D3.6 Project Language Resources Public Release

Partner Responsible	AthenaRC
Other Contributors	UTH-ECE
Document Reference	**D3.6**
Dissemination Level	Public
Version	1.0 (Final)
Type	**Report / Other**
Due Date	July 2023 (**M42**)
Date of Preparation	July 2023

Contract No.: HFRI-FM17-2456

Editor
Eleni Efthimiou (**AthenaRC**)

Contributors
AthenaRC: Eleni Efthimiou, Stavroula-Evita Fotinea
UTH-SED: Galini Sapountzaki
UTH-ECE: Katerina Papadimitriou

SL-ReDu Principal Investigator:
Assoc. Prof. Gerasimos Potamianos
University of Thessaly, Electrical and Computer Engineering Department (**UTH-ECE**)
Volos, Greece 38334
email: gpotamianos@uth.gr (gpotam@ieee.org)

Supporting Documentation of SL-ReDu Data Release

The SL-ReDu project aims to advance the state-of-the-art in the automatic recognition of Greek Sign Language (GSL) from videos, focusing on the novel education use-case of standardized teaching of GSL as a second language. A crucial part of the project constitutes the development of the sign language recognition (SLR) module for GSL, in order to enable the SL-ReDu system to provide binary feedback on produced signs by GSL learners. Necessary for such development is the availability of data resources, matching the education use-case of SL-ReDu.

For this purpose, we have collected a large corpus of GSL data, allowing SLR system development for isolated signing of a large-vocabulary set of lexemes, continuous signing of phrases, as well as continuous fingerspelling of letter sequences. We have reported on that data collection in recent Deliverable D3.4 ("Second Version of Data Resources" – M28), detailing both our recording methodology and the obtained sets. As envisaged in the SL-ReDu Document of Work, we now proceed with the public release of this dataset, in order to foster progress in the research community in the field of SLR both in Greece and abroad. In particular, we release the data that have been collected at the AthenaRC premises (the so-called "studio" dataset), due to the balanced nature of the data across signers and content. We expect that this dataset will be used as standalone in future SLR evaluations of GSL by academic and research / industrial labs from both Greece and abroad, thus providing a golden standard for SLR in GSL. In addition, it could also serve as material to be pooled together with that of other sign languages, aiming to improve SLR model pretraining or even allow multi-lingual SLR development. Thus, this data release is expected to contribute to both project dissemination and exploitation, providing extra visibility and future funding opportunities to the SL-ReDu project partners.

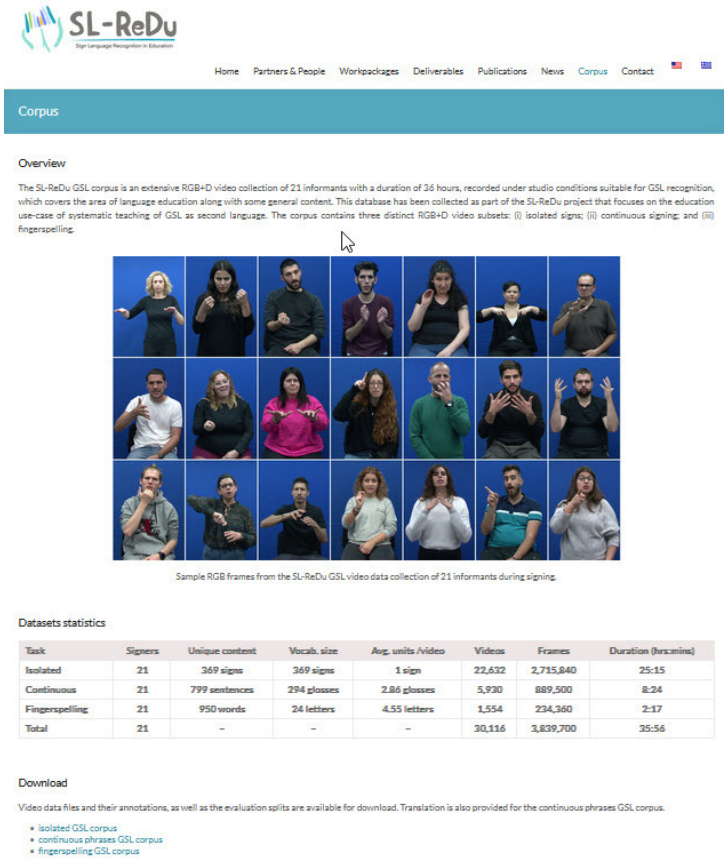

Figure 1: Part of the SL-ReDu website that is being used for the public release of the project's corpus.

The released SL-ReDu dataset can be downloaded from the project's website. For this purpose, a dedicated tab has been created (available in both English and Greek) within the SL-ReDu website, namely:

https://sl-redu.e-ce.uth.gr/corpus (English version, see also Figure 1);

https://sl-redu.e-ce.uth.gr/el/corpus (Greek version).

The website contains a *brief description* of the dataset, including a reference to the *paper* where it was announced, so that proper credit be given to the SL-ReDu project by any database users. The paper appeared at the Eighth International Workshop on Sign Language Translation and Avatar Technology (SLTAT), held as a Satellite Workshop to the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) in Rhodes Island, Greece on June 10th, 2023, and it is indexed in the IEEE Xplore Digital Library (DOI: 10.1109/ICASSPW59220.2023.10193306).

Most importantly, the website contains *download links* for the released data, which are provided separately per SLR task, namely for isolated signs, continuous phrases of glosses, and continuously fingerspelled sequences.

In addition, the website provides *recommended data splits* for training, validating, and testing the developed SLR models, thus fostering comparable and reproducible research on the topic. Specifically, the test set is kept identical under three different experimental frameworks, thus also allowing a fair comparison between them. Namely:

- A *multi-signer* setting, where data from all signers are split between training, validation, and testing (a single fold is used).
- A *signer-independent* setting, where a 7-fold cross-validation framework is adopted. Each fold contains training and validation data from 18 signers, with testing performed on the remaining 3 (and the process repeating over all 7 folds to cover all signers).
- A *signer-adapted* setting, where a similar framework to the signer-independent scheme is used, but an additional set of adaptation data for the 3 test signers is introduced for each fold. This allows for adaptation experiments to be carried out.

The above are available for each of the three SLR tasks. An additional *multi-signer* setting is also proposed that incorporates a more "traditional" data split ratio among the training, validation, and test sets. Additional data splits may also be introduced in the future, following possible suggestions / requests by database users. For example, a phrase-independent setting may be considered for the task of continuous GSL signing of phrases, as well as a framework where this specific task is assisted by incorporating the isolated GSL training data.

Closing this brief report on the data release (note that this deliverable is primarily of type "Other"), we provide in Table 1 an overview of these resources (per SLR task). It should be noted that this database constitutes the largest multi-signer GSL corpus that is suitable for SLR, exceeding the one released by the Information Technologies Institute (ITI) by 3 times in terms of number of signers, as well as in terms of corpus size / duration. Example frames of the 21 signers of our released dataset are also shown in Figure 2.

SLR task	Signers	Unique content	Vocab. size	Videos	Duration
Isolated	21	369 signs	369 signs	22,632	25:15
Continuous	21	799 phrases	294 glosses	5,930	8:24
Fingerspelling	21	950 words	24 letters	1,554	2:17
Total	21	—	—	30,116	35:56

Table 1: Overview statistics of the released dataset for the three SLR tasks. Number of signers, unique content size, vocabulary size, number of videos, and duration (in hr:min format) are shown.

Figure 2: Example frames of the released GSL data, depicting the 21 different signers of the corpus.