Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6040–6048
Marseille, 11–16 May 2020

Dicta-Sign–LSF–v2: Remake of a Continuous French Sign Language Dialogue Corpus and a First Baseline for Automatic Sign Language Processing

Valentin Belissen[1], Annelies Braffort[2], Michèle Gouiffès[3]
[1, 2, 3]LIMSI–CNRS, [1, 3]Université Paris Sud, [1, 3]Université Paris Saclay
[1, 2, 3]LIMSI, Campus universitaire 507, Rue du Belvedère, 91405 Orsay - France
{valentin.belissen, annelies.braffort, michele.gouiffes}@limsi.fr

Abstract

While the research in automatic Sign Language Processing (SLP) is growing, it has been almost exclusively focused on recognizing lexical signs, whether isolated or within continuous SL production. However, Sign Languages include many other gestural units like iconic structures, which need to be recognized in order to go towards a true SL understanding.
In this paper, we propose a newer version of the publicly available SL corpus Dicta-Sign, limited to its French Sign Language part. Involving 16 different signers, this dialogue corpus was produced with very few constraints on the style and content. It includes lexical and non-lexical annotations over 11 hours of video recording, with 35000 manual units.
With the aim of stimulating research in SL understanding, we also provide a baseline for the recognition of lexical signs and non-lexical structures on this corpus. A very compact modeling of a signer is built and a Convolutional-Recurrent Neural Network is trained and tested on Dicta-Sign–LSF–v2, with state-of-the-art results, including the ability to detect iconicity in SL production.

Keywords: Sign Language Recognition, Sign Language Processing, Iconicity, Depicting Signs, French Sign Language

1. Introduction

A great number of sign languages have been naturally developed within Deaf communities, and even the most widely-spoken are still to be thoroughly described in terms of linguistics (Braffort, 2016). They all have their own number of very conventionalized signs as well as highly iconic structures, such as classifiers constructions (Liddell, 1977) used to depict common entities (see Fig. 1). While the research in automatic Sign Language Processing (SLP) has focused on recognizing lexical signs which are by definition very conventionalized, iconicity is usually ignored even though it is crucial for SL understanding.
In this paper, three major contributions to the field of continuous SLP are presented:

1. A public remake of the French Sign Language (LSF) part of the Dicta-Sign Corpus (Matthes et al., 2012), with cleaned and reliable annotations. These annotations include lexical data and more refined linguistic categories. As the corpus is based on dialogue with very loose elicitation guidelines, it is highly representative of natural SL.
2. The manufacturing of a relevant modeling of a signer as a proper input to a convolutional-recurrent neural network (CRNN) for automatic SLP. This modeling is made compact, and easily generalizable to any SL recording. Preprocessed data is also made public;
3. We prove the relevance of our modeling and network on Dicta-Sign–LSF–v2. State-of-the-art accuracy for the classification of video frames into a number of manual unit types is also attained on NCSLGR, a similar – although elicited with more constraints – American Sign Language corpus.

This paper is organized as follows: in Section 2, the linguistics of SLs is discussed and shortcomings of current research are pointed out, then different categories of datasets and associated research in SLP are analyzed. Section 3 introduces Dicta-Sign–LSF–v2: its elicitation and recording conditions, its annotations and key statistics. In Section 4, a first baseline method for automatic SLP is laid out and experimented on Dicta-Sign–LSF–v2 in Section 5. In the same section, the learning model is also tested on another SL corpus for a comparison with existing results.

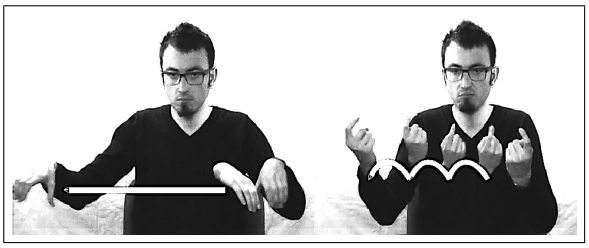


Figure 1: Example of complex linguistic use of space to represent a check-in counter in an airport. The size and location of the counter (left) constraints the placement and orientation of the classifier “persons seated” (right).

2. Related work and limitations

2.1. Sign Language Linguistics

In this section, some fundamental aspects of SL linguistics are discussed.
Conversely to written English for instance, in which every word is conventionalized and the grammar is structured by the use of some specific words, SLs are not fully described by their lexicon. Lexical signs do exist, and can be defined as follows, from Johnston and Schembri (2007): *"fully-lexical signs are highly conventionalised signs in both form and meaning in the sense that both are relatively stable or consistent across contexts. Fully-lexical signs can easily be listed in a dictionary".*
However, the hypothesis – usually used but unstated in the field of SLP – that SL production can be reduced to a se-

quence of lexical signs is wrong. Indeed, some characteristics of SLs make them fundamentally different from unidimensional sequential languages:

- They are **multi-channel**: information is conveyed through hand motion, shape and orientation, body posture and motion, facial expression and gaze;
- They are strongly **spatially organized**: events, objects, people and other entities are placed in the signing space and related to each other in a visual way. The grammar of SLs is structured by the use of space;
- They allow signers to generate new signs – that would not appear in a dictionary – on the go, in an iconic way, or even modify lexical signs. More generally, **SL do not only consist of lexical signs** but they also make use of more complex iconic structures: classifiers, pointing signs, buoys (see Section 3.2 for more detail). As one can appreciate on the random example of Fig. 1, a SL utterance can be completely iconic.

Thus, SL production should not be seen and analyzed as a succession of lexical signs. Classifiers and other gestural units like pointing – used to link entities in a SL production – are crucial in the visual grammar of SLs.

2.2. Sign Language Datasets and Recognition

One of the most basic SLP tasks consists in the recognition of isolated lexical signs. The following datasets have been used for this task, and are all publicly available:

- ASLLVD (Neidle et al., 2012) consists of 2284 ASL isolated lexical signs, realized by 6 signers, with a sign count of 9800. Different views are provided, with RGB recording at 60 fps.
- Devisign-L (Chai et al., 2015) covers 2000 CSL (Chinese Sign Language) lexical signs, realized by 8 signers, with a sign count of 24000. Videos are recorded frontally, with RGBD and skeleton data provided.
- MS-ASL (Joze and Koller, 2018) is a collection of 1000 isolated lexical signs from YouTube videos, with 222 signers and a total sign count of 25513.

Continuous Sign Language Recognition (CSLR) is a more challenging task. Many datasets have been made public, although most of them only include lexical annotations, and are quite artificial in the sense that they consist of simple elicited sentences:

- Purdue RVL-SLLL (Wilbur and Kak, 2006) results from the elicitation of pre-defined "paragraphs" (1-2 sentences) from 14 signers, each of them realizing 10 so-called paragraphs. RGB data is provided, and 104 lexical signs are annotated.
- CSL-25k (Huang et al., 2018) results from the elicitation of pre-defined sentences from 50 signers, with 178 annotated CSL lexical signs and more than 100 hours of RGBD video. Skeleton data is also given.
- Signum (Von Agris and Kraiss, 2007) is a DGS (German Sign Language) dataset, with 5 hours of RGB video from 25 signers. Pre-defined sentences are elicited, with 465 lexical signs.
- RWTH Phoenix (Forster et al., 2014) is made from 11 hours of live DGS interpretation of weather forecast on German TV. It has established itself as a reference dataset for SLR (Koller et al., 2018) even though the language variability and complexity are quite limited. Furthermore, one should note that interpreted SL is different from spontaneous SL: there is a good chance that the translation will be strongly influenced by the original speech (in German), especially in terms of syntax, and make little use of the structures typical of SL (Section 2.1). Also, RGB frame resolution is low at 210×260. A signer-independent version was released in Koller et al. (2017).

At the other end of the spectrum, some datasets are – at least partially – made of natural sign language, often in the form of conversations or narratives, and include annotations other than lexical. For instance:

- Corpus NGT (Zwitserlood et al., 2008, Sign Language of the Netherlands, 72 hours)
- BSL Corpus Project (Schembri, 2008, British Sign Language)
- DGS Korpus (Prillwitz et al., 2008, 50 hours)
- Auslan Corpus (Johnston, 2009, Australian Sign Language, 300 hours)
- NCSLGR (Neidle and Vogler, 2012, ASL, 2 hours)
- Corpus LSFB (Meurant et al., 2016, French Belgian Sign Language, 150 hours)

are continuous SL RGB datasets and include annotations like "Depicting signs", "Pointing signs" and "Buoys" (cf. Section. 3.2). However, annotations are often incomplete and these datasets are usually ignored in the field of automatic SLR. To the authors knowledge, there is only one experiment of extra-lexical automatic Sign Language Processing on a continuous SL dataset in the literature (Yanovich et al., 2016, see Section 5.1 for more detail).
Therefore, a continuous French Sign Language corpus, including very natural conversations on a wide-ranging subject, annotated on lexical *and* extra-lexical levels, is presented in the next section and will be used to evaluate SLP models.

3. Dicta-Sign–LSF–v2

This section presents a remake of the LSF part of the Dicta-Sign Corpus (Matthes et al., 2012).

3.1. Elicitation and recording conditions

Dicta-Sign–LSF–v2 is a dialogue corpus: two signers face each other, with one camera above each of them, and a third camera on the side. In order to ensure consistency, these three views are released in Dicta-Sign–LSF–v2, with identical resolution (720×575 at 25 fps). Other views or better

Figure 2: Dicta-Sign–LSF–v2: setup for the recording of two signers

resolution may exist in the original recording, but only for part of the dataset. The setup can be seen on Fig. 2.
Eight couples were formed, and each couple was given nine different tasks, on the common theme of "travel". Since elicitation was quite loose, conversation was very direct, with pauses, laughter, hesitance etc. occurring naturally.

3.2. Annotations

Focusing on the manual activity, the corpus was annotated within three main annotation categories, according to the classification of Johnston and De Beuzeville (2014): Fully-lexical signs (FLS), Partially-lexical signs (PLS) and Non-lexical signs (NLS). Other structures like constructed action/dialogue[1] (or role shift) were not annotated, or only partially.
An illustration sequence is shown on Fig. 3, with more detail on the annotations in associated Table 1.

3.2.1. Fully Lexical Signs (FLS)

FLS form the basic lexicon of SL, as defined in Section 2.1. They only account for a fraction of what can be analyzed in SL production.

3.2.2. Partially Lexical Signs (PLS)

PLS are annotated into three sub-categories:

- *Pointing signs (PT)*, as mentioned in the name, are used to point towards an entity in the signing space, that is to link what is said to a spatial referent. Since SL are spatially organized, they are of prime importance to understand a discourse.
- *Depicting Signs (DS)* form a broad category of signs, the structure of which is easily identified. They are also referred to as classifier signs, classifier constructions or classifier predicates (see Liddell (1977)). Their definition is close to what is called *highly iconic signs* in Cuxac (2000). They are used to describe the location, motion, size, shape or the action of an entity, along with trajectories in the signing space. They sometimes consist in the tweaking/enrichment of a lexical sign.
- *Fragment buoys* are hand shapes held in the signing space (usually on the weak hand) while signing activity continues on the other hand (Liddell, 2003). They can be seen as a referent, and can be used for specific linguistic functions, like what was called qualification/naming structures in Filhol and Braffort (2012).

3.2.3. Non Lexical Signs (NLS)

NLS comprise fingerspelling (FS), numbering (N) and gestures that are not typically specific to SL and can be culturally shared with non SL signers (*i.e.* speakers).

3.3. Statistics

There are 16 signers in the dataset, with 94 fully annotated videos – one for each signer and each task. In total, 11 hours are annotated, that is 1007593 frames. However, one should note that since each signer is recorded continuously during dialogue, approximately half of these 11 hours correspond to one person looking at the other person signing.
Table 2 presents the detailed frame and manual unit count for each of the main annotated categories, along with their cumulative distribution. Fully-lexical signs, depicting signs and pointing signs account for 93.6% of non blank frames, and 97.6% of all manual units.

3.4. Public access

The whole dataset is made public at `https://www.ortolang.fr/market/item/dicta-sign-lsf-v2` (Belissen et al., 2019), with framewise annotations in `csv` format.

4. Baseline SLP method

In this section, a first Sign Language Processing model is presented. It is based upon a compact and generalizable modeling of a signer, and can be used to learn the detection and recognition of different types of linguistic annotations.

4.1. Generalizable signer modeling

In this section, the manufacturing of relevant upper body pose, face and hand features is described.

4.1.1. Body pose

While previous methods were usually based on optical flow and skin color detection (Gonzalez Preciado, 2012), Convolutional Neural Networks (CNNs) have emerged as a very effective tool to get relevant features from images. OpenPose (Cao et al., 2017) is a powerful open source library, with real-time capability for estimating 2D body, face and hand pose.
Since SL are 3-dimensional, we developed a deep neural network, reproducing the architecture from Zhao et al. (2018), in order to get an estimate on the 3D upper body pose from the 2D upper body pose. This 2D-to-3D model was trained on motion capture data from the LSF corpus MOCAP1 (Benchiheub et al., 2016), only on upper body pose. Finally, body size is normalized in order to increase model generalizability. Instead of raw 3D upper body pose, a few meaningful features are pre-computed – the final body feature vector is of size 160:

[1] From Johnston and De Beuzeville (2014): *enactment of the external physical actions or behaviour of a character*

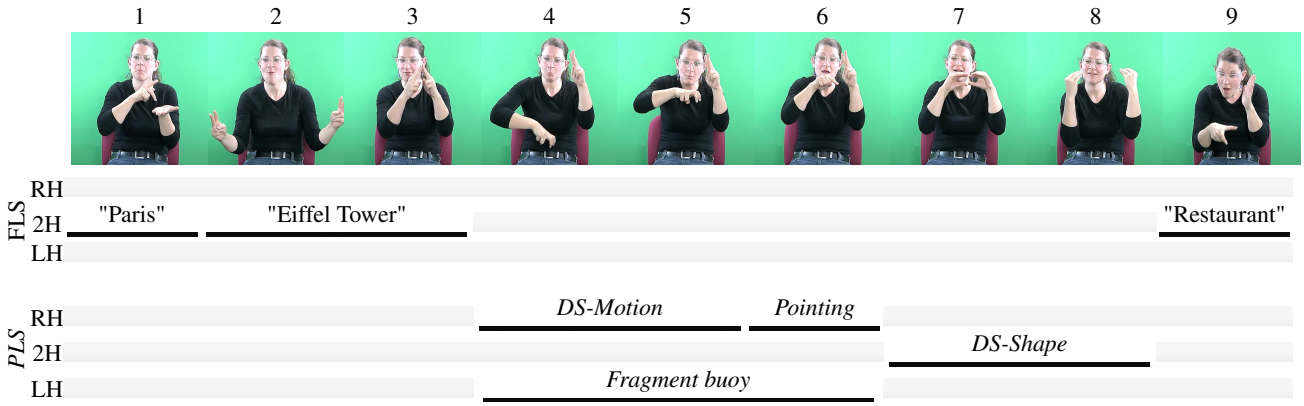


Figure 3: French Sign Language sequence (duration: 4 seconds) from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`). Annotations for right-handed (RH), two-handed (2H) and left-handed (LH) Fully Lexical Signs (FLS) and Partially Lexical Signs (PLS) are given. Non manual activity can be observed but is not annotated.
Possible translation: *In Paris, if you climb the Eiffel Tower, you will find a square-shaped restaurant at the middle floor.*

Frame	Linguistic analysis of the manual activity
1	**Lexical sign** "Paris"
2, 3	**Lexical sign** "Eiffel Tower"
4, 5	The left hand is being used as a ***fragment buoy*** – here a fragment of the tower –, which is a non-lexical SL function helping the interlocutor understand that what is being said still relates to the same scene. The right hand has a typical hand shape for the legs of a person – known as a ***proform*** – and depicts a straight motion from the bottom to the middle of the tower, indicating the action of using the elevator. It is annotated as a ***depicting sign*** of motion type.
7	***Pointing sign*** to a precise location, at the middle of the tower. It indicates the location of what is going to be introduced. The left hand is still used as a ***fragment buoy***.
8, 9	Both hands are used to depict an outer shape – thus annotated as ***depicting sign*** of shape type. Its base is a square, and it is rather slim (which is stressed by the crinkled eyes, even though not in the scope of the annotations).
10	**Lexical sign** "Restaurant". The location and shape that have just been described apply to it.

Table 1: This table is a linguistic description of the manual activity in the SL sequence shown on Fig. 3. It points out many key elements necessary to understand this sequence – namely *lexicon*, *buoys*, *proforms*, *pointing*, *iconic structures* and *spatial structure*.

- Position, speed and acceleration of: each hand w.r.t. body center, parent elbow and face center; one hand w.r.t. other hand; each elbow w.r.t. parent shoulder; each shoulder w.r.t. body center;
- Angle and orientation of elbows/shoulders, and $1^{st}/2^{nd}$ order derivatives.

4.1.2. Head pose

A 3D face estimate is directly obtained from video frames thanks to a CNN model trained on $230,000$ images (Bulat and Tzimiropoulos, 2017). Handcrafted features are used instead of raw data – the head feature vector is of size 16:

- Euler angles, speed and acceleration for axes X, Y and Z of the centroid of the head;
- Horizontal and vertical mouth openness; relative motion of the eyebrows to the eyes;
- Nose to body center distance.

4.1.3. Hand modeling

A lot of information in SL production, if not most of it, is conveyed through the hands. More specifically, the location, shape and orientation of both hands are critical, along with the dynamics of these three variables, that is: hand trajectory, shape deformation and hand rotation.

Ideally, one would greatly benefit from a frame-wise 3D hand pose estimate on RGB images. Although such algorithms have been developed (Xiang et al., 2018), they have not proven reliable on real-life 25 *fps* SL videos – OpenPose, which also provides a *2D* estimate on hand pose, faces the same issue. Indeed, because hands move constantly and relatively fast in SL, motion blur makes the frame-wise estimation of hand pose very difficult. However, a SL-specific model, Deep Hand, was developed in Koller et al. (2016). This CNN model classifies cropped hand images into 61 predefined hand shapes classes. Although this model focuses on hand shape and ignores some information like hand orientation, it is still very valuable. Thus, for each frame and each hand, hand data can be

	Fully-lexical	Depicting	Pointing	Frag. buoys	Numbering	Fingerspelling	Total
Non blank frames	205530	60794	23045	14359	3830	1941	309499
%	66.4%	19.7%	7.5%	4.6%	1.2%	0.6%	
Cumulative %	66.4%	86.1%	93.6%	98.2%	99.4%	100.0%	
Manual units	24939	5289	3899	592	156	122	34997
%	71.3%	15.1%	11.2%	1.7%	0.4%	0.3%	
Cumulative %	71.3%	86.4%	97.6%	99.3%	99.7%	100.0%	
Avg. frames/unit	8.2	11.5	5.9	24.3	24.6	15.9	

Table 2: Frame and sign (manual unit) statistics for the main annotation categories of Dicta-Sign–LSF–v2.

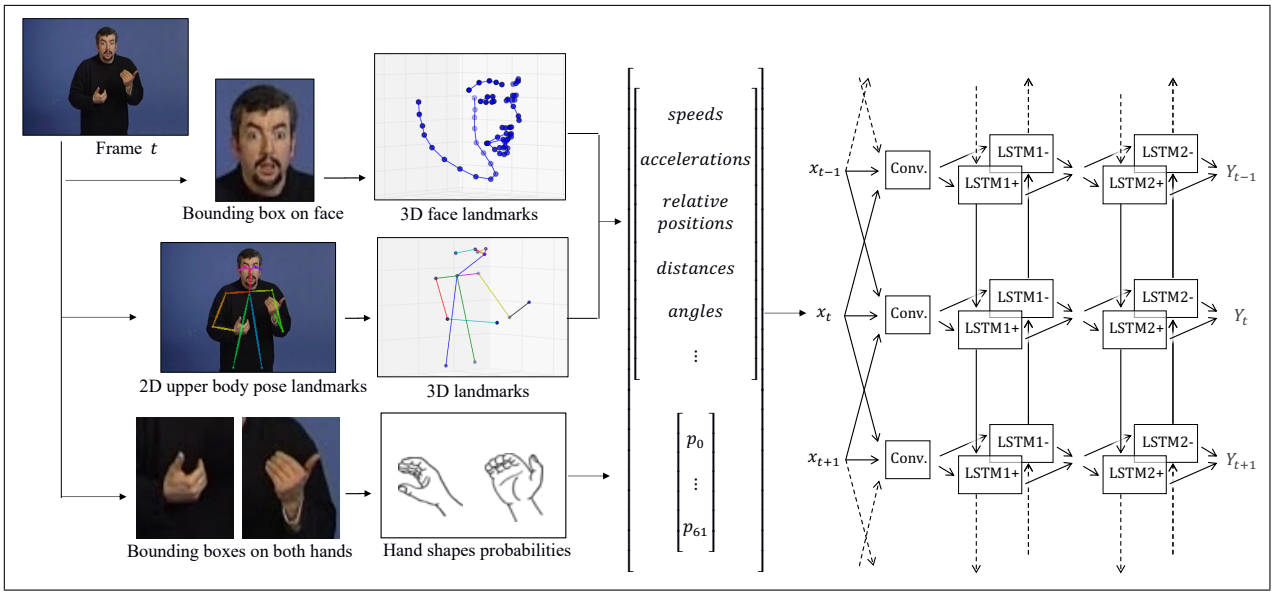


Figure 4: A compact and generalizable signer modeling (see Section 4.1). 3D-features and hand shape data are fed into a CRNN for SL linguistic features learning (see Section 4.2).

scaled down to a vector of size 61 – the final hand feature vector is of size 122.

4.1.4. Public access

Along with video files, metadata and annotation data, all previously defined preprocessed body, face and hand data is publicly accessible at `https://www.ortolang.fr/market/item/dicta-sign-lsf-v2` (Belissen et al., 2019). Some more detail on the computation of this data is included.

4.2. Learning network

This section outlines the architecture of the CRNN that was built to learn a generalizable gesture representation, in order to learn the detection and recognition of different SL features.

With time $t = 1 \dots T$, the input and output sequences are defined as follows: x_t as a flattened input vector, its size N_f^{in} corresponding to the total number of pre-computed motion features (computed from body pose and facial landmarks) and hand features (see Section 4.1); Y_t as the output of the model – it consists of N_f^{out} predictions, one for each output type of the model. The model then learns the frame-wise conditional probabilities

$$f_t^j\left((x_t)_{t=1\dots T}\right) = \mathbb{P}\left(Y_t^j \,|\, (x_t)_{t=1\dots T}\right) \qquad (1)$$

with framewise *categorical cross-entropy* as training loss.

The architecture of the model starts with a convolutional layer applied on the input x_t, and helps build a first set of temporal features. Then, recurrent Long Short-Term Memory (LSTM) layers are added to the network – LSTM units handle vanishing gradient issues (Gers et al., 2000), which in the case of high frequency data is critical. Since this work does not target real-time applications, the recurrent layers are bidirectional.

Dropout is used to prevent overfitting in the LSTM layers (Srivastava et al., 2014). Output layers use softmax activation, while other layers use Rectified Linear Unit (ReLU). RMSProp optimizer is used, and the network is built with Keras (Chollet and others, 2015) on top of Tensorflow (Abadi et al., 2016).

5. Experiments

In this section, the learning model is tested on Dicta-Sign–LSF–v2. When possible, results are also compared with another learning model and dataset (see Section 5.1).

5.1. Manual unit classification

5.1.1. Comparison with a unique reference point

A first experiment consists in the classification of video frames into a small number of types of manual units. Yanovich et al. (2016) developed a multiple instance learning (MIL) model of this kind, and trained and tested it on the NCSLGR ASL corpus (Neidle and Vogler, 2012) – in a signer-dependent fashion. The model is trained on video frames that are classified into three main categories – lexical signs, fingerspelling, depicting signs. For some unknown reason, one shortcoming is that the prediction model never outputs blank frames. That is, frames will automatically be misclassified when there is no annotation or when there is an annotation of another type. The claimed accuracy of the model is 91.27%, but it is computed on non blank frames: when considering the whole dataset it drops to 44.0%.

Using the same three annotation categories while including a fourth one dedicated to blank frames, the CRNN learning model presented in Section 4 is trained and tested on NCSLGR. The dataset was split 70%-20%-10% (training-validation-test) in a random fashion. Results are given on Table 3. The CRNN model gets 87.9% accuracy on the whole NCSLGR dataset (91.8% when ignoring blank frames).

5.1.2. A relevant baseline for LSF

In order to establish a more relevant baseline for LSF, another frame classification model is trained, with three to six annotation categories, following their distribution given in Table 2. For instance, fingerspelling is much more frequent in ASL than in LSF, so it should not be included when considering only three categories.

Models are trained and tested either in a signer-independent fashion or not – signer-independent makes learning more difficult while increasing model generalizability. See Table 3 for an overview of the results.

A test sequence illustration with framewise probabilities is given on Fig. 5. This example shows that the model is able to accurately classify and segment most signs, even when only one blank frame separates them.

5.2. Detection of specific signs

A second experiment tackles the problem of detecting specific signs. The learning model is trained for one annotation type at a time, with *binary cross-entropy* as training loss.

$F1$-score is used, from the evaluation of *true positives*, *true negatives* and *false positives* within a sliding window of four seconds length. The corpus was split 70%-20%-10% respectively for training, validation and testing sets, in a signer-independent fashion. Two lexical signs are considered – "Same" (208 training instances) and "Line" (47 training instances) –, as well as pointing signs, depicting signs, fragment buoys and fingerspelling. Advantage is also taken from this experiment to evaluate the relevance of using the hand shape classifier and 3D modeling that are detailed in Section 4.1.

Results are given on Table 4. The benefit of the modeling presented in Section 4.1 over a simple OpenPose estimate is clear: the performance of every output is improved. The models performs rather poorly on fragment buoys and fingerspelling, which can be explained by the unclear annotation guidelines for the former, and too few training examples for the latter. A few factors are known to lessen model performance, including:

- The chopping of sequences at the middle of signs;
- A non fully optimized network architecture;
- An imperfect modeling, especially regarding hands;
- Annotation errors and bias or subjectivity;
- A great variability between signers;
- A lot of variability between signs because of the continuous nature of the corpus.
- More generally, the continuous nature of Sign Language in itself that makes it very difficult to classify its parameters into a finite number of categories.

5.3. Measuring the iconicity

Although the annotations for depicting signs are binary (signs are either annotated as depicting or they are not), a continuous transition from very conventionalized fully lexical signs to completely iconic structures can be observed in LSF. Cuxac (2000) refers to these two ends as *saying without showing* and *saying with showing*. In this third experiment, we illustrate with a few exemples that the learning model, when fed with annotations corresponding to depicting signs, seems to be learning a measure of iconicity in SL production.

For that purpose, the setting is identical to that of Section 5.2: binary cross-entropy as training loss with depicting signs as the only output of the learning model, signer-independent 70%-20%-10% split of the dataset, *3D+HS* signer modeling.

Fig. 6 presents four excerpts from a LSF sequence of Dicta-Sign–LSF–v2. Most annotated depicting signs are accurately detected, and many "false positives" actually seem to reveal a certain degree of iconicity that was not annotated. Provided these preliminary results are proven true in a future thorough analysis, this model could actually be used to build more appropriate linguistic categories of signs, which are still part of continuing discussions amongst linguists. Although this analysis will require further examination, it can only be conducted on datasets that include such annotations, that is very few of them.

6. Conclusion

In this paper, we introduced Dicta-Sign–LSF–v2, a continuous dialogue French Sign Language corpus. This corpus is finely annotated, with lexicon, classifier constructions and many more manual unit types. A convolutional-recurrent learning network was built, drawing on a compact and generalizable modeling of people signing.

A first baseline was established, showing that video frames can be accurately classified into the main annotation categories of the corpus. The classification network was also tested with state-of-the-art performance on NCSLGR, an ASL corpus. The classifier can be used to segment signs, even though it was not explicitly trained with this objective.

	Blank	FLS	DS	PT	FBUOY	N	FS	NCSLGR		Dicta-Sign	
								SD	SI	SD	SI
Yanovich et al. (2016)		✓	✓				✓	0.913	-	-	-
Ours								**0.918**	-	-	-
Yanovich et al. (2016)	✓	✓	✓				✓	0.440	-	-	-
Ours								**0.879**	-	-	-
Ours	✓	✓	✓	✓				-	-	0.814	0.768
	✓	✓	✓	✓	✓			-	-	0.810	0.761
	✓	✓	✓	✓	✓	✓		-	-	0.806	0.761
	✓	✓	✓	✓	✓	✓	✓	-	-	0.808	0.762

Table 3: Framewise accuracy on test set when classifying into three to six main manual types. A first experiments consists in a comparison with Yanovich et al. (2016) – for a fair comparison accuracy is either computed on all frames or on non blank frames only. The second experiment evaluates the model on Dicta-Sign–LSF–v2, with an increasing number of annotations categories as per Table 2. SD and SI stand for signer-dependent and -independent train/test setup.

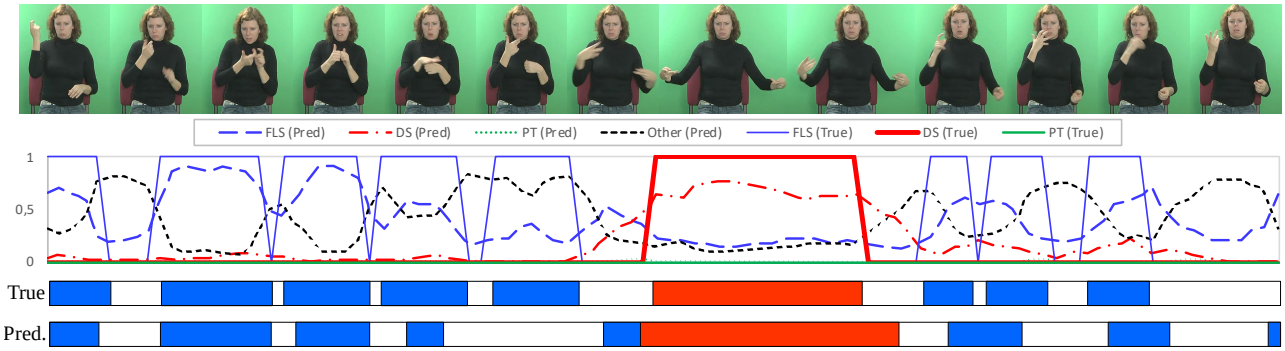


Figure 5: French Sign Language test sequence from Dicta-Sign–LSF–v2 (duration: 4 seconds – video reference: `S3_T6_A2`), with most lexical signs (blue) and one depicting sign (red) accurately detected by the CRNN model. Solid lines are for annotations while dashed lines are for output probabilities (middle graph). At the bottom is shown the argmax of these probabilities (*i.e.* predictions).

	FLS:		PT	DS	FBUOY	FS
	"Same"	"Line"				
2D (OP)	0.517	0.324	0.659	0.573	0.266	0.215
3D+HS	**0.673**	**0.560**	**0.693**	**0.680**	**0.303**	**0.496**

Table 4: Best F1-score on test set (Dicta-Sign–LSF–v2), for the detection of two lexical signs, pointing and depicting signs, fragment buoys and fingerspelling. Two modelings are considered: the one presented in Section 4.1 (3D+HS), and a simple 2D OpenPose modeling of a signer.

Furthermore, specific lexical signs or iconic structures can also be detected with the same learning framework.

Hopefully, this work will raise awareness on the importance of Sign Language Processing as a whole, that is beyond lexical-only perspectives, and lead to progress regarding the automatic analysis of iconic structures. It is intended as a baseline, so that different approaches can be compared.

Future work will focus on a finer analysis of iconicity in Dicta-Sign–LSF–v2, possibly using annotations of the corpus as weak labels.

7. Bibliographical References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.

Braffort, A. (2016). *La Langue des Signes Française (LSF): Modélisations, Ressources et Applications*. Collection Sciences cognitives. ISTE/Hermes Science Publishing.

Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *ICCV*.

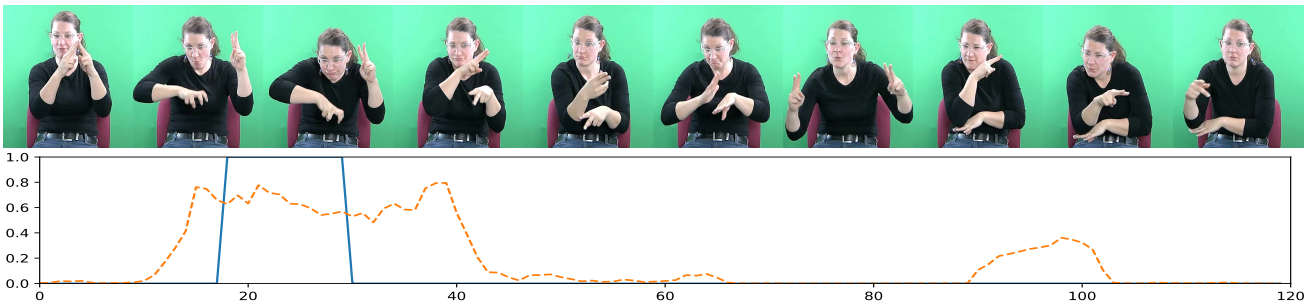
Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

Chai, X., Wanga, H., Zhoub, M., Wub, G., Lic, H., and Chena, X. (2015). DEVISIGN: Dataset and Evaluation for 3D Sign Language Recognition. Technical report, Beijing, Tech. Rep.

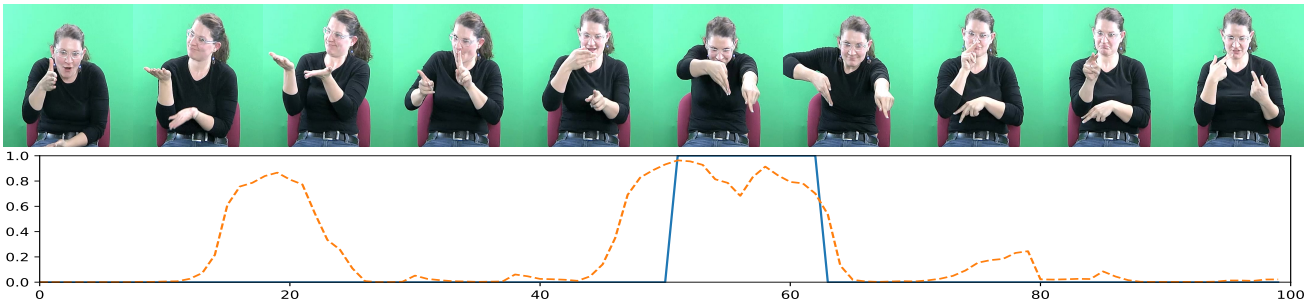
Chollet, F. et al. (2015). Keras.

Cuxac, C. (2000). *La langue des signes française (LSF): les voies de l'iconicité*. Number 15-16. Ophrys.

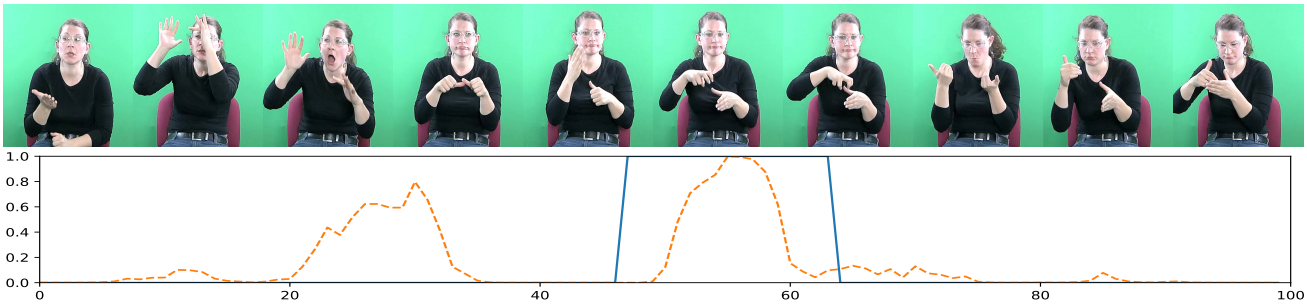
Filhol, M. and Braffort, A. (2012). A Study on Qualification/Naming Structures in Sign Languages. In *5th Workshop on the Representation and Processing of Sign*



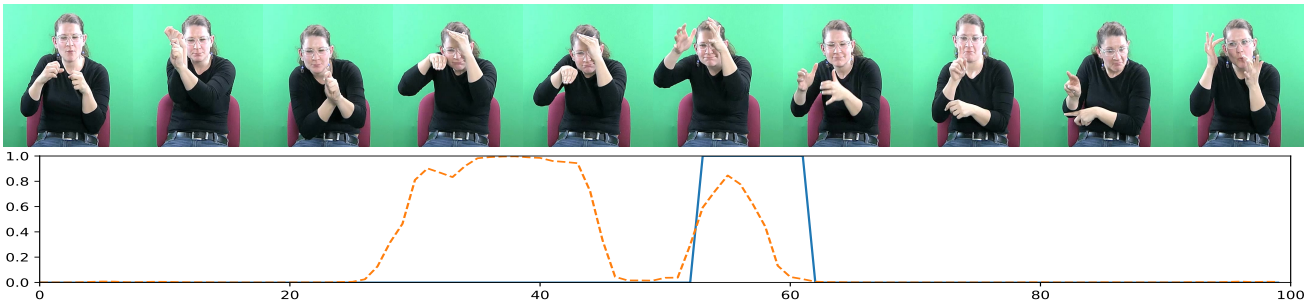
(a) A very iconic situation is detected around frame 25, even though it appears to be longer than annotated. Around frame 100, some unannotated form of constructed action may be detected.



(b) Quite an iconic and spatialized sign is detected around frame 20. The annotated depicting sign around frame 60 is clearly detected.



(c) Around frame 30, even though annotated as several lexical signs, some form of constructed action (role shift) is observed and detected. Around frame 55, the classifier construction is well recognized.



(d) Between frames 30 and 45, the lexical sign "under" is clearly produced in a very iconic way and detected as such. The shape and size classifier construction around frame 55 is accurately detected.

Figure 6: Four excerpts from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`), with annotated (full lines) and predicted (dashed lines) probability of depicting signs. Many false positives actually appear to come from a real but unannotated degree of iconicity in some signs. Only some key frames are shown.

Languages: Interactions between Corpus and Lexicon (LREC 2012), pages 63–66.

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 *LREC*, pages 1911–1916.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10):2451–2471.

Gonzalez Preciado, M. (2012). *Computer Vision Methods for Unconstrained Gesture Recognition in the Context of Sign Language Annotation*. Ph.D. thesis, Université de

Toulouse, Université Toulouse III-Paul Sabatier.

Huang, J., Zhou, W., Zhang, Q., Li, H., and Li, W. (2018). Video-based Sign Language Recognition without Temporal Segmentation. In *32nd AAAI Conference on Artificial Intelligence*.

Johnston, T. and De Beuzeville, L. (2014). Auslan Corpus Annotation Guidelines. *Centre for Language Sciences, Department of Linguistics, Macquarie University*.

Johnston, T. and Schembri, A. (2007). *Australian Sign Language (Auslan): An Introduction to Sign Language Linguistics*. Cambridge University Press.

Johnston, T. (2009). Creating a corpus of Auslan within an Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, pages 87–95.

Joze, H. R. V. and Koller, O. (2018). MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. *arXiv preprint arXiv:1812.01053*.

Koller, O., Ney, H., and Bowden, R. (2016). Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data is Continuous and Weakly Labelled. In *CVPR*, pages 3793–3802, June.

Koller, O., Zargaran, S., and Ney, H. (2017). Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. In *CVPR*, Honolulu, HI, USA, July.

Koller, O., Zargaran, S., Ney, H., and Bowden, R. (2018). Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *IJCV*, 126(12):1311–1325.

Liddell, S. K. (1977). *An investigation into the syntactic structure of American Sign Language*. University of California, San Diego.

Liddell, S. K. (2003). *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge University Press.

Matthes, S., Hanke, T., Regen, A., Storz, J., Worseck, S., Efthimiou, E., Dimou, N., Braffort, A., Glauert, J., and Safar, E. (2012). Dicta-Sign – Building a Multilingual Sign Language Corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon (LREC 2012)*, pages 117–122.

Meurant, L., Sinte, A., and Bernagou, E. (2016). The French Belgian Sign Language Corpus A User-Friendly Searchable Online Corpus. In *7th workshop on the Representation and Processing of Sign Languages: Corpus Mining (LREC 2016)*, pages 167–174.

Neidle, C. and Vogler, C. (2012). A new web interface to facilitate access to corpora: Development of the ASLLRP data access interface (DAI). In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC 2012)*. Citeseer.

Neidle, C., Thangali, A., and Sclaroff, S. (2012). Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC 2012)*. Citeseer.

Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., and Schwarz, A. (2008). DGS Corpus Project-Development of a Corpus Based Electronic Dictionary German Sign Language / German. In *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora (LREC 2008)*, June.

Schembri, A. (2008). British Sign Language Corpus Project: Open Access Archives and the Observer's Paradox. In *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora (LREC 2008)*, pages 165–169.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *JMLR*, 15(1):1929–1958, January.

Von Agris, U. and Kraiss, K.-F. (2007). Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition. *Gesture in Human-Computer Interaction and Simulation*.

Wilbur, R. and Kak, A. C. (2006). Purdue RVL-SLLL American Sign Language Database. Technical report, Purdue University.

Xiang, D., Joo, H., and Sheikh, Y. (2018). Monocular Total Capture: Posing Face, Body, and Hands in the Wild. *arXiv preprint:1812.01598*.

Yanovich, P., Neidle, C., and Metaxas, D. N. (2016). Detection of Major ASL Sign Types in Continuous Signing For ASL Recognition. In *LREC*.

Zhao, R., Wang, Y., and Martinez, A. M. (2018). A Simple, Fast and Highly-Accurate Algorithm to Recover 3D Shape from 2D Landmarks on a Single Image. *PAMI*, 40(12):3059–3066.

Zwitserlood, I., Efthimiou, E., Hanke, T., and Thoutenhoofd, E. (2008). The Corpus NGT: an Online Corpus for Professionals and Laymen. In *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora (LREC 2008)*, pages 44–49. Paris: ELRA.

8. Language Resource References

Belissen, Valentin and Braffort, Annelies and Gouiffès, Michèle. (2019). *Dicta-Sign-LSF-v2*. Limsi, distributed via ORTOLANG (Open Resources and TOols for LANGuage), `https://www.ortolang.fr/market/item/dicta-sign-lsf-v2`, Limsi resources, 1.0, ISLRN 442-418-132-318-7.

Benchiheub, Mohamed-El-Fataf and Braffort, Annelies and Berret, Bastien and Verrecchia, Cyril. (2016). *MOCAP1*. Limsi, distributed via ORTOLANG (Open Resources and TOols for LANGuage), `https://www.ortolang.fr/market/item/mocap1`, Limsi resources, 1.0, ISLRN 502-958-837-267-9.