Sign Language-to-Text Dictionary with Lightweight Transformer Models

Jérôme Fink[1,2,3] , **Pierre Poitier**[1,3] , **Maxime André**[1,3] , **Loup Meurice**[1,3] , **Benoît Frénay**[1,3] , **Anthony Cleve**[1,3] , **Bruno Dumas**[1,3] , **Laurence Meurant**[2,3]

[1]Namur Digital Institute (NaDI)
[2]Namur Institute of Language, Text and Transmediality (NaLTT)
[3]University of Namur
{jerome.fink, pierre.poitier, maxime.andre, loup.meurice, benoit.frenay, anthony.cleve, bruno.dumas, laurence.meurant}@unamur.be

Abstract

The recent advances in deep learning have been beneficial to automatic sign language recognition (SLR). However, free-to-access, usable, and accessible tools are still not widely available to the deaf community. The need for a sign language-to-text dictionary was raised by a bilingual deaf school in Belgium and linguist experts in sign languages (SL) in order to improve the autonomy of students. To meet that need, an efficient SLR system was built based on a specific transformer model. The proposed system is able to recognize 700 different signs, with a top-10 accuracy of 83%. Those results are competitive with other systems in the literature while using 10 times less parameters than existing solutions. The integration of this model into a usable and accessible web application for the dictionary is also introduced. A user-centered human-computer interaction (HCI) methodology was followed to design and implement the user interface. To the best of our knowledge, this is the first publicly released sign language-to-text dictionary using video captured by a standard camera.

1 Introduction

The rise of deep learning [LeCun *et al.*, 2015] led to the creation of successful methods to process unstructured data such as images, videos or texts. These achievements are reflected in sign language recognition (SLR). The field has gained in popularity [Koller, 2020] as it provides a challenging benchmark for gesture or poses recognition. Indeed, to correctly classify signs, a model should be able to grasp facial expressions and precise hand gestures [Stokoe, 1972]. Moreover, there is a clear societal dimension for such technologies, such as the sign language-to-text dictionary which is proposed here to help the deaf community.

Technological advances alone cannot explain the success of SLR. In the past decades, linguists began to have access to affordable storage and recording devices. It facilitated the study of sign languages (SL) and has encouraged several research teams to create digital sign language corpora. In the meantime, the expansion of smartphones and social networks led to the creation of groups on social media platforms in which deaf users can share SL vocabulary or communicate online. The increasing availability of sign language (SL) data allows machine learning (ML) researchers to exploit those corpus [Fink *et al.*, 2021] or crowdsource [Vaezi Joze and Koller, 2019] social media platforms to build large-scale SL datasets suitable for deep learning.

Despite those advances, few tools are available to the deaf community. Initiatives led to the creation of lexicons for sign language enabling to search for a sign corresponding to a written word[1]. However, the opposite is not possible as those tool does not offer a search from a sign to a written word. This work proposes to enhance those tools by providing a dictionary searchable via a webcam recording. This dictionary is, to the best of our knowledge, the first publicly available sign language-to-text dictionary[2] using only video information from a simple webcam to identify the sign.

The overall process leading to the creation and use of our dictionary is summarized in Figure 1. A corpus of French Belgian Sign Language (LSFB) built by a team of linguists from the LSFB laboratory (LSFB Lab) of Namur [Meurant, 2015] is used as a database for the system. A cleaned version of the corpus [Fink *et al.*, 2021] is used as a dataset for the machine learning pipeline. This paper focuses on the creation of a lightweight model for SLR using an architecture similar to the one introduced by Vision Transformer (ViT) [Dosovitskiy *et al.*, 2021]. In addition, the integration of the resulting model into a web application is also presented. A user-centered approach is followed for ensuring the stakeholder's requirements meeting on the resulting dictionary. This ensures that our tool will actually be useful to the deaf community, as confirmed by its quick adoption after its public release in October 2022.

This paper is organized as follows. Section 2 introduces the stakeholders of the SLR system along with its requirements. Then, Section 3 discusses the research in SLR. Section 4 gives more information about the dataset used in this work and its specificities. Section 5 describes the architecture developed for the dictionary and reports results for various architectural choices. A quantitative evaluation of the best-performing model is reported. Section 6 explains how

[1] auslan.org.au

[2] dico.corpus-lsfb.be

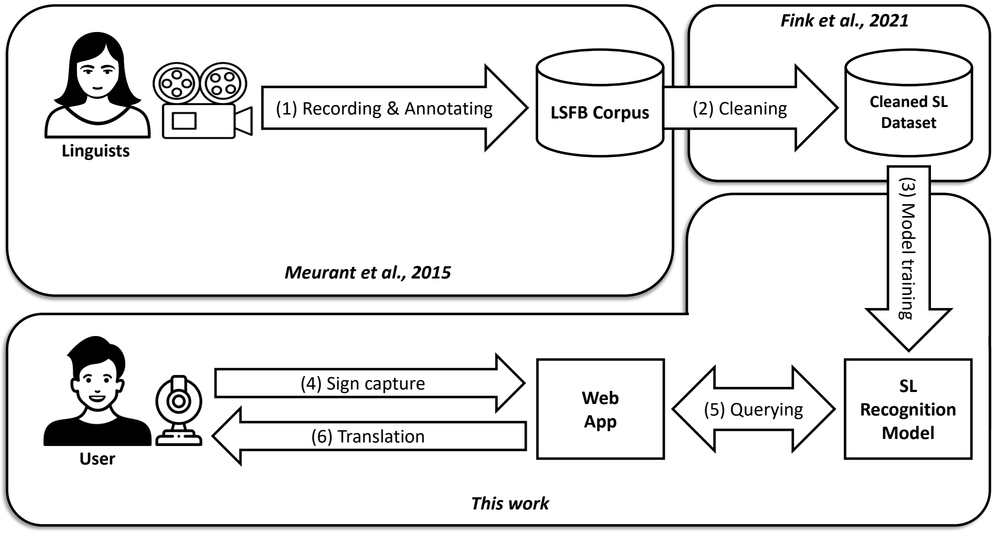


Figure 1: The high-level processes that lead to the creation and manipulation of the bidirectional sign language dictionary. (1) The LSFB Lab collected and annotated a large corpus of French Belgian Sign Language (LSFB) [Meurant, 2015]. (2) The corpus was preprocessed and cleaned to create a sign language dataset [Fink *et al.*, 2021]. (3) The dataset is used to train our SLR model. (4) An interface was built to capture the user's signs and use them to query the dictionary (5). The dictionary proposes possible translations to the user along with definitions and usage examples in text and in video (6).

the web application integrating the model was designed, implemented and evaluated using a user-centered approach. Finally, Section 7 concludes and discusses future works.

2 Stakeholders and Requirements

It is important to notice that sign languages are not universal and may vary depending on the country or region. The system presented in this paper focuses on the French Belgian Sign Language (LSFB). Nevertheless, the overall process followed to build the system is transferable to any sign language (SL), provided that the amount of available data is sufficient.

Our project was initiated by the French Belgian Sign Language Laboratory (LSFB Lab) of Namur, where linguists have been working on the LSFB since early 2000. They collected videos of SL conversations to better study and characterize the language. They also released a text-to-sign language lexicon. The LSFB Lab collaborates with *Sainte-Marie*, a bilingual French and LSFB school located in Namur. The creation of a sign language-to-text dictionary could improve the autonomy of deaf students. Thus, the school was interested and involved in the creation of the interface.

Discussions with the stakeholders allowed us to gather requirements for the application. First, the system should be robust to variations. The users are not expected to stand in a controlled environment with uniform background and lightning or to wear specific clothing. Also, skin color and any other physical characteristics should have no influence.

The system should not rely on expensive, impractical or hard-to-find hardware. Thus, the dictionary should only rely on video captured by a standard webcam that can be found on laptops or smartphones. The association hosting the system cannot afford a server with GPUs. Thus, the algorithm must run efficiently on CPU only. Finally, the system should answer in less than 10 seconds to a query. This ensures that the interface is fluid and not frustrating to use.

3 Related Work

Sign language recognition is gaining in popularity in machine learning [Koller, 2020]. Continuous SLR aims to translate SL sentences directly into text, while isolated SLR focuses on classifying a single sign. This section focuses on isolated sign language recognition using RGB data, as our system can only rely on raw videos for its predictions and its aim is not to recognize and translate entire sentences.

The first vision-based SLR systems relied on handcrafted features like the work of [Huang and Huang, 1998] using Otsu thresholding to isolate the hands. Those methods were only capable of recognizing a limited number of signs (< 100) from a few signers (< 5). The use of sequential models such as Hidden Markov Models led to the first system able to recognize larger sign vocabulary like in the work of [Kadir *et al.*, 2004] that achieved 92% accuracy for 164 signs. By using dynamic time warping, [Wang *et al.*, 2012] achieve impressive results with 78% top-10 accuracy on 1,113 signs using 20 frames and meta-information about the number of hands used to perform the sign and the handedness of the signers. However, those systems are sensitive to changes in lighting, background and signer variations.

The success of convolutional neural networks (CNN) for computer vision along with the development of large public datasets for sign language allowed the creation of algorithms robust to variability in the input data. A CNN-based method [Pigou *et al.*, 2016] was able to classify a vocabulary of 100 signs performed by 78 different signers with a top-

1 accuracy of 60% and a top-10 of 90%. The development of sequential models allows leveraging the temporal information in sign language videos. The MS-ASL dataset was benchmarked [Vaezi Joze and Koller, 2019] on several architectures such as CNN+LSTM and I3D networks with a top-1 accuracy of 81% for 1,000 signs and 222 signers. Recently, transformer networks proved to be efficient in sign language recognition. A transformer-based architecture achieved 73% accuracy on a vocabulary of 100 signs performed by 67 signers by mixing frame information with skeleton metadata extracted from the videos [De Coster *et al.*, 2020].

In parallel, advances in pose estimation led to the creation of valuable tools for preprocessing sign language videos. OpenPose [Cao *et al.*, 2019] and MediaPipe [Lugaresi *et al.*, 2019] provide easy-to-use models to extract skeletons landmarks from raw RGB videos. Those skeletons are often used as a preprocessing step in SLR [Konstantinidis *et al.*, 2018]. This work follows this trend by leveraging landmarks.

Since their creation, transformer-based architectures [Vaswani *et al.*, 2017] have proven successful on tasks such as image classification with the vision transformer (ViT) [Dosovitskiy *et al.*, 2021]. This work investigates the adaptation of such architectures for isolated SLR.

4 Dataset

Our SLR algorithm is trained on one of the largest sign language datasets in the world: the French Belgian Sign Language (LSFB) dataset [Fink *et al.*, 2021]. It is made of 50 hours of video, including 37 hours manually annotated by linguists from the LSFB Lab. Those videos depict natural discussions in LSFB between two individuals. In total, 100 signers participated in the recording sessions. Videos are recorded in a studio with controlled lighting and camera position. For each discussion, two videos are recorded, each focusing on one of the two signers.

LSFB-ISOL. The dataset exists in two versions: (i) LSFB-CONT which contains continuous videos of the whole LSFB discussions and (ii) LSFB-ISOL in which all the signs are isolated in shorter videos extracted from the continuous videos. Only LSFB-ISOL is used here as this paper does not focus on continuous SLR but rather on the recognition of isolated signs. Resulting videos only contain a single sign with an associated label. In total, LSFB-ISOL contains 4,181 different signs that are performed by the 100 signers. In this work, those labels are filtered to only keep the ones associated with French translations in the LSFB dictionary and having more than 20 examples. This leads to a filtered dataset with 700 labels and 77,900 instances.

The LSFB dataset is challenging as signers are free to discuss without vocabulary or rhythm constraints. In this context, signers tend to sign more quickly and signs overlap. Thus, the start position of each sign depends on the previous one.

Pose Features. The dictionary uses pose data extracted from frames with MediaPipe [Lugaresi *et al.*, 2019]. As shown in Figure 2, a pose contains 65 landmarks for the body pose (23) and the hands (2 × 21). As each landmark is made of an x and y component, each pose contains 130 features in total.

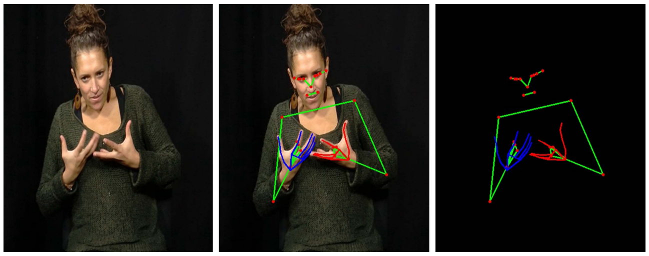


Figure 2: A frame sampled from the LSFB dataset along with its corresponding pose extracted using MediaPipe.

Multiple reasons motivate the use of poses instead of directly using the RGB frames:

(i) Less information is contained in a pose. An RGB frame of size 224x224 contains 150k values while a pose of 65 2D coordinates only contains 130 values. This represents a significantly smaller feature space that is easier to work with.

(ii) Some biases appear in the LSFB datasets, e.g., the uniform background and controlled lightning. This can cause bias if the training is performed directly on the frames. However, the poses are extracted with MediaPipe which is trained with respect to guidelines that prevent issues such as physical biases (background, light condition, etc.) and ethical biases (morphology, gender, skin color, etc.) [Lugaresi *et al.*, 2019]. Therefore, this paper “delegates” some potential biases to MediaPipe by using poses.

(iii) Poses only contains information about the joints of the signer. Therefore, irrelevant information, e.g., the color of the clothes, is not used to make the prediction. This prevents overfitting by filtering information. It also makes the model robust to those variations by design.

Features are processed to avoid a discontinuity in pose sequences and to mitigate vibrations caused by a lack of precision in the pose estimation. Linear interpolation is used to fill in missing values. Then, a filter [Savitzky and Golay, 1964] is used with a moving window of size 7 and a polynomial order of 2 to smooth values and thus mitigate vibrations.

5 Model Design

This section introduces the SLR model integrated to the dictionary. First, the overall architecture is described and results are reported for various meta-parameters. The best-performing model is discussed and other results found in the literature are reported.

5.1 Model Architecture

The success of transformer-based architectures in computer vision motivates their use for the challenging task of SLR. As the target is a specific class (i.e., type of sign) for a sequence of frames constituting a sign, the decoder part of the transformer architecture [Vaswani *et al.*, 2017] is not useful in our case. Instead, the architecture is inspired by the vision transformer (ViT) [Dosovitskiy *et al.*, 2021] for image classification. Figure 3 shows the high-level architecture of our sign language classifier. The linear embedding reduces the dimensionality of the input data before applying a positional encoding on each token. The positional encoding is a 1D trainable vector added to each input token. A classification token is added to the sequence as introduced in the ViT paper. This token is then passed as input to the multilayer perceptron (MLP) containing a normalization layer [Ba *et al.*, 2016] followed by a linear layer in order to predict a label for the sequence. The detailed architecture for the two other components is discussed in the following sections.

5.2 Training Setup

This section presents the training setup used to create our models. The filtered LSFB-Isol dataset presented in Section 4 is used, with a total of 77,900 instances and a vocabulary of 700 signs. The dataset is split into a training set containing 70% of the data and a test set containing the remaining. The signers appearing in the training set are not in the test set, to assess the ability of the model to deal with new signers. The MediaPipe landmarks are extracted from each clip. Only the landmarks are provided as input to our model, i.e., there are 130 input features. The raw video frames are not used.

All the models are trained using the same training scheme. The optimizer is a SGD with a learning rate of 2×10^{-3} and a momentum of 0.9. The loss function is the classical cross-entropy loss. The models are trained for 600 epochs. As recommended by [Vaswani *et al.*, 2017], a warmup phase is performed. A linear warmup is applied during the first 200 epochs. The batch size is set to 128. The metric used to compare each model is the standard accuracy. The clip sequences exceeding the maximal sequence length are cropped and the ones that are shorter are masked.

5.3 Transformer Encoder Architecture

A transformer encoder is made of one or several encoder layers containing a multi-head attention layer and a feed-forward network [Vaswani *et al.*, 2017]. The number of encoder layers and attention heads has an influence on the performance and complexity of the model. To determine the transformer encoder architecture for our SLR model, a grid search on several meta-parameters was performed (see Table 1). The maximal length of signs sequences is set to 50 and the embedding size of the tokens is set to 96. In total, 16 configurations were considered and the results are reported in Table 2.

Number of attention heads	2, 4, 8, 16
Number of encoder layers	1, 2, 4, 6

Table 1: The meta-parameters considered during the grid search for the transformer encoder architecture (see Table 2).

On the training set, the accuracy score rises as the model complexity increases, but it is not the case with the test accuracy. It can be observed that models quickly overfit when they are more complex. The best performances are obtained with a transformer encoder with 2 layers and 4 attention heads. Thus, those meta-parameters were chosen for our model.

Nb. layers	**Nb. heads**	**Train acc.**	**Test acc.**
1	2	61.2%	50.7%
	4	67.2%	51.3%
	8	66.4%	44.9%
	16	68.0%	45.3%
2	2	79.4%	51.6%
	4	80.7%	**51.9%**
	8	81.3%	47.3%
	16	79.8%	41.9%
4	2	93.7%	48.5%
	4	93.8%	45.0%
	8	94.0%	42.1%
	16	94.4%	37.2%
6	2	98.0%	41.1%
	4	98.8%	33.8%
	8	99.1%	35.5%
	16	**99.0%**	26.3%

Table 2: Training and test accuracy for the 16 models trained to find the best meta-parameters for the transformer encoder. The best training and test accuracy are highlighted.

5.4 Embedding Block Architecture

The linear embedding and position encoding block reduce the dimensions of the input and add position information to each token before passing them to the transformer encoder. To find the best sequence length and token size, several architectures are considered for the embedding block. Table 3 summarizes the combinations of meta-parameters. The transformer encoder block is the one selected in the previous section. Once again, a grid search was applied to test all the combinations of those two meta-parameters. Table 4 summarizes the results.

Tokens size	64, 80, 96, 112
Max sequence length	30, 50, 60

Table 3: Summary of the meta-parameters considered during the grid search for the embedding block (see Table 4).

Augmenting the maximal size of the sequence seems to be damageable to the performance, and the embedding size should remain moderate. As in Table 2, too complex models tend to overfit. The best model is obtained with a maximal sequence length of 30 and an embedding size of 80.

5.5 Results and Discussions

Our best-performing architecture uses a transformer encoder with 2 layers and 4 attention heads with a maximal sequence length of 30 frames and a token size of 80. It reaches a top-1 accuracy of 54% and a top-10 accuracy of 83% on the test set. The top-10 accuracy is relevant in our use case as the user of the dictionary could choose the correct sign out of the 10 proposed by the system. The average recall and precision obtained by the model are respectively 43% and 51%. The per-class accuracy shows that classes with more examples are better identified by the model. Due to the unbalanced

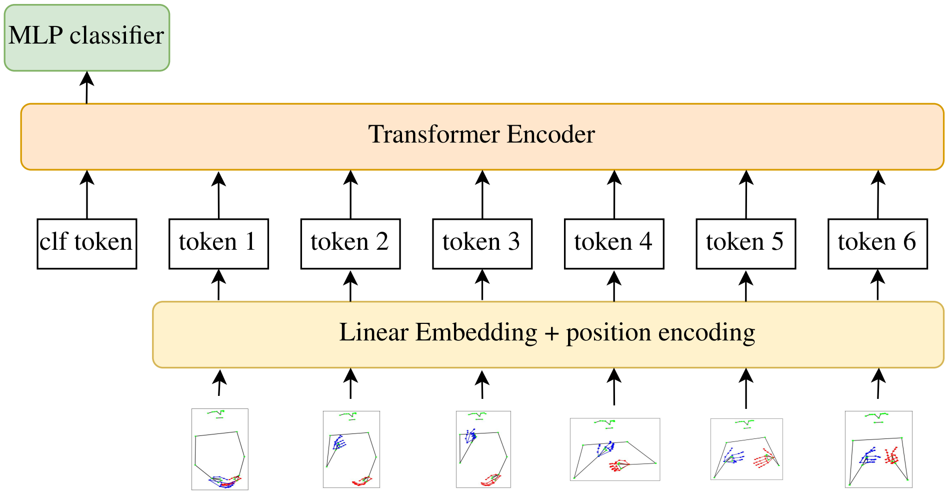


Figure 3: Summary of the architecture used for LSFB recognition. The input is a sequence of skeletons extracted using MediaPipe [Lugaresi *et al.*, 2019]. Each skeleton is embedded using a linear layer and a positional encoding is added to the resulting vector. A classification token is added at the start of the sequence as introduced by ViT. Then, the sequence of resulting tokens is sent to a transformer encoder. The classification token is then used to predict the label for the sign.

Max. seq. length	**Embedding size**	**Training acc.**	**Test acc.**
30	64	70.7%	52%
	80	76.7%	**54.4%**
	96	81.2%	53.6%
	112	**84.2%**	50.5%
50	64	69.9%	48.6%
	80	75.9%	47.9%
	96	79.7%	46.7%
	112	84.0%	49.4%
60	64	68.9%	42.8%
	80	75.3%	44.2%
	96	80.1%	47.0%
	112	83.2%	46.7%

Table 4: Training and test accuracy for the 12 models trained using various sequence lengths and embedding sizes. The best training and test accuracy are highlighted.

nature of the data, the most common signs have hundreds of examples while the least represented appears only 20 times leading to a great disparity in per-sign accuracy. The model also frequently mistakes signs presenting the same hand configuration and gestures.

To better assess the performances of our model regarding previous works, Table 5 reports results obtained by models using RGB video for isolated sign recognition. Only models trained on datasets with a similar number of signers and vocabulary are reported.

Notice that those results should be taken with caution as they are obtained on different datasets captured in different conditions and using distinct sign languages. For instance, the LSFB dataset and the BSL-1K [Albanie *et al.*, 2020] are the only reported datasets containing signs extracted from sentences. Thus signs from those datasets are performed faster and might overlap with the previous sign. It may not be relevant to compare the accuracy obtained on datasets that are so different. It is done here to give an indicative assessment of our system. Actually, the performances in real-world conditions may be radically different and the only relevant indicator of performance is the adoption of the system by users.

A key advantage of our LSFB classifier is that it proposes the lightest architecture for SLR currently available with, at least, 10 times fewer parameters than other methods. It is also lighter than a MobileNet [Sandler *et al.*, 2018] network designed to run on embedded devices. Despite that, the accuracy of our method is in the same range as the performance obtained by other models in the literature. The LSFB classifier is light enough to run on CPU efficiently, which is key for its adoption by non-profit stakeholders that have not enough resources and technical knowledge to maintain a GPU server. Our overarching goal is to maximize its societal impact.

6 System Integration

To achieve tangible societal impact, according to United Nations' Sustainable Development Goals [UN, 2015] and particularly the goal 4 "Quality Education" and the goal 10 "Reduced Inequalities", the model is integrated into a free and accessible system: the sign language-to-text dictionary which has been publicly released and is already used by the deaf community.

As illustrated in Figure 4, the system takes the form of a web application combining the features and appearance inspired by well-established online textual dictionaries such as Google Translate[3] or Linguee[4]. The dictionary allows users to sign in front of their camera to search for the literal translation of a sign in French. Users are invited to sign during a

[3] translate.google.com

[4] www.linguee.com

Authors	Vocabulary	Signers	Parameters	Top-1	Top-10	Dataset	Base architecture
[Izutov, 2020]	500	222	8.3M	63.36	-	MS-ASL	S3D
[Izutov, 2020]	1000	222	8.3M	45.65	-	MS-ASL	S3D
[Li *et al.*, 2020]	1000	116	12M	47.33	84.33	WLASL	I3D
[Albanie *et al.*, 2020]	1000	40	12M	65.57	-	BSL-1K	I3D
[Liao *et al.*, 2019]	500	8	11.4M	89.8	-	DEVISIGN-D	Resnet + LSTM
LSFB classifier (ours)	700	100	782k	54.4	83.4	LSFB-ISOL	ViT

Table 5: This table reports the score obtained by other researchers on various datasets for isolated SLR using only RGB video. The number of parameters for each architecture is reported. Our solution has, at least, 10 times fewer parameters than other methods.

fixed time window. Then, they are able to browse the propositions made by the model to find the corresponding sign in the dictionary. For the selected predicted sign, all the possible French translations are displayed. Moreover, for each translation, the application displays bilingual examples showing how the sign is used in a real SL video sentence alongside its French translation. This allows users to understand the use of the sign in different contexts. The dictionary drastically increases the autonomy of deaf people. It is also a useful tool for French-speaking people learning sign language or sign language interpreters who can perfect their knowledge by browsing contextual examples of signs.

The remaining of this section discusses the design and implementation of the dictionary. The compliance with the requirements elicited by the stakeholders is also assessed.

6.1 Design and Implementation

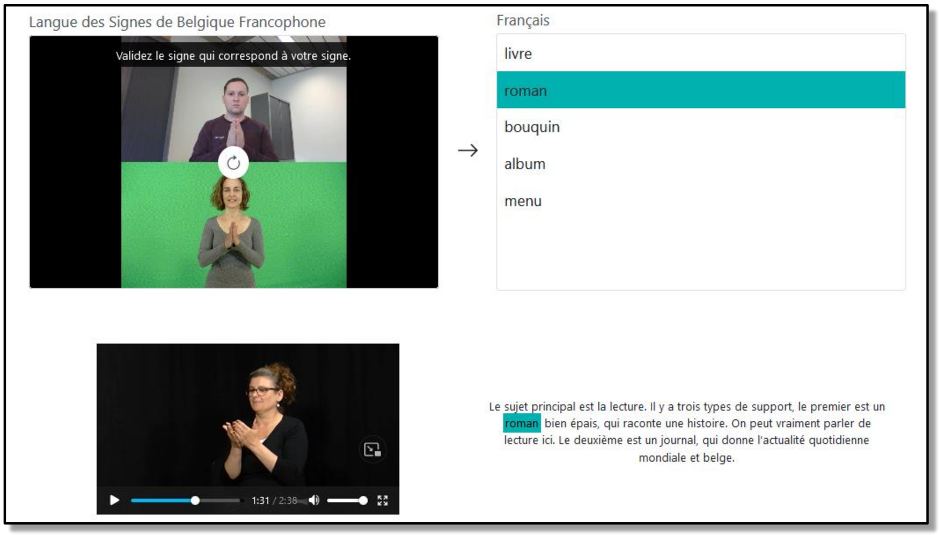
In order to put the user in the center of the process, the design phase started with requirements engineering activities with the stakeholders. First, based on semi-conducted discussions, four personas [Lallemand, 2018] were created (deaf user, deaf student, bilingual teacher, and sign language expert). This HCI good practice helped to identify the target users for the dictionary and the scope of their requirements. Moreover, a comparison of famous online dictionaries or translators (e.g., Google Translate, DeepL, Microsoft Bing) was conducted to confront their features with the needs of the personas. This then initiated the design of low and high-fidelity prototypes [Lallemand, 2018] for the dictionary. Those artifacts were evaluated in a continuous collaboration and validation with the four users representing each persona (2 deaf students, 1 bilingual teacher, 1 sign language expert), stakeholders (2 project leaders), and experts in HCI (1 UX expert and 1 inclusive UX expert). Finally, as the website is used by deaf people, great care has been taken to ensure accessibility. Guidelines for the design of interfaces suited for deaf people were searched. The web content accessibility guidelines (WCAG2) [Caldwell *et al.*, 2008] proposed by the W3C provide some general recommendations to design inclusive websites but nothing specific to the context of deafness. Therefore, the rest of the literature was explored and examined. Among the identified works, the guidelines were sometimes not the primary focus of the study or were too general for our purpose. There was a need for precision, completeness and cohesion. The work by [André, 2022] gathered, classified, and completed the recommendations found in the literature to establish a checklist for the creation of UX adapted to deafness (e.g., transforming all sound signals to visual ones, using icons instead of texts). Those recommendations were applied to the creation of our dictionary.

To transform the prototype into a working web application, all the components were implemented and connected together. The frontend of the application uses MediaPipe to extract the poses on the client side. Thus, only the landmarks extracted on the devices of the users are sent to the server to reduce the bandwidth needs and to preserve the privacy of users. A RESTful API provides endpoints to retrieve the possible translation for a given sign and the video example from the corpus LSFB. The API rely on our model to predict the label of a sign given MediaPipe landmarks. The global architecture is depicted in Figure 5.

6.2 Requirements Assessment

To assess the conformity of the user requirements, a usability testing [Lallemand, 2018] approach was followed. The main goal was to collect qualitative data to improve the system following a feedback loop mechanism. Six realistic usage scenarios mixing success and failure cases were proposed to the four users. It should be noted that tester users were not involved in the dataset creation, few years earlier. Those scenarios forced them to go through all the application functionalities, allowing us to observe their reactions and spot their difficulties. The tests were followed by a survey and a semi-conducted discussion [Lallemand, 2018] to assess the feeling of users about the web application. Each test session was recorded by two cameras and two microphones. An observer took notes on an observation grid to spot all the hesitations or issues encountered by the user during the scenarios. A briefed sign language interpreter assisted the test conductor when the user was deaf. All the materials used during the tests were translated into sign language by the interpreter.

The observations and remarks collected during those tests showed that the users were able to execute all scenarios without major difficulties. The success rate for the scenarios ranges from 87% to 98%. The gap is explained by the variety of users. Indeed, it has been noticed that children took a little more time, due to their distraction. In general, the first scenario also lasted longer, since users were new to the application. Finally, users reported that they appreciated the ease of use, simplicity, guidance, and the contextualized examples. However, they also asked for a better tolerance to their own inaccuracy while signing. Those insights were compiled



Validez le signe qui correspond à votre signe.

Le sujet principal est la lecture. Il y a trois types de support, le premier est un roman bien épais, qui raconte une histoire. On peut vraiment parler de lecture ici. Le deuxième est un journal, qui donne l'actualité quotidienne mondiale et belge.

Figure 4: Screenshot of the dictionary[6] after a successful search. The top of the interface shows the sign performed by the user along with the possible translation in French. The bottom of the interface gives contextual examples of the selected translation in sign language (video) and in French (text). Signers can hence improve themselves based on those examples.

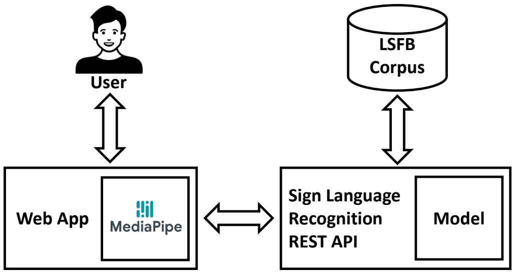


Figure 5: The system is made of three artifacts: (i) the web application that provides an interface for the user and uses MediaPipe JS to preprocess locally the captured video, (ii) an API hosting the SLR model and that is linked to (iii) the corpus database containing lexicon and contextual examples.

to serve as the starting point of the next development iteration [André, 2022], as the dictionary will continue to evolve, so as to better meet the deaf community's needs.

Regarding the requirements elicited by the stakeholders in Section 2, the system is compliant as it has been successfully deployed on the server of the LSFB Lab while responding in less than 10 seconds to a query. Users can use the website in various environments and lighting conditions.

7 Conclusion and Future Work

This work introduces the first dictionary searchable from sign language to text, publicly available through a web interface[7]. It relies on a lightweight sign language recognition model, inspired by the recent advances in transformer networks such as the Vision Transformer architecture introduced by [Dosovitskiy *et al.*, 2021]. This work leverages the progress made in pose estimation to achieve SLR on landmarks extracted from videos instead of the raw frames. This further reduced the complexity of the model and it removes several challenges such as the robustness to changes in the recording environment. Those challenges are delegated to pose estimation libraries such as MediaPipe. Our model is able to classify 700 signs with a top-10 accuracy of 83%, and is light enough to be run on embedded devices if needed. The model achieves competitive results while being 10 times lighter than alternative solutions. The model is integrated into a web dictionary allowing the user to search for the meaning of a sign in French. The dictionary is continuously populated by a team of linguists, the LSFB Lab. A user-centered HCI methodology was followed to design the interface with insights from the stakeholders and future users of the system. An evaluation of the tool was performed with the users to assess its compliance with the requirements identified.

In future work, metrics-based methods will be explored to train models that recognize more signs by predicting the distance between two signs instead of predicting a label directly. Thus, the model might be able to recognize new signs without being retrained. New architectures will be investigated to improve the SLR performance and classification robustness. Online learning methods will be investigated to leverage the input of the users of our website to retrain the model.

A new design iteration for the interface will also be conducted. A survey will be sent to the users to collect their opinions on the UI after a few months of use. Those insights will be considered to upgrade the interface if needed. A browser plugin will also be developed to provide better integration of the tool for the users. The developed dictionary is meant to become a long-lasting tool for the deaf community.

[7] dico.corpus-lsfb.be

Ethical Statement

Our work has no ethical or societal risk. All subjects involved in the dataset agreed to have their video publicly published. Moreover, the developed application does not collect any private data and relies on pose estimation only. Above all, the dictionary improves the autonomy of deaf people and contributes to a more inclusive education system. More generally, it supports a better inclusion of the deaf community in society, according to SDGs 4 and 10 from United Nations.

Acknowledgments

We would like to thank the members of the LSFB Lab for their major contribution and collaboration. Moreover, we express our gratitude to the Baillet Latour Fund, the Walloon region for the Ph.D. grant from FRIA (F.R.S.-FNRS) and the project ARIAC piloted by Trail, an initiative of the Digital4Wallonia for their funding. This work was also funded by the FWO and F.R.S.-FNRS under the Excellence of Science (EOS) program.

References

[Albanie *et al.*, 2020] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer, 2020.

[André, 2022] Maxime André. Recommandations pour des interfaces utilisateurs adaptées à la surdité. Master's thesis, Université de Namur, 2022.

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Caldwell *et al.*, 2008] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. Web content accessibility guidelines (wcag) 2.0. *WWW Consortium (W3C)*, 290:1–34, 2008.

[Cao *et al.*, 2019] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[De Coster *et al.*, 2020] Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. Sign language recognition with transformer networks. In *12th international conference on language resources and evaluation*, pages 6018–6024. European Language Resources Association (ELRA), 2020.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Fink *et al.*, 2021] Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[Huang and Huang, 1998] Chung-Lin Huang and Wen-Yi Huang. Sign language recognition using model-based tracking and a 3d hopfield neural network. *Machine Vision and Applications*, 10(5-6):292–307, April 1998.

[Izutov, 2020] Evgeny Izutov. Asl recognition with metric-learning based lightweight network. *arXiv preprint arXiv:2004.05054*, 2020.

[Kadir *et al.*, 2004] Timor Kadir, Richard Bowden, Eng-Jon Ong, and Andrew Zisserman. Minimal training, large lexicon, unconstrained sign language recognition. In *BMVC*, pages 1–10, 2004.

[Koller, 2020] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.

[Konstantinidis *et al.*, 2018] Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. Sign language recognition based on hand and body skeletal data. In *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4. IEEE, 2018.

[Lallemand, 2018] Carine Lallemand. *Méthodes de Design UX. 30 méthodes fondamentales pour concevoir des expériences optimales. (2e edition).* 09 2018.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[Li *et al.*, 2020] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[Liao *et al.*, 2019] Yanqiu Liao, Pengwen Xiong, Weidong Min, Weiqiong Min, and Jiahao Lu. Dynamic sign language recognition based on video sequence with blstm-3d residual networks. *IEEE Access*, 7:38044–38054, 2019.

[Lugaresi *et al.*, 2019] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. 2019.

[Meurant, 2015] Laurence Meurant. Corpus LSFB. Corpus informatisé en libre acces 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.

[Pigou *et al.*, 2016] Lionel Pigou, Mieke Van Herreweghe, and Joni Dambre. Sign classification in sign language corpora with deep neural networks. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch, editors, *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 175–178, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[Savitzky and Golay, 1964] Abraham. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, July 1964.

[Stokoe, 1972] William C Stokoe. Classification and description of sign languages. *Current trends in linguistics*, 12:345–371, 1972.

[UN, 2015] UN. The 17 goals — sustainable development. https://sdgs.un.org/goals, 2015. (Accessed on 05/15/2023).

[Vaezi Joze and Koller, 2019] Hamid Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. In *The British Machine Vision Conference (BMVC)*, September 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[Wang *et al.*, 2012] Haijing Wang, Alexandra Stefan, Sajjad Moradi, Vassilis Athitsos, Carol Neidle, and Farhad Kamangar. A system for large vocabulary sign search. In *Trends and Topics in Computer Vision: ECCV 2010 Workshops, Heraklion, Crete, Greece, September 10-11, 2010, Revised Selected Papers, Part I 11*, pages 342–353. Springer, 2012.