Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 45–52
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

LSE_UVIGO: A Multi-source Database for Spanish Sign Language Recognition

Laura Docío-Fernández[1], José Luis Alba-Castro[1], Soledad Torres-Guijarro[1], Eduardo Rodríguez-Banga[1], Manuel Rey-Area[1], Ania Pérez-Pérez[1], Sonia Rico-Alonso[2], Carmen García-Mateo[1]

[1]atlanTTic Research Center for Telecommunication Technologies, [2]Ramón Piñeiro Centre for Research in Humanities
[1]Escola de Enxeñaría de Telecomunicación. Universidade de Vigo, [2]Xunta de Galicia
ldocio@gts.uvigo.es, jalba@uvigo.es, soledadtorres@uvigo.es, erbanga@uvigo.es, mreyarea@gts.uvigo.es, aniaperezperez@uvigo.es, sricalo@cirp.es, carmen.garcia@uvigo.es

Abstract

This paper presents LSE_UVIGO, a multi-source database designed to foster research on Sign Language Recognition. It is being recorded and compiled for Spanish Sign Language (LSE acronym in Spanish) and contains also spoken Galician language, so it is very well fitted to research on these languages, but also quite useful for fundamental research in any other sign language. LSE_UVIGO is composed of two datasets: LSE_Lex40_UVIGO, a multi-sensor and multi-signer dataset acquired from scratch, designed as an incremental dataset, both in complexity of the visual content and in the variety of signers. It contains static and co-articulated sign recordings, fingerspelled and gloss-based isolated words, and sentences. Its acquisition is done in a controlled lab environment in order to obtain good quality videos with sharp video frames and RGB and depth information, making them suitable to try different approaches to automatic recognition. The second subset, LSE_TVGWeather_UVIGO is being populated from the regional television weather forecasts interpreted to LSE, as a faster way to acquire high quality, continuous LSE recordings with a domain-restricted vocabulary and with a correspondence to spoken sentences.

Keywords: Spanish Sign Language (LSE), sign language recognition (SLR), LSE_UVIGO, LSE_Lex40_UVIGO, LSE_TVGWeather_UVIGO, Microsoft Kinect v2

1. Introduction

Automatic speech recognition is one of the core technologies that facilitate human-computer interaction. It can be considered a mature and viable technology and is widely used in numerous applications such as dictation tools, virtual assistants and voice controlled systems. However automatic sign language recognition (SLR) is far less mature.

Some reasons for this have to do with the multimodal nature of sign languages, where not just hands, but also face, head, and torso movements convey crucial information. Others are related with the high number of structural primitives used to build the messages. For example, Spanish spoken language has between 22 and 24 phonemes, but Spanish Sign Language (LSE) has 42 hand configurations, 24 orientations (6 of fingers times 4 of palm), 44 contact places (16 in the head, 12 in the torso, 6 in the dominated hand/arm and 10 in space), 4 directional movements and 10 forms of movement (according to (Herrero Blanco, 2009), although there is no unanimity in this classification, see for example CNSE (2008)).

The study of the state of the art suggests that machine learning applied to SLR will be sooner or later able to overcome these difficulties as long as there are adequate sign language databases. Adequate means, in this context, acquired with good quality, carefully annotated, and populated with sufficient variability of signers and visual contexts to ensure that the recognition task is robust to changes in these factors.

Unfortunately, only a few sign languages offer linguistic databases with sufficient material to allow the training of complex recognizers (Tilves-Santiago et al., 2018; Ebling et al., 2018), and LSE is not one of them. There have been some efforts to collect the variety of LSE signs through different recording technologies and with different purposes. The video dataset from Gutierrez-Sigut et al. (2016) contains 2400 signs and 2700 no-signs, grammatically annotated, from the most recent standardized LSE dictionary (CNSE, 2008). Even though this controlled dataset is very useful to study the variability of Spanish signs, the poor variability of signers (a man and a woman signing half dictionary each), the absence of inter-sign co-articulation and the small resolution of the body image, precludes it from its use for training machine learning models for signer-independent continuous Spanish SLR.

The Centre for Linguistic Normalization of the Spanish Sign Language (CNLSE, acronym in Spanish) has been developing a corpus for years in collaboration with numerous associations and research centres in the state. It is composed of recordings of spontaneous discourse, very useful to collect the geographical, generational, gender and type of sign variation of the LSE. However it is not appropriate for SLR training in a first phase, which would require a database with a high number of repetitions per sign and, probably, the temporal segmentation of the signs collected in the recordings.

A completely different LSE dataset (Martinez-Hinarejos, 2017) was acquired with the Leap Motion infrared sensor that captures, at a short distance, the position of the hands and fingers, similarly to a data glove but touchless. This publicly available dataset is composed of a main corpus of 91 signs repeated 40 times by 4 people (3640 acquisitions) and a 274 sentences sub-corpus formed from 68 words of

the main corpus. The technology of Leap Motion limits its use to constrained movements (close to the device and without self-occlusions) and prevents capturing arms, body motion and facial expressions. Therefore, its usefulness to SLR would probably be limited to fingerspelling.

From this review we conclude the need to build a new multi-source database, which we will call LSE_UVIGO, specifically designed to support our ongoing research on SLR, and that of others. Our team is made up of two research groups of the University of Vigo: the Multimedia Technology Group (GTM) and the Grammar, Discourse and Society group (GRADES). GTM has accredited expertise on facial and gesture analysis, and speech and speaker recognition, and GRADES has a longstanding expertise on LSE and interaction with deaf people. With the development of LSE_UVIGO we intend to support fundamental and applied research on LSE and sign languages in general. In particular, the purpose of the database is supporting the following or related lines:

- Analyse the influence of the quality of video footage on the processing of the video stream for segmentation, tracking and recognition of signs.
- Quantify the advantages of including depth information.
- Segment and track upper-body parts in 2D/3D, and quantify the benefits of an accurate segmentation on SLR.
- Develop tools to align continuous speech and LSE.
- Develop signer-independent sign to text/speech translation, both word-based and sentence-based, including fingerspelling.
- Analyse the influence of face expression and body movements on decoding sign language sentences.
- Measure the robustness of sign language processing modules against changes in the scenery.

2. LSE_UVIGO Database

Initially, LSE_UVIGO consist of two different datasets that complement each other to the above purposes: the LSE_Lex40_UVIGO and the LSE_TVGWeather_UVIGO. The first one is intended to support research on LSE through high quality RGB+D video sequences with high shutter speed shooting. The second one is composed of broadcast footage of the weather forecast section in Galician Television (TVG) news programs. Following sections explain with more detail both datasets.

2.1 LSE_Lex40_UVIGO Dataset

This subset is a multi-sensor and multi-signer dataset acquired from scratch. It is thought as an incremental dataset, both in complexity of the visual content and in the variety of signers, most of them deaf.

LSE_Lex40_UVIGO is intended to cover most of the necessities of the research community working in SLR: static and co-articulated sign recordings, both fingerspelled and gloss-based isolated words, and sentences. The recording is done in a controlled lab environment in order to obtain good quality videos with sharp video frames and RGB and depth information, making them suitable to try different approaches to automatic recognition. The RGB and depth information are co-registered in time which allows researchers to work not only on recognition, but also on tracking and segmentation.

In its present form, the contents of LSE_Lex40_UVIGO are organised in three sections:

- The LSE alphabet, composed of 30 fingerspelled letters.
- 40 isolated signs, which can be static or dynamic, in which one or both hands intervene, with symmetric or asymmetric movement, and with different configurations, orientations and spatial-temporal location. They were selected according to linguistic-morphological criteria so as to reflect different modes of articulation that may affect the complexity of SLR (Torres-Guijarro, 2020).
- 40 short sentences related to courtesy and interaction. The sentences were chosen based on vocabulary that is traditionally included in introductory LSE courses. Each sentence ranges from one to five signs in length.

In order to facilitate the labelling process, the signs are performed in a standardized way, trying to avoid dialect variations of glosses as much as possible.

Recording Software and Setup

The UVigoLSE_Lex40 dataset is being recorded with two visual devices: a Microsoft Kinect v2, which captures both RGB video (resolution 1920x1080 pixels @30 FPS) and depth maps (resolution 512x424 pixels @ 30 FPS), and a Nikon D3400 camera which captures high quality RGB video signals (resolution 1920x1080 @ 50 FPS). The shutter speed of the Nikon camera is set to 1/240 sec. to freeze the movement of the signing sequence even for quite fast movements of the signer. Both devices are fitted on a rigid mount on a steady tripod. The mount is placed in front of the signer facing the signing space, and the recording location has been carefully designed to facilitate the framing, focusing, lighting and setting the distance to the signer. Figure 1 shows the recording setting.

Figure 1: Set up of the dataset acquisition. Kinect and Nikon devices are rigidly mounted on a tripod at a fixed distance to the signer, that is uniformly illuminated over a somehow uniform background (location settings vary). No restrictions on clothing are imposed.

To facilitate the introduction of the metadata of the recording session (date and place, operator, recording devices) and the signer self-reported information both written and signed (name, sex, year of birth, school, dominant hand, place of residence, hearing/deaf, at what age she/he started learning LSE, and at what age she/he went deaf), an acquisition platform has been programmed in MatLab®, which also allows simultaneously recording from the two devices.

2.2 LSE_TVGWeather_UVIGO Dataset

Nowadays it is nearly impossible to acquire a large-scale, high-quality LSE dataset which captures all the difficulties of the SLR task. The main reason for this is the high cost of designing, recording and annotating a dataset with a large vocabulary and a sufficient number of signers. To solve this issue, public video sources available in LSE can be used, such as websites dedicated to teaching sign language, and TV programs interpreted in LSE.

Monday through Friday, the midday newscast of the regional television network (TVG) is interpreted in LSE. Both the original broadcast in Galician language and the LSE version, are available on the TVG website. The news domain is too ample for considering the acquisition of a database for continuous SLR. Therefore, inspired by other authors' work (Koller et al., 2015), we decided to focus on a restricted domain: weather forecasts.

LSE_TVGWeather_UVIGO dataset is being populated with weather forecasts from the TVG news on workdays, with a typical duration of 1-2 minutes. The main characteristics of the video codec are: H.264, resolution 1280x720, 50 frames per second. As illustrated in Figure 2, the sign language interpreter occupies about 20% of the image (around 400*470 pixels), a screen portion substantially larger than that used in other TV channels. Every video is automatically annotated at the word level by means of our Galician automatic speech recognizer (ASR) system. This transcription is then manually reviewed at a higher "segment" level (quite similar to a breath-group level) using ELAN, leaving the weather forecast ready for further annotation (as illustrated in Fig. 6; detailed information about annotation is given in Section 4).

3. Video and Depth Signal Post-processing

As explained in Section 2.1, LSE_Lex40_UVIGO recordings are acquired simultaneously with a Nikon camera and a Kinect. The Nikon provides high quality RGB, and the Kinect provides complementary depth information, quite useful for segmenting regions of interest in RGB images, such as hands, arms and face. In the following sections, details are given on the time-alignment of depth and video signals, and on the segmentation process itself. Segmentation will also be applied to LSE_TVGWeather_UVIGO.

Figure 2: weather forecast in the regional TV network (TVG), interpreted to LSE.

3.1 Time-alignment and Transferring of Depth to the RGB Streams

In order to complement Nikon's images with depth information from Kinect, a two-step co-registering and alignment process is needed. This process is outlined in Figure 3.

The first step entails co-registering color and depth from the Kinect sensors. Although RGB and depth information are gathered by the Kinect simultaneously, these two signals are not synchronous because their sensors are initially triggered at different moments, the periodic acquisition has some jitter, and a frame from any of the sensors is occasionally lost. In order to perform a temporal alignment over the whole sequences we have used the skeleton landmarks provided by Kinect software development kit. After calculating the optimal projective transformation between pairs of temporally-aligned frames, we apply a Dynamic Time Warping (DTW) algorithm using the minimum squared error (MSE) of the location of skeleton landmarks among the co-registered pairs as the distance measurement. This last step is avoided if absolute timestamps are preserved during the recording of RGB and depth information[1].

The second step consists in co-registering Kinect RGB and Nikon RGB. In this case we cannot use the Kinect's skeleton landmarks, so we have resorted to OpenPose software to co-locate a set of landmarks in temporally similar frames and calculate a geometrical transformation to co-register the short focal length Kinect RGB+D maps onto the larger focal length Nikon RGB image. Given that the triggering (start, stop and period) and acquisition period are also different, we need to temporarily align the sequences using again a DTW algorithm. Similarly to the previous step, the distance measure between frames is the MSE of the location of OpenPose landmarks.

[1] The first recordings of isolated signs in LSE_Lex40_UVIGO were acquired without the absolute timestamp.

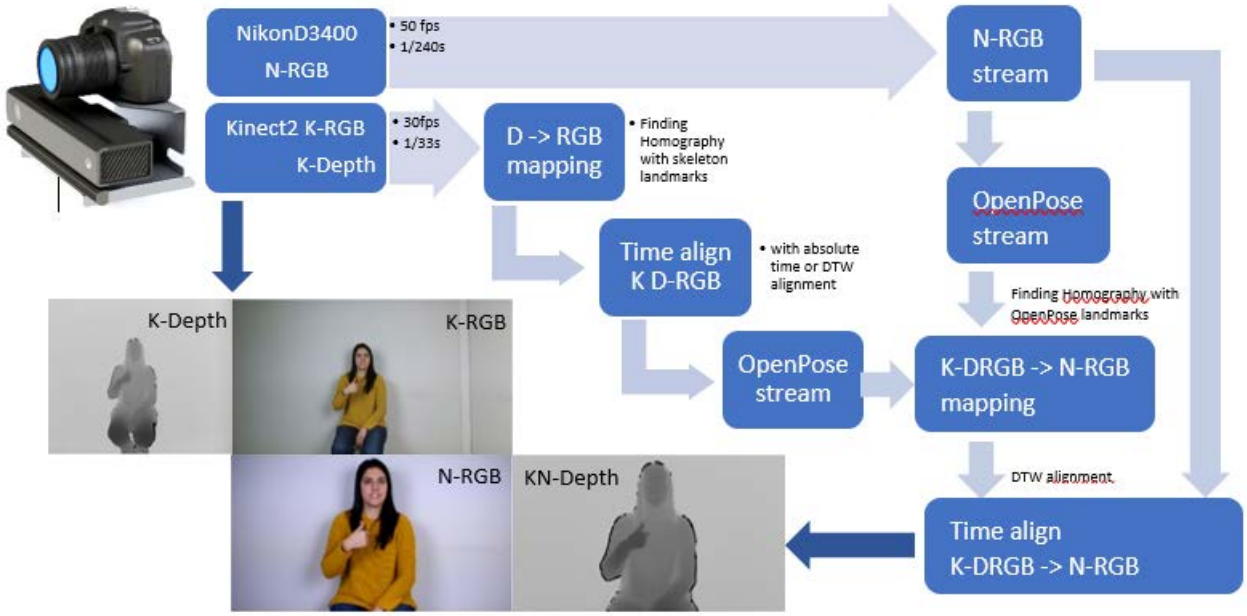

Figure 3: Flow diagram of the post-processing to align all the streams and transfer Kinect depth information to the Nikon acquired RGB stream.

3.1 Hands and Face Segmentation

A recurrent issue in object and instance recognition is the amount of context needed to identify the object or its specific configuration. SLR does not get rid of this problem, and despite some efforts on determining whether perfectly segmented hands and face work better in SLR than the complete image containing the full body context (Huang, 2015; Camgoz, 2017; Koller, 2019), more studies are needed in this field.

Most sign language interpreters use dark clothes to facilitate the contrast of hands over the body, so it seems that an automatic recognition system could benefit from a proper segmentation of the hands. But the relative location of the hands and arms with respect to the body and face is also crucial, so keeping the visual context could help the system. Current techniques using deep neural networks fed by holistic visual appearance seem to digest unsegmented objects properly, but only a large variety of examples (Li, 2019) will help the network to simulate the visual attention made by the brain, and thus to get rid of the non-discriminative surrounding information. Unfortunately, Spanish sign language datasets are still too small to benefit from this approach.

To support research on the influence of segmentation, LSE_UVIGO will also provide a segmentation map, so researchers can directly try their algorithms with or without context information. Figure 4 shows a simplified flow diagram of the image processing, which makes use of colour, OpenPose landmarks of the RGB stream (Cao et al., 2018; Simon et al., 2017), and depth when available. A similar segmentation approach but using just the Kinect sensors was proposed in (Tang, 2014). Image at left shows the result of using a generic skin map. It is clear that colour information alone was not able to eliminate the sweater and the neck information. Picture at right shows the original image filtered by a probability map that takes into account a user-specific skin-map, the depth co-registered image and the distance to the OpenPose landmarks at hands and face. So, instead of providing a final binary mask, we store in the database a probability map with real values between 0 and 1, so researches can choose to threshold at different levels to include more or less body information, or even just use the map as a filter that preserves the information of hands and face and attenuates the rest in a 'saliency-map' way. It is important to highlight that the Kinect RGB stream, as most of the SL videos in other datasets, contains blurred hands when movement is relatively fast because of the shutter speed of 1/33 secs. For this reason we have resorted to the Nikon's streams with shutter speed of 1/240 secs, which allows to freeze most of the very fast hand movements and allows a more accurate segmentation.

4. Database Annotation

We are enriching the database with detailed manual and semi-automatic annotations using the ELAN software package (Brugman & Russell, 2004). The annotation is divided into several parts, similarly to the CORILSE corpus annotation (Cabeza-Pereiro et al., 2016):

4.1 Annotation of Manual Components (MC)

The annotation of MC includes the start and end points of every sign, transition movements between signs and discourse pauses, and the gloss ID with respect to an

annexed lexical database. This annotation phase involves the tiers *MD_Glosa* (Gloss for right hand) and *MI_Glosa* (Gloss for left hand). It is important to highlight that some non-lexical units are also annotated in this phase, the most important one being the buoy (B) hand indicating that one or both hands are paused in a specific position and configuration after (or even before) its participation in a

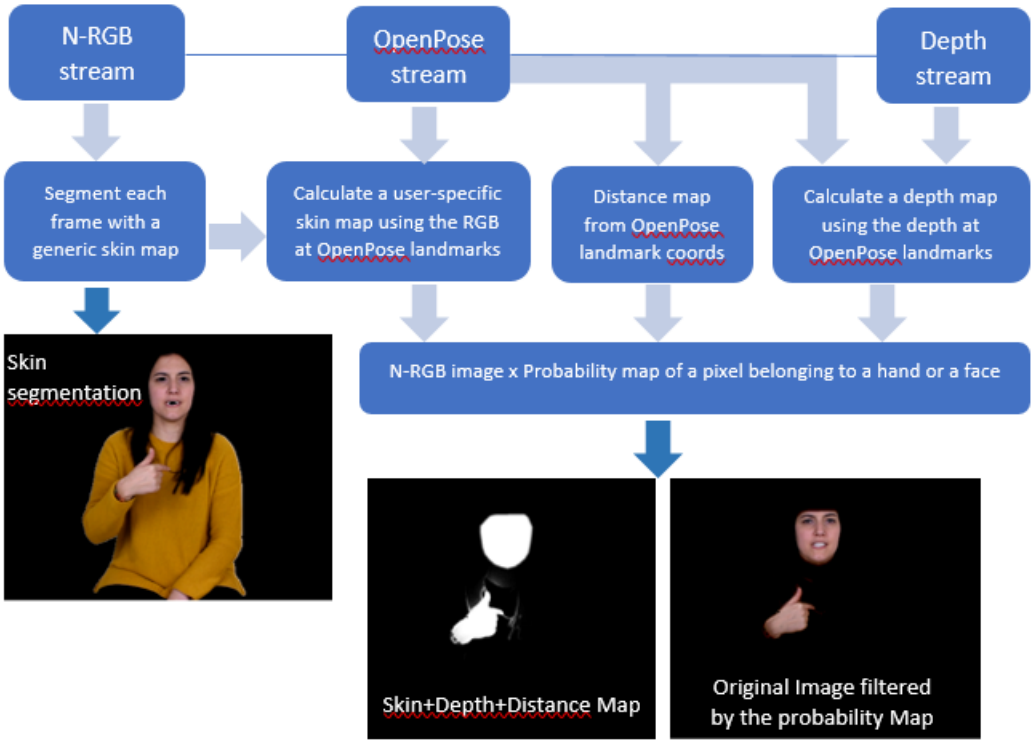

Figure 4: Simplified flow diagram to segment hands and face from the video sequences.

sign. Other non-lexical or semi-lexical units are also annotated like gestures (G) and indexes (INDX) respectively.

4.2 Annotation of Non-Manual Components (NMC)

The annotation of NMC is still under development. The number of components defined by Cabeza-Pereiro et al. (2016) is much larger than needed for the purpose of this database. We will annotate the NMC useful for disambiguation of a sign (like the eyebrows in SWEET and PAIN), those that modulate the discourse (like movement of eyebrows and mouth in a question clause) and those that are modifiers of the sign (like shape of mouth and cheeks when indicating a big amount of people, work, money, etc.). Another type of NMC to annotate is blinking, that helps to determine the end of a clause in LSE. Action Units provided by OpenPose are being used for detecting the NMC in the video stream and will be imported as NMC tiers. Manual revision from an expert LSE signer will be needed to eliminate false positives and add false negatives in these tiers.

4.3 Annotation of Other Linguistic Information

The literal translation to Spanish (tier *Trad*) is annotated, and also a segmentation of each predicative expression (a 'clause-like unit' or CLU, to borrow a term used by Johnston (2013) and Hodge (2013)). Each CLU will have a different reference in the tier *Ref* (Reference) and will facilitate the construction of LSE/Spanish pairs for training and testing end to end translation systems. If the dataset only contains text and signs, as in LSE_Lex40_UVIGO, the *Ref* is aligned with the set of signs that form the CLU (see Fig.5). If the dataset contains also a speech stream simultaneously translated to LSE, as in LSE_TVGWeather_UVIGO, there are two *Ref* tiers; *Ref_LO* for speech CLUs and *Ref_SL* for LSE CLUs. Given that the LSE signer translates from a speech stream in real-time (Galician language in this case), there's a variable amount of time shift between them, so detailed annotation of the spoken-signed CLU pairs is a great help for developing translation systems. Two more tiers are annotated in the LSE_TVGWeather_UVIGO: *Word* and *Segment*. The first one corresponds to the automatic speech recognizer (ASR) output, with timestamps between words, while the second one corresponds to the manual review of the sentences extracted automatically from the sequence of words from the *Word* tier.

Figure 5 shows a screenshot of the annotation of LSE_Lex40_UVIGO dataset and Figure 6 shows the annotation of the LSE_TVGWeather_UVIGO.

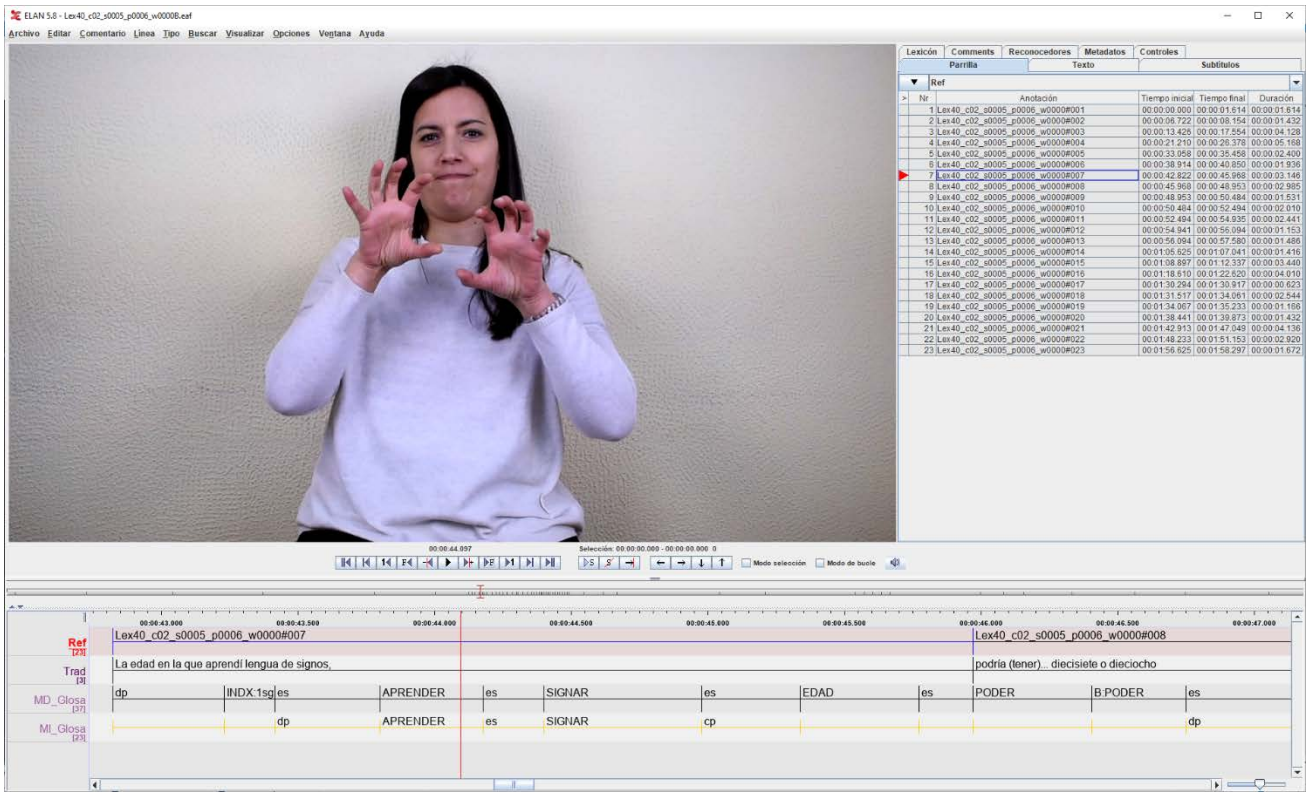

Figure 5: Example of ELAN annotation tiers in LSE_Lex40_UVIGO dataset. *Ref* tier (encapsulates predicative expressions), *Trad* tier (the Spanish translation of the signed sentence), *MD_Glosa* and *MI_Glosa* (the right and left hands lexical signs "APRENDER, SIGNAR, EDAD, PODER"; transitions "dp" -from pause-, "es" -inter sign transition-, "cp" -to pause-; and semi-lexical signs "INDX:1sg" -pointing to subject-, "B:PODER" -buoy sign-).

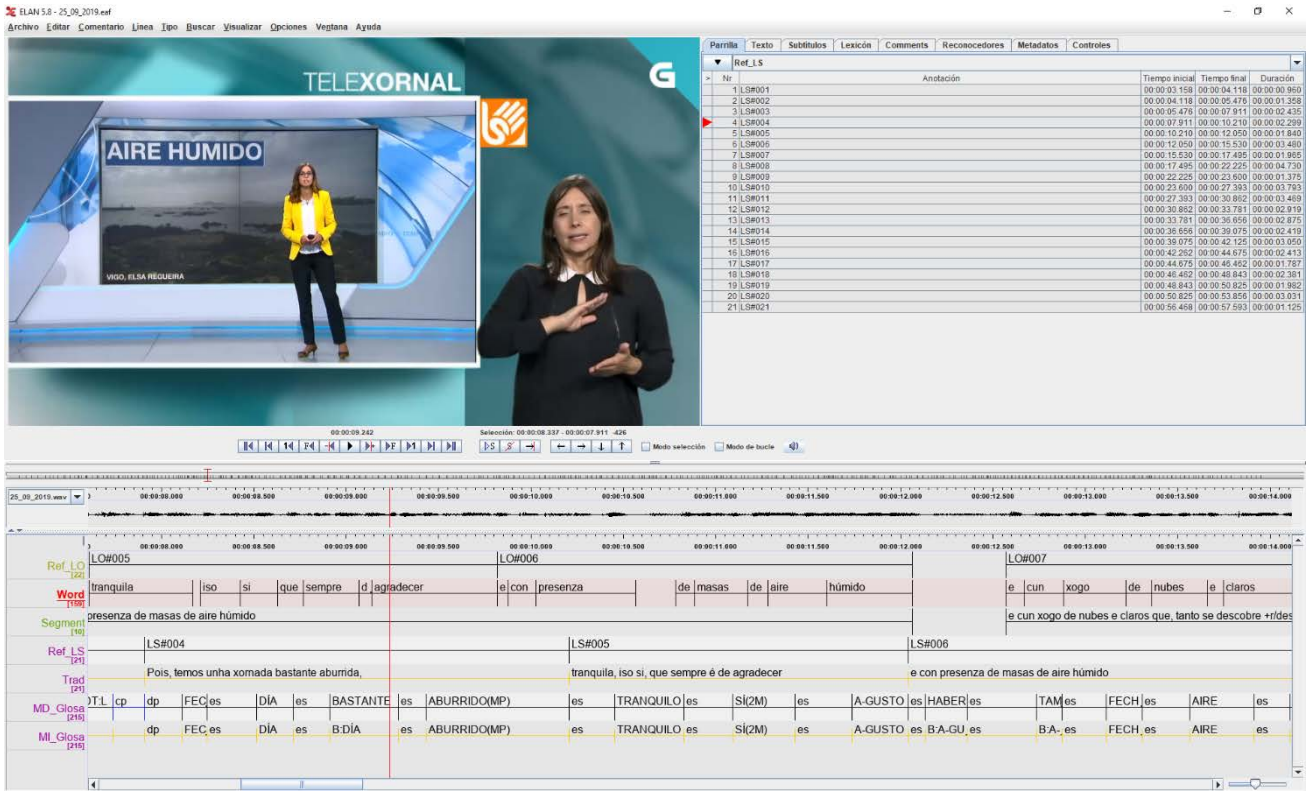

Figure 6: Example of ELAN annotation tiers in LSE_TVGWeather_UVIGO dataset. *Ref_LO* and *Ref_LS* tiers form pairs of spoken-signed CLUs, *Word* and *Segment* tiers come from the ASR and the manual review, respectively, *Trad* tier is the Galician utterance (hopefully quite close to the ASR in the *Word* and *Segment* tiers), but aligned with the LSE stream, and the rest of tiers as in Fig. 5.

5. Current state of the Database and Further Work

We started recording LSE_Lex40_UVIGO in May 2019 and, to this date, 35 signers have contributed to it. They come mostly from the deaf community and display a range of ages and fluency in sign language, and gender parity. So far, most of the videos have been recorded in the Association of Deaf People of Vigo (ASORVIGO), and the rest in the School of Telecommunications Engineering and in the Faculty of Philology and Translation of the University of Vigo. In all three cases the distance to the cameras and the framing was similar, while the background of the image has variations: It is a bare wall painted light in two of the locations, and is covered by a green fabric to eliminate reflections in the third. We did not impose any requirements on signer clothing. In future recordings we will incorporate other locations, lighting conditions and background types to test the robustness of the ASLR against this type of variation in the recording conditions.

Table 2 summarizes the main figures of LSE_Lex40_UVIGO dataset up to now: columns 2 through 5 indicate the number of different items in each section of the dataset (alphabet, isolated signs and sentences), the number of signers that have contributed to each section, the number of available recording of each item, and the total duration of the recordings. We plan to incorporate a new section to the dataset, namely 40 fingerspelled words.

Regarding LSE_TVGWeather_UVIGO dataset, recording started in August 2019 at a rate of about 18-20 videos per month. To this moment, about 100 videos have been recorded, most of which last between 1 and 2 minutes. Usually they are signed by the same person.

We are managing the transfer of the rights of the images by the signers in accordance with the European regulation of the protection of personal data, so a first release of the LSE_UVIGO database may be made available to the research community in the coming months.

Database section	# Items	# Signers	# Recordings	Total recordings duration (hh:mm:ss)
Alphabet	30	3	90	00:03:45
Isolated signs	40	32	1368	01:23:50
Sentences	40	13	493	00:34:46
Total	110	35	1951	02:01:31

Table 2: current state of LSE_Lex40_UVIGO dataset

6. Acknowledgements

This research is funded by the Spanish Ministry of Science, Innovation and Universities, through the project RTI2018-101372-B-I00 *Audiovisual analysis of verbal and nonverbal communication channels (Speech & Signs)*; by the Xunta de Galicia and the European Regional Development Fund through the Consolidated Strategic Group atlanTTic (2016-2019); and by the Xunta de Galicia through the Potential Growth Group 2018/60.

The authors wish express their immense gratitude to the Association of Deaf People of Vigo (ASORVIGO) and the Federation of Associations of Deaf People of Galicia (FAXPG) for their collaboration in the recording of the database LSE_Lex40_UVIGO.

7. References

Brugman, H. & Russell, A. (2004). Annotating multimedia/multi-modal resources with ELAN. Paper presented at the LREC 2004, In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Cabeza-Pereiro, M. C., Garcia-Miguel, J. M., García-Mateo, C., & Alba-Castro, J. L. (2016). CORILSE: a Spanish Sign Language Repository for Linguistic Analysis. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1402-1407).

Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3075-3084.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv preprint arXiv:1812.08008.

CNSE Foundation (2008). Diccionario normativo de lengua de signos española: Tesoro de la LSE [DVD].

Ebling, S., Camgöz, N. C., Braem, P. B., Tissi, K., Sidler-Miserez, S., Stoll, S., ... & Razavi, M. (2018). SMILE Swiss German sign language dataset. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Gutierrez-Sigut, E., Costello, B., Baus, C., & Carreiras, M. (2016). LSE-sign: A lexical database for spanish sign language. Behavior Research Methods, 48(1), 123-137.

Herrero Blanco, Á. L. (2009). Gramática didáctica de lengua de signos española, LSE. Ediciones SM, Madrid.

Hodge, G. (2013). Patterns from a signed language corpus: Clause-like units in Auslan (Australian sign language). Ph.D. thesis, Sydney: Macquarie University.

Huang, J., Zhou, W., Li, H., & Li, W. (2015, June). Sign language recognition using 3d convolutional neural networks. In 2015 IEEE international conference on multimedia and expo (ICME) (pp. 1-6). IEEE.

Johnston, T. (2013). Auslan Corpus Annotation Guidelines. Retrieved from http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers.

Computer Vision and Image Understanding, 141, 108-125.

Koller, O., Camgoz, C., Ney, H., & Bowden, R. (2019). Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. IEEE transactions on pattern analysis and machine intelligence. doi: 10.1109/TPAMI.2019.2911077

Li, D., Rodriguez-Opazo, C., Yu, X. and Li, H. (2019). Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. Accepted at IEEE 2020 Winter Conference on Applications of Computer Vision (WACV '20), March 2020. https://arxiv.org/abs/1910.11006

Martínez-Hinarejos, C. D., & Parcheta, Z. (2017). Spanish Sign Language Recognition with Different Topology Hidden Markov Models. In INTERSPEECH (pp. 3349-3353).

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1145-1153).

Tang, A., Lu, K., Wang, Y., Huang, J., & Li, H. (2015). A real-time hand posture recognition system using deep neural networks. ACM Transactions on Intelligent Systems and Technology (TIST), 6(2), 1-23.

Tilves-Santiago, D., Benderitter, I., & García-Mateo, C. (2018). Experimental Framework Design for Sign Language Automatic Recognition. In IberSPEECH (pp. 72-76).

Torres-Guijarro, S., García-Mateo, C., Cabeza-Pereiro, C., Docío-Fernández, L. (2020). LSE_Lex40_UVIGO Una base de datos específicamente diseñada para el desarrollo de tecnología de reconocimiento automático de LSE. Revista de Estudios de Lenguas de Signos (REVLES), 2.