A Linked Data Approach for linking and aligning Sign Language and Spoken Language Data

Thierry Declerck, Sam Bigeard, Fahad Khan, Irene Murtagh, Sussi Olsen, Mike Rosner, Ineke Schuurman, Andon Tchechmedjiev, Andy Way

A Linked Data Approach for linking and aligning Sign Language and Spoken Language Data

Thierry Declerck[1], **Sam Bigeard**[2], **Anas Fahad Khan**[3], **Irene Murtagh**[4], **Sussi Olsen**[5], **Michael Rosner**[6], **Ineke Schuurman**[7], **Andon Tchechmedjiev**[8], **Andy Way**[9]
[1]DFKI GmbH, Multilingual Technologies, Saarland Informatics Campus, D-66123 Saarbrücken, Germany
[2]Institute of German Sign Language and Communication of the Deaf University of Hamburg, Germany
[3]CNR-ILC, Italy
[4]Technological University Dublin, Ireland
[5]Centre for Language Technology, NorS, University of Copenhagen, Denmark
[6]Dept Artificial Intelligence, University of Malta
[7]Centre for Computational Linguistics, KU Leuven, 3000 Leuven, Belgium
[8]EuroMov-Digital Health in Motion, Univ. Montpellier, IMT Mines Alès, France
[9]ADAPT Centre, Dublin City University, Ireland
declerck@dfki.de, sam.bigeard@uni-hamburg.de, fahad.khan@ilc.cnr.it, irene.murtagh@adaptcentre.ie, saolsen@hum.ku.dk, mike.rosner@um.edu.mt, ineke.schuurman@ccl.kuleuven.be, andon.tchechmedjiev@mines-ales.fr, andy.way@adaptcentre.ie

Abstract

We present work dealing with a Linked Open Data (LOD)-compliant representation of Sign Language (SL) data, with the goal of supporting the cross-lingual alignment of SL data and their linking to Spoken Language (SpL) data. The proposed representation is based on activities of groups of researchers in the field of SL who have investigated the use of Open Multilingual Wordnet (OMW) datasets for (manually) cross-linking SL data or for linking SL and SpL data. Another group of researchers is proposing an XML encoding of articulatory elements of SLs and (manually) linking those to an SpL lexical resource. We propose an RDF-based representation of those various kinds of data. This unified formal representation offers a semantic repository of information on SL and SpL data that could be accessed for supporting the creation of datasets for training or evaluating NLP applications dealing with SLs, thinking for example of Machine Translation (MT) between SLs and between SLs and SpLs.

1 The Linguistic Linked Open Data Cloud and Sign Languages

Proponents of Linguistic Linked Open Data (LLOD) (Cimiano et al., 2020; Declerck et al., 2020) aim towards the representation of linguistic data through a standardised model based on the Resource Description Framework (RDF).[1] OntoLex-Lemon (Cimiano et al., 2016)[2] and its ecosystem (McCrae et al., 2017) are at the core of the LLOD cloud, and follow FAIR principles (Wilkinson et al., 2016)[3] to make linguistic data accessible and interoperable. This semantic interoperability allows for the interlinking of diverse linguistic datasets, establishing a well-connected subset of the Linked Open Data Cloud,[4] and creating avenues for analyses and studies long unattainable due to a history of barely interoperable formats. But the LLOD cloud does not currently include any Sign Language (SL) datasets, establishing the representation of SL data and Multimodality as a frontier for LLOD to accommodate.

[1]a W3C recommendation. See https://www.w3.org/RDF/ for more details.
[2]See the following for the published specifications: https://www.w3.org/2016/05/ontolex/
[3]Where FAIR stands for Findable, Accessible, Interoperable and Reusable and refers to a series of well-known principles for ensuring that datasets can be described by each of the former adjectives.
[4]http://cas.lod-cloud.net/clouds/linguistic-lod.svg

Declerck et al. (2023) discusses an RDF-based representation of the mapping between SL data and Spoken Language (SpL) resources via the Open Multilingual Wordnet (OMW) infrastructure, which is proposed in Bigeard et al. (2022). Elements of OntoLex-Lemon and cross-lingual linking techniques were used to create multilingual SL resources. Such work illustrates the potential to produce parallel training material at scale for MT between SLs or between SLs and SpLs.

These initial efforts have created momentum that has led to the explicit identification of SLs as a target for an extended representation within the OntoLex-Lemon model. This issue is also currently being discussed in the context of the BPMLOD W3C Community Group (detailed further in Section 3), which is producing a survey of existing best-practices to model linguistic (including SL) data as linked data.

One of the ways to ensure the interoperability of these heterogeneous resources, including across language types (SLs and SpLs), is through the use of FAIR principles for all aspects of the production/publication of the datasets (modelling, licensing, deposition in a repository, etc.).

We do not propose any new algorithms in this paper, but advocate for a standardised methodology for producing interoperable high-quality aligned datasets for SL and SpL (SSL) data using linked data and cross-lingual (within and across signed and spoken languages) technologies, as well as best practices and guidelines. For this, we need to involve various communities, and the W3C BPMLOD Community Group could offer a first forum for achieving our joint goals.

In the following, we first summarize the FAIR principles before introducing current, ongoing activities within the W3C BPMLOD Community Group. We then present four research initiatives dealing with the issue of SSL data alignments. For two of them, we already propose an RDF/OntoLex-Lemon modelization (Sections 4 and 5), while work is about to start for the SL data described in Sections 6 and 7.

2 FAIR Data and Linguistic Linked Open Data

FAIR data plays a central role in a number of prominent initiatives which have recently been proposed for the promotion of open science and data by numerous organisations and research funding bodies. We advocate that LLOD models can contribute to the creation of FAIR language resources.

It should come as no surprise, given the growing importance of open science initiatives and in particular those promoting the FAIR guidelines, that shared models and standardized vocabularies have begun to take on an increasingly prominent role within numerous disciplines, not least in the fields of linguistics and language resources. Although the linguistic linked data community has been active in advocating for the use of shared RDF-based vocabularies and models for quite some time now, this new emphasis on FAIR language resources is likely to have a considerable impact in several ways, in terms of the necessity for these models and vocabularies to demonstrate greater coverage with respect to the kinds of linguistic phenomena they can describe, and for them to be more interoperable with each other.

In *The FAIR Guiding Principles for scientific data management and stewardship* (Wilkinson et al., 2016), the article which first articulated the by-now ubiquitous FAIR principles, the authors state that the criteria proposed by those principles are intended both "for machines and people" and that they provide "'steps along a path' to machine actionability", where the latter is understood to describe structured data that would allow a "computational data explorer" to determine:

- the type of "digital research object";
- its usefulness with respect to tasks to be carried out;
- its usability especially with respect to licensing issues with this information represented in a way that would allow the agent to take "appropriate action".

The current popularity of the FAIR principles and, in particular, their promotion by governments, transnational organisations and research funding bodies, such as the European Commission,[5] reflects a wider recognition of the potential of structured, interoperable, machine-actionable data to help effect a major shift in how research is carried out, and in particular, its potential to help underpin open science best practices. The FAIR ideal,

[5] https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283

in short, is to allow machines (non-human software agents) a greater level of autonomy in working with data by rendering as much of the semantics of that data explicit (in the sense of machine-actionable) as possible.

Publishing data using a standardised, general purpose data model such as RDF[6] goes a long way towards facilitating the publication of datasets as FAIR data. RDF, taken together with the other standards proposed in the Semantic Web stack and the technical infrastructure which has been developed to support it, was specifically intended to facilitate interoperability and interlinking between datasets. In order to ensure the interoperability and reusability of datasets within a domain, however, it is vital that in addition to more generic data models such as RDF there also exist domain-specific vocabularies/terminologies/models and data category registries (compatible with the former). Such resources serve to describe, ideally in a machine-actionable way, the shared theoretical assumptions held by a community of domain experts as reflected in the terminology or terminologies in use within that community.

We note here that the emphasis placed on machine actionability in FAIR resources (that is, recall, on enabling computational agents to find relevant datasets and resources and to take "appropriate action" when they find them) gives Semantic Web vocabularies/models/registries a substantial advantage over other (non-Semantic Web-native) standards in the fields of linguistics and language resources. The OntoLex-Lemon ecosystem is to be understood in this light, aiming at enhancing the interoperability and machine actionability of linguistic datasets. It is, therefore, crucial to overcome the one limitation we noticed: there are for now no SL datasets within the LLOD, if we ignore the ongoing experiments in porting to RDF/OntoLex-Lemon the SL datasets (and their linking to OMW or other lexical resources) that are described in Sections 4, 5, 6, and 7.

3 The Best Practices for Multilingual Linked Open Data W3C Group

The BPMLOD W3C community group[7] initially created in 2015 to propose community-sourced guidelines for multilingual linked open data, has recently been resurrected in order to actualise the previously proposed guidelines, as there have been major evolutions in the field.

These renewed efforts have a much broader scope, covering topics such as neurosymbolic approaches to language processing, cross-lingual linking, multi-modality, and the representation of sign languages. The latter two specification efforts are central to establishing the foundational groundwork necessary for representing both SL and SpL data as RDF under the OntoLex family of models, as linked open data. The nature of semantic web technologies is conducive to easily enabling interlinking once both modalities can be represented in one harmonized formal model.

The BPMLOD Community Group has thus the potential of becoming a community nexus to channel work on semantic web models for SLs, SpLs and their linking. We encourage the widespread involvement of both the SL and the SpL communities in this initiative. As mentioned above, BPMLOD is currently working on a survey of existing best-practices to model linguistic (including SL) data as linked data.[8]

4 Aligning several SL Resources via the Open Multilingual WordNet Infrastructure

The work reported on in this section is developed within a research project, which aims to ease the communication between deaf and hearing individuals with the help of MT technologies. As such, linking different SLs through semantics is a priority. We chose to use the Open Multilingual Wordnet (OMW) infrastructure (Bond and Paik, 2012; Bond et al., 2016)[9] as a (semantic) pivot between SL data.

We are dealing with four languages (German, Greek, English and Dutch sign languages). The resources involved in our approach are the DGS corpus (Prillwitz et al., 2008), Noema+ GSL dictionary (Efthimiou et al., 2016), BSL signbank (Jordan et al., 2014), and the NGT global signbank (Crasborn et al., 2020). These resources contain various types of spoken language words associated with each sign. They may be keywords, equivalents, or SL glosses. They are used as a starting point to match with the lemmas present in

[6] https://www.w3.org/TR/rdf-primer/

[7] Best Practices for Multilingual Linked Open Data, see https://www.w3.org/community/bpmlod/

[8] The corresponding reports will be made available at https://github.com/bpmlod

[9] See also https://omwn.org/ for more details.

the corresponding (and aligned) language versions of OMW. Then, native signers manually validate the potential matches. By using the Open Multilingual Wordnet, we aim to identify the signs with the same (or related) senses across languages.

Each resource involved has different structures, and so, the method must be flexible enough to exploit all the data available and avoid mistakes. As an example, the DGS Corpus has a multi-level structure, where each sign can be a type, a subtype, or a variant. Semantics are attached to the sub-type level. If a sense has been associated with a sub-type, it can be spread down to the variants associated with it, but not up to the type. The DGS Corpus also contains synonymy links that can be exploited to spread senses to other signs.

We describe in the following paragraphs elements of SLs that need to and could be (semantically) aligned across languages and language types.

Phonological transcriptions: While in an ideal world, those transcriptions from videos displaying signs could be used for establishing links between SL data for different languages, different SL data sets are transcribed with different transcription systems, e.g. HamNoSys (Hanke, 2004), SignWriting (Sutton, 1991) or others, as in the case of the Swedish SL data[10] or Irish SL, for which an XML-based transcription is under development (see Section 6 for more details).

Besides, even if two resources use the same transcription system, the level of accuracy or precision of the transcription is not the same for all data. In some cases the transcription can be either semi-automatically generated or produced by human transcribers with different skills and views on which phonological elements of a sign should be transcribed.[11]

We are aware of efforts being made toward analysing and processing the videos directly using machine learning, rather than comparing and aligning transcriptions, but those are not in the scope of our current work.

Glosses: Many projects dealing with SL use glosses to identify signs. A gloss is, typically, a spoken language word optionally followed by a sequence of numbers or letters, to allow several signs to share the same word. The word is typically related to the meaning or iconicity of the sign, in the surrounding SpL, for easier identification. But the used word is ultimately somewhat arbitrary. Two unrelated projects working on the same sign language might have different glosses for the same sign, or the same gloss for different signs. This creates an obstacle toward linking resources together.

While many SL resources use glosses for labelling their data, the low accuracy/precision of automated tagging and the low Inter-Annotator Agreement (IAA) between human annotators for such tagging made the glosses difficult to use as a potential cross-language instrument for interlinking SL data in various languages.[12]

For linking to the IDs in OMW, we preferably use keywords and translations as a starting point to approximate the meaning of the sign, and only use glosses as a last resort. However, we use glosses as identifiers.

5 Cross-Linking Nordic SL and SpL Data

We extended our RDF representation of the language coverage described in Section 4 to three Nordic languages: Danish, Icelandic and Swedish.

Troelsgård and Kristoffersen (2018) discuss approaches for ensuring consistency between (Danish) Sign Language corpus data and the Dictionary of Danish signs. This approach aims at delivering a correspondence between the dictionary lemmas and the corpus lexicon, which consists of types introduced for lemmatising the tokens found in the corpus annotations (glosses added to the signs). The strategy is 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 described in Pedersen et al. (2009; Pedersen et al. (2018). This approach is thus very similar to the one described in Bigeard et al. (2022), but is 'limited' to the Danish language. The relations between sign identifiers and lexical elements from both DanNet and other dic-

[10] See (Bergman and Björkstrand, 2015) for a detailed description, and also `https://zrajm.github.io/teckentranskription/intro.html` on recent developments on a tool to support this transcription system.

[11] Power et al. (2022), for example, report in their experiment that the similarity (but not the exact matching) of transcriptions by two undergraduate research assistants working in a related project was 0.69.

[12] Forster et al. (2010) discuss, among others, best practices for gloss annotation, in order to mitigate the issues of divergent tagging results, even in one and the same corpus.

tionary sources are encoded in a database, from which we obtained a TSV export. Luckily for us, the wordnet elements encoded in this TSV export are the subset of DanNet entries that are contained in the Danish section of OMW.

In this export, we first have the signs, which correspond to entries in the Dictionary of Danish Signs (see Figure 1). 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, and it is not displayed at the web page. A third type of information included in the export concerns the senses associated with the signs and their (form) variants.

Our work consists thus in porting all those (interlinked) resources to RDF and OntoLex-Lemon, as we did for the data described in Section 4. 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", thus sharing the same ID for the concept of "protection" in OMW (this holds also for French, etc.). We can therefore add the Danish sign ID (and video), which we obtained from the database, to our RDF-based infrastructure.

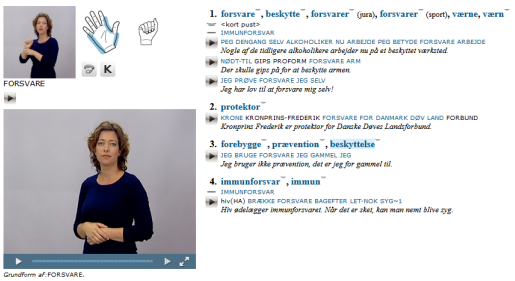

Figure 1: The Danish sign associated with the OMW ID "00817680-n", corresponding to the (highlighted) lemma "beskyttelse", here as one possible lexical realisation of the Danish gloss "FORSVARE" (*defend*)

Using the same strategy of deploying OMW as a pivot between concepts expressed in the videos, we extended our approach to Icelandic and Swedish. Through OMW we can find the lemmas for Icelandic and Swedish associated with the OMW IDs "1128193-v" and "00817680-n" (corresponding to the Danish lemmas). We use these to search in the Icelandic SignWiki,[13] and in the Swedish Sign Language Dictionary, described in Mesch et al. (2012).[14] Icelandic and Swedish glosses can be easily integrated in our RDF-based representation, as can be seen for example in Listing 1, where the gloss for the Danish sign depicted in Figure 1 is augmented with glosses or lemmas from other languages.

```
dts:GLOSS_dts-722
  rdf:type sl:GLOSS ;
  rdfs:label "\"FORSVARE\""@da ;
  rdfs:label "\"PROTEGER\""@fr ;
  rdfs:label "\"SCHUTZ1Aˆ\""@de ;
  rdfs:label "\"protect(v)#1\""@en ;
  rdfs:label "\"beskydd\""@se ;
  rdfs:label "\"Vernda \""@is ;
.
```

Listing 1: The RDF-based representation of the gloss "FORSVARE", with the integration of multilingual labels from corresponding glosses

We further extended this approach to other Nordic languages, as described in Declerck and Olsen (2023).

6 A new Transcription System for the Irish Sign Language

Building on work dealing with linguistic properties of the Irish Sign Language (Murtagh, 2019), a group of researchers was confronted with the question of what is needed for creating an SL lexicon entry, as they wanted to document or "write down" what was being signed or articulated in the videos.

While SpL and SL share fundamental properties in relation to linguistic structure, certain modality-specific linguistic phenomena must be accommodated in computational terms, to allow for the modelling and processing of SLs. A new transcription system was developed for this, which, contrary to HamNoSys or SignWriting, is not based on iconic symbols, but directly encoded in XML.

The Sign_A framework (Murtagh et al., 2022) was developed with a view to providing a definition of linguistically motivated lexicon entries, that were sufficiently robust to accommodate sign language, in particular Irish Sign Language (ISL). Sign_A provides a formal description for the computational phonological parameters of SL. A Sign_A XML specification is provided for manual features (MFs), non-manual features (NMFs), location (both spatial and body anchored) information, and also temporal information. MFs include parameters for Hand ⟨HAND⟩, Handshape ⟨HS⟩, Hand Movement ⟨HM⟩, Palm Orientation

[13] https://is.signwiki.org/index.php/

[14] https://teckenspraklexikon.su.se

⟨PO⟩, Arm Movement ⟨AM⟩, Forearm ⟨FA⟩and Upper arm ⟨UA⟩. In Figure 2, we can see the Sign_A XML representation for the hands, where the "dominant hand" is defined as ⟨dh⟩, and the "non-dominant hand" as ⟨ndh⟩.

```
<MF>
    <HAND>
        <dh>"right"</dh>
        <ndh>"left"</ndh>
    </HAND>
    ...
</MF>
```

Figure 2: Initialising the right hand as the dominant hand

The NMFs include parameters for Eyebrow ⟨EB⟩, Eyelid ⟨EL⟩, Eye Gaze ⟨EG⟩, Cheek ⟨CHEEK⟩, Mouth ⟨MOUTH⟩, Tongue ⟨TNG⟩, Nose ⟨NOSE⟩, Shoulder ⟨SHOULDER⟩, Mouthing ⟨MOUTHING⟩, and Mouth Gesture ⟨MOUTHGESTURE⟩.

The head element ⟨HEAD⟩, contains a ⟨HEADMODE⟩ attribute, which can accept various actions pertaining to the head, e.g. nod, shake, tilt, turn, etc. We provide the XML specification for nodding the head twice in Figure 3.

```
<NMF>
    <HEAD>
        <HEADMODE>"nod"</HEADMODE>
        <TIMES>"2"</TIMES>
        ...
    </HEAD>
</NMF>
```

Figure 3: Specification for nodding the head twice

Sign_A also includes parameters to accommodate the location in space where a sign is articulated. The location parameters can be mapped to spatial locations ⟨LOC⟩ within the signing space and also to locations on the signer's body, referred to as body-anchored ⟨BA⟩ locations. Finally, the formalism also includes an XML specification for temporal information, where each phonological parameter has timing information associated with it, referred to as event duration ⟨ED⟩. Sign_A also includes a timeline parameter ⟨TL⟩, which refers to the overall timing of an utterance. This parameter is used to synchronise the simultaneous and parallel articulation of any given phonological parameter 'event' across an entire SL utterance.

While Sign_A offers a very detailed description (and a taxonomic structure) of articulatory elements of SLs, its XML encoding also eases the conversion of the data into RDF, a task we are starting on now. Another relevant aspect of the work pursued in the context of Sign_A is the attention given to linking the described sign to SpL lexical resources, as can be seen in Figure 4, which is taken from Murtagh et al. (2022).

Gloss	REAL LOVE MY JOB
English Translation	'I really love my job'
RRG+Sign_A Logical Structure	**LOVE'** <TEMPORAL><MF><NMF> (1sg, JOB)
ISL Lexicon XML SL Verb Entry	
<ISLGlossTranslate="LOVE" IPA="/lʌv/" LogicalStructure= "LOVE' <TEMPORAL> <LOCATION><MF><NMF> (1sg, JOB);" NumberVerb="sg" P.O.S="PlainVerb" personVerb="3rd" tenseVerb="PRES" love/>	
Lexeme Repository Sign_A XML description for Manual Features <MF> of SL verb LOVE	
<HAND><dh>"right"</dh><ndh>"left"</ndh></HAND>	
<HS><HSMode>unique</HSMode><HSID><value>24</value></HSID>	
<AM><Spatial><SOURCE>"locus"</SOURCE><GOAL>"locus"</GOAL>EDti></EDti><EDtn></EDtn> <TLti></TLti><TLtn></TLtn> </SPATIAL><AM>	
<PO><p2><p2_i><EDti></EDti><EDtn></EDtn></p2_i><p1_n><EDti></EDti><EDtn></EDtn></p2_n><TLti></TLti><TLtn></TLtn></p1></PO>	
Lexeme Repository Sign_A XML description for Non Manual Features <NMF> of SL verb LOVE	
<MOUTHING><VERB_ONE_TO_ONE><VERBIPA>"/lʌv/"</VERBIPA></VERB_ONE_TO_ONE></MOUTHING	

Figure 4: ISL plain verb "LOVE" lexeme repository and lexicon XML description.

Porting this cross-language type linking to RDF and OntoLex-Lemon will contribute to a full linking between SL and SpL lexical data, beyond the work described in Sections 4 and 5, which focus on the specific multilingual wordnet-based lexical resources for cross-linking SL data. We plan to link the Sign_A SL data to the DBnary resource (Sérasset and Tchechmedjiev, 2014; Sérasset, 2015) which represents lexical information extracted for 23 language editions available from Wiktionary in a way compliant with Linked Open Data.

7 SignNets - WordNets for a specific Type of Natural Language

In the Northern part of Belgium (Flanders), the official language is Dutch; in the Southern part (Wallonia), it is French. There are also two officially acknowledged sign languages, VGT (Flemish Sign Language) and LSFB (French Belgian Sign Language). Dutch is also the official spoken language in the Netherlands, but the officially acknowledged sign language is NGT (Dutch Sign Language). In this section, we concentrate on VGT and NGT, and the link with another natural language: spoken Dutch.[15] VGT and NGT are rather different SLs, having themselves developed quite independently. VGT tends to share characteristics with LSFB, even though they are growing apart. Nev-

[15]There are other sign languages which will have similar issues to solve, like ISL for which the surrounding spoken language is the variant of English used in Ireland. Another characteristic of ISL to be taken into account is that it is a gender-based SL, where men and women have different sign languages.

ertheless, similarities between VGT and NGT are noted especially when dealing with iconic signs, or when mouthing plays an important role, since in both cases the surrounding SpL is Dutch.

When linking via WordNet (OMW) it should be stressed that the glosses assigned to signs in fact represent a (semantic) concept instead of just words, i.e. they represent SpL synsets instead of a word belonging to such a synset. The gloss can even represent several parts of speech. Glosses used for specific concepts[16] may differ in NGT and VGT, but even the two providers of NGT data[17] may use different glosses for one and the same sign. In all these cases, even the part of speech of the chosen gloss may differ. [18]

In contrast, VGT and NGT may use the same gloss for different signs. Within VGT, one and the same gloss often represents a series of signs, all expressing the same concept. This is due to the regional variations of a sign, a property of VGT explicitly preserved by the Deaf Community after the official recognition of VGT. Note that especially older variants may disappear, while new ones pop up. In the Netherlands, the situation was the reverse: one sign per concept was pursued.[19] The VGT gloss will express the common concept. In both the NGT Signbank and the VGT dictionary, indicative translations in spoken Dutch are included to indicate the concept expressed. Quite often these represent several parts of speech like nouns and verbs, nouns and adjectives, etc.[20] We are linking these to the synsets per PoS included in OMW, but are also creating new, broader identifiers to link them to SL concepts, surpassing PoS differences.

In SignNet (Schuurman et al., to be published),[21] VGT thus comes with synsets of signs, whereas NGT usually does not. In SignNet signs (concepts) and words in spoken language are linked, using OMW, and adding hyponyms, hypernyms, homonyms, definitions of the concepts, etc.

There are at least two issues in doing so: first, OMW makes use of Open Dutch Wordnet, and ODW (and OMW) often use the Dutch meaning of a word, not the Flemish one. For example, 'voormiddag' refers to the hours before lunchtime in Flanders, and after lunch in the Netherlands. So we have to adapt ODW (and OMW) to cover such differences. We intend to do so by adding in ODW (and OMW) a 'geography' label "belg" to words that only are used in a specific sense in Flanders ('kleedje' instead of 'jurk' (dress)) or "ned" when the word is only used in the Netherlands ('kinderkopje' instead of 'kassei' (cobblestone)).

A second issue: quite often concepts labelled by one gloss in VGT (and NGT) cover more than one synset in the wordnet of the surrounding language, for example when several parts of speech are involved. However, sometimes also smaller sets are used: artists using voice taken together (singer, actor) vs artists not using voice (ballet dancer, painter, ...). Ebling et al. (2012) describe similar cases for the Swiss-German SL. And for example when the sign is rather iconic, showing a vertical versus a horizontal movement. In Dutch, there is the verb 'aanhaken' (hook on), used both to express hooking a painting on a hook in a wall (vertical) and hooking a trailer on a car (horizontal). In VGT and NGT, there are two different signs that respectively show a more vertically or horizontally oriented movement. Because this difference is not made in SpL, it is neither represented in ODW nor OMW, so we may need to adapt ODW in this respect as well.

Considerations of the similarities and differences between the two variants of the Dutch SLs and of the Dutch SpLs point to the need to properly address linguistic variations, if one wants to adequately interlink or align those variants across languages and language types. It seems that the current status of the OMW infrastructure cannot offer Wordnet IDs to serve as pivot in those cases. We thus need to address those issues in the next steps of our representation work in RDF, and to investigate whether the current "vartrans" module[22] of OntoLex-Lemon is adequately formulated for this

[16] Concept, not sign!

[17] Nederlands Gebarencentrum `https://www.gebarencentrum.nl` and the NGT part of the Global Signbank `https://signbank.cls.ru.nl/datasets/NGT`.

[18] In NGT the gloss for the concept covering 'arm' (poor) is BEHOEFTIG (an adjective), in VGT it is ARMOEDE (a noun).

[19] When in NGT more signs are covered using variants of the same gloss (BEHOEFTIG-A, BEHOEFTIG-B), quite often the coverage of the semantic concept differs. BEHOEFTIG-B can also mean 'broke', not only 'poor', which does not hold for BEHOEFTIG-A.

[20] Vossen (1999) refers to such words as being Near-Synonyms, referring to the EQ NEAR SYNONYM relation between 'aardig' (Adjective) in Dutch and 'to like' (verb) in English.

[21] Based on SL dictionaries, signbanks etc.

[22] See `https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans` for more details.

task. An important lesson we can retain from this section is that the generation of parallel data for SL and SpL language variations is a challenging task.

8 A first Implementation of linking and aligning Strategies in RDF/OntoLex

Listing 1 has already shown how we can encode in RDF a Danish gloss and augment it with glosses or lemmas from other languages, which we extracted via the shared IDs implemented in OMW, pointing back to the Danish video equipped with the corresponding gloss. With the next Listings, we would like to give an idea of how the RDF and OntoLex-Lemon representation ensures the accurate linking of information in a standardized and interoperable way.

Listing 2 shows the encoding of the Danish video already displayed in Figure 1 above, and Listing 3 shows the RDF-based representation of the corresponding gloss.

```
<http://example.org/dts#
    SignVideos_dts-722.mp4>
 rdf:type sl:SignVideos ;
 sl:hasGLOSS dts:GLOSS_dts-722 ;
 sl:hasVideoAdresss "https://www.
    tegnsprog.dk/video/t/t_2162.mp4
    "^^rdf:HTML ;
 rdfs:label "\"Video annotated with
    the gloss 'FORSVARE'\""@en ;
.
```

Listing 2: The video annotated with the gloss "FORSVARE" as an instance of the RDF class "sl:SignVideos"

```
dts:GLOSS_dts-722
  rdf:type sl:GLOSS ;
  rdfs:label "\"FORSVARE\""@da ;
.
```

Listing 3: The RDF-based representation of the gloss "FORSVARE"

Listing 4 shows a corresponding lexical form (in this case a lemma taken from OMW) and links it to the video and to the gloss it is related to, also adding the SiGML notation, which is the XML transcription of the original HamNoSys code (Neves et al., 2020).

```
dts:Form_dts-722
  rdf:type ontolex:Form ;
  sl:hasGLOSS dts:GLOSS_dts-722 ;
  sl:hasVideo <http://example.org/dts#
      SignVideos_dts-722.mp4> ;
  sl:hasVideoAdresss "https://www.
      tegnsprog.dk/video/t/t_2162.mp4"^^
      rdf:HTML ;
  rdfs:label "\"Adding transcription
      information associated with the
      video with the gloss 'FORSVARE'\""
      @en ;
  ontolex:writtenRep "\"<sigml><hns_sign
       gloss='FORSVARE'><hamnosys_manual
      ><hamsymmlr/><hamfist/><hamparbegin
      /><hamextfingeru/><hampalmd/><
      hamplus/><hamextfingerr/><hampalmr
      /><hamparend/><hamparbegin/><
      hammoveu/><hamthumbside/><hamtouch
      /><hamplus/><hamnomotion/><
      hamparend/><hamrepeatfromstart/></
      hamnosys_manual></hns_sign></sigml
      >\"\""@hamnosys-sigml ;
  ontolex:writtenRep "\"beskyttelse\""
      @da ;
.
```

Listing 4: The RDF-based representation of the lexical form related to the gloss "FORSVARE" and the corresponding video

Finally, Listing 5 displays the lexical entry for which the form is a morphological realisation. The lexical entry is pointing to the OMW ID realised as a lexical concept in OntoLex-Lemon, and which itself points to the video annotated by the one gloss.

```
dts:LexicalEntry_722
  rdf:type ontolex:LexicalEntry ;
  rdfs:label "\"forsvare, beskytte,
      beskyttelse\""@da ;
  ontolex:evokes wnid:omw-00817680-n ;
  ontolex:lexicalForm dts:Form_722 ;
.
```

Listing 5: The RDF-based representation of the lexical entry, which relates the concept and the form

The full RDF code will be made available in a GitHub repository, so that interested colleagues can contribute to future developments.

9 Conclusion

We proposed in this paper to investigate the possibilities of a harmonised representation of data from both spoken and sign languages that were originally stored in different formats in different locations. Basing ourselves on the works and issues presented in Sections 4, 5, 6 and 7, we propose the use of RDF and associated standardized vocabularies or models (like OntoLex-Lemon) to support an interoperable encoding for constitutive elements of both SL and SpL resources and their interlinking and alignment, whilst also stressing the importance of following the principles of FAIR data.

We hope in this way to create a semantically organized repository of cross-lingual (both SLs and SpL) data, especially in the field of low-resource SLs, which can be of help for supporting the creation of data sets for training or evaluating NLP applications, thinking in the first place of automated translation.

Acknowledgements: Some of the work described is supported in part by the EASIER (Intelligent Automatic Sign Language Translation) Project. EASIER has received funding from the European Union's Horizon 2020 research and innovation programme, grant agreement n° 101016982. Other parts of the work described is carried out within the SignON (Sign Language Translation Mobile Application and Open Communications Frame) project. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant No. 101017255. Even other parts of the contribution to this work is supported by the LT-BRIDGE project, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 952194. Contributions by many authors are done within the context of the COST Action NexusLinguarum – European network for Web-centred linguistic data science - (CA18209).

We would also like to thank the reviewers for their helpful comments on our submission.

References

[Bergman and Björkstrand2015] Bergman, Brita and Thomas Björkstrand. 2015. Teckentranskription. Technical Report XXV, Stockholm University, Sign Language.

[Bigeard et al.2022] Bigeard, Sam, Marc Schulder, Maria Kopf, Thomas Hanke, Kiki Vasilaki, Anna Vacalopoulou, Theodoros Goulas, Athanasia-Lida Dimou, Stavroula-Evita Fotinea, and Eleni Efthimiou. 2022. Introducing Sign Languages to a Multilingual Wordnet: Bootstrapping Corpora and Lexical Resources of Greek Sign Language and German Sign Language. In Efthimiou, Eleni, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, Johanna Mesch, and Marc Schulder, editors, *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 9–15, Marseille, France, June. European Language Resources Association (ELRA).

[Bond and Paik2012] Bond, Francis and Kyonghee Paik. 2012. A Survey of WordNets and their Licenses. In *Proc. of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 64–71.

[Bond et al.2016] Bond, Francis, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*, volume 2016.

[Cimiano et al.2016] Cimiano, Philipp, John McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report, 10 May 2016. Technical report, W3C, May.

[Cimiano et al.2020] Cimiano, Philipp, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing.

[Crasborn et al.2020] Crasborn, Onno, Richard Bank, Inge Zwitserlood, Els van der Kooij, Ellen Ormel, Johan Ros, Anique Schüller, Anne de Meijer, Merel van Zuilen, Yassine Ellen Nauta, Frouke van Winsum, and Max Vonk. 2020. Ngt dataset in global signbank.

[Declerck and Olsen2023] Declerck, Thierry and Sussi Olsen. 2023. Linked open data compliant representation of the interlinking of nordic wordnets and sign language data. In Ilinykh, Nikolai, Felix Morger, Dana Dannélls, Simon Dobnik, Beáta Megyesi, and Joakim Nivre, editors, *Proceedings of the 2nd Workshop on Resources and Representations for Under-Resourced Languages and Domains*, pages 62–69.

[Declerck et al.2020] Declerck, Thierry, John McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel, Philipp Cimiano, Artem Revenko, Roser Sauri, Deirdre Lee, Stefania Racioppa, Jamal Nasir, Matthias Orlikowski, Marta Lanau-Coronas, Christian Fäth, Mariano Rico, Mohammad Fazleh Elahi, Maria Khvalchik, Meritxell Gonzalez, and Katharine Cooney. 2020. Recent developments for the linguistic linked open data infrastructure. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 5660–5667. European Language Resources Association (ELRA).

[Declerck et al.2023] Declerck, Thierry, Thomas Troelsgård, and Sussi Olsen. 2023. Towards an rdf representation of the infrastructure consisting in using wordnets as a conceptual interlingua between multilingual sign language datasets. In *GWC 2023: 12th International Global Wordnet Conference, Proceedings*, 01. to appear.

[Ebling et al.2012] Ebling, Sarah, Katja Tissi, and Martin Volk. 2012. Semi-Automatic Annotation of Semantic Relations in a Swiss German Sign Language Lexicon. In Crasborn, Onno, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Jette Kristoffersen, and Johanna Mesch, editors, *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 31–36, Istanbul, Turkey, May. European Language Resources Association (ELRA).

[Efthimiou et al.2016] Efthimiou, Eleni, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch, editors. 2016. *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign*

Languages: Corpus Mining, Portorož, Slovenia, May. European Language Resources Association (ELRA).

[Forster et al.2010] Forster, Jens, Daniel Stein, Ellen Ormel, Onno Crasborn, and Hermann Ney. 2010. Best practice for sign language data collections regarding the needs of data-driven recognition and translation. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010)*, 01.

[Hanke2004] Hanke, Thomas. 2004. HamNoSys – representing sign language data in language resources and language processing contexts. In Streiter, Oliver and Chiara Vettori, 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).

[Jordan et al.2014] Jordan, Fenlon, Kearsy Cormier, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam, and Bencie Woll. 2014. Bsl signbank: A lexical database of british sign language (first edition).

[McCrae et al.2017] McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017*, pages 587–597. Lexical Computing CZ s.r.o.

[Mesch et al.2012] Mesch, Johanna, Lars Wallin, and Thomas Björkstrand. 2012. Sign language resources in Sweden: Dictionary and corpus. In Crasborn, Onno, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Jette Kristoffersen, and Johanna Mesch, editors, *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 127–130, Istanbul, Turkey, May. European Language Resources Association (ELRA).

[Murtagh et al.2022] Murtagh, Irene, Víctor Ubieto Nogales, and Josep Blat. 2022. Sign language machine translation and the sign language lexicon: A linguistically informed approach. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 240–251, Orlando, USA, September. Association for Machine Translation in the Americas.

[Murtagh2019] Murtagh, Irene. 2019. *A Linguistically Motivated Computational Framework for Irish Sign Language*. Ph.D. thesis, Trinity College Dublin.

[Neves et al.2020] Neves, Carolina, Luísa Coheur, and Hugo Nicolau. 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.

[Pedersen et al.2009] Pedersen, Bolette Sandford, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet — the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.

[Pedersen et al.2018] Pedersen, Bolette Sandford, Manex Aguirrezabal Zabaleta, Sanni Nimb, Sussi Olsen, and Ida Rørmann Olsen. 2018. Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in danish. In *Proceedings of Global WordNet Conference 2018*. Global WordNet Association. null ; Conference date: 08-01-2018 Through 12-01-2018.

[Power et al.2022] Power, Justin, David Quinto-Pozos, and Danny Law. 2022. Signed language transcription and the creation of a cross-linguistic comparative database. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 173–180, Marseille, France, June. European Language Resources Association.

[Prillwitz et al.2008] Prillwitz, Siegmund, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, and Arvid Schwarz. 2008. DGS Corpus project – development of a corpus based electronic dictionary German Sign Language / German. In Crasborn, Onno, Eleni Efthimiou, Thomas Hanke, Ernst D. Thoutenhoofd, and Inge Zwitserlood, editors, *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 159–164, Marrakech, Morocco, June. European Language Resources Association (ELRA).

[Schuurman et al.to be published] Schuurman, Ineke, Thierry Declerck, Caro Brosens, Margot Jassens, Vincent Vandeghinste, and Bram Vanroy. to be published. Are there just WordNets or also SignNets? In *Proceedings of the 13th Global WordNet Conference*.

[Sérasset2015] Sérasset, Gilles. 2015. DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361. Publisher: IOS Press.

[Sutton1991] Sutton, V. 1991. *Lessons in Sign Writing: Textbook*. Cent. for Sutton Movement Writ.

[Sérasset and Tchechmedjiev2014] Sérasset, Gilles and Andon Tchechmedjiev. 2014. Dbnary : Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, page to appear, Reyjkjavik, France.

[Troelsgård and Kristoffersen2018] Troelsgård, Thomas and Jette Kristoffersen. 2018. Improving lemmatisation consistency without a phonological description. the Danish Sign Language corpus and dictionary project. In Bono, Mayumi, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, Johanna Mesch, and Yutaka Osugi, editors, *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 195–198, Miyazaki, Japan, May. European Language Resources Association (ELRA).

[Vossen1999] Vossen, Piek. 1999. EuroWordNet General Document. Technical report, University of Amsterdam, The Netherlands.

[Wilkinson et al.2016] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March.