

A proof of concept wordnet design for South African Sign Language (SASL)

Prepared by Marissa Griesel

SADiLaR

June 2026

Table of contents

1. Introduction.....	3
2. Background.....	3
2.1 What is a wordnet?.....	3
2.2 Popular methods for developing wordnets.....	4
2.2.1 Expand approach.....	4
2.2.2 Merge approach.....	5
2.2.3. Other approaches and techniques.....	6
Dictionary-based construction.....	6
Corpus-based construction.....	6
Corpus-/data-driven synset induction.....	7
2.3. Related work.....	7
2.3.1 The African Wordnet Project.....	7
2.3.2 The Open Multilingual Signed Wordnet.....	7
3. A proposed design for SASL.....	8
4. Next steps.....	10

A PROOF OF CONCEPT WORDNET DESIGN FOR SOUTH AFRICAN SIGN LANGUAGE (SASL)

1. Introduction

This section presents initial research, hypothetical solutions and a description of what the development of a wordnet for South African Sign Language (SASL) would entail. This proposed design takes into account current international collaboration opportunities, as well as existing resources identified as possible starting points for such a development.

2. Background

2.1 What is a wordnet?

Wordnets are lexical-semantic resources that have applications in both linguistic research and natural language processing (NLP). They provide structured representations of word meanings and semantic relations. Synsets can be seen as the building blocks for wordnets and most simply refers to a group of synonyms that describe and define a single concept. Synsets are linked to each other with semantic relations and enriched with a definition and usage example.

Because wordnets organise lexical items into synsets and encode relations such as hypernymy, hyponymy, meronymy, and antonymy, between those synsets, they can also support the development and evaluation of computational models of lexical semantics. This type of modelling makes wordnets useful for tasks such as word sense disambiguation, semantic similarity and relatedness, information retrieval, text classification, machine translation, and question answering. In multilingual contexts, aligned wordnets facilitate cross-lingual information retrieval, multilingual NLP, and machine translation, while also providing a basis for comparing how different languages lexicalise and organise concepts. Wordnets are further used as linguistic resources in lexicography, language documentation, terminology development, and language teaching. For under-resourced languages in particular, a wordnet can provide valuable structured lexical-semantic data for NLP applications where other annotated resources are scarce. More recently, wordnets have also been used alongside distributional and neural language models, providing interpretable, structured semantic information that can complement corpus-based representations.

Below in Figure 1 is a graphic representation of the meaning network that is created when different synsets are linked via semantic relations. In this example, the relations between different body parts in Sesotho sa Leboa are shown, with English translation equivalents for ease of reference.

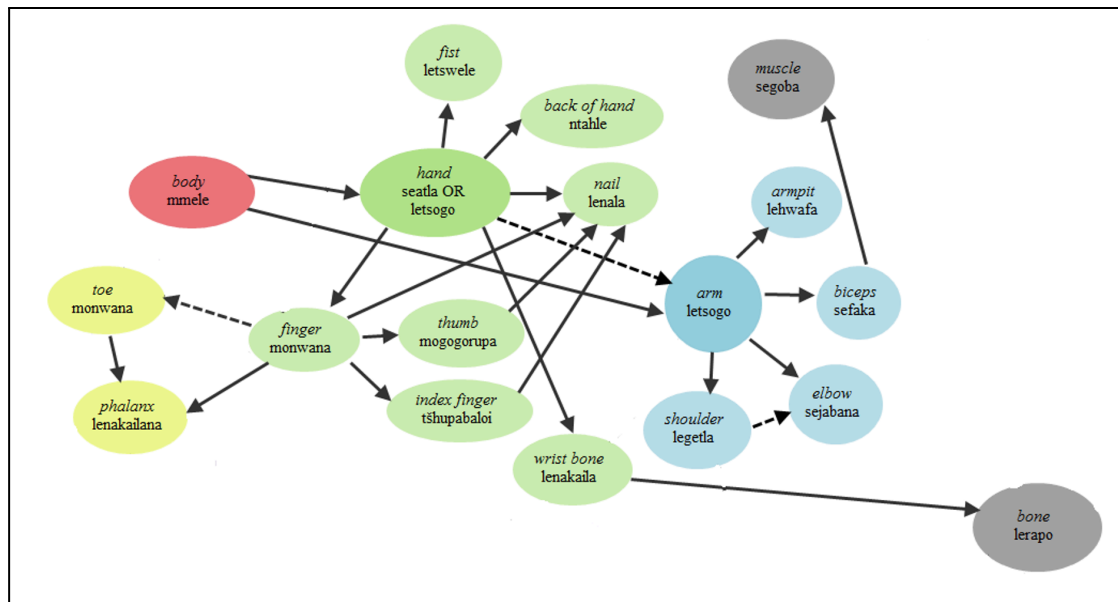


Fig. 1. Synsets to do with different body parts in Sesotho sa Leboa

The English [Princeton WordNet](#) (PWN) was the first large scale development of a wordnet for any written language (Miller, WordNet: A lexical database for English, 1995, p. 39) and currently consists of more than 117 000 synsets. Many wordnets describing different written languages have since been developed and a non-exhaustive list can be found on the Global Wordnet Association website.

Wordnets for signed languages also exist and follow the same conceptual organisation structure and largely the same development methodology. This will be discussed in more detail in Section 2.3.2.

2.2 Popular methods for developing wordnets

Development of wordnets for other languages is still strongly influenced and supported by the original framework of the PWN. Although various approaches exist, two emerge as most popular: the so-called [expand and merge approaches](#). These approaches are best understood as two ends of a continuum concerning the degree to which an existing wordnet, usually Princeton WordNet (PWN), determines the structure of the new resource.

2.2.1 Expand approach

In the expand approach, an existing wordnet provides the conceptual and relational backbone for the new wordnet. Typically, PWN synsets are translated into the target language, while their semantic relations are initially transferred as well. The resulting target-language synsets are then checked and adapted by lexicographers.

For example, if the English synset representing *dog* contains *dog*, *domestic dog*, etc., the developer identifies the corresponding lexicalisations in the target language and initially places

them in the equivalent synset. The target-language synset inherits the relevant relations, such as its hypernym relation to *canine*, subject to later verification.

Advantages

- Much faster than constructing the semantic hierarchy from scratch.
- Provides an immediate conceptual and relational framework.
- Automatically establishes alignment with PWN and, by extension, other wordnets using the same interlingual framework.
- Particularly useful where the target language has limited lexical resources or limited computational resources.

Disadvantages

- It implicitly assumes that the target language has roughly the same conceptual structure and lexicalisation patterns as English.
- English-specific distinctions can be unnecessarily imposed on the target language.
- Concepts lexicalised in the target language but absent from PWN are easily overlooked.
- Translation is not a one-to-one process, so simply translating English synset members does not guarantee an equivalent target-language synset.

The last point is particularly important: expand does not really mean "translate WordNet." A good expand-based project uses the English wordnet as a starting point, but the target-language lexicographer still has to determine whether the proposed lexicalisations and relations actually work in the target language.

The African Wordnet (AfWN) is a useful example. Its development has largely followed the expand model because the African languages involved are comparatively under-resourced and therefore lack the extensive monolingual lexical resources that would make a merge approach practical.

2.2.2 Merge approach

The merge approach reverses the direction of construction. The target-language wordnet is first developed independently, using resources and linguistic evidence from that language. Synsets and their intralingual semantic relations are established without taking PWN as the structural template. Only afterwards is the resulting wordnet mapped to PWN or another interlingual index.

The developer can use dictionaries, corpora, existing lexical databases, semantic resources, and linguistic expertise to determine which senses and semantic relations belong in the wordnet.

Advantages

- Allows the wordnet to reflect the lexicalisation patterns and semantic distinctions of the target language.

- Does not force the target language into the conceptual structure of English.
- Can identify concepts and distinctions that are absent from PWN.
- Produces a more genuinely language-specific semantic resource.
- The resulting wordnet can subsequently be aligned with PWN through equivalence relations.

Disadvantages

- Considerably more labour-intensive.
- Requires substantial linguistic and lexical resources.
- Establishing the semantic hierarchy and relations independently is difficult.
- Mapping the resulting synsets to PWN afterwards is itself a substantial task.

Polish WordNet (PolNet), Danish WordNet (DanNet) and Norwegian WordNet (NorNet) are commonly cited examples of the merge approach.

2.2.3. Other approaches and techniques

The expand/merge distinction describes how the conceptual structure is obtained, rather than whether the actual construction is manual or automatic. This is an important distinction because wordnet development can combine either model with different degrees of automation.

Dictionary-based construction

Bilingual or monolingual dictionaries can be used to identify candidate lexicalisations for synsets. In an expand approach, a bilingual dictionary can provide translations of PWN synset members. In a merge approach, a monolingual dictionary can provide evidence for senses, synonyms and semantic relations.

Corpus-based construction

Corpora can be used to identify:

- candidate lexical items,
- senses,
- synonyms,
- collocations,
- hypernym/hyponym relations,
- and other lexico-semantic relations.

Corpus evidence is particularly useful for validating automatically generated candidates, because dictionary equivalence does not necessarily imply semantic equivalence.

Corpus-based methods have also been used to extract relations directly rather than simply transferring them from English.

Corpus-/data-driven synset induction

A more radical alternative is to use corpus evidence and computational methods to induce synsets and relations, rather than treating PWN synsets as the fundamental units. Such approaches can identify language-specific semantic distinctions and potentially address concepts that are missing from the source wordnet. These approaches are particularly relevant to attempts to reduce the dependence of new wordnets on PWN.

2.3. Related work

2.3.1 The African Wordnet Project

The [African Wordnet](#) (AfWN) is a multilingual lexical-semantic resource developed for the indigenous South African languages. The project originated in a 2007 training workshop for linguists, lexicographers and computer scientists, with the first wordnet prototypes developed from 2008 onwards. Initially covering five languages, the project has subsequently expanded to nine languages: isiZulu, isiXhosa, Setswana, Sesotho sa Leboa, Tshivenda, Siswati, Sesotho, isiNdebele and Xitsonga. The AfWN follows an expand-based approach, using Princeton WordNet as a conceptual scaffold while allowing linguists to account for language-specific lexicalisation and semantic distinctions. Because the languages involved are relatively under-resourced, development has relied heavily on manual work by language experts, supplemented by available dictionaries, corpora and other lexical resources. The project has also incorporated more locally relevant resources, such as the SIL Comparative African Wordlist, to address concepts and lexical distinctions that are poorly represented by an English-based starting point. The resulting wordnets provide valuable resources for NLP, multilingual language technologies, cross-linguistic research, and the digital documentation of African languages and indigenous knowledge.

2.3.2 The Open Multilingual Signed Wordnet

The [Multilingual Sign Language Wordnet](#) (MSL-WN) is the first publicly available multilingual wordnet resource specifically for sign languages. It was initiated within the EU-funded EASIER project and is designed to connect lexical resources for different sign languages through shared semantic representations. The resource currently covers eight sign languages: British, German, Swiss German, Greek, French, Dutch, Polish and Swedish Sign Language, with its sign entries linked to synsets in the [Open Multilingual Wordnet](#) (OMW).

Rather than creating an entirely independent wordnet for each language, MSL-WN builds on existing sign-language dictionaries and lexical resources, linking their signs to existing OMW synsets through a largely semi-automatic annotation process. Candidate sign-synset connections are generated using glosses, translations and other lexical information, and are then validated by human annotators with relevant language expertise. This approach reflects both the advantages and challenges of working with sign languages, which generally lack conventional written forms and for which lexical information is often represented through videos accompanied by glosses or spoken-language keywords. The project has subsequently focused

on improving cross-lingual coverage between sign languages, including methods that use established synonymy between signs to suggest additional semantic connections. The resulting resource supports sign-language processing, cross-lingual linguistic research, and the integration of otherwise disparate lexical resources.

3. A proposed design for SASL

Bringing South African Sign Language into the African Wordnet project seems a natural next step. Some resources and expertise already exist within the community and can act as a starting point for further stakeholder involvement. For a first South African Sign Language (SASL) wordnet, a hybrid expand-plus-language-specific approach is recommended, rather than attempting either a pure expansion of an existing wordnet or a completely independent merge approach. The experience of the Multilingual Sign Language Wordnet (MSL-WN) provides a particularly useful model, while the African Wordnet experience shows the value of starting from an established wordnet for an under-resourced language.

A roadmap for development could follow these steps:

1. Use OMW/MSL-WN as the initial semantic backbone.

Start with the Open Multilingual Wordnet (OMW) synset inventory and identify a manageable core vocabulary, rather than attempting to reproduce the entire PWN inventory. This provides immediate interoperability with existing wordnets and with the growing MSL-WN.

2. Build from existing SASL lexical resources.

(A subsection of) The NID SASL Video Dictionary is an obvious starting point, since it contains thousands of signs, includes synonyms and community variation, and provides video rather than relying solely on spoken-language glosses. Other dictionaries, educational materials, corpora and lexical databases, such as those demonstrated by HandLab and the digitisation of LMS materials, should be incorporated where licensing permits.

3. Generate candidate synset links automatically.

Following the MSL-WN methodology, use English glosses, translations and other available lexical information to propose candidate OMW synsets. Their three-stage process, automatic candidate generation, automatic validation of high-confidence matches, followed by manual validation, seems particularly appropriate for SASL and an initial conversation with the research team by SADiLaR highlighted positive collaboration opportunities.

4. Make Deaf SASL expertise central to validation.

Human validation should determine whether a sign actually expresses the proposed meaning, whether multiple signs belong in the same synset, and whether distinctions in SASL require separate synsets. This is crucial because a gloss is merely a convenient label, not the meaning of a sign. The MSL-WN explicitly uses synsets to avoid treating

translation equivalents as interchangeable. Community led validation of all semi-automatic generated content is therefore vital. The AfWN project has recently partnered with SADiLaR's ESCALATOR programme to engage in a series of capacity development workshops. In these workshops, the importance of language speakers is highlighted and seen as an important aspect of creating digital voices for the South African languages. This should also be a central vision for a SASL wordnet.

5. Prioritise SASL-specific concepts and relations.

The project should not force SASL into the semantic structure of English. Where no suitable OMW synset exists, new synsets should be created and subsequently linked to the wider multilingual network where possible. The current MSL-WN already supports custom synsets, including categories traditionally poorly represented in wordnets and the AfWN project has also taken a strong stance to prioritise lexicalised content over a strict adherence to the framework.

6. Represent variation explicitly.

Regional, generational, community and institutional variation should be treated as linguistic data rather than simply as competing "correct" and "incorrect" signs. The existing NID SASL dictionary's organisation by community provides a useful precedent.

It is therefore recommended that the resulting resource be expand-based in its initial infrastructure, but merge-like in its treatment of SASL-specific semantics. In other words: use the existing network to get started, but allow SASL to tell us where the network should be localised.

In an attempt to visualise what implementation of these steps could look like for South African Sign Language, a very basic proof of concept was created using the available material in the digitisation collaboration between HandLab and SADiLaR.

Figure 2 shows an example of one of the entries. Here, the recorded sign for "tortoise" was enriched with information from the OMW, including the domain, definition, a short example sentence and semantic relations. If synonyms or dialectic variants of the sign as shown in the recording exist, links to those elements could also be added to the entry.

An immediate improvement that could be made is that both the definition (or gloss) and the usage example could also be captured in SASL, thereby showing the use of the sign in context.

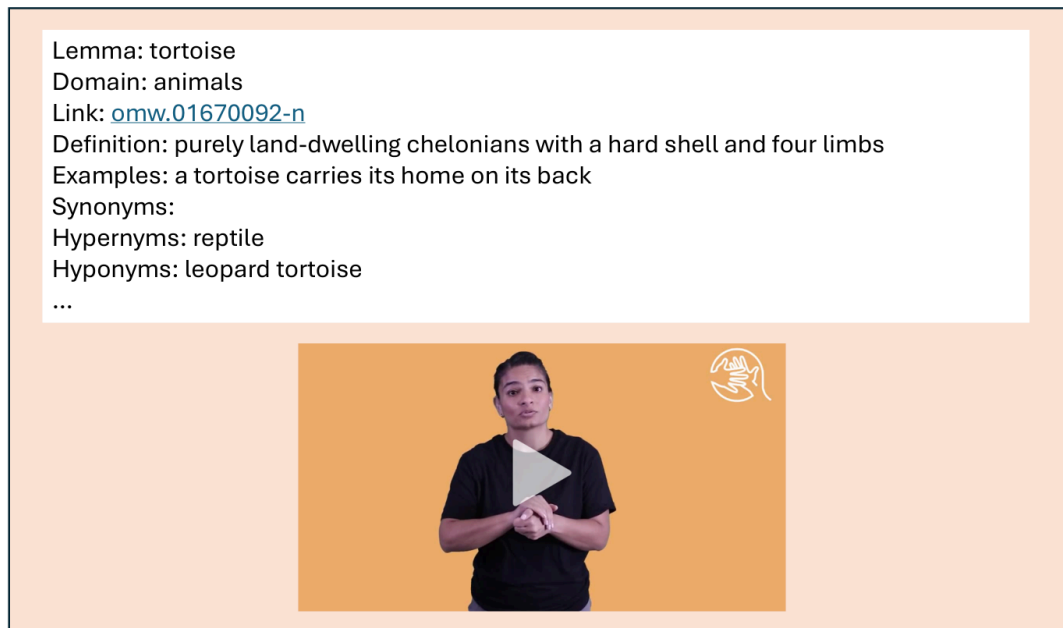


Fig. 2. “Tortoise” as included in a proof of concept synset for SASL

4. Next steps

The development of a South African Sign Language wordnet is both feasible and valuable. Building on existing multilingual wordnet infrastructure and available SASL lexical resources would provide a practical starting point, while expert involvement would ensure that the resulting resource reflects the linguistic structure and variation of SASL. A carefully scoped pilot would allow the methodology, resource requirements and potential challenges to be evaluated before considering a larger-scale wordnet development project.

The AfWN project management has initiated discussion with partners such as the researchers at the University of Hamburg who are involved in development of the MSL-WN, as well as SASL partners at South African universities to form a dedicated project team. The AfWN management will further reach out to stakeholders identified via the event on 29 May 2026. After a project team is established, the use of existing material that could form the basis for synsets needs to be discussed with partners such as the NID and HandLab. It might also be necessary to secure additional funding to support development of the SASL wordnet. SADiLaR and the AfWN project management can support these endeavours and can act as the facilitators for a deaf lead team.

Members of the SASL community who wish to participate in this project are also encouraged to contact the AfWN project manager, Ms Marissa Griesel at [marissa\(dot\)griesel\(at\)nwu.ac.za](mailto:marissa(dot)griesel(at)nwu.ac.za).