

RESEARCH OUTPUTS / RÉSULTATS DE RECHERCHE

Sign Language to Text Dictionary with Lightweight Transformer Models

Fink, Jerome; Poitier, Pierre; ANDRE, Maxime; Meurice, Loup; Frénay, Benoît; Cleve, Anthony; Dumas, Bruno; Meurant, Laurence

Published in:

Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)

Publication date:
2023

[Link to publication](#)

Citation for published version (HARVARD):

Fink, J, Poitier, P, ANDRE, M, Meurice, L, Frénay, B, Cleve, A, Dumas, B & Meurant, L 2023, Sign Language to Text Dictionary with Lightweight Transformer Models. in *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023): AI for Social Good track*.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Sign Language-to-Text Dictionary with Lightweight Transformer Models

Jérôme Fink^{1,2}, 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}

¹Namur Digital Institute (NaDI)

²Namur Institute of Language, Text and Transmediality (NaLTT)

³University of Namur

{jerome.fink, pierre.poitier, maxime.andre, loup.meurice, benoit.frenay, anthony.cleve, bruno.dumas, laurence.meurant}@unamur.be

Abstract

1 The recent advances in deep learning have been
2 beneficial to automatic sign language recognition
3 (SLR). However, free-to-access, usable, and acces-
4 sible tools are still not widely available to the deaf
5 community. The need for a sign language-to-text
6 dictionary was raised by a bilingual deaf school
7 in Belgium and linguist experts in sign languages
8 (SL) in order to improve the autonomy of students.
9 To meet that need, an efficient SLR system was
10 built based on a specific transformer model. The
11 proposed system is able to recognize 700 different
12 signs, with a top-10 accuracy of 83%. Those results
13 are competitive with other systems in the literature
14 while using 10 times less parameters than existing
15 solutions. The integration of this model into a us-
16 able and accessible web application for the dictio-
17 nary is also introduced. A user-centered human-
18 computer interaction (HCI) methodology was fol-
19 lowed to design and implement the user interface.
20 To the best of our knowledge, this is the first pub-
21 licly released sign language-to-text dictionary us-
22 ing video captured by a standard camera.

1 Introduction

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

35 Technological advances alone cannot explain the success
36 of SLR. In the past decades, linguists began to have access
37 to affordable storage and recording devices. It facilitated the
38 study of sign languages (SL) and has encouraged several re-
39 search teams to create digital sign language corpora. In the

41 meantime, the expansion of smartphones and social networks
42 led to the creation of groups on social media platforms in
43 which deaf users can share SL vocabulary or communicate
44 online. The increasing availability of sign language (SL) data
45 allows machine learning (ML) researchers to exploit those
46 corpus [Fink *et al.*, 2021] or crowdsourced [Vaezi Joze and
47 Koller, 2019] social media platforms to build large-scale SL
48 datasets suitable for deep learning.

49 Despite those advances, few tools are available to the deaf
50 community. Initiatives led to the creation of lexicons for sign
51 language enabling to search for a sign corresponding to a
52 written word¹. However, the opposite is not possible as those
53 tools do not offer a search from a sign to a written word.
54 This work proposes to enhance those tools by providing a dictio-
55 nary searchable via a webcam recording. This dictionary
56 is, to the best of our knowledge, the first publicly available
57 sign language-to-text dictionary² using only video informa-
58 tion from a simple webcam to identify the sign.

59 The overall process leading to the creation and use of our
60 dictionary is summarized in Figure 1. A corpus of French
61 Belgian Sign Language (LSFB) built by a team of linguists
62 from the LSFB laboratory (LSFB Lab) of Namur [Meurant,
63 2015] is used as a database for the system. A cleaned version
64 of the corpus [Fink *et al.*, 2021] is used as a dataset for the
65 machine learning pipeline. This paper focuses on the creation
66 of a lightweight model for SLR using an architecture similar
67 to the one introduced by Vision Transformer (ViT) [Dosovitskiy
68 *et al.*, 2021]. In addition, the integration of the result-
69 ing model into a web application is also presented. A user-
70 centered approach is followed for ensuring the stakeholder’s
71 requirements meeting on the resulting dictionary. This en-
72 sures that our tool will actually be useful to the deaf commu-
73 nity, as confirmed by its quick adoption after its public release
74 in October 2022.

75 This paper is organized as follows. Section 2 introduces
76 the stakeholders of the SLR system along with its require-
77 ments. Then, Section 3 discusses the research in SLR. Sec-
78 tion 4 gives more information about the dataset used in this
79 work and its specificities. Section 5 describes the architec-
80 ture developed for the dictionary and reports results for var-
81 ious architectural choices. A quantitative evaluation of the

¹auslan.org.au

²dico.corpus-lsfb.be

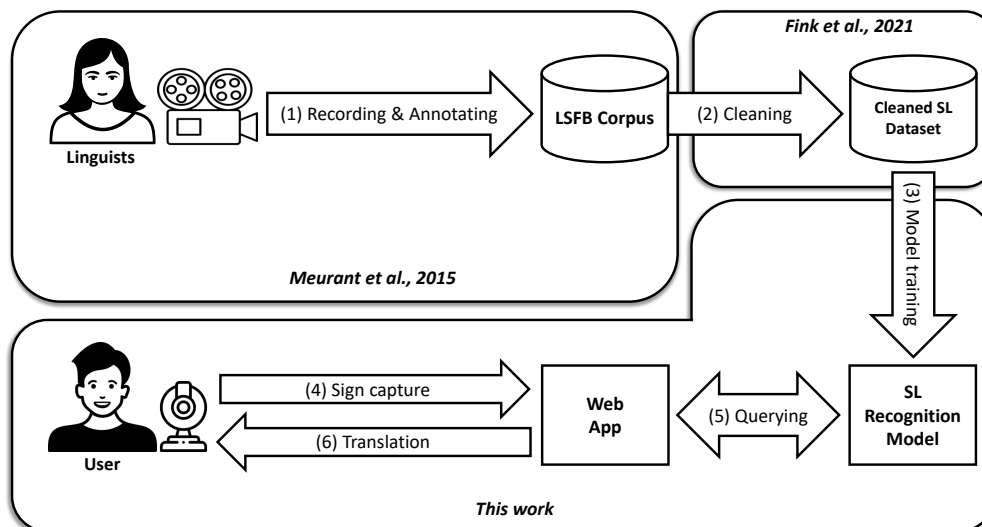


Figure 1: The high-level processes that lead to the creation and manipulation of the bidirectional sign language dictionary. (1) The LSFb Lab collected and annotated a large corpus of French Belgian Sign Language (LSFB) [Meurant, 2015]. (2) The corpus was preprocessed and cleaned to create a sign language 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).

82 best-performing model is reported. Section 6 explains how
 83 the web application integrating the model was designed, im-
 84 plemented and evaluated using a user-centered approach. Fi-
 85 nally, Section 7 concludes and discusses future works.

86 2 Stakeholders and Requirements

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

93 Our project was initiated by the French Belgian Sign Lan-
 94 guage Laboratory (LSFB Lab) of Namur, where linguists
 95 have been working on the LSFB since early 2000. They col-
 96 lected videos of SL conversations to better study and char-
 97 acterize the language. They also released a text-to-sign lan-
 98 guage lexicon. The LSFB Lab collaborates with *Sainte-*
 99 *Marie*, a bilingual French and LSFB school located in Na-
 100 mur. The creation of a sign language-to-text dictionary could
 101 improve the autonomy of deaf students. Thus, the school was
 102 interested and involved in the creation of the interface.

103 Discussions with the stakeholders allowed us to gather re-
 104 quirements for the application. First, the system should be
 105 robust to variations. The users are not expected to stand in a
 106 controlled environment with uniform background and light-
 107 ning or to wear specific clothing. Also, skin color and any
 108 other physical characteristics should have no influence.

109 The system should not rely on expensive, impractical or
 110 hard-to-find hardware. Thus, the dictionary should only rely
 111 on video captured by a standard webcam that can be found on
 112 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 an-
 swer in less than 10 seconds to a query. This ensures that the
 interface is fluid and not frustrating to use.

113 3 Related Work

114 Sign language recognition is gaining in popularity in machine
 115 learning [Koller, 2020]. Continuous SLR aims to translate
 116 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.

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

131 The success of convolutional neural networks (CNN) for
 132 computer vision along with the development of large pub-
 133 lic datasets for sign language allowed the creation of algo-
 134 rithms robust to variability in the input data. A CNN-based
 135 method [Pigou *et al.*, 2016] was able to classify a vocabulary
 136

144 of 100 signs performed by 78 different signers with a top-
145 1 accuracy of 60% and a top-10 of 90%. The development
146 of sequential models allows leveraging the temporal infor-
147 mation in sign language videos. The MS-ASL dataset was
148 benchmarked [Vaezi Joze and Koller, 2019] on several archi-
149 tectures such as CNN+LSTM and I3D networks with a top-1
150 accuracy of 81% for 1,000 signs and 222 signers. Recently,
151 transformer networks proved to be efficient in sign language
152 recognition. A transformer-based architecture achieved 73%
153 accuracy on a vocabulary of 100 signs performed by 67 sign-
154 ers by mixing frame information with skeleton metadata ex-
155 tracted from the videos [De Coster *et al.*, 2020].

156 In parallel, advances in pose estimation led to the creation
157 of valuable tools for preprocessing sign language videos.
158 OpenPose [Cao *et al.*, 2019] and MediaPipe [Lugaresi *et al.*,
159 2019] provide easy-to-use models to extract skeletons land-
160 marks from raw RGB videos. Those skeletons are often used
161 as a preprocessing step in SLR [Konstantinidis *et al.*, 2018].
162 This work follows this trend by leveraging landmarks.

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

168 4 Dataset

169 Our SLR algorithm is trained on one of the largest sign
170 language datasets in the world: the French Belgian Sign
171 Language (LSFB) dataset [Fink *et al.*, 2021]. It is made of
172 50 hours of video, including 37 hours manually annotated by
173 linguists from the LSFB Lab. Those videos depict natural
174 discussions in LSFB between two individuals. In total, 100
175 signers participated in the recording sessions. Videos are
176 recorded in a studio with controlled lighting and camera
177 position. For each discussion, two videos are recorded, each
178 focusing on one of the two signers.

179 **LSFB-ISOL.** The dataset exists in two versions: (i) LSFB-
180 CONT which contains continuous videos of the whole LSFB
181 discussions and (ii) LSFB-ISOL in which all the signs are
182 isolated in shorter videos extracted from the continuous
183 videos. Only LSFB-ISOL is used here as this paper does
184 not focus on continuous SLR but rather on the recognition
185 of isolated signs. Resulting videos only contain a single
186 sign with an associated label. In total, LSFB-ISOL contains
187 4,181 different signs that are performed by the 100 signers.
188 In this work, those labels are filtered to only keep the ones
189 associated with French translations in the LSFB dictionary
190 and having more than 20 examples. This leads to a filtered
191 dataset with 700 labels and 77,900 instances.

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

197 **Pose Features.** The dictionary uses pose data extracted
198 from frames with MediaPipe [Lugaresi *et al.*, 2019]. As

201 shown in Figure 2, a pose contains 65 landmarks for the body
202 pose (23) and the hands (2×21). As each landmark is made
203 of an x and y component, each pose contains 130 features in
204 total.

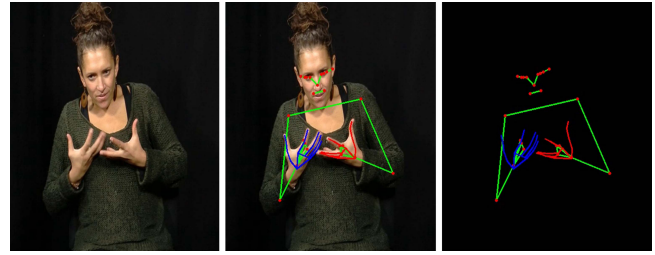


Figure 2: A frame sampled from the LSFB dataset along with its corresponding pose extracted using MediaPipe.

205 Multiple reasons motivate the use of poses instead of di-
206 rectly using the RGB frames:

- 207 (i) Less information is contained in a pose. An RGB frame
208 of size 224×224 contains 150k values while a pose of
209 65 2D coordinates only contains 130 values. This rep-
210 represents a significantly smaller feature space that is easier
211 to work with.
- 212 (ii) Some bias appear in the LSFB datasets, e.g., the uni-
213 form background and controlled lightning. This can
214 cause bias if the training is performed directly on the
215 frames. However, the poses are extracted with Medi-
216 aPipe which is trained with respect to guidelines that
217 prevent issues such as physical biases (background,
218 light condition, etc.) and ethical biases (morphology,
219 gender, skin color, etc.) [Lugaresi *et al.*, 2019]. There-
220 fore, this paper “delegates” some potential biases to Medi-
221 aPipe by using poses.
- 222 (iii) Poses only contains information about the joints of the
223 signer. Therefore, irrelevant information, e.g., the color
224 of the clothes, is not used to make the prediction. This
225 prevents overfitting by filtering information. It also
226 makes the model robust to those variations by design.

227 Features are processed to avoid a discontinuity in pose se-
228 quences and to mitigate vibrations caused by a lack of preci-
229 sion in the pose estimation. Linear interpolation is used to fill
230 in missing values. Then, a filter [Savitzky and Golay, 1964] is
231 used with a moving window of size 7 and a polynomial order
232 of 2 to smooth values and thus mitigate vibrations.

233 5 Model Design

234 This section introduces the SLR model integrated to the dic-
235 tionary. First, the overall architecture is described and re-
236 sults are reported for various meta-parameters. The best-
237 performing model is discussed and other results found in the
238 literature are reported.

239 5.1 Model Architecture

240 The success of transformer-based architectures in computer
241 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 multi-layer 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 accu-

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.

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.

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%.

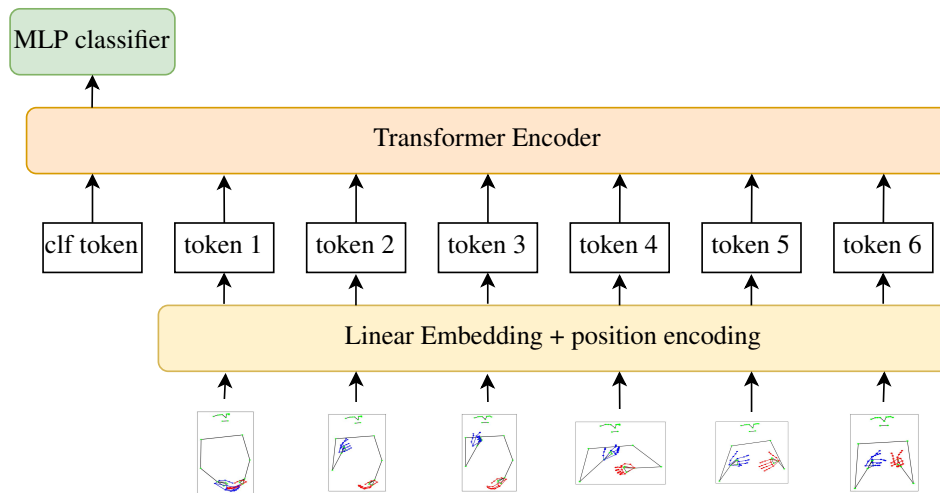


Figure 3: Summary of the architecture used for LSFb recognition. The input is a sequence of skeletons extracted using MediaPipe [Lugaresi *et al.*, 2019]. Each skeleton is embedded using a linear layer and a positional encoding is added to the resulting vector. A classification token is added at the 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.

the only reported datasets containing signs extracted from sentences, making them much more challenging. 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 others 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 maximise 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 Translate³ or Linguee⁴. The dictionary allows users

³translate.google.com

⁴www.linguee.com

320 The per-class accuracy shows that classes with more examples are better identified by the model. Due to the unbalanced 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.
 321
 322
 323
 324
 325
 326
 327 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.
 328
 329
 330
 331
 332 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
 333
 334
 335

336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365

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 <i>et al.</i> , 2020]	1000	116	12M	47.33	84.33	WLASL	I3D
[Albanie <i>et al.</i> , 2020]	1000	40	12M	65.57	-	BSL-1K	I3D
[Liao <i>et al.</i> , 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.

366 to sign in front of their camera to search for the literal trans-
367 lation of a sign in French. Users are invited to sign during a
368 fixed time window. Then, they are able to browse the propo-
369 sitions made by the model to find the corresponding sign in
370 the dictionary. For the selected predicted sign, all the possible
371 French translations are displayed. Moreover, for each trans-
372 lation, the application displays bilingual examples showing
373 how the sign is used in a real SL video sentence alongside
374 with its French translation. This allows users to understand
375 the use of the sign in different contexts. The dictionary dras-
376 tically increases the autonomy of deaf people. It is also a use-
377 ful tool for French-speaking people learning sign language or
378 sign language interpreters who can perfect their knowledge
379 by browsing contextual examples of signs.

380 The remaining of this section discusses the design and im-
381 plementation of the dictionary. The compliance with the re-
382 quirements elicited by the stakeholders is also assessed.

383 6.1 Design and Implementation

384 In order to put the user in the center of the process, the de-
385 sign phase started with requirements engineering activities
386 with the stakeholders. First, based on semi-conducted discus-
387 sions, four personas [Lallemand, 2018] were created (deaf
388 user, deaf student, bilingual teacher, and sign language ex-
389 pert). This HCI good practice helped to identify the tar-
390 get users for the dictionary and the scope of their require-
391 ments. Moreover, a comparison of famous online dictionar-
392 ies or translators (e.g., Google Translate, DeepL, Microsoft
393 Bing) was conducted to confront their features with the needs
394 of the personas. This then initiated the design of low and
395 high-fidelity prototypes [Lallemand, 2018] for the dictionary.
396 Those artifacts were evaluated in a continuous collaboration
397 and validation with the four users representing each persona
398 (2 deaf students, 1 bilingual teacher, 1 sign language expert),
399 stakeholders (2 project leaders), and experts in HCI (1 UX
400 expert and 1 inclusive UX expert). Finally, as the website
401 is used by deaf people, great care has been taken to ensure
402 accessibility. Guidelines for the design of interfaces suited
403 for deaf people were searched. The web content accessibil-
404 ity guidelines (WCAG2) [Caldwell *et al.*, 2008] proposed by
405 the W3C provide some general recommendations to design
406 inclusive websites but nothing specific to the context of deaf-
407 ness. Therefore, the rest of the literature was explored and
408 examined. Among the identified works, the guidelines were
409 sometimes not the primary focus of the study or were too gen-
410 eral for our purpose. There was a need for precision, com-

411 pleteness and cohesion. The work by [André, 2022] gath-
412 ered, classified, and completed the recommendations found
413 in the literature to establish a checklist for the creation of UX
414 adapted to deafness (e.g., transforming all sound signals to
415 visual ones, using icons instead of texts). Those recommen-
416 dations were applied to the creation of our dictionary.

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

428 6.2 Requirements Assessment

429 To assess the conformity of the user requirements, a usabil-
430 ity testing [Lallemand, 2018] approach was followed. The
431 main goal was to collect qualitative data to improve the sys-
432 tem following a feedback loop mechanism. Six realistic us-
433 age scenarios mixing success and failure cases were proposed
434 to the four users. It should be noted that tester users were not
435 involved in the dataset creation, few years earlier. Those sce-
436 narios forced them to go through all the application function-
437 alities, allowing us to observe their reactions and spot their
438 difficulties. The tests were followed by a survey and a semi-
439 conducted discussion [Lallemand, 2018] to assess the feeling
440 of users about the web application. Each test session was
441 recorded by two cameras and two microphones. An observer
442 took notes on an observation grid to spot all the hesitations or
443 issues encountered by the user during the scenarios. A briefed
444 sign language interpreter assisted the test conductor when the
445 user was deaf. All the materials used during the tests were
446 translated into sign language by the interpreter.

447 The observations and remarks collected during those tests
448 showed that the users were able to execute all scenarios with-
449 out major difficulties. The success rate for the scenarios
450 ranges from 87% to 98%. The gap is explained by the vari-
451 ety of users. Indeed, it has been noticed that children took a
452 little more time, due to their distraction. In general, the first
453 scenario also lasted longer, since users were new to the appli-
454 cation. Finally, users reported that they appreciated the ease
455 of use, simplicity, guidance, and the contextualized exam-

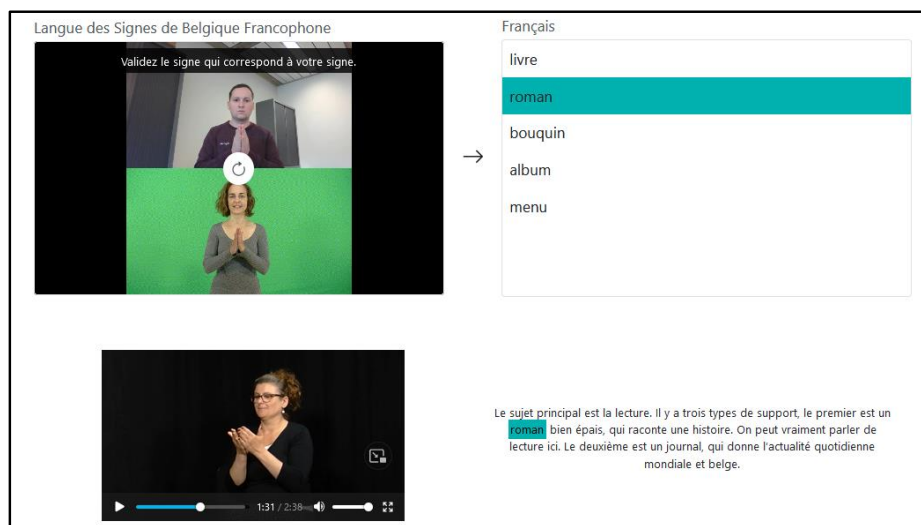


Figure 4: Screenshot of the dictionary⁶ after a successful search. The top of the interface shows the sign performed by the user along with the possible translation in French. The bottom of the interface gives contextual examples of the selected translation in sign language (video) and in French (text). Signers can hence improve themselves based on those examples.

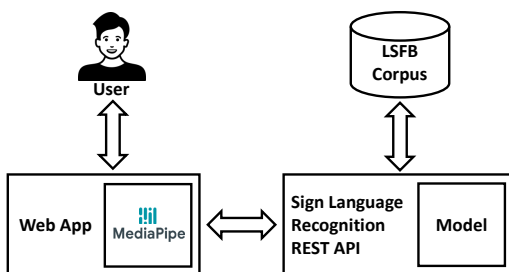


Figure 5: The system is made of three artifacts: (i) the web application that provides an interface for the user and uses MediaPipe JS to preprocess locally the captured video, (ii) an API hosting the SLR model and that is linked to (iii) the corpus database containing lexicon and contextual examples.

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.

7 Conclusion and Future Work

This work introduces the first dictionary searchable from sign language to text, publicly available through a web interface⁷. It relies on a lightweight sign language recognition model, inspired by the recent advances in transformer networks such as

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.

⁷dico.corpus-lsfb.be

503 Ethical Statement

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

512 Acknowledgments

513 We would like to thank the members of the LSFb Lab for their
514 major contribution and collaboration. Moreover, we express our
515 gratitude to the Baillet Latour Fund, the Walloon region for the
516 Ph.D. grant from FRIA (F.R.S.-FNRS) and the project ARIAC
517 piloted by Trail, an initiative of the Digital4Wallonia for their
518 funding. This work was also funded by the FWO and F.R.S.-FNRS
519 under the Excellence of Science (EOS) program.
520

521 References

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

529 [André, 2022] Maxime André. Recommandations pour des
530 interfaces utilisateurs adaptées à la surdit . Master’s thesis,
531 Universit  de Namur, 2022.

532 [Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Ge-
533 offrey E Hinton. Layer normalization. *arXiv preprint
534 arXiv:1607.06450*, 2016.

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

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

544 [De Coster *et al.*, 2020] Mathieu De Coster, Mieke Van Her-
545 reweghe, and Joni Dambre. Sign language recognition
546 with transformer networks. In *12th international
547 conference on language resources and evaluation*, pages
548 6018–6024. European Language Resources Association
549 (ELRA), 2020.

550 [Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer,
551 Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,
552 Thomas Unterthiner, Mostafa Dehghani, Matthias Minder-
553 derer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and
554 Neil Houlsby. An image is worth 16x16 words: Trans-
555 formers for image recognition at scale. In *International
556 Conference on Learning Representations*, 2021.

[Fink *et al.*, 2021] J r me Fink, Beno t Fr nay, Laurence
557 Meurant, and Anthony Cleve. Lsfb-cont and lsfb-isol:
558 Two new datasets for vision-based sign language recog-
559 nition. In *2021 International Joint Conference on Neural
560 Networks (IJCNN)*, pages 1–8, 2021. 561

[Huang and Huang, 1998] Chung-Lin Huang and Wen-Yi
562 Huang. Sign language recognition using model-based
563 tracking and a 3d hopfield neural network. *Machine Vi-
564 sion and Applications*, 10(5-6):292–307, April 1998. 565

[Izutov, 2020] Evgeny Izutov. Asl recognition with metric-
566 learning based lightweight network. *arXiv preprint
567 arXiv:2004.05054*, 2020. 568

[Kadir *et al.*, 2004] Timor Kadir, Richard Bowden, Eng-Jon
569 Ong, and Andrew Zisserman. Minimal training, large lex-
570 icon, unconstrained sign language recognition. In *BMVC*,
571 pages 1–10, 2004. 572

[Koller, 2020] Oscar Koller. Quantitative survey of the state
573 of the art in sign language recognition, 2020. 574

[Konstantinidis *et al.*, 2018] Dimitrios Konstantinidis, Kos-
575 mas Dimitropoulos, and Petros Daras. Sign language
576 recognition based on hand and body skeletal data. In *2018-
577 3DTV-Conference: The True Vision-Capture, Transmis-
578 sion and Display of 3D Video (3DTV-CON)*, pages 1–4.
579 IEEE, 2018. 580

[Lallemant, 2018] Carine Lallemant. *M thodes de De-
581 sign UX. 30 m thodes fondamentales pour concevoir des
582 exp riences optimales. (2e edition)*. 09 2018. 583

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Ge-
584 offrey Hinton. Deep learning. *Nature*, 521(7553):436–
585 444, May 2015. 586

[Li *et al.*, 2020] Dongxu Li, Cristian Rodriguez, Xin Yu, and
587 Hongdong Li. Word-level deep sign language recognition
588 from video: A new large-scale dataset and methods com-
589 parison. In *Proceedings of the IEEE/CVF Winter Confer-
590 ence on Applications of Computer Vision (WACV)*, March
591 2020. 592

[Liao *et al.*, 2019] Yanqiu Liao, Pengwen Xiong, Weidong
593 Min, Weiqiong Min, and Jiahao Lu. Dynamic sign lan-
594 guage recognition based on video sequence with blstm-3d
595 residual networks. *IEEE Access*, 7:38044–38054, 2019. 596

[Lugaresi *et al.*, 2019] Camillo Lugaresi, Jiuqiang Tang,
597 Hadon Nash, Chris McClanahan, Esha Uboweja, Michael
598 Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong,
599 Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg,
600 and Matthias Grundmann. Mediapipe: A framework for
601 building perception pipelines. 2019. 602

[Meurant, 2015] Laurence Meurant. Corpus LSFb. Corpus
603 informatis  en libre acces de vid o et d’annotations de
604 langue des signes de Belgique francophone. Namur: Lab-
605 oratoire de langue des signes de Belgique francophone
606 (LSFB Lab), FRS-FNRS, Universit  de Namur, 2015. 607

[Pigou *et al.*, 2016] Lionel Pigou, Mieke Van Herreweghe,
608 and Joni Dambre. Sign classification in sign language
609 corpora with deep neural networks. In Eleni Efthimiou,
610

- 611 Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochge-
612 sang, Jette Kristoffersen, and Johanna Mesch, editors, *Pro-*
613 *ceedings of the LREC2016 7th Workshop on the Represent-*
614 *ation and Processing of Sign Languages: Corpus Mining,*
615 pages 175–178, Portorož, Slovenia, May 2016. European
616 Language Resources Association (ELRA).
- 617 [Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, Men-
618 glong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen.
619 Mobilenetv2: Inverted residuals and linear bottlenecks. In
620 *Proceedings of the IEEE conference on computer vision*
621 *and pattern recognition*, pages 4510–4520, 2018.
- 622 [Savitzky and Golay, 1964] Abraham. Savitzky and M. J. E.
623 Golay. Smoothing and differentiation of data by sim-
624 plified least squares procedures. *Analytical Chemistry*,
625 36(8):1627–1639, July 1964.
- 626 [Stokoe, 1972] William C Stokoe. Classification and de-
627 scription of sign languages. *Current trends in linguistics*,
628 12:345–371, 1972.
- 629 [UN, 2015] UN. The 17 goals — sustainable develop-
630 ment. <https://sdgs.un.org/goals>, 2015. (Accessed on
631 05/15/2023).
- 632 [Vaezi Joze and Koller, 2019] Hamid Vaezi Joze and Oscar
633 Koller. Ms-asl: A large-scale data set and benchmark for
634 understanding american sign language. In *The British Ma-*
635 *chine Vision Conference (BMVC)*, September 2019.
- 636 [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki
637 Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,
638 Łukasz Kaiser, and Illia Polosukhin. Attention is all you
639 need. In *Advances in Neural Information Processing Sys-*
640 *tems*, volume 30, 2017.
- 641 [Wang *et al.*, 2012] Haijing Wang, Alexandra Stefan, Saj-
642 jad Moradi, Vassilis Athitsos, Carol Neidle, and Farhad
643 Kamangar. A system for large vocabulary sign search.
644 In *Trends and Topics in Computer Vision: ECCV 2010*
645 *Workshops, Heraklion, Crete, Greece, September 10-11,*
646 *2010, Revised Selected Papers, Part I 11*, pages 342–353.
647 Springer, 2012.