Thierry Declerck

INTEGRATION OF SIGN LANGUAGE LEXICAL DATA IN THE OntoLex-Lemon FRAMEWORK

Abstract We describe the status of work intending at including sign language lexical data within the OntoLex-Lemon framework. Our general goal is to provide for a multimodal extension to this framework, which was originally conceived for covering only the written and phonetic representation of lexical data. Our aim is to achieve in the longer term the same type of semantic interoperability between sign language lexical data as this is achieved for their spoken or written counterparts. We want also to achieve this goal across modalities: between sign language lexical data and spoken/written lexical data.

Keywords Sign Languages; OntoLex-Lemon; lexical data

1. Introduction

In the context of work dealing with the integration of multimodal lexical resources into the OntoLex-Lemon framework, which is described in (Cimiano/McCrae/Buitelaar 2016),[1] we investigate how to integrate lexical information included in Sign Language data. OntoLex-Lemon was originally covering only the written and phonetic representations 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.

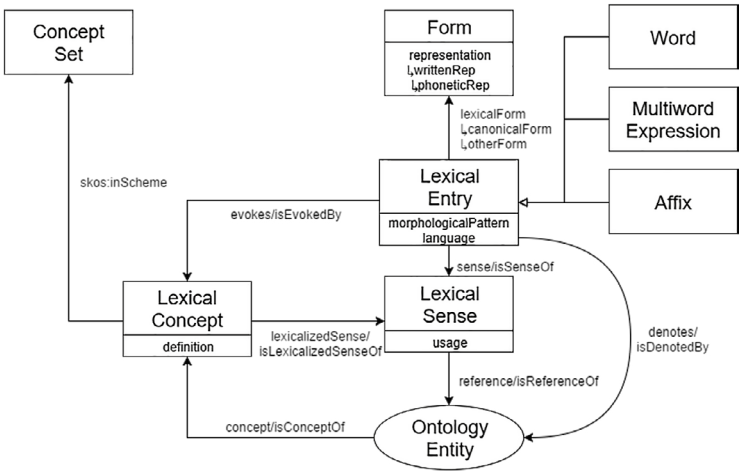

Fig. 1: Lemon_OntoLex_Core, taken from https://www.w3.org/2016/05/ontolex/

2. Consulted sources

We started our work by an extensive overview of the literature dedicated to the properties of sign languages (some of those works are included in the list of references), followed by a study of notational systems used for transcribing signs that mostly available in video or

[1] The full specification of OntoLex-Lemon is also available at https://www.w3.org/2016/05/ontolex/ (last access: 27-05-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

pose streams. We concentrate in this paper on the possible representation of elements of such notational systems in the context of OntoLex-Lemon. Figure 2 gives a good overview of various ways of representing sign language data (here dealing with American Sign Language, taken from (Yin et al. 2021)), with three of them being notational transcriptions of the video or the pose streams: SignWriting,[2] HamNoSys[3] and Glosses.

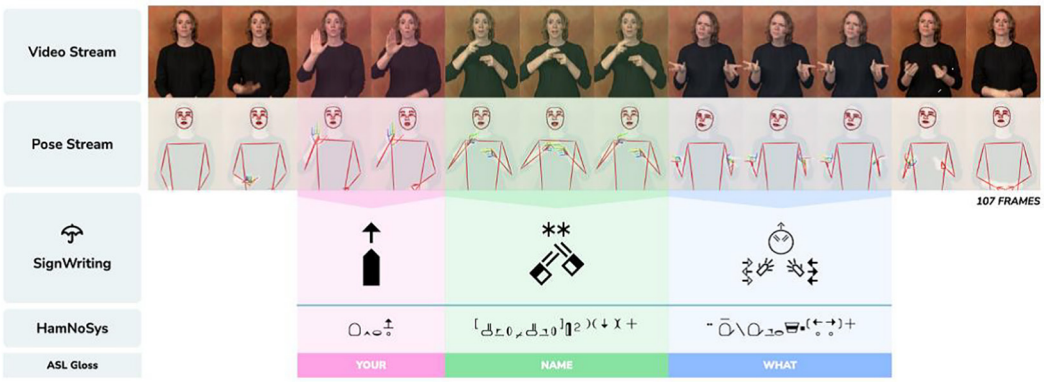

Figure 2: Various representations of American Sign Language. English translation: "What is your name?"

Fig. 2: Taken from Yin et al. (2021)

Glosses can be considered to label a sign (or a sequence of signs), as very often a corresponding (generally accepted) lexicon that could be used for annotating a sign (or a sequence of signs) is lacking. This issue is discussed in detail in (Ormel et al. 2010) and (Crasborn et al. 2012). If the glosses are to be seen more as labels used in the context of a corpus annotation process, it might make sense to consider their encoding within the "FrAC" OntoLex-Lemon extension module.[4]

The two other notational systems are representing (or transcribing) central elements of Sign Languages, like for example the shape and the orientation of the hands used by the signers, the interaction of the hands, their movements, also with respects to parts of the body and their activity, including repetitions, etc. For the time being we do not deal with the representation of facial elements, which left for a next stage of our work.

We focused for now on how to deal with the HamNoSys notational system, which breaks out a sign in four elements: handshape, orientation, location, and actions, as can be seen in Figure 3. But as HamNoSys per se is not machine-readable, we are making use of a conversion of it, called SiGML,[5] which is very often used as the input to avatar generation software. There exists a python implementation that transforms HamNoSys in SiGML, and which is

[2] More information about SignWriting can be found at https://en.wikipedia.org/wiki/SignWriting (last access: 27-05-2022).

[3] More information on HamNoSys can be found at https://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/HamNoSys_2018.pdf (last access:27-05-2022). See also (Hanke 2004).

[4] "FrAC" stands "Frequency, Attestation and Corpus information", and is a potential extension module, that not only covers the requirements of digital lexicography, but also accommodates essential data structures for lexical information in natural language processing. See https://acoli-repo.github.io/ontolex-frac/ (last access:27-05-2022) for more detail.

[5] See https://vh.cmp.uea.ac.uk/index.php/SiGML (last access: 27-05-2022) for more details. See also Jennings et al. (2010):

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

described in Neves/Coheur/Nicolau (2020). The resulting notational code, which is displayed in Figure 4, is the one we use then to be included in OntoLex-Lemon, and from which we will able to link to a pose or video streaming object.

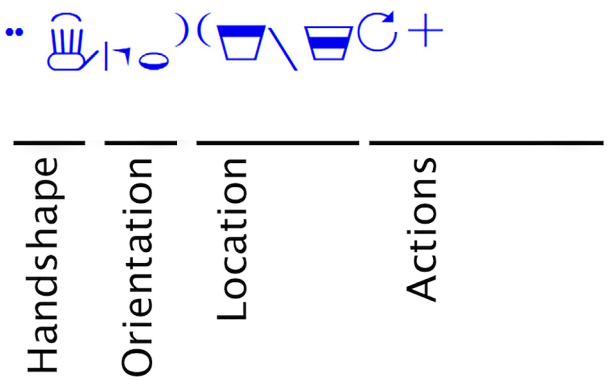

Fig. 3: The sign labelled with the German Word "Busch" in HamNoSys notation, using the four features: Handshape, Orientation, Location and Actions

```
(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>
```

Fig: 4: The SiGML conversion of the HamNoSys notation displayed in Figure 3, and which is used in our OntoLex-Lemon representation of sign language lexical data

3. Our current representation in OntoLex-Lemon

It is the kind of code displayed in Figure 4 that we can straightforwardly add to the OntoLex-Lemon framework, either introducing a new property to the ontolex:Form class (could be named ontolex:signRep) or by considering it as a written representation with a special tag "sigml", which is shown in Figure 5. In this example we can observe the complexity of the representation of such a sign, compared to the encoding for the written and phonetic representations. From this notational code we could link to video or pose streams that are displaying this sequence of signs.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

SignLangDatCat-2.ttl

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

Fig. 5: Inclusion of the SiGML code within an instance of the ontolex:Form class, together with the encoding of the written and phonetic representations

We are currently investigating how the addition of this modality is affecting the representation and the linking its lexical data to lexical senses or lexical concepts. It might be that we need to duplicate lexical entries for being able to fully represent the contributions of sign language lexical data to meanings and concepts. As often stated, sign language is another type of natural language and its full representation (including semantics, etc.) might lead to a specific module extending OntoLex-Lemon. We also need to 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:phonRep properties.

We are also working on establishing an ontology encoding all possible data categories associated with sign language (Declerck 2022). This ontology is re-using elements from 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/) as well as from Institute for German Sign Language and Communication of the Deaf at the University of Hamburg (https://www.idgs.uni-hamburg.de/). This ontology is also reusing elements of a former ontology for the Italian sign language, which is described in (Gennari/di Mascio 2007). Work will consist in linking the more than 250 constitutive elements of Sign Language included in this ontology to lexical descriptions represented in OntoLex-Lemon.

References

Cimiano, P./McCrae, J.-P./Buitelaar, P. (2016): Lexicon model for ontologies: final community report, 10 May 2016. https://www.w3.org/2016/05/ontolex/ (last access: 27-05-2022).

Crasborn, O./de Meijer, A. (2012). From corpus to lexicon: the creation of ID-glosses for the Corpus NGT. In: Crasborn, O. et al. (eds.): Proceedings of the 5th Workshop on the Representation and

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Processing of Sign Languages: Interactions between Corpus and Lexicon, Istanbul, Turkey, May. Istanbul: European Language Resources Association (ELRA), pp. 13–18.

Declerck, T. (2022): Towards a new ontology for sign languages. In: Proceedings of LREC 2022.

Gennari, R./di Mascio, T. (2007): An ontology for a web dictionary of Italian Sign Language. In: Proceedings of the Third International Conference on Web Information Systems and Technologies. Vol. 1: WEBIST, pp. 206–213.

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

Jennings, V./Elliott, R./ Kennaway, R./Glauert, J. (2010): Requirements for a signing avatar. In: Hanke, T. (ed.): 4 th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. Valletta, Malta, 22–23 May 2010. Valletta, pp. 133–136.

Neves, C./Coheur, L./Nicolau, H. (2020): HamNoSyS2SiGML: Translating HamNoSys Into SiGML. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, pp. 6035–6039.

Ormel, E./Crasborn, O./van der Kooij, E./van Dijken, L./Nauta, E. Y./Forster, J./Stein, D. (2010): Glossing a multi-purpose sign language corpus. In: Dreuw, Philippe et al. (eds.): Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, Valletta, Malta. Valletta, pp. 186–191.

Quer, J./Cecchetto, C./Donati, C./Geraci, C./Kelepir, M./Pfau, R./Steinbach, M. (eds.) (2017): 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 (last access: 27-05-2022).

Yin, K./Moryossef, A./Hochgesang, J./Goldberg, J./Alikhani, M. (2021): Including signed languages in natural language processing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, August 1–6, 2021. Vol. 1: Long papers, pp. 7347–7360.

Contact information

Thierry Declerck
DFKI GmbH, Multilinguality and Language Technology, Saarland Informatics Campus D3 2, Saarland, Germany
declerck@dfki.de

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). It is also supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreement no. 731015). We would like to thank the anonymous reviewers for their helpful comments.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.