Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3977–3983
Marseille, 20-25 June 2022

Towards a new Ontology for Sign Languages

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

Abstract

We present the current status of a new ontology for representing constitutive elements of Sign Languages (SL). This development emerged from investigations on how to represent multimodal lexical data in the OntoLex-Lemon framework, with the goal to publish such data in the Linguistic Linked Open Data (LLOD) cloud. While studying the literature and various sites dealing with sign languages, we saw the need to harmonise all the data categories (or features) defined and used in those sources, and to organise them in an ontology to which lexical descriptions in OntoLex-Lemon could be linked. We make the code of the first version of this ontology available, so that it can be further developed collaboratively by both the Linked Data and the SL communities.

Keywords: Sign Languages, Ontology, Linguistic Linked Open Data, OntoLex-Lemon

1. Introduction

There is in the field of electronic lexicography an increasing interest in offering ways to represent and interlink lexical data originating from different modalities. This topic is particularly discussed within initiatives and projects[1] concerned with the representation of lexical data in a Linked Data (LD) compliant format, so that they can be published within the Linguistic Linked Open Data (LLOD) cloud.[2] In this context, we could observe that Sign Language (SL) lexical data are not represented in the 227 datasets included by now in the LLOD cloud. Also looking at the "Overview of Datasets for the Sign Languages of Europe" (Kopf et al., 2021) published by the "Easier" European project,[3] we do not see any mention of a dataset being available in an LD compliant format.

A first intuition for helping to remedy the situation was 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 used in ontologies, 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 2.

But we soon realized that considering only an extension to OntoLex-Lemon would not do justice to the specificity and richness of SL lexical data and that a thorough LD compliant description of the constitutive elements of SLs would be needed, so that those can be linked at various levels to lexical descriptions in OntoLex-Lemon.

In this paper, we describe first briefly the LLOD cloud and the OntoLex-Lemon framework. We follow by detailing our approach for building a new ontology for constitutive elements of sign languages, before sketching in the conclusion possible ways to interlinking elements of this ontology with lexical data as they are represented in the OntoLex-Lemon model.

2. Linguistic Linked Open Data Cloud

The Linguistic Linked Open Data (LLOD) cloud[4] is an initiative started in 2012 by a group of the Open Knowledge Foundation.[5] The aim was to break the data silos of linguistic data and thus encourage NLP applications that need to use data from multiple languages, categories (e.g., lexicon, corpora, etc.) and develop novel algorithms for Semantic Web applications involving natural language data. The LLOD cloud is one of the largest subsets of the Linked Open Data (LOD) cloud, with 227 datasets out of a total 1301 in the whole LOD.[6]

Looking at the current state of the LLOD, displayed in Figure 1, one can see that the data sets published in this cloud are classified along the lines of seven categories:

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries

[1] For example the H2020 Elexis (`https://elex.is/`) and Prêt-à-LLOD (`https://pret-a-llod.github.io/`) projects, or the COST Action CA18209 "NexusLinguarum - European network for Web-centred linguistic data science" (`https://nexuslinguarum.eu/`). See also (Declerck et al., 2020) for details on the contributions of those projects to the LLOD framework.

[2] See `http://www.linguistic-lod.org/`.

[3] See `https://www.project-easier.eu/` for more details on this project.

[4] `https://linguistic-lod.org/llod-cloud`

[5] See (McCrae et al., 2016).

[6] The total number of datasets (and links between those) is available at `https://lod-cloud.net`.

- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases
- Other

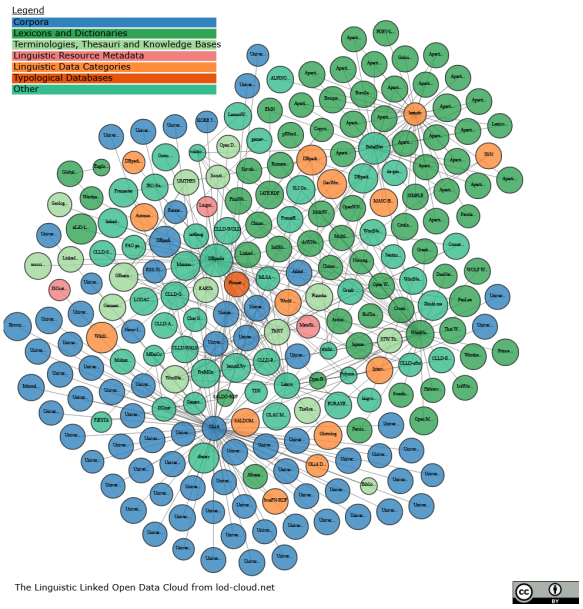

Figure 1: The Linguistic Linked Data Cloud, taken from `http://linguistic-lod.org/llod-cloud` [2022.04.23]

Not all the data sets are equally linked to each other, and some initiatives are contributing, for example. in better linking the data sets in the fields of Terminologies, Thesauri and Knowledge Bases and those in the fields of Lexicons and Dictionaries.

The goal of our current work is to support the representation and the interlinking of language data of different modalities, investigating for now how to represent SL lexical data so that it can be linked to the OntoLex-Lemon framework, which is briefly introduced in the next section.

3. OntoLex-Lemon

The OntoLex-Lemon model, which is resulting from a W3C Community Group,[7] was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the labels, definitions or comments of ontology elements are equipped with an extensive linguistic description.[8] This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as their syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialized vocabularies.

The main organizing unit for those linguistic descriptions is the *LexicalEntry* class, which enables, among others, the representation of morphological patterns for each entry (a multi-word expression, a word or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the *denotes* property or is mediated by the *LexicalSense* or the *LexicalConcept* classes, as this is represented in Figure 2, which displays the core module of the model.

OntoLex-Lemon builds on and extends the *lemon* model (McCrae et al., 2012). A major difference is that OntoLex-Lemon includes an explicit way to encode conceptual hierarchies, using the SKOS[9] standard. As can be seen in Figure 2, lexical entries can be linked, via the *ontolex:evokes* property, to such SKOS concepts, which can represent WordNet synsets. This structure is paralleling the relation between lexical entries and ontological resources, which is implemented either directly by the *ontolex:reference* property or mediated by the instances of the *ontolex:LexicalSense* class.

More recently, OntoLex-Lemon has been used also as a de-facto standard in the field of digital lexicography and is being applied for example in the European infrastructure project Elexis (European Lexicographic Infrastructure),[10] for which the interlinking of lexical data from different modalities is of high relevance.

Discussions about the representation of multimodal language data, with a first focus on SLs, both for lexicons and corpora, have been initiated within the "Ontology Lexica" W3C Community, in the context of an extension of the OntoLex-Lemon model dealing with frequency, attestation and corpus information of the lexicon model for ontologies.[11]

But, as stated already, it turned out that there is a need for a thorough LD compliant description of the constitutive elements of sign languages, in order to be able to link them to lexical descriptions of the written or spoken language modalities. Therefore, we started our work in building an ontology for those constitutive elements of sign languages. We describe the current state of this ontology after giving a brief (and certainly oversimplified) description of constitutive elements of sign languages.

[7] See `https://www.w3.org/2016/05/ontolex/`

[8] See (McCrae et al., 2012) and (Cimiano et al., 2016).

[9] SKOS stands for "Simple Knowledge Organization System". SKOS provides "a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary" (`https://www.w3.org/TR/skos-primer/`).

[10] See (Krek et al., 2018) and `http://www.elex.is/` for more details.

[11] This potential extension module is therefore called "FrAC". See (Chiarcos et al., 2020) and `https://acoli-repo.github.io/ontolex-frac/formoredetails`.

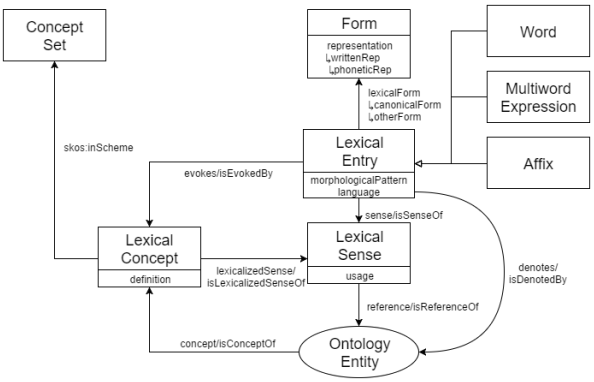

Figure 2: The core module of OntoLex-Lemon, taken from `https://www.w3.org/2016/05/ontolex/`

4. A brief Characterization of Sign Languages

Sign Language is a type of natural language with distinctive properties.[12] Those distinctive properties pose a challenge for the sign language lexical data to be linked to OntoLex-Lemon, as SL representation and interpretation involve a huge number of descriptors, including information about "physical" (body parts), spatial (orientation, movements, etc.) and temporal (duration of a sign) elements, which are usually not playing a role when it comes to represent the "classical" lexical data in the spoken and written languages.

This complexity of the SL lexical data and the challenges it poses for its full formal representation in the OntoLex-Lemon lexical framework is leading to both an LD-compliant representation of the constitutive elements of sign languages and to the design of a specific module extension of OntoLex-Lemon, 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, which are displayed in Figure 2.

Figure 3 gives a good overview of various ways of representing sign language data (here dealing with the American Sign Language (ASL), taken from (Yin et al., 2021)), with three of them being notational representations of the video or the pose streams: SignWriting (Bianchini, 2021), HamNoSys (Hanke, 2004)[13] and glosses.

Glosses are in most cases looking like normal (capitalized) words, with some additional notational conventions, but lacking in general detailed lexical information. We are planing to introduce a special class in OntoLex-Lemon for encoding the glosses, as those play a specific role in the annotation of signs, although they can not be considered as an accurate (semantic) interpretation of the sign (or sequence of signs) they are associated with. Rather, glosses can be considered as a way to label a sign (or a sequence of signs), as very often a corresponding lexicon that could be used for annotating a sign (or a sequence of signs) is lacking.[14]

Central elements of sign languages we want to focus on are 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 duration and repetitions, etc.[15] Those features are taken in consideration within the SignWriting and HamNoSys notational systems. We are currently investigating HamNoSys and describe its influence on the building of our ontology in Section 5.

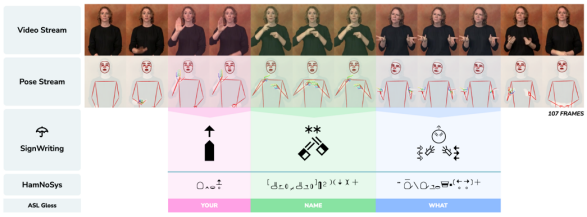

Figure 2: Various representations of American Sign Language. English translation: "What is your name?"

Figure 3: Various representations of American Sign Language, taken from (Yin et al., 2021)

5. The current Status of the Sign Language Ontology

(Gennari and di Mascio, 2007) already proposed an ontology for the Italian Sign Language, but this ontology is no longer available, apart from its description in the cited paper. We therefore propose a new ontology, going beyond the one described in (Gennari and di Mascio, 2007), including more recent descriptions of Sign Languages, and not limited to the Italian case, and also with the goal of supporting the linking of the new ontology to formal descriptions of SL lexical data in OntoLex-Lemon.

We started our work by an in-depth study of the literature dedicated to the properties of sign languages and of sites offering descriptions of and applications for SLs. We could extract a huge number of features (or data categories) for our work. We are reusing elements from, among others:

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

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

[14] See (Ormel et al., 2010) and (Crasborn and de Meijer, 2012), among others, for more details on issues related to glossing.

[15] And for sure, one also needs to take into account the representation of facial features, like eyebrow raise or mouthing.

- the CLARIN concept repository (https://www.clarin.eu/content/clarin-concept-registry), with 115 concepts related to Sign Language.
- the ASL-LEX database and its visualization tool (https://asl-lex.org/visualization/), with ca 95 features distributed over 7 main classes: Frequency Properties, Iconicity Properties, Lexical Properties, Sign Duration, Phonology, Phonological Calculations, and Acquisition Information.
- the British Sign Language Dictionary (https://www.british-sign.co.uk/british-sign-language/dictionary/), which contains, among others, a detailed textual descriptions for 484 signs.
- the DGS-Korpus project (https://www.idgs.uni-hamburg.de/en/forschung/forschungsprojekte/dgs-korpus.html), with a focus on HamNoSys, which breaks out a sign in four classes: hand shape, orientation, location, and actions, as can be seen in Figure 4. This SL transcription scheme was very helpful for our ontological classification of features (or data categories) for Sign Languages.
- 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. A very rich and long document (close to 900 pages), with a huge number of features distributed over 5 overarching sections: Phonology, Lexicon, Morphology, Syntax, and Pragmatics. For now, we focused on the Phonology and Lexicon sections of this monumental work.
- published grammars of sign languages (CA, DE, IT, FR, NL, TR, SP) resulting from the SIGN-HUB project (https://www.sign-hub.eu/project), which are following the SignGram Blueprint recommendations.

As already mentioned, our ontology is also reusing elements of the former ontology for the Italian Sign Language, which is described in (Gennari and di Mascio, 2007).
Our approach consisted mainly in proposing an initial harmonisation 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 ontology elements, organised in a tentative hierarchy. The ontology is imported in the

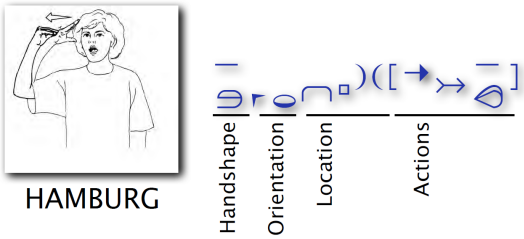

Figure 4: The four main classes of HamNoSys for the sign glossed with "HAMBURG". Taken from https://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/HamNoSys_2018.pd

OntoLex-Lemon ontology, which is also including the LexInfo[16] and SKOS[17] vocabularies. This step was done for preparing the possible linking of SL elements to OntoLex-Lemon lexical descriptions for spoken languages.
In this hierarchy, we propose the following top-level classes (where the prefix "sldc" stands for "Sign Language Data Category":

- sldc:ConstructedAction, which is dealing with the constructed repetition of elements of multiple character's behaviour, including communicative actions.
- sldc:NotationsForSignLanguages, which is for the time being only concerned with notational elements of HamNoSys.
- sldc:Sign, collecting information about the sign itself, as a single morpheme. This class is subdivided in subclasses about Duration, Fingerspelling, Lexical Properties (for example if a sign is a compound, a loan sign, or member of the core lexicon, etc.), and Phonology (which is including 239 elements, being subclasses or instances). The Phonology class is divided in two sub-classes, one on Manual and one on NonManual signs. All the SL relevant body or facial parts, their orientation, interaction and movements are included as instances of one of those two classes.
- sldc:SignLanguage, which is about definitions of various types of sign languages, for example AuxiliarySignLanguage, HomeSign, SignSupportedSpeech, etc.

[16]LexInfo is an ontology that was defined to provide data categories for the Lemon model and has been updated, in its version 3.0, for compliance with the new OntoLex-Lemon model. See https://lexinfo.net/ or (Cimiano et al., 2011) for a description of the first version of LexInfo.

[17]SKOS stands for "Simple Knowledge Organization System", and is used for encoding light-weight ontologies, like thesauri, taxonomies, etc. See https://www.w3.org/TR/skos-reference/ for more details.

- `sldc:SpokenLanguage`, which is about an auditory-visual language used primarily by a hearing community.

The ontology is for the time being discussed within the W3C Ontolex Community Group, in cooperation with members of the NexusLinguarum COST Action,[18] and this first version of the ontology is publicly available on Github.[19] Figure 5 gives a partial view on the hierarchy of classes, as currently suggested.

6. Related Work

(Elsayed and Fathy, 2020) make use of an ontology (and deep learning approaches) for supporting Sign Languages machine translation. But the ontology is in this case referring to a wordnet extension to include signs.

The EasyTV project has developed an ontology, which is "oriented to the inclusion and link of concepts representing Sign Language videos with the linguistic information ..." and provides for "A review of existing ontologies for linguistic information ... to annotate Sign Language videos. ... The current ontology is based on a core module in which sub-models from SKOS, BabelNet, lemon and LexInfo ontologies are reused."(Konstantinidis et al., 2020). EasyTV is thus primarily considering linguistics ontologies for the annotation of SL videos, but not an ontology for describing the Sign Languages.

(Sugandhi et al., 2020) propose a very shallow top level classification of signs used in Indian Sign Language, which we combined with the deeper ontology proposed for the Italian Sign Language and other resources, for inclusion in our ontology.

(Bonial et al., 2016) discuss the potential advantages of an event ontology for supporting multimodal applications. We will investigate if and how it is possible to combine our ontology data with the proposals described in this paper.

7. Conclusion

We presented ongoing work on including multimodal lexical data into the OntoLex-Lemon framework. We started this endeavour by considering the integration of Sign Languages lexical data. This type of data is challenging, as it includes the description of physical and spatial elements we do not have in the lexical data transported by the spoken or written languages. Our study of numerous sources and resources dealing with Sign Languages led us to the development of an ontology containing and describing many harmonised data categories for (or features of) SL lexical data, which are used in a disparate way in all those sources and resources.

This ontology should work as a reference point describing those physical and spatial elements of SL, as well as their transcriptions and glosses, to which lexical descriptions of spoken languages will be linked. We need to establish a way to relate the transcribed signs and their labels to OntoLex-Lemon lexical representations. Current work is therefore dedicated to the design of an extension module to OntoLex-Lemon that would allow this kind of linking.

8. 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 Horizon 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. We would also like to thank the anonymous reviewers for their helpful comments on the first submitted version of this paper.

9. 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.

Bonial, C., Tahmoush, D., Brown, S. W., and Palmer, M. (2016). Multimodal use of an upper-level event ontology. In Martha Palmer, et al., editors, *Proceedings of the Fourth Workshop on Events, EVENTS@HLT-NAACL 2016, San Diego, California, USA, June 17, 2016*, pages 18–26. Association for Computational Linguistics.

Chiarcos, C., Ionov, M., de Does, J., Depuydt, K., Khan, A. F., Stolk, S., Declerck, T., and McCrae, J. P. (2020). Modelling frequency and attestations for OntoLex-lemon. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 1–9, Marseille, France, May. European Language Resources Association.

Cimiano, P., Buitelaar, P., McCrae, J. P., and Sintek, M. (2011). Lexinfo: A declarative model for the lexicon-ontology interface. *J. Web Semant.*, 9(1):29–51.

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

Crasborn, O. and de Meijer, A. (2012). From corpus to lexicon: the creation of ID-glosses for the Corpus NGT. In Onno Crasborn, et al., editors, *Proceedings*

[18] More specifically in the context of its Task 3.4: `https://nexuslinguarum.eu/the-action/working-groups`

[19] See https://github.com/Declerck/sl-onto

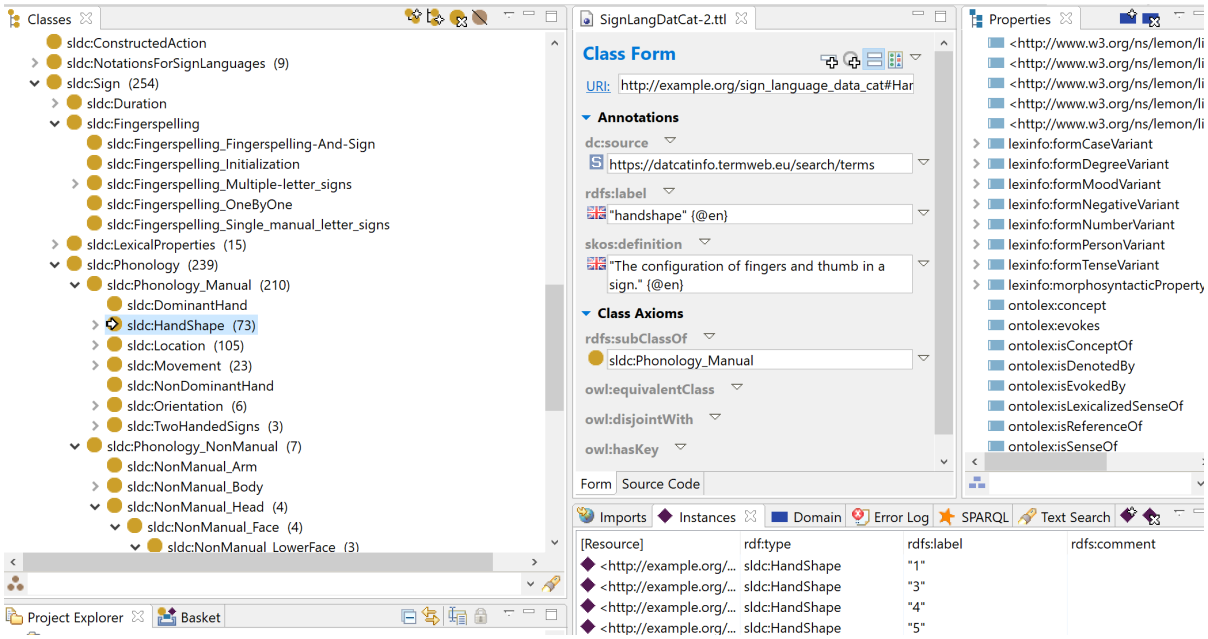

Figure 5: A screenshot of the ontology, displaying parts of its tentative hierarchy of classes

of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, pages 13–18, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Declerck, T., McCrae, J., Hartung, M., Gracia, J., Chiarcos, C., Montiel, E., Cimiano, P., Revenko, A., Sauri, R., Lee, D., Racioppa, S., Nasir, J., Orlikowski, M., Lanau-Coronas, M., Fäth, C., Rico, M., Elahi, M. F., Khvalchik, M., Gonzalez, M., and Cooney, K. (2020). Recent developments for the linguistic linked open data infrastructure. In Nicoletta Calzolari, et al., editors, *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 5660–5667. ELRA, ELRA, 5.

Elsayed, E. K. and Fathy, D. R. (2020). Sign language semantic translation system using ontology and deep learning. *International Journal of Advanced Computer Science and Applications*, 11(1).

Gennari, R. and 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 - Volume 2: WEBIST,*, pages 206–213. INSTICC, SciTePress.

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).

Konstantinidis, D., Dimitropoulos, K., Stefanidis, K., Kalvourtzis, T., Gannoum, S., Kaklanis, N., Votis, K., Daras, P., Rovira-Esteva, S., Orero, P., Uribe, S., Moreno, F., Llorente, Á., Calleja, P., Poveda-Villalón, M., Andriani, P., Vitolo, G., Caruso, G., Manes, N., Giacomelli, F., Fabregat, J., Mas, F., Mata, J., Skourtis, S., Bourlis, C., Frittelli, G., Lago, E. F., and Alvarez, F. (2020). Developing accessibility multimedia services: the case of easytv. In Fillia Makedon, editor, *PETRA '20: The 13th PErvasive Technologies Related to Assistive Environments Conference, Corfu, Greece, June 30 - July 3, 2020*, pages 38:1–38:8. ACM.

Kopf, M., Schulder, M., and Hanke, T. (2021). Overview of Datasets for the Sign Languages of Europe. https://doi.org/10.25592/uhhfdm.9561, July.

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B. S., Tiberius, C., and Wissik, T. (2018). European lexicographic infrastructure (elexis). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*, pages 881–892.

McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–719.

McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A., and Pool, J. (2016). The open linguistics working group: Developing the linguistic linked open data

cloud. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 9, rue des Cordelières, 75013 Paris, 5. ELRA, ELRA.

Ormel, E., Crasborn, O., van der Kooij, E., van Dijken, L., Nauta, E. Y., Forster, J., and Stein, D. (2010). Glossing a multi-purpose sign language corpus. 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 186–191, Valletta, Malta, May. European Language Resources Association (ELRA).

Sugandhi, Kumar, P., and Kaur, S. (2020). Sign language generation system based on indian sign language grammar. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(4), apr.

Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., and 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 (Volume 1: Long Papers)*, pages 7347–7360, Online, August. Association for Computational Linguistics.