The Cologne Corpus of German Sign Language as L2 (C/CSL2): Current Development Stand

Alejandro Oviedo, Thomas Kaul, Leonid Klinner, Reiner Griebel
Department of Deaf Education, University of Cologne
{alejandro.oviedo, thomas.kaul, l.klinner, reiner.griebel}@uni-koeln.de

Abstract

The primary data of our learner corpus consists of 60 hours of videos produced by 350 L2/M2 learners of German Sign Language. The videos are monologues and dialogues that serve as tests at the end of each CEFR level (A1 to C1). An important part of the data includes videos produced by the same students at different times of their acquisition of the DGS over more than two years. Up to now, approximately 3% of the primary data have been transcribed (5,021 tokens; 281 lemmas). The corpus has already offered data for two studies on fluency in the DGS as L2. The biggest challenge we face is that most students refuse to grant permission for free access to their data in the corpus. Looking for solutions, we have made good experiences in obtaining those permits from students that have been directly linked to the research work related to the corpus.

Keywords: learner corpus, German Sign Language, fluency, resource availability.

1. Introduction

The University of Cologne offers courses of German Sign Language (DGS, henceforth) for about 900 L2 bachelor and master students at the Faculty of Human Sciences. Every year around 64 courses are taught from the beginners to the advanced CEFR-levels (Council of Europe, 2001) A1-C1. Upon completion of a level, students perform a series of tests of reception, interaction and production in DGS. Since summer 2015 a part of these tests has been video-recorded and archived with the aim of creating a learner corpus (Granger et al., 2015) of DGS as L2/M2[1] (Universität zu Köln, 2018).

2. Primary Data, Metadata

Currently, we have gathered about 60 hours of video, in more than 1,250 individual files (429 GB). They constitute the primary data of our learner corpus. Among the video files, about two thirds correspond to monologues (average duration 2.5 minutes) induced by an instruction (like "tell me what you did this week"), an illustration or a video. The rest of the files contains dialogues between the informant and a Deaf teacher (tests corresponding to the levels A1 to B2) or between two students (for level C1). Interaction videos have an average duration of 8 minutes. All test videos are archived, both approved and unapproved (students who do not pass an exam receive up to two new opportunities to present it. Every time, the test is video-recorded and archived). The data include a total of 350 informants (312 female and 38 male).
Metadata linked to these videos include age, gender and hearing status of the informants as well as the proof level and semester of data collection.

[1]A small group of students has reported having a significant hearing loss. However, all of them have German as L1 and are therefore also included in the L2/M2 setting.

Part of the data is of a longitudinal nature, dealing with students who have visited our courses and presented the corresponding tests at various CEFR- levels between mid-2015 and the end of 2017. Currently, the most common settings correspond to A1-A2 and B1-B2. A small group of students have recorded videos from B1 to C1.

3. Transcription

So far, only a sample of 23 videos (17 monologues and 6 dialogues) has been transcribed and translated in ELAN (Osborn & Slotjes, 2008). Annotations consist of parent tiers containing German translations as well as ID-Glossing for every sign appearing in video[2]. The ID-Glossings are surrogates of the citation form of a sign and mostly assume the written form of the German word(s) historically related to the basic meaning of that sign. In the corpus, ID-Glossings are contrasted with the WebDGS (Universität zu Köln, 2008), a glossary of around 8,000 entries developed in our University. So far, we have registered 5,021 tokens and elaborated a list of 281 lemmas.
Special attention has been paid to segmentation. Following Hanke et al. (2012), transitions between two signs are not included as part of any of them. These moments of non-significant activity remain empty.
In addition to the annotation lines mentioned above, our transcription template includes five more series of tiers:

- **Deviations from the lexical norm:** These consist of variations observed in any manual parameter with respect to the native model (defined by the signs of our glossary WebDGS). There is one tier for each manual parameter (i.e. handshape, orientation,

[2]A second line, ID-Glossing2, is available for cases in which each hand simultaneously articulates different signs.

location and movement). The activity of non-manual articulators can be transcribed using a tier for every articulator (e.g. head, eyebrows, nose, eyes, etc.). Most tiers mentioned above are attached to controlled vocabularies.

- **Elements of the utterance:** This annotation line include a controlled vocabulary comprising subject, object, nucleus of predication and predicate complements.
- **Fluencemes** (Götz, 2013), i.e., phenomena which interrupt the flow of lexical information. This group of lines includes pauses (empty and filled), repetitions and false starts.
- **Type of discourse:** It includes narration, explanation, description and argumentation (Grimes, 1975).
- **Paragraph limits**, in the sense of the border between the end of a thematic unit and the beginning of the next. (Longacre, 1979).

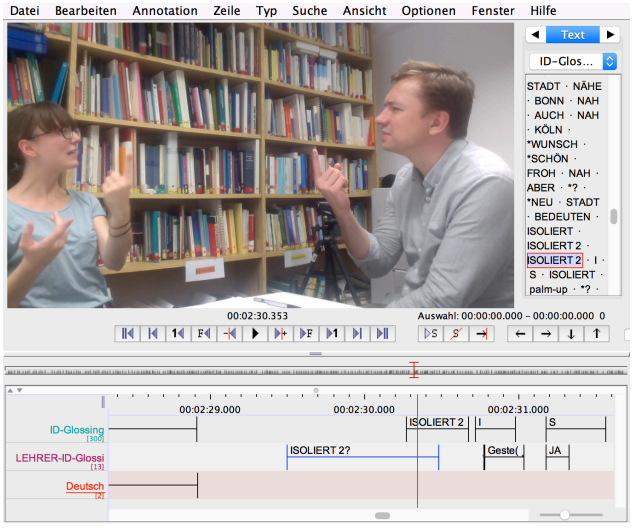

Figure 1. Screenshot of a C/CSL2 file in ELAN

4. Research Carried on C/CSL2 Data

4.1. The Problem of Assessing Fluency in DGS-Courses

Although little progress has been made in the elaboration of the corpus, available data have already served as the basis for an investigation on fluency, i.e., the capacity to produce complex chains of significant units with few interruptions (Fillmore, 1979). Fluency is seen as an inherent property of native discourse, and it offers a criterion for ranking learners at advanced L2 levels (Chambers, 1997). Since the introduction of the CEFR standards, the concept of fluency is a mandatory part of the curriculum. Regarding DGS, there is no research that allows us to know what makes a signed discourse fluent or not-fluent, what hampers the teaching and assessment of fluency in DGS courses.

4.2. First Survey

Looking for a solution to that problem, we have used our first corpus data (CEFR-level B1), as a basis for an experiment related with perceptions of fluency in L2 discourse (Kaul et al., 2017). The two selected videos shared the same stimulus and have a similar duration (3 min). But scores obtained differed: while the student of video 1 received the highest grade, the student of video 2 obtained the minimum one. The teachers who evaluated both tests, themselves part of our research team, thought that while video 1 showed a very fluent production, video 2 was less fluent. This difference had influenced the assigned grades.

Following a first analysis, the most relevant difference between both videos were the amount, form and distribution of pauses:

- Video 1 contained less pauses than video 2, and they were shorter (Video 1: Ø 743 ms vs. video 2: Ø 1326 ms).
- All pauses of video 1 were filled (with meaningless gestures, false starts or lengthened transitions between signs[3]), while most of the pauses in video 2 were empty (moments of no activity).
- Pauses in video 1 did not apparently have a fixed context of occurrence. Most pauses of video 2 appeared between sentences.

Additionally, both videos contained a similar number of repetitions and self-repairs. Regarding grammar, video 1 was free of errors, while video 2 contained some errors related to word order and non-manuals.

Based on these analysis, we developed a questionnaire in which we asked Deaf DGS users to rate both videos by assigning scores to seven items: fluency, pauses, repetitions, self-repairs, meaningless gestures, grammar correctness and intelligibility. Each aspect was rated in a 5-level-scale. 31 people answered the questionnaire after seeing each video twice.

Results: video 1 was rated as being more fluent than video 2. But both videos were rated as little fluent (between 3 and 4 at the given scale) and formal differences between both videos were perceived as quite smaller than expected. Correlations to fluency were found in both videos as highly significant for intelligibility and pauses and significant for repetitions. The regression analysis revealed that self-repairs (video 1) and intelligibility (video 2) were predictors of fluency. Summarizing, the indicators included in the questionnaire allowed to predict perceptions of fluency in video 1, but perceptions of fluency in video 2 seem to be partially determined by aspects not taken into consideration. For instance: Self-repairs played a role in video 1, but not in video 2. This could not be explained by our linguistic analysis.

[3]We did not have any reference to decide that the duration of a transition was normal or markedly longer than normal. Therefore, the decision about that was always taken in an impressionistic way by a Deaf researcher. Transitions marked as lengthened had an average duration of 900 ms.

4.3. Second Survey

The first study suggested that we were not considering the necessary indicators to determine why native users rate a non-native discourse as fluent or non-fluent. To advance in the solution of this problem, we carried on a study on possible indicators of fluency in native DGS discourse (Oviedo et al. to appear). We have recorded and transcribed eight videos produced by five Deaf native signers. Afterwards, we have asked a second group of users to indicate at what moments of the videos some type of interruption or disturbance in the signing flow was perceived. The indicated moments were marked and later used as orientation for the linguistic analysis.

Most marks corresponded to pauses, but the shape and distribution of them differed from the pauses we previously found. In native signing, the number of pauses was linked to the type of discourse. In argumentative or explicative fragments there were at least twice as many pauses as in narratives. Additionally, native signers gave empty pauses a regular distribution, namely in paragraph boundaries (Longacre, 1979). Filled pauses, however, did not have a predictable distribution. They seemed to serve as planning indicators. Finally, the pauses in L1 were, on average (Ø 550 ms), shorter than in our L2 sample.

Empty pauses of average duration were considered by the evaluators as fluent. On the contrary, almost all filled pauses were considered non-fluent. An exception was made when they marked paragraph boundaries and were accompanied by a clear interruption of the gaze contact (the gaze directed to a high point of the room).

Following the conclusions of the studies described above, we decided to add a new series of lines to our transcription template, in order to generate data regarding fluency in DGS as L2. These lines correspond to *fluencemes*, type of discourse and paragraph limits. When such data is available, we are planning to replicate the study by Kaul et al. (2017).

5. Challenges

5.1. No funding available

The work of organizing and transcribing the data has been done so far by the teachers and researchers of our work team in their free time. We do not yet have a specific funding to cover the costs of building up the corpus, so the project is progressing very slowly so far. We are working on applications for external funding.

5.2. Restricted access

The permissions to access and use the corpus remain, however, our biggest limitation. Before performing their tests, students are asked to sign a written form allowing the use of the videos as research data and / or as material for public presentations. So far, most students have granted authorization for restricted use of the videos (exclusively in our research group), but just a small group of 17 students have also authorized the full use of their data for both research and dissemination.

In order to expand access to the data, we are trying ways to convince our students to change their opinion about the necessary permits. Linking students actively in transcription and research work has proved to be promissory. We will summarize two experiences we have had in this sense:

Shortly after finishing the first study (Kaul et al., 2017), we asked a group of advanced DGS students to make some videos narrating in sign language the same story previously seen in a computer-animated film. We explained them that we wanted to analyze the videos with the collaboration of some Deaf teachers who did not belong to the university staff. The teachers, we added, wanted to improve the assessment of their own courses. All students agreed to permit this use of their videos. Subsequently, a group of external Deaf teachers held a workshop to seek criteria for evaluation and assessment of fluency in their DGS courses. Afterwards, our team discussed with the students the conclusions reached in the workshop and suggested strategies to increase their own signed fluency. All students expressed their understanding about providing data for studies that would improve the quality of teaching and authorized us to use their C/CSL2 data in an unrestricted manner.

Our second experience occurred within the framework of a linguistics seminar for M.A. students. We have presented them our corpus project and introduced them into the basics of sign language transcription. As final work of the course students could choose between a written essay or the transcription of a series of self-produced videos. The majority of students chose the second option. These transcriptions are currently in progress and should be incorporated into the corpus at the beginning of 2018. Most of this group of students have agreed that all their videos can be freely accessed. A repetition of this experience is planned for the next summer semester.

6. References

Chambers, F. (1997). What do we mean by fluency? In *System* 25(4), pp. 535–544.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Press Syndicate of the University of Cambridge.

Crasborn, O., Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In Crasborn, O, Hanke, Th., Efthimiou, E., Zwitserlood, I., Thoutenhoofd, E. (Eds.). In *Sixth International Conference on Language Resources and Evaluation*, Paris: ELDA, pp. 39-43.

Fillmore, Ch. J. (1979). On fluency. In Fillmore, Ch.J., Kempler, D., Wang, W.Y. (Eds.). *Individual differences in language ability and language behavior*. New York: Academic Press, pp. 85–102.

Götz, S. (2013). *Fluency in Native and Nonnative English Speech*. Amsterdam: John Benjamins.

Granger, S., Guilquin, G., Meunier, F. (Eds.). (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

Grimes, J. E. (1975). *The thread of discourse*. The Hague: Mouton.

Hanke, T., Matthes, S., Regen, A., Worseck, S. (2012). Where does a sign start and end? Segmentation of

continuous signing, In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon Language Resources and Evaluation Conference (LREC)* (Istanbul), pp. 69–74.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. In *International Journal of Corpus Linguistics*, 15(1), pp. 106-131.

Kaul, Th., Oviedo, A., Griebel, R., Klinner, L., Prüfer, T., Krumpen, M. (2016). C/CSL2, The Cologne Corpus of Sign Language as a Second Language. Poster presented at TaLC 12, University of Giessen, 2/07/2016.

Kaul, Th., Oviedo, A., Griebel, R., Klinner, L., Glatz, A. (2017). How do Deaf native signers perceive fluency in adult L2/M2 learners of German Sign Language? Paper presented at the International Conference Fluency and Disfluency Across Languages and Language Varieties. Université Catholique de Louvain, 15/02/2017.

Longacre, R. E. (1979). The Paragraph as a Grammatical Unit. In Givón, T. (Ed.). *Syntax and Semantics*. New York: Academic Press, pp. 115–134.

Oviedo, A., Kaul, Th., Urbann, K., Griebel, R., Klinner, L. (to appear). Eine Annäherung zu den Pausen als Flüssigkeitsfaktoren in Deutscher Gebärdensprache als L1. In *Das Zeichen*, 32(108).

7. Language Resource References

Universität zu Köln (2018). C/CSL2. The Cologne Corpus of German Sign Language as L2/M2. Cologne: University of Cologne.

Universität zu Köln (2008). WebDGS: E-Learning für Gebärdensprache. Accessed at: http://webdgs.uni-koeln.de/ [20/02/2018]