







# Mediapi-RGB: Enabling Technological Breakthroughs in French Sign Language (LSF) Research Through an Extensive Video-Text Corpus

Yanis Ouakrim<sup>1,2,\*</sup><sup>a</sup>, Hannah Bull<sup>1,\*</sup><sup>b</sup>, Michèle Gouiffès<sup>1</sup><sup>c</sup>, Denis Beautemps<sup>2</sup><sup>d</sup>  
Thomas Hueber<sup>2</sup><sup>e</sup> and Annelies Braffort<sup>1</sup><sup>f</sup>

<sup>1</sup>LISN, Univ. Paris-Saclay, CNRS, 91405 Orsay, France

<sup>2</sup>Univ. Grenoble Alpes, GIPSA-Lab, CNRS, F-38000 Grenoble, France

**Keywords:** Sign Language Processing, Sign Language Translation, Sign Language Corpora, French Sign Language, LSF.


**Abstract:** We introduce Mediapi-RGB, a new dataset of French Sign Language (LSF) along with the first LSF-to-French machine translation model. With 86 hours of video, it is the largest LSF corpora with translation. The corpus consists of original content in French Sign Language produced by deaf journalists, and has subtitles in written French aligned to the signing. The current release of Mediapi-RGB is available at the Ortolang corpus repository (<https://www.ortolang.fr/workspaces/mediapi-rgb>), and can be used for academic research purposes. The test and validation sets contain 13 and 7 hours of video respectively. The training set contains 66 hours of video that will be released progressively until December 2024. Additionally, the current release contains skeleton keypoints, sign temporal segmentation, spatio-temporal features and subtitles for all the videos in the train, validation and test sets, as well as a suggested vocabulary of nouns for evaluation purposes. In addition, we present the results obtained on this corpus with the first LSF-to-French translation baseline to give an overview of the possibilities offered by this corpus of unprecedented caliber for LSF. Finally, we suggest potential technological and linguistic applications for this new video-text dataset.


## 1 INTRODUCTION


In comparison to written and spoken corpora, there is a lack of sign language (SL) corpora, particularly large corpora with native or near-native signers in natural settings. Unfortunately, the existence of such large annotated corpora is a prerequisite for the automatic processing of sign languages. This is especially true for translation (Camgoz et al., 2018a) or sentence alignment (Bull et al., 2021), where deep learning models need to build an understanding of both the source sign language and the target vocal language.


To this day, there are very few sign language corpora translated into vocal languages. And when they do exist, they are often very small (Braffort,


2022) due to the fact that they are often produced in laboratories, such as (Belissen et al., 2020a; Fink et al., 2021), and are therefore expensive to acquire, translate and annotate. To face this challenge, many researchers rely on interpreted sign language data, i.e. a reverse translation from a vocal language to a sign language., e.g. Phoenix-14-T (Forster et al., 2012), SWISSTXT-WEATHER and SWISSTXT-NEWS (Camgoz et al., 2021), Content4all (Camgoz et al., 2021) and BOBSL (Albanie et al., 2021) datasets. These corpora are derived from television programs whose vocal language has been transcribed into subtitles and also interpreted into a given sign language. Researchers using these corpora are working on the assumption that the sign language interpretation retains enough information so that it is possible for a translation model to recover the subtitle text from the interpreted sign language. Moreover, there are differences between interpreted and non-interpreted language (Dayter, 2019) due to source language interference and time constraints. SL content from real-time interpretation tends to closely follow the grammatical structure of the spoken language


<sup>a</sup> <https://orcid.org/0009-0006-5147-8823>

<sup>b</sup> <https://orcid.org/0000-0002-9649-357X>

<sup>c</sup> <https://orcid.org/0000-0002-7152-4640>

<sup>d</sup> <https://orcid.org/0000-0001-9625-3018>

<sup>e</sup> <https://orcid.org/0000-0002-8296-5177>

<sup>f</sup> <https://orcid.org/0000-0003-4595-7714>

\*Equal contribution

due to strong time constraints (Leeson, 2005). There is also some evidence of differences between hearing and deaf interpreters (Stone and Russell, 2013), and more generally differences between hearing non-native signers, deaf non-native signers and deaf signers (Morford and Carlson, 2011). Therefore, the lack of representation of native or near-native deaf signers outside of laboratory conditions is also a key issue.

Datasets of non-interpreted sign language along with translation are very rare; for a recent SL translation challenge, (Müller et al., ) release a dataset with 19 hours of original Swiss-German SL content. MEDI-API-SKEL, a first French SL (LSF) dataset of 27 hours, has been released with 2D keypoints data aligned with subtitles (Bull et al., 2020a). Mediapi-**RGB** aims to complement existing corpora by providing 86 hours of video of SL content outside of laboratory conditions with a high representation of deaf native signers.

Existing translated French sign language corpora are of very limited duration (Braffort, 2022). Mediapi-**RGB** is therefore the largest translated LSF corpus, enabling to train machine translation models. Along with the presentation of the corpus we apply a simple transformer based machine translation baseline to give an overview of the translation possibilities with this new corpus following existing works on German Sign Language (DGS) (Camgoz et al., 2018b), British Sign Language (BSL) (Albanie et al., 2021) and American Sign Language (ASL) (Tarrés et al., 2023). The two main contributions of this paper are the following:

- The provision of the largest translated LSF corpus to date and one of the largest compared to other sign language corpora around the world.
- The first machine translation model from LSF video to French text.

The remainder of the paper is organized as follows. Section 2 provides an overview of the dataset and section 3 gives information about its production. Then, section 4 proposes a translation baseline and its results on Mediapi-**RGB**. Opportunities and limitations of this new dataset are discussed in Section 5.

## 2 DATASET OVERVIEW

Mediapi-**RGB** is a large corpus (86 hours) of original LSF content available for academic research purposes. This corpus can be used for numerous research applications, including training or evaluating SL retrieval, recognition or translation models.

We provide a summary of the Mediapi-**RGB**



Figure 1: **Mediapi-**RGB** source data**. The source data consists of journalistic content in LSF, along with translated French subtitles aligned to the signed content. (Subtitles are not embedded in the video pixels).

dataset and compare this corpus with the MEDI-API-SKEL dataset.

### 2.1 Dataset Content and Statistics

The source of the data is the French media association *Média'Pi!*<sup>1</sup>, the same source of the data used to create MEDI-API-SKEL (Bull et al., 2020a). *Média'Pi!*, is a generalist and independent online media launched in April 2018. It offers articles and reports on national and international news as well as topics related to the deaf community in LSF and French. It operates on a subscription basis. The quantity and quality of original content produced by *Média'Pi!* (noted episodes or original videos in the paper) is difficult to find outside of laboratory-produced corpora, and the diversity of scenarios better reflects the diversity of real-world SL videos. The LSF produced by the deaf journalists follows a natural structure. In a second stage, the LSF content is translated into written French and aligned to the LSF video in the form of subtitles. The alignment between written subtitles and LSF video is accurate. This is in contrast to interpreted data, where there is often a lag between the written subtitles (aligned to the audio) and the sign language interpretation. Figure 1 shows a screenshot of a video from the *Média'Pi!* website along with its associated subtitle in French ("Last Saturday, 19,000 doses of vaccine were released to vaccinate the entire major population of the area").

The dataset contains 1,230 videos, representing a total of 86 hours of LSF produced between 2017 and 2022. In order to protect the economic model of *Média'Pi!*, we are only able to publish videos of information and sports programmes with a time delay of three years, and these videos may only be used for academic research purposes. We have decided to use videos dating from 2017-2018 in the validation set,

<sup>1</sup><https://media-pi.fr/>

and videos from 2019 in the test set, so that models trained on other corpora can already be evaluated on Mediapi-RGB. Other videos from 2020-2022 will be progressively released in the training set until December 2024. We are nevertheless able to release keypoints data, video features, automatic sign segmentation and subtitles of the training set videos. We also propose a 4k vocabulary of common and proper nouns in the Mediapi-RGB subtitles for evaluation purposes. Table 1 gives information on the split between train, validation and test sets.

Table 1: **Train, validation and test split.** The RGB training videos will be released progressively until Dec. 2024. Features and skeleton keypoints on the training set are available.

	Train	Val	Test	Total
Fully released	No	Yes	Yes	-
# videos	950	74	206	1230
# subtitles	37651	4373	8060	50084
# hours (full video)	66.1	7.2	12.5	85.9
# hours (subtitles)	52.3	4.9	10.8	68.0

## 2.2 Relationship to MEDIAPI-SKEL

The Mediapi-RGB dataset is intended as a replacement rather than a complementary dataset to MEDIAPI-SKEL. This is because the training set of MEDIAPI-SKEL partially overlaps with the test partition of Mediapi-RGB. The current release of the training set of Mediapi-RGB contains skeleton keypoints for the train, validation and test sets, and so skeleton-based models can currently be trained and evaluated on Mediapi-RGB rather than MEDIAPI-SKEL. The main advantages of Mediapi-RGB are that: 1) it’s a bigger corpus (86 hours vs. 27 hours), 2) the original RGB videos are available on the validation and test sets, and will soon be released for the training set, and 3) additional data such as Mediapipe keypoints (Lugaresi et al., 2019), Video Swin Transformer features (Liu et al., 2022) and automatic sign temporal segmentations are available. The main disadvantage of Mediapi-RGB in comparison to MEDIAPI-SKEL is the fact that there are fewer signers, due to the exclusion of certain types of programmes including interviews. A comparison of the statistics of MEDIAPI-SKEL and Mediapi-RGB can be found in Table 2.

## 3 DATASET COLLECTION

This section describes the dataset collection, as well as various post-processing steps to facilitate usage of

Table 2: Descriptive statistics of our datasets MEDIAPI-SKEL and Mediapi-RGB.

	MEDIAPI-SKEL	Mediapi- RGB
<b>Global statistics</b>		
# subtitled episodes	368	1230
# hours	27	86
# frames	2.5 million	7.7 million
<b>Video statistics</b>		
Resolution (# vids.)	1080p (327) 720p (41)	2160p (28) 1080p (1182) 720p (18) 480p (2)
Framerate (# vids.)	30 fps (111) 25 fps (242) 24 fps (15)	25 fps
Avg. length of vid.	4.5 minutes	4.2 minutes
# signers	> 100	> 10
<b>Text statistics</b>		
# subs.	20 187	50 084
Avg. length of sub.	4.2 seconds 10.9 words	4.9 seconds 12.2 words
Vocab. (tokens)	17 428	35 599
Vocab. (nouns+verbs+adj.)	14 383	27 343

Mediapi-RGB. We temporally crop the videos using the start and end times of each subtitle (Sec. 3.1) and spatially crop the signers (Sec. 3.2). This creates video-text sentence-like pairs, with the signer centered within a square crop of  $444 \times 444$  pixels at 25fps. We remove duplicates to ensure that there are no identical videos across the train, validation and test splits (Sec. 3.3). Although we can not currently release the training set videos, we can release the corresponding processed data, including OpenPose (Cao et al., 2019) and Mediapipe Holistic (Lugaresi et al., 2019) keypoints (Sec. 3.4), Video Swin Transformer features (Sec. 3.5), sign segmentation (Sec. 3.6), original and processed subtitles (Sec. 3.7), as well as a proposed noun vocabulary for evaluation purposes (Sec. 3.8).

### 3.1 Temporally Cropping Subtitles

In Mediapi-RGB, the subtitles are aligned to the signing. Almost all of the *Média’Pi!* episodes are produced in LSF, then subsequently subtitled in written French. There are rare cases where a hearing person is interviewed in spoken language and this is interpreted into LSF. In these cases, the subtitles correspond to the original audio, but are aligned to the signing, and the audio track is removed. We temporally crop the videos using the subtitle timestamps, adding 0.5s on each side as padding. This creates sentence-

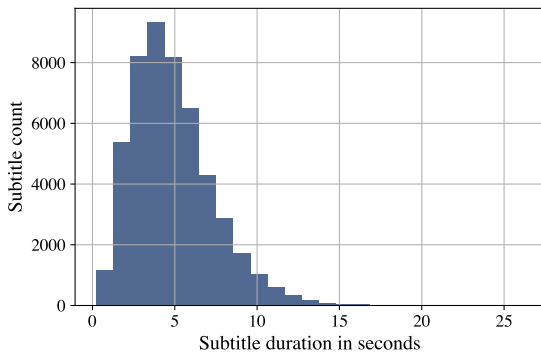


Figure 2: **Duration of subtitles.** The distribution of the durations of the extracted subtitle clips (without 0.5s padding).

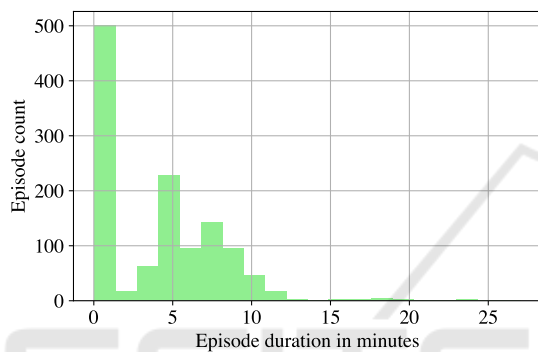


Figure 3: **Duration of episodes.** The distribution of the durations of *Média'Pi!* original episodes (without any modification).

like video-text pairs. The distribution of subtitle durations is shown in Fig. 2, and the distribution of the duration of the original episodes is shown in Fig. 3.

### 3.2 Spatially Cropping Signers

As the *Média'Pi!* videos capture signers in different positions and orientations, we automatically extract bounding boxes around the most likely signer in each of the extracted subtitle crops (Sec. 3.1). To do this, we follow the methodology described in (Bull et al., 2020b) with available online code<sup>2</sup>. We extract the 2D OpenPose (Cao et al., 2019) keypoints of each person in the videos, omit the legs and feet keypoints, track each person between consecutive frames, impute missing skeleton keypoints using past or future frames, temporally smooth keypoints using a Savitzky-Golay filter and omit unlikely signers such as people with occluded hands, people with hands that hardly move or people in the background. In the case of multiple potential signers, we then choose the most likely signer based on a metric computed by multiply-

<sup>2</sup>[https://github.com/hannahbull/clean\\_op\\_data.sl](https://github.com/hannahbull/clean_op_data.sl)

ing the hand size and the variation of wrist movement of the dominant hand. We then use a static square crop around this most likely signer for the duration of each subtitle. This square crop is then resized to  $444 \times 444$  pixels.

### 3.3 Removing Duplicate Videos

In the initial dataset, there were a large amount of duplicate videos, generally some extracts from an original video (for advertisement or quoting) or recurrent opening/closing titles. These duplications have to be removed to avoid biases. To that purpose, we select the videos that share the same subtitle and measure the spatio-visual similarity between them using the Video Swin features, described further in Sec. 3.5. Letting  $\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$  be the sequence of visual feature vectors of a video  $A$  and  $\{\mathbf{b}_1, \dots, \mathbf{b}_m\}$  be the sequence of visual feature vectors of a video  $B$ , we compute:

$$S_{A,B} = \max_{i \in 1, \dots, n, j \in 1, \dots, m} \frac{\mathbf{a}_i \cdot \mathbf{b}_j}{\|\mathbf{a}_i\| \|\mathbf{b}_j\|}. \quad (1)$$

If the maximum cosine similarity  $S_{A,B}$  in (1) is above a threshold  $T = 0.95$ , the two videos are considered to be duplicates. Once duplicated crops  $A$  or  $B$  are detected, the longer one is kept since it is more likely to correspond to the original video. Crops with a similarity between  $T = 0.85$  and  $T = 0.95$  are considered as *potential* duplicates and have therefore been grouped in the same split to avoid biases. Thresholds were set on the basis of manual observations.

### 3.4 Skeleton Keypoints Extraction

Skeleton keypoints are useful inputs for many automatic SL processing tasks, such as hand or face cropping (Huang et al., 2018; Shi et al., 2019), generating SL (Ventura et al., 2020; Saunders et al., 2020), or as inputs to improve recognition methods (Belissen et al., 2020b; Jiang et al., 2021b). Thus, OpenPose (Cao et al., 2019) and Mediapipe Holistic (Lugaresi et al., 2019) keypoints are provided for the body, the hands and the face. Fig. 4 shows an example of the face, body and hand keypoints detection with both Mediapipe Holistic and OpenPose.

### 3.5 Spatio-Temporal Features Extraction

Features are a useful input for automatic SL processing systems, because they allow models to be trained more rapidly than end-to-end systems using RGB video frames as input. For example, features are



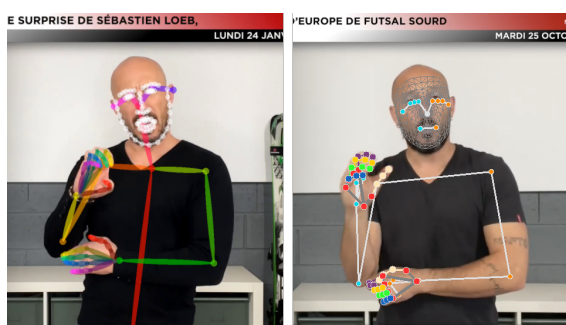


Figure 4: **Illustration of OpenPose and Mediapipe Holistic keypoints.** Illustration of the face, body and hand keypoints detection using both OpenPose (Cao et al., 2019) (left) and Mediapipe Holistic (Lugaresi et al., 2019) (right).

used directly to train an automatic alignment model in (Bull et al., 2021) and to search for lexical signs in (Momeni et al., 2022).

To extract features, we use the Video Swin transformer model (Liu et al., 2022) trained for the task of sign recognition using dense automatic annotations acquired using the methods described in (Momeni et al., 2022). Superior performance of this model over I3D models is shown in (Prajwal et al., 2022) for sign classification. The input to the Video Swin transformer model is a temporal context window of 16 frames, and the output is a vector of features of dimension 768. We extract these features at stride 1. Due to 0.5s padding applied on each of the subtitle timestamps during temporal cropping, all of the extracted videos contain at least 16 frames.

### 3.6 Sign Temporal Segmentation

In (Renz et al., 2021), the authors train a model to automatically segment signs using annotated sign glosses from BSL Corpus (Schembri et al., 2017). This model inputs I3D features from (Albanie et al., 2021) and outputs change-points between signs. The model tends to recognise changes in handshape, and thus over-segments fingerspelling. We use available online code<sup>3</sup> to segment signs in the Mediapi-RGB subtitle crops. Although trained on BSL data, this model provide valuable approximations of sign boundaries in LSF. Fig. 5 shows an illustration of the sign segmentation model on a Mediapi-RGB video. The sign segmentation model recognises transitions between signs due to visual cues such as changes in hand shape.

<sup>3</sup><https://github.com/RenzKa/sign-segmentation>

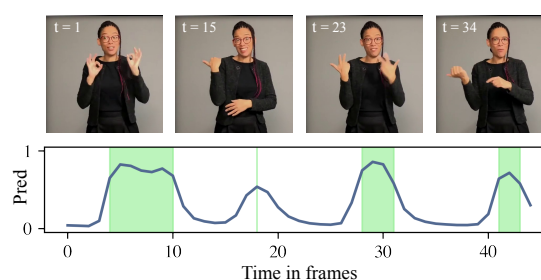


Figure 5: **Sign temporal segmentation.** Example of sign segmentation on an extract of video from Mediapi-RGB. The blue line denotes the predicted output scores and the green blocks denote a binary threshold for scores above 0.5.

### 3.7 Text Processing

We provide the raw subtitle texts for all of the videos in the train, validation and test sets. Additionally, we extract the part-of-speech of each subtitle word<sup>4</sup>. The total vocabulary size is 35,599 and the vocabulary size of nouns, verbs and adjectives is 27,343 (Tab. 2). Other parts of speech such as prepositions, determinants and adverbs do not have associated lexical signs. Linguistic elements of this type are carried by non-manual components or movement features.

### 3.8 Noun Vocabulary

We propose a vocabulary of 3,894 common and proper nouns for evaluation purposes. This vocabulary corresponds to a list of nouns from the Mediapi-RGB subtitles, appearing at least 5 times. These nouns were lemmatized using the spaCy lemmatizer trained on French news text<sup>5</sup>. Fig. 6 plots the frequency of the top 20 most commonly used nouns in this corpus. Fig. 7 shows the distribution of the number of words in each subtitle, as well as the number of words in the vocabulary in each subtitle.

## 4 TRANSLATION BASELINE

Following work conducted on other sign languages (Camgoz et al., 2020; Tarrés et al., 2023), we apply a transformer (Vaswani et al., 2017) based translation baseline to Mediapi-RGB for LSF-to-French translation. An overview of the architecture of this baseline is given in Fig. 8. On the sign language video side, we use the features obtained with a frozen Video Swin transformer model (Liu et al., 2022) trained on British Sign Language (BSL) presented in section 3.5. On the

<sup>4</sup><https://huggingface.co/gilf/french-postag-model>

<sup>5</sup>[https://spacy.io/models/fr#fr\\_core\\_news\\_md](https://spacy.io/models/fr#fr_core_news_md)

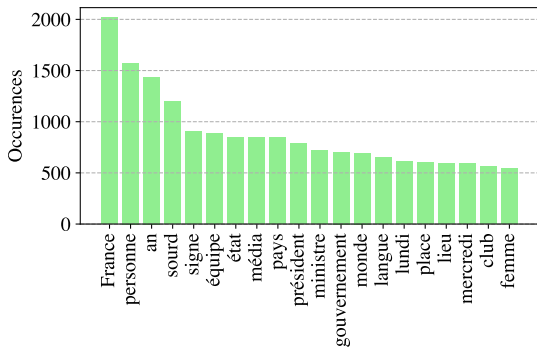


Figure 6: **Top 20 most frequent words in noun vocabulary.** The frequency of the top 20 noun occurrences.

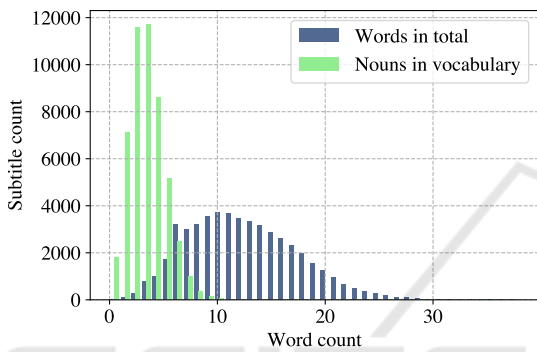


Figure 7: **Number of words in subtitles.** In blue, the distribution of the total number of words in the Mediapi-RGB subtitles; and in green, the distribution of the number of the nouns in our vocabulary in each subtitle (see Sec. 3.8 for details on this vocabulary).

text side, we use a tokenizer adapted to French<sup>6</sup>. Finally, the model is case-sensitive.

### 4.1 Implementation Details

Our architecture comprises two transformer encoder layers and one transformer decoder layer. Each transformer block has 8 attention heads. Output of the encoder block is a vector of 768 values. The fully connected layer at the end of the decoder part has a dimension of 64. We train the model with a batch size of 32 and a RMSProp optimizer with a learning rate of  $10^{-3}$ . We use learning rate scheduling: the learning rate is halved with no improvement in validation loss over 7 epochs. Similarly, we use early stopping: training stops after 27 epochs after no improvement in validation loss for 15 epochs. With this mechanism, the result shown in the table corresponds to weights obtained after the 12th epoch. The dropout is set to 0.5. We provide source code for the training and eval-

<sup>6</sup>CamemBERTTokenizer made available by Hugging face based on SentencePiece (Kudo and Richardson, 2018)

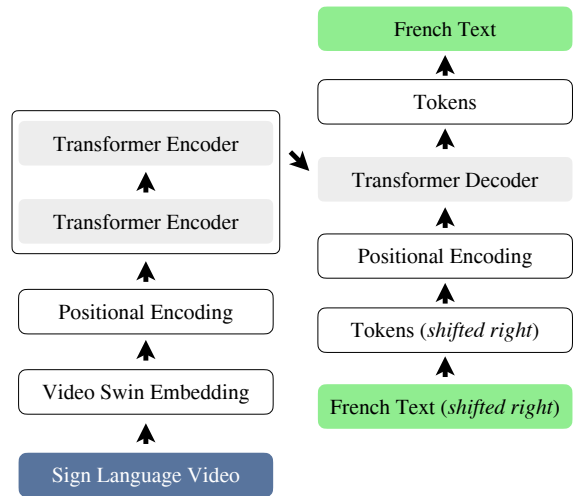


Figure 8: **Translation baseline architecture.** Based on the *Transformer* encoder-decoder architecture presented in (Vaswani et al., 2017).

uation of this model<sup>7</sup> as well as the weights of trained models. Future work could focus on hyperparameter tuning and improving this baseline architecture.

### 4.2 Evaluation and Results

Table 3: BLEU scores obtained with our translation baseline on the Mediapi-RGB test set.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Ours	21.56	10.76	6.40	4.14

Table 3 gives obtained BLEU score of the proposed baseline on the Mediapi-RGB test set. We use the standard SacreBLEU (Post, 2018) parameters for the BLEU score computation (tokenizer 13a). As BLEU scores depend on language structure, meaningful comparison can only be achieved with the same target language<sup>8</sup>. To the best of our knowledge, this work is the first to adress LSF-to-French translation, therefore, no comparison to results from other work or method is possible so far. Table 4 provides translation examples with associated ground truth along with English translations obtained with online translator DeepL<sup>9</sup>. We can note that the model has learnt semantic links: predictions mostly employ words from

<sup>7</sup>[https://github.com/youakrim/sign\\_language\\_translation\\_Mediapi-RGB](https://github.com/youakrim/sign_language_translation_Mediapi-RGB)

<sup>8</sup>English translations given purely for information purposes in Table 4, show this strong link between BLEU score and target language. For the four given examples, the obtained BLEU score if the target language was English would have been 11.75 instead of 0.0.

<sup>9</sup>[www.deepl.com](http://www.deepl.com)

Table 4: **Prediction examples with associated ground truth.** Text in brackets corresponds to English translations obtained using online translator DeepL. These examples all obtained a BLEU-4 score of 0.

GT	La Première ministre britannique Theresa May a annoncé sa démission vendredi, (British Prime Minister Theresa May announced her resignation on Friday.)
Pred	Le Premier ministre britannique a annoncé vendredi, le Premier ministre britannique, Jean Castex, (The British Prime Minister announced on Friday, the British Prime Minister, Jean Castex.)
GT	Bonjour, nous sommes le mercredi 11 mars 2019. (Hello, today is Wednesday, March 11, 2019.)
Pred	Bonjour! Nous sommes le lundi 11 mars 2019, (Hello! Today is Monday, March 11, 2019.)
GT	Ehpad : intoxication alimentaire mortelle. (French nursing home: fatal food poisoning.)
Pred	L'assurance-nil est un peu plus de retraite. (The insurance-nil is a little more retirement.)
GT	fait de nouveau l'objet d'une enquête, ouverte par le parquet de Paris. (is once again under investigation by the Paris public prosecutor's office.)
Pred	Le procès s'est ouvert le nouveau, le nouveau, en justice de Paris. (The trial opened on the new, new, Paris court.)

the same lexical field than the corresponding ground truth.

## 5 OPPORTUNITIES AND LIMITATIONS

In this section, we discuss opportunities and limitations of the Mediapi-RGB dataset for technological applications (Sec. 5.1) and for research in SL linguistics (Sec. 5.2).

### 5.1 Applications Perspective

We note that there are historical examples of SL technological applications with little consultation of the deaf community nor practical value (Bragg et al., 2019; Erard, 2017). Research projects using Mediapi-RGB should take into consideration input from deaf researchers and members of the deaf community in order to ensure practical benefits and to avoid harm. *Média'Pi!* is a project created by deaf journalists to increase access to information in the deaf commu-

nity. In that regard, we note three potential applications that could take advantage of Mediapi-RGB: anonymization, sign retrieval and automatic translation/subtitling.

One key characteristic of written language is the ability to share information anonymously. It is onerous to present information in SL without also representing the identity of the signer. Solutions such as computer-generated avatars are difficult to implement. MOCAP methods for animating SL avatars are not necessarily anonymous, as individuals can be recognised by their movements and signing style (Bigand et al., 2019). In (Bigand et al., 2022), the authors discuss methods to remove identity cues from SL production. Such methods could be combined with techniques of realistic sign generation (Saunders et al., 2020; Ventura et al., 2020). Anonymous representations of SL can then be used to communicate factual information such as legal or administrative information, governmental documents or weather reports, or to represent signers who wish to conceal their identity.

Search engines are very efficient for finding written information based on textual queries. Searching for information in SL using text queries or sign queries is very difficult due to the challenges of recognising signs and clustering similar topics in SL videos without written translations. In (Duarte et al., 2022), the authors present a method for searching for SL sequences using free-form textual queries. In (Jiang et al., 2021a), the authors search for isolated signs in continuous SL video on BSL Corpus (Schembri et al., 2017) and PHOENIX14-T (Forster et al., 2012; Forster et al., 2014). The Mediapi-RGB corpus can be used to train and evaluate models for retrieval tasks on continuous SL videos using textual or sign queries, as the subtitles of the videos may be used as weak annotation.

Adding subtitles to SL video improves accessibility and comprehension, but is a time-consuming task. Various ways to simplify this task are automatic segmentation into subtitle-units (Bull et al., 2020b), automatic alignment of subtitles to video (Bull et al., 2021), and eventually automatic translation of SL video to text. These tasks can be trained and evaluated using the videos in the Mediapi-RGB corpus as demonstrated in section 4.

Users of Mediapi-RGB should be aware of the specificities of this dataset. Models trained on other datasets may not necessarily perform well on Mediapi-RGB, and models trained on Mediapi-RGB may not generalise well to other situations. The *Média'Pi!* videos are of professional quality and are unlikely to be representative of spontaneous conversations in daily life. This can be considered both

an advantage and a limitation of Mediapi-RGB. The signers are highly skilled deaf journalists producing examples of eloquent and formal LSF, but models trained on Mediapi-RGB may be less applicable to signers with lower levels of fluency or a more informal register.

## 5.2 A Sign Linguistics Perspective

One common critique of current large SL corpora for automatic SL processing is the over-representation of interpreted TV data (Bragg et al., 2019). Nevertheless, the differences between interpreted SL data and non-interpreted SL data are not well understood. Comparing Mediapi-RGB with the other available journalistic SL content from interpreted journalistic data sources such as SWISSTXT-NEWS (Camgoz et al., 2021), VRT-NEWS (Camgoz et al., 2021) and BOBSL (Albanie et al., 2021) may provide some clues on how SL production from deaf journalists differs from interpreted journalistic content from both hearing and deaf interpreters. Increased linguistic knowledge about this differences would help acknowledge and alleviate the biases of models trained on interpreted SL data.

The number of signers is lower in Mediapi-RGB than in MEDIAPI-SKEL, due to the removal of numerous videos involving interviews with members of the public at events. Nevertheless, there are over 10 different signers in Mediapi-RGB, allowing for linguistic comparisons across signers.

## 6 CONCLUSION

We presented a summary of the new dataset Mediapi-RGB, a large dataset with 86 hours of content from deaf journalists at the French online media *Média’Pi!*. The videos are in LSF with aligned subtitles in written French. Mediapi-RGB can be used for academic research purpose. We presented a first LSF-to-French translation baseline, showing that Mediapi-RGB is suitable for this research task. This is a stepping stone towards future improvements. We hope that Mediapi-RGB can be used to train and evaluate more automatic SL processing tasks, as well as contribute to research in SL linguistics.

## ACKNOWLEDGEMENTS

We thank *Média’Pi !* for providing the data. This work has been funded by the Bpifrance investment “Structuring Projects for Competitiveness” (PSPC), as part of the Serveur Gestuel project

(IVès and 4Dviews Companies, LISN — University Paris-Saclay, and Gipsa-Lab — University Grenoble Alpes).

## REFERENCES

- Albanie, S., Varol, G., Momeni, L., Afouras, T., Bull, H., Chowdhury, H., Fox, N., Cooper, R., McParland, A., Woll, B., and Zisserman, A. (2021). BOBSL: BBC-Oxford British Sign Language Dataset. *arXiv preprint arXiv:2111.03635*.
- Belissen, V., Braffort, A., and Gouiffès, M. (2020a). Dicta-Sign-LSF-v2: Remake of a continuous french sign language dialogue corpus and a first baseline for automatic sign language processing. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC’20)*, pages 6040–6048, Marseille, France. European Language Resource Association (ELRA).
- Belissen, V., Braffort, A., and Gouiffès, M. (2020b). Experimenting the automatic recognition of non-conventionalized units in sign language. *Algorithms*, 13(12):310.
- Bigand, F., Prigent, E., and Braffort, A. (2019). Retrieving human traits from gesture in sign language: The example of gestural identity. In *Proceedings of the 6th International Conference on Movement and Computing*, pages 1–4.
- Bigand, F., Prigent, E., and Braffort, A. (2022). Synthesis for the kinematic control of identity in sign language. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 1–6.
- Braffort, A. (2022). Langue des Signes Française : Etat des lieux des ressources linguistiques et des traitements automatiques. In Becerra, L., Favre, B., Gardent, C., and Parmentier, Y., editors, *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, pages 131–138, Marseille, France. CNRS.
- Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.
- Bull, H., Afouras, T., Varol, G., Albanie, S., Momeni, L., and Zisserman, A. (2021). Aligning subtitles in sign language videos. In *ICCV*.
- Bull, H., Braffort, A., and Gouiffès, M. (2020a). MEDIAPI-SKEL - a 2D-skeleton video database of french sign language with aligned french subtitles. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC’20)*, pages 6063–6068, Marseille, France. European Language Resource Association (ELRA).



- Bull, H., Gouiffès, M., and Braffort, A. (2020b). Automatic segmentation of sign language into subtitle-units. In *ECCVW, Sign Language Recognition, Translation and Production (SLRTP)*.
- Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018a). Neural sign language translation. In *CVPR*.
- Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018b). Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.
- Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Camgoz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., and Bowden, R. (2021). Content4all open research sign language translation datasets. *arXiv preprint arXiv:2105.02351*.
- Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.
- Dayter, D. (2019). Collocations in non-interpreted and simultaneously interpreted english: a corpus study. In *New empirical perspectives on translation and interpreting*, pages 67–91. Routledge.
- Duarte, A., Albanie, S., Giró-i Nieto, X., and Varol, G. (2022). Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14094–14104.
- Erard, M. (2017). Why sign-language gloves don't help deaf people. *The Atlantic*, <https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/>.
- Fink, J., Frénay, B., Meurant, L., and Cleve, A. (2021). Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition.
- Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J. H., and Ney, H. (2012). Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).
- Forster, J., Schmidt, C., Koller, O., Bellgardt, M., and Ney, H. (2014). Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916.
- Huang, J., Zhou, W., Zhang, Q., Li, H., and Li, W. (2018). Video-based sign language recognition without temporal segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Jiang, T., Camgoz, N. C., and Bowden, R. (2021a). Looking for the signs: Identifying isolated sign instances in continuous video footage. *IEEE International Conference on Automatic Face and Gesture Recognition*.
- Jiang, T., Camgoz, N. C., and Bowden, R. (2021b). Skeletor: Skeletal transformers for robust body-pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.
- Leeson, L. (2005). Making the effort in simultaneous interpreting. In *Topics in Signed Language Interpreting: Theory and Practice*, volume 63, chapter 3, pages 51–68. John Benjamins Publishing.
- Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. (2022). Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211.
- Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C., Yong, M. G., Lee, J., Chang, W., Hua, W., Georg, M., and Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172.
- Momeni, L., Bull, H., Prajwal, K., Albanie, S., Varol, G., and Zisserman, A. (2022). Automatic dense annotation of large-vocabulary sign language videos. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 671–690. Springer.
- Morford, J. P. and Carlson, M. L. (2011). Sign perception and recognition in non-native signers of asl. *Language learning and development*, 7(2):149–168.
- Müller, M., Ebling, S., Avramidis, E., Battisti, A., Berger, M., Zurich, H., Bowden, R., Braffort, A., Camgöz, N. C., España-Bonet, C., et al. Findings of the first wmt shared task on sign language translation.
- Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Prajwal, K., Bull, H., Momeni, L., Albanie, S., Varol, G., and Zisserman, A. (2022). Weakly-supervised fingerspelling recognition in british sign language. In *British Machine Vision Conference (BMVC) 2022*.
- Renz, K., Stache, N. C., Fox, N., Varol, G., and Albanie, S. (2021). Sign segmentation with changepoint-modulated pseudo-labelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3403–3412.
- Saunders, B., Camgoz, N. C., and Bowden, R. (2020). Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*.
- Schembri, A., Fenlon, J., Rentelis, R., and Cormier, K. (2017). British Sign Language corpus project: A corpus of digital video data and annotations of

- British Sign Language 2008–2017 (Third Edition). <http://www.bsllcorpusproject.org>.
- Shi, B., Rio, A. M. D., Keane, J., Brentari, D., Shakhnarovich, G., and Livescu, K. (2019). Fingerspelling recognition in the wild with iterative visual attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5400–5409.
- Stone, C. and Russell, D. (2013). Interpreting in international sign: Decisions of deaf and non-deaf interpreters.
- Tarrés, L., Gállego, G. I., Duarte, A., Torres, J., and Giró-i Nieto, X. (2023). Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5624–5634.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Ventura, L., Duarte, A., and Giró-i Nieto, X. (2020). Can everybody sign now? exploring sign language video generation from 2d poses. *arXiv preprint arXiv:2012.10941*.

