easier
intElligent Automatic
SIgn languagE tRanslation

D6.4 INTERLINGUAL INDEX FOR THE PROJECT'S CORE SIGN LANGUAGES

Revision: v1.0

Work Package	WP6
Task	T6.2
Due date	31/05/2023
Submission date	30/06/2023
Deliverable lead	Universität Hamburg (UHH)
Version	1.0
DOI (latest version)	10.25592/uhhfdm.12675
Authors	Sam Bigeard, Maria Kopf, Marc Schulder, Thomas Hanke (UHH), Kiriaki Vasilaki, Anna Vacalopoulou, Theodor Goulas, Athanasia–Lida Dimou, Stavroula–Evita Fotinea, Eleni Efthimiou (Athena), Neil Fox (UCL), Onno Crasborn, Lianne Westenberg (RU), Sarah Ebling and Laure Wawrinka (UZH)
Reviewers	Onno Crasborn (RU)

Abstract	The purpose of the interlingual index is to link the lexical resources of all sign languages of the project. This is the second version of the index, which covers all the languages of the project.
Keywords	multilingual wordnet, lexical resource, crosslingual resource, semi- automatic resource creation

Grant Agreement No.: 101016982
Call: H2020-ICT-2020-2
Topic: ICT-57-2020
Type of action: RIA

Document Revision History

Version	Date	Description of change	List of contributors
V0.1	21/06/2023	First draft	Sam Bigeard, Maria Kopf, Marc Schulder, Thomas Hanke (UHH), Kiriaki Vasilaki, Anna Vacalopoulou, Theodor Goulas, Athanasia–Lida Dimou, Stavroula–Evita Fotinea, Eleni Efthimiou (Athena), Neil Fox (UCL), Onno Crasborn, Lianne Westenberg (RU), Sarah Ebling and Laure Wawrinka (UZH)
V0.2	30/06/2023	Internal Review	Onno Crasborn (RU)
V1.0	30/06/2023	Camera-ready submission DOI: 10.25592/uhhfdm.12676	

DISCLAIMER

The information, documentation and figures available in this deliverable are written by the "Intelligent Automatic Sign Language Translation" (EASIER) project's consortium under EC grant agreement 101016982 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable	**OTHER**	
Dissemination Level		
PU	Public, fully open, e. g., web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to EASIER project and Commission Services	

R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

* OTHER: Software, technical diagram, etc

EXECUTIVE SUMMARY

The purpose of the inter-lingual index is to link the lexical resources from the different languages of the project and make them machine-readable. The earlier deliverable D6.3 was the first version of this index. It included German Sign Language (DGS) and Greek Sign Language (GSL). This deliverable is the second version of the index. It covers further core sign languages of the project: British Sign Language (BSL), Sign Language of the Netherlands (NGT), French Sign Language (LSF) and Swiss-German Sign Language (DSGS). The next version will be deliverable 6.5 and will include languages beyond the project's core languages.

The deliverable is the index itself. This report provides background information on wordnet research, explains our method and choices, and presents the resulting dataset.

Our interlingual index uses the wordnet concept of synonym sets (synsets), which define concepts by gathering signs and words that can represent the same meaning. This approach is more resistant to translation mistakes stemming from translation pairs being only valid for certain word/sign meanings. It also provides a new way to define sign types that does not rely on approximate translations to a single spoken language word, the way glosses do. As a basis for our index, we build on the synset inventory of Open Multilingual Wordnet (OMW).

We use a three-step method: The first step is automatically matching candidate synsets to signs using the keywords and glosses associated with the sign. The second step is automatically validating links that are most likely to be correct. The final step is manual validation of the remaining links, prioritising the most useful signs.

This work has resulted in a dataset of 7929 signs in 6 sign languages linked to 11806 synsets. Additionally, a web interface has been launched to make the index accessible for the general public.

Funded by the Horizon 2020
Framework Programme of the European Union

CONTENTS

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF FIGURES

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF TABLES

Funded by the Horizon 2020
Framework Programme of the European Union

ABBREVIATIONS

SL	sign language
OMW	Open Multilingual Wordnet
PWN	Princeton Wordnet

Sign Languages

ASL	American Sign Language
BSL	British Sign Language
DGS	German Sign Language / Deutsche Gebärdensprache
DSGS	Swiss-German Sign Language / Deutschschweizer Gebärdensprache
GSL	Greek Sign Language / Ελληνική νοηματική γλώσσα (Elleniké Noematiké Glossa)
LIS	Italian Sign Language / Lingua Italiana dei Segni
LSF	French Sign Language / Langue des Signes Française
NGT	Sign Language of the Netherlands / Nederlandse Gebarentaal
PJM	Polish Sign Language / Polski Język Migowy
STS	Swedish Sign Language / Svenskt Teckenspråk

1 INTRODUCTION

The purpose of the inter-lingual index is to link the lexical resources from the different languages of the project and make them machine-readable. This deliverable follows D6.3 (Bigeard et al., 2022)[1], which was the first version of the index, covering German Sign Language (DGS) and Greek Sign Language (GSL).

The version of the index presented in this deliverable also covers other core languages of the project: BSL, NGT, LSF and DSGS.

The index uses the wordnet concept of synonym sets (synsets), which define concepts by gathering signs and words that can represent that meaning. By equipping a synset with signs/words from different languages, cross-lingual semantic information is established that can be used for translation and other linguistic tasks. This approach is more resistant to translation mistakes stemming from choosing the wrong meaning of a polysemous word/sign when deciding how to translate it. It also provides a new way to define sign types that does not rely on approximate translations to a single spoken language word, the way glosses do, but rather on (largely) language-agnostic concept representations. As a basis for our index, we build on the synset inventory of Open Multilingual Wordnet (OMW)[2].

We present our approach and results so far. We use a combination of automatic and manual methods to integrate sign languages into a multilingual wordnet.

[1] https://doi.org/10.25592/uhhfdm.10170

[2] https://omwn.org

2 BACKGROUND

2.1 WORDNETS

A more complete state of the art on Wordnets, and more particularly Wordnets for sign languages, can be found in D6.3 (Bigeard et al., 2022). Key information is repeated below.

The concept of a wordnet was first introduced by Miller et al. (1990) as the idea of a dictionary based on psycholinguistic principles. While the original Princeton Wordnet (PWN) was designed for English, wordnets for many different languages have since been created. Several efforts to interconnect these into a multilingual wordnet have been undertaken. The most prominent such resource that is still actively supported is the Open Multilingual Wordnet (OMW) (Bond and Paik, 2012).

Most wordnet projects use Princeton Wordnet as a basis to expand upon, rather than developing their own wordnet from scratch (Bond et al., 2016). This approach is known as the *expand model*. While this creates a bias toward English, it significantly reduces the amount of work needed to create a new wordnet and connect existing ones.

Work on creating wordnets for individual sign languages has been reported for DSGS (Ebling et al., 2012), Italian Sign Language (LIS) (Shoaib et al., 2014) and American Sign Language (ASL) (Lualdi et al., 2021), although no publicly available resources have yet been released. All of these works have in common that they seek to link wordnet structures to existing lexical resources of the respective sign language.

Other works do not seek to publish full signed language wordnets, but rather use existing wordnets for a spoken language as an aid to internal work.

2.2 PREVIOUS WORK

This deliverable follows D6.3 (Bigeard et al., 2022), where we started the index with DGS and GSL. We set up the method which we now use to add more languages to the index.

When D6.3 was submitted, the index contained 1819 GSL signs with validated synsets, 2230 DGS signs with validated synsets, and 11856 DGS signs yet to be validated. Since then, more manual validation has been done on DGS, and the following languages have been added to the index: BSL, NGT, LSF and DSGS.

3 RESOURCES

3.1 WORDNET RESOURCES

3.1.1 Open Multilingual Wordnet

We use OMW's pre-existing list of synsets. A synset corresponds to a single meaning or sense and is very fine-grained. It is identified by a numerical ID independent from any particular language. It contains in several languages: a definition, words having this meaning, and example sentences. Synsets are semantically linked, thus forming the "net" part of a wordnet. As an example, the synset `07739125-n`[3] represents an apple in the sense of the fruit. Apple in the sense of the tree species is instead represented by `12633994-n`[4], a different synset.

OMW by itself is a collection of individually built wordnets that share the same identifiers and overall structure. All the spoken language of the project exist in OMW except German.

3.1.2 GermaNet

Since OMW doesn't natively support German, we need to link it to a distinct resource. The largest wordnet for German is GermaNet (Hamp and Feldweg, 1997). It contains 151843 synsets. While it is inspired by PWN, it was built independently, from German resources. Due to licence restrictions it is not directly integrated into OMW. However, for 28564 of its synsets a mapping to PWN exists, from which OMW identifiers can be inferred. For our multilingual wordnet we decided to use GermaNet and expand the connections to OMW. We use GermaNet when working on DGS and DSGS, where our lexical resources include German words.

3.2 SIGN LANGUAGE RESOURCES

3.2.1 DGS Corpus

The DGS Corpus (Hanke et al., 2020; Konrad et al., 2020) is our resource for DGS. It is presented in detail in D6.3. This resource differs from the others we use by its implementation of a type hierarchy, called 'double glossing' (Konrad et al., 2012, p. 88). Each *type* represents a distinct sign realisation. It is further subdivided into *subtypes*, each of which represents a lexicalised meaning of that sign which typically also determines potential mouthing. Glosses for types and subtypes are available in English and German. In the interlingual index, we operate at the level of the subtype. The double glossing system allows to distinguish keywords that represent different meanings, and synonyms.

[3] https://compling.upol.cz/ntumc/cgi-bin/wn-gridx.cgi?gridmode=grid&synset=07739125-n

[4] https://compling.upol.cz/ntumc/cgi-bin/wn-gridx.cgi?gridmode=grid&synset=12633994-n

Funded by the Horizon 2020 Framework Programme of the European Union

3.2.2 Polytropon and Noema+ dictionary

The Polytropon (Efthimiou et al., 2016; Efthimiou et al., 2018) and the Noema+ dictionary are our resources for GSL. They are presented in details in D6.3.

3.2.3 BSL Signbank

The BSL signbank (Fenlon et al., 2014)[5] contains 3566 entries at time of writing. The data for each sign includes video, ID-glosses, and a list of English keywords. Keywords may represent each a different meaning, or synonyms with the same meaning. We use this resource as basis for BSL.

3.2.4 NGT signbank

For NGT we use the NGT dataset in Global Signbank (Crasborn et al., 2016)[6]. It contains 4454 entries at time of writing. It has videos, as well as ID-glosses and keywords in both Dutch and English. Keywords may represent each a different meaning, or synonyms with the same meaning.

3.2.5 DSGS database

We have access to an internal database of 3755 signs with video, ID-glosses and German keywords, kindly provided by Penny Boyes Braem. Keywords may represent each a different meaning, or synonyms with the same meaning. This dataset is not currently public data, but efforts are being made toward making the videos shareable in the future. Once the dataset becomes public data, the wordnet will be updated accordingly.

3.2.6 Dicta-Sign

Dicta-Sign (Efthimiou et al., 2010)[7] was a project that aligned 1046 concepts to signs from BSL, DGS, LSF and GSL. The concepts were linked to PWN synsets where available. This dataset provides links between signs and wordnet synsets, making their inclusion into the index trivial. 3847 signs were imported, each linked to one synset.

This is the only resource we have available for LSF. However, for DGS, GSL and BSL we also have signs from other resources. The same sign might appear in two resources, creating duplicates. For this reason, in this report, Dicta-Sign entries for languages other that LSF are not counted in the total of signs, unless clearly stated otherwise.

[5] https://bslsignbank.ucl.ac.uk/

[6] https://signbank.cls.ru.nl/datasets/NGT

[7] https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/

3.2.7 Other languages

The aim of this version of the index was to cover all core languages of the project. However, in the case of LIS, while corpus resources are available, we could not identify a publicly available lexical resource that would have been suitable for inclusion in the index. We are aware of (Shoaib et al., 2014), but the resource is not currently available. Therefore, LIS is absent from the current version of the index. A decision was reached to replace it with another sign language (SL) outside of the project's core languages. Efforts have been made to reach out to other resources. Owners of the Corpus of Polish Sign Language (Kuder et al., 2022; Wójcicka et al., 2020) and of the Swedish Sign Language Dictionary (Svenskt teckenspråkslexikon, 2023) have expressed interest, but at the time of writing this report, we don't have yet confirmation of a possible collaboration.

Funded by the Horizon 2020
Framework Programme of the European Union

4 WORDNET CREATION

We follow the same general method for each language, with adaptations to address specific problems or better make use of extra data available. This is a three-step method.

The first step is to automatically match a sign's spoken language equivalents (glosses and keywords) to lemma present in this language's pre-existing wordnet. Words in contact language are preferred to English words when both are available, assuming that the annotators who associated this word with the sign are more familiar with the contact language language.

The second step is applied if this process results in a one-to-one match. The link is then assumed correct, and automatically validated. This is the case for single-sense words such as *refrigerator*.

In most cases, several candidates are associated with each sign. This leads to the third step: manual validation. Ideally, native signers would perform this work. But due to limited resources, fluent non-native signers were also deployed as annotators.

The annotators are as follows:

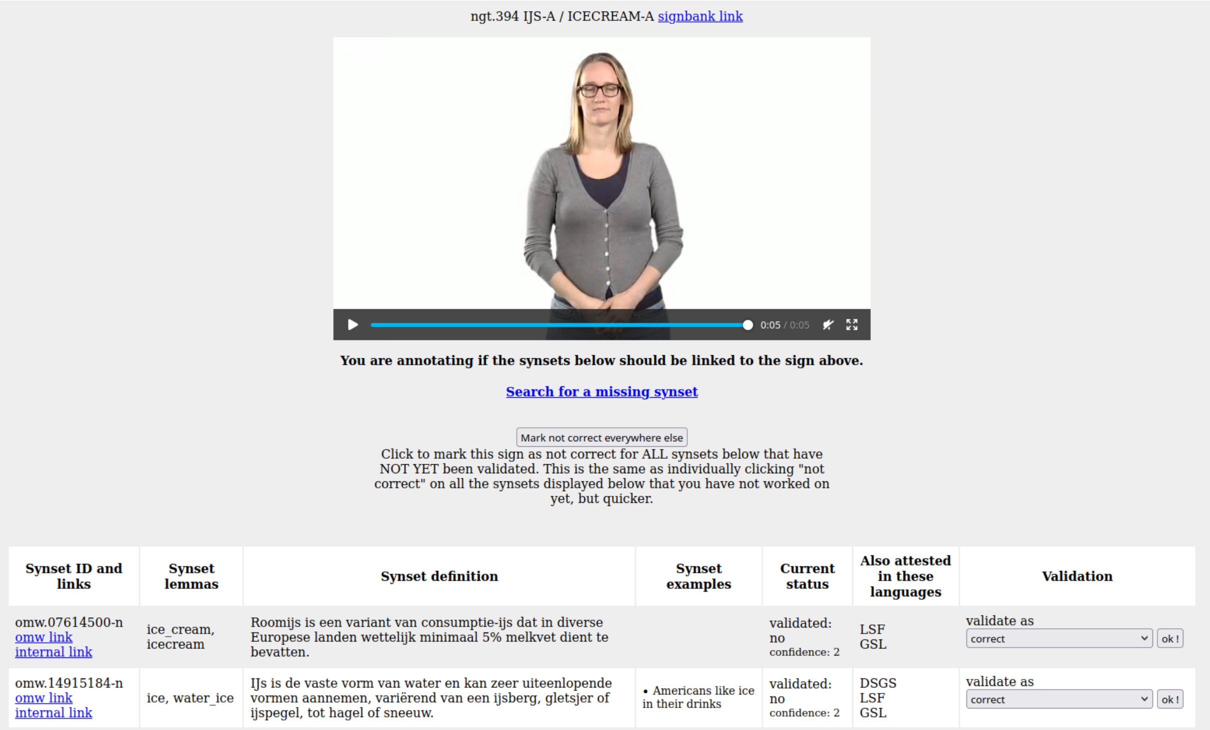
- **DGS:** Maria Kopf
- **GSL:** Kiriaki Vasilaki, Anna Vacalopoulou, Theodor Goulas, Athanasia–Lida Dimou, Stavroula–Evita Fotinea and Eleni Efthimiou
- **BSL:** Neil Fox
- **NGT:** Onno Crasborn and Lianne Westenberg
- **DSGS:** Laure Wawrinka

The annotation interface is shown in Figure 4.1. To better identify the meaning of synsets, annotators have access to lemmas, definitions and examples in their preferred language when available, English otherwise. Concerning signs, video and gloss are displayed. They can also open the page of the sign in its original resource to have access to more information, such as examples of use in corpus.

We do not have the resources for an exhaustive manual annotation of each language. This leads to the question of priority. Annotators were instructed to first cover synsets that are already in use and validated in other sign languages. This leads to the creation of a core set of senses covered in multiple SLs.

Then, signs in the language of the annotator are displayed in frequency order, to prioritise the most useful signs. If the lexical database of this language is linked to a corpus, we import frequencies from it. Otherwise, we derive frequencies from the synsets.

Annotators may be confronted with long lists of synsets per sign, some close in meaning, needing to spend extra time to understand the difference between them. This is especially


ngt.394 IJS-A / ICECREAM-A signbank link

You are annotating if the synsets below should be linked to the sign above.

Search for a missing synset

Mark not correct everywhere else

Click to mark this sign as not correct for ALL synsets below that have NOT YET been validated. This is the same as individually clicking "not correct" on all the synsets displayed below that you have not worked on yet, but quicker.

Synset ID and links	Synset lemmas	Synset definition	Synset examples	Current status	Also attested in these languages	Validation
omw.07614500-n omw link internal link	ice_cream, icecream	Roomijs is een variant van consumptie-ijs dat in diverse Europese landen wettelijk minimaal 5% melkvet dient te bevatten.		validated: no confidence: 2	LSF GSL	validate as correct ok !
omw.14915184-n omw link internal link	ice, water_ice	IJs is de vaste vorm van water en kan zeer uiteenlopende vormen aannemen, variërend van een ijsberg, gletsjer of ijspegel, tot hagel of sneeuw.	• Americans like ice in their drinks	validated: no confidence: 2	DSGS LSF GSL	validate as correct ok !

Figure 4.1: *Screenshot of the annotation interface*

the case for the most frequent signs which link to lemmas such as *to have* (20 linked synsets) or *good* (27 linked synsets). In such cases, annotators are encouraged to limit the time they spend on each sign to increase coverage, which leads to only some of the synsets of these signs being validated.

OMW, by design, doesn't have synsets for function words, such as pronouns. To allow more complete linking of lexical resources and open up use of our wordnet resource for tasks that also require sense-disambiguated information for function words, we expand the scope of our resource accordingly. We created 39 custom synsets to cover high frequency cases. This includes personal pronouns, interrogative pronouns, conjunctions and items specific to sign languages such as pointing and the palm-up gesture.

Below we describe steps specific to certain sign languages. The specificities of DGS and GSL are already discussed in D6.3.

4.1 BSL

For BSL the equivalents are given only in English. While this removes one level of possible mistranslation, it also increases the noise: The basis for OMW is Princeton Wordnet, which is in English. The English part of OMW contains far more synsets per lemma, many of which are too specific for our purpose. This increases the time annotators must spend per sign. Because of that, the coverage of the manual validation is smaller for BSL compared to other languages.

Funded by the Horizon 2020
Framework Programme of the European Union

4.2 DSGS

In the lexical resource used for DSGS, spoken language equivalents are only available in German. This is a difficulty, as OMW doesn't natively support German, and links between Germanet and OMW are not always available. In those cases, we fall back on automatic translation of the German equivalents into English, which results in low quality candidates. For the case of DSGS, we decided to only keep automatically translated candidates if the automatic translation suggests a single possibility, or the translation was identical to the source lemma (case of named entities). In other cases, no candidates were kept, even if this results in signs with no candidates at all. This avoids long lists of irrelevant candidates.

4.3 LSF

As Wordnet links were already present in Dicta-Sign, there was no need to create or validate new links.

Funded by the Horizon 2020
Framework Programme of the European Union

5 RESULTS

We will first present our work through a few examples to better show what the data looks like, and how it can be browsed. Statistics on the size of the index are given afterwards.

5.1 EXAMPLES

Figure 5.1 shows an example of a sense which has been linked to all the sign languages we cover so far.[8] This sense typically links to mono-sense spoken lemmas, which makes it easy to auto-validate. In our index, the senses with the most languages coverage tend to be this kind of sense.

The complete list of correct lemmas for this synset is as follows. For each sign, we show the ID of the sign, local language gloss, English gloss if available, and the webpage of that sign if it's in a public dataset.

- `gsl.6902` φθινόπωρο video
- `lsf.56` AUTOMNE webpage
- `ngt.438` HERFST / AUTUMN webpage
- `bsl.2695` AUTUMN webpage
- `bsl.5948` AUTUMN02 webpage
- `dgs.9761` HERBST1A / AUTUMN1A doi, webpage
- `dgs.76225` HERBST1B / AUTUMN1B
- `dgs.13038` HERBST2A / AUTUMN2A
- `dgs.58031` HERBST2B / AUTUMN2B
- `dgs.74076` HERBST2C/ AUTUMN2C
- `dgs.58320` HERBST3 / AUTUMN3 doi, webpage
- `dgs.72471` HERBST4 / AUTUMN4
- `dgs.73085` HERBST5 / AUTUMN5 doi, webpage
- `dgs.74097` HERBST6A / AUTUMN6A
- `dgs.74117` HERBST6B / AUTUMN6B
- `dsgs.1354` HERBST_1B
- `dsgs.1355` HERBST_1C
- `dsgs.1356` HERBST_1E

This sense also is linked to the following spoken language lemmas, and lemmas in many other languages not displayed here:

- English: fall, autumn
- German: Herbst
- Dutch: herfst
- Greek: φθινόπωρο

[8] As we use non-public DSGS data, the DSGS entry shown in Figure 5.1 instead uses a public recording provided by the Swiss Association of the Deaf at `https://signsuisse.sgb-fss.ch/lexikon/114810/herbst`

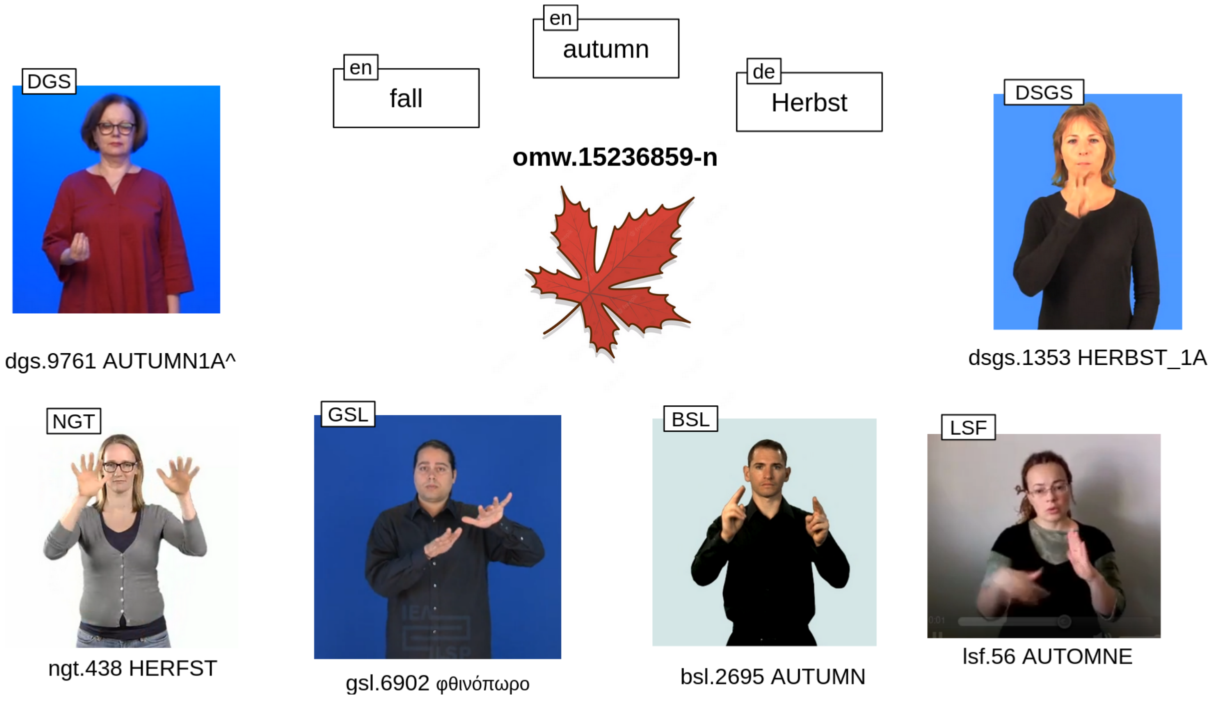


Figure 5.1: *Example of synset for autumn with some of its spoken and sign lemmas*

The next example illustrates how the index disambiguates between possible senses of the keywords and glosses present in the original resources. The sign `dgs.13544` ARBEITEN1ˆ / TO-WORK1ˆ[9] is linked to a number of senses. For each sense are indicated its id, lemmas, then definition. Notice that some of the correct sense are not directly linked to the lemma *work* and would not have been discovered by simply researching for it in wordnet.

This example also shows a sense that was created to fill a gap in OMW, as shown by its distinct `ea` prefix. The lemma *what do* is intended as a tool for the annotators to more easily find this synset.

The following senses are correct:

- `ea.0036` *what do*: question word, what activity a person does
- `omw.00584367-n` *employment, work*: the occupation for which you are paid
- `omw.00620752-n` *labor, labour, toil*: productive work (especially physical work done for wages)
- `omw.02410855-v` *work, do work*: be employed
- `omw.04602044-n` *workplace, work*: a place where work is done
- `omw.13968092-n` *employment, employ*: the state of being employed or having a job
- `omw.13541167-n` *processing*: preparing or putting through a prescribed procedure

[9] Entry in DGS Corpus: `https://doi.org/10.25592/dgs.corpus-3.0-type-13544`
Entry in sign wordnet: `https://www.sign-lang.uni-hamburg.de/easier/sign-wordnet/sign/dgs.13544.html`

Funded by the Horizon 2020
Framework Programme of the European Union

The following senses were candidates that were marked as incorrect:

- `omw.00100551-v` *exercise, work, work out*: give a workout to
- `omw.00634906-v` *solve, work out, figure out, puzzle out, lick, work*: find the solution to (a problem or question) or understand the meaning of
- `omw.01162754-v` *exploit, work*: use or manipulate to one's advantage
- `omw.01235355-v` *knead, work*: make uniform
- `omw.01659248-v` *shape, form, work, mold, mould, forge*: make something, usually for a specific function
- `omw.03841417-n` *oeuvre, work, body of work*: the total output of a writer or artist (or a substantial part of it)

5.2 PUBLIC WEBSITE

A public web interface has been launched to make the resulting data easier to browse, and make it available to the public. It contains data for DGS, GSL, BSL, NGT and LSF. DSGS will be added as soon as the data becomes public. Screenshots of the interface are shown in Figures 5.2 and 5.3. It is hosted at the following address:
`https://www.sign-lang.uni-hamburg.de/easier/sign-wordnet/`

Only synsets that have been validated as correct, automatically or manually, are visible to the public. In the case of DGS a choice had to be made between displaying sub-types separately, or merging to the level of the type. The other lexical resources we work with operate on a level more similar to the DGS types. To match it, we decided to display types for DGS.

In total, the website displays 6343 signs linked to 10217 synsets.

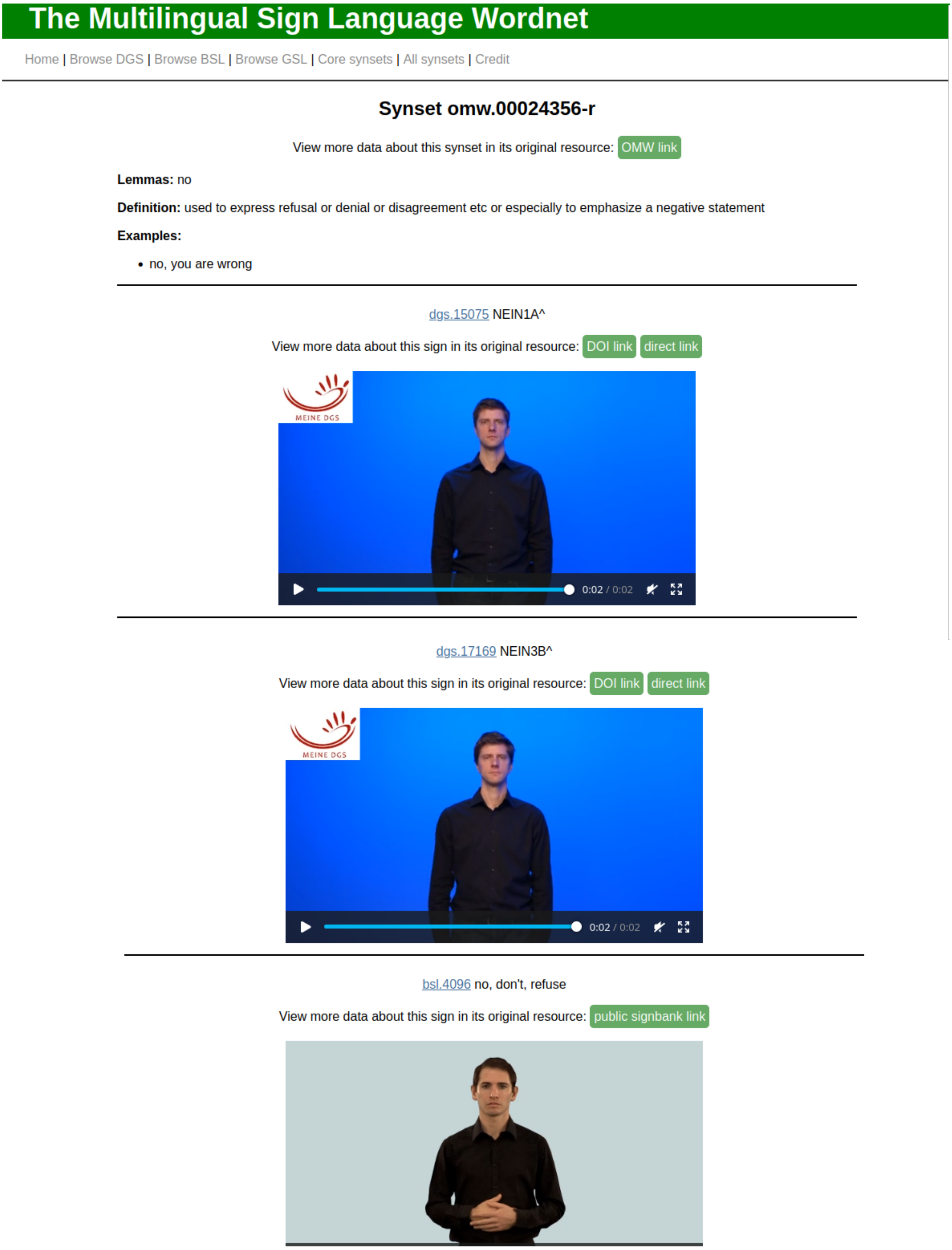
5.3 STATISTICS

Table 5.1 shows the number of positively validated entries per language. This represents senses that have been manually or automatically validated as correct senses for their sign. They are counted per sign (covering one or several synsets) and per individual sign-synset link.

Tables 5.2 and 5.3 show current progress on the manual validation, counted by sign (Table 5.2) or by sign-synset link (Table 5.3). For easier visual inspection, the statistics of both tables are also provided as stacked bar graphs in Figures 5.4 and 5.5, respectively.

	NGT	DGS	GSL	BSL	LSF	DSGS	Total
valid distinct signs	1501	2482	1816	260	1013	1699	8771
valid distinct links	1768	2748	4293	604	1013	1755	12181

Table 5.1: *Number of signs per language that have at least one correct sense, and number of links per languages that are correct.*



The Multilingual Sign Language Wordnet

Home | Browse DGS | Browse BSL | Browse GSL | Core synsets | All synsets | Credit

Synset omw.00024356-r

View more data about this synset in its original resource: OMW link

Lemmas: no

Definition: used to express refusal or denial or disagreement etc or especially to emphasize a negative statement

Examples:

- no, you are wrong

dgs.15075 NEIN1A^

View more data about this sign in its original resource: DOI link direct link

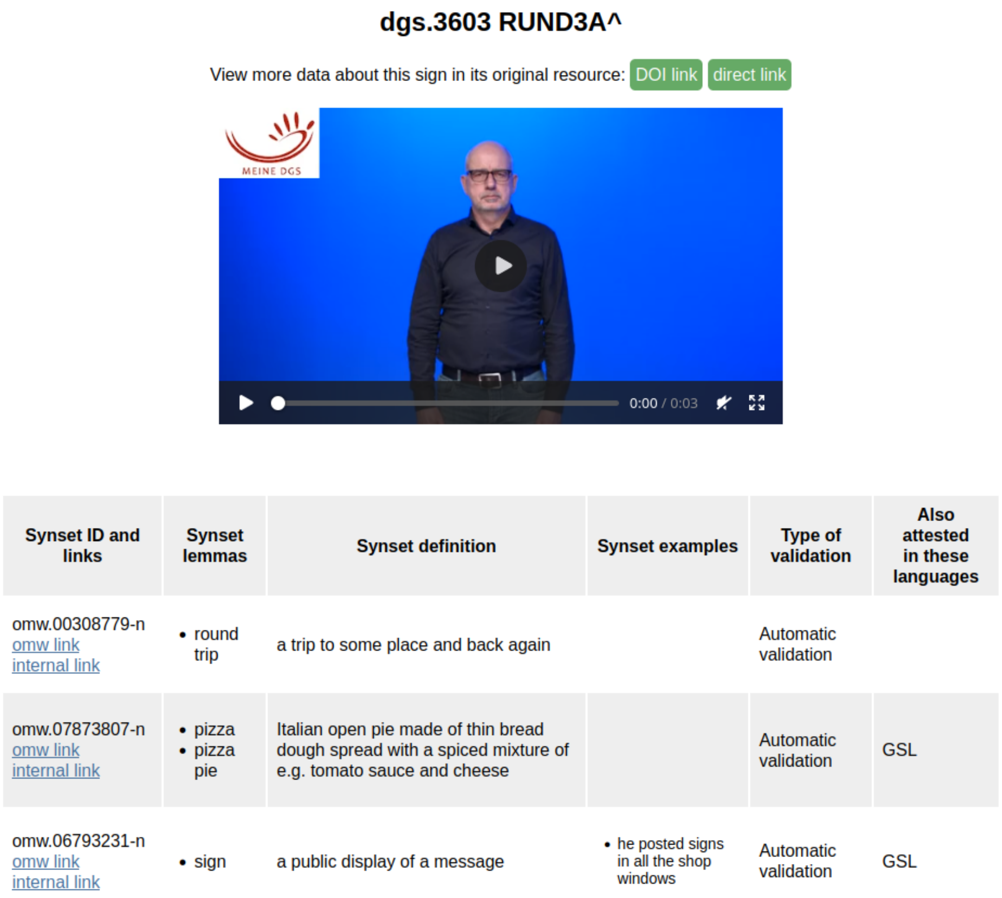
dgs.17169 NEIN3B^

View more data about this sign in its original resource: DOI link direct link

bsl.4096 no, don't, refuse

View more data about this sign in its original resource: public signbank link

Figure 5.2: *Screenshot of the public website showing the data for one synset*



dgs.3603 RUND3A^

View more data about this sign in its original resource: DOI link direct link

Synset ID and links	Synset lemmas	Synset definition	Synset examples	Type of validation	Also attested in these languages
omw.00308779-n omw link internal link	- round trip	a trip to some place and back again		Automatic validation	
omw.07873807-n omw link internal link	- pizza - pizza pie	Italian open pie made of thin bread dough spread with a spiced mixture of e.g. tomato sauce and cheese		Automatic validation	GSL
omw.06793231-n omw link internal link	- sign	a public display of a message	- he posted signs in all the shop windows	Automatic validation	GSL

Figure 5.3: *Screenshot of the public website showing the data for one sign*

sign validation status	NGT	DGS	GSL	BSL	LSF	DSGS
no validation yet	783	8896		3240		1876
all links incorrect	693	257		12		179
some links correct, some incorrect	539	132		28		7
all links correct	270	547	1816	10	1013	289
auto-validated	700	1795		223		1405

Table 5.2: *Current progress of manual validation by number of signs.*

link validation status	NGT	DGS	GSL	BSL	LSF	DSGS
no validation yet	4127	11601		3566		3744
validated as incorrect	3458	969		531		195
validated as correct	1068	943	4293	381	1013	350
auto-validated	700	1805		223		1405

Table 5.3: *Current progress of manual validation by number of sign-synset links.*

Funded by the Horizon 2020
Framework Programme of the European Union

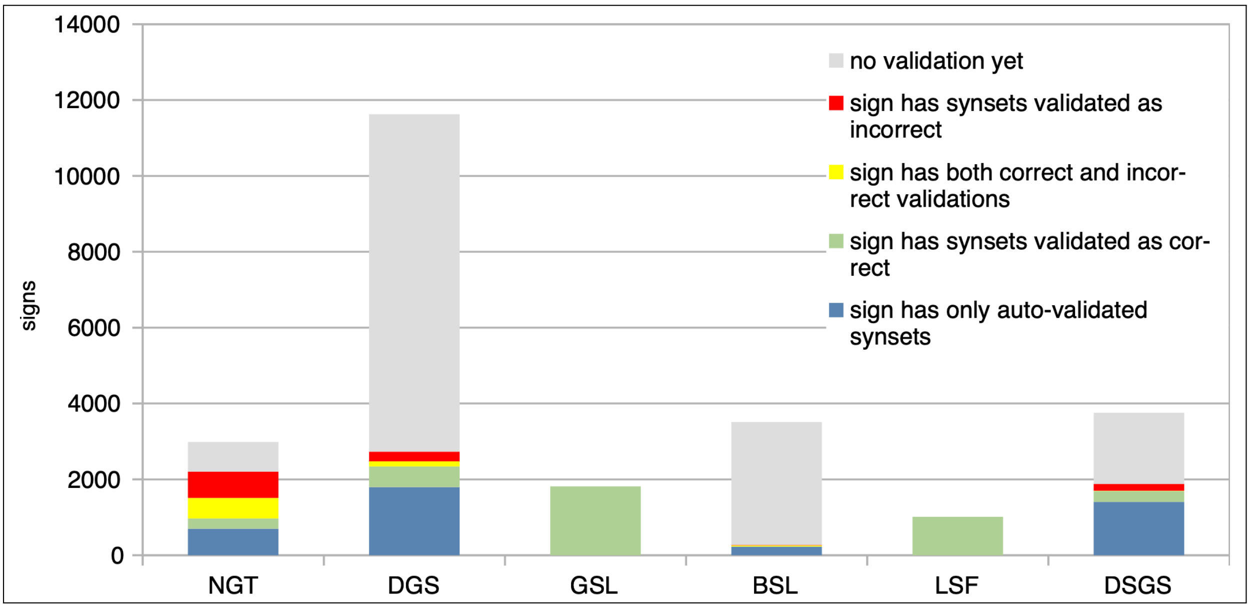


Figure 5.4: *Stacked bar graph representation of Table 5.2, showing the manual validation progress by number of signs.*

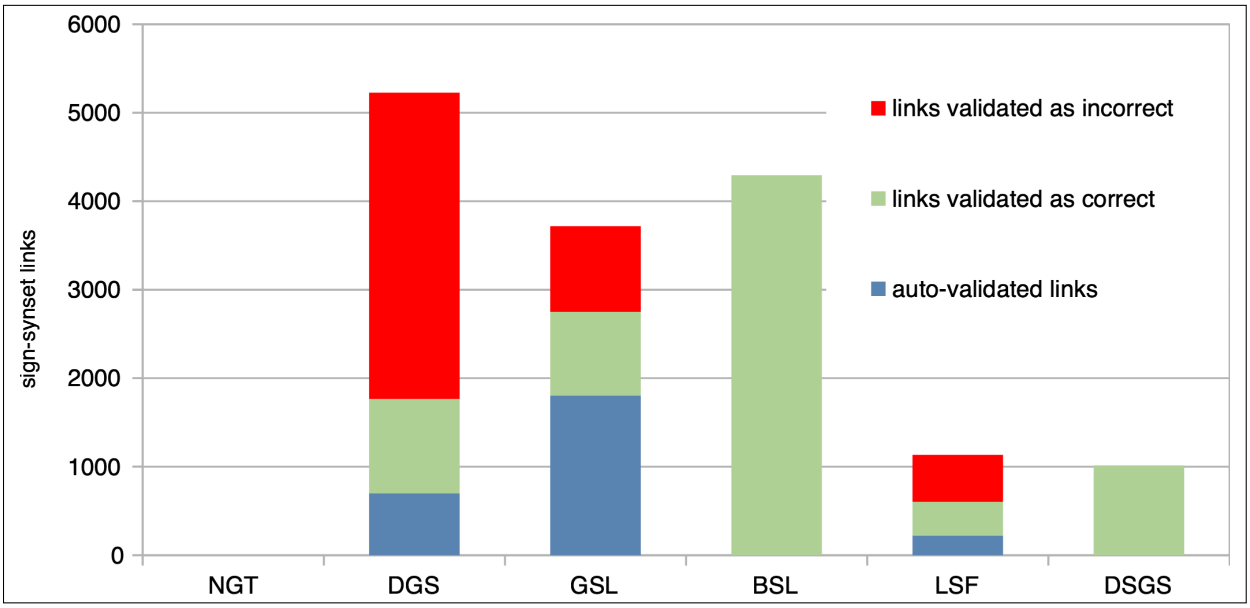


Figure 5.5: *Stacked bar graph representation of Table 5.3, showing the manual validation progress by number of sign-synset links. The count of links that were not yet validated is omitted to ensure the readability of the graph.*

Funded by the Horizon 2020
Framework Programme of the European Union

The work on GSL was done internally by ATHENA team early in the project before automatic candidates were generated. For LSF we imported existing links from Dicta-Sign. This is why for these two languages, only entries validated as correct are given. Annotation for BSL focussed on providing full synset coverage of highly polysemous signs, resulting in a large number of verified links for a small number of signs (compare Figures 5.4 and 5.5).

For GSL, BSL and DSGS the manual validation is finished, as no more human resources for manual annotation are available. The work is still ongoing for DGS and NGT, and more links will be validated before the end of the project. Currently, 1394 synsets have validated links to two or more languages, forming a common interlingual index.

5.4 USE CASES

The index data is available in CSV files, which provide links between glosses, videos, synsets ID, and spoken language lemmas. An API is also under consideration to make the index even easier to include in the translation pipeline.

The main usage for the index is within spoken-to-sign translation. This provides an alternative to deep-learning approaches that better suits low-resource languages. The index has been used this way in conjunction with a tool to detect signs in continuous video to help correct alignment issues generated by the tool.

The index can also be used in sign-to-sign translation, with the assumption that token-by-token substitution is more acceptable while doing sign-to-sign than spoken-to-sign or vice-versa. This method can be used in combination with machine-learning: Machine learning is used as a first step for spoken-to-sign, with the restriction that the target sign language be comparatively well resourced. The index can then be used to transfer the translation to a second, less-resourced sign language. Token-by-token substitution and chaining translation systems result in lower quality translation, and should be seen as a fall-back or exploratory method, rather than a fully functional translation of acceptable quality. Nevertheless, this can provide a pipeline for languages without the resources to train a machine learning model.

Another possible use of the index in combination with spoken-to-sign machine learning translation is when machine learning outputs the gloss of a sign that the avatar cannot produce. Then the index can suggest synonymous signs that the avatar may be able to produce.

More anecdotally, it has been suggested that when using sign-to-sign translation for anonymisation, switching a sign to a synonym found in the index can create greater entropy.

The index is also a convenient repository of lexical data, that has been used by other work packages to save time on preparing and compiling resources themselves.

Outside the EASIER project, the data and the website provide value to sign language projects and projects wishing to include sign languages, as they can use and build on the sign wordnet. For example, a collaboration is in progress toward adding sign language tokens to Ontolex, which is a semantic web structure for lemmas in any language (cf. Declerck and Siegel, 2019).

6 CONCLUSION

We have presented our work on connecting lexical resources of BSL, NGT, LSF and DSGS semantically through a multilingual wordnet. This complements earlier work done for DGS and GSL and reported on in deliverable D6.3 (Bigeard et al., 2022). All core languages of the project are now covered, except for LIS, for lack of usable resources.

This work has resulted in a dataset of 7929 signs in 6 sign languages linked to 11806 synsets.

Additionally, a web interface has been launched to make the index accessible for the general public at https://www.sign-lang.uni-hamburg.de/easier/sign-wordnet/

The next deliverable within this task will be D6.5, where we will work on languages from outside the project. We have already contacted owners of resources for Polish Sign Language (PJM) and Swedish Sign Language (STS) to this end.

REFERENCES

Bigeard, Sam, Marc Schulder, Maria Kopf, Thomas Hanke, Kiriaki Vasilaki, Anna Vacalopoulou, Theodor Goulas, Athanasia-Lida Dimou, Stavroula-Evita Fotinea, and Eleni Efthimiou (2022). *Initial Interlingual Index for DGS and GSL*. Project deliverable D6.3. Version 1.0. EASIER Consortium. DOI: 10.25592/uhhfdm.10171.

Bond, Francis and Kyonghee Paik (2012). "A Survey of WordNets and their Licenses". In: *Proceedings of the 6th Global WordNet Conference*. Matsue, Japan, p. 8. ISBN: 978-80-263-0244-5. URL: https://bond-lab.github.io/pdf/2012-gwc-wn-license.pdf (visited on 12/09/2021).

Bond, Francis, Piek Vossen, John Philip McCrae, and Christiane Fellbaum (2016). "CILI: the Collaborative Interlingual Index". In: *Proceedings of the Eighth Global WordNet Conference*. Bucharest, Romania: University of Iasi, pp. 50–57. ISBN: 978-973-0-20728-6. URL: https://aclanthology.org/2016.gwc-1.9.

Crasborn, Onno, Richard Bank, Inge Zwitserlood, Els van der Kooij, Anique Schüller, Ellen Ormel, Ellen Yassine Nauta, Merel van Zuilen, Frouke van Winsum, and Johan Ros (2016). "Linking Lexical and Corpus Data for Sign Languages: NGT Signbank and the Corpus NGT". In: *10th International Conference on Language Resources and Evaluation (LREC 2016). Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining* (Portorož, Slovenia). Ed. by Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch. Paris, France: European Language Resources Association (ELRA), pp. 41–46. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/16023.pdf.

Declerck, Thierry and Melanie Siegel (2019). "OntoLex as a possible Bridge between WordNets and full lexical Descriptions". In: *Proceedings of the 10th Global Wordnet Conference*. GWC 2019. Wroclaw, Poland: Global Wordnet Association, pp. 264–271. URL: https://aclanthology.org/2019.gwc-1.34 (visited on 06/30/2023).

Ebling, Sarah, Katja Tissi, and Martin Volk (2012). "Semi-Automatic Annotation of Semantic Relations in a Swiss German Sign Language Lexicon". In: *8th International Conference on Language Resources and Evaluation (LREC 2012). Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon* (Istanbul, Turkey). Ed. by Onno Crasborn, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Jette Kristoffersen, and Johanna Mesch. Paris, France: European Language Resources Association (ELRA), pp. 31–36. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/12010.pdf.

Efthimiou, Eleni, Stavroula-Evita Fotinea, Athanasia-Lida Dimou, Theodoros Goulas, Panagiotis Karioris, Kiki Vasilaki, Anna Vacalopoulou, and Michalis Pissaris (2016). "From a Sign Lexical Database to an SL Golden Corpus – the POLYTROPON SL Resource". In: *10th International Conference on Language Resources and Evaluation (LREC 2016). Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining* (Portorož, Slovenia). Ed. by Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch. Paris, France: European Language Resources Association (ELRA), pp. 63–68. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/16003.pdf.

Efthimiou, Eleni, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Goudenove (2010). "DICTA-SIGN: Sign Language Recognition, Generation and Modelling with application in Deaf Communication”. In: *7th International Conference on Language Resources and Evaluation (LREC 2010). Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies* (Valletta, Malta). Ed. by Philippe Dreuw, Eleni Efthimiou, Thomas Hanke, Trevor Johnston, Gregorio Martínez Ruiz, and Adam Schembri. Paris, France: European Language Resources Association (ELRA), pp. 80–83. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/10027.pdf.

Efthimiou, Eleni, Kiki Vasilaki, Stavroula-Evita Fotinea, Anna Vacalopoulou, Theodoros Goulas, and Athanasia-Lida Dimou (2018). “The POLYTROPON Parallel Corpus”. In: *11th International Conference on Language Resources and Evaluation (LREC 2018). Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community* (Miyazaki, Japan). Ed. by Mayumi Bono, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, Johanna Mesch, and Yutaka Osugi. Paris, France: European Language Resources Association (ELRA), pp. 39–44. ISBN: 979-10-95546-01-6. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/18043.pdf.

Fenlon, Jordan, 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)*. URL: https://bslsignbank.ucl.ac.uk/.

Hamp, Birgit and Helmut Feldweg (1997). “GermaNet - a Lexical-Semantic Net for German”. In: *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, Spain: Association for Computational Linguistics, pp. 9–15. URL: https://aclanthology.org/W97-0802/ (visited on 07/21/2021).

Hanke, Thomas, Marc Schulder, Reiner Konrad, and Elena Jahn (2020). “Extending the Public DGS Corpus in Size and Depth”. In: *12th International Conference on Language Resources and Evaluation (LREC 2020). Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives* (Marseille, France). Ed. by Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch. Paris, France: European Language Resources Association (ELRA), pp. 75–82. ISBN: 979-10-95546-54-2. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/20016.pdf.

Konrad, Reiner, Thomas Hanke, Susanne König, Gabriele Langer, Silke Matthes, Rie Nishio, and Anja Regen (2012). “From form to function. A database approach to handle lexicon building and spotting token forms in sign languages”. In: *8th International Conference on Language Resources and Evaluation (LREC 2012). Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon* (Istanbul, Turkey). Ed. by Onno Crasborn, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Jette Kristoffersen, and Johanna Mesch. Paris, France: European Language Resources Association (ELRA), pp. 87–94. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/12023.pdf.

Konrad, Reiner, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, and Marc Schulder (2020). *MY DGS – annotated. Public Corpus of German Sign Language, 3rd release*. Dataset. DGS-Korpus project, IDGS, Hamburg University. DOI: 10.25592/dgs.corpus-3.0.

Kuder, Anna, Joanna Wójcicka, Piotr Mostowski, and Paweł Rutkowski (2022). “Open Repository of the Polish Sign Language Corpus: Publication Project of the Polish Sign Language Corpus”. In: *Proceedings of the LREC2022 10th Workshop on the Representation and Pro-*

Funded by the Horizon 2020
Framework Programme of the European Union

cessing of Sign Languages: Multilingual Sign Language Resources. Marseille, France: European Language Resources Association, pp. 118–123. URL: https://aclanthology.org/2022.signlang-1.18.

Lualdi, Colin P., Elaine Wright, Jack Hudson, Naomi K. Caselli, and Christiane Fellbaum (2021). "Implementing ASLNet V1.0: Progress and Plans". In: *Proceedings of the 11th Global Wordnet Conference*. Potchefstroom, South Africa: South African Centre for Digital Language Resources (SADiLaR), pp. 63–72. URL: https://aclanthology.org/2021.gwc-1.8 (visited on 07/21/2021).

Miller, George A., R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller (1990). "Introduction to WordNet: An On-line Lexical Database". In: *International Journal of Lexicography* 3. DOI: 10.1093/ijl/3.4.235.

Shoaib, Umar, Nadeem Ahmad, Paolo Prinetto, and Gabriele Tiotto (2014). "Integrating MultiWordNet with Italian Sign Language lexical resources". In: *Expert Systems with Applications* 41.5, pp. 2300–2308. ISSN: 09574174. DOI: 10.1016/j.eswa.2013.09.027.

Svenskt teckenspråkslexikon (2023). *Swedish Sign Language Dictionary online*. Lexical resource. Stockholm, Sweden: Department of Linguistics, Stockholm University. URL: https://teckensprakslexikon.su.se (visited on 06/30/2023).

Wójcicka, Joanna, Anna Kuder, Piotr Mostowski, and Paweł Rutkowski, eds. (2020). *Open Repository of the Polish Sign Language Corpus*. Dataset. Warsaw, Poland: Faculty of Polish Studies, University of Warsaw. URL: https://www.korpuspjm.uw.edu.pl (visited on 06/30/2023).

Funded by the Horizon 2020
Framework Programme of the European Union