An Extension of the NGT Dataset in Global Signbank

Ulrika Klomp, Lisa Gierman, Pieter Manders, Ellen Nauta, Gomèr Otterspeer, Ray Pelupessy, Galya Stern, Dalene Venter, Casper Wubbolts, Marloes Oomen, Floris Roelofsen[1]

SignLab, University of Amsterdam;
Kloveniersburgwal 48, 1012 CS Amsterdam, the Netherlands;
{u.klomp, l.gierman, p.w.j.manders, e.nauta, g.otterspeer, t.pelupessy, g.stern, d.venter2, c.wubbolts, m.oomen2, f.roelofsen}@uva.nl

Abstract

To support language documentation, linguistic research, and acquisition of Sign Language of the Netherlands (NGT), we are expanding the NGT dataset in the lexical database Global Signbank. Our most prioritized goal is to add ca. 11,000 glosses (entries). We further aim at adding ca. 3,000 example sentences and to provide linguistic information with as many glosses as possible. As for linguistic information, Signbank allows for extensive phonological descriptions of signs, and the addition of multiple senses per sign, which we would like to connect to synsets in the Multilingual Sign Language Wordnet. Additionally, we are recording extra video data: we make multiple videos of the same sign, taken from different angles, and videos with non-manual expressions. Furthermore, we are collecting motion capture data, for improved (automatic) sign language recognition and production in the future. In this paper, we describe how we proceed, the decisions that have been made so far, and future uses of the dataset.

Keywords: data collection, Signbank, sign language, NGT, documentation, motion capture

1. Introduction

The online lexical database Global Signbank (Crasborn et al., 2018) includes datasets from various sign languages, Sign Language of the Netherlands (*Nederlandse Gebarentaal*, NGT) being one of them. The NGT dataset was composed from 2007 to 2023 (Crasborn et al., 2020) and originated from the need to store and access glosses during corpus annotation work. At the end of 2023, the NGT dataset consisted of ca. 4,100 glosses, where each gloss has its own entry (see Section 2 for more information about entires). The main source of this dataset were annotations within the Corpus NGT (Crasborn, Zwitserlood and Ros, 2008). In 2023, responsibility for the NGT dataset and for changes in Global Signbank were transferred from the Radboud University Nijmegen to the University of Amsterdam. In 2024, a team of mostly Deaf NGT signers (henceforth: the NGT expert team) was composed to work on the Signbank project at the University of Amsterdam. This project runs till December 2024, and aims at extending the NGT dataset in multiple ways (as outlined in the following sections): 1. adding approximately 11,000 glosses; 2. adding example sentences; 3. adding and systematizing senses;

[1] Inspired by Nyst et al. (2022), we provide drawings of the name signs of our project members (following the author order). Illustrations by Casper Wubbolts.

4. adding more video data; 5. adding linguistic information; 6. collecting (and potentially adding) motion capture data. By expanding the NGT dataset in these ways, we envision to support the documentation of NGT, linguistic research into sign languages, and support learners of NGT. In this paper, we report on the current progress in this project, motivate our decisions so far and discuss potential ways of moving forward.

2. Adding Entries

Let us first go into the most significant extension of the NGT dataset; the addition of new glosses. Every gloss receives its own entry. With an entry, we mean a gloss with the video, its meanings and all additional information, visible in one webpage – see Figure 1 below.

When signs are encountered in corpus data and do not have an entry in the NGT dataset in Signbank yet, it is relatively easy to gloss them and include them in the database. But the Corpus NGT is no longer being actively annotated. Furthermore, since we aim to add thousands of lexical items in a short period of time, annotating corpus data is not efficient for gaining so many new entries, as annotation work is highly time-consuming in itself. The question then arose: how do we expand the dataset? We decided to let the Deaf NGT signers in our team be the data source, and document their knowledge of NGT. As inspiration for concepts to add, we are currently using: 1. themed word lists (e.g. on food or crafting); 2. the gloss list from the Flemish SignBank (Vlaams GebarentaalCentrum, 2024); 3. a list of words from the Corpus Spoken Dutch (CGN Version 2.0.3, 2014[2]). We collected about 6,000 potentially useable concepts until now. We are still thinking of efficient ways to collect 5,000 more concepts to reach our goal.

An important decision that was made to get from concept to entry, is that we only collect signs that are used in the Deaf community, instead of developing or making up signs ourselves. The main reason for this, is that we want to document the language as it is. Thus, if the team has not found an existing NGT sign for a certain concept, the concept is then removed from our list of potentially new entries, and thus not included in the database at this point. Originally, we made the decision that multiple people from the NGT expert team should know a certain sign before it could be included, but this was strikingly unworkable – multiple team members experienced that they were often the only team member who used a specific variant of a sign, due to the signers' different linguistic backgrounds (e.g. different schools and ages).

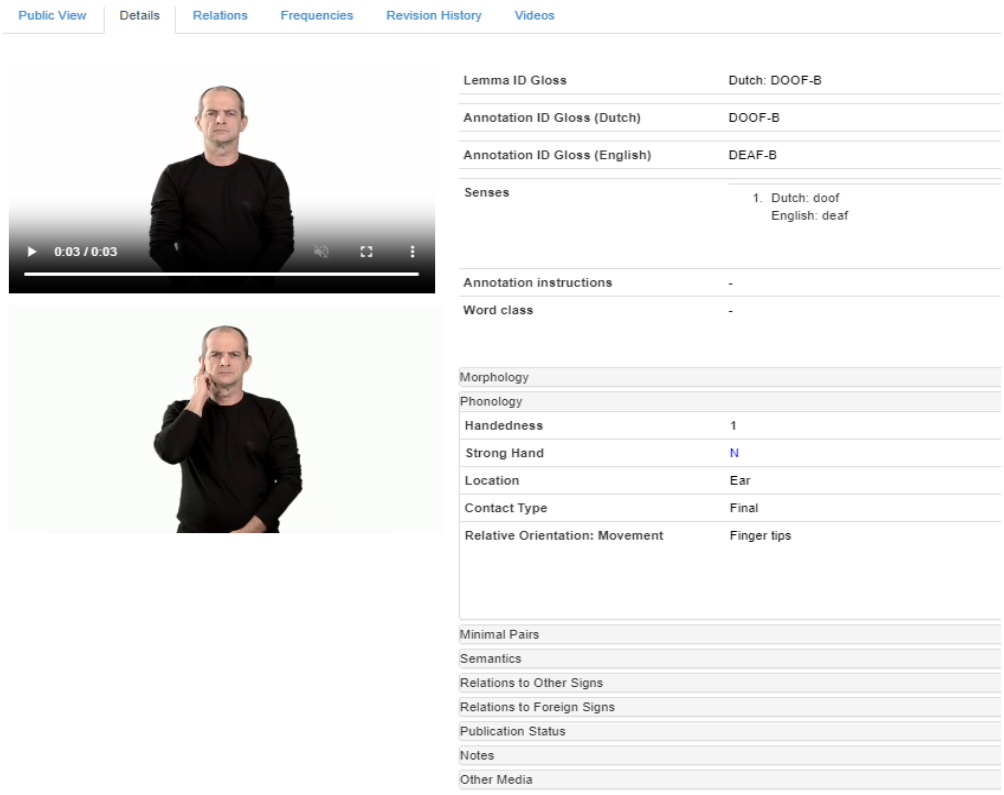

Figure 1: The entry of the gloss 'DEAF-B' in the NGT dataset on Global Signbank, with the phonological panel opened to show specifications (Crasborn et al., 2020).

[2] http://hdl.handle.net/10032/tm-a2-k6

To give the team more space, and at the same time guard that not (too many) idiosyncratic forms would be included, we therefore decided that a sign should be used by at least one team member and that this team member should know of at least one other deaf signing person that uses this sign. This did not only speed up the process of deciding upon signs that could be added to the NGT dataset, but was also more in line with the composition of the first dataset, where glosses came from signs that were simply encountered in the Corpus NGT data – sometimes only signed by one signer – and then added. This approach might still change in the future, or might be complemented with other collection projects. An example of an approach that we could take inspiration from for future work is to use an app like SignHunter, as described by Hanke et al. (2020).

When we create a new entry, we add the following information: Annotation gloss IDs, Lemma IDs, senses (possible meanings), a quick webcam recording of the sign and basic phonological information. For the NGT dataset, the decision has (previously) been made to use meaningful Annotation ID glosses (vs. a meaningless code or number, for example), where the ID gloss represents a common meaning. Soon after creating this entry, we expand the senses and (other) linguistic information, and replace the quick webcam video with a high-quality video, made in a professional recording studio. The video of the sign shows what we call a "phonological form", which represents the manual sign without mouthings, body movements or facial expressions. This is done so that the same manual form always receives the same Annotation ID gloss, even if the form has multiple, very different meanings (see also Section 4 below). The form is therefore articulated in the most neutral way. Since one phonological form can easily represent multiple concepts, we make sure the phonological forms of the proposed signs are not already in the database – perhaps under a different gloss than expected (see Section 6 on how to search for a phonological form). To clarify, it is important here that the intended *phonological form* is not represented in the database yet – the *meaning*, however, may be represented by another form. For example, a commonly used sign for the Dutch island 'Texel' refers to (the wool of) sheep. When the NGT expert team considers to add this sign to the database, we would first search for 'Texel' as a sense in the NGT dataset. We would then see that this sense is not in the database yet, meaning that the concept is not covered in the dataset. However, when we search for the phonological specifications, we see that the sign is already there, under the gloss SHEEP. We then add the sense 'Texel' to this phonological form, and do not make a new entry.

It is also possible, and even desirable for the purposes of our project, to add multiple signs for one concept, as variants. These different variants are likely to receive similar Annotation ID glosses, but are distinguished by different suffixes. For instance, the signs for 'dog' in NGT are currently represented by three different manual forms, with the Annotation ID glosses DOG-A, DOG-B and DOG-C (see Figure 2a, 2b and 2c, respectively).

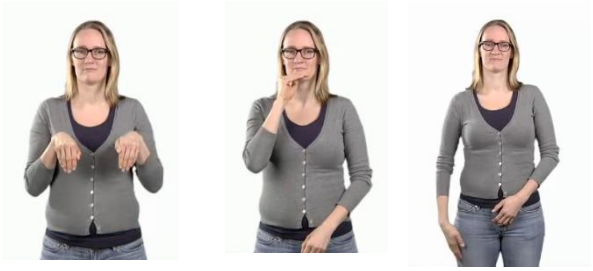

Figure 2a, 2b, 2c: The signs DOG-A (left), DOG-B (middle), DOG-C (right) in the NGT dataset in Global Signbank (Crasborn et al., 2020)

At the moment of writing[3], we added 1,600 glosses. One can imagine this process of going from a concept to a full-fledged entry is quite time consuming. One team member therefore developed the signCollect platform in which the work is automatized as much as possible (see Otterspeer, Klomp and Roelofsen (accepted) for a more elaborate explanation of this system). Through this platform, the team can propose glosses, keep track of signs that need consultation, check who will record the signs, make professional recordings and save everything together. We therefore expect to be able to speed up the process and to need less time for the next additions.

3. Adding Example Sentences

To provide use-in-context information, we will create example sentences for at least 3,000 glosses. Each example sentence will be accompanied by a gloss-by-gloss representation and a Dutch translation. These sentences will be linked to all the glosses it contains. Some of the 3,000 sentences will be developed in collaboration with a different project, where natural sentences for learners of NGT will be created with help of NGT teachers and parents of deaf children. To join forces, our team supports the recording and annotation process of these sentences, after which we may publish applicable sentences on the Signbank website. Other sentences will be taken from the Corpus NGT (Crasborn et al., 2008; Crasborn et al., 2015). Where possible, the original corpus fragment will

[3] April 4, 2024

be included on the Signbank website; otherwise, the sentence will be refilmed.

4. Adding and Systematizing Senses

When Global Signbank was developed, it included a functionality to add translation equivalents or keywords to a gloss (Cassidy et al., 2018). In 2023, keywords have been replaced by senses. A sense is a conceptual meaning and signs may easily have multiple senses – either because multiple distinct meanings are involved (as in homonyms), or because several related concepts apply (as in polysemes). The senses can then be grouped so that senses with a similar meaning are mentioned together. The change of providing (groups of) senses instead of keywords, has, however, not systematically been executed for the NGT dataset. Additionally, many English translations of the senses are still lacking. We therefore have several goals for the upcoming year: 1. add Dutch and English senses to the new and existing glosses; 2. systematically group the senses per concept; 3. connect the senses to synsets in the Multilingual Sign Language Wordnet (Bigeard et al., 2022).

So far, for the entries that also had English translations of the senses already, we checked the translation and regrouped the senses when necessary. For example, the Dutch/English groups of senses that are now available for the gloss PT:down (point down, see Figure 3) are: 1. in/in; 2. nu/now; 3. hier/here; 4. zuid/south; 5. daar/there; 6. dit/this. For every new gloss that we add, we immediately add the most salient sense in Dutch and English. We are developing guidelines to add Dutch and English senses and to group them systematically. Apart from the senses that we added to the new glosses, we added approximately 200 senses to already existing glosses.

Figure 3: The sign ‘PT:down’ in the NGT dataset on Global Signbank (Crasborn et al., 2020).

5. Adding More Video Data

So far, every entry has one video of the sign connected to the Annotation ID gloss, and one picture. The picture is usually the automatically derived middle frame of the video. In the video, the focus is on the plain articulation of the manual form without non-manual expressions (e.g. facial expressions, mouth actions) (see also Section 2 above). Since these plain signs are considered very unnatural, we aim to add one to three videos per entry where facial expressions and/or mouth actions are included in a natural way. These videos are not meant as replacements for the plain signs and they will not receive their own Annotation ID gloss. Instead, they should be seen as additional material that exemplifies possible natural articulation forms of this basic phonological form.

Both the neutral phonological video and the videos with possible articulation forms are recorded with three different cameras, to provide visual information from three different angles. The different angles will help human recognition of the sign, particularly if a handshape is difficult to perceive from the front angle, but can also be used to train automatic recognition by artificial intelligence. In our current set-up, one camera is situated to have the standard front perspective (similar to the perspective in Figure 3). The other two cameras are in a ca. 25-degree angle from the signer on the left and right of the middle camera, as we discovered these are optimal camera positions to capture multiple perspectives.

6. Adding Linguistic Information

Global Signbank allows for extensive description of linguistic information on different levels for every glossed sign – although the different datasets in Signbank vary in the extent to which they make use of these possibilities. For the NGT dataset, it has been a specific goal to collect phonological information (Cassidy et al., 2018) and therefore the possibilities to describe phonological characteristics of signs are quite elaborate. For each entry, one can fill out several fields on handshape(s), location, movement, orientation and, if necessary, other additional information about the sign. See, as an example, Figure 1 for the phonological description of the sign DEAF-B.

The description of phonology is mostly done through the selection of features in drop-down menus, to make the process easier, more standardized and less prone to typos. An advantage of this standardization is that it makes it easier to look up whether a phonological form is already in the dataset. When looking for a phonological form, one can fill out the relevant phonological information and find any relevant

sign without having to know the possible senses of the sign. Furthermore, Global Signbank allows for automatic searches for minimal pairs, for which the phonological information is used.

Note that, which phonological information is considered relevant, is also depending on the theoretical framework one is working with. The current structure of the phonology fields in Global Signbank reflects the line of work performed and followed at the Radboud University Nijmegen – and now by our team –, i.e., based on the work of e.g. Crasborn (2001) and van der Kooij (2002). At the moment of writing, we added phonological information for the majority of the 1,600 newly added signs.

Another section in the detailed view of an entry is related to morphological information, where one can describe if a sign is a compound, and if yes, what the individual compounded parts are. Within our project, we will add phonological information on the newly added glosses, and potentially investigate possibilities to describe compounds more elaborately. We will also look into the descriptions made by other datasets in Global Signbank, to enhance comparability among the datasets.

7. Collecting Motion Capture Data

To enhance and support developments in automatic sign language recognition and production, we are collecting motion capture data. By collecting data from the same signers and on the same signs that we collect for the NGT dataset, we create a big dataset with datapoints from different types (2D video data, 3D motion capture data) that all relate to the same concepts. In our current set-up, we use 12 infrared cameras, most of which are located on the ceiling to record from above, and a few on the ground to record from below (see Figure 4). We use the motion capture suit of Vicon, where we reconstruct a scene in 3D through the markers on this suit. Facial movements are captured with Live Link of Unreal, supported by ARKit of Apple. Additionally, we use StretchSense gloves to measure hand and finger movements (position and configuration of the hand and fingers). In Figure 4, one of our team members is preparing to produce the sign presented on the left screen, while wearing the motion capture suit and the StretchSense gloves. The screen on the right in Figure 4 shows an avatar, reconstructed from the signer in real time.

Processing of the data is done with Unreal Editor 5.3 to combine the data stream in a so-called FBX file. We use the signCollect system (Otterspeer, Klomp and Roelofsen, accepted) for directing the systems, collecting, saving and labelling the data. We are still practicing and experimenting with this set-up, but the results so far are promising: we have been able to record 1,000 glosses in this set-up by now. If it seems useful, the motion capture data will also be added to the NGT dataset in Signbank.

Figure 4: The set-up for recording motion capture data for lexical signs.

8. Future Directions

Global Signbank already has the possibility of performing automated searches and basic analyses. It is, for example, possible to automatically look for minimal pairs, or provide a distribution of the most frequently occurring handshapes. The more data in the NGT set, and the more precise their description, the more reliable these outcomes will be. Additionally, if one has access to multiple datasets, one can easily make cross-linguistic comparisons with these tools. Thus, expanding the NGT dataset supports linguistic research.

The original NGT dataset has frequency data for occurrence of the signs in the Corpus NGT available. In future research, we would like to collect frequency data on newly added signs as well (see e.g. Johnston (2012) on why lexical frequency data is relevant) – either by taking frequencies from the corpus, or by eliciting frequency measures from a large group of Deaf NGT signers.

With the extension of the NGT dataset, it will also be a richer platform for learners of NGT. The addition of signs, senses, examples sentences and videos from different angles support in acquiring a rich vocabulary and in understanding the different meanings a sign may have. The database could at some point also function as a dictionary. This is important, because not many sign language dictionaries exist for NGT. Furthermore, Signbank is freely accessible, and allows for searching from Dutch or English words (senses) to signs, but also the other way around, by searching with the phonological specifications.

Lastly, the video data and motion capture data will be used for automatic recognition and production of sign languages. By providing language models with our extensive dataset, we support the development of automatic translations from written language to sign language and vice versa.

9. Acknowledgments

This research is funded by Platform Digital Infrastructure Social Science and Humanities (PDI-SSH).

10. Bibliographical References

Bigeard, S., Schulder, M., Kopf, M., Hanke, T., Vasilaki, K., Vacalopoulou, A., Goulas, T., Dimou, A.-L., Fotinea, S.-E., and Efthimiou, E. (2022). Introducing Sign Languages to a Multilingual Wordnet: Bootstrapping Corpora and Lexical Resources of Greek Sign Language and German Sign Language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 9–15, Marseille, France. European Language Resources Association (ELRA).

Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E., and Johnston., T. (2018). Signbank: Software to Support Web Based Dictionaries of Sign Language. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2359–2364, Miyazaki, Japan. European Language Resources Association (ELRA).

Crasborn, O. (2001). *Phonetic implementation of phonological categories in Sign Language of the Netherlands.* PhD dissertation, Leiden University.

Crasborn O., Bank, R., Stoop, W., Komen, E., Hulsbosch, M., and Even, S., (2018). *Global Signbank source code*. Radboud University, Nijmegen.

Crasborn, O., Bank, R., Zwitserlood, I., van der Kooij, E., Schüller, A., Ormel, E., Nauta, E., van Zuilen, M., van Winsum, F., and Ros, J. (2016). Linking Lexical and Corpus Data for Sign Languages: NGT Signbank and the Corpus NGT. In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining,* pages 41–46, Portorož, Slovenia. European Language Resources Association (ELRA).

Hanke, T., Jahn, E., Wähl, S., Böse, O., and König, L. (2020). SignHunter – A sign elicitation tool suitable for deaf events. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages,* pages 83-88, Marseille, France. European Language Resources Association (ELRA).

Johnston, T. (2012). Lexical frequency in Sign Languages. *Journal of deaf studies and deaf education,* 17(2):163–193.

van der Kooij, E. (2002). *Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity.* PhD dissertation, Leiden University.

Nyst, V., Morgado, M., Mac Hadjah, T., Nyarko, M., Martins, M., van der Mark, L., Burichani, E., Angoua, T., Magassouba, M., Sylla, D., Admasu, K., and Schüller, A. (2022). Object and handling handshapes in 11 sign languages: towards a typology of the iconic use of the hands. *Linguistic Typology,* 26(3):573-604.

Otterspeer, G., Klomp, U., and Roelofsen, F. (Accepted). SignCollect: A 'touchless' pipeline for constructing large-scale sign language repositories. To appear in: *Proceedings of the LREC2024 11th Workshop on the representation and processing of sign languages: Evaluation of sign language resources.*

11. Language Resource References

Corpus Gesproken Nederlands - CGN (Version 2.0.3). 2014. Data set. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-k6

Crasborn, O., Bank, R., Zwitserlood, I., van der Kooij, E., Ormel, E., Ros, J., Schüller, A., de Meijer, A., van Zuilen, M., Nauta, E., van Winsum, F., and Vonk, M. 2020. Nederlandse Gebarentaal (NGT) dataset in Global Signbank. In O. Crasborn et al., (Eds.) *Global Signbank*. Radboud University, Nijmegen. ISLRN: 976-021-358-388-6, DOI: 10.13140/RG.2.1.2839.1446.

Crasborn, O., Zwitserlood, I., and Ros., J. 2008. *Het Corpus NGT. Een digitaal open access corpus van filmpjes en annotaties van de Nederlandse Gebarentaal [The Corpus NGT. A digital open access corpus of movies and annotations of Sign Language of the Netherlands].* Nijmegen: Centre for Language Studies, Radboud University.

Vlaams GebarentaalCentrum. 2024. VGT Signbank [dataset]. https://vlaamsegebarentaal.be/signbank.