Towards an RDF Representation of the Infrastructure consisting in using WordNet(s) as a conceptual Interlingua between multilingual SignLanguage Datasets

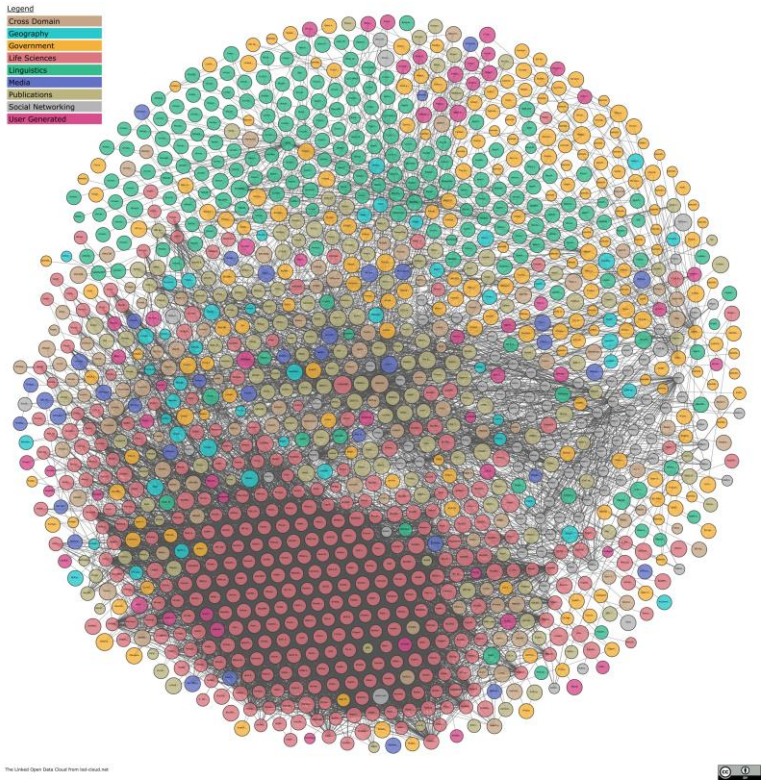
Thierry Declerck[1], Thomas Troelsgård[2] and Sussi Olsen[3]

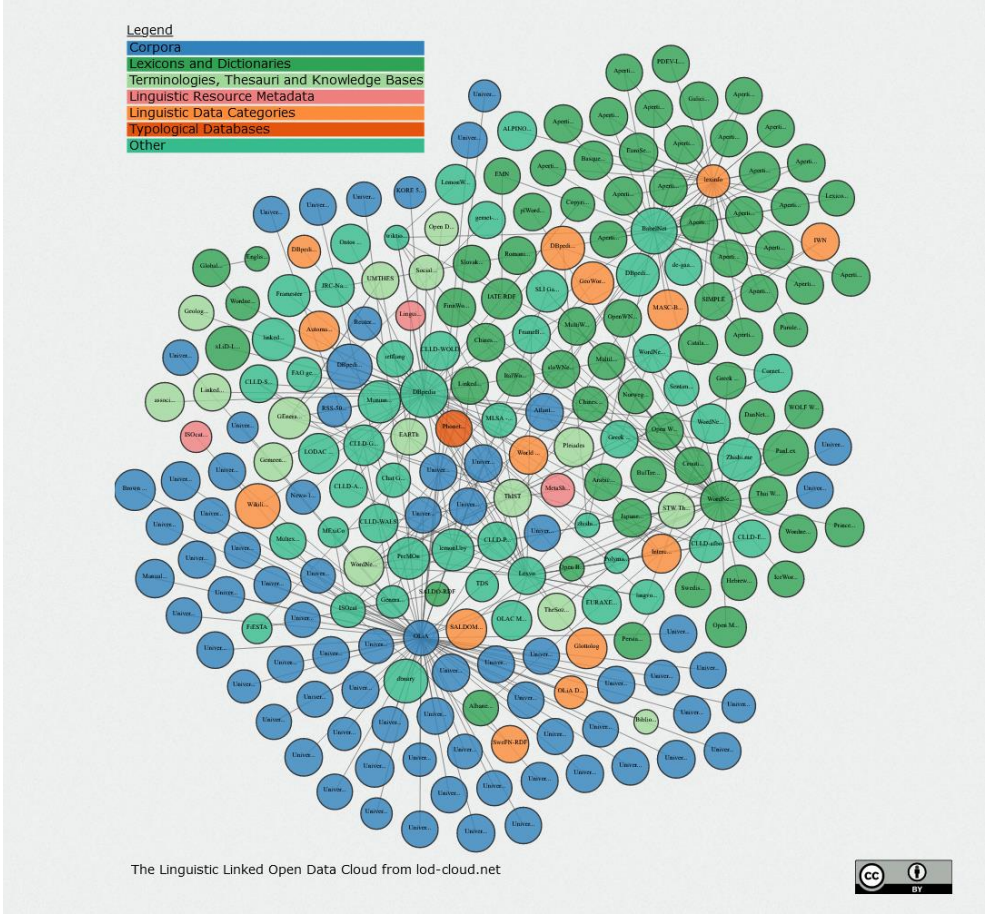
[1]German Research Center for Artificial Intelligence (DFKI GmbH)

[2]University College Copenhagen (KP)

[3]Centre for Language Technology,NorS, University of Copenhagen

Goal: Representing and Publishing Sign Language Data Sets in the Linguistic Linked Data (LLOD) Cloud, which is a subset of the Linked Data (LD) Cloud

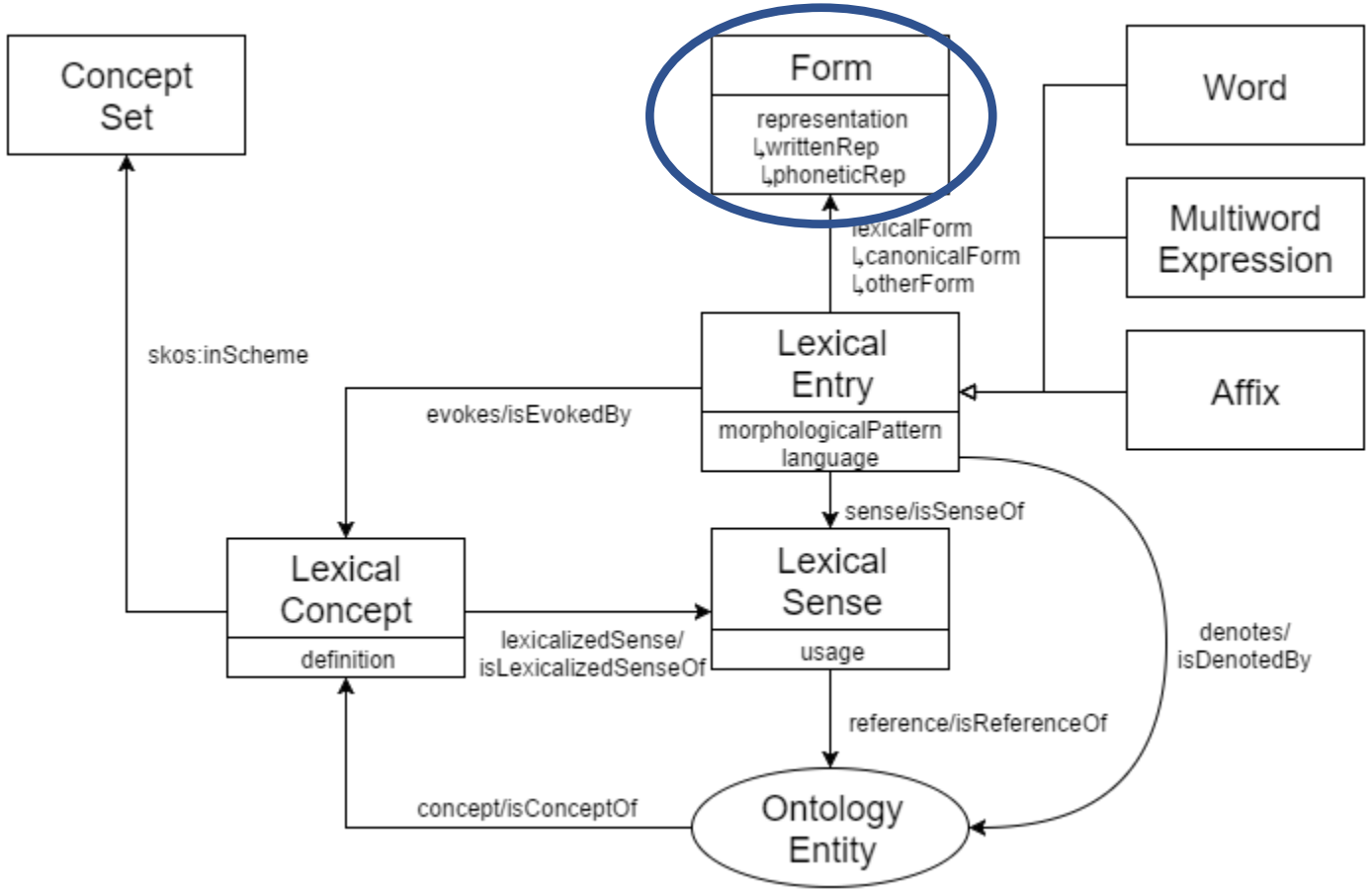




Current status of the relation between Linked Data and Sign Languages

- We could observe that Sign Language (SL) lexical data are not represented in the datasets included by now in the LLOD cloud.
- Also looking at the “Overview of Datasets for the Sign Languages of Europe” published by the “Easier” European project (https://www.project-easier.eu/wp-content/uploads/sites/67/2021/08/EASIER-D6.1-Overview-of-Datasets-for-the-Sign-Languages-of-Europe.pdf) we do not see any mention of a dataset being available in an Linked Data compliant format.

Checking the usability of Ontolex-Lemon for encoding Sign Language data – Ontolex-Lemon covers only writtenRep and phoneticRep. How to include the Representation of SLs?



The Features of Sign Language Data we want to integrate in the LLOD cloud? An example from the American Sign Language

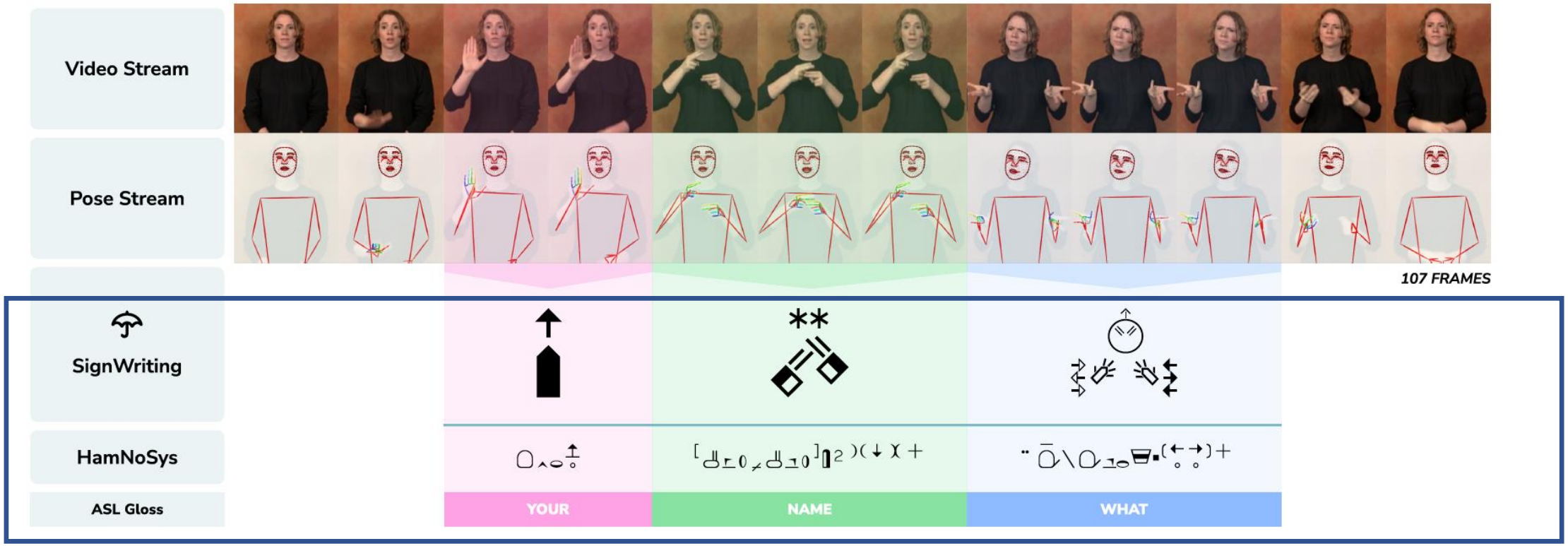


Figure 2: Various representations of American Sign Language. English translation: "What is your name?"

Taken from: Including Signed Languages in Natural Language Processing. Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, Malihe Alikhani. https://arxiv.org/abs/2105.05222 (ACL-IJCNLP 2021)

About the Phonology of Sign Languages

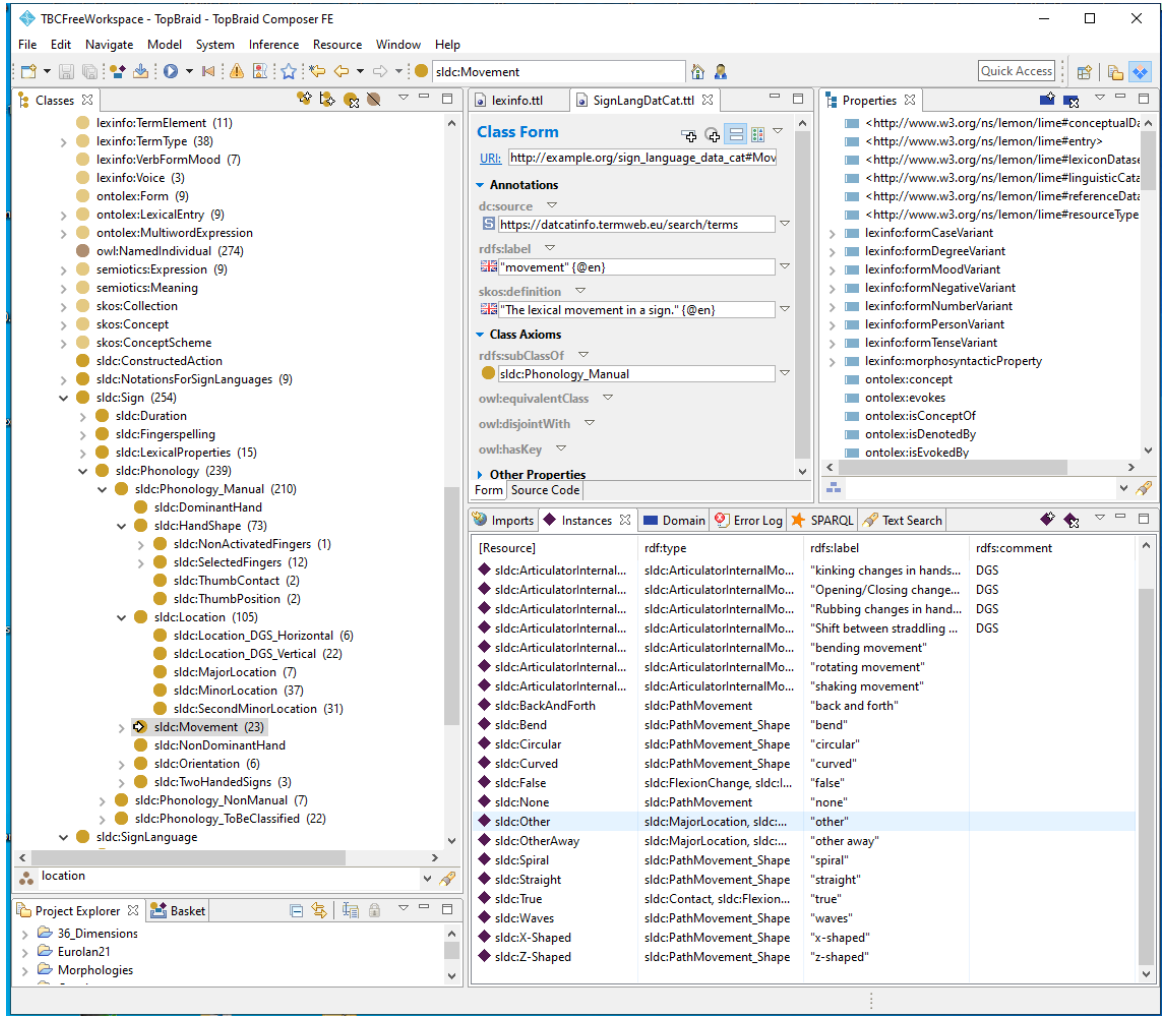
"Sign languages such as American Sign Language (ASL) are characterized by phonological processes analogous to, yet dissimilar from, those of oral languages. Although there is a qualitative difference from oral languages in that sign-language phonemes are not based on sound, and are spatial in addition to being temporal, they fulfill the same role as phonemes in oral languages.

Basically, three types of signs are distinguished: one-handed signs, symmetric two-handed signs (i.e. signs in which both hands are active and perform the same or a similar action), and asymmetric two-handed signs (i.e. signs in which one hand is active [the 'dominant' or 'strong' hand] and one hand is held static [the 'non-dominant' or 'weak' hand]). The non-dominant hand in asymmetric signs often functions as the location of the sign. Almost all simple signs in ASL are monosyllabic."

(https://en.wikipedia.org/wiki/American_Sign_Language_phonology)

More details in specialized literature.

A first Step for encoding the Phonology: A (first and tentative) Ontology of constitutive Elements of SLs



See also Thierry Declerck Towards a new Ontology for Sign Languages. LREC 2022

About the status of “Glosses” used for labelling Sign Language Data

- “Glosses” used in Sign Language resources are not originated in lexical entries of the spoken language or from specific vocabularies. More a kind of informal annotation of a performed sign:
  - “In ASL it is an English word or words that we use to name ASL signs so that we can talk about these signs. The word or words associated with that sign do not relay the signs meaning.” (Signs & their Glosses - Bellevue College)

The “GLOSSING” pages of IDGS point to Corpora where the signs are occurring, but also to more detailed phonological and lexical information

Pointing to corpus attestations

ABWEHR1

Frankfurt | dgskorpus_fra_06 | 31-45f Sie wollten mich mit Gewalt ausziehen, aber ich habe mich dagegen gewehrt.

r	AUF-PERSON1*	AUSZIEHEN-UNTERKÖRPER1A*	ICH1	ABWEHR1
l				
m	auf	ausziehen	ich	

Münster | dgskorpus_mst_16 | 61+m Da habe ich mich gewehrt.

r	ICH1*	ABWEHR1
l		
m	[MG]	[MG]

Münster | dgskorpus_mst_16 | 61+m Ich habe mich gewehrt und gewehrt.

r	ICH1	ABWEHR1
l		
m	[MG]	wehren wehren

Münster | dgskorpus_mst_16 | 61+m Ich habe mich gewehrt und an meiner Mutter festgehalten und gerufen „Mama, Mama.“

r	ABWEHR1	FESTHALTEN2^*
l		
m	[MG]	mama mama

Pointing to more detailed information (transcription, phonology, and lexical)

Pointing to more detailed information (transcription/notation, phonology, and lexical) https://www.sign-lang.uni-hamburg.de/galex/glossen/g13990.html

SCHUTZ1A

Pointing also to signs having a similar had configuration

BEDEUTUNGEN	geschützt, Schutz, schützen
BESCHREIBUNG	Mögliche Assoziation: jemandem oder etwas eine Blockade entgegensetzen oder jemanden von sich wegschieben (stilisiert); Bildelemente mit möglichen Assoziationen: Bewegung: von sich weg; Handform: Kraft, Stärke, Gewalt, massiv, kompakt.
GEBÄRDENRAUMNUTZUNG	Variierbare Gebärde: Ausgangspunkt: Person, die sich schützt; Bewegung ausgerichtet auf Gefahr, Person oder Sache, gegen die man sich schützt; andere Verwendung: Ausgangspunkt: Person, die eine andere Person oder Sache (Endpunkt) schützt.
AUSFÜHRUNGSVARIANTEN	SCHUTZ1B
FORMÄHNLICH	BLEIBEN2
GEBÄRDENUMGEBUNG	SCHUTZ2A
	SCHUTZ2B
	SCHUTZ2C
VORKOMMEN	Konventionell verwendet für „geschützt, Schutz, schützen“:
	32: Anwenderschutz
	46: Atemschutzmaske

About the Status of Transcriptions

- We focus on HamNoSys („Hamburg Sign Language Notation System”), which is a phonetic transcription system.
- There is a conversion mechanism of HamNoSys to XML, resulting in a SiGML machine-readable format (example for the sign labelled with the gloss „BUSCH“):

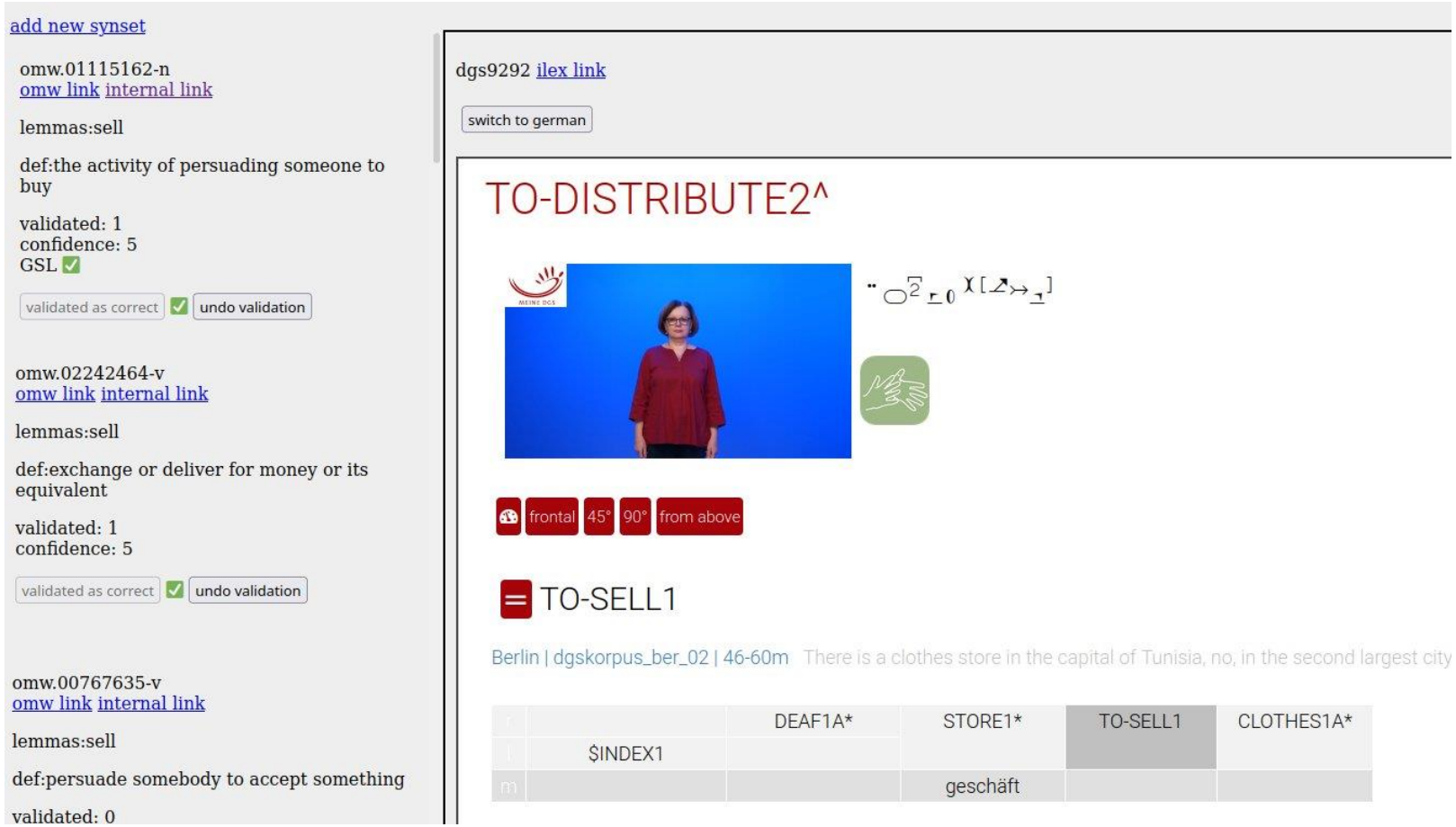

```
(base) C:\Users\thde00\.spyder-py3\python_script\SignLanguage\HamNoSys2SiGML-
master\Original>python HamNoSys2SiGML.py "□□□□□□□□□□□□" (Busch)
<?xml version="1.0" encoding="UTF-8"?>
<sigml>
        <hns_sign gloss="(Busch)">
                <hamnosys_nonmanual/>
                <hamnosys_manual>
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

HamNoSys or ist SiGML transcription can be encoded as a value of ontolex:Form

Linking values of ontolex:Form to instances of ontolex:LexicalConcept

- Current (very promising) work is dealing with linking SL data to WordNet data included in the Open Multilingual Wordnet infrastructure.
  - Contribution on NexusLinguarum (and the projects SignOn and Easier), in cooperation with the University of Hamburg, the Institute for Language and Speech Processing (Greece), and the Fran Ramovš Institute for the Slovene Language: offering a Linked Data compliant representation of this type of linking: ontolex:Form to ontolex:LexicalConcept (where WordNet data is encoded), especially for the many cases for which we are lacking a lexical entry or a lexical sense.

The Linking of SL data and OMW (for English and Greek), as proposed in the EASIER project (https://www.fdr.uni-hamburg.de/record/10169#.Y01WXExBzmE)



The Linking between SL Data and OMW in an Excel File (for Greek and English) – Cross-Lingual Linking via WordNet IDs Conceptually, more precise then linking via Glosses

The Greek data

970	gsl326,omw.00988028-v,manual accept
971	gsl1049,omw.00362103-n,manual accept
972	gsl1050,omw.00377364-n,manual accept
973	gsl1050,omw.07308563-n,manual accept
974	gsl2592,omw.05128519-n,manual accept

The English data

6833	dgs67339,omw.00568430-n,auto accept
6834	dgs10875,omw.14449405-n,auto accept
6835	dgs10040,omw.00377364-n,manual accept
6836	dgs10481,omw.00377364-n,manual accept
6837	dgs9882,omw.04228054-n,auto accept
6838	dgs73480,omw.07349299-n,auto accept

(from https://www.fdr.uni-hamburg.de/record/10169#.Y01WXExBzmE)

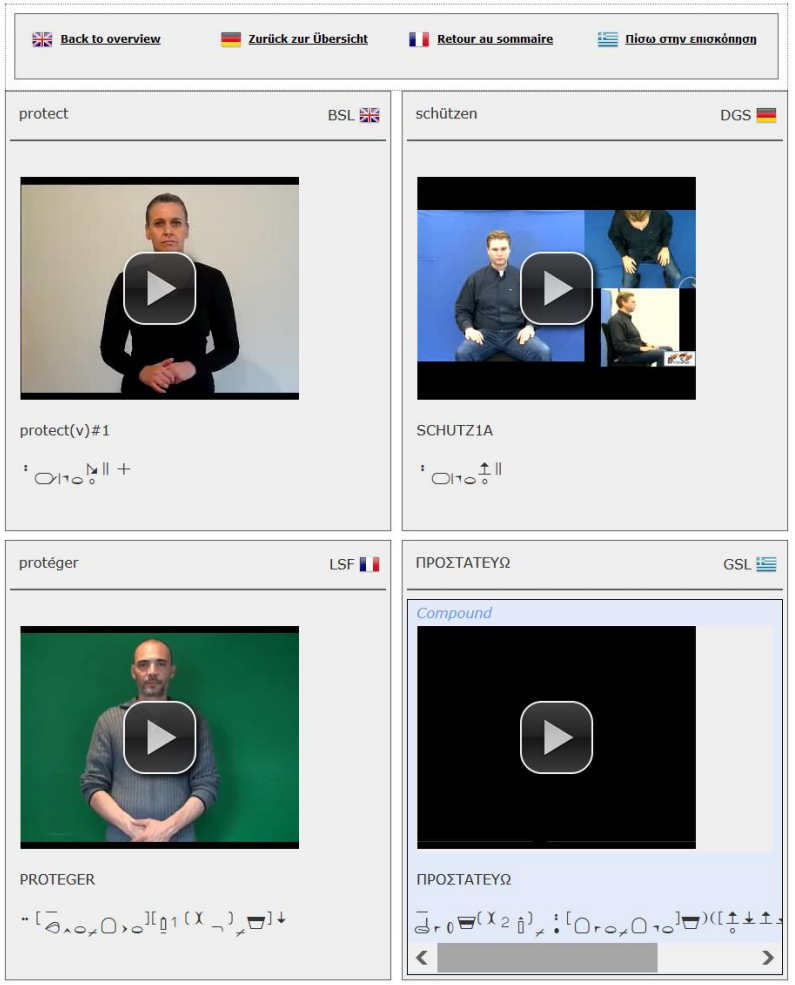
The entry 00377364-n in the Greek Wordnet

00377364-n	ell:lemma	έκρηξη	
00377364-n	ell:lemma	σκάσιμο	
00377364-n	ell:def	0	η ενέργεια του εκρήγνυται ή σκάει κάτι

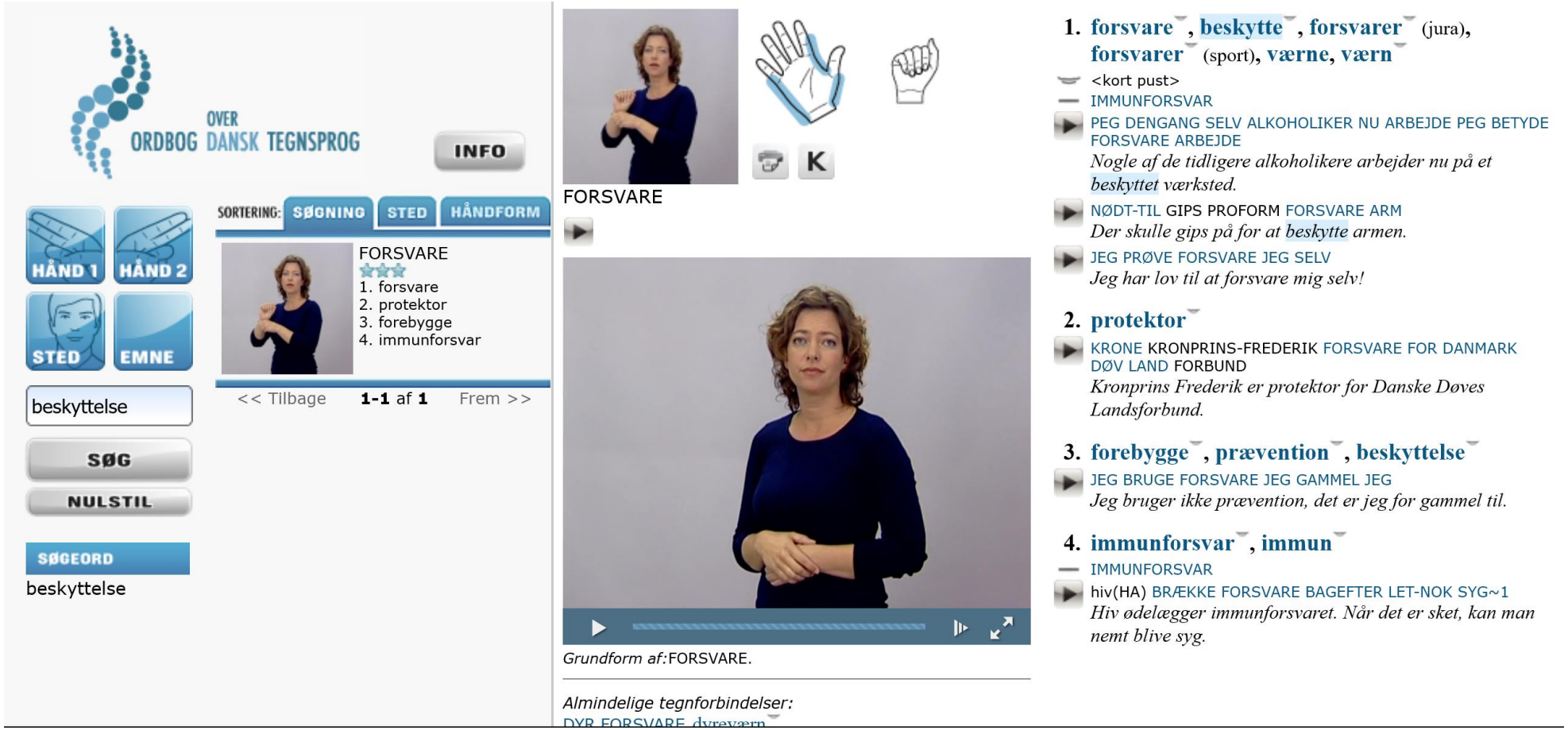
The entry 00377364-n in the English Wordnet

00377364-n	lemma	explosion
00377364-n	lemma	burst

Another IDGS Resource: Dicta-Sign (https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/concepts/cs/cs_688.html): 1000 concepts with Sign Realizations in 4 languages, using single Glosses and other videos



“Beskyttelse” in the Danish SL Portal



Transcriptions of Danish SL

FORSVARE•VAR

□□□□□□□□□□□□□□□□□□

<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>

https://www.tegnsprog.dk/video/t/t_2162.mp4

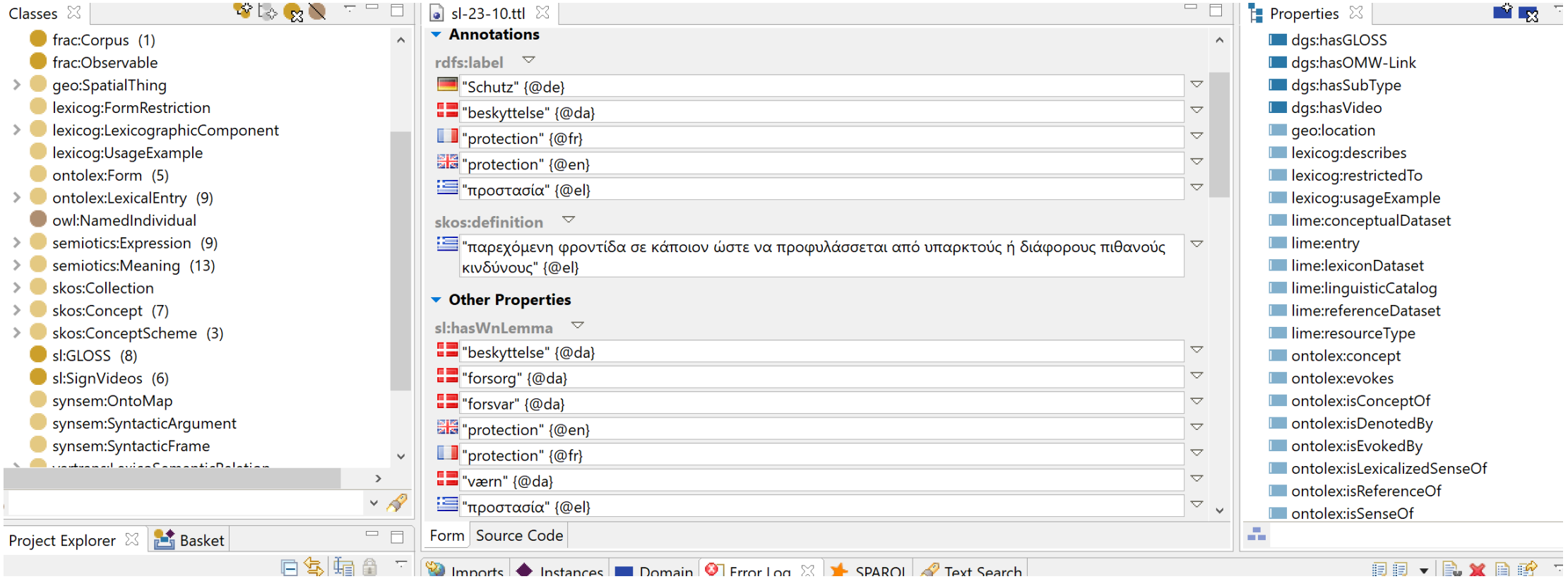
DanNet – PWN core (taken from https://github.com/omwn/omw-data/blob/main/wns/dan/wn-data-dan.tab

4534	13244109-n	lemma	besiddelse
4535	13244109-n	lemma	eje
4536	13244109-n	lemma	ejendom
4537	05091194-n	lemma	proportion
4538	07161429-n	lemma	forslag
4539	14489699-n	lemma	velstand
4540	01128193-v	lemma	beskytte
4541	01128193-v	lemma	forsvare
4542	01128193-v	lemma	værne
4543	00817680-n	lemma	beskyttelse
4544	00817680-n	lemma	forsorg
4545	00817680-n	lemma	forsvar
4546	00817680-n	lemma	værn
4547	14539960-n	lemma	ly
4548	14539960-n	lemma	skjul

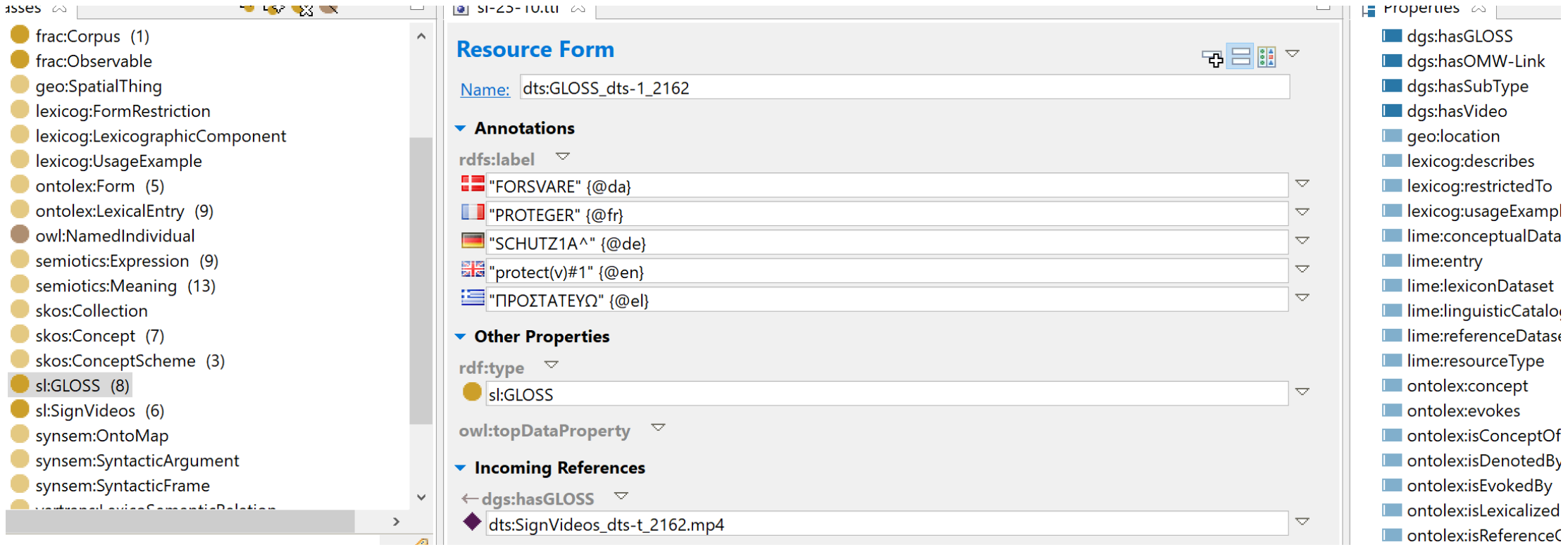
ODT-ID 3844:
https://www.tegnsprog.dk/#%7Csoeg%7C'tekst'beskyttelse%7Cresultat%7C1%7Ctrestjerner%7C1%7Ctegn%7C837

Signs/Videos/Glossesof Signs,Transcriptions of Signs In other Sign Languages, like English, French, German, Greek, etc..

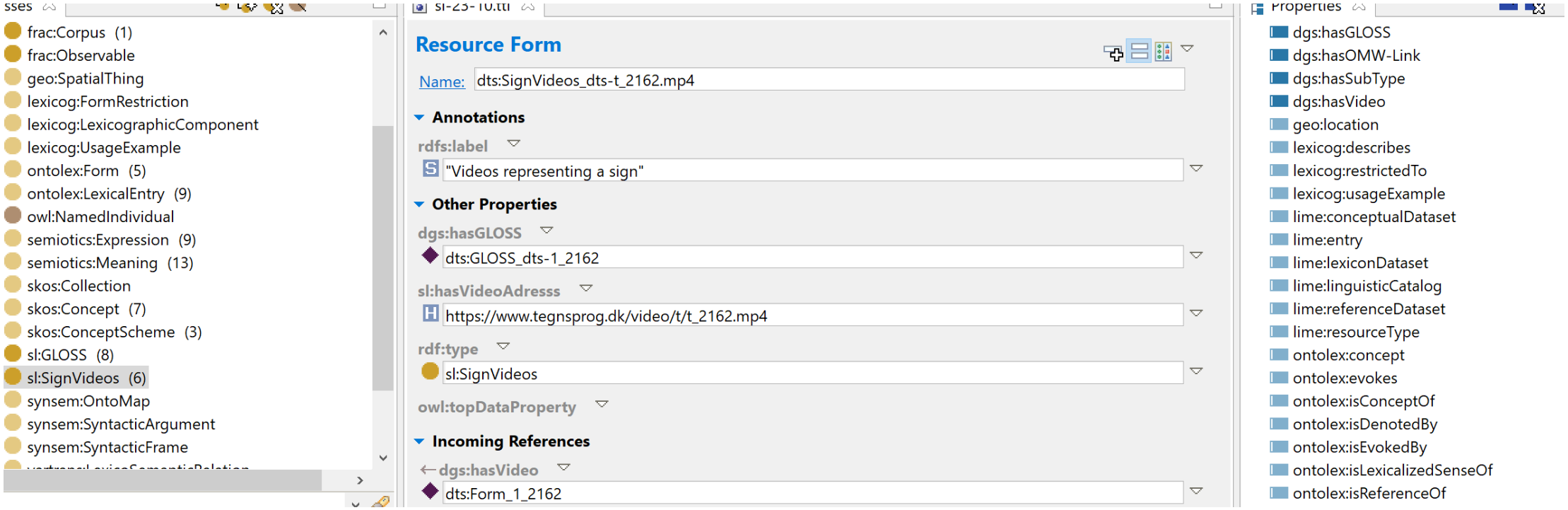
Our RDF/OntoLex-Lemon Encoding of the conceptual Information



Our RDF/OntoLex-Lemon Encoding of the SL Glosses



Our RDF/OntoLex-Lemon Encoding of the Locations of the SL Videos



Conclusions

- Promising on-going work with an exciting cross-disciplinary cooperation!
- The linking of SLs data to Wordnet seems to be an optimal anchor point for linking SL Data to spoken language data.
- Integrating a number of disparate SL, conceptual and lexical resources under one “umbrella”, realising a dense linking of this distinct information sources.
- Current work in applying the approach across Nordic Sign Languages (Danish, Swedish, Icelandic for the time being)

Thanks for your Attention!