PaSCo1: A Parallel Video-SiGML Swiss French Sign Language Corpus in Medical Domain

David, Bastien; Bouillon, Pierrette; Mutal, Jonathan David; Strasly, Irene; Gerlach, Johanna; Spechbach, Hervé

PaSCo1: A Parallel Video-SiGML Swiss French Sign Language Corpus in Medical Domain

Bastien David[1], Pierrette Bouillon[1], Jonathan Mutal[1], Irene Strasly[1], Johanna Gerlach[1] and **Hervé Spechbach[2]**

[1] TIM/FTI, University of Geneva, Geneva, Switzerland
[2] DMPR-HUG, Geneva University Hospitals, Geneva, Switzerland

Correspondence: bastien.david, pierrette.bouillon, jonathan.mutal, irene.strasly, johanna.gerlach@unige.ch
herve.spechbach@hug.ch

Abstract

This article introduces the parallel sign language translation corpus, PaSCo1, developed as part of the BabelDr project, an automatic speech translation system for medical triage. PaSCo1 aims to make a set of medical data available in Swiss French Sign Language (LSF-CH) in the form of both videos signed by a human and their description in G-SiGML mark-up language. We describe the beginnings of the corpus as part of the BabelDr project, as well as the methodology used to create the videos and generate the G-SiGML language using the SiGLA platform. The resulting FAIR corpus comprises 2 031 medical questions and instructions in the form of videos and G-SiGML code.

1 Introduction

Today, there are few corpora available for sign languages (SLs), which slows the development of data-driven systems (Table 1). The existing corpora also remain marginal compared with those developed for spoken languages. In the context of neural machine translation, Vandeghinste et al. (2024, p.122) mention that data available for the largest SL corpus (Prillwitz et al., 2008) is still 10 times smaller than its Europarl equivalent (Koehn, 2005). Some SLs are also very poorly represented, such as Swiss French Sign Language (LSF-CH, *Langue des signes française de Suisse romande*)[1].

There are several problems that make the development of SL corpora difficult. SLs are not written languages and SL corpora are mainly stored in video format. Also, many of these corpora contain interpreted speeches. They are therefore rarely parallel to the originals, which makes it difficult to align the source and the target texts. In addition, the large number of signers in general leads to dialectal variation and a lack of uniformity in the language. The collection and annotation of SL corpora is also longer than for their spoken equivalents with the lack of flexibility in the editing and post-production work on some videos further extending the working time. Finally, the recording format used may become obsolete after a few years and conversion to another standard is not always possible, leading to the loss of recorded data (Chiriac et al., 2016).

To annotate videos, a writing system that is descriptive and machine-readable has several advantages. It can be easily manipulated, adapted, and interpreted, unlike data from video recordings. G-SiGML (Elliott et al., 2004), for example, is an XML mark-up language based on the Hamburg Notation System for Sign Languages (HamNoSys) (Hanke, 2004). It is composed of several levels of information specific to SLs and is able to control a JASigning virtual animation (Ebling and Glauert, 2013). This code has also been used as a pivot for the development of machine translation and annotation systems, for example in Skobov and Lepage (2020) or Mutal et al. (2024).

In this paper, we present PaSCo1[2], a corpus translating French medical triage questions and instructions into LSF-CH, composed of human-signed videos and the corresponding G-SiGML code. We introduce the French source corpus (section 2) and the different SL translation methodologies (section 3) before describing the content of the corpus (section 4) and its public metadata.

[1] In 2023, there were almost 20 000 deaf people will live in Switzerland, 20 % of them in French-speaking cantons (Boyes-Braem and Rathmann, 2010). The canton of XXX is currently the only French-speaking canton that recognizes LSF-CH in its constitution.

[2] PaSCo1 repository: https://doi.org/10/gqnbps, consulted on April 24, 2025.

Proceedings of the Third International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), pages 37–43, June 24, 2025

Dataset	Reference		Recording		Area
	Date	Author	Human	G-SiGML	
GSLC	2007	Efthimiou and Fotinea	Yes	No	Education
NGT-Corpus	2008	Crasborn and Zwitserlood	Yes	No	N/S
DGS-Korpus	2008	Prillwitz et al.	Yes	No	N/S
RWTH-Phoenix	2012	Forster et al.	Yes	No	Weather
Dicta-Sign	2012	Matthes et al.	Yes	No	Travel
DGS-Corpus	2020	Hanke et al.	Yes	No	N/S
GSL Dataset	2020	Adaloglou et al.	Yes	No	Service
BOBSL	2021	Albanie et al.	Yes	No	N/S
OpenASL	2022	Shi et al.	Yes	No	N/S
Youtube-ASL	2023	Uthus et al.	Yes	No	N/S
PaSCo1	**2024**	**David et al.**	**Yes**	**Yes**	**Health**

Table 1: Some Examples of Public SL Corpora

2 Source corpus

The source corpus (language: French) was developed for the BabelDr[3] translation application as part of a collaboration between the Faculty of Translation and Interpreting (FTI) and the Outpatient emergency unit of the primary care medicine ward in Geneva University Hospitals (HUG), supported by the HUG private foundation (Rayner et al., 2016).

BabelDr is a fixed-sentence translation system. It is based on 11 089 sentences pre-translated by humans linked to more than a million variants using a grammar (Bouillon et al., 2021). The translation system works in four stages: (1) the doctor asks a question orally, (2) the system uses speech recognition to recognize the sentence, (3) as in a translation memory, the result of the speech recognition is linked to the closest sentence in the database using neural methods trained on the synthetic corpus generated by the grammar: (4) if the doctor validates this result, the sentence is then finally shown to the non-native speaker patient.

The source corpus has already been translated into written and spoken forms in 11 different languages (Arabic, Algerian Arabic, Moroccan Arabic, Tunisian Arabic, Dari, Farsi, Russian, Simple English, Spanish, Tigrigna and Ukrainian). A partial version in LSF-CH has been added using human-recorded videos and avatar animations. The aim was to be able to compare users' perceptions of these two modalities, in terms of usability and more specifically patient satisfaction, a very important criterion in medicine for adherence to treatment, for example Janakiram et al. (2020) and David et al. (2022b). The following section describes the translation methodology used to translate the French source corpus into SL.

[3]BabelDr website: https://babeldr.unige.ch/, consulted on Avril 24, 2025

3 Translation methodology

The translation methodology follows a two-step process. First, a set of reference recordings is produced by human translators (section 3.1). These reference translations are then used to generate the G-SiGML code using a rule-based approach (section 3.2).

3.1 Reference translation

The LSF-CH reference translations were produced in a recording studio at the University of Geneva. The team consisted of a deaf nurse, a hearing doctor, a hearing interpreter and two deaf LSF-CH experts who all worked collaboratively to produce the final videos. Following team discussions, the deaf nurse was filmed for the final version of the translations (Strasly et al., 2018).

The recording was done using the LiteDevTool online platform, which enables the video content to be stored immediately and avoids any post-production work. The captured video stream is displayed, validated and then recorded in real time (Gerlach et al., 2018). During the translation process, three deaf individuals from the local deaf community—who are also LSF-CH teachers—regularly came to the university to ensure the

translated content was accurate and easy to understand. After the initial set of translations was filmed, the project coordinator held seven focus groups with members of the local deaf community to gather feedback, which was then used to refine existing translations and adapt the additional content that had to be added to the existing corpus (Strasly, 2024).

3.2 Translation into animation

The second phase of the project was to develop the G-SiGML code for the reference corpus. This mark-up language can be used to generate a fully synthesized animation.

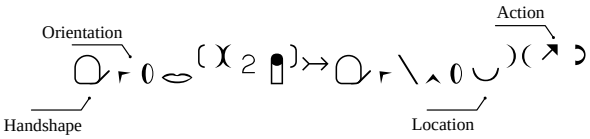

Figure 1: Lexical Resource: HamNosys Notation [HELLO - LSF-CH]

The code was generated using SIGLA,[4] a web-based application developed to generate G-SiGML code from a glossary and translation grammar (David et al., 2022c).

SIGLA has a storage function (GLOSSARY and GRAMMAR) and a code generation and animation function (GENERATE).

The GLOSSARY feature loads and stores lexical data written using the Hamburg Notation System for Sign Languages (HAMNOSYS) (Prillwitz et al., 1989), a phonological language notation system describing the physical components of each hand gesture. Figure 1 shows the phonological composition (hand shape, palm and finger orientation, location and movement) of the sign HELLO (LSF-CH) in HAMNOSYS. For the BabelDr project, 608 glosses/HAMNOSYS entries were manually produced: 370 nouns, 82 actions, 57 adjectives, 36 adverbs, 19 transfer signs, 15 pronouns, 8 prepositions, 5 forms of punctuation, 3 interjections and 3 conjugation terms.

The GRAMMAR feature loads and stores synchronous context-free grammar rules. A rule is multi-channel and maps sentences to the appropriate sequence of glosses/ HAMNOSYS entries. Each gloss is synchronized to different non-manual channels and lip expressions that are pre-registered in G-SiGML. Each rule can also introduce terminal or non-terminal variables. Our grammar resource contains nearly 450 rules, 115 non-terminal symbols and 608 terminals. Several grammatical and lexical sets can be loaded onto SIGLA.

Users can load the stored lexical and grammatical content they require, as well as the rule they wish to translate, using the GENERATE functionality. SIGLA then transforms the rule into the sign table (Rayner et al., 2016), the intermediate representation of the synchronized signed sentence. The matrix in figure 2 shows an example of a grammar rule with the corresponding sign table for the sentence "Hello, I am the nurse". This sign table is then translated into G-SiGML notation (Elliott et al., 2004). The gloss encodes individual sign features from the HAMNOSYS, while the other lines represent pre-registered non-manual features. In addition to the G-SiGML code corresponding to the rule, the generation output includes a JASigning animation, the translation in written format and the sign table.

Once this process is completed, the G-SiGML codes are imported into the BabelDr in CSV format. Each new import overwrites the previous one in order to match the latest corrections made in the initial resources. The grammar can now generate 1 234 828 synthetically signed sentences, 6 200 of which have been imported into BabelDr.

4 Parallel Sign language Corpus (PaSCo1)

The PaSCo1[5] corpus has been available since August 2022 though the institutional repository YARETA. Respecting the FAIR principles of access to information, the data can be downloaded easily and securely.

This corpus makes our medical data available in formats adapted to sign language. It is characterized by two features:

- **Domain**: Unlike many sign language corpora, PaSCo1 specializes in the medical field. It translates a set of questions and instructions related to the medical emergency context.
- **Composition**: PaSCo1 is a parallel corpus composed of French triage questions, LSF-CH videos and the corresponding descriptions in G-SiGML.

[4]SIGLA application: https://babeldr.unige.ch/demos-and-resources#sigla, consulted on May 24, 2025.

[5]PaSCo1 repository: https://doi.org/10/gqnbps, consulted on April 24, 2025.

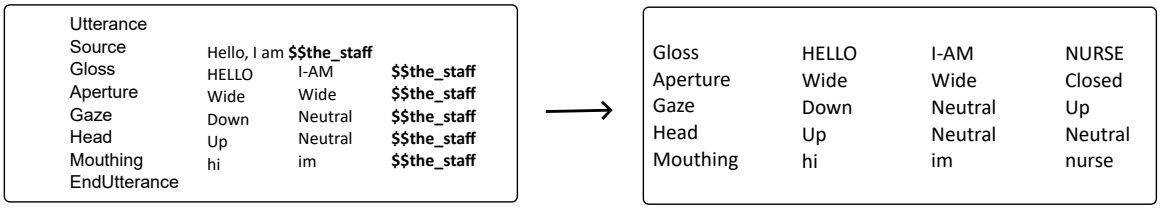

Figure 2: Grammar Resource: Translation Rule and Sign Table [I am the nurse - LSF-CH]

Corpus	1	2	3
BabelDr	N:11 089	N:82 152	N:2 260
PaSCo1	N:2 031 %:18,3	N:17 023 %:20,7	N:833 %:36,8

Table 2: Number (N) of Sentences (1), Tokens (2) and Types (3) in the BabelDr and PaSCo1 Source Corpus

PaSCo1 consists of two sub-folders containing 2 031 MP4 files. The first sub-folder contains the reference translations which correspond to almost 5 hours of video recordings, while the second contains the corresponding G-SiGMLs. A README file is attached to the main folder and provides the correspondences between the source sentences ("*Do you take vitamins every day?*"), the standardized names of the video files (*BabelDr_LSFCH_1804*), the G-SiGML notations and subtitle files.

18,3 % of the sentences in the BabelDr source corpus are now available in PaSCo1 (Table 2). At the lexical level, this represents 20,7 % of all words (tokens) and 36,8 % of unique words present (types). The corpus contains 20 181 gloss tokens, which means an average of 9,9 glosses per sentence.

Table 3 shows the distribution of the 2 031 sentences in PaSCo1 according to BabelDr's domains. Some sentences may belong to several domains. For example, PaSCo1 translated nearly 62,71 % of the sentences in the COVID domain, 38,87 % in checkup and 32,51 % in traumatology.

Comparing file sizes, the size of the video recordings is 5,6 GB, while the file containing the G-SiGML animation codes is just 0,02 GB. The total size needed to store the recordings should reach 30,5 GB, while a complete file of G-SiGML translations should not exceed 0,1 GB.

5 Conclusion

PaSCo1 is a French sign language medical translation corpus from French-speaking Switzerland (LSF-CH). It provides access to a set of phrases commonly used by emergency doctors when triaging patients, which are available in both video and G-SiGML formats. The latter was produced using the SIGLA platform, which generates the code, as well as the animation with a grammar and lexicon.

This corpus enables several research possibilities, including the descriptive analysis of videos in SLs, the automatic construction of virtual avatars driven by G-SiGML, the comparison of human recordings and virtual animation, the evaluation of virtual animation in the medical context and the automation of annotation in G-SiGML.

The remainder of the G-SiGML codes are already available on the BabelDr platform and will be available on YARETA soon. Currently, 1 730 selected new sentences are being translated by a team of deaf students from the University of Geneva's LSF-CH academic translation programme and will be added to the reference corpus.[6]

References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th. Papadopoulos, Vassia Zacharopoulou, George J. Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2022. A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition. *IEEE Transactions on Multimedia*, 24:1750 – 1762.

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BBC-Oxford British Sign Language Dataset.

Pierrette Bouillon, Johanna Gerlach, Jonathan Mutal, Nikos Tsourakis, and Hervé Spechbach. 2021. A

[6]UNIGE page: https://perma.cc/ZD3B-3E22, consulted on April 24, 2025

Domain	Source Corpus			PaSCo1		
	1	2	3	1	2	3
Checkup	N:4 784	N:39 471	N:1 737	N:1 862 %:38,87	N:15 647 %:39,64	N:747 %:43,00
Chest	N:4 721	N:39 217	N:1 720	N:1 269 %:26,87	N:10 530 %:26,85	N:714 %:41,51
Covid	N:236	N:1 435	N:348	N:148 %:62,71	N:934 %:65,08	N:293 %:84,19
Traumatology	N:3 325	N:26 958	N:1 524	N:1 081 %:32,51	N:8 745 %:32,43	N:643 %:42,19
Follow-up	N:1 551	N:13 316	N:1 051	N:242 %:15,60	N:1 583 %:11,88	N:412 %:39,20
Dermatology	N:3 682	N:31 082	N:1 594	N:940 %:25,52	N:7 622 %:24,52	N:660 %:41,40
Habits	N:1 273	N:11 440	N:745	N:60 %:4,71	N:431 %:3,76	N:146 %:19,59
Abdomen	N:7 094	N:59 219	N:1 846	N:1 987 %:28,00	N:16 728 %:28,24	N:823 %:44,58
Head	N:5 399	N:43 403	N:1 781	N:1 255 %:23,24	N:10 270 %:23,66	N:723 %:40,59

Table 3: Number (N) of Sentences (1), Tokens (2) and Types (3) in the BabelDr Source Corpus and PaSCo1

Speech-enabled Fixed-phrase Translator for Healthcare Accessibility. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, NLP4PI 2021, pages 135 – 142, Online. Association for Computational Linguistics (ACL).

Penny Boyes-Braem and Christian Rathmann. 2010. Transmission of Sign Languages in Northern Europe. In Diane Brentari, editor, *Cambridge Language Surveys: Sign Languages*, pages 19 – 45. Cambridge University Press, Cambridge.

Ionut A. Chiriac, Lăcrămioara Stoicu-Tivadar, and Elena Podoleanu. 2016. Romanian Sign Language Oral Health Corpus in Video and Animated Avatar Technology. In Valentina E. Balas, Lakhmi C. Jain, and Branko Kovačević, editors, *Soft Computing Applications*, volume 356 of *Advances in Intelligent Systems and Computing*, pages 279 – 293. Springer, Cham.

Onno Crasborn and Inge Zwitserlood. 2008. The Corpus NGT: An Online Corpus for Professionals and Laymen. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages*, LREC 2008, pages 44 – 49, Marrakech, Morocco. European Language Resources Association (ELRA).

Bastien David, Pierrette Bouillon, Irene Strasly, Jonathan Mutal, Johanna Gerlach, and Hervé Spechbach. 2022a. Parallel Sign Language Corpus.

Bastien David, Jonathan Mutal, Irene Strasly, Pierrette Bouillon, and Hervé Spechbach. 2022b. Babeldr: un système de traduction du discours médical vers l'animation virtuelle signée. In *12e conférence de l'IFRATH sur les technologies d'assistance*, Handicap 2022, Paris, France. IFRATH.

Bastien David, Jonathan Mutal, Irene Strasly, Johanna Gerlach, and Pierrette Bouillon. 2022c. SIGLA: une plateforme de développement d'animations en langue des signes. In *Technologies du Langage Humain et Accès Interactif à l'Information*, TLH-JAII 2022, pages 22 – 24, Paris, France.

Sarah Ebling and John Glauert. 2013. Exploiting the Full Potential of JASigning to Build an Avatar Signing Train Announcements. In *Third International Symposium on Sign Language Translation and Avatar Technology*, SLTAT 2023, Chicago, IL, USA. s.n.

Eleni Efthimiou and Stavroula-Evita Fotinea. 2007. GSLC: Creation and annotation of a Greek sign language corpus for HCI. In Constantine Stephanidis, editor, *Universal Access in Human Computer Interaction*, volume 4554 of *Lecture Notes in Computer Science*, pages 657 – 666. Springer, Berlin, Heidelberg.

Ralph Elliott, John Glauert, Vince Jennings, and Richard Kennaway. 2004. An Overview of the SiGML Notation and SiGMLSigning Software System. In *Proceedings of the 1st Workshop on the Representation and Processing of Sign Languages*, LREC 2004, pages 98 – 104, Lisbon, Portugal. European Language Resources Association (ELRA).

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. 2012. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, LREC 2012, pages 3785 – 3789, Istanbul, Turkey. European Language Resources Association (ELRA).

Johanna Gerlach, Hervé Spechbach, and Pierrette Bouillon. 2018. Creating an Online Translation Platform to Build Target Language Resources for a Medical Phraselator. In *40th Conference on Translating and the Computer*, TC40, pages 60 – 65, London, UK. AsLing.

Thomas Hanke. 2004. HamNoSys: Representing sign language data in language ressources and language processing contexts. In *Proceedings of the 1st Workshop on the Representation and Processing of Sign Languages*, LREC 2004, Lisbon, Portugal. European Language Resources Association (ELRA).

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in size and depth. In *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages*, LREC 2020, pages 75 – 82, Marseille, France. European Language Resources Association (ELRA).

Antony A. Janakiram, Johanna Gerlach, Alyssa Vuadens-Lehmann, Pierrette Bouillon, and Hervé Spechbach. 2020. User Satisfaction with a Speech-Enabled Translator in Emergency Settings. *Studies in Health Technology and Informatics*, 270:1421 – 1422.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X*, MTSummit 2005, pages 79 – 86, Phuket, Thailand.

Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. 2012. Dicta-Sign: Building a Multilingual Sign Language Corpus. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages*, LREC 2012, pages 117 – 122, Istanbul, Turkey. European Language Resources Association (ELRA).

Jonathan D. Mutal, Raphael Rubino, Pierrette Bouillon, Bastien David, Johanna Gerlach, and Irene Strasly. 2024. Improving sign language production in the healthcare domain using UMLS and multitask learning. In *Proceedings of the 1st Workshop on Patient-Oriented Language Processing*, LREC-COLING 2024, pages 1–7, Torino, Italia. European Language Resources Association (ELRA).

Siegmund Prillwitz, 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 *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages*, LREC 2008, pages 159 – 164, Marrakech, Morocco. European Language Resources Association (ELRA).

Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, and Jan Henning. 1989. *Hamburg Notation System for Sign Languages: An Introductory Guide*, volume 5 of *International Studies on Sign Language and the Communication of the Deaf*. Signum, Hamburg, Germany.

Manny Rayner, Alejandro Armando, Pierrette Bouillon, Sarah Ebling, Johanna Gerlach, Sonia Halimi, Irene Strasly, and Nikos Tsourakis. 2016. Helping Domain Experts Build Phrasal Speech Translation Systems. In José F. Quesada, Francisco-Jesús Martín Mateos, and Teresa Lopez-Soto, editors, *Future and Emergent Trent in Language Technology*, volume 9577 of *Lecture Notes in Artificial Intelligence*, pages 41 – 52. Springer, Cham.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Open-Domain Sign Language Translation Learned from Online Video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365 – 6379, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics (ACL).

Victor Skobov and Yves Lepage. 2020. Video-to-HamNoSys Automated Annotation System. In *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages*, LREC 2020, pages 209 – 216, Marseille, France. European Language Resources Association (ELRA).

Irene Strasly. 2024. *De la traduction médicale sur le terrain à la formation universitaire : étude de cas par la recherche-action sur la communauté sourde et la langue des signes française de Suisse romande*. Ph.D. thesis, Université de Genève.

Irene Strasly, Tanya Sebaï, Evelyne Rigot, Valentin Marti, Jesus Manuel Gonzalez, Johanna Gerlach, Hervé Spechbach, and Pierrette Bouillon. 2018. Le projet BabelDr: Rendre les informations médicales accessibles en langue des signes de Suisse romande (LSF-SR). In *Proceedings of the 2nd Swiss Conference on Barrier-free Communication*, BfC 2018, pages 92 – 96, Geneva.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A Large-Scale, Open-Domain American Sign Language-English Parallel Corpus. In *Proceedings of the 37th International Conference on*

Neural Information Processing Systems, NIPS 2023, Red Hook, NY, USA. Curran Associates Inc.

Vincent Vandeghinste, Mirella De Sisto, Santiago Egea Gómez, and Mathieu De Coster. 2024. Challenges with Sign Language Datasets. In Andy Way, Lorraine Leeson, and Dimitar Shterionov, editors, *Sign Language Machine Translation*, volume 5 of *Machine Translation: Technologies and Applications*, pages 117 – 139. Springer, Cham.