Developing Bilingual English-Setswana Datasets for Space Domain

Tebatso G. Moape[1], **Sunday O. Ojo**[2], **Oludayo O. Olugbara**[3]
University of South Africa[1], Durban University of Technology[2,3]
28 Pioneer Ave, Florida Park, 1704, South Africa[1],
43 M L Sultan Rd, Greyville, Durban, 4001, South Africa[2,3]
moapetg@unisa.ac.za[1], sundayo1@dut.ac.za[2], oludayoo@dut.ac.za[3]

Abstract

In the current digital age, languages lacking digital presence face an imminent risk of extinction. In addition, the absence of digital resources poses a significant obstacle to the development of Natural Language Processing (NLP) applications for such languages. Therefore, the development of digital language resources contributes to the preservation of these languages and enables application development. This paper contributes to the ongoing efforts of developing language resources for South African languages with a specific focus on Setswana and presents a new English-Setswana bilingual dataset that focuses on the space domain. The dataset was constructed using the expansion method. A subset of space domain English synsets from Princeton WordNet was professionally translated to Setswana. The initial submission of translations demonstrated an accuracy rate of 99% before validation. After validation, continuous revisions and discussions between translators and validators resulted in a unanimous agreement, ultimately achieving a 100% accuracy rate. The final version of the resource was converted into an XML format due to its machine-readable framework, providing a structured hierarchy for the organization of linguistic data.

Keywords: Digital language resources, Setswana bilingual dataset, Space domain translation

1. Introduction

The Princeton Wordnet (PWN) is an English lexical database formally introduced by Miller (1995) and developed at Princeton University. It has served as the primary lexical semantic resource for numerous researchers in the field of Natural Language Processing (NLP) and computational linguistics (Batsuren et al., 2019). The presence of this resource has facilitated the development of various NLP applications such as machine translation, information retrieval, and tools for word sense disambiguation. Additionally, the availability of PWN has provided researchers with the capability to evaluate and compare the effectiveness of various language models and applications.

However, languages such as Setswana face a scarcity of resources (Sebolela, 2009, Marivate et al., 2020), resulting in limited availability of linguistic tools and applications. Furthermore, the current tools are frequently created within isolated projects, each with curated data tailored to its particular scope. This fragmentation poses a challenge for researchers in effectively collaborating and comparing their work.

Apart from unannotated parallel-aligned corpora and word list dictionaries extracted from government websites, the only available resource comparable to the PWN is the African Wordnet (AWN). The AWN project was initiated with the aim of promoting multilingualism and facilitating the development of language tools and resources for South African (SA) languages (Bosch and Griesel, 2017). Currently, wordnets have been developed for Setswana, isiXhosa, isiZulu, Sesotho sa Leboa, and Tshivenda. The AWN holds significance due to the scarcity of data in South African languages. This makes the AFW a crucial resource.

In an effort to contribute to the development of resources for SA languages in general and Setswana in particular, this paper presents a Setswana lexicon. The lexicon was developed by translating a subset of the PWN through expert translation, expansion, and domain adaptation methods. The chosen domain for the translation focused on space-related concepts. The outcome of this project is a Setswana lexicon comprising of 6016 synsets, with lemmas, glosses, and usage examples. The use of expert-driven translation was to ensure the generation of high-quality translations, and the decision to focus on a specific domain was made to enable the Setswana lexicon's relevance and applicability in the targeted context.

The rest of the paper is structured as follows: section 2 presents relevant literature related to the development of semantic resources across languages and the state of the art of the available language resources for Setswana. Section 3 outlines the techniques and methodologies used for the resource development. Resource evaluation and results are presented in section 4. Section 5 concludes the paper.

2. Related Works

This section is divided into two subsections. The first sub-section focuses on relevant literature related to the development of semantic resources across languages. This literature provides a foundation and context for the methodology employed in developing the resource presented in this paper. The second sub-section provides a

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 32–36
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

high-level overview of available resources for Setswana.

2.1 Development of Language Resources

Monolingual lexicons are constructed using two methods, namely, the expansion method and the merge method (Bosch and Griesel, 2018). In the expansion method, developers translate a subset of English synsets from PWN. The merge method involves the creation of synsets for the target languages, which are then merged with PWN synsets. The key distinction between the two methods is that the expansion method results in the target language inheriting PWN's semantic structure, while the merge method entails the creation of a new semantic structure for the target language, which is subsequently merged with PWN's semantic network. Resources conducted using these methods include (Batsuren et al., 2019, Bella et al., 2020).

In focusing on the space domain, this study used the expansion method for Setswana. The rationale behind this choice stems from the existence of similar field concepts within both English and Setswana space domains. To ensure that all the Setswana space concepts were included, concepts present in Setswana but absent from the translated set were added to the dataset and subsequently lexicalized into English. This guarantees a comprehensive coverage of space-related concepts in Setswana.

2.2 State of the Art on Setswana Resources

The importance of the availability of language resources cannot be overstated, as they play a crucial role in the preservation of languages and serve as an enabler for the advancement and development of NLP tools. In efforts to create, consolidate, and disseminate language resources for diverse SA languages, the South African Centre for Digital Language Resources (SADiLaR), in collaboration with various universities was established for this purpose. Supported by the Department of Science and Innovation (DSI), SADiLaR plays a significant role in facilitating the centralization of these language resources, contributing to their accessibility and use. This section outlines the text resources accessible for Setswana on SADiLaR, providing an overview of the presently available resources accessible to researchers.

In summary, currently, including the AFW (Bosch and Griesel, 2017), there is a total of 24 various types of text corpora. This includes multilingual word and phrase translations, phrase chunk annotated corpus, monolingual corpora, test suite, data treebank, named entity annotated corpus, annotated text corpora, and English-Setswana parallel corpora (Lastrucci et al., 2023, McKellar and Puttkammer, 2020). There is also domain-based data where English data from specific domains were translated into multiple SA languages, including Setswana. This encompasses data from domains such as soccer, mathematics, technology, health, natural sciences, arts and culture, government, elections, and parliamentary proceedings. The dataset presented in this paper specifically falls within the space domain, further expanding the scope of available data resources for the Setswana language.

3. Methodology

This paper adopted the expert-sourced expand approach to develop the presented resource. A subset of words, glosses, and examples from the PWN were translated and validated by Setswana language experts. The methodology consists of four phases, namely, translation data generation, translation, reformatting, and validation. The translations were conducted on a Microsoft Excel Spreadsheet. The following sub-sections expand on the structure of the spreadsheet, data generation, translation, validation, and reformatting phases.

3.1 Translation via Microsoft Excel Spreadsheets

The Microsoft Excel Spreadsheet contains source and target lemmas, synsets, glosses, and examples. These are defined as:

3.1.1 Lemma

A lemma is the canonical form (dictionary form, citation form) of a set of words (word forms). For example, *tsamaya (go)* is the lemma of the words *tsamaya (go)*, *wa tsamaya(goes)*, and *o tsamaile (went)*.

3.1.2 Synsets

A synset is a set of synonyms that represent a single concept or idea in linguistics which consists of lemmas. Each synset represents a unique concept, and words within the same synset are considered synonymous with one another. Synsets provide a way to organize and understand the relationships between words and their meanings in a structured format.

3.1.3 Gloss

A text or sentence that describes the concept, i.e., a lemma.

3.1.4 Example

A text or sentence(s) that clarify the exact meaning of the described concept. Examples are also used to clarify and demonstrate how the lemma/concept is used in a sentence.

3.2 Translation Data Generation

For the translation data generation phase, a domain-specific adaptation method was used.

This method focuses on the creation of language resources based on a specific domain or subject area. The following criteria were used when selecting the subset of English synsets to be translated.

3.2.1 Domain Identification

The space domain dataset was selected.

3.2.2 Data Extraction

Lemmas, glosses, and examples were extracted and transferred to an Excel file.

Wordnet data is normally divided into four parts of speech categories, nouns, verbs, adjectives, and adverbs. To narrow the focus, this study focused on data in these four categories that are in the space domain.

3.3 Translation

The translations were conducted on a Microsoft Excel file. The Excel sheet contains a number of synsets in the source language to be translated into the target language. In this case, English was the source language, and Setswana was the target language. The file is organized in a pair-wise format (source language column - target language column). The translator fields consist of the following:

3.3.1 Synset lemmas columns

Column C: Contains a comma-separated list of lemmas of the source language.

Column D: The translator provides a comma-separated list of the synset lemmas of the target language.

3.3.2 Synset gloss columns

Column F: Contains the synset gloss in the source language.

Column G: The translator provides the synset gloss in the target language.

3.3.3 Synset examples columns

Column I: Contains the synset examples in the source language.

Column J: The translator provides the synset examples in the target language.

3.3.4 Translator notes columns

Column L: The translator can provide notes related to the synset translation if there are any.

3.4 Validation

The same Excel sheet used for translation was used for validation. The validator fields consist of the following:

3.4.1 Synset lemmas validation

The validator provides his validation on the lemmas in the target language in column E. The validator can choose between:

- Accepted: If the validator finds that the lemmas are complete and do not contain any errors such as spelling errors, she/he writes A (for accepted).

- Rejected: If the validator finds that the lemmas are not correct, or there are missing lemmas, or lemmas that do not belong to the synset, she/he writes R (for rejected) and provides justification for the decision in the validator notes column.

3.4.2 Synset gloss validation

The validator provides his validation on the synset gloss column H. The validator can choose between :

- Accepted: If the validator finds that the synset gloss describes the synset correctly and does not contain errors such as spell errors, she/he writes A.

- Rejected: If the validator finds that the synset gloss does not describe the synset or it contains errors, she/he writes R and provides justification.

3.4.3 Synset examples validation

The validator provides his decision on the synset examples in Column K. The validator can choose between:

- Accepted: If the validator finds that the synset examples are correct and they do not contain errors, and if there are no synset examples that may be necessary to describe how to use the lemmas, she/he writes here A. It is possible to accept synsets without examples if the translator did not provide them and the validator accepts the translator decision.

- Rejected: If the validator did not provide examples and the validator does not accept the translator's decision, or if he finds errors in the examples, she/he writes here R and provides justification.

3.4.4 Validator notes validation

The validator provides justification for his rejection of any of the previous synset translations in this column. Validator comments are optional in case of acceptance, but they are mandatory in case of rejection.

In cases where translations were rejected by the validator, the Excel sheet was returned back to the translator with the validator's notes for clarification. The identified mistakes were corrected, or the translators provided reasons for the chosen translations. This process continued until the translators and validators reached an

agreement, and all translations were accepted, making this a high-quality resource.

3.5 Reformatting

The developed resource can be used for the development of NLP applications. However, data in an Excel format is not suitable for programming Integrated Development Environments (IDEs) and computational linguistics fields where these applications are developed (Suárez et al., 2007). This is due to its memory-intensive nature which results in inefficiency. The use of such data in IDEs could lead to increased memory consumption, longer execution time, and reduced performance, making it less than ideal for application development.

To address these limitations, the developed resource was reformatted to Extensible Markup Language (XML) format. XML format is a widely accepted standard for representing structured data (Bourret, 1999). Its standardization ensures consistency and compatibility across different development software applications and platforms. Furthermore, XML provides a machine-readable framework that allows the representation of linguistic data in a hierarchical order and the inclusion of metadata and semantic annotations (Kroeze et al., 2010). The data was grouped according to the parts of speech and converted to XML files.

4. Evaluation and Results

Our validation method explicitly and formally evaluated individual lemma, examples and definitions translations, and their quality. The evaluation was carried out by a group of native Setswana speakers who are proficient in English. They determined the validity of translations by marking them as "accept" if accurate and "reject" if incorrect or lacking in translation equivalents. To calculate the accuracy of the translations, the following metric was used to measure accuracy (A). This is calculated by dividing the number of correct translations i.e. “accept” by the number of total translations using the following equation:

A= (correct translations)/(total number of translations) * 100
A= 5981/6016
A=0.99

There were 5981 synsets correctly translated out of a total of 6016 translated synsets, thus substituting these values into the equation above. The initial submission achieved a 99% accuracy rate before undergoing validation. After validation, continuous revisions and discussions between translators and validators resulted in a unanimous agreement, ultimately achieving a 100% accuracy rate. The presented resource in this study consists of the following translated lexicon in Table 1.

File	Synsets	Number of Words
Setswana-nouns-1	1004	15970
Setswana-nouns-2	1001	15724
Setswana-verbs-1	1001	15643
Setswana-verbs-2	1006	15595
Setswana-adjectives-1	1009	11005
Setswana-adjectives-2	1001	19393
Total number	**6016**	**93330**

Table 1: Translated lexicon statistics.

The number of the lemmas, glosses, and examples are the same for both Setswana and English as all the source data in all rows of all the files were translated.

5. Conclusion

5.1 Bibliographical References

This paper presented a new English-Setswana bilingual dataset that was professionally translated with a specific focus on the space domain. The dataset was developed using the expansion approach, involving the translation of a subset of synsets from PWN into Setswana. The translation process was undertaken by professional Setswana translators specifically contracted for this task. Following the translation process, native Setswana speakers, who also possessed proficiency in English, validated the translated content.

For validation, iterative assessments and discussions between translators and validators confirmed the accuracy of all translations, achieving a 100% accuracy rate. The current funding covered the translation of specified files presented in this paper within this domain, there are still pending files to be translated. Our future endeavours entail securing further funding to significantly enhance the dataset. Additionally, as the current translation process was manual, we aim to semi-automate both translation and validation procedures by leveraging computer-aided translation software. Once completed, the dataset will be openly accessible to researchers for application in linguistic and NLP research.

6. Bibliographical References

Batsuren, K., Ganbold, A., Chagnaa, A. & Giunchiglia, F. Building the mongolian wordnet. Proceedings of the 10th global WordNet conference, 2019. 238-244.

Bella, G., Mcneill, F., Gorman, R., Donnaíle, C. Ó., Macdonald, K., Chandrashekar, Y., Freihat, A. A. & Giunchiglia, F. A major wordnet for a minority language: Scottish gaelic. 12th Language Resources and Evaluation Conference, 2020. European Language Resources Association (ELRA), 2812-2818.

Bosch, S. & Griesel, M. African Wordnet: facilitating language learning in African languages. Proceedings of the 9th Global Wordnet Conference, 2018. 306-313.

Bosch, S. E. & Griesel, M. 2017. Strategies for building wordnets for under-resourced languages: The case of African languages. Literator (Potchefstroom. Online), 38, 1-12.

Bourret, R. 1999. Xml And Databases.

Kroeze, J. H., Bothma, T. J. D. & Matthee, M. C. 2010. Constructing An Xml Database Of Linguistics Data. Td: The Journal For Transdisciplinary Research In Southern Africa, 6, 139-174.

Lastrucci, R., Dzingirai, I., Rajab, J., Madodonga, A., Shingange, M., Njini, D. & Marivate, V. 2023. Preparing The Vuk'uzenzele And Za-Gov-Multilingual South African Multilingual Corpora. Arxiv Preprint Arxiv:2303.03750.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R. & Modupe, A. 2020. Low Resource Language Dataset Creation, Curation And Classification: Setswana And Sepedi. Arxiv Preprint Arxiv:2004.13842.

Mckellar, C. A. & Puttkammer, M. J. 2020. Dataset For Comparable Evaluation Of Machine Translation Between 11 South African Languages. Data In Brief, 29, 105146.

Miller, G. A. 1995. Wordnet: A Lexical Database For English. Communications Of The Acm, 38, 39-41.

Sebolela, F. 2009. The Compilation Of Corpus-Based Setswana Dictionaries. University Of Pretoria.

Suárez, O. S., Riudavets, F. J. C., Figueroa, Z. H. & Cabrera, A. C. G. 2007. Integration Of An Xml Electronic Dictionary With Linguistic Tools For Natural Language Processing. Information Processing & Management, 43, 946-957.

Superman, S., Batman, B., Catwoman, C., and Spiderman, S. (2000). *Superheroes experiences with books*. The Phantom Editors Associates, Gotham City, 20th edition.

7. Language Resource References

Bosch, Sonja and Griesel, Marissa. 2017. Strategies for building wordnets for under-resourced languages: the case of African languages