Proceedings of the Globalex Workshop on Linked Lexicography @LREC2022, pages 6–9
Marseille, 20 June 2022

Towards the Linking of a Sign Language Ontology with Lexical Data

Thierry Declerck
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus D3 2
66123 Saarbrücken, Germany
declerck@dfki.de

Abstract

We describe our current work for linking a new ontology for representing constitutive elements of Sign Languages with lexical data encoded within the OntoLex-Lemon framework. We first present very briefly the current state of the ontology, and show how transcriptions of signs can be represented in OntoLex-Lemon, in a minimalist manner, before addressing the challenges of linking the elements of the ontology to full lexical descriptions of the spoken languages

Keywords: Linked Data, Sign Languages, OntoLex-Lemon

1. Extended Abstract

The final goal of our work is to provide for a multimodal extension to the OntoLex-Lemon framework (Cimiano et al., 2016), which was originally conceived for covering the written and phonetic representation of lexical data, as can be seen in the relation existing between the `ontolex:LexicalEntry` and `ontolex:Form` classes, which are displayed with the core module of OntoLex-Lemon in Figure 1.

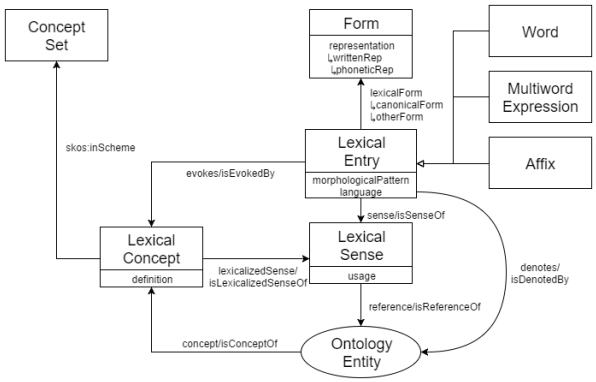

Figure 1: The core module of OntoLex-Lemon, taken from `https://www.w3.org/2016/05/ontolex/`

Thereby, we aim at supporting the same type of semantic interoperability between Sign Language(s) (SL) lexical data as this is achieved in OntoLex-Lemon for the written or phonetic representations of lexical data.

Sign Language is a type of natural language with distinctive properties.[1] It poses a challenge for its integration in OntoLex-Lemon, as SL descriptions and interpretations involve a huge number of descriptors (or data categories), including information about "physical" (body parts) and spatial (orientation, movements, etc.) elements, which are not playing any role when it comes to represent the "classical" lexical data in the spoken or written language. This complexity of the SL lexical data and the challenges it poses for its full formal representation in the OntoLex-Lemon lexical framework might lead to the design of a specific module extension, in which we can also address the issue on how to represent cross-modal relations, as this was not needed in the case of the values of only the `ontolex:writtenRep` and `ontolex:phoneticRep` properties (see Figure 1).

One aspect of our work was to design and implement an ontology of the data categories used for describing Sign Languages, including the already mentioned "physical" (body parts) and spatial (orientation, movements, etc.) elements, but also classifications of different types of sign languages, the phonological properties of SL, etc. The current status of this ontology is presented in a paper ("Towards a new Ontology for Sign Language") to be presented at the LREC conference, and which we briefly summarise in this extended abstract.

We built the ontology on the basis of a number of available SL resources, like the CLARIN concept repository (`https://www.clarin.eu/content/clarin-concept-registry`), the American Sign Language lexicon (`https://asl-lex.org/visualization/`), the British Sign Language dictionary (`https://www.british-sign.co.uk/british-sign-language/dictionary/`) or the Institute for German Sign Language and Communication of the Deaf at the University of Hamburg (`https://www.idgs.uni-hamburg.de/`), and the "SignGram Blueprint. A Guide to Sign Language Grammar Writing" publication, resulting from the SignGram COST Action: `https://parles.upf.edu/llocs/cost-signgram/node/18`.

Our approach consisted mainly in proposing an har-

[1] Specifics of Sign Languages and the challenges for defining a corresponding writing system are described in depth in (Bianchini, 2021)

monisation of all the features (or data categories) introduced and explained in those different highly relevant sources, and to organise this harmonised set of descriptors into an ontology, while conserving the information on the origin of the data. We have for now more than 260 harmonised ontology elements, organised in a (tentative) hierarchy. Figure 2 is displaying aspects of the current state of the SL ontology.

Parallel to this work, we started to investigate the encoding of transcriptions of Sign Language data in OntoLex-Lemon. For this purpose, we studied the type of transcription offered by the HamNoSys notational system (Hanke, 2004).[2] Figure 3 displays the sign labelled with the German word "Busch".

As HamNoSys per se is not machine-readable, we are making use of a conversion of it into an XML format called SiGML, which is very often used as the input to avatar generation software, as described in (Jennings et al., 2010). There exists a python implementation that transforms HamNoSys in SiGML, which is described in (Neves et al., 2020). The resulting notational code, an example of which is displayed in Figure 4, is the one we use to be included in OntoLex-Lemon, and from which we can link to elements of the ontology, or to a pose or video streaming object.

We tentatively represent this SiGML code as a value of the OntoLex-Lemon "writtenRep" property, with a special tag "sigml", as can be seen in Figure 5. We need to stress here that the string "Busch" associated with the HamNoSys notation of the sign is to be considered as a label, and not as a lexical entry. In our suggested representation, we can see how three encodings for "Busch" are representing three different modalities, with different types of information. But other options are under discussion within the Ontolex community.

An alternative solution could consist in introducing a specific lexical entry for the "word" used for labelling the sign, and to "loosely" relate it to the lexical entry that is encoding the word "Busch" as used in the spoken language. Another option would be to consider the label "Busch" rather as a conceptual entity, which can be linked to a number of lexical entries that could be a lexical realisation of this conceptual "tagging", as we can think that the annotators of SL corpora are rather using concepts instead of specific lexical entries of the spoken language. In this we would orient ourselves towards a WordNet like representation of the semantics of signs.

2. Current Work

While the solution presented in the former section for encoding transcriptions of SL data in OntoLex-Lemon seems to be relatively straightforward, it does ignore many aspects of Sign Languages, which are encoded in our ontology. Our current work, to be made soon available in a first version, consists in implementing a strategy for linking the descriptors included in the SL Ontology with the OntoLex-Lemon representation of HamNosys/SiGML encodings, maybe also including videos sequences as external references.

We need for this to take into account a variety of descriptor types, some of which we summarise in this section.

The ASL-LEX (`https://asl-lex.org/visualization/`) resource uses for describing a sign ca 95 features distributed over 7 main classes: Frequency Properties, Iconicity Properties, Lexical Properties, Sign Duration, Phonology, Phonological Calculations, and Acquisition Information. As we can see, some of those data categories are not included in the HamNoSys/SiGML set of features. We will need to include the "Acquisition Information" within the Metadata Module for OntoLex (LIME), which might need to be extended. This high number of descriptors is challenging, as it makes it difficult to link them in a consistent way to the HamNoSys/SiGML representation in OntoLex-Lemon, also with the question if all the 95 features are equally relevant for this linking task.

The British Sign Language dictionary (`https://www.british-sign.co.uk/british-sign-language/dictionary/`) has an interesting approach, as it offers textual descriptions of the sign used for a concept. For example for "aeroplane", the site is providing this information: " **Description**: Thumb and little finger of primary hand extended with palm facing downwards. Hand starts in front of body and moves up at an angle across body. **Definition**: A machine that can fly. It has wings and engines. **Also Means**: plane, flight". The text included in the "Description" section is very interesting and very specific to Sign Language (or for describing gestures in general), and for which we have no field in OntoLex-Lemon. It will be challenging to link this kind of information to an HamNoSys/SiGML representation in OntoLex-Lemon, as the text has to correspond to the features used in the XML code. Also interesting in the "Aeroplane" example is the fat that various meanings are given to the sign. This calls also for a WordNet like representation in OntoLex-Lemon, and linking thus the set of features used for describing the sign to an ontolex:LexicalConcept instance.

We also need to handle multilingual aspects. The Dicta-Sign project is offering a list of 1000 concepts realised in 4 languages (German, Greek, English and French), with videos and HamNoSys transcriptions. As the "words" used to label the concepts (like "abandon") can not be considered as lexical entries, we will integrate those labels as instances of the ontolex:LexicalConcept. It remains unclear to how many lexical entries those concepts can be linked.

As a consequence of this preliminary study, we see that

[2] See also `https://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/HamNoSys_2018.pdf` for a detailed graphical representation of HamNoSys

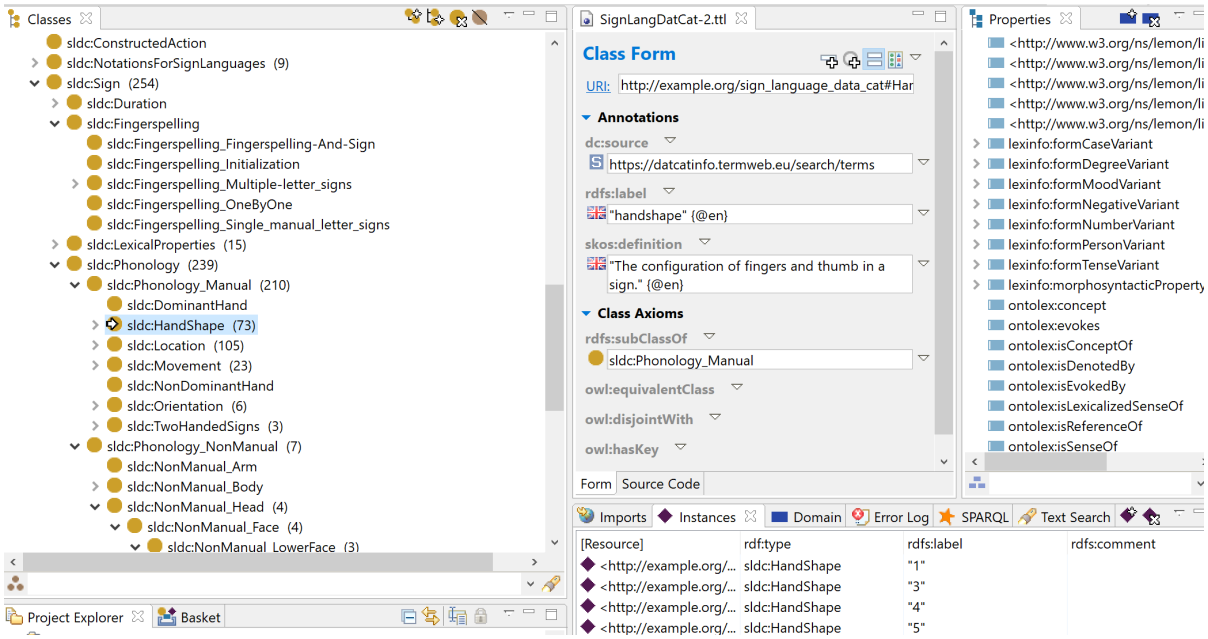

Figure 2: A screenshot of the ontology, displaying parts of its tentative hierarchy of classes

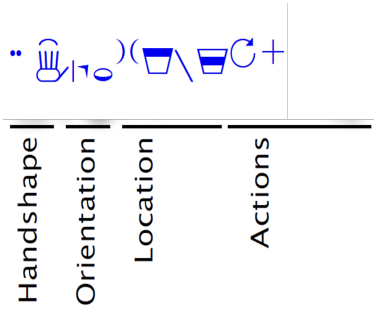

Figure 3: The sign labelled with the German Word “Busch” in HamNoSys notation, using the four features: Handshape, Orientation, Location and Actions.

```
(base) C:\Users\thde00\.spyder-py3\python_script\SignLanguage\HamNoSys2SiGML-
master\Original>python HamNoSys2SiGML.py "□□□□□□□□□□□□" (Busch)
<?xml version="1.0" encoding="UTF-8"?>
<sigml>
        <hns_sign gloss="(Busch)">
                <hamnosys_nonmanual/>
                <hamnosys_manual>
                        <hamsymmlr/>
                        <hamfinger2345/>
                        <hamfingerbendmod/>
                        <hamthumboutmod/>
                        <hamextfingeruo/>
                        <hampalmd/>
                        <hamclose/>
                        <hamshoulders/>
                        <hambetween/>
                        <hamchest/>
                        <hamcircleo/>
                        <hamrepeatfromstart/>
                </hamnosys_manual>
        </hns_sign>
</sigml>
```

Figure 4: The Transformation of an HamNoSys notation for the German label ”Busch” in SiGML code

linking a set of features describing signs to a lexical entry of the spoken language might not always be possible, but rather to instances of ontolex:LexicalConcept. An other consequence seems to be that we might need a specific module for describing dictionaries or lexicons of sign languages.

SignLangDatCat-2.ttl

```
ontolex:Busch_Sg
  rdf:type ontolex:Form ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:number lexinfo:singular ;
  rdfs:label "\"Busch\""@de ;
  ontolex:phoneticRep "\"bʊʃ\""@IPA ;
  ontolex:writtenRep "\"Busch\""@de ;
  ontolex:writtenRep "\"hamsymmlr
            hamfinger2345
            hamfingerbendmod
            hamthumboutmod
            hamextfingeruo
            hampalmd
            hamclose
            hamshoulders
            hambetween
            hamchest
            hamcircleo
            hamrepeatfromstart\""@de-DE-sigml ;
.
```

Figure 5: Inclusion of the SiGML code as an instance of the `ontolex:Form` class

3. Acknowledgements

This paper is based upon work from the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). The article is also supported by the Hori-

zon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreeement no. 731015).
We thank the members of the W3C Ontolex Community Group for their contributions to the discussions on this extension work.

4. Bibliographical References

Bianchini, C. S. (2021). How to improve metalinguistic awareness by writing a language without writing: Sign Languages and SignWriting. In Y. Haralambous, editor, *Proceedings of Grapholinguistics in the 21st Century, 2020*, volume 5 of *Grapholinguistics and Its Applications*, pages 1037–1063. Fluxus Editions.

Cimiano, P., McCrae, J. P., and Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report. W3C community group final report, World Wide Web Consortium.

Hanke, T. (2004). HamNoSys – representing sign language data in language resources and language processing contexts. In Oliver Streiter et al., editors, *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*, pages 1–6, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Jennings, V., Elliott, R., Kennaway, R., and Glauert, J. (2010). Requirements for a signing avatar. In Philippe Dreuw, et al., editors, *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 133–136, Valletta, Malta, May. European Language Resources Association (ELRA).

Neves, C., Coheur, L., and Nicolau, H. (2020). HamNoSyS2SiGML: Translating HamNoSys into SiGML. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6035–6039, Marseille, France, May. European Language Resources Association.