Thursday, April 30, 2026

TReC 2026 Project Proposal Submission Form

Submit your project proposal for the 7th TRAIL Research Camp (August 24th - September 4th, 2026, Lausanne, Switzerland). Please complete all required sections and submit your proposal before April 30th, 01:00 PM (CET).

Administrative Data

Full Name of Team Leader Pierre Poitier

Contact Email pierre.poitier@unamur.be

Project Information

Project Title Leveraging Unannotated Data with Self-Supervised Learning for Continuous Recognition: Application to French Belgian Sign Language (LSFB)

Profile of the Team Leader(s) & Expected Team Composition

Pierre Poitier: PhD Student (UNamur, HuMaLearn, LSFB-Lab, TRAIL, NaDI), thesis on Better Transition Modeling in Sign Language Segmentation, expertise in sign language processing (both segmentation and recognition) and computer vision.

Ariel Basso Madjoukeng: PhD Student (UNamur, HuMaLearn, TRAIL, NaDI), thesis on Self-supervised and continual learning for sign language recognition, expertise in sign language recognition, self-supervised, metric and weakly supervised learning.

Adélaïde Couplet: PhD Student (UNamur, HuMaLearn, LSFB-Lab, TRAIL, ARIAC, NaDI, NaLTT), thesis on Sign Language Processing and LSFB Linguistics, expertise in linguistics and NLP.

Elisa De Coster: PhD Student & Teaching Assistant (UNamur, HuMaLearn, NaDI), thesis on the Rashomon effect for variability forecasting, expertise in XAI and deep learning.

Have you already identified potential team members for your project? No

Domain of Application Media/Culture/Digital Humanities

Scientific Theme NLP

Proposal Content

Abstract

Continuous Sign Language Recognition (CSLR) for French Belgian Sign Language (LSFB) is constrained by annotation cost: roughly two-thirds of the available LSFB corpus is unannotated, and only one-third carries the manual annotations needed to train continuous recognition models directly. Self-supervised learning (SSL) is a natural fit for this setting, as it can exploit unlabeled signing data to learn representations that transfer to downstream tasks. Yet, to our knowledge, no SSL method has been applied to LSFB for continuous recognition, and the few SSL approaches proposed for sign language more broadly predate

recent advances in generative modeling of video, which we believe offer strong opportunities for learning from unannotated signing data. We propose to pretrain a CSLR model on the unannotated portion of LSFB using a generative self-supervised objective inspired by recent video generation work, and to fine-tune the resulting encoder on the annotated portion for evaluation of CSLR. The expected contribution is twofold: a first self-supervised baseline for continuous recognition on LSFB, and an empirical comparison between earlier SSL recipes for sign language and a more recent generative approach.

Background Information & Problem Statement

Sign language processing (SLP) suffers from a data scarcity problem [6]. Modern deep learning models require large amounts of labeled data, but sign language annotation is prohibitively expensive: Renz et al. [14] estimate that producing gloss-level annotations for one hour of signing requires roughly 100 hours of expert work. Although several large-scale corpora have been collected, e.g., BOBSL [1], AUSLAN [10], and How2Sign [5], only a small fraction of the datasets has been manually annotated. LSFB-CONT [7] follows the same pattern: around one-third of its ~130 hours of continuous SL videos has gloss-level supervision, while the remaining two-thirds is left unused by standard supervised pipelines and can be leveraged by unsupervised methods.

Self-supervised learning (SSL) has become the dominant strategy for leveraging such unannotated data, with successes ranging from BERT [4] in natural language processing (NLP) to MAE [8] and DINO [2] in computer vision (CV). SLP inherits properties from both fields and is a natural target for SSL. Several works have indeed applied SSL to sign language, e.g., SignBERT [19], SignBERT+ [9], BEST [18], but model weights and implementations are often non-publicly available and have not been evaluated on LSFB.

Recent advances outside SLP suggest concrete ways to strengthen these methods. Video masked generative models such as MAGVIT [16] and MAGVIT-v2 [17] combine a temporal VQ-VAE tokenizer [15] with a MaskGIT-style bidirectional transformer [3], achieving state-of-the-art video generation while scaling more gracefully than autoregressive alternatives and capturing richer structure than pixel-level reconstruction. Adapting this paradigm to SLP, i.e., discretizing LSFB clips with a temporal VQ-VAE and pretraining a masked transformer, i.e., Mask-GIT, over the resulting tokens, is therefore a promising route to modernize BERT-style SSL for sign languages.

To assess the resulting representations, we adopt Continuous Sign Language Recognition (CSLR) [6] as the downstream evaluation task. CSLR jointly probes visual recognition and temporal modeling and provides a basis for comparison with results obtained on other sign languages. To the best of our knowledge, neither continuous SSL pretraining nor CSLR has previously been applied to LSFB, and this proposal aims to close that gap.

Project Objectives & Concrete Implementation

The main goal of this project is to develop a continuous sign language recognition model for French Belgian Sign Language (LSFB) by leveraging unannotated data. BERT-style pretraining [4], which relies on masking and reconstructing latent representations, has proven effective for other sign languages (e.g., SignBERT [19]), and recent work shows that reconstructing discrete tokens rather than continuous features further improves representation quality. Our first objective is therefore to train an autoencoder such as a Vector Quantized Variational Autoencoder (VQ-VAE) [15] to provide a temporally rich, discretized representation of LSFB videos (O1); its codebook will serve as the tokenizer for the subsequent masked pretraining stage. In addition, the quality of more classical Bidirectional Encoder Representations from Transformers (BERT)-style representations is known to depend heavily on the masking ratio and frame selection strategy, both of which vary across datasets. We will therefore empirically determine the masking configuration best suited to LSFB and use it together with the VQ-VAE embedding space to pretrain a BERT model on unannotated data (O2). We will also experiment with more recent variants better suited to visual data, such as MaskGIT [3] (O3). Finally, the pretrained encoder will be fine-tuned on the annotated data available and compared against an equivalent model trained from scratch, to quantify the benefit of self-supervised pretraining (O4).

Do you plan to deliver, as an outcome of your project, a reusable “brick” for the TRAIL Factory (https://factory.trail.ac/en/home_page) that could later be transferred and converted into a company process?

Not sure

Briefly describe what the brick would be and its intended users.

Post-workshop, the findings will be published, and all corresponding assets (documentation, source code, and results) will be open sourced on the TRAIL Factory.

Project Dataset

The French Belgian Sign Language (LSFB) corpus [13] is one of the largest annotated sign language corpora in the world. It was originally compiled by linguists for linguistic research on LSFB; the raw data have been processed to make it suitable for deep learning models. As a result, the LSFB-CONT dataset [7] is publicly available, comprising both an annotated portion (one third) with gloss-level supervision (a gloss is a label referring to a semantically meaningful sign [11]) and a larger unannotated portion (two thirds) of raw signing video. Pose data have been extracted for all videos using MediaPipe [12], providing a lightweight modality alongside the raw video. We have full rights to use the dataset for research purposes. The corpus has already been packaged in WebDataset format with a companion data loader, making it straightforward to integrate into training pipelines. Together, these properties make the LSFB dataset well suited to a wide range of sign language processing tasks, including continuous recognition and self-supervised pretraining.

Detailed Work Plan

The first phase of this project focuses on training a Vector Quantized Variational Autoencoder (VQ-VAE) to extract a temporally rich, discretized representation of LSFB videos (O1). The preliminary task will consist of training a baseline VQ-VAE, followed by rigorous hyperparameter tuning to ensure high-fidelity spatial and temporal reconstruction. The resulting discrete codebook will effectively serve as the tokenizer for all subsequent self-supervised pretraining stages. Leveraging the codebook generated in O1, the teams will pretrain a BERT model on unannotated data (O2). After deploying a baseline model, the goal will be to identify the optimal masking configuration. This will specifically involve tuning the masking ratio and frame selection strategies to best accommodate the specific characteristics of the LSFB dataset. As a more advanced alternative, or complement to BERT, MaskGIT will be explored (O3), as its bidirectional transformer decoder architecture is best suited for visual data. Similar to the previous phase, this will require establishing a baseline MaskGIT model, followed by the tuning of its masking scheduling and hyperparameters. Whether the BERT (O2) or MaskGIT (O3) strategy is favored, or combined for a hybrid approach, the final objective is to evaluate the resulting models on Continuous Sign Language Recognition (CSLR) tasks (O4).

The strength of this work plan lies in its highly modular design, where each objective (O1-O4) yields clear, achievable deliverables that ensure logical continuity and regular milestones throughout the project. Because the BERT (O2) and MaskGIT (O3) architectures can be explored as standalone substitutes or complementary models, the pipeline supports a fully parallelized workflow. O2 and O3 being independent of one another, once the tokenizer (O1) is established, multiple teams can concurrently tackle different objectives. This parallelism allows for a broader exploration of architecture choices within the constrained two-week timeframe. Furthermore, because the foundational representation (O1) and the final evaluation protocol (O4) remain strictly identical regardless of the intermediate pretraining strategy, the framework guarantees a straightforward and robust comparative analysis. This will yield insights into which model performs best for unannotated LSFB data, and the potential synergies of combining them. Finally, this parallel structure empowers teams to align their specific tasks with their personal preferences, methodological expertise, and research motivations.

As for the project schedule, day one will focus on establishing a shared vision to strategically allocate tasks based on team expertise; subsequently, the first two-thirds of the workshop will be dedicated to active implementation, with the final third reserved for rigorous evaluation.

Bibliographic References

[1] S. Albanie, G. Varol, L. Momeni, et al. BBC-Oxford British Sign Language Dataset. 2022.

[2] M. Caron et al., “Emerging Properties in Self-Supervised Vision Transformers”, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.

[3] H. Chang, H. Zhang, L. Jiang, et al. “Maskgit: Masked generative image transformer”, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 11315-11325.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: ACL, 2019, pp. 4171–4186.

[5] A. Duarte et al., “How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language”, in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE, 2021, pp. 2734–2743.

[6] J. Fink, M. De Coster, J. Dambre, and B. Frénay, “Trends and Challenges for Sign Language Recognition with Machine Learning”, in ESANN 2023 proceesdings, Bruges (Belgium), 2023, pp. 561–570.

[7] J. Fink, B. Frenay, L. Meurant, and A. Cleve, “LSFB-CONT and LSFB-ISOL: Two New Datasets for Vision-Based Sign Language Recognition”, in 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China: IEEE, Jul. 2021, pp. 1–8.

[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked Autoencoders Are Scalable Vision Learners”, presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[9] H. Hu, W. Zhao, W. Zhou, and H. Li, “SignBERT+: Hand-Model-Aware Self-Supervised Pre-Training for Sign Language Understanding”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 9, pp. 11221–11239, 2023.

[10] T. Johnston, “The lexical database of Auslan (Australian Sign Language)”, Sign Language & Linguistics, vol. 4, no. 1–2, pp. 145–169, 2001.

[11] T. Johnston, “From archive to corpus: Transcription and annotation in the creation of signed language corpora”, International Journal of Corpus Linguistics, vol. 15, no. 1, pp. 106–131, Jan. 2010.

[12] C. Lugaresi, J. Tang, H. Nash, et al., “Mediapipe: A framework for perceiving and processing reality”, in: Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR). Long Beach, CA, 2019. p. 2.

[13] L. Meurant, “Corpus LSFB”, Corpus informatisé en libre accès de vidéo et d’annotations de langue des signes de Belgique francophone. Namur : Laboratoire de langue des signes de Belgique francophone (LSFB Lab), FRS-FNRS, Université de Namur, 2015.

[14] K. Renz, N. C. Stache, S. Albanie, and G. Varol, “Sign Language Segmentation with Temporal Convolutional Networks”, in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 2135–2139.

[15] A. Van Den Oord, O. Vinyals, et al., “Neural discrete representation learning”, in Advances in Neural Information Processing Systems, 2017, vol. 30.

[16] L. Yu et al., “MAGVIT: Masked Generative Video Transformer”, presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10459–10469.

[17] L. Yu et al., “Language Model Beats Diffusion - Tokenizer is Key to Visual Generation”, 12th International Conference on Learning Representations, ICLR 2024, 2024.

[18] W. Zhao, H. Hu, W. Zhou, J. Shi, and H. Li, “BEST: BERT Pre-training for Sign Language Recognition with Coupling Tokenization”, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 3, pp. 3597–3605, Jun. 2023.

[19] Z. Zhou, V. W. L. Tam, and E. Y. Lam, “SignBERT: A BERT-Based Deep Learning Framework for Continuous Sign Language Recognition”, IEEE Access, vol. 9, pp. 161669–161682, 2021.

Eligibility & Evaluation

Does the project include multidisciplinarity between STEM & SSH?

Yes

How?

This project is interdisciplinary; although the work realized during the workshop will be very technical and focused on advanced deep learning techniques, the output will directly be beneficial to SSH studies. Team leaders are used to working with the linguists of the LSFB-Lab and know from experience that linguists are looking forward to using models developed by our team in order to help their research. Furthermore, we have a linguist in the team who brings linguistic expertise and a human-centered point of view.

We confirm that the Team Leader will be present for the full duration of TReC’26 if the project is selected (August 24th - September 4th, 2026, Lausanne, Switzerland)

I/We agree and confirm

Additional Comment

For this project, we will have access to high-performance computing via the institutional PTCI platform for large scale experiments (including Lucia with 200 GPU nodes). Also, most of the team members are from the HuMaLearn Team (AI and deep learning Lab of UNamur’s computer science faculty), and all members have access to three high-performance workstations, each equipped with an RTX 5090, remotely available for training AI models. All these resources will be sufficient to process large amounts of skeleton-based sign language data.