Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 83–88
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

SignHunter – A Sign Elicitation Tool Suitable for Deaf Events

Thomas Hanke, Elena Jahn, Sabrina Wähl, Oliver Böse, Lutz König
Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Germany
{thomas.hanke, elena.jahn, sabrina.waehl, oliver.boese, lutz.koenig}@uni-hamburg.de

Abstract

This paper presents *SignHunter*, a tool for collecting isolated signs, and discusses application possibilities. *SignHunter* is successfully used within the DGS-Korpus project to collect name signs for places and cities. The data adds to the content of a German Sign Language (DGS) – German dictionary which is currently being developed, as well as a freely accessible subset of the *DGS Corpus*, the *Public DGS Corpus*. We discuss reasons to complement a natural language corpus by eliciting concepts without context and present an application example of *SignHunter*.

Keywords: Elicitation tools, Mixed methods, Sign names for places and cities

1. Introduction

The use of corpora as language resources in the scientific exploration of natural language is commonly considered as state-of-the-art. However, building corpora, signed or spoken, large enough for the proficient analysis of a specific research question, is very costly in time, effort and budget. Building reference corpora that mirror natural language is even more challenging, as these corpora need to be of considerable size. As the size of a corpus determines to what extent low-frequency concepts are included in the corpus, and as the size of a corpus is limited at cost, some low-frequency concepts (e. g. discipline-specific vocabulary, regional specifications, cultural characteristics or names of cities, places and locations) will naturally be missing in each corpus.

This becomes an issue especially in the context of corpus-based dictionary creation where the user expects the semantic ‘neighborhood’ of each entry to be included as well. Therefore, lexicographers sometimes need to complement corpus-based dictionary entries with other low-frequency concepts that are not included (in enough quantity) in the corpus. In order to base such entries on data and not on the lexicographer’s language intuition, supplementary data collection is needed.

In Langer et al. (2016), we introduced a system to collect data from members of the language community via a web-based application, the *Feedback System* (Wähl et al., 2018). This paper introduces *SignHunter*, a tool to elicit isolated concepts in any sign language at community events. *SignHunter* is used in the DGS-Korpus project to enhance the *DGS Corpus*, as well as a corpus based dictionary DGS-German with less-frequent sign names for places and cities.

2. The DGS-Korpus Project

The DGS-Korpus project is a long term project of the Academy of Sciences and Humanities in Hamburg (Prillwitz et al., 2008). The project’s aims are:

- to build a reference corpus of DGS and to publish a subset of this corpus to be freely accessible online,
- to compile and publish a dictionary DGS-German that is based on and linked with the *DGS Corpus*.

2.1. The DGS Corpus

The *DGS Corpus* is designed as a reference corpus that displays the natural everyday language of deaf persons in Germany and is composed of 560 hours of signed narrations and dialogues. Parts of it have been translated, others have been annotated in detail. The elicitation took place in different regions across Germany to cover regional variants. The corpus is balanced for the sex of the informants, four age groups and the regions in which recordings took place. For the elicitation of the corpus, the tasks were deliberately designed to cover a broad variety of topics with as little influence on the informants as possible (Nishio et al., 2010). A translated and annotated subset of 50 hours of the *DGS Corpus* is published as the *Public DGS Corpus* (Jahn et al., 2018). The *Public DGS Corpus* is a research resource for natural DGS that is freely accessible online via two different portals, *MY DGS* (Hanke et al., 2020) for the DGS community and *MY DGS – annotated* (Konrad et al., 2020) for the research community.

One of the most important underlying motives that affected decisions regarding the design of the publication formats of the corpus was that the resource built should account for the needs of different user groups: persons who use DGS as their main language, interpreters, students, teachers and researchers interested not only in linguistic research but also in the history, culture and sociology of deaf persons across Germany, as well as many others. [1] This motive remains a driving factor in the improvement and enhancement of the resources published by the DGS-Korpus project

2.2. The DGS-German Dictionary

The “*Digitales Wörterbuch der Deutschen Gebärdensprache*” (*DW-DGS*) [Digital Dictionary of German Sign Language] is being compiled on the basis of the *DGS Corpus* data. Its final version is to be published in 2023, with the first pre-release made available in 2020. Information given on the signs include variants, typical mouthings, sense definitions, German translational equivalents, exam-

[1] For a more detailed description of the selection process, the data contained in the corpus and choices regarding the design of the two different portals, see Jahn et al. (2018).

ple sentences taken directly from the corpus (Langer et al., 2018), synonyms, antonyms, information on regionality, collocations and compound-like structures (Langer et al., 2019), as well as signs with a similar form and related signs.

Currently only signs that can be attested in the *DGS Corpus* are included and described in the dictionary. As the corpus is relatively large, it can be assumed that many of the commonly used signs and their meanings can be found in the corpus. Still some low-frequency signs will not be attested for in the corpus.

However, it would be desirable to have them listed (e. g. name signs for cities) in the upcoming dictionary as the information may be handy and interesting for the future user. This is where additional data collection methods can be used as long as it is transparent to the user which information is not based on the corpus.

3. SignHunter

SignHunter is an app created by the DGS-Korpus project that enables the user to collect isolated signs, the semantics of which are easy to communicate on a computer screen. It presents the informants a set of concepts they may choose from. This is the main difference between *SignHunter* and typical word list elicitations: Informants are free to choose what items to answer and how many.

Typically, a word list task requires the informant to answer all items on the list. This procedure may result in heavily skewed data as with sign languages informants could spontaneously invent new signs or fingerspell terms as they feel the pressure to give answers when prompted.

On the other hand, the approach taken here shares limitations with word list tasks, namely that it is crucial that there is no doubt about what concept the informant has in mind when producing a sign. We share concerns about word lists as the only basis for a dictionary (Brien and Brennan, 1995; Johnston and Schembri, 1999), therefore data from *SignHunter* can only be a supplement to the corpus data that were collected for the DGS-Korpus project.

In *SignHunter*, the set of concepts of interest to the researchers may be presented as a word list, a word list combined with visual stimuli or any other graphical representation of a set of concepts, such as a map showing geographical entities.

3.1. Data collection with SignHunter

Users are seated in front of the computer. Having identified herself by providing an id number, the system presents a set of concepts to the user. Having chosen an item, the user is invited to contribute her sign(s) for the respective item by signing into the camera.

Optionally, the user can playback the recording and then choose to either delete or keep the recording. In the latter case, the recording is automatically annotated with the concepts the user herself had identified. Of course this annotation needs to be examined and verified manually by human annotators later, but the automatic link between item and answer eases the annotation process considerably.

Once the user has provided one or more signs for the concept chosen, she can choose another concept to sign or just quit. Thus the informant herself operates the recording session.

SignHunter does not collect metadata, only the informant's id – a running number provided by the data collector team. Any metadata needed for further analyses needs to be collected separately. If used in public events, obviously the tool is best used in contexts where minimal data on the informants are sufficient. Depending on the program used for further processing the recordings, extra metadata need to be manually linked with the respective informants' ids.

3.2. Elicitation Setup

While the concepts available are simply pairs of ids and text labels in the *SignHunter* database, the system allows most flexible presentations for both the selection of concept and the prompt pages for concepts chosen by the informant: For this purpose, *SignHunter* just displays HTML pages that can be customized in the data collection preparation phase as needed.

3.3. Technical Details

SignHunter is an app that runs on macOS and Windows desktop or laptop systems and, once installed, can be used both online (i. e. with access to a database server) or offline. Obviously, the computer needs to have a camera built-in or attached.

During the elicitation sessions, all media files collected with *SignHunter* are stored locally, however in order to connect the media files to their metadata (signer, concept, date) *SignHunter* needs to connect to a database. If the computer *SignHunter* is running on cannot connect to the internet, a database can be installed locally. *SignHunter* uses PostgreSQL as the database machine which is easy and quick to install locally. It is also possible to record more than one person at a time by using multiple computers. In that case, a network needs to be set up., e. g. by connecting two computers with a network cable or a network switch or wifi router for more than two computers.

By using a second screen connected to one of the data collection computers (or another machine also in the network), the system can be used to compute and display output graphics at regular intervals. This allows the team to communicate the data collection progress to the audience, potentially attracting more informants to take part. An example of this is shown in Section 4 in Figure 1.

3.4. Further Processing

Once the recordings are finished, the video files collected need to be transferred to the central repository used in the annotation environment. In addition, the recordings table (holding the ids of the informants, the signed concepts, the name of the computer where the movie was recorded, as well as the ids of the movie files) needs to be transferred from the first machine's database server to your central server. *SignHunter* data collected by the DGS-Korpus project is stored in the annotation tool and lexical database *iLex*[2], a multi-user application for annotation and lemmatisation of sign language data (Hanke, 2002; Hanke and

[2] https://www.sign-lang.uni-hamburg.de/ilex/

Auf den Kulturtagen wurden schon 1403 Städtenamen aufgenommen.

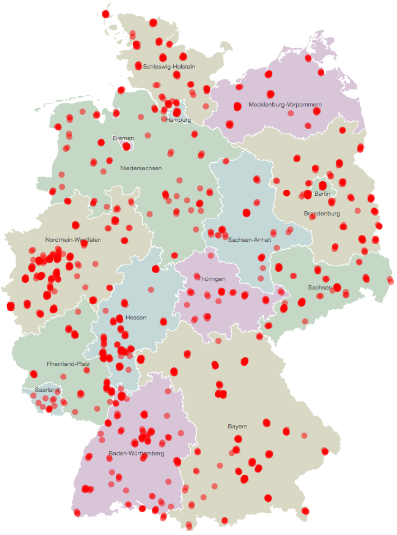


Figure 1: Count of recordings on the third/last day of the event. German caption: 'During the culture days, 1403 name signs for cities have already been collected'.

Storz, 2008). The fact that informants select the terms they want to sign and thus implicitly create an annotation of the target concept facilitates the annotation process to a great extent.

4. A Use Case: Name Signs for Cities

In addition to providing data to complement a corpus, collecting isolated signs with *SignHunter* is an excellent opportunity to directly engage a larger part of the language community beyond the group of 330 informants who participated in the corpus data collection. *SignHunter* made its first appearance at the the 6. Deutsche Kulturtage der Gehörlosen (6th German Culture Days of the Deaf) in 2018, a large event organized by and for the German Deaf Community. Name signs for cities and locations were collected with *SignHunter* on the three days of the event on two computers in parallel. A large screen showed the number of signs already collected during the event as well as dots for all cities on a map of Germany, as can be seen in Figure 1. The screen not only served to catch interest of bystanders, but was also an opportunity for team members to explain the data collection.

4.1. Elicitation Procedure

During the event, the name signs were collected by staff members of the DGS-Korpus project and student co-workers. The two computers were placed inside booths to guarantee good video quality without too much visual noise from bystanders and to allow some privacy. Informants were explained the goal and proceedings of the *SignHunter* elicitation and possible further uses of their data and had to read and sign an informed consent document. They were also explicitly told that they were free to chose what concepts and how many they want to record. Informants only needed to fill in their full name and address and sign the document. Further metadata like sex, age or others were not collected.

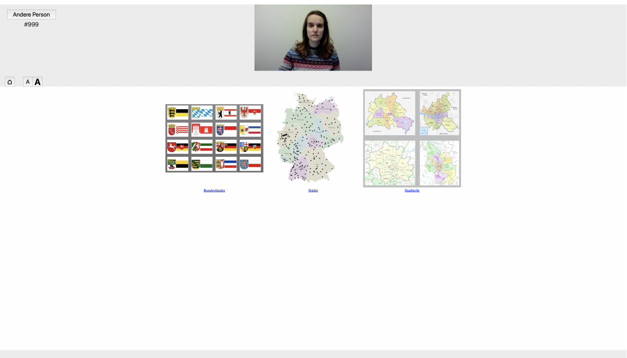


Figure 2: Selection of either federal states, German cities or districts of four large German cities (Berlin, Cologne, Hamburg, Munich).

Each informant received a personalized id for logging in. The informed consent forms that were used for the elicitation were numbered, with the running number used as the informant ids. The concept chooser was a two-steps approach, in which the first step was for informants to choose between three options: whether they would like to record sign names for either federal states of Germany, German city names or the names for districts of four large German cities (Berlin, Cologne, Hamburg, Munich), as shown in Figure 2. Clicking on one of the options lead to a site on which the concepts were presented. In the case of German city names, items were presented both as a list of city names in German as well as a map of the state with the cities superimposed as dots (see Figure 3 The prompt page then contained the German name as well as some typical landmark, as exemplified in Figure 4.

The possible items for the cities were pre-selected, with cities with more than 100.000 inhabitants at the elicitation time being included in the item list as well as cities that have or had a Deaf school, Hard of Hearing school, a Deaf club or a Hard of Hearing Club.

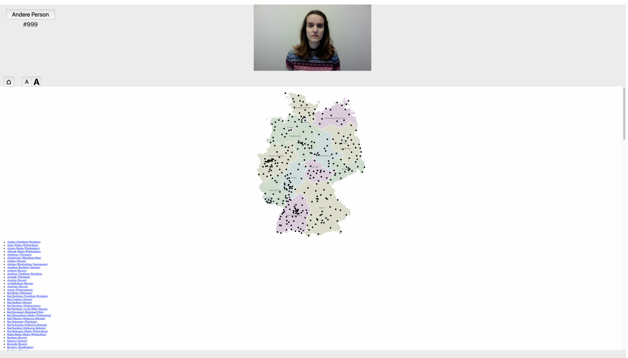
In total, informants could choose from 470 items, out of which 363 were city names, 16 were names of federal states, 19 were districts of Munich and Cologne each, 26 were districts of Berlin and 27 were districts of Hamburg.

The recording was operated by the informants themselves. When finished with the recordings, the informant logged out. During the elicitation, a staff member or a student co-worker was always present to answer (technical) questions. Due to the personalized id, it was possible for informants to return any time later and resume the recording.

4.2. Results

All in all, 135 persons took part in the *SignHunter* elicitation of name signs for German cities and locations. 404 distinct concepts were recorded. These 135 informants recorded 1978 answers in total, out of which 47 answers had to be excluded from analysis [3]. Answers were excluded for different reasons, mostly (in 35 cases) because the in-

[3] This as well as all following numbers are as evaluated in February 2020.



- Aachen (Nordrhein-Westfalen)
- Aalen (Baden-Württemberg)
- Achern (Baden-Württemberg)
- Albstadt (Baden-Württemberg)
- Altenburg (Thüringen)
- Altenkirchen (Rheinland-Pfalz)
- Amberg (Bayern)
- Anklam (Mecklenburg-Vorpommern)
- Annaberg-Buchholz (Sachsen)
- Ansbach (Bayern)
- Arnsberg (Nordrhein-Westfalen)
- Arnstadt (Thüringen)
- Arnstein (Bayern)
- Aschaffenburg (Bayern)
- Augsburg (Bayern)
- Aurich (Niedersachsen)
- Bad Berka (Thüringen)
- Bad Berleburg (Nordrhein-Westfalen)
- Bad Camberg (Hessen)
- Bad Endbach (Hessen)
- Bad Harzburg (Niedersachsen)
- Bad Homburg vor der Höhe (Hessen)
- Bad Kreuznach (Rheinland-Pfalz)
- Bad Mergentheim (Baden-Württemberg)
- Bad Oldesloe (Schleswig-Holstein)
- Bad Salzungen (Thüringen)
- Bad Schwartau (Schleswig-Holstein)
- Bad Segeberg (Schleswig-Holstein)
- Bad Säckingen (Baden-Württemberg)
- Baden-Baden (Baden-Württemberg)
- Bamberg (Bayern)
- Bautzen (Sachsen)
- Bayreuth (Bayern)
- Beeskow (Brandenburg)

Figure 3: Selection of German cities either by dots on a map or by selecting city names from an alphabetical list.

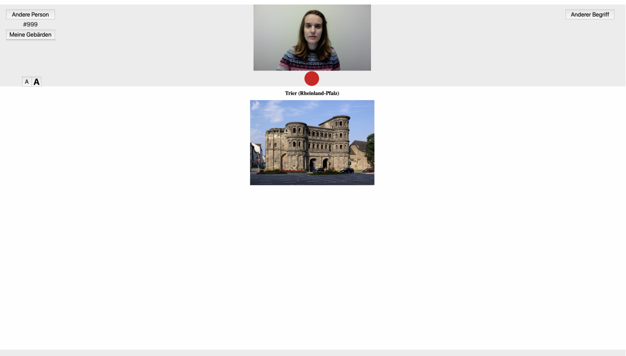


Figure 4: Example 'Trier'. German name of the city combined with a typical landmark as visual stimulus.

formant repeated the same answer twice or more. Another reason to exclude answers from further analysis was that the human annotators were unsure with respect to the assessment of the answer. From all 1978 recorded answers in total, in 47 cases the stimulus did not match the answer the informant gave. From these cases, 4 were excluded for other reasons from analysis, the other 43 cases were lemmatised nevertheless. In many mismatch cases, informants selected a federal state as stimulus and then recorded name signs for places in that respective federal state. In some cases, informants wanted to record sign names for places that were not included in the list of stimuli and thus used this little detour. As the answers recorded were nevertheless judged to be actual name signs and the informants seemed to have understood the task, these mismatch cases were not excluded from analysis. So, from the 1978 recorded answers, this leaves 1931 concepts recorded for lemmatisation and further analysis.

Out of these 1931 concepts, in 1558 cases the informants' answers were lemmatised with a single token. In 329 cases, there are two tokens per answer. This is predominantly due to the selected city names being compounds in German. In some cases the lemmatisation of two tokens per answer can also be traced back to reduplication. In a few cases, the answers were lemmatised into up to five tokens. For example, one informant signed every word of the German city name which is 'Brandenburg an der Havel'. The informant signed 'Brandenburg' as a compound-like structure [4] consisting of two signs and then added signs for 'an der Havel' [at the (river) Havel].

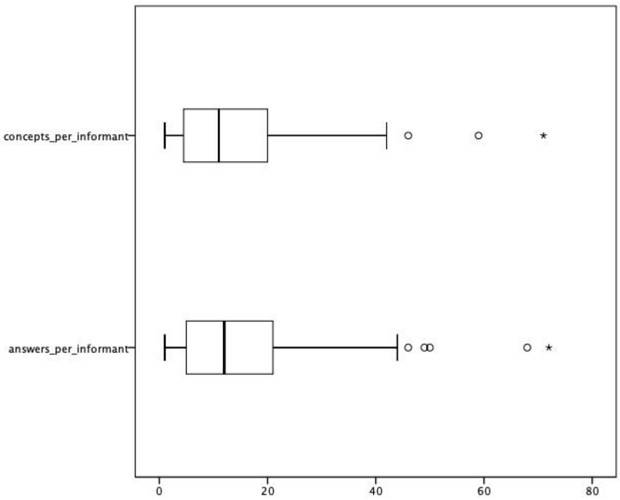


Figure 5: Number of recorded answers and recorded concepts per informant, with a total of 135 informants.

During the event, informants recorded between 1 and 72 answers, with the mean being 14.65 and the median being 12 answers per informant. The number of recorded concepts per informant ranges between 1 to 71, with the mean being 13.56 and the median being 11. As Figure 5 shows, both the total number of answers per informant and the recorded concepts per informant are distributed quite similarly, showing that informants recorded almost as many different concepts as answers. [5] The length of stay of informants in the recording cabins was not measured, but this data helps to predict the informants behavior for further elicitation at similar events.

As DGS is a sign language known to display a high number of variants (lexical as well as phonological) it could be expected that this might also show in the *SignHunter* recordings. However, the number of different variants attested for one concept was between 1 and 22, with the mean being 2.59 and the median being 2. As can be seen in Figure 6, for most concepts 1 to 7 variants were recorded, with more than 7 different variants per concepts being outliers in the data. The extreme outlier of 22 variants was attested for a concept that is a compound in German ("Baden-Württemberg") and was signed by most informants as a compound-like structure of which each part could have variants, e.g. for the first part "Baden" 3 possible variants were attested. This result mostly brings new challenges to light with respect to the representation of these units in the *DW-DGS*[6] as well as on *MY DGS* that need to be dealt with in the future.

The analysis of the collected data resulted in 123 new type

[4] As these constructions are loans from German they are called compound-like here because they are not DGS compounds in the narrow sense (cf. Becker (2003)).

[5] For Figure 5 and Figure 6 the small circles represent outliers with a distance of 1.5 - 3.0*iqr (interquartile range) from the first or third quartile, stars represent extreme outliers with a distance of more than 3.0*iqr to the first or third quartile.

[6] Two questions arise here. One, how to represent the compound-like structure? Two, how to interlink them within the dictionary?

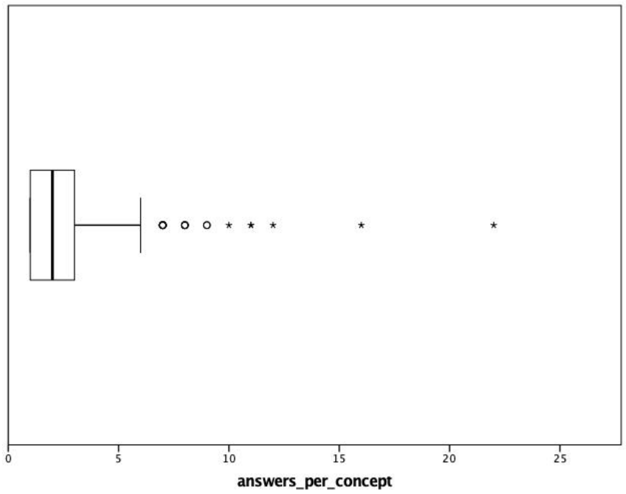


Figure 6: Number of recorded answers and recorded concepts per informant, with a total of 135 informants.

entries in our database. Thus our data could be supplemented and enriched by the data collection.

5. Conclusions and Outlook

The first application of *SignHunter* fulfilled our expectation for the data collection point of view, allowing us to include lists of city name signs in the dictionary on a solid data basis. In addition, the city name signs will be featured in the *MY DGS* community portal.

At the same time, feedback given during the event showed that many participants enjoyed the format. Setting up the HTML files for comparable cases is a straight-forward task. We include the project's focus group [7] into the decision process which vocabulary domains will be collected with *SignHunter* prospectively. By doing so we hope for future data collections to be relevant and interesting for the community as well as useful additions to the *DGS Corpus* and the *DW-DGS*. An interesting collection for future *SignHunter* recordings are for example name signs of famous deaf persons.

The focus group has also been trained in managing data collection sessions with *SignHunter*, allowing them to use the tool at events across Germany. As *SignHunter* runs on laptops, the equipment is relatively easy to transport and thus can be sent from one member of the focus group to another.

A feedback that was given during the event and that needs to be discussed for further elicitations is that some names informants would have liked to provide sign names for were not included in the list of concepts to choose from. In some cases, informants selected federal states as stimuli and then provided name signs for the concepts missing in the list. One possible solution would be to allow for free text input. However this would raise new difficulties, as the entered names may be ambiguous or may contain typos and thus would thus increase the annotation effort. Whether this is a useful addition should be checked by a test run first.

6. Acknowledgments

We are thankful for the support of the student co-workers during the elicitation. We would also like to thank the members of the focus group who were engaged in the process of developing *SignHunter* and gave valuable feedback with respect to the user interface and the operation of *SignHunter*. This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

7. Bibliographical References

Becker, C. (2003). *Verfahren der Lexikonerweiterung in der Deutschen Gebärdensprache*. Number 46 in International studies on sign language and the communication of the deaf. Signum, Seedorf, Germany.

Brien, D. and Brennan, M. (1995). Sign Language Dictionaries: Issues and Developments. In *Sign Language Research 1994: Proceedings of the Fourth European Congress on Sign Language Research*, pages 313–338, Munich, Germany. Hamburg : Signum.

Hanke, T. and Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 64–67, Marrakech, Morocco. European Language Resources Association.

Hanke, T. (2002). iLex – A tool for Sign Language Lexicography and Corpus Analysis. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 923–926, Las Palmas, Canary Islands, Spain. European Language Resources Association.

Jahn, E., Konrad, R., Langer, G., Wagner, S., and Hanke, T. (2018). Publishing DGS Corpus Data: Different Formats for Different Needs. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 107–114, Miyazaki, Japan. European Language Resources Association.

Johnston, T. and Schembri, A. C. (1999). On Defining Lexeme in a Signed language. *Sign Language & Linguistics*, 2(2):115–185.

Langer, G., König, S., Matthes, S., Groß, N., and Hanke, T. (2016). What Sign Language Lexicography Can Gain from a Mixed Method Approach: Corpus Data Supplemented by Crowd Sourcing. *Poster presented at the International Conference on Theoretical Issues in Sign Language Research*, Melbourne, Australia.

Langer, G., Müller, A., Wähl, S., and Bleicken, J. (2018). Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 483–497, Ljubljana, Slovenia. Ljubljana University Press.

[7] The focus group is a group of deaf individuals actively taking part in the Deaf community. Members of the focus group are from different regions across Germany. The focus group cooperates with the DGS-Korpus project as advisers as well as multipliers, maintaining the contact with the DGS community.

Langer, G., Müller, A., Wähl, S., and Hanke, T. (2019). The DGS-Korpus approach to including frequent sign combinations in a corpus-based electronic sign language dictionary. *Poster presented at the International Conference on Theoretical Issues in Sign Language Research*, Hamburg, Germany.

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., and Rathmann, C. (2010). Elicitation Methods in the DGS (German Sign Language) Corpus Project. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 178–185, Valletta, Malta. European Language Resources Association.

Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., and Schwarz, A. (2008). DGS Corpus Project – Development of a Corpus Based Electronic Dictionary German Sign Language / German. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 159–164, Marrakech, Morocco. European Language Resources Association.

Wähl, S., Langer, G., and Müller, A. (2018). Hand in Hand – Using Data from an Online Survey System to Support Lexicographic Work. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 7–12, Miyazaki, Japan. European Language Resources Association.

8. Language Resource References

Hanke, Thomas and König, Susanne and Konrad, Reiner and Langer, Gabriele and Barbeito Rey-Geißler, Patricia and Blanck, Dolly and Goldschmidt, Stefan and Hofmann, Ilona and Hong, Sung-Eun and Jeziorski, Olga and Kleyboldt, Thimo and König, Lutz and Matthes, Silke and Nishio, Rie and Rathmann, Christian and Salden, Uta and Wagner, Sven and Worseck, Satu. (2020). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release*. DGS-Korpus project, IDGS, Hamburg University, DOI `10.25592/dgs.meinedgs-3.0`.

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