Towards an RDF Representation of the Infrastructure consisting in using Wordnets as a conceptual Interlingua between multilingual Sign Language Datasets

Thierry Declerck  
DFKI GmbH  
Saarland Informatics Campus  
Stuhlsatzenhausweg, 3  
66123 Saarbrücken  
declerck@dfki.de

Thomas Troelsgård  
University College  
Copenhagen (KP)  
Humletorvet 3  
1799 Copenhagen V  
ttro@kp.dk

Sussi Olsen  
Centre for Language Technology, NorS, University of Copenhagen  
Emil Holms Kanal 2  
2300 Copenhagen S  
saolsen@hum.ku.dk

Abstract

We present ongoing work dealing with a Linked Data compliant representation of infrastructures using wordnets for connecting multilingual Sign Language data sets. We build for this on already existing RDF and OntoLex representations of Open Multilingual Wordnet (OMW) data sets and work done by the European EASIER research project on the use of the CSV files of OMW for linking glosses and basic semantic information associated with Sign Language data sets in two languages: German and Greek. In this context, we started the transformation into RDF of a Danish data set, which links Danish Sign Language data and the wordnet for Danish, DanNet. The final objective of our work is to include Sign Language data sets (and their conceptual cross-linking via wordnets) in the Linguistic Linked Open Data cloud.

1 Introduction

A final goal of our work is to represent and publish Sign Language (SL) data sets in the Linguistic Linked Data (LLOD) cloud, which is a subset of the Linked Data (LD) cloud. We can observe that SL data are not represented in the data sets currently included in the LLOD cloud. We look at the “Overview of Datasets for the Sign Languages of Europe” published by the “EASIER” European project (Kopf et al., 2022) where we do not see any mention of a data set being available in a Linked Data compliant format.

This shortcoming is a problematic issue, as an important type of natural language is missing from the LLOD, while the motivation behind the creation of the LLOD is that it can ease the linking of all types of natural language resources.

The prerequisite for publishing linguistic data in the LLOD cloud is to have it formally represented within the Resource Description Framework (RDF). And as an RDF-based de facto standard for representing lexical information, the OntoLex-Lemon specifications, already exists, we investigate as a first step the re-use of this model in order to accommodate the description of Sign Language data sets. But as we can see in Figure 1, the class ontolex:Form covers only the representation of written languages (with the addition of the associated pronunciation information), so that there is a need to think about possible adaptations or extensions of OntoLex-Lemon.

At the same time, the OntoLex-Lemon model supports the representation of WordNet data, which are typically encoded with the SKOS vocabulary, where the synsets are represented as instances of the ontolex:LexicalConcept subclass of the skos:Concept class. This feature is offering us a good starting point for transforming into RDF (and OntoLex-Lemon) recent work by the EASIER project dealing with the use of shared IDs of the Open Multilingual Wordnet (OMW) infrastructure for interlinking SL data sets for two languages: German and Greek, as described in (Bigeard et al., 2022).
OMW (as the GermaNet resource (Kunze and Lemnitzer, 2002) used in EASIER is not included in OMW). It can also be seen in Figure 2 that a “GSL” box is being positively checked. “GSL” stands for “Greek Sign Language”, and the positively checked abbreviation in the screenshot means that there is a corresponding synset in the Greek Wordnet available in OMW. This way, a DGS sign can be linked to a GSL sign, based on a shared OMW ID, which is much more accurate than linking only via translation of glosses.

The links between the one OMW ID and the two signs/videos IDs are available in Excel files. The corresponding CSV lines are displayed in Figure 3 and Figure 4, where we can see that one OMW ID (omw.00377364-n, with the English lemma “explosion”, translated to German “Explosion”, and with the Greek lemma “εκρηξη”) is associated with both the German and the Greek SL resources, thus establishing a conceptual link between those.

OMW (as the GermaNet resource (Kunze and Lemnitzer, 2002) used in EASIER is not included in OMW). It can also be seen in Figure 2 that a “GSL” box is being positively checked. “GSL” stands for “Greek Sign Language”, and the positively checked abbreviation in the screenshot means that there is a corresponding synset in the Greek Wordnet available in OMW. This way, a DGS sign can be linked to a GSL sign, based on a shared OMW ID, which is much more accurate than linking only via translation of glosses.

The links between the one OMW ID and the two signs/videos IDs are available in Excel files. The corresponding CSV lines are displayed in Figure 3 and Figure 4, where we can see that one OMW ID (omw.00377364-n, with the English lemma “explosion”, translated to German “Explosion”, and with the Greek lemma “εκρηξη”) is associated with both the German and the Greek SL resources, thus establishing a conceptual link between those.

Those elements: videos, glosses, phonetic transcriptions, links to OMW, are the elements we are encoding in a unified and harmonized Linked Data compliant format.

---

10 As stated in the web page, from which this screenshot is taken, the online interface is not yet live. But the displayed screenshot represents clearly how the linking of a German sign and an OMW ID is (will be) represented on the web.

11 On the specificity of glosses used for naming (or labelling) SL data in corpora, see (Ormel et al., 2010)

12 DGS stands for “Deutsche Gebärdensprache”, German Sign Language
2 Linked Data compliant Encoding of the Infrastructure using shared OMW IDs

As stated in the introduction, we need to transform into RDF the different types of data used for representing signs for their future publication in the LLOD cloud. We also make use of RDF(S) and OWL representation languages, as those are constitutive parts of the OntoLex-Lemon specifications and of the building of ontologies.¹⁴

For the RDF representation of videos included in our data set, we just introduce a class and have all videos encoded as instances of this class.

Listing 1 displays the RDF-based encoding of a video containing a German sign.¹⁵

Listing 1: The RDF-based encoding of a video containing a sign

```xml
<http://example.org/dgs#SignVideos_40085921.mp4>
  rdf:type sl:SignVideos ;
  dgs:hasGLOSS dgs:GLOSS_13990 ;
  dgs:hasGLOSS dgs:GLOSS_13990−2966 ;
  sl:hasVideoAddress
    "https://www.sign-lang.uni-hamburg.de/korpusdict/clips/40085921.mp4"^^rdf:HTML ;
  rdfs:label "Videos representing a sign" ;
```

Listing 2 displays the corresponding glosses (as instances of a specific class).

Listing 2: The RDF-based encoding of glosses

```xml
dgs:GLOSS_13990
  rdf:type dgs:GLOSS ;
  rdfs:label "PROTECTION1A"^^en ;
  rdfs:label "SCHUTZ1A"^^de ;

dgs:GLOSS_13990−2966
  rdf:type dgs:GLOSS ;
  rdfs:label "PROTECTION1A"^^en ;
  rdfs:label "SCHUTZ1A"^^de ;
```

The subclass/subtype relation between the glosses displayed in Listing 2 is encoded in a specific class, called “Type”, as can be seen in Listing 3, which displays the subclass hierarchy between glosses (here class and subclass instances are linking to the same video), and linking the instance of the subclass to an OMW element (as instance of the class ontolex:LexicalConcept), establishing thus the link to the WordNet world, and to the corresponding video(s), as we can collect more than one video representing a sign.

Listing 3: Subclass hierarchy of glosses linking the subclass to OWM and to videos

```xml
dgs:Type_13990
  rdf:type dgs:Type ;
  dgs:hasGLOSS dgs:GLOSS_13990 ;
  dgs:hasSubType
dgs:Subtype_13990−2966 ;
  dgs:hasVideo
    <http://example.org/dgs#SignVideos_40085921.mp4> ;
    rdfs:label "SCHUTZ"^^de ;
    rdfs:label "PROTECTION"^^en ;

dgs:Subtype_13990−2966
  rdf:type dgs:Subtype ;
  dgs:hasGLOSS dgs:GLOSS_13990−2966 ;
  dgs:hasOMW−Link wnid:omw−00817680−n ;
  dgs:hasVideo
    <http://example.org/dgs#SignVideos_40085921.mp4> ;
    rdfs:label "SCHUTZ"^^de ;
    rdfs:label "PROTECTION"^^en ;
```

¹⁴RDF(S) stands for “RDF-Schema”, see https://www.w3.org/TR/rdf-schema/ for more details. OWL stands for “Web Ontology Language”, see https://www.w3.org/TR/2012/REC-owl2-primer-20121211/ for more details.

¹⁵The original sign and all the related information are accessible at https://www.sign-lang.uni-hamburg.de/meinedgs/types/type13990_de.html (for the English translation of the page: https://www.sign-lang.uni-hamburg.de/meinedgs/types/type13990_en.html)
The OMW synset linked to in Figure 3 has the internal organisation displayed in Figure 6. Here we didn’t include links to glosses or videos, as the relations to OMW described in listing 3 are inverse.

Finally, the representation of the form(s) of the sign is performed for the time being as instances of `ontolex:Form` (mediated, also for the time being, by an underspecified instance of `ontolex:LexicalEntry`). This representation, displayed in Figure 7, includes the machine-readable transcription of the HamNoSys code, in the so-called SiGML XML format (Neves et al., 2020). It also includes potential keywords or lexical entries.

3 Extending the EASIER Approach with additional Sign Videos per Language

We searched for other Sign Language resources in order to extend the approach described in (Bigeard et al., 2022), thus linking SL data and wordnets, and then transforming those SLs-wordnets combinations into RDF and OntoLex-Lemon.

We found a basic lexicon of 1000 concepts associated with SL data in 4 languages, English, French, German and Greek, an outcome of the past Dicta-Sign project (Matthes et al., 2012), which is available at the University of Hamburg at https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/concepts/concepts_eng.html. This resource is directly relevant to our purposes, as the included videos are equipped with glosses and HamNoSys transcriptions, as shown in Figure 8.

In Figure 8, we observe that the gloss and the HamNoSys transcription for the German video are identical with those deployed in the data used by the EASIER project for linking SL data and wordnets, as can be seen at https://www.sign-lang.uni-hamburg.de/meinedgs/types/type13990_de.html, and which is also shown in Figure 5.

This concordance of gloss and HamNoSys transcriptions16 not only allows for the association of two videos representing this German sign to one OWM ID, but it also permits the addition of signs in two additional languages, English and French, extending thus the multilingual coverage of the approach described by (Bigeard et al., 2022). We just need to introduce in our RDF representation new video instances (one per language) and to link them to the same OWM ID.

But we can observe that in the one case the gloss is realised as a noun and in the second case as a verb. Signs are often ambiguous with respect to PoS, and in the future we will link the videos to both the nominal and verbal synsets, if both are available in the corresponding wordnet.

As the page https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/concepts/cs/cs_688.html is linking to a more detailed lexical description of the sign, with the same gloss and HamNoSys transcription (see https://www.sign-lang.uni-hamburg.de/galex/glossen/gl3990.html), with another video for the sign, we can in fact have 3 videos for this German sign associated with one OWM ID.
Thus, the transformation of this additional data into our RDF and OntoLex-Lemon representation means organising those originally disparate and heterogeneous data sources in one harmonised representation format, with OMW as the central component for the interlinking of the different data types and sources.

4 Extending the Approach to Danish WordNet and Sign Language Data

(Troelsgård and Kristoffersen, 2018) discusses approaches for ensuring consistency between (Danish) Sign Language corpus data and the dictionary of Danish signs that is described in (Kristoffersen and Troelsgård, 2010). This approach aims at getting a correspondence between the dictionary lemmas and the corpus lexicon, which consists of types introduced for lemmatising the tokens found in the annotations (glosses added to the signs) of the corpus.

The strategy being to use words and their equivalents (also found in the dictionary) to search for signs in the corpus. In order to extend the list of potential Danish equivalents that could be used for a word-based search of signs in the corpus, (Troelsgård and Kristoffersen, 2018) suggest using the Danish wordnet, DanNet, which is briefly described below and in more details in (Pedersen et al., 2009) and (Pedersen et al., 2018).

DanNet was constructed using the merge approach where the wordnet is built on a monolingual resource, in this case on the corpus-based Danish dictionary Den Danske Ordbog (DDO, Lorentzen, 2004), see also https://ordnet.dk/ddo, and subsequently linked to PWN. For DanNet this linking was based on the Princeton Core wordnet, a subset containing 5,000 central concepts of English (see the file “corewordnet.txt” in the folder “LFglosses_standoff_files.zip” under https://wordnetcode.princeton.edu/morpholinks) that were semi-automatically linked to DanNet. These linked elements constitute the part of DanNet available in OMW.

Since 2018, there has been an ongoing effort to link a larger part of DanNet’s more than 65,000 synsets to PWN, this time taking departure in the core Danish vocabulary, see (Pedersen et al., 2019).

The relations between sign identifiers and lexical elements from both DanNet and other dictionary sources are encoded in a database, out of which we got a CSV export.

In this export, we first have the signs, which are corresponding to entries in the dictionary of signs available at www.tegnsprog.dk.

A second type of data available in the export holds video links and information about the sign form (HamNoSys/SiGML). The HamNoSys notation, though, is rather coarse, as it is generated automatically from the dictionary’s phonological descriptions.

A third type of information included in the export is dealing with the senses associated with the signs and their (form) variants.

Lastly, we also have the information from DanNet and PWN. Our work consists thus in porting all those (interlinked) resources to RDF and OntoLex-Lemon, as we did for the German and the Greek data, as presented in Section 2. In the OMW version of DanNet, we find for example the following information “00817680-n lemma beskyttelse”, where the lemma corresponds to the OMW English wordnet “00817680-n lemma protection”, sharing thus the same ID for the concept of “protection” as the English and Greek wordnets we have in OMW. So that we can add the Danish sign ID (and video), which we got from the database, to our infrastructure. The Danish sign associated with the wordnet

Figure 8: The concept “protect” as realised in 4 different Sign Languages. Taken from https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/concepts/cs/cs_688.html
lemma “beskyttelse” is displayed in Figure 9.

![Figure 9: The Danish sign associated with the OMW ID “00817680-n”, corresponding to the (highlighted) lemma “beskyttelse”, here as possible lexical realisation of the Danish gloss “FORSVARE” (defend)](image)

It is then straightforward to encode all those types of information on the relation between Danish SL data and DanNet into our RDF-based model. We need only to add an instance for the video displaying the sign, and its associated gloss (with language equivalents), as shown in Figure 10. The language equivalents are included, so that a Danish sign can be cross-lingually searched for, using glosses in other languages.

![Figure 10: The Danish gloss (with language equivalents) associated with the video with ID dts-1_2162](image)

Then we just have to add an `ontolex:Form` instance for the Danish sign, displayed in Figure 11 and which is linked via its corresponding lexical entry to the corresponding OMW instance, which are shown in Figure 12.

Finally, Figure 13 displays (partially) the current encoding of the OMW ID, showing the central and pivotal role of this ID for interlinking the various types of resources involved in our work.

5 Current Results

Our encoding results in a harmonised representation of data that was originally stored in different formats in different locations. Taking advantage of the work proposed by (Bigeard et al., 2022), (Troelsgård and Kristoffersen, 2018) and others, we can include the links between SL data and wordnets in a harmonised representation, under the umbrella of RDF and by re-using elements of OntoLex-Lemon. The Open Multilingual Wordnet infrastructure is playing a central role in this work, as the shared OMW IDs across various languages are at the core of the interlinking of the distinct data types and sources. The resulting unified representation supports a dense linking of different types of information. Our model will be made available on Github (<https://github.com/Declerck/sl-wn-rdf-ontolex>).

6 Future Work

The next steps of our work will consist in automatising the transformation into RDF and aspects of OntoLex-Lemon so that we have all the data in the harmonised representation space. We are also planning to investigate a transformation of ASLNet (Lualdi et al., 2021) into RDF. We continue to extend our work with more data in more languages, starting with Maltese,

19For example, a useful dictionary resource for Maltese Sign Language is available at [https://mlrs.research](https://mlrs.research).
Figure 13: The encoding of the OWM ID, linking to corresponding lexical entries, which again are linking to other elements of our data set, as can be seen in 12 for the Danish case.

guage. Finally, we aim at adding other types of visual lexical data, like pictograms, as the links between such data and wordnet have been already investigated, for example in (Schwab et al., 2020).

Acknowledgements

The presented work is pursued in the context of the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), 731015). We thank Thomas Hanke and Sam Bigeard from the Institute of German Sign Language and Communication of the Deaf (IDGS) at the University of Hamburg for providing links and explanations to data developed in the context of the EASIER project (https://www.project-easier.eu/de/). The work we started dealing with Maltese language is pursued in the context of the LT-BRIDGE project, which has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No 952194.

um.edu.mt/resources/lsm. This resource is interesting to us, as it makes use of another transcription system than HamNoSys, the SignWriting system (Sutton, 1991), so that our model will deal also with more than one transcription system.

References


Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. D6.1 overview of datasets for the sign languages of europe.


