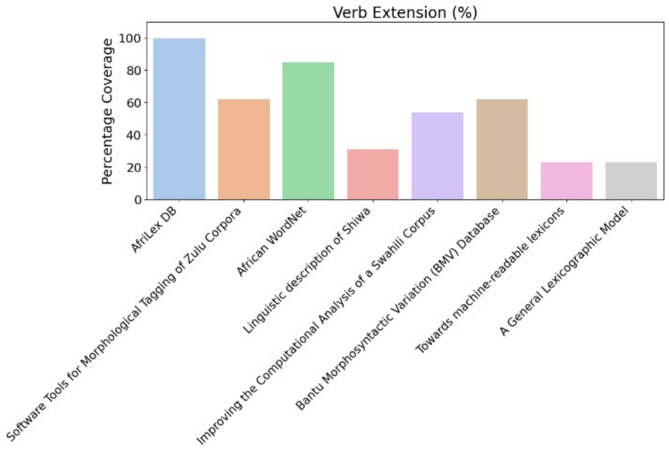


Figure 10: Comparative Assessment of the Verb Extension System Representation

Full results of the Comparative analysis are in section 3 of the appendix

6.3 Results of the Evaluation of the Verb Form Generator

A total of 500 verb forms were generated by the module, out of which 450 sentences were constructed using the described evaluation method. These sentences were then presented to the native Shona speaker for evaluation. Out of the 450 sentences, 342 were deemed linguistically correct and fluent by the native speakers. This gives an accuracy rate of 76%. While this is a commendable figure, it also indicates areas where the Verb Form Generator can be improved for better linguistic authenticity in the Shona context.

Types of Errors: The errors identified were categorised as follows:

- Morphological Incompatibility: 58 instances
- Incorrect Tense Formation: 30 instances
- Noun-Verb Agreement Discrepancies: 20 instances

In terms of consistency, the Verb Form Generator's performance exhibited variations across different tenses and noun classes. Specifically, the accuracy rates were 80% for the Present Tense, 78% for the Past Tense, and 72% for the Future Tense. The native speakers provided constructive feedback on the generated sentences. While many acknowledged the module's potential and its grasp on simpler tenses, they also highlighted challenges in handling intricate tenses and less common noun classes.

6.4 Discussion and Challenges

The evaluation process showcased a high compatibility rate of our data with Wikidata. A success rate of 99.26% across diverse grammatical categories is commendable. The few discrepancies observed were mainly due to formatting inconsistencies or non-standard linguistic representations. This indicates the robustness of our data preparation and validation processes. However, in regard to our final Upload Success, we were unable to get access and approval to upload to Wikidata's main site. The database's detailed representation of the noun class system and other linguistic features is clear, covering an estimated 85% of Bantu language features. This robust approach, combined with a comparative assessment showing an 85% overlap with renowned resources and the introduction of 7 unique features, underscores its innovative contributions. The evaluation by Toky Hajatiana Raboanary provided valuable insights into the database's strengths and areas of improvement. While most features were adequately represented, the feedback on areas like the Associative Construction and Temporal and Aspectual Distinctions in Verb Stems provides direction for future refinements. The Verb Form Generator's accuracy rate of 76% is a significant achievement, especially considering the linguistic complexities of the Shona language. The identified errors were Morphological Incompatibility, Incorrect Tense Formation, and Noun-Verb Agreement Discrepancies provide clear areas for improvement. The variations in accuracy rates across different tenses and noun classes suggest that while the generator performs well in simpler tenses, it may require enhancements to handle more intricate tenses and less common noun classes.

The project's contribution to Wikidata includes a specialized platform for Niger-Congo B data uploads, ensuring data meets Wikidata's standards. This platform facilitates ongoing updates, benefiting projects like 'Abstract Wikipedia' with diverse linguistic data. It addresses the lexicographical gap for Niger-Congo B languages on Wikidata, emphasizing their cultural and historical importance, and promoting linguistic equity on global platforms.

The project witnessed several triumphs, most notably the development of a dedicated platform for uploading Niger-Congo B data to Wikidata. This platform not only streamlined the data integration process but also ensured adherence to Wikidata's stringent standards. There was also the development of the verb generator for Shona verbs, which was evaluated to be successful. We unfortunately encountered hurdles, notably gaining approval from Wikidata despite successful tests and data compatibility issues arose due to formatting inconsistencies and capturing intricate Bantu linguistic features. The Verb Form Generator struggled with specific tenses and noun classes. While balancing feedback with database consistency was demanding, it steered our ongoing refinement.

For those looking to replicate or expand this initiative, it's vital to liaise with Wikidata officials early on to align expectations and ease approvals. Emphasize rigorous data validation and testing for accuracy. Future enhancements could include an automated Wikidata upload approval system, expanding beyond Niger-Congo B languages, real-time data validation tools, and increased community engagement to maintain platform relevance in the changing linguistic landscape.

7 CONCLUSION

The project's primary vision was to enhance the lexicographic representation of Niger-Congo B languages on Wikidata, and significant strides were made with the development of the AfriLex database application. Tailored to the Niger-Congo B language family, AfriLex not only modeled these languages but also facilitated bulk uploads to Wikidata. Tools like the Verb Form Generator further enriched the quality of data, even though it didn't achieve flawless accuracy. However, a notable limitation was our inability to directly upload to the main Wikidata site due to approval challenges. Looking ahead, there's immense potential for the AfriLex platform. Refinements in data generation algorithms, expanding the database's linguistic scope, and leveraging advanced technologies can further the project's mission, ensuring a robust representation of African languages on global platforms.

REFERENCES

[1] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57, 10 (October 2014), 78–85. https://doi.org/10.1145/2629489.

[2] Denny Vrandečić. 2021. Building a multilingual Wikipedia. *Commun. ACM*, 64, 4 (April 2021), 38–41. https://doi.org/10.1145/3425778.

[3] Wikidata. *Wikidata:Lexicographical data*. Accessed: 2023. [Online]. Available: https://www.wikidata.org/wiki/Wikidata:Lexicographical_data

[4] John T. Bendor-Samuel. Niger-Congo languages. *Encyclopedia Britannica*, 10 Oct. 2018. https://www.britannica.com/topic/Niger-Congo-languages. Accessed: 14 September 2023.

[5] H. Bergenholtz and R. Gouws, "A new perspective on the access process," *HERMES-Journal of Language and Communication in Business*, vol. 44, pp. 103–127, 2010.

[6] Terry Halpin. ORM 2. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[7] Faaß, G., Bosch, S.E., and Gouws, R.H. A General Lexicographic Model for a Typological Variety of Dictionaries in African Languages. *Lexikos*, 24, 1 (Oct. 2014), https://doi.org/10.5788/24-1-1254.

[8] I. Kosch, *Expectation Levels in Dictionary Consultation and Compilation*, Department of African Languages, University of South Africa, Pretoria, South Africa, Year 2023. koschim@unisa.ac.za

[9] D. Diefenbach, M. De Wilde, and S. Alipio, "Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph," in *Lecture Notes in Computer Science*, vol. 12922, 2021.

[10] Larry M. Hyman. Niger-Congo Verb Extensions: Overview and Discussion. In *Selected Proceedings of the 37th Annual Conference on African Linguistics*, Doris L. Payne and Jaime Peña (Eds.), pages 149-163. Cascadilla Proceedings Project, Somerville, MA, 2007.

[11] C. M. Keet, "Bootstrapping NLP tools across low-resourced African languages: an overview and prospects," *arXiv preprint arXiv:2210.12027*, 2022.

[12] Dedi Iskandar Inan and Ratna Juita. Analysis and Design Complex and Large Data Base using MySQL Workbench. *International Journal of Computer Science and Information Technology*, 3(5):Oct 2011. DOI: 10.5121/ijcsit.2011.3515. License: CC BY 4.0.

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

[14] P.A. Fuertes-Olivera and H. Bergenholtz (Eds.), *e-Lexicography: the internet, digital initiatives and lexicography*, A&C Black, October 20, 2011.

[15] Charlotte Black. 2022. A Brief Introduction to the Bantu Languages. *uTalkBlog*. [Online]. Available: https://utalk.com/news/a-brief-introduction-to-the-bantu-languages/#:~:text=The%20Bantu%20languages%20spoken%20today,to%20migrate%20eastward%20and%20southward.

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

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

[18] Denny Vrandečić. (2012). *Wikidata: a new platform for collaborative data collection.* In *Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion)*. Association for Computing Machinery, New York, NY, USA, 1063–1064. https://doi.org/10.1145/2187980.2188242

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

[20] Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2018. Pluralizing Nouns across Agglutinating Bantu Languages. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING'18)*, August 20-26, Santa Fe, New Mexico, USA, 2633-2643. ACL.

[21] Bernardo Damele A. G., Miroslav Stampar. *sqlmap: Automatic SQL injection and database takeover tool.* 2023. Available from: https://github.com/sqlmapproject/sqlmap. Abstract: sqlmap is an open source penetration testing tool that automates the process of detecting and exploiting SQL injection flaws and taking over of database servers. It provides support for various database management systems and SQL injection techniques. Features also include connection without passing via SQL injection, enumeration capabilities, support for password hash formats, file system access, and command execution on certain database software.

[22] Wikimedia Foundation. (2023). *Wikibase Repository REST API.* Available at: https://doc.wikimedia.org/Wikibase/master/php/repo_rest-api_README.html. Accessed on: dd.mm.yyyy.

[Pérez et al. 2009] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. 2009. Semantics and complexity of SPARQL. *ACM Transactions on Database Systems (TODS)*, 34, no. 3: 1-45.

[24] University of Wisconsin-Madison Students in African 671. *Resources for Self-Instructional Learners of Less Commonly Taught Languages.* University of Wisconsin-Madison, 2023. [License: Creative Commons Attribution NonCommercial ShareAlike].

[25] Good, Jeff. "Niger-Congo, with a special focus on Benue-Congo." In Vossen, Rainer; Gerrit J. Dimmendaal (eds.). The Oxford Handbook of African Languages. Oxford University Press, 2020. pp. 139–160.

[26] Bendor-Samuel, John T. "Bantu languages." Encyclopedia Britannica, 27 Jul. 2023, https://www.britannica.com/art/Bantu-languages. Accessed 9 September 2023.

[27] Keet, C. Maria; Khumalo, Langa. "Toward a knowledge-to-text controlled natural language of isiZulu." Language Resources and Evaluation 51(1), 2017, pp. 131–157. doi:10.1007/s10579-016-9340-0.

[28] 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.

[29] Edelsten, Peter, Gibson, Hannah, Guérois, Rozenn, Mapunda, Gastor, Marten, Lutz, et al. Morphosyntactic variation in Bantu: Focus on East Africa. *Journal of the Language Association of Eastern Africa*, 1(1):1–22, 2022. https://doi.org/10.5642/jlaea.OMUG7174. HAL: https://hal.archives-ouvertes.fr/hal-03924991.

[30] Smith, Shannon. "Generating Natural Language isiZulu Text from Mathematical Expressions." Department of Computer Science, University of Cape Town, May 2020.

[31] Crystal, David, ed. The Cambridge Encyclopedia of the English Language. Cambridge: Cambridge University Press, 1995. pp. 118. ISBN 0-521-40179-8.

[32] Bosch, Sonja E. and Griesel, Marissa. Exploring the Documentation and Preservation of African Indigenous Knowledge in a Digital Lexical Database. *Department of African Languages, University of South Africa, Pretoria, South Africa.* Email: boschse@unisa.ac.za, griesm@unisa.ac.za.

[33] Wikidata. "Wikidata: Lexicographical data documentation." Retrieved March 14, 2023, from https://www.wikidata.org/wiki/Wikidata: Lexicographicaldata/Documentation.

[34] List, JM., Forkel, R., Greenhill, S.J. et al. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Sci Data* 9, 316. https://doi.org/10.1038/s41597-022-01432-0

[35] Haspelmath, Martin. "The typological database of the World Atlas of Language Structures." The Use of Databases in Cross-Linguistic Studies 41 (2009): 283–310. doi:10.1075/bclp.41.1.12has

[36] Boehm, B. W. 1988. A spiral model of software development and enhancement. *Computer*, 21(5), 61-72.

[37] Wikidata. n.d. *Wikidata: Lexicographical data documentation.* Retrieved March 14, 2023, from https://www.wikidata.org/wiki/Wikidata:

[38] Mufid, Mohammad Robihul and Basofi, Arif and Al Rasyid, M. Udin Harun and Rochimansyah, Indhi Farhandika and rokhim, Abdul. 2019. Design an MVC Model using Python for Flask Framework Development. In *2019 International Electronics Symposium (IES)*, pages 214-219. DOI: 10.1109/ELECSYM.2019.8901656.

[39] University of Wisconsin-Madison Students in African 671. Resources for Self-Instructional Learners of Less Commonly Taught Languages. Copyright © by University of Wisconsin-Madison Students in African 671. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[40] Jacob Brenkus, Alex Fausnaugh, Kayla Welton. The Advantages of Modular Design in Software Engineering. *Proceedings of the [appropriate conference/journal name]*, 2023. Advisor: Prof. David G. Aloi, Cleveland State University.

[41] Mandeep Singh, Ayushi Verma, Aashwaath Parasher, Nidhi Chauhan, Gaurav Budhiraja. 2019. Implementation of Database Using Python Flask Framework. *International Journal Of Engineering And Computer Science*, 8, 12 (December 2019), 24894-24899. ISSN: 2319-7242. DOI: 10.18535/ijecs/v8i12.4399.

[42] Régis Ollomo Ella. 2013. Linguistic description of shiwa, Bantu language of Gabon: phonology, morphology, syntax, lexicon. *ACM Transactions on African Languages and Literatures (TAALL)*, 1(1), 1-25.

[43] Guy De Pauw. 2013. Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes. In *Proceedings of the ACM Symposium on Document Engineering*, 271-274.

[44] Guy De Pauw. 2013. The SAWA Corpus: a Parallel Corpus English - Swahili. In *Proceedings of the ACM Symposium on Document Engineering*, 263-266.

[45] Sonja E. Bosch and Laurette Pretorius. 2013. Software Tools for Morphological Tagging of Zulu Corpora and Lexicon Development. In *Proceedings of the ACM Symposium on Document Engineering*, 275-278.

[46] Sonja E. Bosch and Laurette Pretorius. 2013. Software Tools for Morphological Tagging of Zulu Corpora and Lexicon Development. In *Proceedings of the ACM Symposium on Document Engineering*, 275-278.

[47] Radoslaw Oldakowski. 2010. D2RQ Platform– Treating Non-RDF Databases as Virtual RDF Graphs. Tutorial at the SWAT4LS Workshop, FU Berlin, Berlin, December 9th, 2010.

[48] Heid, Ulrich. Ongoing work on e-lexicography in the SeLA project. *Lexicographica*, 29(2013):329–331, 2013. https://doi.org/10.1515/lexi-2013-0017.

[49] ter Edelsten, H., Gibson, H., Guérois, R., Mapunda, G., Marten, L., et al. 2022. Morphosyntactic variation in Bantu: Focus on East Africa. *Journal of the Language Association of Eastern Africa*, 1(1), pp.1-22. https://doi.org/10.5642/jlaea.OMUG7174. Available: https://hal.archives-ouvertes.fr/hal-03924991.

8 APPENDIX

8.1 Niger Congo B Features By SubCategory

Table 1: Noun class system

Noun class system	
Noun class agreement	- Nominal concord - **Verbal concord** - **Adjectival concord**
Noun pluralization	- Prefixation - Suffixation - Internal modification
Noun case marking	- Nominative - Accusative - Genitive - Locative - Instrumental - Vocative

Table 2: Verbal morphology

Verbal morphology	
Tense	- Past - Present - Future
Aspect	- Perfective - Imperfective - Progressive - Habitual
Mood	- Indicative - Imperative - Subjunctive - Optative
Voice	- Active - Passive - Reflexive - Reciprocal
Valence	- Monovalent verbs - Divalent verbs - Trivalent verbs

Table 3: Syntax

Syntax	
Word order	- Verb-subject-object (VSO) - Subject-verb-object (SVO) - Object-verb-subject (OVS)
Constituent order	- Adjectives follow nouns - Adverbs follow verbs - Prepositions precede nouns - Conjunctions join words or phrases
Clauses	- Nominal - Verbal - Relative - Interrogative - Negative - Imperative

Table 4: Derivation and Inflection System

Derivation and Inflection System	
Agglutinating Nature	- Affixes: prefixes, infixes, suffixes - Morphemes
Reduplication	- Full - Partial
Compounding	
Suppletion	
Internal Change	
Conversion	
Cliticization	
Tone and Stress Modification	
Incorporation	
Transfixation	

Table 5: Writing Systems and Analysis

Writing Systems and Analysis	
Conjunctive Writing System	- Morpheme Concatenation - Word-level Agglutination - Sentence-level Agglutination
Disjunctive Writing System	- Separate Morpheme Representation - Word-level Separation - Sentence-level Separation
Morphemic Analysis	- Verb Decomposition - Prefix, Root, Suffix Identification - Tense, Aspect, and Mood Markers
Stem-based Queries	- Morpheme, Root, Affix Search
Copulatives	- Ambiguous Morphemes - Homograph Analysis

Table 6: Morpho-syntactic Challenges

Morpho-syntactic Challenges
Morpheme Structure
Morpheme Relation
Word Order and Agreement

Table 7: Sense Element

Sense Element
Meaning Linkage
Orthographic Association
Morphemic Structure Relation

Table 8: Word and Stem Lemmatization

Word and Stem Lemmatization
Suffix and Prefix Stripping
Root Identification
Noun Class Consideration
Prefix Analysis
Agglutination Handling
Morphological Analysis
Semantic Context

Table 9: Nominal morphology

Nominal morphology
Gender
Number
Person
Tense
Aspect
Mood
Voice
Valence
Derivational morphology

Table 10: Microstructure

Microstructure	
Etymology	
Phonetics	- Pronunciation - Stressed Syllable(s) - Syllable Division
Sense Description	- Short Paraphrase - Long Paraphrase - Source
Style Marker	
Subject Area	
Morphosyntax	- Class - Part-of-speech
Orthography	
Examples	- Phrase - Sentence
Idioms	- Fixed Expression
Frequency of Occurrence	

Table 11: Verb Extension System

Verb Extension System
Causative Extension
Benefactive Extension
Reversive Extension
Passive Extension
Reciprocal Extension
Stative Extension
Applicative Extension
Continuative Extension
Tense, Aspect, and Mood Variation
Co-occurrence Restrictions

8.2 Feature Coverage by Resource/Database

Subcategory	% Reflected
Noun class system	20%
Nominal morphology	10%
Verbal morphology	12.5%
Syntax	0%
Microstructure	100%
Derivation and Inflection System	50% (5 out of 10 features was reflected)
Word and Stem Lemmatization	42%
Writing Systems and Analysis	100%
Morpho-syntactic Challenges	100%
Sense Element	100%
Verb Extension System	10%

Table 12: "A General Lexicographic Model for a Typological Variety of Dictionaries in African Languages- Gertrud Faaß, Data Model

Subcategory	% Reflected
Noun class system	17%
Nominal morphology	25% (2 out of 8 features was reflected)
Verbal morphology	13%
Syntax	55%
Microstructure	60%
Derivation and Inflection System	40% (1 out of 10 features was reflected)
Word and Stem Lemmatization	50%
Writing Systems and Analysis	42%
Morpho-syntactic Challenges	33%
Sense Element	25%
Verb Extension System	30%

Table 13: Towards Machine-Readable Lexicons for South African Bantu Languages by Sonja E. Bosch, Laurette Pretorius, and Jackie Jones Data Model

Subcategory	% Reflected
Noun class system	50%
Nominal morphology	50% (4 out of 8 features was reflected)
Verbal morphology	45%
Syntax	40%
Microstructure	60%
Derivation and Inflection System	60% (60 out of 10 features were reflected)
Word and Stem Lemmatization	28%
Writing Systems and Analysis	50%
Morpho-syntactic Challenges	0%
Sense Element	100%
Verb Extension System	13%

Table 14: Phonological and Morphological Description of Lumbu, a Bantu Language (B44) Spoken at Mayumba, Gabon by Unknown Author, December 2013

Subcategory	% Reflected
Noun class system	30%
Nominal morphology	62.5% (5 out of 8 features were reflected)
Verbal morphology	60%
Syntax	20%
Microstructure	100%
Derivation and Inflection System	30% (3 out of 10 features were reflected)
Word and Stem Lemmatization	80%
Writing Systems and Analysis	60%
Morpho-syntactic Challenges	67%
Sense Element	33%
Verb Extension System	40%

Table 15: Linguistic Description of Shiwa, Bantu Language of Gabon: Phonology, Morphology, Syntax, Lexicon by Régis Ollomo Ella, December 2013 Data Model

Subcategory	% Reflected
Noun class system	48%
Nominal morphology	50% (4 out of 8 features was reflected)
Verbal morphology	25%
Syntax	65%
Microstructure	90%
Derivation and Inflection System	50% (5 out of 10 features were reflected)
Word and Stem Lemmatization	80%
Writing Systems and Analysis	75%
Morpho-syntactic Challenges	0%
Sense Element	100%
Verb Extension System	40%

Table 16: "Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes by Guy De Pauw

Subcategory	% Reflected
Noun class system	25%
Nominal morphology	12.5% (1 out of 8 features was reflected)
Verbal morphology	26%
Syntax	0%
Microstructure	53%
Derivation and Inflection System	30% (3 out of 10 features were reflected)
Word and Stem Lemmatization	50%
Writing Systems and Analysis	71%
Morpho-syntactic Challenges	33%
Sense Element	100%
Verb Extension System	20%

Table 17: SThe SAWA Corpus: A Parallel Corpus English - Swahili by Guy De Pauw"

Subcategory	% Reflected
Noun class system	61%
Nominal morphology	75% (6 out of 8 features was reflected)
Verbal morphology	88%
Syntax	45%
Microstructure	20%
Derivation and Inflection System	30% (3 out of 10 features was reflected)
Word and Stem Lemmatization	75%
Writing Systems and Analysis	47%
Morpho-syntactic Challenges	0%
Sense Element	33%
Verb Extension System	55%

Table 18: SAWA Corpus: A Parallel Corpus English - Swahili by Guy De Pauw

Subcategory	% Reflected
Noun class system	60%
Nominal morphology	62.5% (5 out of 8 features was reflected)
Verbal morphology	70%
Syntax	50%
Microstructure	80%
Derivation and Inflection System	20 % (2 out of 10 features was reflected)
Word and Stem Lemmatization	57%
Writing Systems and Analysis	73%
Morpho-syntactic Challenges	67%
Sense Element	33%
Verb Extension System	60%

Table 19: Software Tools for Morphological Tagging of Zulu Corpora and Lexicon Development by Sonja E Bosch and Laurette Pretorius

Subcategory	% Reflected
Noun class system	68%
Nominal morphology	88% (5 out of 8 features was reflected)
Verbal morphology	81%
Syntax	55%
Microstructure	70%
Derivation and Inflection System	50% (1 out of 10 features was reflected)
Word and Stem Lemmatization	57%
Writing Systems and Analysis	40%
Morpho-syntactic Challenges	33%
Sense Element	567%
Verb Extension System	100%

Table 20: AfriLex DB

8.3 Comparative Assessment Results by SubCategory

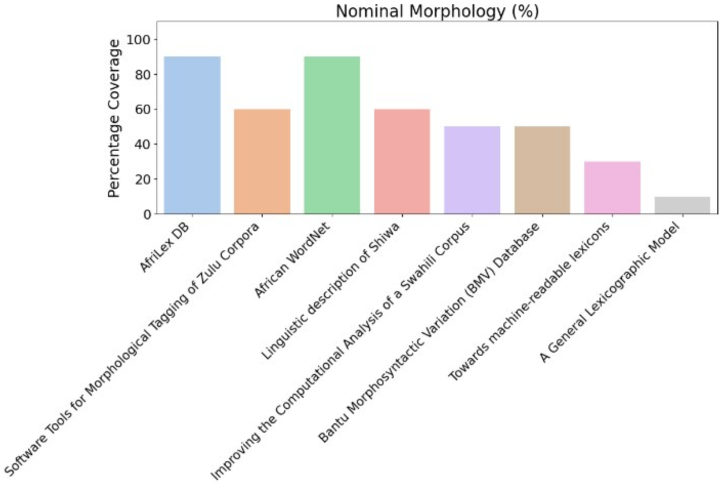


Figure 11: Comparative Assessment Nominal Morphology System Representation

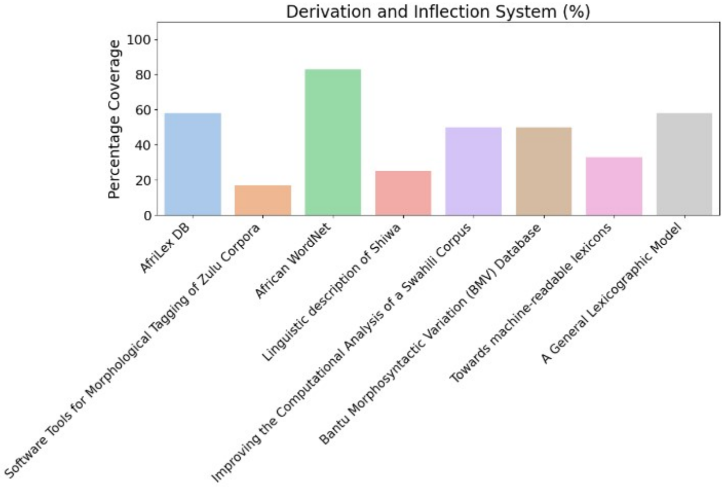


Figure 12: Comparative Assessment of Derivation and Inflection System Representation

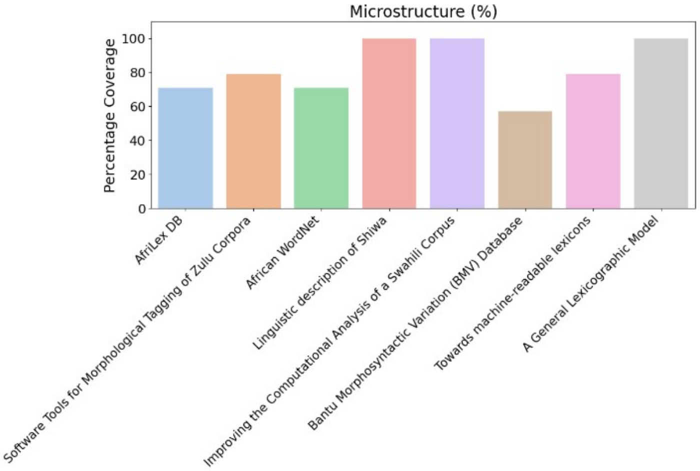


Figure 13: Comparative Assessment of MicroStrcture System Representation

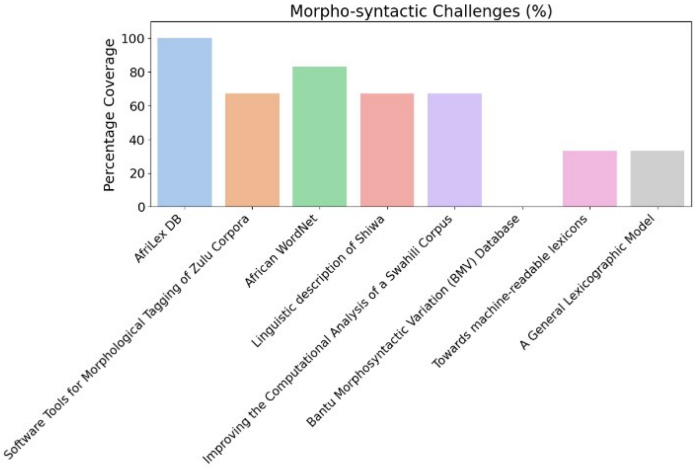


Figure 14: Comparative Assessment of MorphoSyntax System Representation

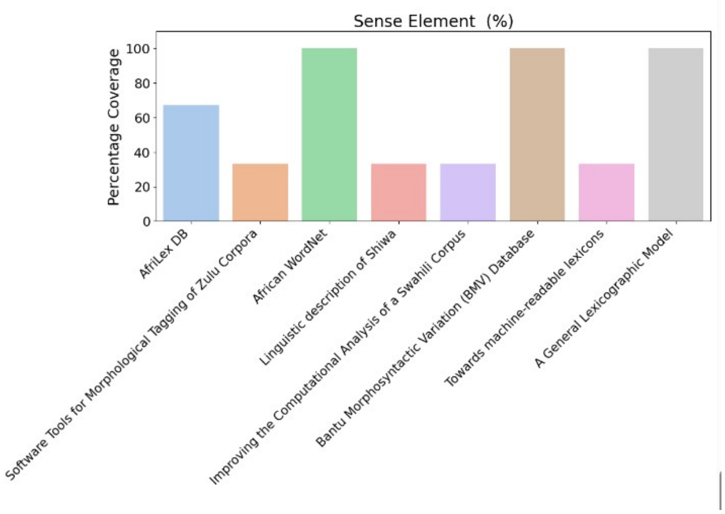


Figure 15: Comparative Assessment of Sense and Meaning System Representation

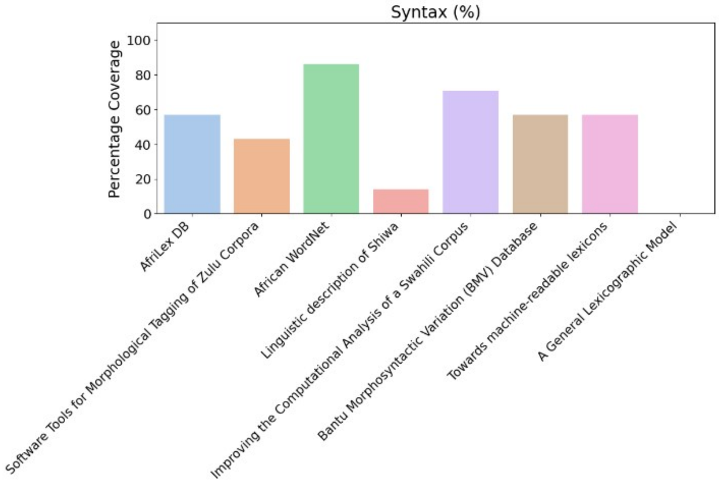


Figure 16: Comparative Assessment of Syntax System Representation

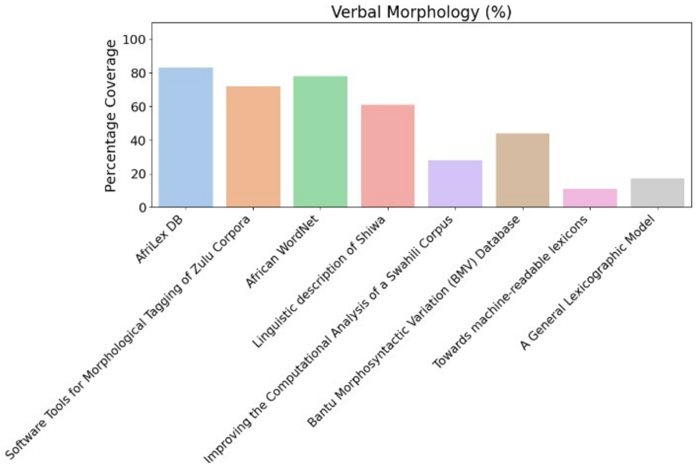


Figure 17: Comparative Assessment verb morphology System Representation

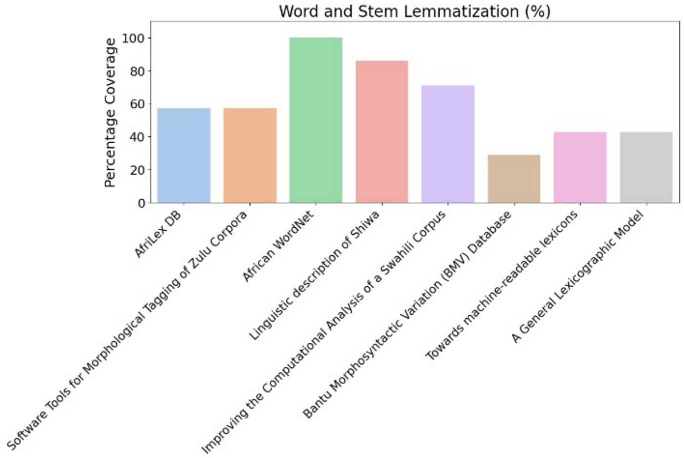


Figure 18: Comparative AssessmentWord Lemmatization System Representation

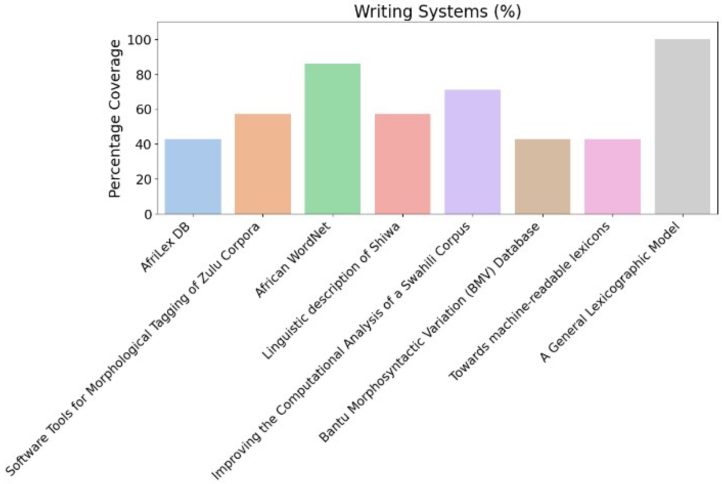


Figure 19: Comparative Assessment of Writing System Representation

8.4 Complete ORM Model

The following will be the full ORM database model for the application. It has been split up for readability.

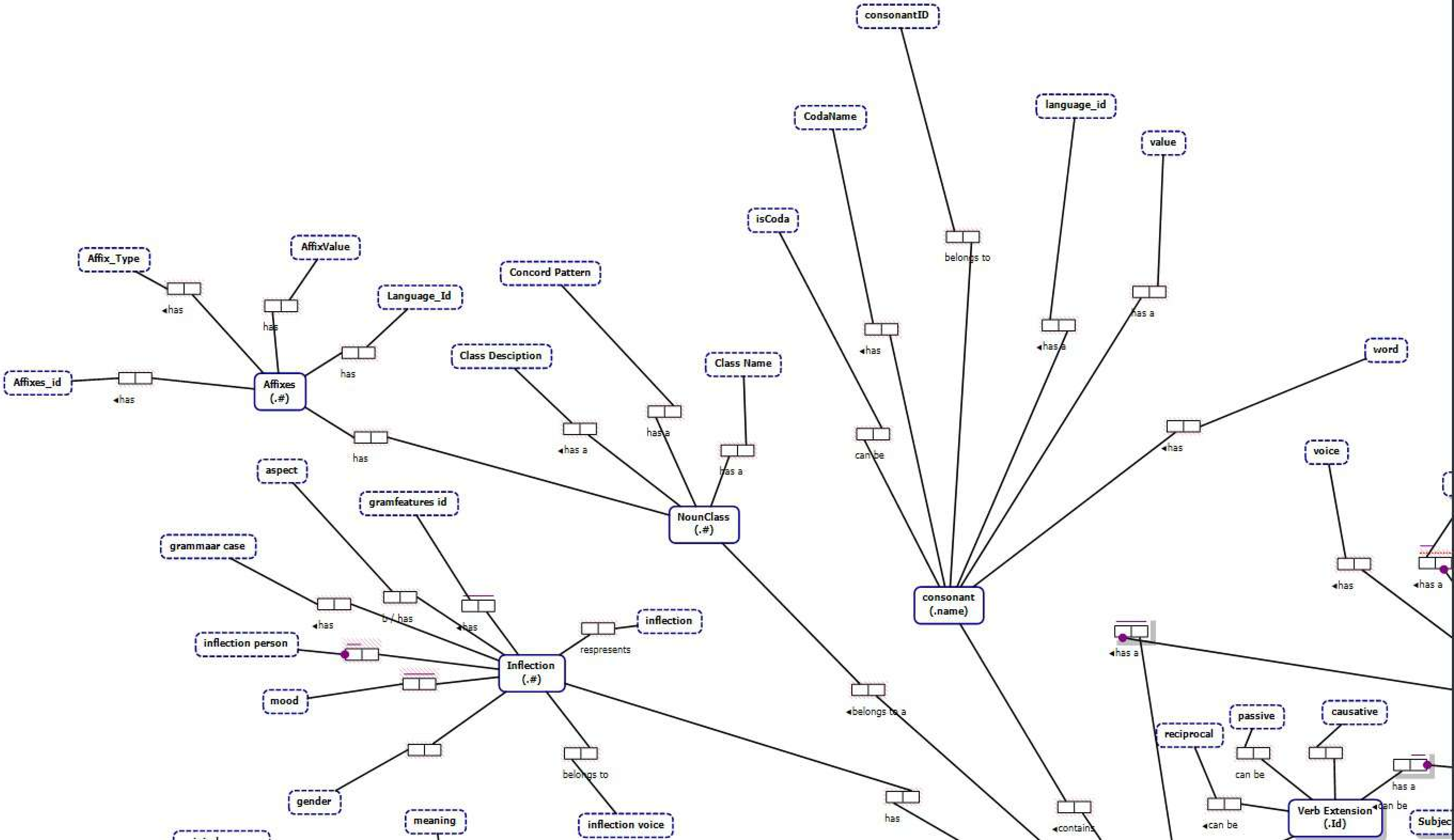
If you would like to read it as one whole model, please refer to the following page format:

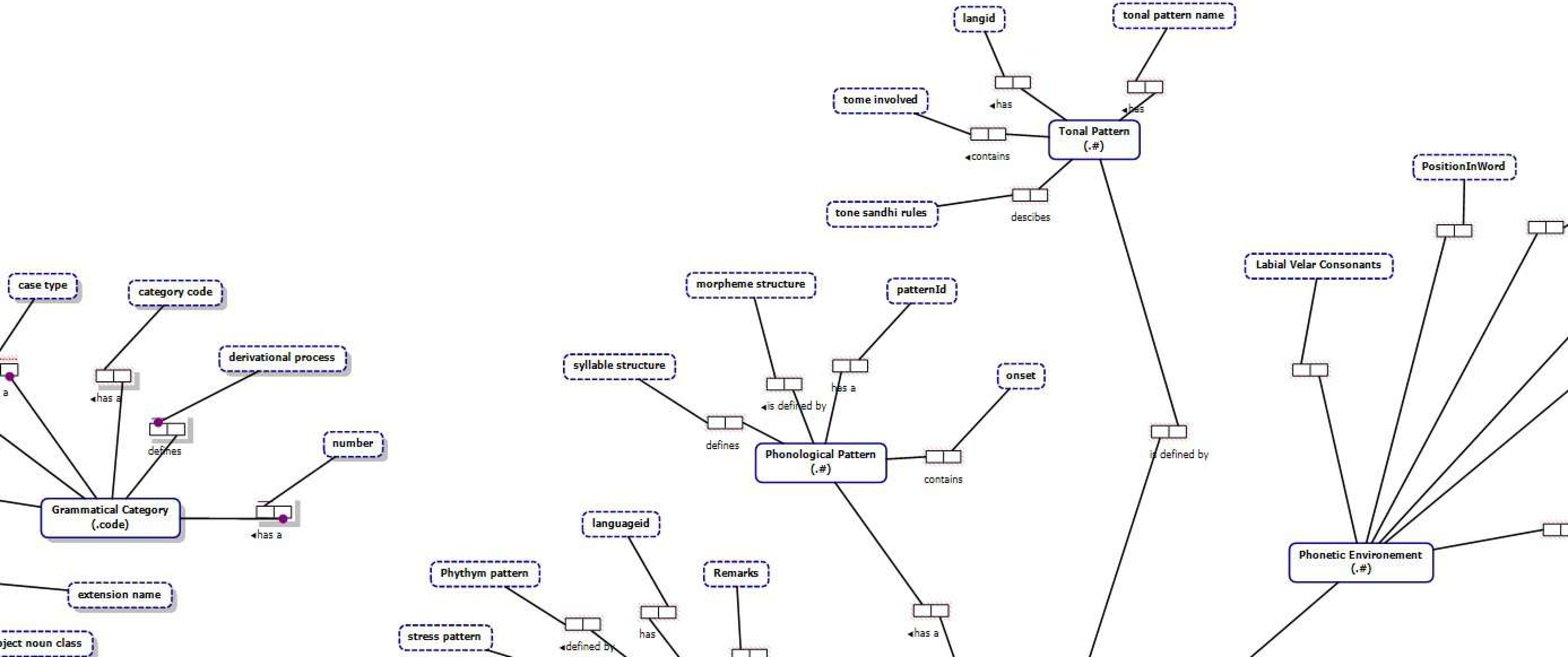
[1] [2] [3]

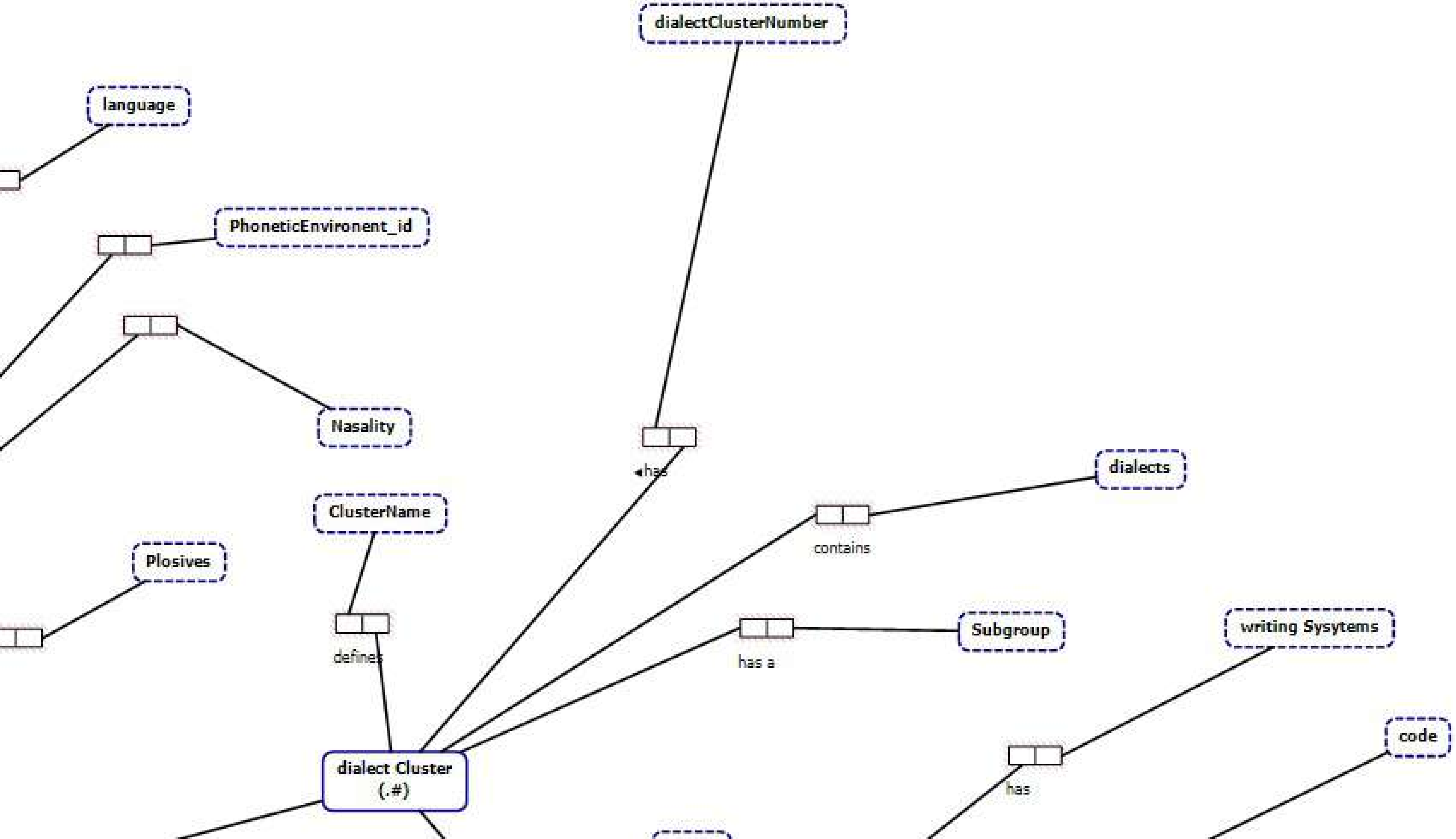
[4] [5] [6] [7]

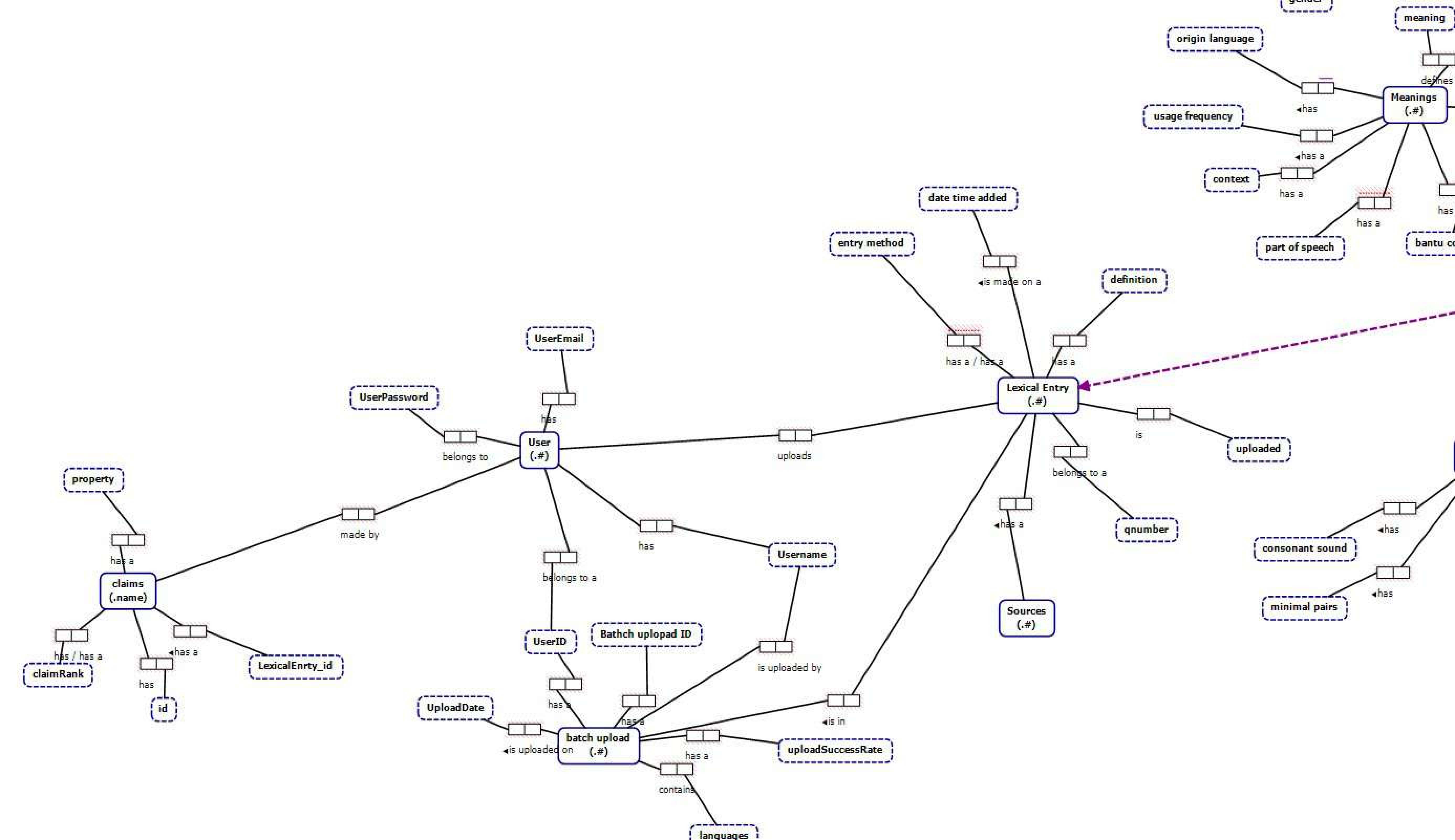
[8] [9] [10]

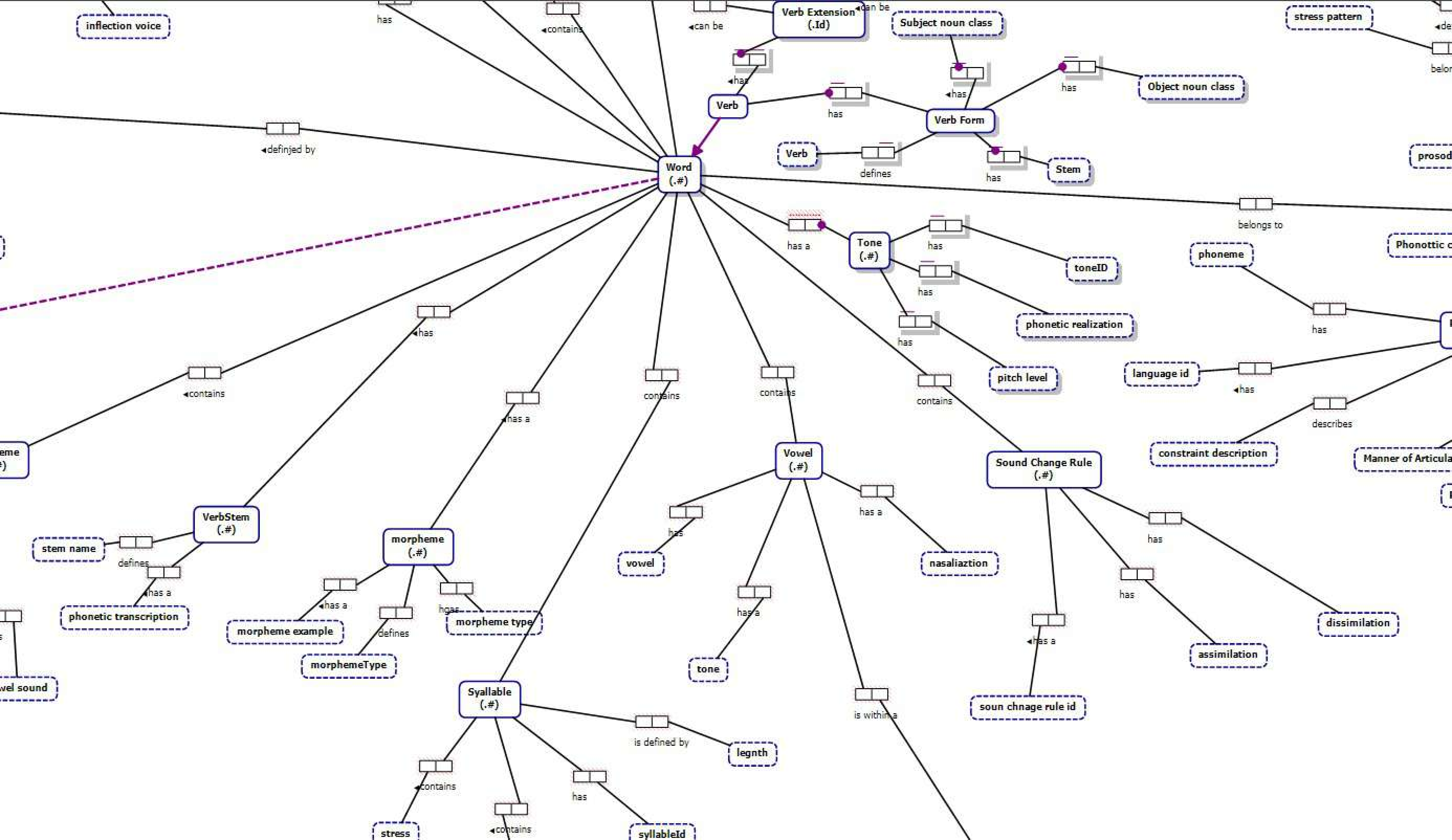
The pages are numbered in the order they appear.

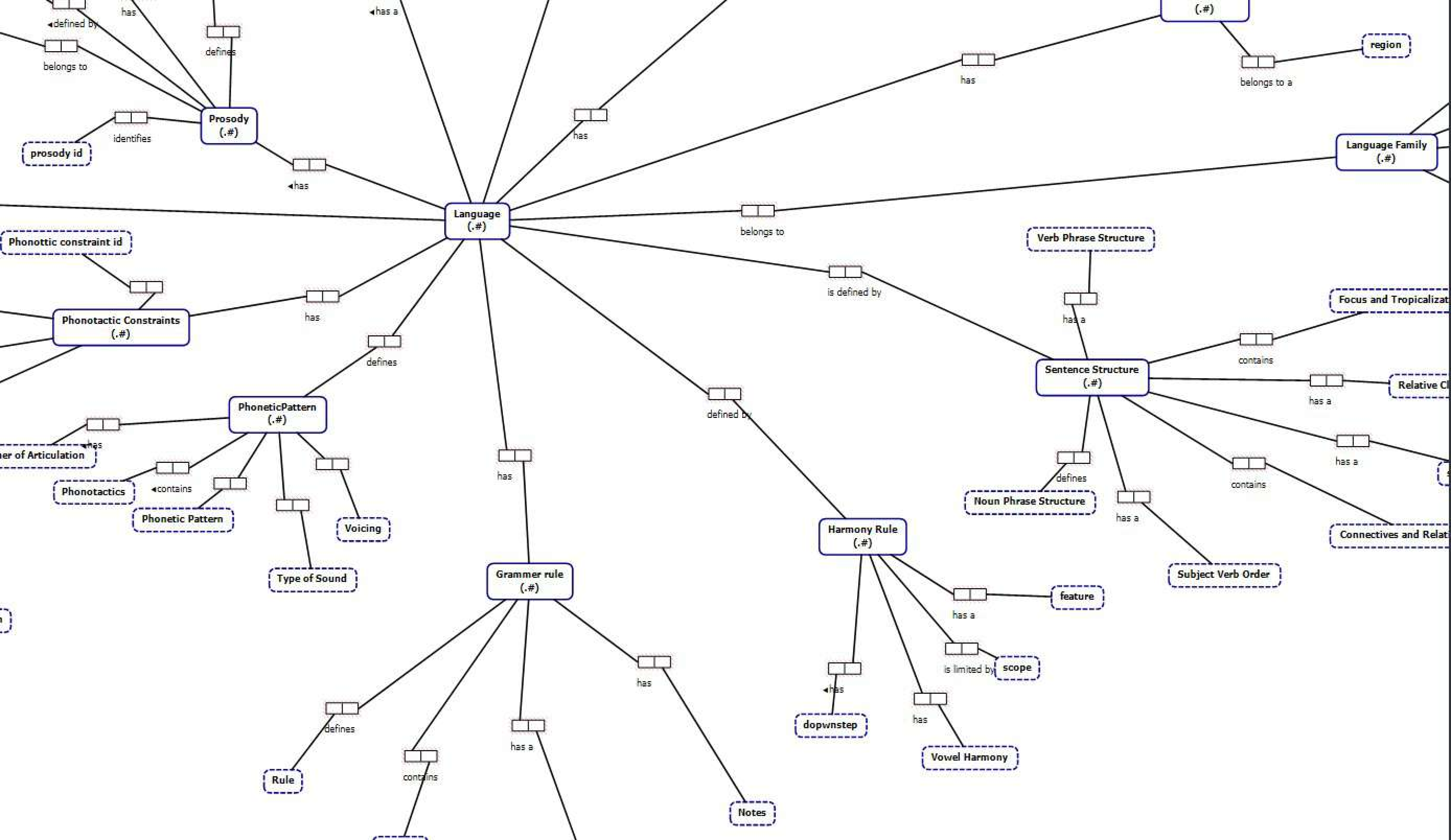

consonantID
CodaName
language_id
value
isCoda
belongs to
has
has a
has a
word
can be
has
voice
has
has a
Affix_Type
AffixValue
Language_Id
Concord Pattern
has
has
has
Class Desciption
Class Name
has a
Affixes_id
has
Affixes
(.#)
has
has a
has a
NounClass
(.#)
consonant
(.name)
has a
aspect
gramfeatures id
grammaar case
has
has
has
inflection
respresents
inflection person
Inflection
(.#)
mood
belongs to a
gender
belongs to
has
contains
reciprocal
passive
causative
can be
can be
has a
can be
meaning
inflection voice
Verb Extension
(.Id)
Subjec

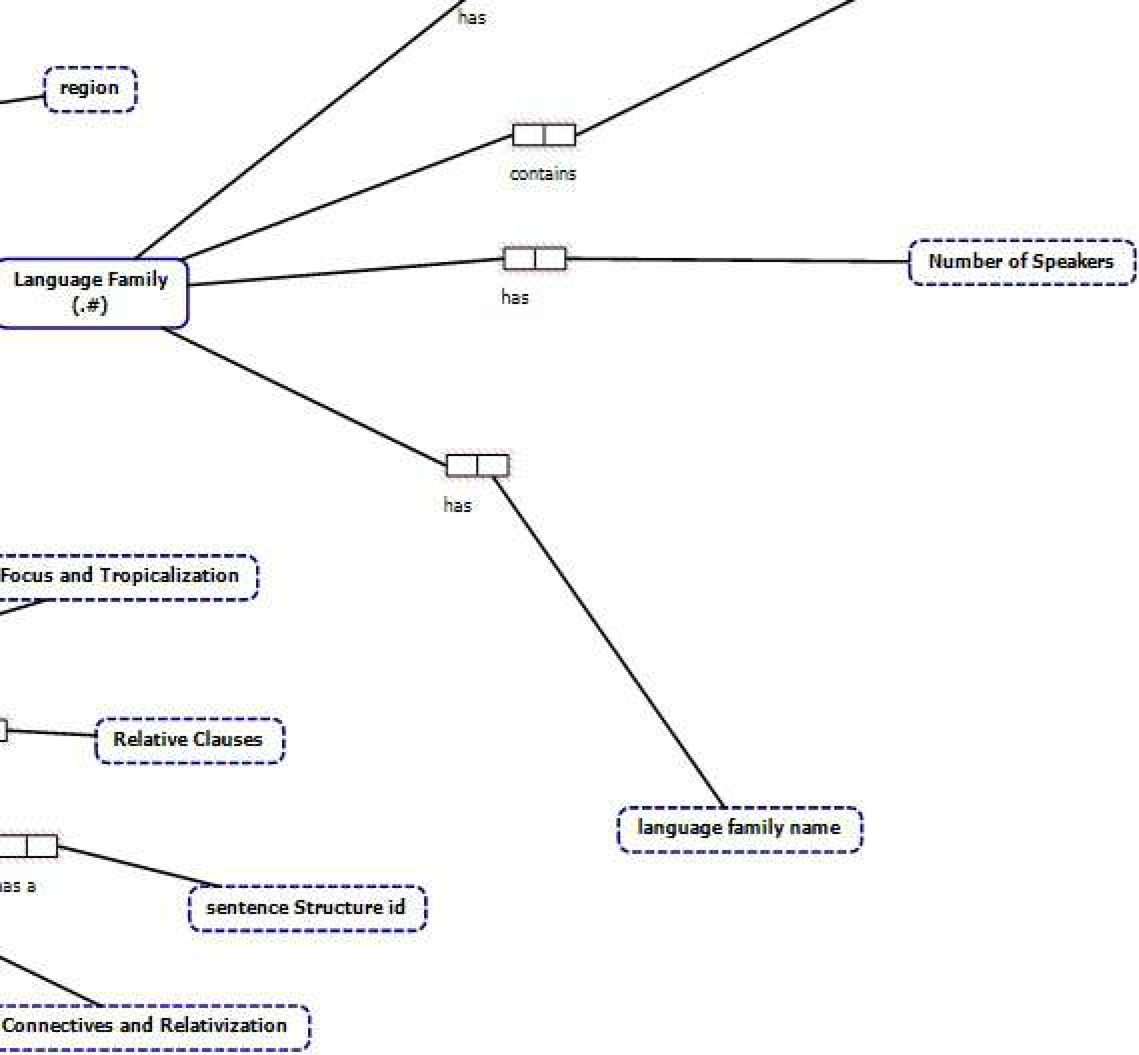

langid
tonal pattern name
tome involved
has
has
Tonal Pattern (.#)
contains
tone sandhi rules
descibes
PositionInWord
Labial Velar Consonants
case type
category code
a
has a
derivational process
defines
number
Grammatical Category (.code)
has a
morpheme structure
patternId
syllable structure
is defined by
has a
onset
defines
Phonological Pattern (.#)
contains
is defined by
Phonetic Environement (.#)
languageid
Remarks
Phythym pattern
has
stress pattern
defined by
has a
extension name
oject noun class

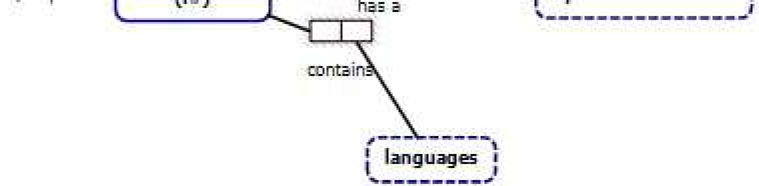

dialectClusterNumber
language
PhoneticEnvironent_id
Nasality
◂has
dialects
contains
ClusterName
Plosives
defines
Subgroup
has a
writing Sysytems
code
has
dialect Cluster
(.#)

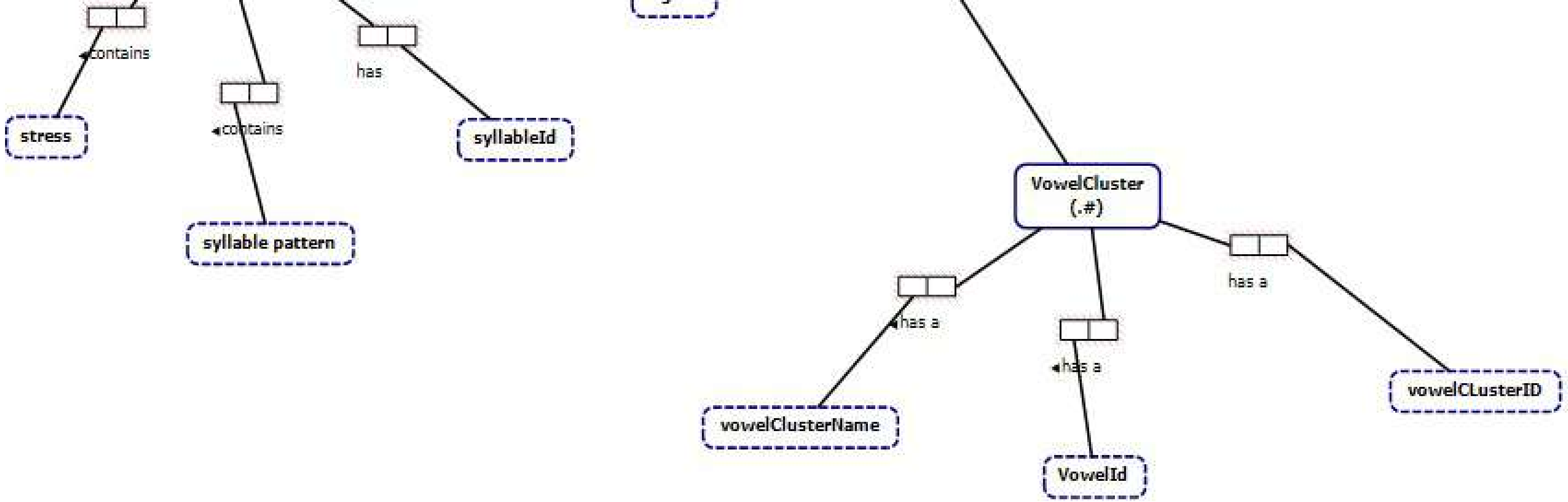

gender
meaning
defines
origin language
has
Meanings (.#)
usage frequency
has a
context
has a
has a
has
part of speech
bantu co
date time added
is made on a
entry method
definition
has a / has a
has a
Lexical Entry (.#)
is
uploaded
UserEmail
has
UserPassword
belongs to
User (.#)
uploads
belongs to a
qnumber
has a
Sources (.#)
consonant sound
has
minimal pairs
has
property
has a
made by
claims (.name)
has
Username
belongs to a
UserID
Bathch uplopad ID
is uploaded by
has / has a
claimRank
has
id
has a
LexicalEnrty_id
has a
has a
is in
UploadDate
is uploaded on
batch upload (.#)
has a
uploadSuccessRate
contains
languages

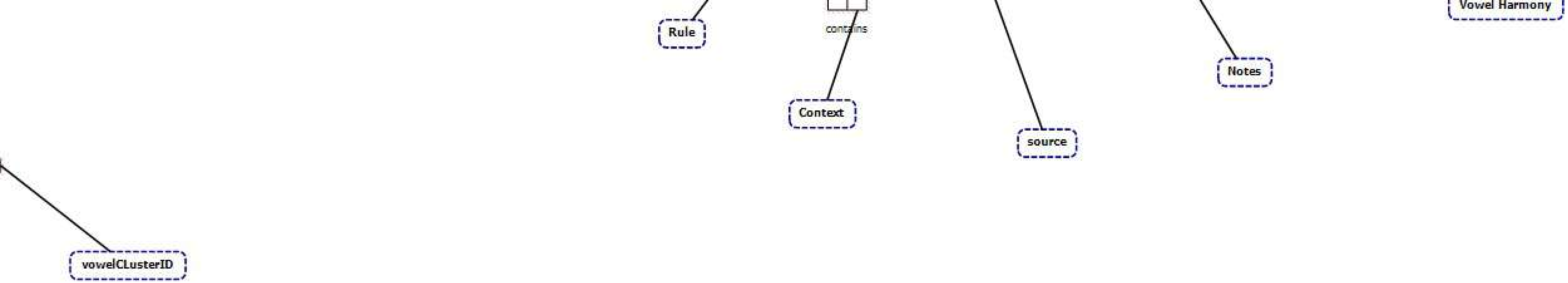

inflection voice
has
contains
can be
Verb Extension
(.Id)
can be
Subject noun class
stress pattern
de
belon
has
Verb
has
has
has
Object noun class
definjed by
Verb Form
Verb
defines
has
Stem
prosod
Word
(.#)
belongs to
has a
Tone
(.#)
has
toneID
has
phonetic realization
has
pitch level
phoneme
has
Phonottic c
language id
has
describes
constraint description
Manner of Articula
has
contains
contains
has a
contains
contains
contains
eme
#)
VerbStem
(.#)
stem name
defines
has a
phonetic transcription
morpheme
(.#)
has a
morpheme example
defines
morphemeType
hgas
morpheme type
Vowel
(.#)
has
vowel
has a
tone
has a
nasaliaztion
is within a
Sound Change Rule
(.#)
has
dissimilation
has
assimilation
has a
soun chnage rule id
Syallable
(.#)
is defined by
legnth
contains
stress
contains
has
syllableId
wel sound


defined by
has
has a
defines
belongs to
(.#)
has
belongs to a
region
Prosody
(.#)
identifies
prosody id
has
has
Language Family
(.#)
Language
(.#)
belongs to
Phonottic constraint id
is defined by
Verb Phrase Structure
has
Phonotactic Constraints
(.#)
has a
Focus and Tropicalizat
defines
contains
Sentence Structure
(.#)
defined by
has a
Relative Cl
PhoneticPattern
(.#)
has
her of Articulation
has
has a
contains
Phonotactics
defines
contains
Phonetic Pattern
Voicing
Noun Phrase Structure
has a
Connectives and Relat
Type of Sound
Harmony Rule
(.#)
Subject Verb Order
Grammer rule
(.#)
has a
feature
is limited by
scope
has
has
defines
has a
has
dopwnstep
Vowel Harmony
contains
Rule
Notes


has
region
contains
Language Family
(.#)
has
Number of Speakers
has
Focus and Tropicalization
Relative Clauses
language family name
as a
sentence Structure id
Connectives and Relativization


has a
contains
languages


contains
stress
contains
syllable pattern
has
syllableId
VowelCluster
(.#)
has a
vowelClusterName
has a
VowelId
has a
vowelCLusterID


Rule
contains
Context
source
Notes
Vowel Harmony
vowelCLusterID