Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 7–12
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Improving and Extending Continuous Sign Language Recognition: Taking Iconicity and Spatial Language into account

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

Abstract

In a lot of recent research, attention has been drawn to recognizing sequences of lexical signs in continuous Sign Language corpora, often artificial. However, as SLs are structured through the use of space and iconicity, focusing on lexicon only prevents the field of Continuous Sign Language Recognition (CSLR) from extending to Sign Language Understanding and Translation. In this article, we propose a new formulation of the CSLR problem and discuss the possibility of recognizing higher-level linguistic structures in SL videos, like classifier constructions. These structures show much more variability than lexical signs, and are fundamentally different than them in the sense that form and meaning can not be disentangled. Building on the recently published French Sign Language corpus Dicta-Sign–LSF–v2, we also discuss the performance and relevance of a simple recurrent neural network trained to recognize illustrative structures.
Keywords: Continuous Sign Language Recognition, Iconicity, Annotation

1. Introduction

In a series of recent papers focused on Continuous Sign Language Recognition (CSLR), the performance seems to be getting better and better and one may wonder what remains to be done before Sign Language Translation (SLT) can be envisioned. However, we want to show that the current trend in CSLR has inherent limitations, which prevent it from extending to SL understanding, *a fortiori* to SLT. Besides, extending on previous work, we try to highlight the value of recognizing illustrative elements in SL for the task of CSLR, even though few corpora are available for this research. This paper will also discuss the related challenges and the difficulties in evaluating the recognition results.

In Section 2, we start with a summary on the current CSLR paradigm, including linguistic assumptions, prevalent corpora and performance metrics. Then, the formulation we propose for a broader CSLR is presented in Section 3, with a discussion on some important linguistic properties of SLs, interesting corpora and a formal description. Building on this more general acceptation of CSLR, we end in Section 4 with an analysis and discussion of the results in the automatic recognition of illustrative structures within the recently published corpus Dicta-Sign–LSF–v2.

2. The current paradigm: Continuous *Lexical Sign* Recognition

In this section, we formalize the current CSLR paradigm. It focuses on the recognition of specific elements called *lexical signs*, therefore we choose to call it Continuous *Lexical Sign* Recognition (C*LS*R). Three main corpora are used, with a similar annotation scheme and types of discourse with limited variability, that do not allow for an easy generalization.

2.1. Lexical signs

In this paper, we refer to lexical signs following the definition of 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"*.

Whereas it is fairly safe to draw a parallel between common words and lexical signs, it only goes so far. Indeed, it is important to note that the grammar of SL is realized through the use of very different and more complex structures like described in Section 3.1.2.

2.2. Prevalent corpora

Since learning models are determined by the data they are trained with, it is necessary to discuss the nature of prevalent datasets in the field of CLSR. Three corpora stand out, and have been extensively used for training CLSR models:

- Signum (Von Agris and Kraiss, 2007) is a German Sign Language (DGS) dataset, with 5 hours of RGB video from 25 signers. Pre-defined sentences are elicited, with 465 lexical signs.
- RWTH Phoenix Weather 2014 (Forster et al., 2014; Koller et al., 2017) is made from 11 hours of live DGS interpretation of weather forecast on German TV.
- CSL-25k (Huang et al., 2018) results from the elicitation of 100 pre-defined sentences from 50 signers, with 178 annotated lexical signs, in Chinese Sign Language (CSL).

Table 1 presents a few random elicitation or transcribed sentences from these three corpora.

2.3. Type of discourse – elicitation material

Since they consist in elicited pre-defined sentences, it is important to realize that Signum and CSL-25k are, to some extent, artificial and with limited generalizability. In the CSL-25k corpus, the elicited sentences follow a simple syntactic structure, whereas the sentences in Signum show a

Signum	- Excuse me, could you help me? - The teacher will guide a tour through the church on Wednesday.
Phoenix	- The week starts unpredictable and cooler. - On Sunday spreads partly heavy shower and thunderstorm.
CSL-25k	- My dad is a businessman. - The cup is orange.

Table 1: Random elicited or transcribed sentence examples from three common SLR corpora. Signum and CSL-25k are elicited from pre-written sentences, while Phoenix is a live interpretation from weather forecast on German TV.

little more variability, with statements and questions, possibly a few subordinate clauses. Phoenix is more natural than CSL-25k and Signum, although the language variability and complexity are modest. Indeed, it is safe to assume that the live interpretation can be influenced by the original speech, especially in terms of syntax, and will make little use of the structures typical of SL like iconicity and space (Section 3.1).

2.4. Annotation and performance metric

The three aforementioned corpora all share the same annotation scheme: for each SL sequence, the annotation Y_{CLSR} consists in the sequence of elicited or observed lexical signs[1]:

$$Y_{\text{CLSR}} = [g_1, \ldots, g_N], g_i \in \mathscr{G} \quad (1)$$

where $\mathscr{G} = \{g^1, \ldots, g^G\}$ is a dictionary of lexical signs, and N is the number of annotated lexical signs in the sequence. A straightforward performance metric for recognition is then the word error rate (WER), also referred to as Levenshtein Distance, applied to the expected sequences of lexical sign glosses. WER measures the minimal number of insertions I, substitutions S and deletions D to turn the recognized sequence to the expected sequence of length N:

$$\text{WER} = (I + S + D)/N. \quad (2)$$

Table 2 summarizes the WER achieved by best recent CLSR models on Signum, Phoenix and CSL-25k[2].

Recognizing the sequence of produced lexical signs in a SL utterance has undeniable values. For instance, it can help getting a grasp of the general topic of a SL discourse. However, as we will discuss in the next section, because of specificities of SLs including iconicity and the use of space, more natural corpora with finer annotations are needed to get closer to automatic SL understanding.

[1]It is to be noted that temporal information is lost in this annotation scheme. For Phoenix only, Koller et al. (2017) released estimated frame alignments from a hybrid model.

[2]On the CSL-25k corpus, Pu et al. (2019) have also considered a signer-independent dataset split, but all test sequences are seen in training. As this is formally equivalent to the problem of recognizing isolated gestures, we did not include their results in the table.

	Signum		Phoenix		CSL-25k	
	SD	SI	SD	SI	SD	SI
Koller et al. (2017)	4.8%	-	26.8%	44.1%	-	-
Cui et al. (2019)	2.8%	-	22.9%	39.8%	-	-
Pu et al. (2019)	-	-	36.7%	-	32.7%	-

Table 2: Word Error Rates of most recent lexical sign recognition models on three prevalent SL corpora, with signer-dependent (SD) and -independent (SI) settings.

3. Continuous *Sign Language* Recognition: a better consideration for linguistics

In this section, rather than a thorough description of the linguistics of SLs, we want to highlight some fundamental properties, arguing for a necessary redefinition of the CSLR problem with appropriate corpora.

3.1. Fundamental linguistic properties

3.1.1. Simultaneity

Although SL has often been described as a hand-articulated language, the linguistic role of non-manual articulators – including facial expressions, eye gaze, mouth, and body posture (Baker and Padden, 1978) – is actually as relevant as that of manual ones.

Notably, this great number of articulators make it possible to convey various information simultaneously (Vermeerbergen et al., 2007). This is illustrated on the SL sequence of Fig. 1, where expert annotations are given below video thumbnails (see Section 3.2.2 for more detail on annotation categories). Indeed, on frames 7, 8 and 9 of the sequence, the left hand represents part of a previously instantiated building, the right hand locates several restaurants while the facial expression insists on their important number.

3.1.2. Iconicity and visual grammar

Too often overlooked in the field of CSLR, iconicity is nonetheless a major SL feature. For Cuxac (2000), iconicity even has a structuring role in the linguistics of SL: building on the visual modality, it enables to *show while saying*. Using the signing space in a visual way to structure discourse is also fundamental, and forms the core of the visual grammar of SL.

Johnston and De Beuzeville (2014) draw a distinction between *Fully Lexical Signs* (FLS) and *Partially Lexical Signs* (PLS). In this classification, PLS include Pointing signs (PT), Fragment buoys and Depicting Signs (DS) (see Fig. 1 and associated Table 3 for a detailed example). DS are sometimes referred to as classifier constructions, classifier signs or illustrative signs. Sometimes building on purely lexical signs, they use proforms[3] to visually describe the location, motion, size, shape or the action of referrents, along with trajectories in the signing space.

As one can notice on the annotations of Fig. 1, a SL utterance can be mostly made of illustrative structures.

[3]Often referred to as classifiers. They are standard hand shapes used to represent a variety of common entities (Collomb et al., 2018).

Figure 1: LSF utterance from Dicta-Sign–LSF–v2 (Belissen et al., 2019), with a predominant use of space and iconicity (video reference: `S7_T2_A10` – duration: 4 seconds). From top to bottom: thumbnails, detailed annotation for the manual activity: fully lexical signs (FLS) and partially lexical signs (PLS), each on three tracks (right handed (RH), two handed (2H), left handed (LH)). More detail is given in Table 3 and Section 3.2.2.
Possible translation: *At the very center of this area, there is a large building surrounded by restaurants*.

Frame	Linguistic analysis of the manual activity
1, 2	**Depicting sign** construction, with the right hand localizing an area at the middle of the signing space, while the left hand helps representing its limit in space.
3	**Pointing sign** to the middle of the area in question. The left hand is static and maintains a fragment of the previous sign for spatial coherence, which is called a **fragment buoy**.
4	**Lexical sign** "Middle/center", insisting on the fact that what is going to be said is at the *very* center of the area.
5	**Depicting sign** representing the shape of a building, with a facial expression highlighting its massive size and central position.
6	The left hand has a **fragment buoy** function, from the building sign at the center of the setting. The right hand produces a one-handed version of the **lexical sign** "Restaurant" (its standard form is two-handed).
7, 8, 9	A standard **classifier** that can be understood as a smaller building is successively placed all around the area. Three instances are placed, but the face expression suggest that there are many of them, probably more than just three. The left hand still maintains the **reference point** to the large building.

Table 3: This table is a linguistic description of the manual activity in the SL sequence shown on Fig. 1, including *lexicon*, *buoys*, *proforms*, *pointing*, *iconic structures* and *spatial structure*.

3.2. Alternative corpora

Conversely to the corpora presented in Section 2.2, NCSLGR (Neidle and Vogler, 2012, ASL) and Dicta-Sign–LSF–v2 (Belissen et al., 2019, LSF) are two public corpora made of or including very natural SL and frame-aligned annotation on lexical and non-lexical levels.

3.2.1. NCSLGR

NCSLGR includes two categories of discourse. Most videos are made of elicited utterances, similar to that of Signum. However, the corpus also includes spontaneous short stories, with a lot more language variability.

Manual activity is annotated on two fields, one for the dominant hand and the other for the non-dominant hand. Annotations follow the conventions from Baker and Cokely (1980) and Smith et al. (1988), with: *lexical sign glosses*, *fingerspelling*, *hold signs* (hand position held at the end of a sign, not necessarily with a linguistic function), *pointing signs*, *depicting signs* (7 categories) with proforms and *gestures*. Non-manual activity is also annotated, with head movement and eye gaze among others.

3.2.2. Dicta-Sign–LSF–v2

Dicta-Sign–LSF–v2 is a public remake of the French Sign Language (LSF) part of the Dicta-Sign Corpus (Matthes et al., 2012), with cleaned and reliable annotations. The corpus is based on dialogue with very loose elicitation guidelines, it is thus highly representative of natural SL. The annotated manual activity is inspired from the convention of Johnston and De Beuzeville (2014), with:

- Fully Lexical Signs (FLS) on three tracks (dominant hand, two-handed, non-dominant hand):
- Partially Lexical Signs (PLS) on three tracks (dominant hand, two-handed, non-dominant hand):
  - Depicting Signs with proforms, under 7 types:
    location (of an entity, DS-L),
    motion (of an entity, DS-M),

size and shape (of an entity, DS-SS),
ground (spatial or temporal reference, DS-G),
action (handling of an entity, DS-A),
trajectory (in signing space, DS-T),
deformation (of a standard lexical sign, DS-X)

  - Pointing signs (PT)
  - Fragment buoys

- Non Lexical Signs (NLS), with fingerspelling, numbering and gestures.

Constructed actions, also referred to as role shifts or personal transfers were not annotated, even though they share some of the properties of DS.

As the two underlying linguistic models are different, the annotations do not follow the same conventions. However, one will notice that the difference is not so significant. With spontaneous SL and fine annotations on lexical and non-lexical levels, these two corpora pave the way for a newer and broader acceptation of CSLR.

In the next section, we discuss appropriate metrics and a possible formalization for CSLR that could include FLS, PLS and NLS.

3.3. C*SLR*: formalization

Let us consider a CSLR system dealing with M different linguistic descriptors $d^i, i \in \{1, \dots, M\}$, such that the annotation for a sequence of length T can be written as:

$$Y_{\text{CSLR}} = \begin{bmatrix} d^1 \\ \vdots \\ d^M \end{bmatrix} = \begin{bmatrix} d_1^1 & \cdots & d_T^1 \\ \vdots & \ddots & \vdots \\ d_1^M & \cdots & d_T^M \end{bmatrix}. \tag{3}$$

Each of these descriptors can be binary, categorical or continuous, depending on the encoded information. For instance, d^1 could encode recognized lexical signs (categorical), d^2 the presence/absence of a pointing sign (binary), etc. They could also include spatial information.

For a general CSLR model, each descriptor d^i must be assigned a specific performance metric. For a categorical descriptor like the temporal recognition of lexical signs, the accuracy Acc defined as the ratio of correctly labeled frames over the total number of frames T looks like a good candidate. For binary outputs like the presence/absence of a depicting sign, we have found frame-wise F1-score to be an informative metric. From the count of true/false positives/negatives, F1-score is defined as the geometric mean of precision P and recall R, that is:

$$\text{F1} = 2\left(P^{-1} + R^{-1}\right)^{-1}. \tag{4}$$

However, it is important to realize that even very good prediction models may not get close to F1 $= 1$. Amongst many reasons is the fact that the beginning and end of any linguistic phenomenon can be difficult to assess with precision. For very short realizations, a discrepancy of 1-2 frames at the beginning and end between predictions and annotations may worsen the score dramatically.

In order to reduce the impact that the subjectivity of the temporal localization of signs can have on the performance measure, true/false positives/negatives and thus F1-score can be evaluated on sliding windows as opposed to frame-wise. For instance in Belissen et al. (2020), all metrics are computed on a sliding window of four seconds length. Although F1-score is very informative, whether frame-wise or computed on sliding windows, a better performance metric is still to be engineered.

In Section 4, we analyze the recognition results of a first CSLR attempt on the depicting signs of Dicta-Sign–LSF–v2, and discuss the relevance and interest of this analysis.

4. Recognizing depicting signs in Dicta-Sign–LSF–v2

Belissen et al. (2020) developed a modern learning framework for the recognition of many linguistic descriptors. A simple representation of a signer is obtained by separately processing the head, body pose and hand shapes from any SL RGB video. A convolutional and recurrent neural network is then built on top of this representation, and trained to recognize lexical signs, depicting signs and others in a supervised learning fashion.

4.1. Analyzing a few sequences

In this section, we return to this work and focus on the recognition of depicting signs (DS) on Dicta-Sign–LSF–v2, with a finer discussion on the prediction of the trained model. Fig. 2a, Fig. 2b and Fig. 2c are three excerpts from one of the test videos of Dicta-Sign–LSF–v2. For each sequence, along with video thumbnails are given:

- Model predictions (dashed lines) compared to annotations (full lines) for the recognition of the broad category "Depicting signs" (F1-score in the caption),
- Fine annotations of the manual activity annotated on three tracks (right handed (RH), two handed (2H), left handed (LH)), both for fully lexical signs (FLS) and partially lexical signs (PLS).

The selected excerpts show good prediction performance, with F1-scores between 49% and 86%:

Fig. 2a The two depicting signs are almost perfectly recognized in this sequence, even though one will notice that F1-score is only 86%, due to slight temporal shifts.

Fig. 2b The unique depicting sign is detected, although the prediction lasts longer, lowering the F1-score to 62%. As a matter of fact, the previous sign ("Eiffel Tower") and the next one ("Visit") are somehow included in the illustrative setting so that it could make sense to recognize iconicity outside the annotated depicting sign of motion type. Close to frame 100, a form of constructed action could be recognized, even though not annotated.

Fig. 2c The only annotated depicting sign is recognized, although the F1-score is quite low at 49%: between frames 20 and 35, the models recognizes somethings that could look like unannotated constructed action.

4.2. Benefitting from this analysis

Based on these three different examples, a first analysis suggests that some *false positives* could actually make sense. Indeed, the relevance of a clear separation between lexical and illustrative levels has been discussed for a long time (Cuxac, 2000). A finely annotated corpus like Dicta-Sign–LSF–v2 could enable researchers to extend our work and question the relevance of prevalent linguistic descriptions of SLs. Conversely, the usual SLR setting, with lexical annotations and WER metric prevents one from conducting this type of research. It implicitly uses the hypothesis that SL discourse can be described with sequences of lexical signs, which we have shown is far from sufficient.

Highlighting the subjectivity in the annotation, these examples show that an appropriate metric is still to be designed.

Finally, annotation for constructed action would have been a great help for the analysis of the results, so it might be a future addition to this corpus. Indeed, the results might suggest that constructed action and depicting signs also have a lot in common.

5. Conclusion and perspectives

In this paper, we have insisted on the central role of iconicity and spatial structure in Sign Language discourse, highlighting the fact that Lexical Sign Recognition is only a part of the Continuous Sign Language Recognition task.

Since prevalent SL corpora have intrinsic limits in terms of generalizability and do not include annotations outside lexicon, we felt it was important to point out that richer corpora do exist, with fine temporal annotations.

As a first attempt on the French Sign Language corpus Dicta-Sign–LSF–v2, we have trained a recurrent neural network to recognize depicting signs. While noting the limits of F1-score as a metric, model performance was carefully analyzed. Decent scores are met, especially when considering the unclear boundary between lexical and depicting signs. Indeed, this frontier is dependent upon the chosen linguistic model, with no clear consensus on the matter.

Beside more analysis on the performance metric and linguistic model, future work will include further reflection on the ways spatial information can be annotated and included in automatic recognition models. On a long-term basis, we will also reflect on how to go from the detection of important discourse elements like illustrative structures to global Sign Language Understanding.

6. Bibliographical References

Baker, C. and Cokely, D. (1980). American Sign Language. *A Teacher's Resource Text on Grammar and Culture. Silver Spring, MD: TJ Publ*.

Baker, C. and Padden, C. (1978). Focusing on the nonmanual components of American Sign Language. Understanding language through sign language research.

Belissen, V., Gouiffès, M., and Braffort, A. (2020). Dicta-Sign-LSF-v2: Remake of a Continuous French Sign Language Dialogue Corpus and a First Baseline for Automatic Sign Language Processing. In *LREC*.

Collomb, A., Braffort, A., and Kahane, S. (2018). L'anatomie du proforme en langue des signes française:. quand il sert à introduire des entités dans le discours. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, (34).

Cui, R., Liu, H., and Zhang, C. (2019). A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia*.

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

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.

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.

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.

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.

Pu, J., Zhou, W., and Li, H. (2019). Iterative alignment network for continuous sign language recognition. In *CVPR*, pages 4165–4174.

Smith, C., Lentz, E., and Mikos, K. (1988). Vista American Sign Language series: Signing naturally.

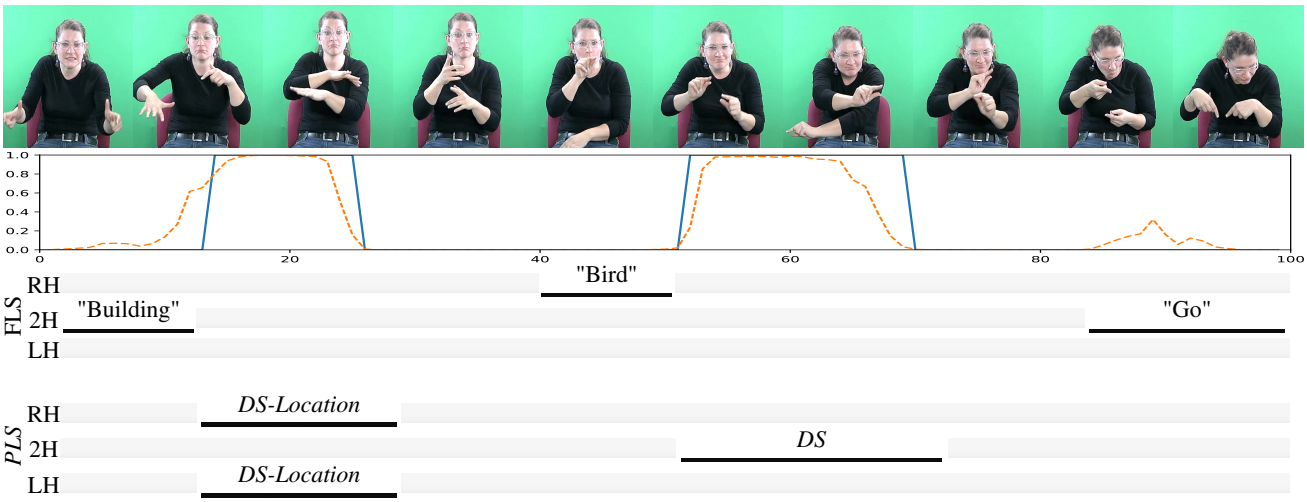
Vermeerbergen, M., Leeson, L., and Crasborn, O. (2007). *Simultaneity in Signed Languages: Form and Function*. Amsterdam studies in the theory and history of linguistic science. John Benjamins.

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*.

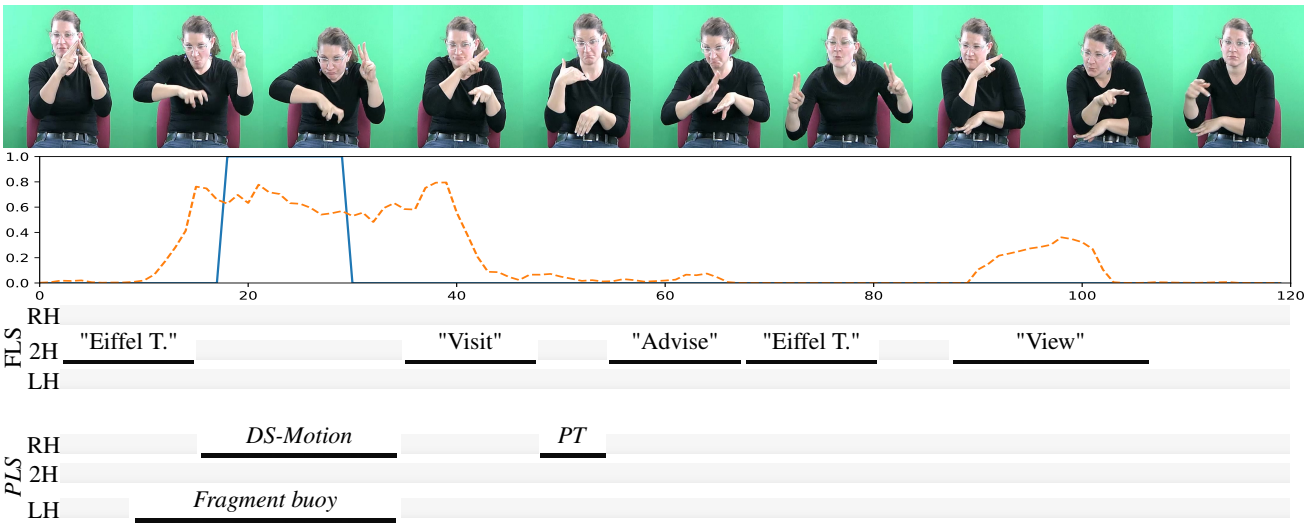
7. Language Resource References

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.

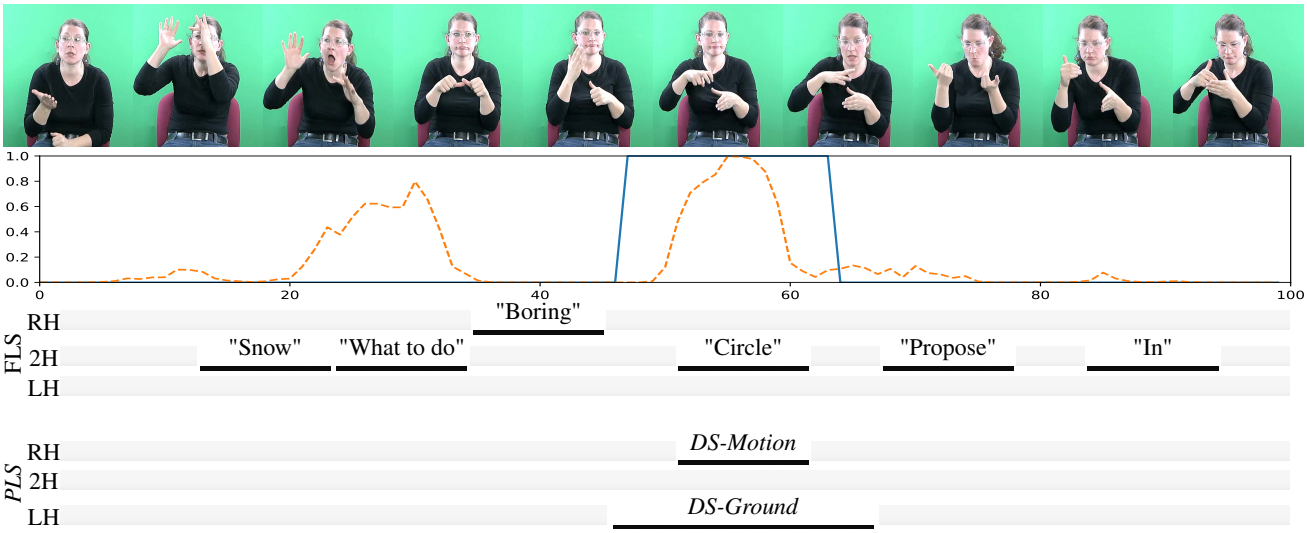
Neidle, Carol and Vogler, Christian. (2012). *NCSLGR*. American Sign Language Linguistic Research Project, `http://www.bu.edu/asllrp/ncslgr.html`.



(a) F1-score: 86% – *You should definitely go see this place where birds fly all around buildings.*



(b) F1-score: 62% – *I advise you to climb up the Eiffel Tower, you will then get a very nice panoramic view.*



(c) F1-score: 49% – *If you are stuck inside because of snow and you get bored, here is what I propose you to do.*

Figure 2: Three excerpts from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`). For both sequences, from top to bottom: thumbnails, ground truth (solid) and predictions (dashed) for the recognition of Depicting signs, detailed annotation for the manual activity: fully lexical signs (FLS) and partially lexical signs (PLS), each on three tracks (right handed (RH), two handed (2H), left handed (LH)). Frame-wise F1-score is indicated in the caption, next to a proposed translation.