Formal and Content-related Checks for Corpus Release

Reiner Konrad, Uta Salden

DOI (latest version): 10.25592/uhhfdm.838

Releases:

Version 3	2021-02-08	First English Version

Abstract

This working paper describes checks that were carried out for quality assurance purposes before the release of the videos and their associated annotations. The checks are summarized and grouped by content.

Content of this working paper

The Public DGS corpus[1] consists of approximately 50 hours of annotated DGS texts. The selection was mainly based on elicitation tasks (discussion formats), which reflect the experiences of D/deaf people and the D/deaf culture. Dialogues and free conversation were favored. In addition to the videos, the German translation of the DGS texts, the lemmatization (by means of glosses), and the annotations of mouthings/mouth gestures are published. Furthermore, DGS texts are annotated with keywords to allow a thematic access. Translations as well as glosses and keywords are translated into English. The data are edited and displayed in different formats on various websites to better meet the expectations of diverse user groups (for more details see Jahn et al. 2018).

Between 2015 and 2017, various checks were established to ensure the quality of the translations and annotations to be published. These checks are grouped by content and described in more detail below.

Checks

Sequences not to be published (blackening)

After completion of the data elicitation, the informants received their video recordings with the request to mark those passages they did not want to see published.[2] These sequences were tagged in the transcript. Furthermore, the translation tags were adjusted accordingly with the aim of losing as little material as possible for publication by blackening, since all sequences that contain a blackening tag or overlap with it are excluded from publication. The completion of this check is documented for the subtasks chosen for release by the metadata entry *TK_07: Blackening*.

[1] See MY DGS (https://meine-dgs.de/ resp. DOI 10.25592/dgs.meinedgs) and MY DGS – annotated (https://ling.meine-dgs.de/ resp. DOI 10.25592/dgs.corpus).

[2] In total, 76 informants requested blackening of certain sequences. Most of them were under one minute long. Most sequences were released for research. Only two informants did not consent to releasing the videos for publication; one informant released her recordings only for research purposes.

Furthermore, annotators and staff members could mark sequences they considered unsuitable for publication. The team discussed and agreed on the blackening of these sequences. In only a few cases did we decide to ask the informant. Sequences were blackened as a precautionary measure if their publication could possibly have negative consequences for an informant or the language community, in cases inappropriate information was given about third parties, or if statements could be interpreted in an ethically questionable manner. For all blackenings whether requested by the informants or the project, it was checked whether the segmentation of the affected translation tags could be adjusted accordingly so that as little material as possible had to be blackened.
Preregistration tags with the value "irrelevant for subcorpus[3]" were excluded from publication. Checking these sequences and the creation of blackening tags is documented with the metadata entry *TK_15: Preregistrations*.

Spelling

To support spell checking when aligning the German translations, an automatic spell checking has been integrated into iLex so that incorrect or questionable German text is marked and can be changed immediately if necessary. Once the alignment is complete – which is noted by the value of the metadata entry *Processing status of translation* in the subtask[4] – the German translations are proofread for correct spelling. This check is documented by an additional metadata entry (*TK_01: Spell checking German*).
There is also an automatic spell checking for the English texts, which marks incorrect or questionable spelling during input. After completion of the English translations, these texts are also proofread. The subtask then receives the metadata entry *TK_02: Spell checking English* with the value "completed".
Mouthing annotations are proofread at regular intervals. For this purpose, the mouthings created at a certain moment in time are extracted and proofread with the help of an SQL query. All errors are corrected directly in iLex using queries.
The German and English glosses are also proofread, and newly added glosses are checked at regular intervals.

Thematic access

When aligning the translations, annotators create a tag with keywords that summarize the essential contents of the DGS text of the respective subtask. This information forms the basis for systematic keywording using a controlled vocabulary (564 keywords in total). These keywords are assigned to 35 superordinate terms, whereby multiple assignments have also been done. For example, the keyword *pension* is assigned to the superordinate terms *work and profession* as well as *finance*. The generic terms contain the 26 subject areas that were used in the elicitation: *Work and profession, authorities, GDR, energy and the environment, food and drinks, family and relatives, feast – celebration – party, finances, feelings, house and apartment, cinema – theater – museum – art, clothing and fashion, communication, personal hygiene – health – illness, nature, relationships – love and sexuality, politics, law and justice, religion, school and education, sport and games, city and country, vacation – leisure time – travel, traffic, weather, the economy*. This list was extended by 9 terms to cover all keywords: *Do-It-Yourself, sign language, D/deaf culture, society, information and entertainment, stage of*

[3] Subcorpus = Public DGS-Corpus
[4] Subtasks are DGS sequences in which both informants work on the assigned task after receiving a set of instructions and, in some cases, one or more stimuli. A task (also referred to as discourse format in the Public DGS Corpus) may contain one or more subtasks. Only subtask sequences were translated, annotated and prepared for publication. For each subtask, the status of translations, transcription, as well as various checks was documented by metadata.

life, pregnancy and birth, terror – war – catastrophes, science and technology. Using these superordinate terms, texts can be selected by topic in the Public DGS corpus (http://.meine-dgs.de). When selecting a DGS text, the keywords are listed under the category *topics*, preceded by the generic terms such as *sign language: finger alphabet, sign language instructor*.
The keywords and generic terms were translated into English and proofread. Once the systematic assignment of the content-related keywords to the general keywords is completed, the metadata entry *TK_05: Keywords* is created, then these keywords are time aligned. If, for example, the informants talk exclusively about one topic for a certain period of time, this section is assigned the corresponding keyword. If the informants switch between different topics, the whole section is tagged with several keywords. To complete this step, there is the metadata entry *TK_06: Keyword time alignment*. An additional control step is to check that keyword tags are placed only within one subtask and cover it completely.

Content review and revision

In the usual workflow, a consistency check between the transcription and the German translation is done before starting with the English translation. The German translation is compared with the lemmatization and the annotated mouthings (metadata entry *TK_04: Matching lemmatization – translation*). This check is crucial for the consistency of the annotations and especially important if many different people are involved.

Anonymization

- During anonymization, references to third parties in videos, lemmatizations, and translations may have to be removed. The following criteria were applied:
  - Names of third parties are anonymized when talking about private subject matters.
  - Names of third parties are not anonymized if they hold an official office (e.g., in a club or a theater) and if they are talked about in a positive way.
  - Names of well-known persons are not anonymized, even if they are talked about in a negative context. A person is considered to be well-known if he or she is known beyond the borders of a federal state.
  - Furthermore, place names are anonymized if a specific person can be inferred directly from the name of the place (e.g., a small village).

This is done by checking the translations that have been marked by Named Entity Recognition (see Bleicken et al. 2016) as containing names, and by manually deciding whether or not to apply the automatically proposed anonymization. If yes, we check whether the lemmatization must also be anonymized. We thereby determine the temporal extent of the sequences in the video that are to be blackened and for which a blackening area is automatically calculated in image coordinates (the same procedure as for sequences marked for blackening due to other reasons – see above).
Furthermore, the use of name signs is checked to see if they have already been anonymized by the procedure described above. If necessary, the anonymization will be done at a later stage. The metadata entry *TK_08: Names* indicates that the manual review of the anonymization proposals is completed.
iLex offers a list of the automatically created blackening sequences for manual checking. These are marked as checked directly in the data set.

Tag boundaries

After being automatically imported into iLex, translations are split across several tags. This often involves correcting their temporal extension/scope, but not at a level of precision where translations could end in the middle of a sign. Further checks therefore consist of testing translations against subtask boundaries and token or mouthing tags in order to adjust the translation tag boundaries appropriately. Furthermore, we check whether translation tags of the

two informants overlap without both of them actually signing at the same time. In such cases, tag boundaries are moved appropriately.
The completion of the corresponding work steps is marked by assigning the metadata entry *TK_16: Tag boundaries: Subtask translation, TK_17: Tag boundaries: Translation token/MB or TK_18: Overlaps translation tags*.

Consistency

Using SQL queries, we detect sequences in which the length ratios of German translation vs. English translation and German translation vs. number of signs are conspicuous. These are checked to see if the translation and the lemmatization is complete.
Further queries are used to check before production whether or not the glossing conventions are followed. (Not all glosses used in the database should appear in published annotations).

Technical quality

Based on the trims (the video sections of the subtasks intended for publication) we check if the material to be published contains image disturbances etc. Such sequences are flagged if cuts etc. are not possible.

References

Jahn, Elena / Konrad, Reiner / Langer, Gabriele / Wagner, Sven / Hanke, Thomas (2018). Publishing DGS Corpus Data: Different Formats for Different Needs. In Bono, Mayumi et al. (eds.): Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 12.5.2018. Paris, Frankreich: European Language Resources Association (ELRA), pp. 83–90. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/18018.html.