From Sign Recognition to Automatic Sign Language Understanding : Addressing the Non-Conventionalized Units

Valentin Belissen

université
PARIS-SACLAY

From Sign Recognition to Automatic Sign Language Understanding: Addressing the Non-Conventionalized Units

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580, sciences et technologies de l'information et de la communication (STIC)

Spécialité de doctorat : Informatique

Unité de recherche : Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France

Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 10 novembre 2020, par

Valentin BELISSEN

Composition du jury :

François Yvon Directeur de recherche, CNRS, LIMSI	Président
Mounim El Yacoubi Professeur, Télécom SudParis – Institut Mines-Télécom	Rapporteur & Examinateur
Benoit Favre Maitre de conférences HDR, Université Aix-Marseille, CNRS, LIS	Rapporteur & Examinateur
Laurence Meurant Chercheur qualifié FNRS, Université de Namur	Examinatrice
Annelies Braffort Directrice de recherche, CNRS, LIMSI	Directrice de thèse
Michèle Gouiffès Maitre de conférences HDR, Université Paris-Saclay, CNRS, LIMSI	Co-encadrante de thèse

Thèse de doctorat

NNT: 2020UPASG064

Abstract

Sign Languages (SLs) have developed naturally in Deaf communities. With no written form, they are oral languages, using the gestural channel for expression and the visual channel for reception. These poorly endowed languages do not meet with a broad consensus at the linguistic level. They make use of lexical signs, i.e. conventionalized units of language whose form is supposed to be arbitrary, but also – and unlike vocal languages, if we don't take into account the co-verbal gestures – iconic structures, using space to organize discourse. Iconicity, which is defined as the existence of a similarity between the form of a sign and the meaning it carries, is indeed used at several levels of SL discourse.

Most research in automatic Sign Language Recognition (SLR) has in fact focused on recognizing lexical signs, at first in the isolated case and then within continuous SL. The video corpora associated with such research are often relatively artificial, consisting of the repetition of elicited utterances in written form. Other corpora consist of interpreted SL, which may also differ significantly from natural SL, as it is strongly influenced by the surrounding vocal language.

In this thesis, we wish to show the limits of this approach, by broadening this perspective to consider the recognition of elements used for the construction of discourse or within illustrative structures.

To do so, we show the interest and the limits of the corpora developed by linguists. In these corpora, the language is natural and the annotations are sometimes detailed, but not always usable as input data for machine learning systems, as they are not necessarily complete or coherent. We then propose the redesign of a French Sign Language dialogue corpus, Dicta-Sign-LSF-v2, with rich and consistent annotations, following an annotation scheme shared by many linguists.

We then propose a redefinition of the problem of automatic SLR, consisting in the recognition of various linguistic descriptors, rather than focusing on lexical signs only. At the same time, we discuss adapted metrics for relevant performance assessment.

In order to perform a first experiment on the recognition of linguistic descriptors that are not only lexical, we then develop a compact and generalizable representation of signers in videos. This is done by parallel processing of the hands, face and upper body, using existing tools and models that we have set up. Besides, we preprocess these parallel representations to obtain a relevant feature vector. We then present an adapted and modular architecture for automatic learning of linguistic descriptors, consisting of a recurrent and convolutional neural network.

Finally, we show through a quantitative and qualitative analysis the effectiveness of the proposed model, tested on Dicta-Sign-LSF-v2. We first carry out an in-depth analysis of

the parameterization, evaluating both the learning model and the signer representation. The study of the model predictions then demonstrates the merits of the proposed approach, with a very interesting performance for the continuous recognition of four linguistic descriptors, especially in view of the uncertainty related to the annotations themselves. The segmentation of the latter is indeed subjective, and the very relevance of the categories used is not strongly demonstrated. Indirectly, the proposed model could therefore make it possible to measure the validity of these categories. With several areas for improvement being considered, particularly in terms of signer representation and the use of larger corpora, the results are very encouraging and pave the way for a wider understanding of continuous Sign Language Recognition.

Keywords: Sign Language Recognition, Continuous Sign Language, Iconicity, Sign Language Linguistics, Signer Representation, Recurrent Neural Networks

Résumé

Les langues des signes (LS) se sont développées naturellement au sein des communautés de Sourds. Ne disposant pas de forme écrite, ce sont des langues orales, utilisant les canaux gestuel pour l'expression et visuel pour la réception. Ces langues peu dotées ne font pas l'objet d'un large consensus au niveau de leur description linguistique. Elles intègrent des signes lexicaux, c'est-à-dire des unités conventionnalisées du langage dont la forme est supposée arbitraire, mais aussi – et à la différence des langues vocales, si on ne considère pas la gestualité co-verbale – des structures iconiques, en utilisant l'espace pour organiser le discours. L'iconicité, ce lien entre la forme d'un signe et le sens qu'il porte, est en effet utilisée à plusieurs niveaux du discours en LS.

La plupart des travaux de recherche en reconnaissance automatique de LS se sont en fait attelés à reconnaitre les signes lexicaux, d'abord sous forme isolée puis au sein de LS continue. Les corpus de vidéos associés à ces recherches sont souvent relativement artificiels, consistant en la répétition d'énoncés élicités sous forme écrite, parfois en LS interprétée, qui peut également présenter des différences importantes avec la LS naturelle.

Dans cette thèse, nous souhaitons montrer les limites de cette approche, en élargissant cette perspective pour envisager la reconnaissance d'éléments utilisés pour la construction du discours ou au sein de structures illustratives.

Pour ce faire, nous montrons l'intérêt et les limites des corpus de linguistes : la langue y est naturelle et les annotations parfois détaillées, mais pas toujours utilisables en données d'entrée de système d'apprentissage automatique, car pas nécessairement cohérentes. Nous proposons alors la refonte d'un corpus de dialogue en langue des signes française, Dicta-Sign-LSF-v2, avec des annotations riches et cohérentes, suivant un schéma d'annotation partagé par de nombreux linguistes.

Nous proposons ensuite une redéfinition du problème de la reconnaissance automatique de LS, consistant en la reconnaissance de divers descripteurs linguistiques, plutôt que de se focaliser sur les signes lexicaux uniquement. En parallèle, nous discutons de métriques de la performance adaptées.

Pour réaliser une première expérience de reconnaissance de descripteurs linguistiques non uniquement lexicaux, nous développons alors une représentation compacte et généralisable des signeurs dans les vidéos. Celle-ci est en effet réalisée par un traitement parallèle des mains, du visage et du haut du corps, en utilisant des outils existants ainsi que des modèles que nous avons développés. Un prétraitement permet alors de former un vecteur de caractéristiques pertinentes. Par la suite, nous présentons une architecture adaptée et modulaire d'apprentissage automatique de descripteurs linguistiques, consistant en un réseau de neurones récurrent et convolutionnel.

Nous montrons enfin via une analyse quantitative et qualitative l'effectivité du modèle proposé, testé sur Dicta-Sign-LSF-v2. Nous réalisons en premier lieu une analyse approfondie du paramétrage, en évaluant tant le modèle d'apprentissage que la représentation des signeurs. L'étude des prédictions du modèle montre alors le bien-fondé de l'approche proposée, avec une performance tout à fait intéressante pour la reconnaissance continue de quatre descripteurs linguistiques, notamment au vu de l'incertitude relative aux annotations elles-mêmes. La segmentation de ces dernières est en effet subjective, et la pertinence même des catégories utilisées n'est pas démontrée de manière forte. Indirectement, le modèle proposé pourrait donc permettre de mesurer la validité de ces catégories. Avec plusieurs pistes d'amélioration envisagées, notamment sur la représentation des signeurs et l'utilisation de corpus de taille supérieure, le bilan est très encourageant et ouvre la voie à une acception plus large de la reconnaissance continue de langue des signes.

Mots-clefs : Reconnaissance de langue des signes, Langue des signes continue, Iconicité, Linguistique des langues des signes, Représentation du signeur, Réseaux de neurones récurrents

Acknowledgements

First of all, I would like to express my deepest gratitude to my two supervisors, Senior Scientist Annelies Braffort and Assistant Professor Michèle Gouiffès. Besides your consistent scientific input, moral support and valuable guidance during the course of this thesis, your everyday energy and enthusiasm made this research a truly enjoyable experience. I am thus indebted to you and I look forward to pursuing this work.

Relatedly, I wish to thank the M&TALS team for the constant insightful interactions, as well as the whole LIMSI staff for ensuring such a pleasant and joyful work environment.

I would also like to give special thanks to Professor Mounim El Yacoubi and Assistant Professor Benoit Favre for reviewing this thesis, which is a great deal of work I am sincerely grateful you accepted undertaking. Thank you very much to Research Associate Laurence Meurant and Senior Scientist François Yvon too for your highly appreciated participation in the jury.

On a more personal note, I wish to thank my parents, especially my father, for giving me the chance to get into the world of the Deaf and develop a fascination for Sign Language at a very early age. Last but no least thanks go to my love Hélène: throughout your constant support, you actually contributed to this work in a substantial way, in addition to making these years beautiful in so many ways.

Contents

List of Figures

List of Tables

List of Terms and Acronyms

Glossary

Buoy A hand posture held during some time of a discourse, usually on the weak hand, that is used as a physical reference point for a referent. *See* LBuoy, FBuoy, TBuoy. 50, 86

Citation-form lexical sign A lexical sign produced in isolation, such as when cited for purposes of illustration, as distinguished from the form it would take when produced in the normal stream of SL. 43, 45, 48, 54, 55, 71, 87

Depicting sign *see* illustrative structure and DS. 75, 83, 96

Dominant hand For a right-handed signer, the right hand. For a left-handed signer, the left hand. 39, 41, 42, 44, 55, 62, 96, 111

Iconicity The conceived similarity or analogy between the form of a sign and its meaning. 27, 34, 39–41, 43, 44, 50, 66, 79, 87, 91, 125, 155

Illustrative structure A manual unit that includes conventional and non-conventional elements that are highly context-dependent. These units cannot be listed in a dictionary in a simple way. 17, 49, 90, 91

Isolated lexical sign *see* citation-form lexical sign. 53–55, 60, 61, 66, 68, 112

Lexical sign A highly conventional manual unit, both in its form and meaning. It is relatively stable or consistent from one context to another. Lexical signs can easily be listed in a dictionary. 27, 36, 43, 49, 50, 80, 83, 86, 87, 89–91, 99, 144, 152, 153

Manual unit A meaningful gestural activity carried out by the hands and arms, whatever its linguistic function. 39, 45, 46, 50, 89–91, 93, 94, 96, 97, 121, 125, 127, 139

Signer-dependent Cross-validation setting of a prediction model for Sign Language Recognition, in which signers in test videos also appear in training videos. 54, 60–63, 65, 72–74, 78

Signer-independent Cross-validation setting of a prediction model for Sign Language Recognition, in which signers in test videos are excluded from training videos. 54, 57, 60, 61, 63–65, 68, 72–74, 78, 126, 152

Weak hand For a right-handed signer, the left hand. For a left-handed signer, the right hand. 17, 41, 42, 44, 50, 55, 62, 72, 90, 96, 111

Machine Learning and Image Processing

AAM Active Appearance Model. 60, 72

BLSTM Bidirectional LSTM. 63, 72, 118, 119, 126, 128–131

BW Black-White. 84

CNN Convolutional Neural Network. 54, 62, 63, 72, 73, 107, 112–115

CRF Conditional Random Field. 62, 78, 117

CTC Connectionist Temporal Classification. 72

DNN Deep Neural Network. 108–111

DTW Dynamic Time Warping. 72

EM Expectation Maximization. 72

FC Fully Connected. 113, 114, 119, 120

fps frames per second. 55, 59, 60, 68, 108, 111, 112, 128

GAN Generative Adversarial Network. 75

GCM Grassmann Covariance Matrix. 63

HMM Hidden Markov Model. 61–63, 71–73, 77, 117

HOG Histogram of Oriented Gradients. 61, 63, 72

LSTM Long Short-Term Memory. 72, 73, 117–119, 121, 126, 128, 130, 131

ML Machine Learning. 61

NLP Natural Language Processing. 27, 154, 155

NMT Neural Machine Translation. 75, 76

NN Neural Network. 64, 105

OP OpenPose [Cao et al., 2017]. 55, 57, 75, 86, 96, 108–112, 114, 115, 133, 135, 136

PCA Principal Component Analysis. 60

RCNN Recurrent Convolutional Neural Network. 63

RGB Red-Green-Blue. 58–60, 63, 68, 71, 84, 87, 110–112

RGBD RGB-Depth. 59–61, 63, 71, 84, 87

RNN Recurrent Neural Network. 63, 107, 117, 118, 121, 128

SLVM Spectral Latent Variable Model. 61

ST-GCN Spatial-Temporal Graph Convolutional Network. 63

SVM Support Vector Machine. 61, 63, 77, 78

VAE Variational Auto-Encoder. 75

WER Word Error Rate. 66, 67, 72–74, 78, 100

Sign Language

ACS Arbitrary Conventional Symbol. 40

ASL American Sign Language. 40–42, 46, 47, 55, 57, 60, 61, 77, 84, 86

Auslan Australian Sign Language. 83, 84

BSL British Sign Language. 41, 42, 84, 86, 87

ChSL Chinese Sign Language. 59, 60, 71, 72, 84

CLexSR Continuous Lexical Sign Recognition. 27, 66–68, 71–74, 76, 78–80, 99–101, 152–154

CSL Continuous Sign Language. 53, 55, 66, 68, 69, 71, 73, 75, 83–85, 87, 99, 101, 112

CSLR Continuous Sign Language Recognition. 45, 64, 66, 73, 79, 84, 99–101, 106, 107, 116, 117, 120, 121, 137, 153–155

DGS German Sign Language (*Deutsche Gebärdensprache*). 46, 47, 59, 60, 68, 70–72, 84, 86, 87, 112, 153

DTS Danish Sign Language (*Dansk Tegnsprog*). 72, 112

GR Gesture Recognition. 117

GSL Greek Sign Language. 87

HamNoSys Hamburg Notation System for Sign Languages [Hanke, 2004]. 46, 47, 87

HS Hand shape. 133, 135, 136

ILSR Isolated Lexical Sign Recognition. 27, 54, 55, 60–65, 67, 79

IPA International Phonetic Alphabet. 46

KSL Korean Sign Language. 41, 42

LSF French Sign Language (*Langue des Signes Française*). 27, 28, 33–35, 40, 45, 47, 48, 51, 79, 83, 84, 87, 100, 109, 111, 121, 143, 145–147, 149, 150, 153

LSFB French Belgian Sign Language (*Langue des Signes Francophone de Belgique*). 84, 86

MF Manual Feature. 60

NGT Dutch Sign Language (*Nederlandse Gebarentaal*). 84, 86

NMF Non-Manual Feature. 60, 78

NZSL New Zealand Sign Language. 72, 112

RGS Richly Grounding Symbol. 40

SD Signer-Dependent (*see* signer-dependent). 61, 65, 74, 78, 139, 140

SGI Highly Iconic Structure (*Structure de Grande Iconicité*) [Cuxac, 2000]. 40, 42, 43, 45, 47, 49, 50

SI Signer-Independent (*see* signer-independent). 61, 65, 74, 78, 139–142, 152

SL Sign Language. 27, 28, 33–37, 39–41, 43–50, 53–55, 60, 66–68, 71–73, 75, 77, 79, 80, 83, 84, 87, 88, 91, 96, 101, 106–108, 111, 112, 116, 121, 125, 152–155

SLG Sign Language Generation. 73, 87

SLP Sign Language Processing. 27, 33, 36, 37, 39, 73

SLR Sign Language Recognition. 27, 28, 36, 37, 39, 44–46, 50, 53, 54, 61, 66, 68, 73, 77, 79, 80, 83, 84, 87, 96, 100, 107, 108, 115, 117, 152–155

SLT Sign Language Translation. 27, 36, 44, 73, 75–77, 79, 80, 96, 99, 153

SLU Sign Language Understanding. 27, 44, 77, 79, 96, 99, 121, 153, 155

T-FS Transfer of Form and Size. 41, 42, 50

T-P Transfer of Persons. 41, 42, 144, 148

T-S Situational Transfer. 41, 42, 50

TD Task Dependent. 139, 140, 142

TI Task Independent. 140, 141, 152

Sign Language Corpora

ASLLVD American Sign Language Lexicon Video Dataset [Neidle et al., 2012]
cited p. 55, 56, 60–65

Auslan Corpus Australian Sign Language Corpus [Johnston, 2009]
cited p. 49, 83–85

BSLCP British Sign Language Corpus Project [Schembri, 2008]
cited p. 84–86

Corpus NGT Corpus *Nederlandse Gebarentaal* (Dutch Sign Language) [Crasborn and Zwitserlood, 2008; Crasborn et al., 2008]
cited p. 84–86

CSLR100 Continuous Sign Language Recognition 100 [Huang et al., 2018]
cited p. 71, 73, 74, 80, 84

DEVISIGN-L [Chai et al., 2014]
cited p. 58–60, 63, 65

DGS Korpus *Deutsche Gebärdensprache* (German Sign Language) Corpus [Prillwitz et al., 2008]
cited p. 84–86

Dicta-Sign [Efthimiou et al., 2010]
cited p. 83, 87–89, 91, 97, 155

Dicta-Sign–LSF–v2 [LIMSI and IRIT, 2019; Belissen et al., 2020a]
cited p. 28, 47, 50, 51, 84, 87, 88, 90–96, 100, 121, 127, 130, 131, 135, 136, 141–147, 149, 150, 152–155

ISLR500 Isolated Sign Language Recognition 500 [Pu et al., 2016]
cited p. 59–61, 63, 65, 71

LSFB Corpus *Langue des Signes Francophone de Belgique* (French Belgian Sign Language) Corpus [Meurant et al., 2016]
cited p. 84–86

MS-ASL Microsoft American Sign Language [Joze and Koller, 2018]
cited p. 57, 58, 60, 63–65

NCSLGR National Center for Sign Language and Gesture Resources [Neidle and Vogler, 2012]
cited p. 77, 78, 83–86

Purdue RVL-SLLL Purdue Robot Vision - Sign Language Linguistics Labs [Martínez et al., 2002]
cited p. 61

RWTH-Boston-50 [Zahedi et al., 2005]
cited p. 61

RWTH-PW RWTH-Phœnix-Weather [Forster et al., 2012, 2014]
cited p. 68, 70–74, 77, 80, 84, 153

SIGNUM Signer-Independent Continuous Sign Language Recognition for Large Vocabulary Using Subunit Models [Von Agris and Kraiss, 2007]
cited p. 59–61, 65, 68, 69, 71–74, 78, 80, 84

WLASL Word-Level American Sign Language [Li et al., 2020]
cited p. 57

Trevor Johnston's and Dicta-Sign–LSF–v2 annotation categories

FLS Fully Lexical Sign (*see* lexical sign). 46, 48, 49, 51, 90, 92–95, 100–102, 121, 125, 127–130, 132, 134, 135, 139, 141–152, 154

PLS Partially Lexical Sign. 49, 51, 90, 93, 94, 100, 142, 143, 145–147, 149, 150

DS Depicting Sign (*see* depicting sign). 17, 49, 83, 84, 86, 90, 91, 93–95, 97, 100, 101, 121, 125, 127–130, 132, 134, 135, 139, 141–152, 154

DS-L DS-Location (of an entity). 49, 91, 93, 96, 97, 100, 149, 150

DS-M DS-Motion (of an entity). 49, 51, 91, 93, 96, 97, 147

DS-SS DS-Size&Shape (of an entity). 49, 51, 91, 93, 97, 100, 144, 146, 149

DS-G DS-Ground (spatial or temporal reference). 49, 91, 93, 96, 97, 100, 148, 149

DS-H DS-Handling (an entity). 49, 91

DS-A DS-Action (*see* DS-H). 91, 93, 97

DS-T DS-Trajectory (shown in the signing space). 91, 97

DS-X DS-Other (any other deformation of a standard lexical sign). 91, 97

PTS Pointing Sign. 50, 51, 83, 84, 86, 90, 93–95, 100, 101, 103, 121, 125, 127–129, 131, 132, 136, 139, 141–152, 154

LBuoy List Buoy (*see* Section 2.3.3.2). 17, 50, 90

FBuoy Fragment Buoy (*see* Section 2.3.3.2). 17, 50, 51, 84, 90, 93–95, 100, 101, 125, 127–129, 131, 132, 136, 139, 141–152, 154

TBuoy Theme Buoy (*see* Section 2.3.3.2). 17, 50, 90

NLS Non Lexical Sign. 49, 50, 90, 93, 94

FS Fingerspelled Sign. 50, 84, 86, 91, 93–95

NS Numbering Sign. 84, 91, 93–95

G Gesture. 50, 84, 86, 91, 93

Introduction

This thesis concludes a research project in a sub-domain of Artificial Intelligence, namely Natural Language Processing (NLP). It focuses on Sign Languages (SLs) with most experiments carried out on French Sign Language (LSF).

Contrary to popular belief, SLs are in no way *coding systems* and have not been *invented* by anyone. They have instead developed naturally inside Deaf communities and, like other natural languages, they fit into historical, geographical and political environments. Granted, since these environments are partly shared with hearing people, each SL is, to some extent, influenced by the co-occurrent vocal language(s). However, because of the very specific visual-gestural modality, SLs hardly fall into the linguistic frameworks used to describe vocal languages.

Perhaps partly because of misconceptions about SLs and the fact that SLs are poorly endowed languages, the field of Sign Language Recognition (SLR) has mostly focused on the recognition of lexical signs, which are conventionalized units that could loosely be compared to words. Yet, if SLR is considered as a step towards Sign Language Understanding (SLU) and Sign Language Translation (SLT), this approach is bound to be ineffective. Indeed, SLs are much more than sequences of signed words: they are iconic languages, that use space to organize discourse benefitting from the use of multiple simultaneous language articulators.

Therefore – and bearing in mind that the recognition of lexical signs alone remains a rather complicated task – to what extent can we propose and experiment a redefinition of SLR, so that it effectively points towards SLU and SLT? This is the main question addressed in this thesis, that we have organized into three parts.

Part I aims to introduce the context behind this research, and present the state of the art in the field of SLR. In Chapter 1, we elaborate on what SLs are, through their specific modality, sociolinguistic matters and different levels of observed variety. Then, we discuss Sign Language Processing (SLP) in a general way, especially the objectives the implications of processing poorly endowed languages. In Chapter 2, we dive into the linguistics of SLs. The evolution in the linguistic theories is first presented with and emphasis on iconicity, then we highlight the challenges these theories imply for the field of SLR. Relatedly, we discuss how SLs can be transcribed and annotated. Chapter 3 then details the state of the art in automatic SLR, outlining the general framework then the usual acceptation of SLR, that we call Isolated Lexical Sign Recognition (ILSR) and Continuous Lexical Sign Recognition (CLexSR). This chapter ends with a discussion of the few experiments that fall outside this framework.

Our proposed approach to answer the above research question is then developed in Part II. Valuable

natural SL corpora made by linguists that include detailed annotations are first presented in Chapter 4 along with their intrinsic limits, then a SLR-oriented remake of a LSF dialogue corpus that we propose, Dicta-Sign–LSF–v2, is introduced. The strong assets of this corpus for SLR purposes are highlighted in the same chapter. In Chapter 5, we reformulate the problem of SLR as the recognition of parallel linguistic descriptors. We also discuss appropriate performance metrics and loss functions for training models based on this approach. Motivated proposals for a compact learning framework and a generalizable signer representation are then introduced in Chapter 6.

Finally, Part III is aimed at the validation and evaluation of the proposed approach, including the redefined SLR acceptation, the signer representation and learning model, all based on the Dicta-Sign–LSF–v2 corpus. This validation is first carried out quantitatively in Chapter 7, using metrics developed in Part II, then more qualitatively in Chapter 8 through a detailed analysis of the predictions of our model. Finally, perspectives for the future of SLR are drawn, hopefully accounting for a larger part of SL linguistics.

Qu'importe la surdité de l'oreille quand l'esprit entend ?
La seule surdité, la vraie surdité, la surdité incurable,
c'est celle de l'intelligence.

Victor Hugo
Correspondence with Ferdinand Berthier (1843)

Part I

Sign? Language? Recognition?

Chapter 1

Sign Languages and Sign Language Processing

This first chapter consists in a short and general introduction on Sign Languages (SLs) and on Sign Language Processing (SLP), leaving out detailed linguistic considerations for Chapter 2.

In the current chapter, the specific signed modality that SLs involve is first discussed in Section 1.1.1, then ethical and sociolinguistic aspects are considered in Section 1.1.2 and the question of variety and variability in SL is analyzed in Section 1.1.3. Finally, we reflect on the sense, objectives and current limits of SLP in Section 1.2.1.

1.1 Sign Languages

First and foremost, we want to discuss a few generalities with respect to the specificities of SLs, which can be defined as natural languages that originate in the communication between Deaf people. Although we will sometimes refer to the particular case of French Sign Language (LSF), most observations can actually be extended to all SLs.

1.1.1 A specific modality: *signed* languages

One of the most obvious ways to start the description of SLs is to analyze their specific modality. They can be considered as a form of oral language, in the sense that they include both expressive and receptive language. The expressive and receptive parts of traditional vocal languages are usually referred to as *speaking* and *listening*. For Sign Languages they can respectively be referred to as *signing* and *watching*. Also, *oral* may be understood as the opposite of *written* language, which is a further argument for classifying sign languages as oral languages.

Thus, SLs form a face-to-face oral language, with a specific expressive-receptive gestural-visual modality (Figure 1.1). The many language articulators comprise hands, arms, shoulders, torso, head, face and eyes.

In light of this discussion on the modality of SLs, we would like to raise a note of caution. Although the name *Sign Language* might suggest a language made of signs – or gestures – with similar linear structure as that of common vocal languages –, we feel that *signed language* could be more informative. Indeed, as a result of their specific modality and the number of visual articulators, SLs hardly fit into the formal framework used for the description of vocal languages. As this will be detailed in the next chapter, SLs allow for a simultaneous use of different articulators while they make a strong use of

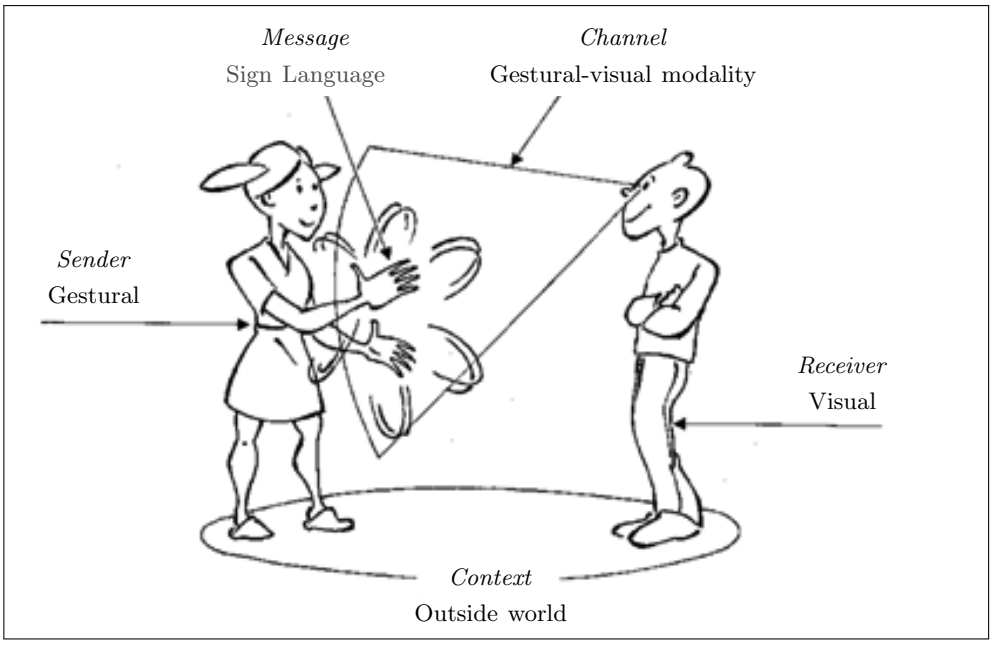


Figure 1.1: Sign Language: a face-to-face oral language, with a gestural-visual modality. Illustration of Laurent Verlaine, from [Guitteny, 2006].

space and iconicity to construct discourse. We will however keep on using the name *Sign Language* for the rest of this thesis, since it is predominantly used in the literature and the Deaf community.

1.1.2 Sociolinguistics of Sign Languages and Deaf identity

Because SLs have naturally developed within Deaf communities, originating in the communication between Deaf people, they each have historical, geographical and political properties, like other natural languages. What makes them fundamentally different is their intrinsic relationship to the absence of the hearing modality.

Therefore, while ***d****eafness* is clinically defined as the state of hearing loss – that is an audiological condition –, ***D****eafness* is rather the affiliation to a linguistic, thus social, cultural, geographical, historical and political community. In this sense, sociologists Bernard Mottez and Harry Markowicz have shown that hearing loss is a form of *shared handicap* that SL contributes to dissolve [Mottez et al., 1990]. We have thus chosen to use the capital letter D when referring to Deaf people.

In fact, the Deaf identity described by Mottez and Markowicz [1979] has been shaped by and through SL. This foundation is nevertheless very fragile. Indeed, SLs have not been officially recognized and accepted until very recently, and only in some countries. In France, LSF has long been banned as a language of instruction for deaf children (from 1880 to 1991) and its official recognition as a language of France is very recent (2005)[1].

As a result, studies on SL, whether on a linguistic level or in terms of automatic processing, are understandably a sensitive subject. In this respect, we take a strong stance: our research is focused on the language of the Deaf, rather than on an abstract and disembodied apprehension of language.

[1]Loi n°2005-102 du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées https://www.legifrance.gouv.fr/eli/loi/2005/2/11/2005-102/jo

M A X

(a) Fingerspelling for "Max"

(b) The sign for FRIDAY (*VENDREDI* in French) initialized with the "V" hand shape

Figure 1.2: A few examples of the influence of French on French Sign Language (LSF): fingerspelling and initialization with a letter from the French alphabet.

1.1.3 Variety of Sign Languages

Because Deaf people belong to national communities, it makes sense that SLs are somehow influenced by – or, in a more neutral perspective, related to – the official language(s) of the associated countries. Fingerspelling for proper nouns, variable levels of mouthing or the initialization of some signs by using the manual alphabet of the official language are three examples of this influence (see Figure 1.2). More generally, the Deaf and hearing people share numerous aspects of the national culture, which obviously results in a certain degree of proximity between languages.

At the intra-national level, because of the lack of standardization in the national education of the Deaf, regional varieties are easily observed, mostly with respect to the lexicon, with a strong relationship to the history of important schools for the Deaf. As far as the grammar is concerned, the variability is significantly lower [Schembri et al., 2010]. The same line of argument car be pursued regarding the international comparison of different SLs, with even larger disparities on the lexical level, and different

grammar usage, still within the linguistic constraints developed in Section 2.2.

1.2 Automatic Sign Language Processing

1.2.1 What, and for whom?

In Section 1.1.2, we insisted on the fact that language and identity are inherently related. Noting that identity is never a neutral subject, research on SL – especially with applicative claims – should not be conducted without questioning one's objectives and their limitations.

A first goal for developing SLP systems is obviously to focus on applications, like Sign Language Translation (SLT). As we will develop in Chapter 3, the best performing[2] models for this task process SL videos with a *black box* architecture.

On the other hand, one may consider the development of SLP systems as a way to advance both applications *and* language descriptions, thus excluding such black box frameworks. Indeed, determining the key characteristics for Sign Language Recognition (SLR), for instance, can bring light on what should be further analyzed by linguists – and *vice versa*. This is the direction that is followed in this thesis.

Last, the issue of *standardization of language* is not to be overlooked. SLP systems are trained and tested with SL corpora, that are not necessarily representative of the inner variability of SLs, and with linguistic elicitation that can be highly constrained. Care should then be taken when extrapolating trained SLP systems or making them publicly available. Typically, most SLR models detailed in Chapter 3 focus on the recognition of lexical signs, which are fully conventional signs, equivalent to common words in English. Potential users (signers) of such applicative models would thus be discouraged to integrate natural illustrative structures – distinct from lexical signs – into their discourse, for instance. More generally, it is clear that such systems are not neutral with respect to the use and evolution of language.

1.2.2 Automatic Processing of Poorly Endowed Languages

Many difficulties faced by the field of automatic SLP actually stem from the fact that SLs are poorly endowed languages. Indeed, research on SL in general is quite recent, whether on linguistic, historical or sociological bases.

Because of the novelty of this research, linguistic descriptions of SLs are unfortunately lacking consensus and robustness. SL corpora are few and limited in size and scope, which also adds to the absence of written form of SLs. Even more significantly, a lot of research on machine intelligence and SLP has been conducted by hearing researchers. At best, even the most experienced of them remain people whose mother language is not a SL – with very rare exceptions like Professor Christian Vogler. At worst, many of them are unaware of the linguistic complexity of these SLs and may design processing systems with very poor – if any – linguistic relevance or real effective usability.

Conclusion

SLs are thus a very specific family of languages, with an original signed modality, no written form and small communities of signers. Some SLs have been officially recognized and accepted very recently,

[2] There is still a very long way to go before reaching quality SLT.

and few resources are available. The variety of SLs is also important, which is often neglected. For these reasons, automatic SLP is still in its early stages.

In the next chapter, we will dive into the main key linguistics properties of SLs, and try to identify the main challenges for SLR.

Chapter 2

Linguistics of Sign Languages

In this chapter, we emphasize the different linguistic descriptions of SLs, and highlight the relationship to automatic Sign Language Processing (SLP). In Section 2.1, we present three main approaches for the description of SL linguistics, that differ from each other in their relationship with the linguistics of traditional vocal languages and in the role they give to iconicity. The implications of the choice of linguistic description on Sign Language Recognition (SLR) is outlined in Section 2.2, then we discuss how SLs can be transcribed and annotated in Section 2.3.

2.1 Evolution in the description of Sign Languages

2.1.1 Signs as lexical units and arbitrariness – a parametric convergent approach

Historically, lexical units (or words) have been defined as the conventionalized minimal form-meaning pairings found in a language, also referred to as *free morphemes*[1]. More precisely, they are the combination of meaningless units (phonemes), and their form is arbitrary, *i.e.* not motivated [De Saussure, 1916].

Stokoe [1960, 1972], in an effort to show that manual languages used by the Deaf communities in the United States were proper languages with the same expressive power as spoken languages, and drawing inspiration from the precursor work of Bébian [1825] with respect to the Deaf communities in France, derived a phonology of SLs. His *parametric* theory indeed describes signs, that is meaningful gestural units, as manual units, combination of the following three sub-lexical units, or parameters, or components: hand location (at the beginning of the sign), hand shape and hand movement – trajectory or hand transition. Minimal pairs of signs are then formed by the modification of a single component, like minimal pairs of words are formed by changing a single phoneme. Signs can be bimanual (two-handed) or unimanual (one-handed). In the latter case, the sign is then realized with the dominant hand (the right – resp. left – hand for a right-handed – resp. left-handed – person).

A fourth parameter, the hand orientation, is then included by Battison [1974]. Battison also develops on the specificities of bimanual signs *versus* unimanual signs, observing that the relationship between the parameters of the two hands in bimanual signs is subject to either symmetry or dominance constraints.

[1] Morphemes are the first level of meaningful units in language, and free morphemes are morphemes that can stand alone as words.

(a) Vehicle proform, used to represent a car moving

(b) Person proform, used to show a person diving

(c) Animal proform, used to represent an animal hoping

Figure 2.1: Very common proforms (classifiers) used in American Sign Language [Supalla, 1986]

This parametric approach, which is consistent with the structuralist theory, was subsequently adopted by others, like Nève [1996] in France. In the words of Millet and Colletta [2002], it is referred to as *convergent approach*, in the sense that it is based on the same theoretical concepts as those developed for analyzing vocal languages – especially phonology and syntax. Because it is reliant on the Saussurian definition of language, where arbitrariness is key, iconicity is set aside.

2.1.2 Iconic Sign Languages – a differentialist approach

Unlike the convergent approach, *differentialists* – as per Millet and Colletta [2002] – like Liddell [1998] then Cuxac [1999, 2000]; Cuxac and Sallandre [2007]; Pizzuto et al. [2007], have chosen to describe the linguistics of SLs without resorting to the epistemological tools usually employed for describing vocal languages. In those models, a key concept is iconicity, that is the direct form-meaning association in which the linguistic sign resembles the denoted referent in form.

Related claims had actually been made earlier for American Sign Language (ASL). The structuring aspect of iconicity in ASL is for instance noted by DeMatteo [1977], who insisted that iconicity should be taken into account in the grammar of ASL, not only as a component of it, but as its basis. Macken et al. [1993] also proposed that ASL is a *dual-representation language*, with both Arbitrary Conventional Symbols (ACSs) and Richly Grounding Symbols (RGSs).

The work of Cuxac is nonetheless unique, in that iconicity holds a pivotal role at every level of his *semiological* model. On the one hand, iconicity is analyzed through a theory of intent – the illustrative and non illustrative intents ; on the other hand, three types of iconicity are described, on a grammatical level – imagic as opposed to degenerated iconicity, as well as diagrammatic iconicity.

2.1.2.1 The illustrative intent (imagic iconicity)

When there is a deliberate intent of showing, illustrating and demonstrating while telling, very iconic structures are used in SL discourse. These Highly Iconic Structures (*Structures de Grande Iconicité*, SGIs) are at the core of Cuxac's theory. They are often referred to as *transfers*, in the sense that they are assumed to result from the transfer of cognitive operations in the mental universe of imagery to the three-dimensional signing space.

These transfers generally use *proforms*, that are very iconic hand shapes and generic forms used to represent the form[2] of a referent entity, similarly to a pronoun. Figure 2.1 shows three examples

[2]Either the general form of the referent entity – a flat hand for the "car" proform in French Sign Language (LSF) –, or some salient features of it – for instance the legs of a person in Figure 2.1b.

of proforms in ASL. Very often, they are referred to in the literature as *classifiers*, while some authors propose the more indicative expression *property marker* [Slobin et al., 2003]. Within transfers, proforms are used together with other manual and non-manual parameters.

Three main categories of transfers are described by Cuxac, and illustrated in Figure 2.2:

Transfers of Form and Size (Ts-FS) use proforms to describe the shape, size and position of entities. In Figure 2.2a, the signer draws a kind of sketch in space, representing the surface of an object with her hands. The shape of her lips and cheeks, her partially closed eyes and her lowered head emphasize the imposing character of this object.

Situational Transfers (Ts-S) are used to describe the movement and/or action of entities, usually with the dominant hand, while the weak hand is used as a reference point. In frames 3 and 4 of Figure 2.2b, the weak (left) hand depicts a reference point, like the corner of a room, while the dominant (right) hand uses a specific proform to represent a person, which is rotated inside the depicted room to illustrate a person going round and round in circles.

Transfers of Persons (Ts-P) go one step further, with the signer playing the role of the described entity, using many body articulators. This is the case in the four frames of Figure 2.2b, where the signer enacts a bored person, which is particularly visible on her face expression (cheeks and lips), her gaze looking away and her head moving side to side.

Ts-FS, Ts-S and Ts-P can then be combined into many sub-types of transfers, according to the signers's point of view and intent [Sallandre and Cuxac, 2002]. The eye gaze moving away from the addressee to the signing space is also shown to be a property of transfers [Cuxac, 1999].

Other widely used expressions for transfers are:

- For Ts-FS: *size-and-shape specifiers* or *size-and-shape classifier constructions* [Supalla, 1986].
- For Ts-S: *polycomponential signs* [Slobin et al., 2003] or *classifier predicates* [Supalla, 1986].
- For Ts-P: *role playing*, *referential shifts*, *role shifts* or *constructed action* [Taub, 2001].

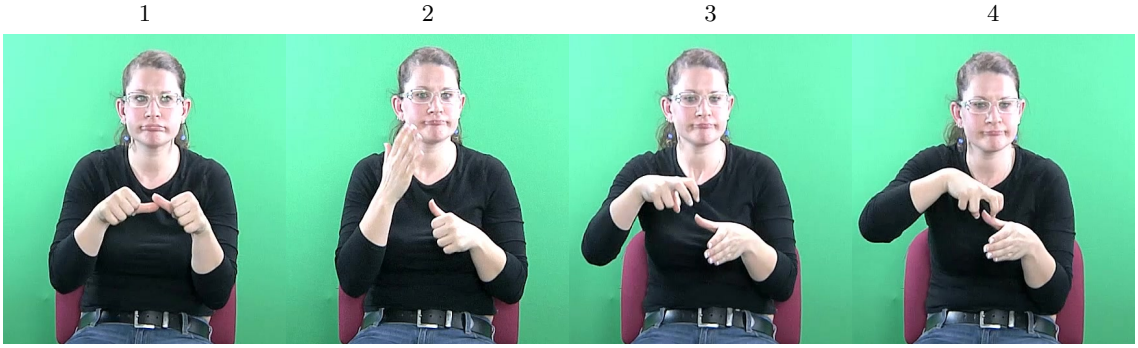
Because some authors may assimilate *classifiers* to *classifier constructions*, we prefer to use the distinct expressions *proforms* and *transfers* in the rest of this thesis, as suggested by Sallandre [2006].

2.1.2.2 The non-illustrative intent (degenerated iconicity)

Without reducing the role of iconicity in SL discourse, early research [Frishberg, 1975] has shown that, although many signs may originate from imitating or miming, their conventionalization is such that they progressively lose their *readily inferable meaning* – that is their immediate iconic property [Macken et al., 1993]. More recent research [Ortega, 2017; Östling et al., 2018], however, brings out clear evidence of iconic origin in both SLs *and* vocal languages. They suggest the presence of a continuum of iconicity across the lexicon of SLs, with various degrees. One should note that if a wide proportion of conventional signs can be iconically motivated – by the physical structure of their referents –, their form can still be seen as arbitrary. For instance, the lexical sign `PLANE` is iconically motivated in British Sign Language (BSL), ASL and Korean Sign Language (KSL), but the three SLs use different hand shapes to represent the form of a plane [Ortega, 2017] (Figure 2.3).

For Cuxac [2000], such units signed without iconic intent but with an iconic origin fall within the category of degenerated or downgraded iconicity. They are units that have stabilized – *i.e.* lexicalized – through time such that they are semantically *dormant*, but which iconic properties can easily be

(a) Transfer of Form and Size. The signer draws a sketch in space, representing the surface of an object with her hands. The shape of her lips and cheeks, her partially closed eyes and her lowered head emphasize the imposing character of this object.



(b) Transfer of Persons (bored person) mixed with a Situational Transfer in frames 3 and 4 (*going round and round in circles*). The signer enacts a bored person, which is particularly visible on her face expression (cheeks and lips), her gaze looking away and her head moving side to side. In frames 3 and 4, the weak (left) hand depicts a reference point (corner of a room), while the dominant (right) hand uses a specific proform to represent a person, which illustrates a person going round and round in circles.

Figure 2.2: Examples of transfers in Highly Iconic Structures (*Structures de Grande Iconicité*), according to the typology of Cuxac [1999]

(a) BSL (b) ASL (c) KSL

Figure 2.3: Hand shapes used in the lexical sign PLANE in British Sign Language (BSL), American Sign Language (ASL) and Korean Sign Language (KSL) [Ortega, 2017]. Although iconically motivated by the shape of the fuselage, some arbitrariness in the precise choice of hand shape can be observed.

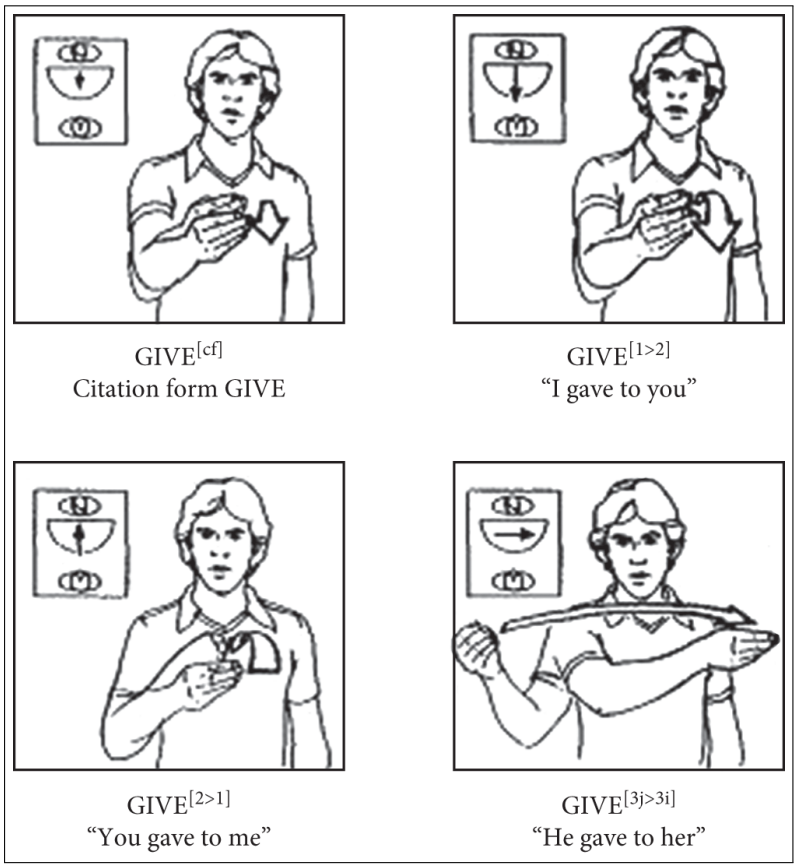


Figure 2.4: Citation form and several variations of the directional verb `GIVE` [Meier, 1987]

reactivated, for instance in SGIs. In other words, even fully conventionalized signs can be modified, in a way that is readily inferable, thanks to the visual modality: iconicity can be *reactivated* in conventionalized signs, if needed [Cuxac, 2000]. An example is given in Figure 2.8 – with more explanations in Section 2.3.4 – in which the degenerated iconicity of the lexical sign `EIFFEL TOWER` is reactivated into a SGI. Individual components of signs – hand shape, orientation, location, movement and facial expression – can thus be individually meaningful, even in conventionalized units. For Cuxac [2000], these components should thus be considered as morphemes instead of phonemes.

2.1.2.3 Diagrammatic iconicity

Present both in the illustrative and non-illustrative intents, the diagrammatic iconicity is a form of syntactical iconicity and consists in using the signing space as a diagram where space, time or entities referents can be built and subsequently reactivated, with a constant use of eye gaze and pointing signs. The iconic use of space thus plays a major role in the construction of SL discourse.

The role of this syntactic iconicity is for instance clear when analyzing what has been called *directional verbs*, which are verbs that vary depending on person, number and location [Padden, 1986]. A very typical example is the verb `GIVE`, that varies depending on person and number, like `SHOW`, `TELL`, `SEND`, etc. Other directional verbs, like `MOVE`, rather depend on location only. An illustration for `GIVE` is shown in Figure 2.4, with the citation form (top left) and three different use cases.

2.1.3 Iconic dynamics – an intermediate approach

Besides the convergent and differentialist approaches, a third and intermediate direction for describing the linguistics of SLs consists in integrating iconicity into traditional parametric linguistic models. This is for instance the case of Millet et al. [2015], who integrates iconicity at the lexical, syntactic and discursive levels in a dynamic fashion. Related work includes [Risler, 2007; Van der Kooij, 2002]. This approach is compatible with the concept of *semantic phonology* [Stokoe, 1991].

2.2 Key Sign Language challenges for recognition

In Section 2.1, we have outlined different approaches for the description of SLs. Building on the main findings of these theories, we want to highlight the key SL properties that must be accounted for in the case of automatic SLR.

2.2.1 General properties

Although there are different conceptions for the linguistics of SLs, we can unequivocally mention three properties of SLs that stem from the visual-gestural modality and require special care in the case of SLR: the *multilinearity*, the specific use of *signing space* and the *iconicity* of SLs [Filhol and Braffort, 2012]. For sake of clarity, we will develop them separately, even though they are inter-dependent.

2.2.1.1 Multilinearity

The property of multilinearity is related to the ability of signers to convey information through the simultaneous use of different language articulators, possibly in an asynchronous way. These articulators include that of the hands, arms, upper body, head and face. Typically, the weak hand might represent a static passive entity, the dominant hand a second entity interacting with the first one while the face and the body posture convey additional information, which is exactly the case of Figure 2.2b. For more detail on the role of simultaneity in SL, refer to [Vermeerbergen et al., 2007].

In terms of SLR, it implies that the temporal correlation between articulators is not guaranteed – *a fortiori* it should not be assumed.

2.2.1.2 Use of space

At the discursive level, the use of space is fundamental. Elements of discourse are indeed introduced spatially, then related to each other visually – in this sense, the signing space is part of the iconic structure of SLs. Previously introduced elements can also be reactivated if needed, for instance using pointing signs. This corresponds to the diagrammatic iconicity (Section 2.1.2.3). A good example is the concept of directional verb (see for instance Figure 2.4), or a detailed sequence analysis in Section 2.3.4.

If SLR is considered as a step towards Sign Language Understanding (SLU) or even Sign Language Translation (SLT), the signing space must obviously be accounted for, as it structures SL discourse.

2.2.1.3 Iconicity

Iconicity is certainly a major challenge for SLR. As described in Section 2.1.2, the iconic properties of SLs disqualify the vision of language as a sequence of conventional units that can be listed. Although

conventional units do exist in SL, many of them possess an iconic origin that can be reactivated if needed, by adding sense to a form that may initially have seemed arbitrary (Section 2.1.2.2).

Furthermore, very complex illustrative structures referred to as transfers or classifier constructions make use of proforms to visually represent the size, shape, movement or action of entities, possibly with strong role shift (Section 2.1.2.1).

2.2.2 Movement epenthesis

An important challenge for Continuous Sign Language Recognition (CSLR) is the presence of movement epenthesis, also called co-articulation. In a comparable way to what happens is natural speech, the transition between signs in natural SL is continuous, with a modification of the beginning and end of signs with respect to their citation form. Therefore, the segmentation of any manual unit can be somehow unclear, especially with low frequency video recordings or fast signers. For short units, the observed form within continuous signing can also be quite different from the citation form.

2.2.3 Discourse types

As it will be developed in Section 2.3, no writing system exists for SLs. However, this does not mean that only one discourse type can be observed, obviously. Sallandre [2003] analyzed the semantic and enunciative categories of LSF, showing that proportions of SGI are highly dependent upon the type of discourse.

Because they are often used as resources for training SLR systems, we also want to discuss the specificities of translated and interpreted SL.

Interpretation consists in translating an utterance *while* it is being produced, either from a spoken language to a SL – the way we focus on – or the opposite direction. The interpreter has to speak both languages, thus this task is realized by hearing people, which first language is generally the spoken language and not the SL.

On the contrary, translation – from a written language to a SL – is usually realized by professional Deaf people, who are not subject to the same time constraints that interpreters have to respect. Practically speaking, translated SL is bound to be more *natural* than interpreted SL, in the sense that the latter results from hard time stresses and the simultaneous presence of the spoken language, which influence on the interpreted message may be strong. Metzger [1999] recalls that there are many differences between translated and interpreted SL, among which the fact that

> *while [translation and interpretation] both aim to convey an equivalent sense of the source message, translators have the time to address linguistic meaning whereas interpreters do not.*

Janzen [2005] also warns about the risk of a reduced usage of visual structures like classifier constructions in interpreted SL. In some extreme cases, it has been suggested that the TV news interpreters use signed English rather than natural SL [Wehrmeyer, 2014, South African TV].

It is important to note that the nature and type of SL discourse will impact the variety and types of linguistic structures used, which may have an effect on the representativeness of the corpus used for SLR, and therefore on the generalizability of the system to other types of discourse. Another

important aspect to take into account for SLR is the annotation, which is used as ground truth during the learning phase.

Depending on the underlying linguistic theory and the type of discourse, the annotation of SL corpora may indeed vary significantly. This is an important aspect to keep in mind for the remainder of this thesis, as SLR systems are trained with and are thus highly dependent upon these annotations. We will develop this point in the next section.

2.3 Transcription and annotation of Sign Languages

SLs have no written forms. Although it is also the case for many vocal languages, no straightforward transcription like the International Phonetic Alphabet (IPA) can be used for SLs.

A few graphical transcription codes have been developed, like HamNoSys or SignWriting (Section 2.3.1). Most frequently, corpora are annotated with a *gloss*-based system that focuses on lexical units (Section 2.3.2). Nevertheless, some SL corpora are annotated with much more detail and account for non-lexical structures (Section 2.3.3).

2.3.1 Transcription into graphical forms

In the case of SL, *transcription* refers to the action of writing down the gestural flow of a SL discourse. It is thus a specific form of annotation. The existing codes are based on the system of Stokoe [1972], either at the phonologic or phonetic level.

The Hamburg Notation System for Sign Languages [Hanke, 2004] (HamNoSys) encodes – in a linear and quite rigid fashion – the parameters described by Stokoe. Although it is relatively effective for the representation of isolated signs, HamNoSys does not allow for a straightforward integration of multilinear and/or non-manual structures of SL. Therefore, it is not adapted to the representation of discourse, characterized by a great use of space and the simultaneous involvement of the different body articulators.

SignWriting [Sutton, 1995] consists in a system of *glyphs*, that are iconic symbols possibly combined to represent manual units. However, its writing rules are known to lack standardization while its implementation in terms of annotation software is tricky.

Other systems have been developed but are not detailed here for sake of brevity. Figure 2.5 presents the HamNoSys and SignWriting encodings for the German Sign Language (DGS) or ASL sign `HOUSE`, along with four other codes.

In conclusion, there is to date no usable effective system for the transcription of SLs. However, this is an active research field with ongoing experiments, like Typannot [Boutet et al., 2020].

2.3.2 Gloss-based annotation for lexical units

By far the most popular annotation system, *glosses* are used to annotate lexical units in SL corpora.

From Section 2.1.2.2, it appears that the arbitrariness of lexical signs of SLs, in the sense of De Saussure, is not necessarily verified (nor is it for other vocal languages, actually). However, this does not mean that conventional signs do not exist. These conventional units, shared across the signers of a Sign Language, form what is usually called a lexicon. In the classification of Johnston and De Beuzeville [2016] (see Section 2.3.3), they are referred to as Fully Lexical Signs (FLSs):

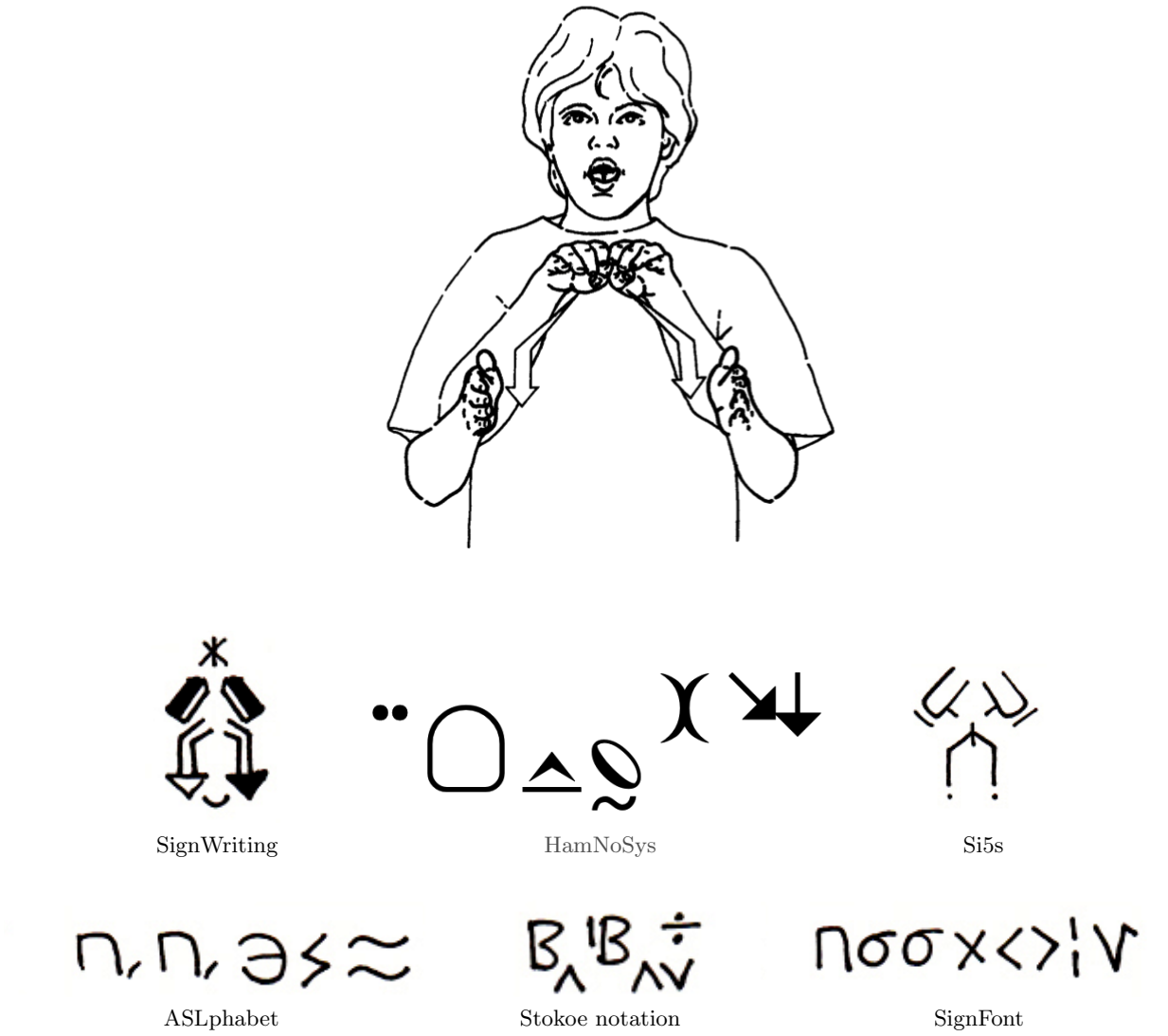


Figure 2.5: Six Sign Language transcription codes for the DGS or ASL sign HOUSE. From [Hanke, 2004] and https://en.wikipedia.org/wiki/American_Sign_Language.

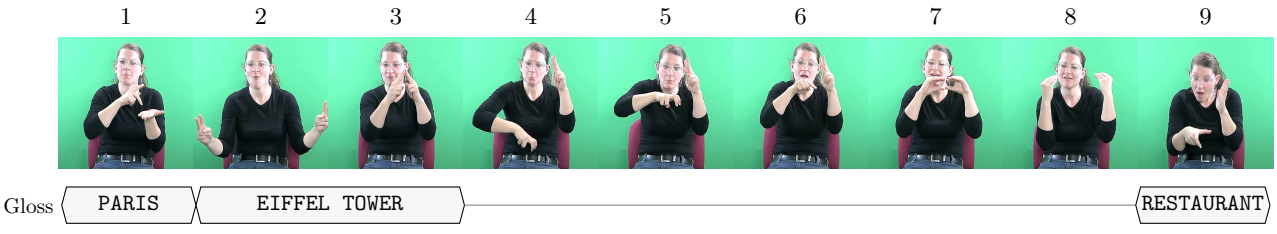
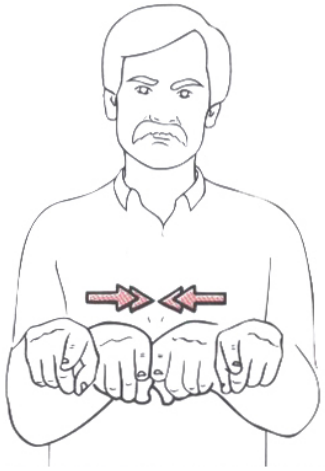


Figure 2.6: LSF sequence from Dicta-Sign–LSF–v2 (video reference: S7_T2_A10, see Chapter 4), with gloss annotations for lexical signs. Frames 4 to 8 can not be described with gloss annotations, as they correspond to SGIs, or classifier predicates (see Section 2.3.4 for a detailed analysis).



(a) LIKE/SAME/ALSO

(b) GLASS1 (c) GLASS2

Figure 2.7: LSF glosses: on top, one gloss corresponds to different meanings (LIKE/SAME/ALSO). On the bottom, two glosses (two forms) for one meaning (GLASS – the solid material). From [Moody et al., 1997].

> *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.*

FLSs are identified by *glosses*, more exactly by *ID-glosses*, which are unique identifiers, related to the form of the sign only, without consideration for meaning. As a result, two signs with the same form but a different meaning will have the same ID-gloss. It is important to note that ID-glosses are chosen somehow arbitrarily and do not represent the meaning of the sign in a given context. They can be composed of a succession of words separated by '/' in order to indicate different possible meanings, without looking for completeness (e.g. LIKE/SAME/ALSO). If different signs express the same concept, they are identified differently (e.g. GLASS1, GLASS2). See Figure 2.7 for an illustration.

Because FLSs can be defined outside of any context, they can be signed in an isolated fashion, which is usually the case in SL dictionaries. This specific case is usually referred to with the term citation-form

lexical sign, corresponding to the most standard and isolated form of a lexical sign.

The main limit of gloss-based annotation systems is that lexical units only account for a fraction of the content of SL discourse: in Figure 2.6, frames 1, 2, 3 and 9 can be *glossed*, as these frames correspond to lexical signs. Frames 4 to 8, however, do not correspond to any known conventional unit, but are well described by SGIs or classifier predicates. These illustrative structures are by nature infinite, and thus can not be annotated using a lexicon.

In order to annotate more content in SL discourse in a consistent fashion, detailed schemes have been developed like that of Johnston and De Beuzeville [2016]. This is presented below.

2.3.3 Classification of units based on the degree of lexicalization

We have chosen to present the classification of Johnston and De Beuzeville [2016], used for the annotation of the Auslan Corpus [Johnston, 2009]. This classification has since been used to finely annotate many SL corpora (see Section 4.1). In this classification, two main categories are listed:

- Lexicalized signs, referred to as Fully Lexical Signs (FLSs)
- Non-conventional, highly context-dependent signs, referred to as Partially Lexical Signs (PLSs)

A third category, referred to as Non Lexical Signs (NLSs), is also included.

2.3.3.1 Fully Lexical Signs

Annotated Fully Lexical Signs (FLSs) correspond to the core of popular annotation systems that use a gloss-based coding (see Section 2.3.2). They are conventionalized units ; a FLS may either be a content sign or a function sign (which roughly correspond to nouns and verbs in English).

2.3.3.2 Partially Lexical Signs

PLSs are formed by the combination of conventional and non-conventional elements, the latter being highly context-dependent. Thus, they can not be listed in a dictionary. For Johnston and De Beuzeville [2016], PLSs are defined by having one or two of the following characteristics:

> *(i) they have little or no conventionalised or language-specific meaning value in addition to that carried by their formational components (e.g. handshape, location, orientation etc.)*
> *(ii) they have a meaning that is incomplete in some way – one needs to refer to the context of utterance [...] in a non-trivial way to 'complete' the meaning of the sign.*

In the PLS category are listed:

Depicting Signs (DSs) or illustrative structures. DSs generally use proforms, and correspond to what is often called *classifier constructions*, or SGIs in the typology of [Cuxac, 1999].

DS-Location (DS-L)	for the location of an entity
DS-Motion (DS-M)	for the motion of an entity
DS-Size&Shape (DS-SS)	for the size and shape of an entity
DS-Ground (DS-G)	for a spatial or temporal reference (ground)
DS-Hold (DS-H)	for the handling of an entity

Pointing Signs (PTSs) or indexing signs

Buoys are hand postures held during some time of a discourse, usually on the weak hand, that are used as physical reference points for a referent [Liddell, 2003].

List Buoy (LBuoy) for a maintained hand posture, with fingers stretched out, that is used to list a certain number of entities

Fragment Buoy (FBuoy) for the hold of a fragment or the final posture of a two-handed lexical sign, usually on the weak hand

Theme Buoy (TBuoy) for an extended finger to mark a *theme* or subject, or even a moment in time

2.3.3.3 Non Lexical Signs

NLSs include:

Fingerspelled Signs (FSs) for proper names or when the sign is unknown

Gestures (Gs) for non-lexicalized gestures, which may be culturally shared or idiosyncratic – these gestures are not assigned an ID-gloss

2.3.4 An annotated example

In order to illustrate the different annotation categories, let us look at the sequence example of Figure 2.8.

This sequence from the Dicta-Sign–LSF–v2 corpus [Belissen et al., 2020a, see Chapter 4] is annotated with a similar code as that of [Johnston and De Beuzeville, 2016].

Three lexical signs are produced (thumbnails 1, 2-3, 9), while thumbnails 4-8 correspond to a Highly Iconic Structure (*Structure de Grande Iconicité*), which is related to the illustrative intent. According to the typology of Cuxac [1999], thumbnails 4 and 5 correspond to a Situational Transfer – representing someone climbing up to the middle of the tower, using a proform on the right hand –, while thumbnails 6 and 7 would be accurately described by a Transfer of Form and Size – representing the shape of a restaurant.

Interestingly, the degenerated iconicity of `EIFFEL TOWER`, which is a conventional unit signed in thumbnails 2-3, is reactivated by thumbnail 4 (a fragment of the tower is maintained so that the visual scene is kept coherent).

Conclusion

Following this linguistic description of SLs, it should be clear that detailed annotation systems are needed to accurately describe SL discourse. Unfortunately most corpora are only transcribed through glosses, that correspond to the very conventionalized manual units of SL, leaving out all other types of structures. Non-manual features are generally ignored, while spatial information is at best over-simplified, when it is accounted for. In light of this conclusion, the next chapter will be dedicated to reviewing the state of the art in automatic SLR.

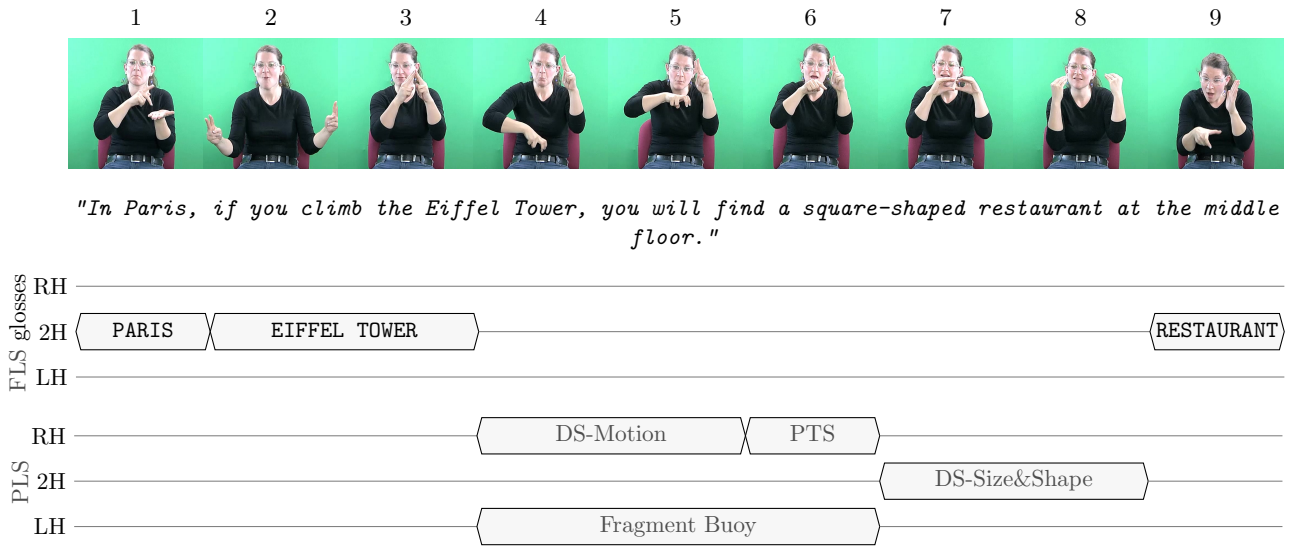


Figure 2.8: LSF sequence from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`, see Chapter 4). Expert annotations for right-handed (RH), two-handed (2H) and left-handed (LH) Fully Lexical Signs (FLSs) and Partially Lexical Signs (PLSs) are given.

Chapter 3

Automatic Sign Language Recognition: state of the art

In this chapter, we aim to review the state of the art in automatic Sign Language Recognition (SLR), that is to describe and analyze the common acceptation of SLR. Building on the linguistic findings presented in Chapter 2, we formalize the problem, first introducing a general framework (Section 3.1), then distinguishing between isolated signs (Section 3.2) and Continuous Sign Language (Section 3.3). In each section, we introduce some popular and important corpora, then we discuss common learning frameworks and associated results. Finally, we discuss a few works that have gone beyond this common acceptation of SLR.

3.1 General framework

Data acquisition often represents the first step of a SLR system. Although early SLR methods used data gloves and accelerometers [Murakami and Taguchi, 1991; Kadous et al., 1996; Braffort, 1996], vision based approaches have progressively become more popular. Whereas gloves offer the clear advantage of a low-dimensional, directly usable feature vector, they are nonetheless very intrusive and strongly restrict the field of application of SLR. On the contrary, there are many publicly available SL video corpora, with different kinds of annotation.

In this thesis, the starting point is assumed to be a SL video recording. Indeed, we have insisted in Section 1.2.1 that our main concern is that SLR systems can widely benefit the Deaf. By *recognition*, we mean the recognition of *elements* within such a video recording, which we will detail later.

In order to formalize the general problem of SLR, let:

- $X = [f_1, \dots, f_T]$ a SL video sequence of T frames
- $\mathscr{R}$ an intermediate representation of X, often called *features*
- $\mathscr{M}$ a learning and prediction model
- Y the element(s) of interest from X
- $\widehat{Y}$ an estimation of Y

The process of SLR can be seen as a function, or model, using $\mathscr{R}$ and $\mathscr{M}$ to estimate Y:

$$X \xrightarrow{\mathscr{R},\mathscr{M}} \widehat{Y} \tag{3.1}$$

The performance of such a model is then evaluated through a function $\mathscr{P}$ that measures the difference between Y and $\widehat{Y}$, the objective being of course that $\widehat{Y}$ is as close as possible to Y:

$$\mathscr{P}(Y, \widehat{Y}). \tag{3.2}$$

Obviously, the performance is always evaluated on videos unseen during training of both $\mathscr{R}$ and $\mathscr{M}$. A crucial setting is the choice of signer-dependency: a signer-independent setting, in which tested signers are excluded from training videos, makes learning a much harder task than a signer-dependent training, but also drastically increases the generalizability of the trained model.

The different categories of SLR rely in the form and content of X and Y. Within each category, different options can be considered for $\mathscr{R}$, $\mathscr{M}$ and $\mathscr{P}$. It is important to note that $\mathscr{R}$ and $\mathscr{M}$, that is the representation of data and the learning-prediction model, are usually chosen in conjunction. Some learning architectures are indeed better adapted to some representations than others. Also, $\mathscr{R}$ and $\mathscr{M}$ are sometimes one and the same, for instance in the case of Convolutional Neural Networks (CNNs) like detailed later.

The next sections formalize these differences and present a state of the art for the past and current research.

3.2 Isolated Lexical Sign Recognition

The case of isolated signs does not actually involve language processing, and is usually considered without referring to SL linguistics. Indeed, we have shown in Section 2.2 that SL is much more complex than a simple process of aligning standard signs. Nevertheless, since a substantial fraction of SLR research focuses on this case, we have decided to include it in this chapter.

Also, it is important to note that the recognition of isolated signs is actually focused on lexical signs only. Non lexical signs are usually context-dependent, hence, the recognition of isolated non lexical signs would hardly make sense. Still, we will use the expression Isolated Lexical Sign Recognition (ILSR) for clarity. It is sometimes referred to as the recognition of citation-form lexical signs (see Section 2.3.2).

In this section, we formalize (Section 3.2.1) the problem of ILSR – objective and performance metric – as well as present common corpora (Section 3.2.2) and associated experiments (Section 3.2.3).

3.2.1 Formalization

The case of models focusing on the recognition of isolated lexical signs falls within the framework of isolated gesture recognition. It is a classification problem, and requires a dictionary of lexical sign glosses, *i.e.* a list of words.

Formally, let $\mathscr{G}$ a dictionary of G lexical sign glosses:

$$\mathscr{G} = \{g^{(1)}, \ldots, g^{(G)}\}. \tag{3.3}$$

In this framework, a SL video i is supposed to contain a unique annotated element of interest $g_i \in \mathscr{G}$ that the prediction model is aimed at recognizing:

$$\begin{cases} Y_{i,\text{ILSR}} &= g_i \in \mathscr{G} \\ \widehat{Y}_{i,\text{ILSR}} &= \widehat{g}_i \in \mathscr{G}. \end{cases} \tag{3.4}$$

The individual recognition performance $\mathscr{P}_i$ is either 0 or 1:

$$\mathscr{P}_i(g_i, \widehat{g}_i) = \mathbb{1}(g_i, \widehat{g}_i) = \begin{cases} 1 & \text{if } \widehat{g}_i = g_i \\ 0 & \text{if } \widehat{g}_i \neq g_i. \end{cases} \tag{3.5}$$

Usually, for a test dataset comprising N videos, the global performance metric $\mathscr{P}$ is then defined as the accuracy, that is:

$$\mathscr{P} = \text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(g_i, \widehat{g}_i) = \frac{\#\text{ correctly recognized signs}}{N}. \tag{3.6}$$

Sometimes, top-n accuracy is also used, with different values of n. Top-n accuracy counts how often the correct class falls in the top n predicted values. By definition, accuracy Acc and top-1 accuracy are equal.

Obviously, the performance is highly dependent upon the corpus type and variability (size of lexicon, number of signers, ...), which is addressed in the following section.

3.2.2 Corpora

In this section, we aim to give an overview of various corpora of isolated lexical signs, popular in the field of ILSR. We do not look for completeness, but rather try to present different types of corpora and associated experiments for ILSR. Some of them also include a Continuous Sign Language (CSL) part, which will be detailed in Section 3.3.

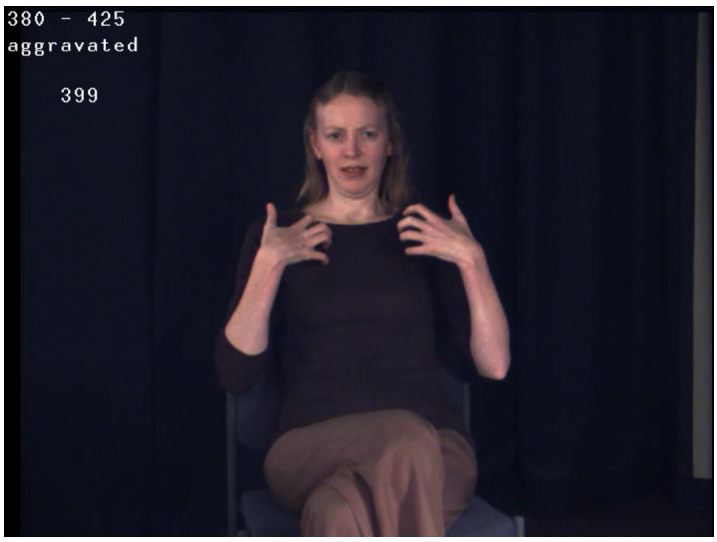
3.2.2.1 American Sign Language Lexicon Video Dataset

The American Sign Language Lexicon Video Dataset (ASLLVD) [Neidle et al., 2012] is a video collection of 2284 American Sign Language (ASL) citation-form lexical signs – 2793 when including variants –, each produced by one to six native signers. The total number of instances reaches 8585.

Although the number of instances per sign is rather low – 3.8, 3.1 when accounting for variants – the size of the lexicon makes ASLLVD a valuable corpus. Furthermore, the recording setup is very well thought (see Figure 3.1), with two front view cameras (640 × 480 at 60 frames per second (fps) and 1600 × 1200 at 30 fps), as well as a side view and head region close-up (640 x 480 at 60 fps). The quite high frame rate and carefully adjusted lighting widen the range of possibilities in terms of image processing methods.

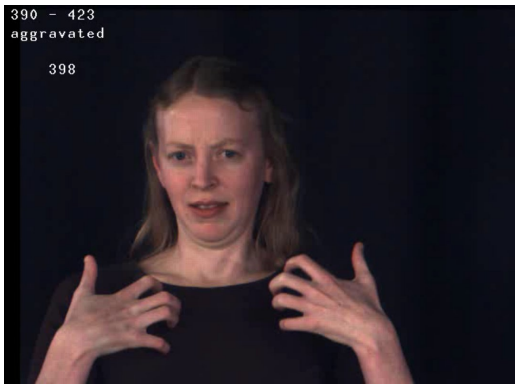
Apart from the lexical sign glosses, the annotations also include information on dominant and weak hand shapes, at the beginning and end of each sign.

Finally, we mention that 2D skeleton estimates calculated with OpenPose (OP) [Cao et al., 2017] have



Front view

Side view



Head region close-up

Figure 3.1: Recording setup for the ASLLVD corpus [Neidle et al., 2012]

Figure 3.2: A YouTube lexical sign video included in the MS-ASL corpus [Joze and Koller, 2018]

been publicly released[1] by de Amorim et al. [2019]. They include body, head and hand joint locations (see Section 6.1.1 for more detail on OP).

3.2.2.2 MS-ASL

The Microsoft American Sign Language (MS-ASL) corpus [Joze and Koller, 2018] is one of a kind, as it is made of a partially automatic gathering of ASL lessons on YouTube. In detail, the authors start with a selection of ASL lessons videos – more precisely, a selection of ASL lexicon lessons. Some shorter ones only teach one sign, while longers ones are aimed at teaching many signs. For instance, the video shown on Figure 3.2 includes 45 food signs. The authors then automatically extracted the glosses from video titles, description or subtitles, depending on the type of video.

As a result, MS-ASL comprises four sets, ASL1000, ASL500, ASL 200 and ASL100, with respectively 1000, 500, 200 and 100 sign classes; 222, 222, 196 and 189 signers; 25513, 17823, 9719 and 5736 instances; 11, 20, 34 and 47 instances per sign class on average. A signer-independent training/validation/test split is released for each of the four sets.

Because of the nature of the corpus, there is a great variability in terms of recording conditions, video quality, lighting, clothing and camera view point. Also, one should note that the number of sign instances per signer, which on average is about 115 in ASL1000, is actually very variable too. This is visible on Figure 3.3, and results from the fact that some teachers are much more active than others on YouTube. In this very uneven distribution, three signers have more than 1000 video samples and ten signers only appear once in the corpus.

A very similar corpus – the Word-Level American Sign Language (WLASL) dataset – is proposed by Li et al. [2020], with 2000 lexical sign classes and four subsets from online ASL lessons.

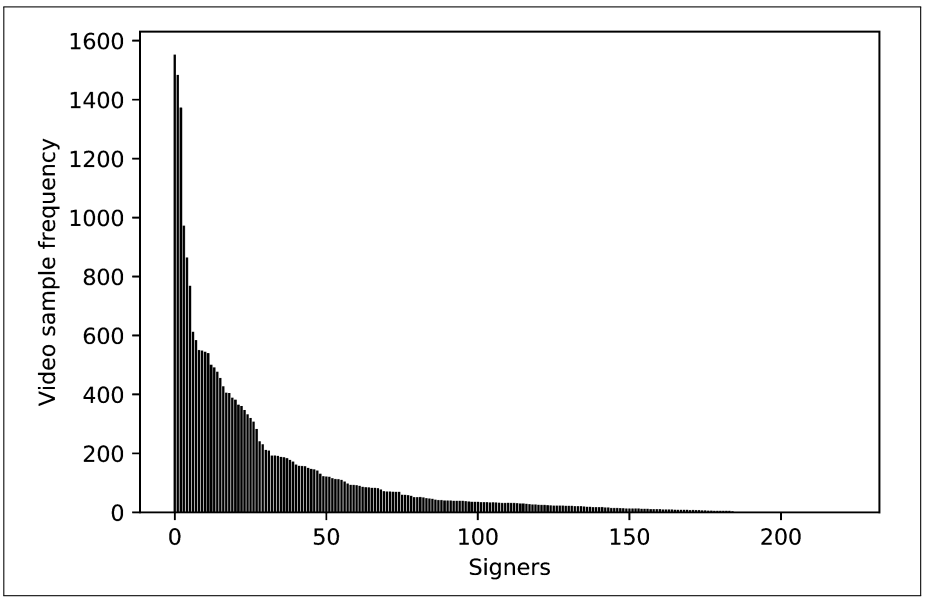


Figure 3.3: Distribution of the number of sign instances per signer in the MS-ASL corpus [Joze and Koller, 2018]. A few signers account for more than 1000 sign instances each, and many signers only have a few instances.

Figure 3.4: Recording setup for the DEVISIGN-L corpus [Chai et al., 2014]. The Kinect setup provides RGB, depth and 2D skeleton data.

Figure 3.5: Random frame from the SIGNUM Database [Von Agris and Kraiss, 2007]

3.2.2.3 DEVISIGN-L

The DEVISIGN-L corpus [Chai et al., 2014] is a Chinese Sign Language (ChSL) corpus recorded with a RGB-Depth (RGBD) setup, covering 2000 lexical signs. The recording is realized with a Microsoft Kinect, with a Red-Green-Blue (RGB) resolution of 1280×720, and a depth resolution of 512×424, both at 30 fps. A 2D skeleton for the upper body is also provided.

Eight signers are involved, with a total number of instances equal to 24000. The recording setup is shown on Figure 3.4. Two smaller subsets are also released under the names DEVISIGN-G and DEVISIGN-D.

3.2.2.4 Isolated SLR500

The Isolated SLR500 (ISLR500) corpus [Pu et al., 2016] is, like DEVISIGN-L (Section 3.2.2.3), a ChSL with RGBD Kinect recording, thus with identical resolution and frame rate, as well as skeleton data.

This dataset comprises 500 lexical signs, each produced five times by each of the 50 signers. The total number of instance is, thus, 125000.

3.2.2.5 SIGNUM Database

The SIGNUM Database [Von Agris and Kraiss, 2007] is a German Sign Language (DGS) corpus, including 450 lexical signs and 25 signers. A reference signer performs the signs three times, whereas the 24 others only realize them once. The total number of instances is equal to 12150. The recording format is RGB, with 776×578 resolution at 30 fps.

As can be seen in Table 3.1 which provides a brief overview of the characteristics in popular datasets

[1] https://www.cin.ufpe.br/~cca5/asllvd-skeleton/

	SL	Data	Signers	Lexicon	Instances
ASLLVD [Neidle et al., 2012]	ASL	RGB 30-60 fps	6	2284	8585 ($\simeq$ 1 per signer per sign)
DEVISIGN-L [Chai et al., 2014]	ChSL	RGBD skeleton	8	2000	24000 (1 or 2 per signer per sign)
ISLR500 [Pu et al., 2016]	ChSL	RGBD skeleton 30 fps	50	500	125000 (5 per signer per sign)
MS-ASL [Joze and Koller, 2018]	ASL	RGB	222	1000	25513 (from 1500 per signer per sign to 1 per signer per sign)
SIGNUM [Von Agris and Kraiss, 2007]	DGS	RGB 30 fps	25	450	12150 (1 per signer per sign)

Table 3.1: Overview of the main characteristics in popular datasets of isolated lexical signs: Sign Language, data formatting, number of signers, size of lexicon and number of sign instances.

used for ILSR, the field of ILSR still lacks large scale multi-signer corpora. The most recent corpus, MS-ASL, shows an interesting variability in terms of signers and lexicon, however it is highly imbalanced (a few signers account for most sign instances).

3.2.3 Experiments: signer representation, learning frameworks and results

In this section, we focus on the past and recent trends in ILSR, in terms of signer representation $\mathscr{R}$ and learning-prediction model $\mathscr{M}$. While some surveys describe the two components separately [Ong and Ranganath, 2005; Cooper et al., 2011], we prefer to insist on the importance of pairing them accordingly. Results of these models are then outlined, mostly on the representative corpora listed in Section 3.2.2. This section is organized in five sub-sections, corresponding to different types of approaches to ILSR.

3.2.3.1 Early vision-based approaches

Early vision-based methods focused on so-called Manual Features (MFs), compatible with the model of Stokoe [1972]: they include hand pose – shape or configuration and orientation – as well as hand position. The dynamics of these parameters can also be included. Hand tracking methods were first employed, with colored gloves, then skin segmentation techniques, which also led to accounting for Non-Manual Features (NMFs).

In [Von Agris et al., 2008], the authors use a generic skin color model to detect and track the face and hands. Static and dynamic MFs are derived, with computations like area, compactness or eccentricity of each hand. An Active Appearance Model (AAM) [Edwards et al., 1998] is used to model the face, using Principal Component Analysis (PCA). A 2D estimate on the facial pose is then obtained, which enables the calculation of facial features like eyebrows or lips shape. On the 450 isolated lexical signs of the SIGNUM Database (Section 3.2.2.5), they get respectively 96.9% and 80.2% accuracy when training in a signer-dependent and -independent fashion, with both manual and facial features. More details are given in Table 3.2.

Setting	Features		
	Manual	Facial	Combined
Signer-Dependent (SD)	94.4%	37.1%	**96.9**%
Signer-Independent (SI)	78.7%	10.0%	**80.2**%

Table 3.2: Detailed Isolated Lexical Sign Recognition results (accuracy) of [Von Agris et al., 2008] on the 450 isolated lexical signs of the SIGNUM Database [Von Agris and Kraiss, 2007]. Signer-dependent and -independent settings are evaluated, with manual features, facial features or both as input.

Setting	Top-1	Top-5	Top-10
Signer-Dependent (SD)	67.3%	86.6%	89.8%
Signer-Independent (SI)	54.4%	77.3%	82.7%

Table 3.3: Detailed Isolated Lexical Sign Recognition results of [Pu et al., 2016] on 100 isolated lexical signs from the ISLR500 corpus. Signer-dependent and -independent settings are evaluated, with top-1, top-5 and top-10 accuracy.

Traditional vision-based methods, using features like optical flow are still being used. In the work of Lim et al. [2016], background subtraction is used to help hand detection, then a block-based histogram of optical flow is computed on the hands region. A simple histogram distance is used for ILSR. Their model is evaluated on small subsets of three ASL corpora: RWTH-Boston-50 [Zahedi et al., 2005], Purdue RVL-SLLL [Martínez et al., 2002] and ASLLVD [Neidle et al., 2012]. The performance is moderate, with for instance 85% accuracy on a subset of 20 lexical signs from ASLLVD, in a signer-independent setting.

3.2.3.2 The importance of 3D

The progressive use of 3D data from RGBD recording or 3D reconstruction methods led to more consideration for the importance of trajectory. Pu et al. [2016] built the ISLR500 corpus (Section 3.2.2.4) and used the depth data from a Kinect recording to analyze the trajectories of hands during production of signs. First, curve features are derived from a codebook training with K-means algorithm. Curve segmenting is realized with a Discrete Contour Evolution algorithm, then a Hidden Markov Model (HMM) model[2] is trained for classification. On a subset of 100 isolated lexical signs from the ISLR500 corpus, they get respectively 67.3% and 54.4% accuracy when training in a signer-dependent or -independent fashion. More details are given in Table 3.3

The work of Dilsizian et al. also highlighted the importance of accounting for 3-dimensionality in ASL. Dilsizian et al. [2014] used the data from the ASLLVD dataset to develop and train a hand tracker and 3D handshape classifier, with 87 hand shapes, using Histogram of Oriented Gradients (HOG) features and a Spectral Latent Variable Model (SLVM). Dilsizian et al. [2016] also analyzed the importance of 3D motion trajectory on a small RGBD ASL corpus, in which they trained a SVM-HMM (Support

[2]HMMs have been very popular in the Machine Learning (ML) community and broadly applied to speech recognition, handwriting recognition and SLR – among others. HMMs use the assumption that a signing sequence is a Markov process describing how hand locations change through the sign production, and a certain number of states are used to represent different parts of the signing action.

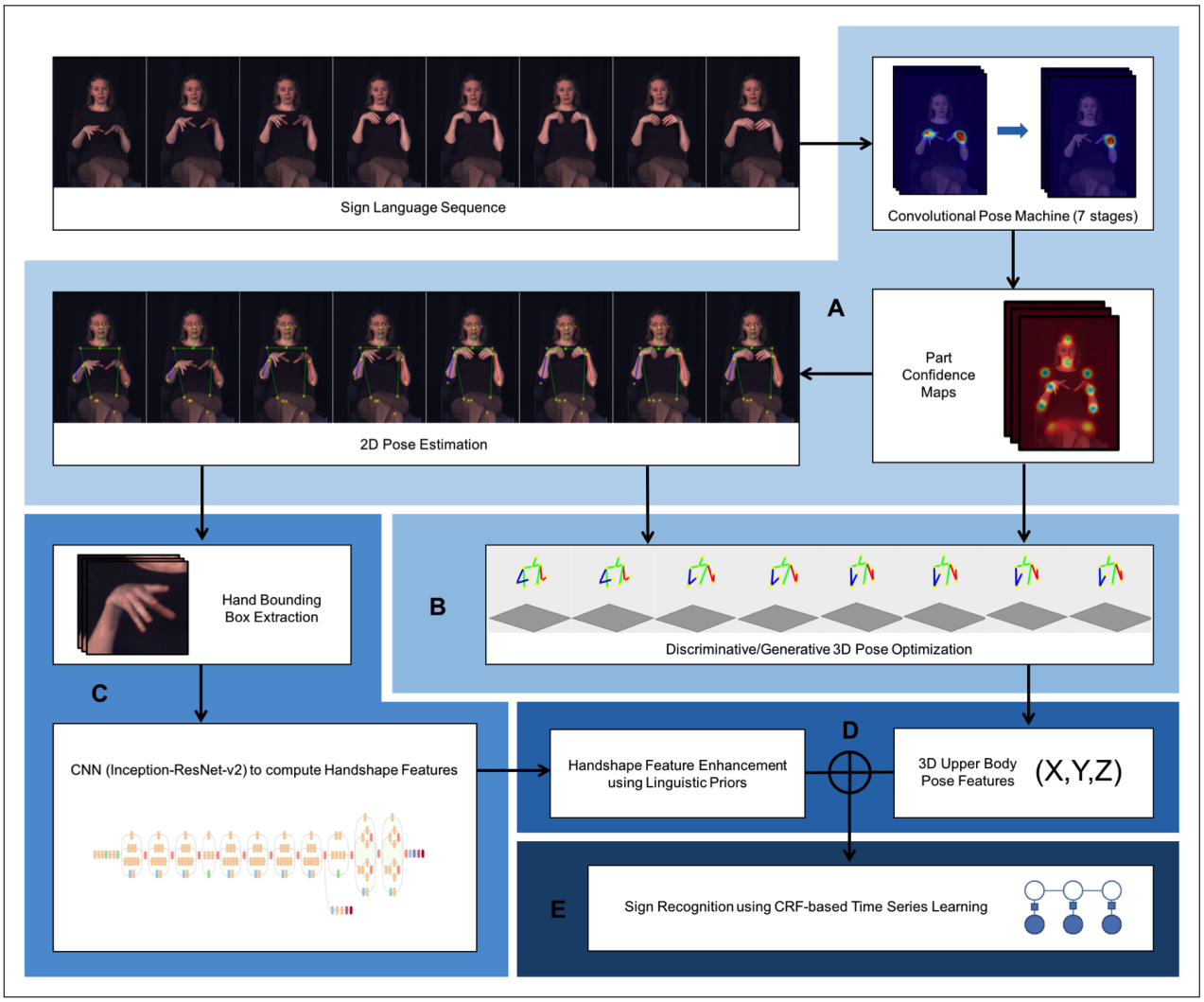


Figure 3.6: Isolated Lexical Sign Recognition framework in [Dilsizian et al., 2018]

Vector Machine-Hidden Markov Model) to distinguish between signs with similar handshapes but different hand trajectories.

3.2.3.3 The importance of linguistic input

More recently, Dilsizian et al. [2018] used their previous work to build a complete ILSR framework, shown on Figure 3.6. In this model, a first CNN model inspired by [Wei et al., 2016] is trained and used to build a discriminative/generative 3D upper body pose and head pose estimator. Another CNN model learns hand shape features, which are then completed by linguistic-inspired sign-level features like the relationship between start and end hand shapes, as well as that between dominant and weak hands. Probabilities of relevant contacts are also computed thanks to the 3D body pose estimate. In the learning phase, they use Conditional Random Fields (CRFs). Their recognition results on a subset of 350 signs from the ASLLVD corpus are very good, given the very low number of instances per lexical sign on the corpus. Top-1 and top-5 accuracy amount respectively to 93.3% and 97.9%, with a signer-dependent setting. The benefit of including the different features can be seen in Table 3.4.

Features	Top-1	Top-5
3D pose	64.3%	81.3%
2D pose + hand shape	80.0%	89.8%
3D pose + hand shape	90.0%	92.9%
3D pose + hand shape + contact events	92.1%	94.6%
3D pose + hand shape + all linguistic parameters	**93.3**%	**97.9**%

Table 3.4: Detailed Isolated Lexical Sign Recognition results of [Dilsizian et al., 2018] on a subset of 350 signs from the ASLLVD corpus [Neidle et al., 2012], in a signer-dependent setting, with top-1 and top-5 accuracy.

3.2.3.4 Covariance matrix–based representation

Wang et al. [2016] used a simple representation of signers in videos, applied to the RGBD and skeleton data of the DEVISIGN-L dataset. Hand features are derived by using HOG, while body features are obtained by pairwise relative position on the skeleton data. Then, a Grassmann Covariance Matrix (GCM) is computed for the feature vector, and the Grassmann metric is used as a kernel in a SVM classifier for ILSR. Experiments are conducted on the DEVISIGN-L dataset (Section 3.2.2.3) with 1000 lexical signs. Reported accuracy reaches respectively 92.4% and 70.9% when training in a signer-dependent or -independent fashion. Wang et al. [2019] presented a refined model, with hierarchical GCM, but no performance improvement on DEVISIGN-L are reported.

3.2.3.5 Most recent methods: Convolutional and Recurrent Neural Networks

RGB input

Recent methods for action recognition have given a lot of importance to CNNs. They can be used and trained to detect body keypoints, like in [Dilsizian et al., 2018], or directly to learn global signer features on images, although they require a lot of training data, like in the gesture recognition task of [Pigou et al., 2014]. Regarding the case of gesture recognition, CNNs are known to be a good match with Recurrent Neural Networks (RNNs), which are very effective at learning temporal dependencies. Pigou et al. [2018] showed that temporal convolutions could significantly improve recognition performance in a Recurrent Convolutional Neural Network (RCNN) framework.

In the specific case of isolated gestures or signs, 3DCNNs, that is networks computing convolutions both in space and time, have also become quite popular. They are used by Wu et al. [2016] inside a HMM framework, to enable fusion of other features. In the work of Joze and Koller [2018], the 3DCNN architecture I3D [Carreira and Zisserman, 2017] is compared to many different others on the MS-ASL dataset. I3D is shown to be the best performing method, with detailed results presented in Table 3.5. We can also cite Liao et al. [2019], who have demonstrated the effectiveness of mixing the Faster R-CNN algorithm [Ren et al., 2015] for hand tracking, 3D convolutions and Bidirectional LSTM (BLSTM) layers for ILSR. Their results on ISLR500 and a subset of DEVISIGN-L are included in Table 3.6.

Skeleton input

Instead of realizing convolutions between close pixels on RGB frames, Spatial-Temporal Graph Convolutional Networks (ST-GCNs) [Yan et al., 2018] use a graph representation of skeletons to compute

Subset	Top-1	Top-5
ASL100	81.8%	95.2%
ASL200	82.0%	93.8%
ASL500	72.5%	89.8%
ASL1000	57.7%	81.1%

Table 3.5: Detailed Isolated Lexical Sign Recognition results of [Joze and Koller, 2018] on the four subsets of the MS-ASL corpus. All results are signer-independent, with top-1 and top-5 accuracy.

convolutions between adjacent body joints, both in space and time. They have been applied to ILSR on the *skeleton* version of ASLLVD [de Amorim et al., 2019], with mediocre results compared to Lim et al. [2016] and *a fortiori* to [Dilsizian et al., 2018] (see Table 3.6).

In conclusion, although we have insisted on the fact that ILSR is a very different task from Continuous Sign Language Recognition (CSLR), it appears that some features and frameworks can be very effective for sign recognition, and may also be used for CSLR – among others, the 3D representation of signers, hand shapes, sign-level linguistic features as well as convolutional and recurrent Neural Networks (NNs).

Corpus	Paper	N	SD	SI
ASLLVD	[Lim et al., 2016]	20 (subset)	-	85.0%
	[de Amorim et al., 2019]⋆	20 (subset)	61.0%⋆⋆	
		2745	16.5%⋆⋆	
	[Dilsizian et al., 2018]	350 (subset)	93.3%	-
MS-ASL	[Joze and Koller, 2018]	100 (subset)	-	81.8%
		200 (subset)	-	82.0%
		500 (subset)	-	72.5%
		1000	-	57.7%
DEVISIGN-L	[Liao et al., 2019]	500 (subset)	89.8%	-
	[Wang et al., 2016, 2019]	1000	92.4%	70.9%
ISLR500	[Pu et al., 2016]	100 (subset)	67.3%	54.4%
	[Liao et al., 2019]	500	86.9%	-
SIGNUM	[Von Agris et al., 2008]	450	96.9%	80.2%

Table 3.6: Reported accuracy of methods detailed in Section 3.2.3 applied to Isolated Lexical Sign Recognition on the corpora presented in Section 3.2.2. N is the number of lexical signs in the corpus, and SD and SI stand for Signer-Dependent and Signer-Independent.
⋆ Using the ASLLVD-skeleton data.
⋆⋆ Unclear whether SD or SI.

3.3 Continuous Lexical Sign Recognition

The case of Continuous Sign Language (CSL) is of different nature to the case of isolated lexical signs, discussed in Section 3.2. Indeed, the production of isolated signs is analogous to the production of isolated gestures, in the sense that language and linguistics do not play any major role, if any. As a matter of fact, we can only regret that many research articles mention Sign Language Recognition in their title, abstract or claimed objectives, while they focus on isolated lexical signs only.

The study of CSL is, on the other hand and as in the case of any language, an extremely difficult task. Language recognition problems involve linguistic issues, as well as difficulties related to signal processing. In the case of SLs, these two questions are crucial and require special care, as they must be dealt with in a very different manner than usual vocal or written languages.

In this section, we focus on the most common acceptation of CSLR, that is better qualified as Continuous Lexical Sign Recognition (CLexSR). After formalizing this framework (Section 3.3.1), popular corpora (Section 3.3.2) and different types of experiments (Section 3.3.3) are presented.

3.3.1 Formalization

In the case of CSL, most research has focused on the recognition of lexical signs within continuous signing. It sets aside the spatial grammar and iconicity of SLs, but still enables to understand simple utterances. It can also be seen as an efficient way to build towards sign spotting models. Recognition can be aligned or not, as described in the next sections.

3.3.1.1 Unaligned recognition

A first setting for CLexSR – by far the most popular – is, for a SL sequence, to aim at the recognition of the sequence of produced lexical signs, omitting temporal information. In this case, the recognized lexical signs are not aligned with the original video frames, therefore we choose to call it *unaligned recognition*.

Formally, let $\mathscr{G}_U$ a dictionary of G lexical sign glosses:

$$\mathscr{G}_U = \{g^{(1)}, \dots, g^{(G)}\}. \tag{3.7}$$

In this framework, a SL video X is supposed to contain N consecutive lexical signs ($N \geq 1$). We assume $\widehat{N}$ lexical signs are recognized, such that:

$$\begin{cases} Y_{\text{CLexSR,U}} &= \begin{bmatrix} g_1 & \cdots & g_N \end{bmatrix}, g_i \in \mathscr{G}_U \\ \widehat{Y}_{\text{CLexSR,U}} &= \begin{bmatrix} \widehat{g}_1 & \cdots & \widehat{g}_{\widehat{N}} \end{bmatrix}, \widehat{g}_i \in \mathscr{G}_U. \end{cases} \tag{3.8}$$

Note than in general, $N \neq \widehat{N}$, so $Y_{\text{CLexSR,U}}$ and $\widehat{Y}_{\text{CLexSR,U}}$ have different lengths.

The sequence-wise recognition performance $\mathscr{P}_{\text{CLexSR,U}}$ is then usually defined as 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 into the expected sequence of length N:

$$\mathscr{P}_{\text{CLexSR,U}}(Y, \widehat{Y}) = \text{WER} = \frac{I + S + D}{N}. \tag{3.9}$$

Let us note that Gloss Error Rate would be a more appropriate naming of this metric. In any case, whether *Word* or *Gloss*, we insist on the fact that the choice of metric says a lot on the – often implicit – linguistic assumptions on SL. In this particular setting, a SL production is always assumed to be reducible to a sequence of standard signs, seen as words. Based on the linguistic description of Chapter 2, the serious limits of this assumption should appear clearly.

Furthermore, even in the field of Automatic Speech Recognition, the use of WER is questioned and better alternatives are developed that are better correlated with human performance [Favre et al., 2013].

For information, a related metric is sometimes used, called *Word Accuracy*, and defined as:

$$\text{WAcc} = 1 - \text{WER}. \tag{3.10}$$

However, WER can be greater than 1, and WAcc smaller than 0, so the term *Accuracy* is not exactly appropriate.

3.3.1.2 Aligned recognition

The case we choose to call *aligned recognition* is slightly more informative than that of *unaligned recognition*. Indeed, every video frame is then assigned a label, which can be either a sign from the dictionary or nothing (the null class).

Formally, let $\mathscr{G}$ a dictionary of G lexical sign glosses, plus a null class $g^{(0)}$:

$$\mathscr{G}_A = \{g^{(0)}, g^{(1)}, \dots, g^{(G)}\} = \mathscr{G}_U \cup \{g^{(0)}\}. \tag{3.11}$$

If the model is able to align predicted lexical signs with video frames, the output can be expressed as a sequence, the length of which is equal to the original sequence length T:

$$\begin{cases} Y_{\text{CLexSR,A}} &= \begin{bmatrix} g_1 & \cdots & g_T \end{bmatrix}, g_t \in \mathscr{G}_A \\ \widehat{Y}_{\text{CLexSR,A}} &= \begin{bmatrix} \widehat{g}_1 & \cdots & \widehat{g}_T \end{bmatrix}, \widehat{g}_t \in \mathscr{G}_A \end{cases} \tag{3.12}$$

In this case, the accuracy Acc defined as the ratio of correctly labeled frames over the total number of frames T is generally used as a straightforward performance metric:

$$\mathscr{P}_{\text{CLexSR,A}}(Y, \widehat{Y}) = \text{Acc} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(g_t, \widehat{g}_t) = \frac{\#\text{ correctly labeled frames}}{T} \tag{3.13}$$

where $\mathbb{1}$ is the identity function, defined as:

$$\mathbb{1}(g_t, \widehat{g}_t) = \begin{cases} 1 & \text{if } \widehat{g}_t = g_t \\ 0 & \text{if } \widehat{g}_t \neq g_t. \end{cases} \tag{3.14}$$

As for the case of ILSR, the recognition performance is necessarily highly dependent upon the corpus type. Different types of corpora are thus detailed in the next section.

3.3.2 Corpora

In this section, we have chosen to focus on the three corpora mainly used in the field of CLexSR.

3.3.2.1 SIGNUM Database

Adding to the isolated lexical signs corpus presented in Section 3.2.2.5, the DGS SIGNUM Database [Von Agris and Kraiss, 2007] also integrates a CSL part. More precisely, 780 *sentences* are elicited, based on the 450 lexical signs contained in the isolated part of the database. In total, approximately five hours of video are recorded.

The elicited sentences are shown to the signers in the form of a gloss sequence, along with a video reference, which is visible on Figure 3.7. Because of the rigorous elicitation procedure, it is safe to say that the level of spontaneity as well as the interpersonal variability in the observed SL are very low. Signers indeed repeat the original reference *sentences* with no initiative. The elicited gloss sequences have lengths ranging from two to eleven glosses, with a few random examples given in Table 3.7.

3.3.2.2 RWTH-Phœnix-Weather

The RWTH-Phœnix-Weather (RWTH-PW) corpus [Forster et al., 2012, 2014] is made from 11 hours of live DGS interpretation of weather forecast on German television. The 2012 release contains a single signer, therefore, we will focus on the 2014 version, also referred to as RWTH-Phœnix-Weather-2014-Multisigner. It is made public by Koller et al. [2015], with gloss annotation for the lexical signs observed in the signed recordings. Koller et al. [2017] then released a signer-independent version, named RWTH-Phœnix-Weather-2014-Signerindependent (RWTH-PW-2014-SI5). Last, a release with German translations was published by Camgoz et al. [2018] (RWTH-PW-2014-T).

Nine signers are present in the corpus, wearing dark clothes in front of an artificial grey background (see Figure 3.8) and the video format is RGB, with a quite low resolution of 210×260 pixels at 25 fps.

This corpus has established itself as a reference dataset for SLR, with many experiments detailed in Section 3.3.3.2. Conversely to many corpora produced in laboratory conditions and/or strong elicitation rules, RWTH-PW has been described by its authors as *real life data* [Forster et al., 2012].

However, because of the specific topic, the language variability and complexity are bound to be limited. Furthermore, it is crucial to note that interpreted SL is necessarily a specific type of SL, quite 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 (see Section 2.2.3).

This observation is actually shared by the original authors of the RWTH-PW corpus [Forster et al., 2012]:

> *Moreover, the domain of weather forecasting features a limited vocabulary and a restricted use of specific sign language phenomena such as classifier signs.*

Even though we are not able to qualify the quality of the interpretation in the RWTH-PW corpus, one should be careful regarding the generalizability of the results for SLR models trained on this corpus.

Figure 3.7: Elicitation procedure for the Continuous Sign Language part of the SIGNUM Database [Von Agris and Kraiss, 2007]. The signers are shown a gloss sequence along with a reference video.

Ref	GERMAN GLOSS SEQUENCE *ENGLISH GLOSS SEQUENCE* German translated sequence *English translated sequence*
116	ICH SCHOKOLADE KAUFEN *I CHOCOLATE BUY* Ich kaufe Schokolade. *I am buying chocolate.*
165	ICH WEIN UND BIER neg-MÖGEN *I WINE AND BEER neg-LIKE* Ich mag keinen Wein und kein Bier. *I like neither wine nor beer.*
418	WIR-beide JETZT FRISCH LUFT BRAUCHEN *WE-both NOW FRESH AIR NEED* Wir brauchen jetzt frische Luft. *We need some fresh air now.*
713	FREIZEIT DU OFT-häufig TANZEN KOMMEN-nach? *LEISURE-TIME YOU OFTEN-frequently DANCE COME-to?* Gehst du in der Freizeit oft tanzen? *Do you often go dancing in your leisure time?*

Table 3.7: Random elicitation sequences from the SIGNUM Database [Von Agris and Kraiss, 2007], with German and English translations.

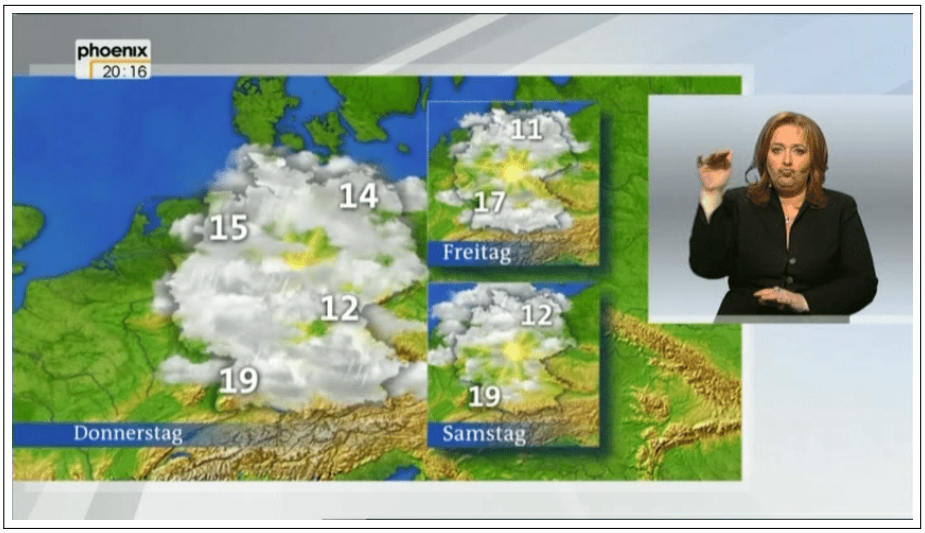


Figure 3.8: Weather forecast on German television, with live DGS interpretation that forms the RWTH-Phœnix-Weather corpus [Forster et al., 2012, 2014]

Ref	GERMAN GLOSS SEQUENCE *ENGLISH GLOSS SEQUENCE* German translated sequence *English translated sequence*
-	SAMSTAG WECHSELHAFT *SATURDAY CHANGING* Am Samstag ist es wieder unbeständig. *On Saturday it is changing again.*
-	BESONDERS FREUNDLICH NORDOST BISSCHEN BEREICH *ESPECIALLY FRIENDLY NORTH-EAST LITTLE-BIT AREA* Am freundlichsten ist es noch im Nordosten sowie in teilen Bayerns. *It is friendliest still in the north-east as well as parts of Bavaria.*
-	SONNTAG REGEN TEIL GEWITTER *SUNDAY RAIN PART THUNDER-STORM* Am Sonntag ab und an Regenschauer teilweise auch Gewitter. *On Sunday rain on and off and partly thunderstorms.*

Table 3.8: Random annotated gloss sequences from the RWTH-Phœnix-Weather corpus [Forster et al., 2012, 2014], with German and English translations.

Ref	Chinese elicitation sequence	English translated sequence
16	他的外祖母是园丁	His grandmother was a gardener.
33	你婆婆是盲人	Your mother-in-law is blind.
74	他的盆是绿的	His pot is green.
96	我们的婚姻是幸福的	We are happily married.

Table 3.9: Random elicitation sequences from the Continuous SLR100 corpus [Huang et al., 2018], with English translations.

3.3.2.3 Continuous SLR100

The Continuous SLR100 (CSLR100) dataset [Huang et al., 2018] is a continuous ChSL corpus, associated to the database of isolated signs ISLR500 presented in Section 3.2.2.4. The format is RGBD, with a recording setup similar as that of Figure 3.4. 2D skeleton data is also provided.

In terms of duration, this corpus is definitely major. Indeed, more than 100 hours are recorded. However, the level of variability and spontaneity in the produced language is very low: the corpus is based on 100 pre-defined *sentences*, that are repeated five times by 50 signers. In total, 25000 sentences are thus recorded. The lexicon only amounts to 178 different lexical signs. A few examples are given in Table 3.9.

Corpus	SL	Source	Signers	Hrs.	Discourse type	Translation
CSLR100	ChSL	RGBD skeleton	50	100	Artificial	-
RWTH-PW	DGS	RGB	9	11	Interpreted	German
SIGNUM	DGS	RGB	25	5	Artificial	German/English

Table 3.10: Three main corpora used for Continuous Lexical Sign Recognition.

Although the three corpora we have described are not representative of the diversity of discourse types in SL, they have been the testing ground for many experiments of CLexSR, which we present below.

3.3.3 Experiments: frameworks and results

In this section, we focus on the experiments of CLexSR, mainly on the three corpora presented in Section 3.3.2. We start by going back to the way movement epenthesis was dealt with by the first CLexSR architectures, then we develop the recent developments with more complex neural architectures, mainly applied to the RWTH-PW corpus.

3.3.3.1 Developing architectures to account for movement epenthesis

A first notable difference between the recognition of citation-form lexical signs and signs within continuous signing is the movement epenthesis, also called co-articulation (see Section 2.2.2).

HMMs enable to model explicitly the transition between signs in CSL. This is demonstrated by Braffort [1996], using data gloves, then by Vogler and Metaxas [1997], with RGB video input. Fang

et al. [2001] show that embedding a Simple Recurrent Network into a HMM framework can deal with co-articulation, and trained their model for the recognition of continuous ChSL in a signer-independent fashion. In [Gao et al., 2004], an even more refined HMM-based transition model is proposed, with a modified k-Means algorithm for spatio-temporal clustering.

3.3.3.2 Modern architectures for CLexSR on bigger corpora

An ongoing competition for the best CLexSR results on the RWTH-Phœnix-Weather corpus was initiated by a sophisticated model presented by Koller et al. [2015]. In their model, the authors use dynamic programming to track hands, then extract HOG-3D features, imitating the work of Dilsizian et al. [2014], as well as inter-hand features. High-level facial features are obtained by using an AAM [Edwards et al., 1998]. A class language model is used in addition to a more classical HMM framework, and the inter-signer variability is accounted for with a constrained maximum likelihood linear regression. They tested their model on both RWTH-PW and SIGNUM.

Thereafter, CNNs have become more and more popular and predominantly used as an effective way to derive visual features. Koller et al. [2016a] embedded a CNN into an iterative Expectation Maximization (EM) algorithm in order to train Deep Hand, a powerful hand shape classifier, on weak labels. Training is realized on data from three SLs, namely DGS, New Zealand Sign Language (NZSL) and Danish Sign Language (DTS). Interestingly, for each lexical sign, weak hand labels are obtained by parsing SignWriting[3] dictionaries, then extracting information on hand shape and leaving out hand pose related data. Finally, the authors used Deep Hand instead of the HOG-3D for hand features in the model of [Koller et al., 2015], with improved results. Later, the authors built a unified CNN-HMM model, trained in an end-to-end fashion [Koller et al., 2016b].

Similarly to [Koller et al., 2016a], Camgoz et al. [2017] trained SubUNets, a CNN-BLSTM network trained for hand shape recognition and CLexSR, in an end-to-end fashion, with a Connectionist Temporal Classification (CTC) loss. The same kind of model is proposed by Cui et al. [2017]. Koller et al. [2017] then released a new model, consisting of embedding a CNN-BLSTM into a HMM, and treat the annotations as weak labels. Thanks to several EM re-alignments, the performance improves significantly, both on RWTH-PW and on SIGNUM, with WER of 26.8% and 4.8% on the respective signer-dependent test sets. Moreover, they also tested their model on the signer-independent version of RWTH-PW and obtained a WER of 44.1%, that is a relative 65% higher, which shows that the signer-independence is a challenge that should not be overlooked.

Using two CNN streams – one for the hands and a global one – for feature extraction, Huang et al. [2018] used a combination of Long Short-Term Memory (LSTM) and Attention [Luong et al., 2015] to tackle the temporal modality, with an encoder-decoder architecture, along with a Dynamic Time Warping (DTW) algorithm. They published results on RWTH-PW and the signer-independent version of the Continuous SLR100 dataset. Note that this version consists of testing on unseen signers, but on sentences seen during training, which makes the learning task much easier.

Recently, 3DCNNs – that is convolution blocks taking videos as input – have proven effective for action recognition, and have progressively replaced traditional 2D convolutions. Applied to CLexSR on the RWTH-PW and Continuous SLR100 datasets, this is the case for Guo et al. [2018, 2019a,b]; Pu et al. [2019]; Yang et al. [2019]; Zhou et al. [2019]. The LSTM encoder-decoder architecture with Attention is used by Guo et al. [2018] and Pu et al. [2019], and CTC decoding by Guo et al. [2019a,b]; Pu et al.

[3] SignWriting [Sutton, 1995] is a simplified pictorial representation of signs, with open online resources. See Section 2.3.1.

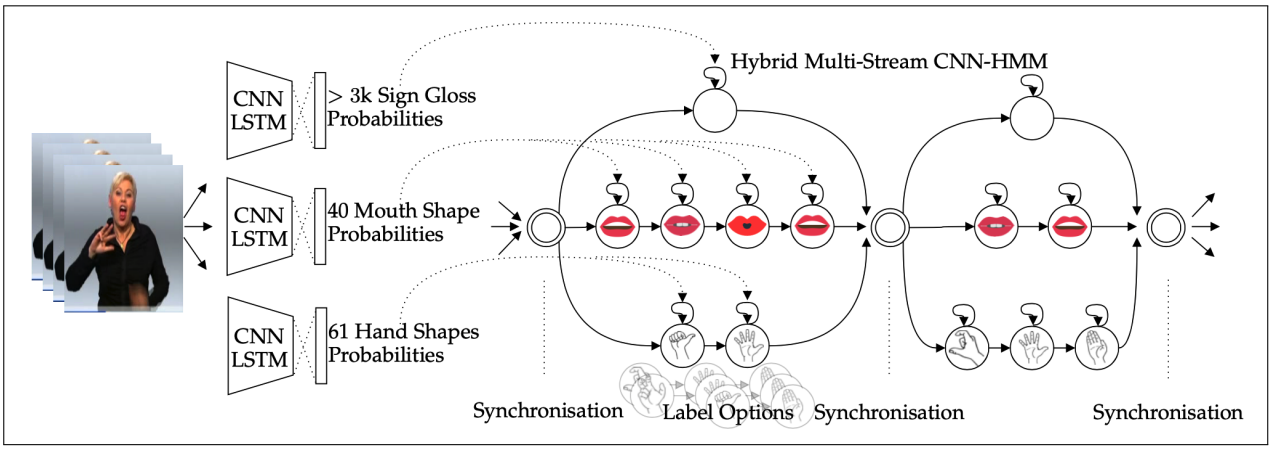


Figure 3.9: Continuous Lexical Sign Recognition framework in [Koller et al., 2019], using a 3-stream CNN-LSTM-HMM

[2019]; Yang et al. [2019]; Zhou et al. [2019]. Guo et al. [2019a,b] also compute temporal convolutions.

Recently, Koller et al. [2019] used different sources of data to train a sophisticated multi-stream CNN-LSTM embedded into a HMM framework. They indeed trained the network to recognize lexical sign glosses, mouth shapes and hand shapes, in a weakly supervised fashion, with the three HMMs having to synchronize at the end of each sign, as shown in Figure 3.9. The originality of their work also stems from the way weak labels were obtained. Hand shape labels are estimated by parsing SignWriting dictionaries, as for Deep Hand [Koller et al., 2016a], and mouth shape labels are derived by using a phonetic model on the German translations of the RWTH-PW corpus [Camgoz et al., 2018].

Table 3.11 summarizes most CLexSR results on RWTH-PW, SIGNUM and CSLR100. From this table, it appears that most experiments are conducted in a signer-dependent fashion. Signer independence appears to be quite a challenge, with a best result of 44.1% WER on RWTH-PW. This table confirms that RWTH-PW corresponds to the most difficult CLexSR task, whereas some models yield WERs lower than 5% on SIGNUM and CSLR100.

3.4 Going past Continuous Lexical Sign Recognition: a few perspectives

As exemplified by Table 3.11, competition in the field of SLR is highly focused on the task of CLexSR, especially on the RWTH-PW corpus. However, a few works have tried to explore the task of end-to-end Sign Language Translation (Section 3.4.1), while others have attempted to tackle grammar-oriented issues in CSL (Section 3.4.2).

3.4.1 A new trend towards Sign Language Translation?

Sign Language Translation (SLT) is obviously a long-standing objective in the field of Sign Language Processing (SLP) while being strongly related to the field of CSLR. A few attempts have been proposed, all on the 2018 release of the RWTH-PW corpus [Camgoz et al., 2018].

3.4.1.1 `Text` → `Sign` [Stoll et al., 2018]

A first Sign Language Generation (SLG) model is proposed by Stoll et al. [2018] addressing the task of `Text` → `Sign`, where *Text* and *Sign* respectively stand for English sentences and SL sequences. More

Paper	RWTH-Phœnix-Weather		SIGNUM		Continuous SLR100[†]	
	SD	SI	SD	SI	SD	SI[⋆]
[Von Agris et al., 2008]	-	-	12.7	**34.9**	-	-
[Koller et al., 2015]	53.0	-	10.0	-	-	-
[Koller et al., 2016a]	45.1	-	7.6	-	-	-
[Koller et al., 2016b]	38.8	-	7.4	-	-	-
[Camgoz et al., 2017]	40.7	-	-	-	-	-
[Cui et al., 2017]	38.7	-	-	-	-	-
[Koller et al., 2017]	26.8	**44.1**	**4.8**	-	-	-
[Koller et al., 2018]	32.5	-	7.4	-	-	-
[Huang et al., 2018]	38.3	-	-	-	-	17.3
[Guo et al., 2018]	-	-	-	-	63.0	10.2
[Pu et al., 2019]	36.7	-	-	-	**32.7**	-
[Guo et al., 2019a]	38.7	-	-	-	61.9	-
[Guo et al., 2019b]	36.5	-	-	-	44.7	14.3
[Zhou et al., 2019]	34.5	-	-	-	-	4.5
[Yang et al., 2019]	34.9	-	-	-	-	**3.8**
[Koller et al., 2019]	26.0	-	-	-	-	-
[Camgoz et al., 2020]	**24.5**	-	-	-	-	-

Table 3.11: Reported Word Error Rate (WER) (%) of methods detailed in Section 3.3.3 applied to Continuous Lexical Sign Recognition on the corpora presented in Section 3.3.2. SD and SI stand for Signer-Dependent and Signer-Independent.

[⋆]The exact same annotated *sentences* are present in training and test sets.

[†]It is unclear whether the training/test splits of the different papers are comparable.

precisely, the authors break the problem down into three sub-problems, with the following pipeline: `Text` → `Gloss` → `Skeleton` → `Sign`.

`Text` → `Gloss` is realized by a Neural Machine Translation (NMT) encoder-decoder architecture;

`Gloss` → `Skeleton` is simply the result of averaging the OP skeleton data for each gloss instance of the training set;

`Skeleton` → `Sign` is dealt with by a VAE-GAN (Variational Auto-Encoder - Generative Adversarial Network).

One of the major drawbacks of this architecture is that it assumes SL can be represented by sequences of lexical sign glosses.

3.4.1.2 `Sign` → `Text` [Camgoz et al., 2018]

Symmetrically and at the same time, another model is proposed by Camgoz et al. [2018], aiming at SLT, also using the NMT architecture. Several models are actually trained:

`Sign` → `Text` with no intermediate supervision;

`Sign` → `Gloss` & `Gloss` → `Text` two models trained independently[4] and used to form a complete `Sign` → `Gloss` - `Gloss` → `Text` model (with no additional training);

`Sign` → `Gloss` → `Text` a unified model with intermediate gloss supervision, trained in an end-to-end fashion.

This model, which is closer to the object of this thesis than that of Stoll et al. [2018], actually exhibits the same limitations. It explicitly assumes that the *gloss sequence* representation is a satisfactory way of analyzing CSL discourse. This can be read in [Camgoz et al., 2018, p. 1-3]:

> *This translation task is illustrated in Figure 1, where the sign language glosses give the meaning and the order of signs in the video, but the spoken language equivalent (which is what is actually desired) has both a different length and ordering.*

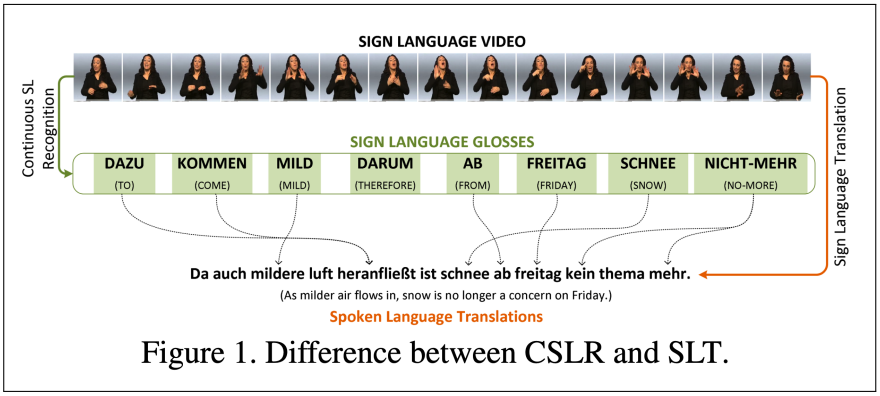


Figure 1. Difference between CSLR and SLT.

> *[...] Translating sign videos to spoken language is a seq2seq learning problem by nature.*

Yet, as we detailed in Section 2.2.1, the *gloss sequence* representation misses the multilinearity of conveyed information, the use of space and the common use of non-conventional depicting signs, that is to say the three major characteristics that are usually highlighted by linguists. Thus, we can only disagree with the following quote, where the authors put forward the idea that the `Gloss` → `Text`

[4] The `Sign` → `Gloss` part of their architecture uses that of [Koller et al., 2017].

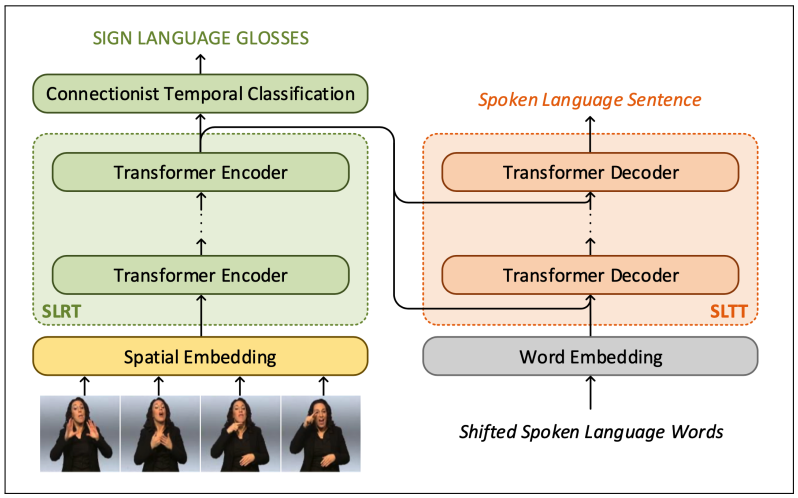


Figure 3.10: `Sign` → `(Gloss+Text)` architecture of Camgoz et al. [2020], with two simultaneous training objectives: CLexSR and SLT.

model, when fed with ground truth gloss labels, would yield an upper limit for the performance of SLT. From Camgoz et al. [2018, p. 5-6]:

> *We categorize our experiments under three groups:*
>
> 1. *Gloss2Text (G2T), in which we simulate having a perfect SLR system as an intermediate tokenization.*
> 2. *Sign2Text (S2T) which covers the end-to-end pipeline translating directly from frame level sign language video into spoken language.*
> 3. *Sign2Gloss2Text (S2G2T) which uses a SLR system as tokenization layer to add intermediate supervision.*
>
> *[...] G2T experiments [...] create an upper bound for end-to-end SLT.*

3.4.1.3 `Sign` → `Text` [Camgoz et al., 2020]

In a more recent publication, Camgoz et al. [2020] have somehow revised their position with respect to the assumed superiority of the gloss representation for SLT:

> *glosses [...] represent an information bottleneck for any translation system. This means that under ideal conditions, a Sign2Text system could and should outperform Gloss2Text.*

In this newer paper, Camgoz et al. experiment a state-of-the-art NMT architecture, namely transformer networks [Vaswani et al., 2017], instead of the legacy attention based encoder-decoder approach of [Camgoz et al., 2018]. Interestingly, in addition to the model already presented in Section 3.4.1.2, Camgoz et al. introduce a different model, called `Sign` → `(Gloss+Text)`, illustrated on Figure 3.10, in which the encoder-decoder architecture simultaneously aims at recognizing the sequence of glosses and at outputting the translated text sequence, without using glosses as an intermediate representation.

The effectiveness of the transformers architecture can be seen in Table 3.12, with improved performance for all model types – the chosen performance metric is BLEU score [Papineni et al., 2002], which is the most common metric for machine translation, corresponding to a form of modified precision using n-grams. Furthermore, the simple `Sign` → `Text` model of [Camgoz et al., 2020] – that is, with no

Release	Model type	BLEU-1	BLEU-2	BLEU-3	BLEU-4
[2018]	`Gloss` → `Text`	44.1	31.5	23.9	19.3
	`Sign` → `Gloss` → `Text`	43.3	30.4	22.8	18.1
	`Sign` → `Gloss - Gloss` → `Text`	41.5	29.5	22.2	17.8
	`Sign` → `Text`	32.2	19.0	12.8	9.6
[2020]	`Gloss` → `Text`	48.9	36.9	29.5	24.5
	`Sign` → `Gloss` → `Text`	48.5	35.4	27.6	22.5
	`Sign` → `Gloss - Gloss` → `Text`	47.7	34.4	26.6	21.6
	`Sign` → `(Gloss+Text)`	46.6	33.7	26.2	21.3
	`Sign` → `Text`	45.3	32.3	24.8	20.2

Table 3.12: BLEU scores (%, unigrams (BLEU-1) to 4-grams (BLEU-4) – higher is better) for the different Sign Language Translation models of Camgoz et al. [2018, 2020], applied to the test set of the RWTH-Phœnix-Weather corpus. `Gloss` → `Text` results are also computed, based on ground truth gloss annotations. BLEU score [Papineni et al., 2002] is the most common metric for machine translation, corresponding to a form of modified precision using n-grams.

gloss supervision – performs better than the `Gloss` → `Text` model of [Camgoz et al., 2018]. However, within the transformers architecture, the `Gloss` → `Text` model remains the most effective. This shows that either the interpreted SL of the RWTH-PW corpus makes little use of the structures typical of SL (refer to the discussion of Section 3.3.2.2), or the translation performance is actually rather poor.

3.4.2 Realistic expectations involving linguistics

As emphasized in the previous section, acceptable SLT performance – which can be seen as a long-term goal of SLR – is not nearly achieved. Other SLR works have opted for more realistic goals and yet tackling some complex linguistics processes of SL.

Very early on, Braffort [1996] proposed that the recognition of spatial information should be integrated into SLR systems, in order to go towards Sign Language Understanding (SLU). In this work, that relies on the use of data gloves, a SLR system integrates the recognition of standard lexical signs, proforms and directional verbs (see Section 2.1). This HMM-based model was tested on a small self-made corpus, with encouraging results, although scaling up to bigger corpora with coarser annotation schemes is not straightforward.

More recently, the NCSLGR corpus was released by Neidle and Vogler [2012]. With more details given in Section 4.1, this ASL corpus is made of short elicited utterances and longer spontaneous narratives, with grammar-related annotations. Using a stochastic face tracker, Metaxas et al. [2012] trained a HMM-SVM model to recognize five nonmanual markers – in this case, face expressions – on a subset of the NCSLGR corpus[5]. These markers are relevant at the syntactic level, namely: *Negation*, *Wh-questions*, *Yes/no questions*, *Topic or focus* and *Conditional or 'when' clauses*.

[5] It is not clear if the chosen utterances in the training and test sets belong to the more artificial part of the corpus, to the more spontaneous or both.

Setting	Features		
	Manual	Facial	Combined
Signer-Dependent (SD)	19.2%	87.7%	**12.7**%
Signer-Independent (SI)	39.4%	94.6%	**34.9**%

Table 3.13: Detailed Continuous Lexical Sign Recognition results (WER) of [Von Agris et al., 2008] on the SIGNUM Database [Von Agris and Kraiss, 2007]. Signer-dependent and -independent settings are evaluated, with manual features, facial features or both as input.

Still on the NCSLGR corpus[5], Yanovich et al. [2016] trained and tested a *sign type* classifier. Their model, based on optical flow and a CRF architecture, classifies any frame into one of three main sign types: *Lexical sign*, *Fingerspelled sign* and *Classifier sign* (see Section 2.1). The advertised accuracy is high (91.3% at the frame level), but it is computed on frames that belong to the three categories only.

We can also mention the punctual but effective inclusion of so-called Non-Manual Features (NMFs). On the SIGNUM Database, Von Agris et al. [2008] used the model described in Section 3.2.3 and analyzed the importance of NMFs for CLexSR. Their results in terms of WER are given in Table 3.13. On a much smaller dataset – 98 *sentences* and 24 lexical signs – Yang and Lee [2013] built a hybrid model, with a hierarchical CRF for segmentation of the manual activity, a BoostMap embedding for hand shape analysis and a SVM for the classification of the facial expression into five different categories. The final WER is 15.9%. However, as pointed out by Cooper et al. [2011], NMFs should not be restricted to facial expression.

At this stage, the state of the art we have outlined in this chapter enables us to derive some general conclusions as well as lay out the objectives and position of this thesis.

Problem statement

In Chapter 3, we have sought to provide a current state of the art in the field of Sign Language Recognition (SLR). Because Isolated Lexical Sign Recognition (ILSR) is formally equivalent to gesture recognition, we have chosen to insist on Continuous Sign Language Recognition (CSLR), which is a research domain that involves language-related questions.

It appears that the current acceptation of CSLR is what we refer to as Continuous Lexical Sign Recognition (CLexSR), that is the recognition of lexical sign glosses within continuous signing (Section 3.3). On the basis of the arguments we have developed in Chapter 2, it appears that this direction is strongly biased, and will not make it possible to go towards Sign Language Understanding (SLU) and *a fortiori* to Sign Language Translation (SLT). Indeed, the *gloss sequence* description misses the three main SL characteristics that we have detailed in Section 2.2: the multilinearity, that makes it possible to convey several types of information at once; the prevalent use of space, that structures SL discourse; the iconicity, that enables to show while saying.

Our point is hardly new, and has been put forward early on, by a few researchers in the field of SLR. For instance, Braffort [1996] insisted on the importance of space as a grammar component of French Sign Language (LSF):

> *Since the order of the signs is much less significant in LSF than their relative spatial arrangement, statistical grammars (based on succession frequencies of symbols) are not sufficient to deal with sentences in which spatial information is fully utilized* (translated from French, p. 164).

In another work, Edwards [1997] mentioned complementary arguments. He observed that the focus had been on conventional signs – *gestures* –, leaving out the grammar of SL. In terms of grammar, Edwards primarily referred to the iconic characteristic of SLs, for which he used the expression *property of richly grounding*. Also, he insisted on considering multilinear aspects of SL, like facial expression or body posture:

> *To date the research emphasis has been on the capture and classification of the gestures of sign language [...]. However, it is suggested that there are some greater, broader research questions to be addressed before full sign language recognition is achieved. The main areas to be addressed are sign language representation (grammars) and facial expression recognition [...], though there are others [...], such as body posture.*
> *[...]*
> *Most researchers are either unaware of or have chosen to ignore [these areas].*

Related to the fact that CLexSR has been the main concern of researchers, leaving out the three linguistic characteristics we have just mentioned, specific types of SL corpora have become popular. Many of them consist of artificial elicited sentences, repeated several times, with eliciting material – and annotation schemes – consisting of sequences of glosses, precisely. This is for instance the case of the SIGNUM Database (Section 3.3.2.1) and the Continuous SLR100 (CSLR100) corpus (Section 3.3.2.3).

Of a somewhat different type, RWTH-Phœnix-Weather (RWTH-PW) is another very popular corpus, probably the most used to this day for performance assessment and comparison of CLexSR systems. Although the annotation scheme is similar to that of SIGNUM and CSLR100, the discourse type is necessarily more spontaneous, as it consists of interpreted SL. However, we have pointed out in Sections 2.2.3 and 3.3.2.2 the fact that interpreted SL is inevitably influenced by the original speech, especially in terms of syntax, which means that the use of structures typical of SL is definitely restricted – an observation that is shared by the authors of this corpus. The generalizability of this corpus is thus limited, which is not always acknowledged by SLR researchers.

Another limitation of using glosses as training objective of SLR systems that should be noted, is the fact that glosses do not necessarily represent the meaning of signs they are associated to. This is highlighted by Johnston and De Beuzeville [2016]:

> *ID-glosses are likely to confuse a general audience because they might not closely reflect (literally "gloss") the meaning of the sign. That is not their purpose or function. A gloss which is the best translation equivalent for a given context is much more appropriate in other cases. One of the keywords associated with an ID-gloss is probably going to be the most suitable word to use in these cases. [...] Used alone like this, glosses almost invariably distort face-to-face SL data. Their use may well be counter-productive.*

We appreciate that a few leads have been initiated towards different directions than CLexSR. SLT is one of them, although it has been mostly driven by gloss supervision, on the limitative RWTH-PW corpus (Section 3.4.1).

On the other hand, focusing on SLT or on the recognition of lexical signs only, with *black box* architectures, may prevent developments in the linguistic description and automatic analysis of SL. A few approaches have actually chosen to deal with linguistic matters, yet on very small corpora or only superficially (Section 3.4.2).

In light of all these observations, it seems necessary to find a new way for automatic SLR that would take into account the specificities of SL as a visual-gestural language. In the next chapters, we will propose to drive research in SLR in this direction, with a relevant corpus and a methodology that do not rely on lexical signs only.

Part II

A more general paradigm for SLR

Chapter 4

Towards better corpora for SLR

Sign Language Recognition (SLR) systems are highly dependent on and driven by SL corpora. In this chapter, we discuss possible improvements with respect to the corpora largely used by the SLR community.

A first group of corpora, developed by linguists, is presented in Section 4.1. They are carried out with particular care to the linguistic quality and generalizability of the studied SL. However, the annotations are not always consistent, with many videos only partially annotated. Then, a remake of the French Sign Language (LSF) part of the Dicta-Sign corpus, that is intended to be both linguistically relevant and interesting for the SLR community, is detailed in Section 4.2.

4.1 Linguistic-driven corpora

In this section, we introduce six Continuous Sign Language (CSL) corpora realized by linguists (see Figure 4.1). Their respective annotation guidelines are usually quite detailed, with lexical signs, depicting signs and more. Except for NCSLGR, these corpora are very large in terms of duration, they include many signers and are made of dialogues, narratives and conversations. Undoubtedly, they can be considered very representative of natural SL. However, because these corpora have been made by linguists and intended for linguistic analyzes, using them for SLR tasks is not straightforward. The main reason for this is the lack of consistency in the annotations across the corpora: most of them are still ongoing work, with annotations being updated continuously.

An overview of these corpora, along with those presented in Section 3.3.2, is given in Table 4.1. In this table, we include the number of signers, total duration, discourse type, whether a written translation is included in the annotation as well as the annotation categories (besides lexical sign glosses).

Auslan Corpus

The Auslan Corpus [Johnston, 2009], belonging to the Endangered Languages Archive, consists of 300 hours of video recording – 150 hours of usable language production – from 100 signers of Australian Sign Language (Auslan). The elicitation procedure was extensive, with an interview, the production of narratives, responses to questions, free conversation among others.

The annotations are rich and include lexical sign glosses, detailed grammatical class (Depicting Signs (DSs), Pointing Signs (PTSs), *etc.*), gaze, constructed action (refer to Section 2.1 for more detail).

Corpus	SL	Source	Signers	Hrs.	Discourse type	Translation	Annotation outside lexicon		Used for SLR or linguistic studies
							Categories	Consistent	
CSLR100	ChSL	RGBD skeleton	50	100	Artificial	-	-		SLR
RWTH-PW	DGS	RGB	9	11	Interpreted	German	-		SLR
SIGNUM	DGS	RGB	25	5	Artificial	Ger./Eng.	-		SLR
NCSLGR	ASL	BW/RGB	7	2	Mixed	-	PTSs, DSs, FSs	Yes	Both (mainly linguistics)
Auslan Corpus	Auslan	RGB	100	150	Natural	-	PTSs, DSs, Constructed action	No	Linguistics
BSLCP	BSL	RGB	249	180	Natural	English	PTSs, DSs, FBuoys	No	Linguistics
DGS Korpus	DGS	RGB skeleton	330	50-300	Natural	Ger./Eng.	Mouthing	No	Linguistics
LSFB Corpus	LSFB	RGB	100	150	Natural	French	DSs	No	Linguistics
Corpus NGT	NGT	RGB	92	72	Natural	Dutch	DSs, Mouthing	No	Linguistics
Dicta-Sign–LSF–v2	LSF	RGB	16	11	Natural	French	PTSs, DSs, FSs, FBuoys, NSs, Gs	Yes	Both (mainly linguistics)

Table 4.1: Continuous Sign Language datasets. The top corpora have been developed and used by the SLR community (Section 3.3.2), but they are either artificial or not representative of natural Sign Language. Other have been built by linguists, with natural discourse and detailed annotation, yet not always consistent. To the best of our knowledge, Dicta-Sign–LSF–v2 and NCSLGR are the only two corpora built by linguists that have been used for *beyond gloss-level* CSLR experiments.

Auslan Corpus

BSLCP

DGS Korpus

NCSLGR

LSFB Corpus

Corpus NGT

Figure 4.1: Random frames from the six Continuous Sign Language corpora presented in Section 4.1

Only a fraction of the videos have been annotated at this time.

BSL Corpus

The BSL Corpus (BSLCP) [Schembri, 2008] contains video data recordings of dialogue in British Sign Language (BSL) from 249 deaf signers, for a total duration of 180 hours. Only a part of the corpus is publicly available.

Annotations are progressively released, with lexical sign glosses and English translations, as well as – for a fraction of released annotations – labels for PTSs, DSs and different types of buoys.

DGS Korpus

The DGS Korpus [Prillwitz et al., 2008] is an ongoing long-term project aiming at collecting 500 hours of German Sign Language (DGS) narratives and conversations, from more than 300 signers. A fraction of this corpus – about 50 hours of video – is released as the Public DGS Korpus [Jahn et al., 2018]. The public corpus also contains 2D skeleton data, computed by OpenPose (OP) [Cao et al., 2017]. A rich DGS dictionary with local variants should be released by 2023. As of now, it seems that the data collection is completed, although annotation tasks are still ongoing.

The annotation scheme covers lexical sign glosses, translation and mouthing.

LSFB Corpus

The LSFB Corpus [Meurant et al., 2016] is a dialogue corpus, with 100 signers of French Belgian Sign Language (LSFB) and 150 hours of recording. The type of discourse is natural and spontaneous, with general discussions and narratives.

As of now, ten hours are annotated, with lexical sign glosses and a general DS label[1]. Two hours of video are also translated in French.

NCSLGR

The NCSLGR corpus [Neidle and Vogler, 2012] includes two categories of discourse, in American Sign Language (ASL), for a total of two hours from seven signers. Most videos are made of artificial elicited utterances, in a similar way to Signum (Section 3.3.2.1). However, the corpus also includes spontaneous short stories, with more language variability.

Annotations for the manual activity follow the conventions from Baker and Cokely [1980]; Smith et al. [1988], with: lexical sign glosses, Fingerspelled Signs (FSs), *hold signs* (hand position held at the end of a sign, not necessarily with a linguistic function), PTSs, DSs (seven categories) with proforms and Gestures (Gs). Non-manual activity is also annotated, with head movement and eye gaze among others.

Corpus NGT

The Corpus NGT [Crasborn and Zwitserlood, 2008; Crasborn et al., 2008] is an online open archive corpus including 72 hours of – partially – annotated data from 92 native signers of Dutch Sign Language (NGT). This corpus is mainly a dialogue and conversation corpus, supplemented by a few elicited and spontaneous narratives.

[1] The authors envision to further specify the different DS categories in the future.

Annotations mostly cover lexical signs (glosses) and Dutch translations, but sometimes include information on mouthing or classifier constructions.

4.2 The example of Dicta-Sign–LSF–v2

This section focuses on Dicta-Sign–LSF–v2, a remake of the LSF part of the Dicta-Sign corpus. This new corpus is especially intended for assessing the automatic recognition of elements within the iconic and spatial modalities.

In this section, we discuss the relationship between Dicta-Sign–LSF–v2 and Dicta-Sign (Section 4.2.1), the recording conditions (Section 4.2.2), elicitation material (Section 4.2.3) and annotation categories as well as guidelines (Section 4.2.4), then we present some statistics (Section 4.2.5).

4.2.1 From the Dicta-Sign project to Dicta-Sign–LSF–v2

With an initial objective of improving web-based human-computer interfaces for the Deaf, the European project Dicta-Sign [Efthimiou et al., 2010] focused on four different SLs: British (BSL), Greek (GSL), German (DGS) and French (LSF). Targeting both SLR and Sign Language Generation (SLG), this project also emerged from the shared interest of researchers in the cross-comparison of different SLs. In terms of produced data, we can mention :

- a shared lexicon with 1000 concepts and a citation-form realization for each SL, transcribed with the Hamburg Notation System for Sign Languages [Hanke, 2004] (HamNoSys);
- a comparable CSL dialogue corpus for the four SLs, with, for each SL, at least 14 signers and 10 dialogue tasks all related to the concept of *travel*.

Recently, we published Dicta-Sign–LSF–v2 [Belissen et al., 2020a], a remake of the French part for the CSL corpus from Dicta-Sign. Video data was standardized, annotations were cleaned, synchronized and made reliable, and the redesigned corpus was published on the language platform Ortolang[2][3]. We also published fully preprocessed signer data, following the proposed signer representation features developed in Section 6.1.

The rest of this chapter is focused on Dicta-Sign–LSF–v2, that is the redesigned LSF part of Dicta-Sign.

4.2.2 Recording setup

For each of the eight couples in the corpus, the two signers faced each other. Two recording cameras were placed just above their head, so that they were facing the person being recorded almost perfectly, while a third camera recorded the scene from the side. The setup can be seen on Figure 4.2.

These three views were released in Dicta-Sign–LSF–v2, with identical resolution (720 × 575 at 25 fps). Other views, like a top-down view, a frontal RGB-Depth (RGBD) camera recording, or better resolution may have existed in the original corpus. They are not part of this release for various reasons, including the absence of consistency across all videos, and our intent to stimulate research and applications on common front-view RGB recordings.

[2] https://www.ortolang.fr/market/corpora/dicta-sign-lsf-v2 [LIMSI and IRIT, 2019]

[3] Ortolang is a platform for language, which aims at constructing an online infrastructure for storing and sharing language data (corpora, lexicons, dictionaries etc.) and associated tools for its processing

Figure 4.2: Recording setup in Dicta-Sign–LSF–v2, with two frontal cameras and a side one.

4.2.3 Elicitation material and type of discourse

Since the original multilingual corpus Dicta-Sign was made to be comparable, nine common *tasks* were shared, all related to the concept of *travel in Europe*:

Task 1 Public Transportation

Task 2 Travel Agency

Task 3 Planning a Holiday

Task 4 Airport

Task 5 City Map

Task 6 Expectation & Reality

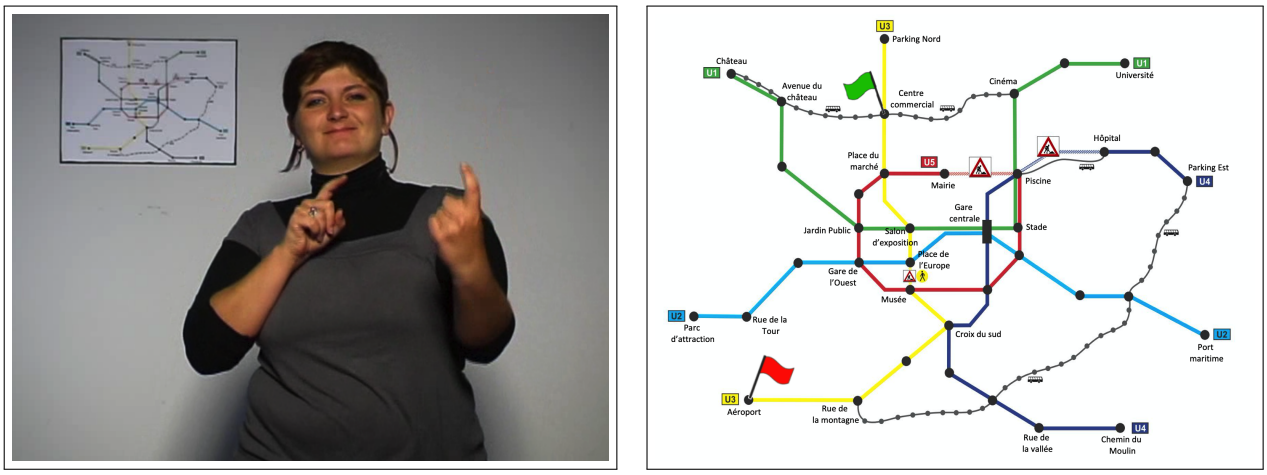
Task 7 Travel then & now

Task 8 Signed Story/Picture Story

Task 9 Dates, Isolated Signs

Elicitation methods consisted in having the participants watch videos and slides describing the purpose of the task. For instance in task 1, the elicitation video and the slides ask a signer to explain to the other signer of the couple how to get from A to B, using public transportation. See Figure 4.3 for an illustration.

Because all tasks were related to the theme of travel, it is obvious that the lexical field is somehow biased in the corpus. However – except for task 9 –, as elicitation guidelines were sufficiently loose, the type of discourse was very natural and spontaneous, exhibiting specific structures of SL and a significant variability between signers (see Section 4.2.5).



(a) Elicitation video

(b) Elicitation slide

Figure 4.3: Elicitation video (a) and slide (b) for the task 1 of the Dicta-Sign corpus. Signers are asked to explain how to go from point A (green flag) to point B (red flag) using public transportation.

4.2.4 Annotations

In order to make the annotation as consistent and comprehensive as possible, an annotation guide has been written and made public. In this section, we present the main guidelines on the tracks used, the segmentation rules for manual units and the categories used.

Beforehand, we remind that although the chosen guidelines correspond to commonly used decision rules amongst linguists, there is no such thing as annotation standards on this subject, especially on the number of tracks or the method for segmenting units. For a detailed discussion, see [Crasborn, 2010].

4.2.4.1 Tracks

Manual units

Three tracks are used for manual units. The goal is to annotate the units carried by each of the left and right hands independently and those for which both hands carry the unit in such a way that they are considered inextricable:

LH for manual units made with the left hand. This also includes lexical signs usually performed only with the left hand in a given context.

RH for manual units made with the right hand. This also includes lexical signs usually performed only with the right hand in a given context.

2H for two-handed manual units, i.e. units that have a global meaning and for which the initial and final postures of each hand are approximately aligned with each other. This also includes lexical signs that are usually one-handed and are performed with both hands in a given context.

Translation

A French translation was realized orally by an interpreter, who then transcribed and aligned the text to the video using annotation software. The style is therefore close to oral French, as in the case of real-time interpretation.

4.2.4.2 Segmentation

Start and end of a manual unit
A manual unit starts when a key posture is identified, that is when the hand parameters that constitute the beginning of the unit are met: configuration, orientation, location. Note that the parameters are defined *a priori* for lexical signs but not for other manual units. As in Dicta-Sign–LSF–v2 the frame rate is 25 fps, this often corresponds to a sharp image. If this posture seems to fall between two frames, the unit starts at the next frame.

A unit ends when at least one of the hand parameters that make up the end of the unit is no longer in place: configuration, orientation, location.

Case of repeated signs
When a manual unit consists in the repetition of subparts, it is considered a single unit only when transitions are very short and hand parameters are unchanged. Repeated manual units are thus divided into subunits when transitions are longer or when at least one hand parameter (usually the location) is changed.

Case of hold signs
Sometimes two-handed units end but one of the hands – usually the weak hand – maintains the final posture of the two-handed unit in a more or less committed fashion.

If an interaction with the other hand is observed while the hold is active, the hold is annotated with the Fragment Buoy (FBuoy) category. This annotation segment begins at the image following the end of the annotation segment for the two-hand unit and ends as described earlier. It should be noted that this type of unit often dies out slowly over time. As a result, the end of the unit is more frequently indicated by a change in orientation or configuration: a simple change of location is not always sufficient. If the hold only lasts a few frames, the posture dies out rapidly and there is no interaction between the two hands, the hold is not annotated.

4.2.4.3 Categories and values

The annotation categories are strongly influenced by the guidelines of [Johnston and De Beuzeville, 2016] that are detailed in Section 2.3.3. Once segmented, each manual unit was assigned one of the categories listed below. All annotations are binary, except for FLS which are annotated as a categorical variable.

Fully Lexical Signs (FLSs) for conventional units. The value is an identifier (*ID*), in the form of an integer, associated to an ID-gloss in French.

Partially Lexical Signs (PLSs) for non-conventional, context-dependent units, including:

Depicting Signs (DSs) for illustrative structures,

Pointing Signs (PTSs) or indexing signs,

Fragment Buoys (FBuoys) for the hold of a fragment or the final posture of a two-handed lexical sign, usually on the weak hand [Liddell, 2003]. List Buoys (LBuoys) and Theme Buoys (TBuoys) are not part of the annotation guidelines of Dicta-Sign–LSF–v2.

Non Lexical Signs (NLSs) for units that are neither lexicalized nor illustrative, including:

Numbering Signs (NSs) for numbers greater than 10 (no values),

Fingerspelled Signs (FSs) for proper names or when the sign is unknown,

Gestures (Gs) for non-lexicalized gestures, which may be culturally shared or idiosyncratic – these gestures are not assigned an ID-gloss.

These categories are considered mutually exclusive, although one should note that ambiguity is often present. This is the case for some very iconic units that can be categorized as lexical signs but also as illustrative structures. Cases that could have been annotated into more than one category should be further investigated in the future.

Because it is defined as an elimination category, category G also raises some difficulties. It may well be that an annotated unit G can be considered a lexical sign if its systematic use is found in the same form to convey the same meaning.

Also, the original corpus Dicta-Sign included sub-categories[4] for DSs that will be proposed in a future release:

DS-Location (DS-L)	for the location of an entity,
DS-Motion (DS-M)	for the motion of an entity,
DS-Size&Shape (DS-SS)	for the size and shape of an entity,
DS-Ground (DS-G)	for a spatial or temporal reference (ground),
DS-Action (DS-A)	for the handling of an entity – originally DS-H in [Johnston and De Beuzeville, 2016],
DS-Trajectory (DS-T)	for a trajectory shown in the signing space,
DS-Other (DS-X)	for any other deformation of a standard lexical sign.

4.2.4.4 Annotation procedure and limits

Without going into too much detail, the annotation procedure consisted in having two experts annotate in parallel each video recording. The two experts then monitored each other and discussed disagreements. In the case a disagreement could not be resolved this way, a supervisor was called in.

In spite of this solid supervision, annotation errors may remain, for instance omissions.

More generally, the linguistic theories for describing SLs are relatively recent and lack consensus – see for instance the developments on the role of iconicity in Section 2.1. The relevance of different types of annotation categories is still subject to debate amongst linguists, and more research is needed. Practically speaking, assigning a category to a manual unit is a complicated and sometimes ambiguous task.

4.2.5 Statistics

This section presents some statistics on Dicta-Sign–LSF–v2, in order to get a grasp on the variability between considered gestural units, signers and tasks of the corpus.

[4] These sub-categories are actually not mutually exclusive

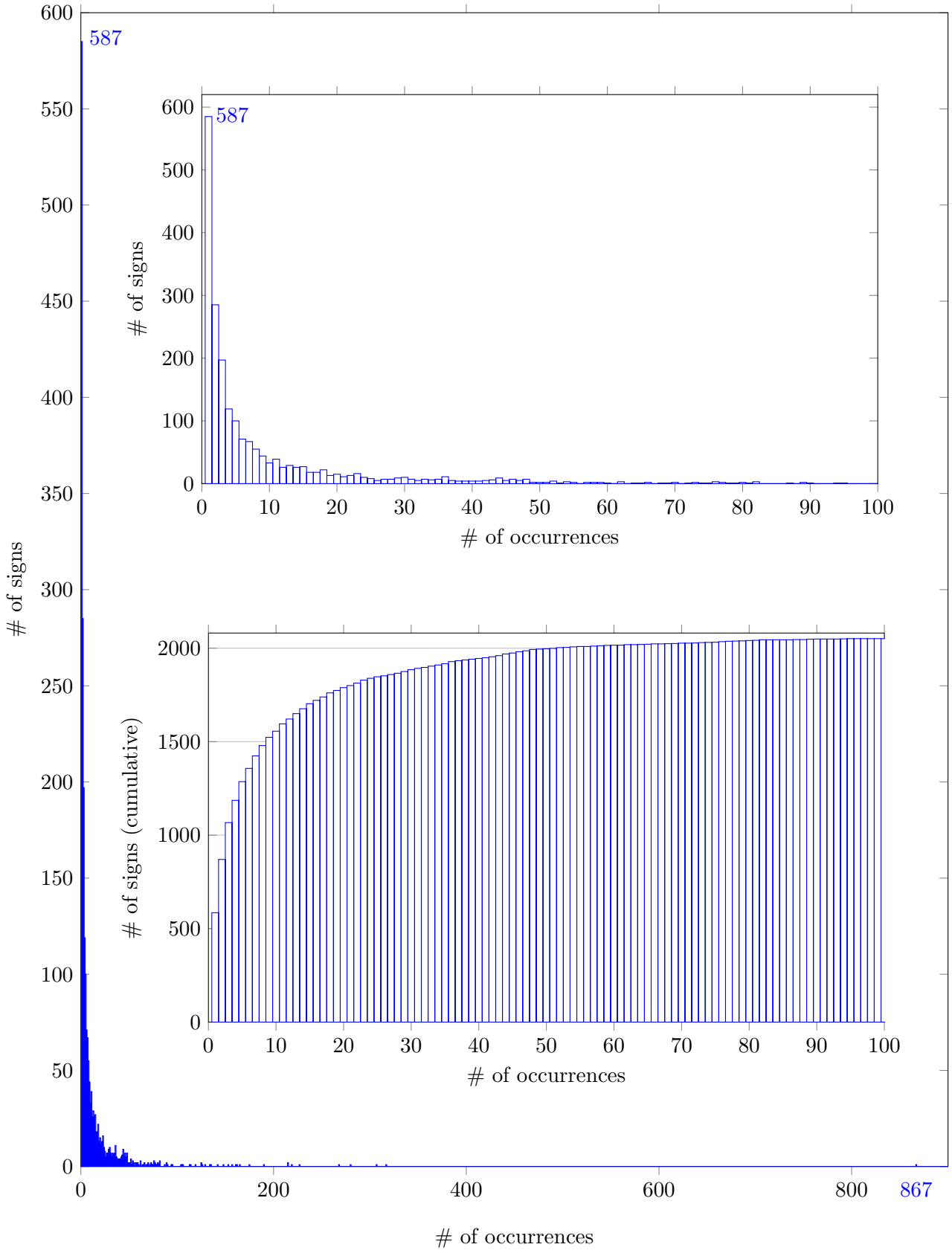


Figure 4.4: Distribution of the number occurrences of Fully Lexical Signs in the Dicta-Sign–LSF–v2 corpus. A detailed view is given for signs with less than 100 occurrences (top figure), along with the cumulative distribution (bottom figure). 587 signs only appear once in the corpus, whereas the sign `YES` has 867 occurrences. The total number of FLSs is 2081.

# of occurrences	# of signs with a smaller or equal # of occurrences	# of signs with a greater # of occurrences
0	0	2081
1	585	1496
10	1556	525
20	**1789**	**292**
50	1997	84
100	2051	30
200	2072	9
400	2080	1

Table 4.2: Numbers derived from the cumulative distribution of the number of occurrences for the Fully Lexical Signs of Dicta-Sign–LSF–v2. The way this table can be read is for instance: 1789 signs have less or exactly 20 occurrences, while 292 signs have more than 20 occurrences.

Dicta-Sign–LSF–v2 is made up of more than 11 hours of video recording (1007593 annotated frames), with 16 signers. Figure 4.4 presents the distribution of the number of occurrences for the 2081 FLS of the corpus. 587 signs only appear once in the corpus, whereas the sign `YES` has 867 occurrences. More detail is given in Table 4.2.

Table 4.3 and associated Figure 4.5 present the frame-wise and unit-wise statistics for the main annotation categories, that is FLS, DS, PTS, FBuoy, NS and FS. The category G is not included, because of the very low number of instances and the fact that it is defined by elimination (see Section 4.2.4.3).

In terms of annotated frames, the ratio FLS : PLS is about 2 : 1, and reaches 3 : 1 with respect to manual units. One will notice that PTSs are usually very short (252 ms per unit on average) whereas DSs and FBuoys are longer, with respectively 674 ms and 975 ms per unit on average. NLSs – which are relatively long units too – only account for 0.9% (resp. 1.8%) of the total annotated units (resp. frames). Results from Table 4.3 are actually quite consistent with [Sallandre et al., 2019], which includes a fine analysis of the distribution of units according to different discourse genres, including dialogue.

Figure 4.6 presents a finer analysis on the distribution of frame counts, for the main annotation categories. Figure 4.6a illustrates that signers have significantly different distributions in terms of FLSs, DSs, etc. For instance, the frame-wise distribution for Signer 4 is 81% for FLS, 11% for DSs and 1% for FBuoys, whereas for Signer 6 these categories respectively amount to 57%, 28% and 8%. In terms of tasks, the variation in distribution is also significant, with some tasks exhibiting particularly high or low values for certain categories. For instance, task 9 mainly consisted in signing dates, so the very high value in the NS category is unsurprising – as well as the very low value in DS. Also, task 8, consisting in a short signed story of a narrative type, has a quite high value for DSs at 32%.

Even though we have not released corresponding annotations yet, detailed statistics for sub-categories of DSs are also presented in Table 4.4 and associated Figure 4.7.

The original DS categories from the guidelines of Johnston and De Beuzeville [2016], that is DS-L, DS-M, DS-SS, DS-G and DS-A amount to 92.4% of the annotated DS frames.

	FLS	PLS			NLS		Total
		DS	PTS	FBuoy	NS	FS	
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	24565	3606	3651	589	155	118	32684
%	75.2%	11.0%	11.2%	1.7%	0.5%	0.4%	
Cumulative %	75.2%	86.2%	97.4%	99.1%	99.6%	100.0%	
Avg. frames/unit	8.4	16.8	6.3	24.4	24.7	16.4	
Avg. duration (ms)	335	674	252	975	988	658	

Table 4.3: Frame count and sign count (manual unit) statistics for the main annotation categories of Dicta-Sign–LSF–v2

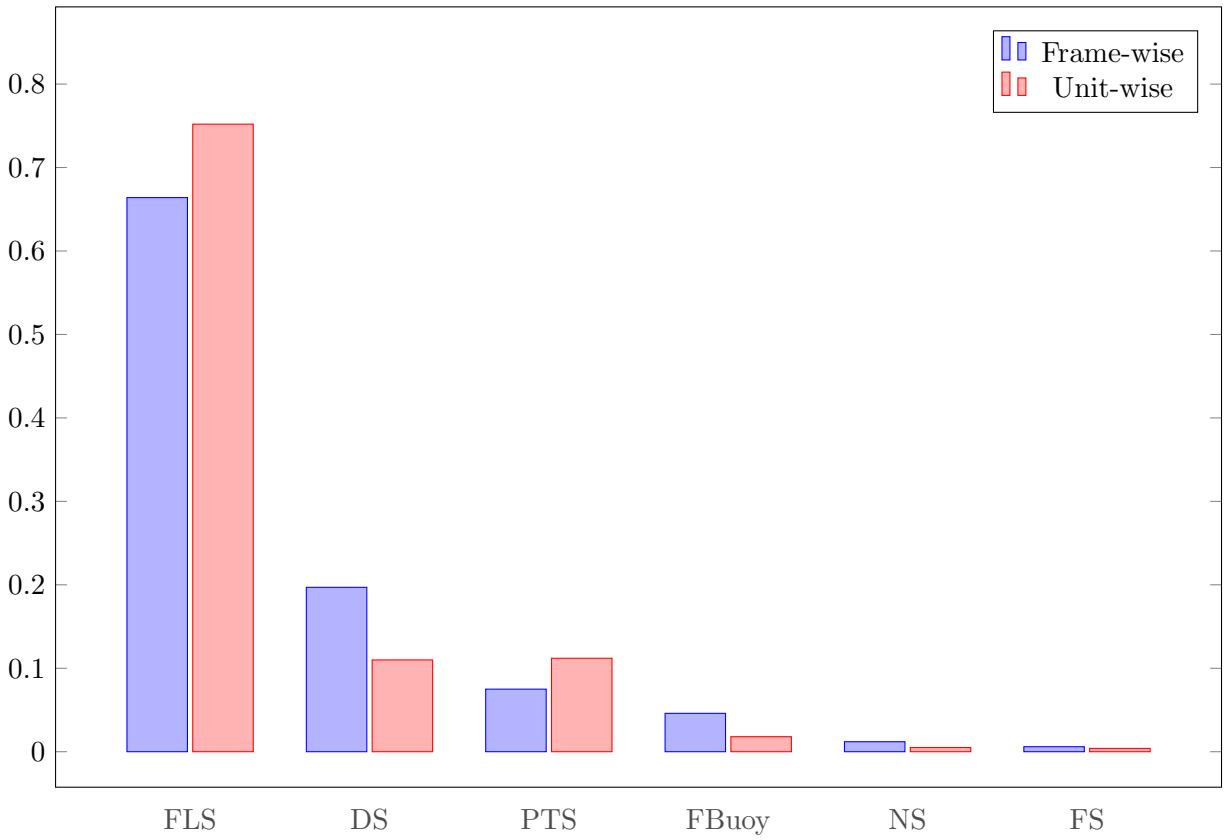
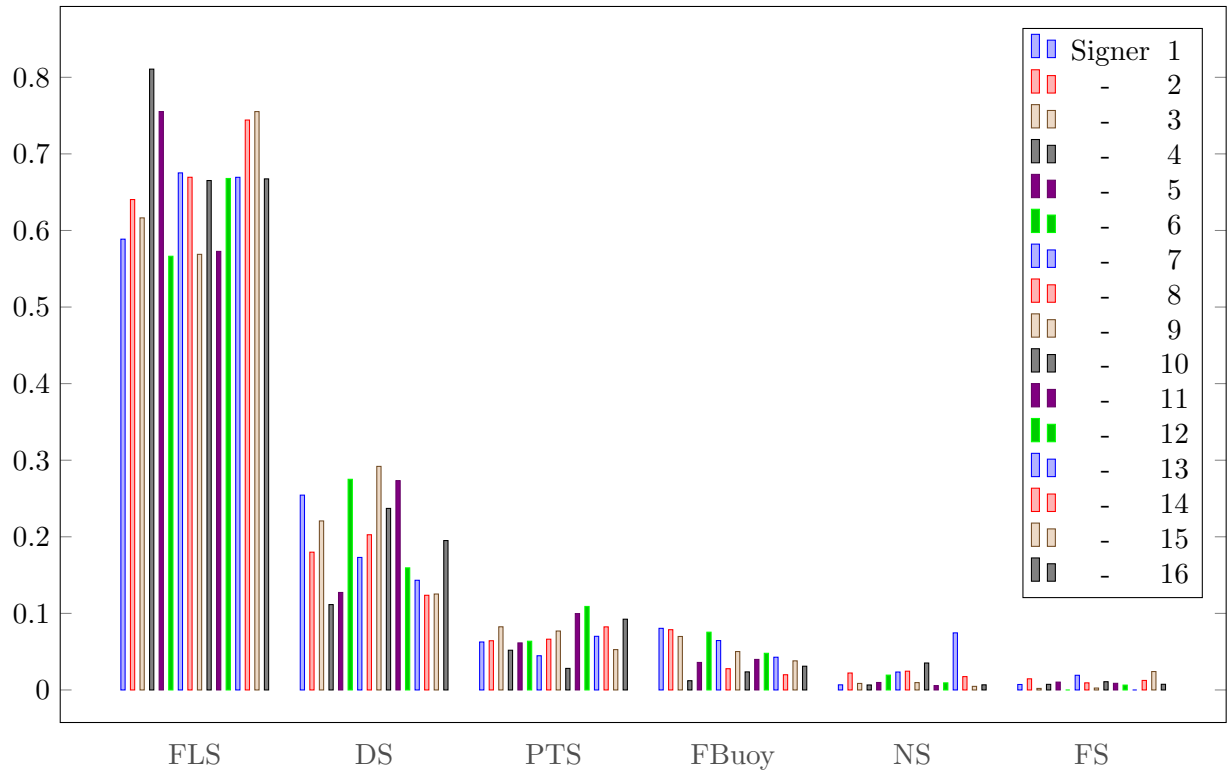
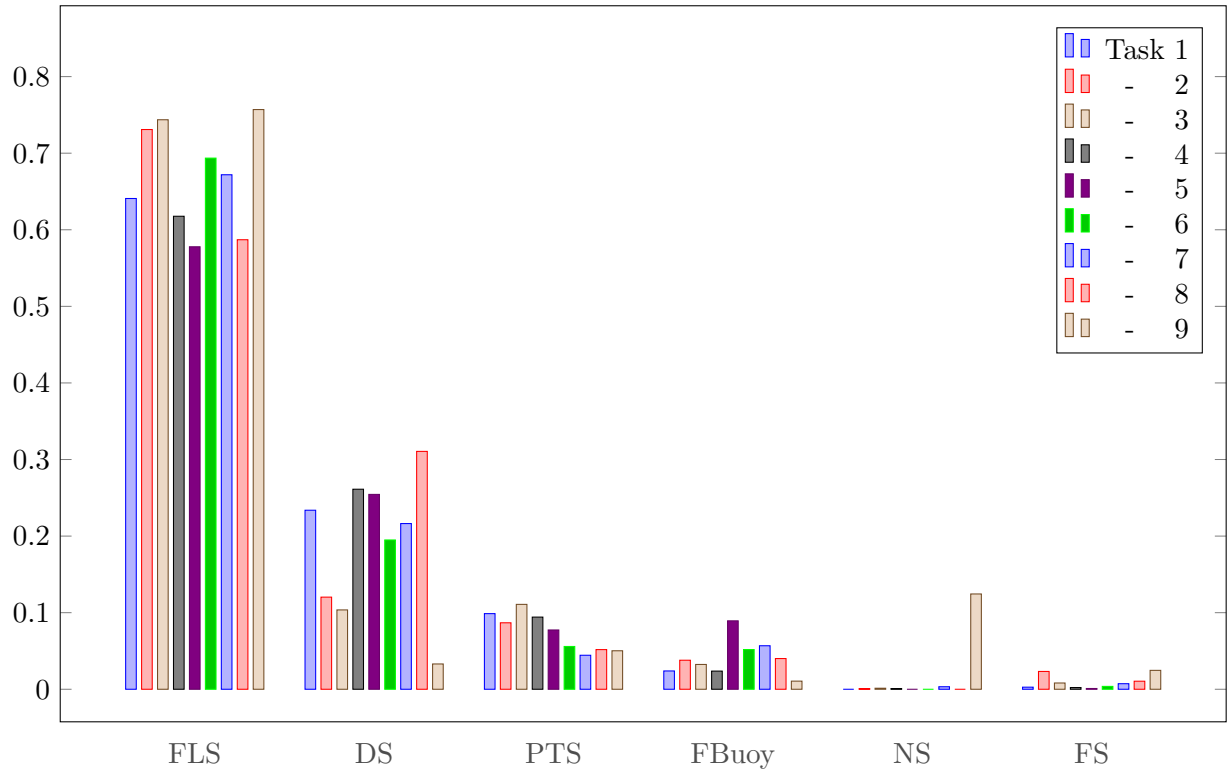


Figure 4.5: Frame count and sign count (manual unit) distribution for the main annotation categories of Dicta-Sign–LSF–v2



(a) Statistics per signer



(b) Statistics per task

Figure 4.6: Frame count statistics for the main annotation categories of Dicta-Sign–LSF–v2, for each signer (a) and for each task (b) of the corpus.

In terms of unit count, DS-L and DS-M account for the majority of manual units. Frame-wise, the unit length is quite uniform, except for DS-G that are about twice longer than other depicting signs. These units are indeed used as reference points with the weak hand, held for a long time while other lexical or depicting signs are produced on the dominant hand.

Conclusion

In light of the above description for the redesigned corpus Dicta-Sign–LSF–v2, its relevance for SLR is established.

Conversely to the common corpora that are used to train SLR systems (Section 3.3.2), Dicta-Sign–LSF–v2 is made of natural and spontaneous SL, in the form of dialogue. With varied tasks, it is fairly representative of the different linguistic phenomena one can observe in SL dialogue. Furthermore, it is annotated very finely in a consistent manner, with three hand tiers (LH, RH, 2H) and many categories following the guidelines of Johnston and De Beuzeville [2016].

Thus, this corpus should be a very good base for experimenting generalizable automatic SLR systems, also enabling to go towards Sign Language Understanding (SLU) and Sign Language Translation (SLT). The limitations we can mention include the modest size of the corpus (11 hours) and the fact that a better way to annotate the discourse in terms of the linguistically relevant use of signing space remains to be defined.

All of the data composing this corpus can be downloaded from the Ortolang website: video and OP preprocessed data, elicitation material, annotation guide and annotation files. This corpus constitutes the first contribution of this thesis.

In the next chapter, we will propose an appropriate framework for SLR experiments.

	DS-L	DS-M	DS-SS	DS-G	DS-A	DS-T	DS-X	Total
Non blank frames	22725	16378	7044	16286	9056	5081	626	77196
%	29.4%	21.2%	9.1%	21.1%	11.7%	6.6%	0.8%	
Manual units	1389	1170	535	581	590	310	41	4616
%	30.1%	25.3%	11.6%	12.6%	12.8%	6.7%	0.9%	
Avg. frames/unit	16.4	14.0	13.2	28.0	15.3	16.4	15.3	
Avg. duration (ms)	654	560	527	1121	614	656	611	

Table 4.4: Frame count and sign count (manual unit) statistics for the Depicting Sign categories of Dicta-Sign

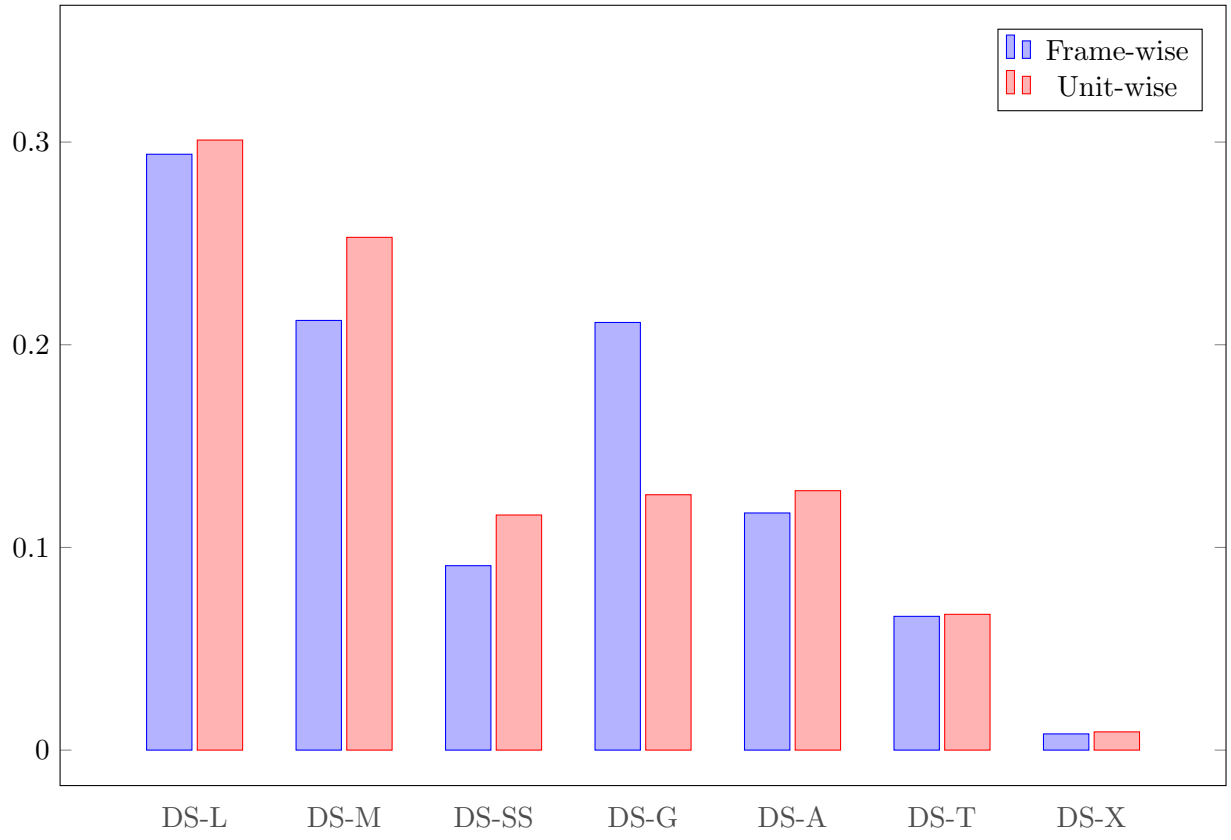


Figure 4.7: Frame count and sign count (manual unit) statistics for the Depicting Sign categories of Dicta-Sign

Chapter 5

A broader definition of Continuous Sign Language Recognition

Previously (Section 3.3), we have shown that the current acceptation of Continuous Sign Language Recognition (CSLR) should actually be referred to as Continuous Lexical Sign Recognition (CLexSR), since it is focused on the recognition of fully conventionalized signs (lexical signs) within Continuous Sign Language (CSL) videos. In this chapter, we propose a much more general definition of CSLR, as the simultaneous recognition of several linguistic *descriptors*.

In order to define CSLR in a very general way, we go back to the formalization of Sections 3.1 and 3.3.1. With $X = [f_1, \dots, f_T]$ as a CSL video made of T frames (f), the purpose of this chapter is to define a general form for Y_{CSLR}, the recognition objective, such that the process of CSLR consists in computing estimates $\widehat{Y}_{\text{CSLR}}$:

$$X \xrightarrow{\mathscr{R},\mathscr{M}} \widehat{Y}_{\text{CSLR}} \tag{5.1}$$

where $\mathscr{R}$ – an intermediate representation of X (*features*) – and $\mathscr{M}$ – a learning and prediction model – will be discussed in Chapter 6.

The quality of the estimates must then be assessed with a performance metric $\mathscr{P}(Y, \widehat{Y})$.

An introduction to the descriptors is the purpose of Section 5.1, then we discuss associated performance metrics in Section 5.2 and adapted error functions for training neural networks on this task in Section 5.3.

5.1 Linguistic descriptors

Generally speaking, we propose interpreting CSLR as the continuous recognition of different linguistic descriptors. By using the most linguistically relevant descriptors, one can hope to progress towards Sign Language Understanding (SLU) and Sign Language Translation (SLT).

Let us consider such a CSLR acceptation with M different linguistic descriptors $d^m, m \in \{1, \dots, M\}$,

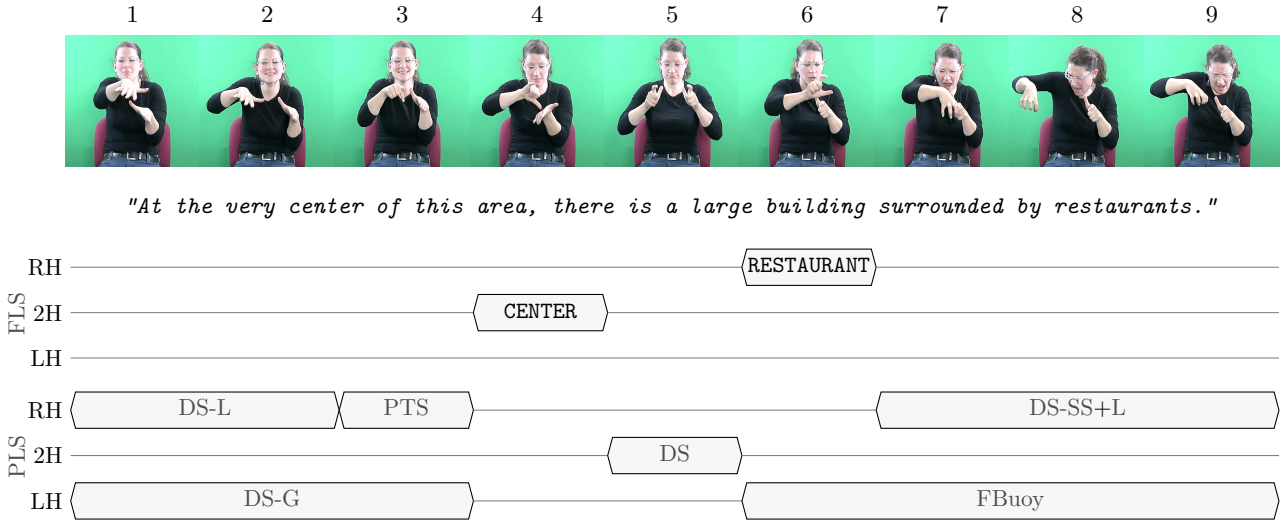


Figure 5.1: French Sign Language sequence from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`, see Chapter 4). Expert annotations for right-handed (RH), two-handed (2H) and left-handed (LH) Fully Lexical Signs (FLSs) and Partially Lexical Signs (PLSs) are given.

SLR type	Y	$\mathscr{P}$
Unaligned CLexSR (Section 3.3.1.1)	$\begin{bmatrix} g^1 & g^2 \end{bmatrix}$	WER
Aligned CLexSR (Section 3.3.1.2)	$\begin{bmatrix} g^0 & g^0 & g^0 & g^1 & g^0 & g^2 & g^0 & g^0 & g^0 \end{bmatrix}$	Acc
CSLR: $\begin{cases} d^1: \text{FLSs} \\ d^2: \text{DSs} \\ d^3: \text{PTSs} \\ d^4: \text{FBuoys} \end{cases}$	$\begin{bmatrix} g^0 & g^0 & g^0 & g^1 & g^0 & g^2 & g^0 & g^0 & g^0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$	$\begin{bmatrix} \mathscr{P}^1: \text{Acc} \\ \mathscr{P}^2: \text{F1} \\ \mathscr{P}^3: \text{F1} \\ \mathscr{P}^4: \text{F1} \end{bmatrix}$

Table 5.1: Illustration of unaligned and aligned Continuous Lexical Sign Recognition (CLexSR) on the sequence example from Figure 5.1, as well as a proposal for Continuous Sign Language Recognition (CSLR), including aligned Fully Lexical Signs (FLSs), and binary prediction for the presence or absence of Depicting Signs (DSs), Pointing Signs (PTSs) and Fragment Buoys (FBuoys).
Here, the lexicon is $\mathscr{G} = \{(g^0 : \texttt{NULL},)g^1 : \texttt{CENTER}, g^2 : \texttt{RESTAURANT}, \dots\}$.
Y and $\mathscr{P}$ respectively stand for the recognition objective and the performance metric.

so that Y_{CSLR} can be written as:

$$Y_{\text{CSLR}} = \begin{bmatrix} d^1 \\ \vdots \\ d^M \end{bmatrix} \tag{5.2}$$

with performance metric as a vector of size M, each descriptor having its own metric:

$$\mathscr{P}(Y, \widehat{Y}) = \begin{bmatrix} \mathscr{P}^1 \\ \vdots \\ \mathscr{P}^M \end{bmatrix}. \tag{5.3}$$

One may notice that unaligned and aligned CLexSR, as defined in Sections 3.3.1.1 and 3.3.1.2, correspond to the continuous recognition of one linguistic descriptor ($M = 1$). The form of the unique descriptor d^1 is detailed in Equations 3.8 and 3.12.

Because we are considering *continuous* recognition, the temporal dimension is necessarily present in Y_{CSLR}. Without loss of generality, we will suppose that all descriptors $d^m, m \in \{1, \dots, M\}$ have a temporal dimension of length T, that is the original number of video frames, like in the case of *aligned* CLexSR (going from $Y_{\text{CLexSR,A}}$ to $Y_{\text{CLexSR,U}}$ is straightforward, as it consists in removing duplicates and frames with no label). With this assumption, we can write:

$$Y = \begin{bmatrix} d_1^1 & \cdots & \cdots & \cdots & d_T^1 \\ \vdots & & \ddots & & \vdots \\ d_1^M & \cdots & \cdots & \cdots & d_T^M \end{bmatrix}. \tag{5.4}$$

As SLs are four-dimensional languages [Vermeerbergen et al., 2007, (Sallandre, p. 103)], with signs and realizations located not only in time but also in the three dimensions of space, each d_t^m ($m \in \{1, \dots, M\}, t \in \{1, \dots, T\}$) could also include spatial information – for instance they could be described by a vector of size 3, indicating the location of each sign realization. However, for sake of simplicity, and because we have no knowledge of a CSL corpus that would be annotated both in space and time, we will consider each d_t^m as a scalar. Each of these scalars can be binary, categorical or continuous, depending on the associated information.

In Table 5.1, we give an example of fine CSLR, with d^1 encoding recognized Fully Lexical Signs (FLSs) – categorical –, d^2 the presence/absence of Depicting Signs (DSs) – binary –, d^2 the presence/absence of Pointing Signs (PTSs) – binary – and d^4 the presence/absence of Fragment Buoys (FBuoys) – binary.

For categorical (including binary) descriptors, these scalars can be encoded by *one-hot* vectors for the annotations, while prediction models usually estimate probabilities for each category. For instance, for a classification problem with five categories $c \in \{1, 2, 3, 4, 5\}$, $d_t^m = 4$ is equivalent to $d_t^m = \begin{bmatrix} 0 & 0 & 0 & \mathbf{1} & 0 \end{bmatrix}$, while a (good) prediction could look like $\widehat{d_t^m} = \begin{bmatrix} 0.13 & 0.01 & 0.19 & \mathbf{0.62} & 0.31 \end{bmatrix}$.

5.2 Performance metrics

As mentioned earlier, each descriptor d^m should be assigned an adapted performance metric $\mathscr{P}^m$, which is developed in this section.

5.2.1 Frame-wise performance metrics

5.2.1.1 Accuracy

Each categorical descriptor d^m, like the continuous – aligned – recognition of FLS glosses, can be analyzed with a simple accuracy metric Acc^m:

$$\mathrm{Acc}^m = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(d_t^m, \widehat{d_t^m}) = \frac{\#\text{ correctly labeled frames}}{T} \tag{5.5}$$

where $\mathbb{1}$ is the identity function (see Equation 3.14). Of course, accuracy can also be used for binary descriptors, which is a specific case of categorical descriptor with two categories.

5.2.1.2 Precision, recall and F1-score

Binary descriptors, which can be seen as categorical with two possible values, often correspond – in our case – to relatively *rare* events, such that predicting the value "0" for all frames may correspond to a very high accuracy. In order to address this issue, one may resort to the calculation of precision P and recall R:

$$P^m = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = \frac{\sum_t d_t^m \widehat{d_t^m}}{\sum_t \widehat{d_t^m}} \tag{5.6}$$

$$R^m = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{\sum_t d_t^m \widehat{d_t^m}}{\sum_t d_t^m} \tag{5.7}$$

where TP, FP and FN respectively stand for true positives, false positives and false negatives. The precision is then the ratio of correct positive predictions with respect to the total number of positive predictions, whereas the recall is the ratio of correct positive predictions with respect to the total number of positive annotations. The F1-score, defined as the harmonic mean of precision and recall, is then used as a trade-off metric for binary classification:

$$\mathrm{F1}^m = 2\left((P^m)^{-1} + (R^m)^{-1}\right)^{-1}. \tag{5.8}$$

One advantage of F1-score is that the minimum of the two performance values is emphasized, which can be seen on a graph for this function, shown on Figure 5.2.

As a matter of fact, precision and recall – and thus F1-score – can actually be generalized to non-binary values, with the following definitions:

$$P^m = \frac{\sum_t \mathbb{1}(d_t^m = \widehat{d_t^m} \text{ and } d_t^m \neq 0)}{\sum_t \mathbb{1}(\widehat{d_t^m} \neq 0)} \tag{5.9}$$

$$R^m = \frac{\sum_t \mathbb{1}(d_t^m = \widehat{d_t^m} \text{ and } d_t^m \neq 0)}{\sum_t \mathbb{1}(d_t^m \neq 0)} \tag{5.10}$$

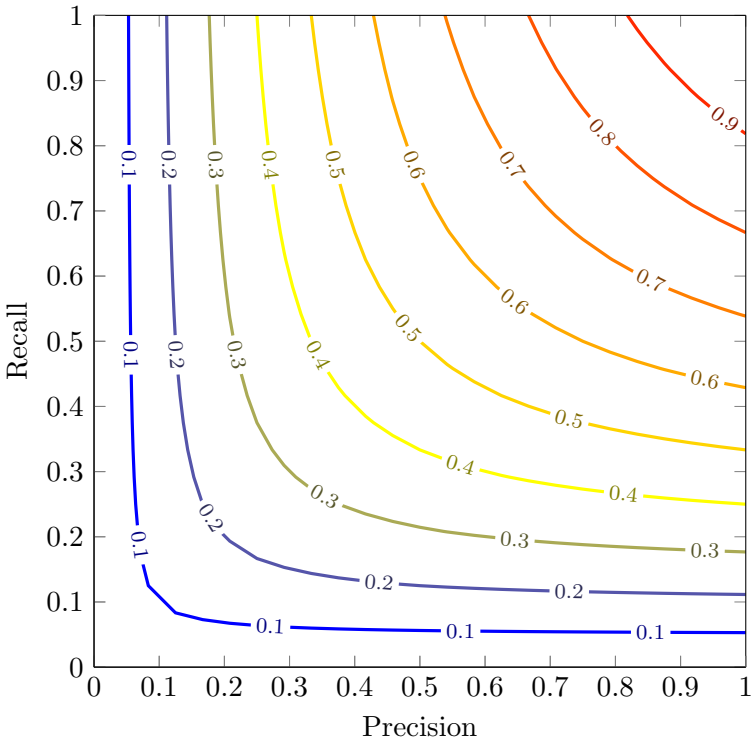


Figure 5.2: F1-score as a function of precision and recall (contour plot)

5.2.2 Unit-wise refined metrics for activity recognition and localization

Although accurate temporal localization is aimed for, frame-wise performance metrics may not be perfectly informative. Indeed, because the start and end of each unit can be quite subjective, even a good recognition model can get poor frame-wise Acc, P, R, F1 *etc.*, especially if the units are short, like in the case of PTSs (*cf.* Table 4.3). Unit-level metrics are then needed to get a better perspective on a system performance. They require to define what a good prediction is, with variants detailed in the next paragraphs. A detailed calculation for these metrics is given in Appendix A, with an example in the case of binary classification.

The starting point is to list U_G, corresponding to the set of all ground-truth annotated units, and U_D, corresponding to that of all detected units. The notion of precision and recall for categorical values in a temporal sequence format can then be extended to units. True and false positives and negatives are counted with respect to two points of view: either analyzing each annotated unit – and deciding whether it is sufficiently *close* to any detected unit –, or each detected unit – and deciding whether it is sufficiently *close* to any annotated unit. Modified versions of precision and recall are defined as follows:

$$P^\star = \frac{\#\text{ of correctly detected units w.r.t. } U_D}{\#\text{ of detected units}} = \frac{1}{|U_D|} \sum_{u_d \in U_D} \text{IsCorrectlyPredicted}(u_d, U_G) \quad (5.11)$$

$$R^\star = \frac{\#\text{ of correctly detected units w.r.t. } U_G}{\#\text{ of annotated units}} = \frac{1}{|U_G|} \sum_{u_g \in U_G} \text{IsCorrectlyPredicted}(u_g, U_D) \quad (5.12)$$

where IsCorrectlyPredicted is a counting function (values are 0 or 1). The F1-score is defined as in Equation 5.8.

5.2.2.1 Counting units within a certain temporal window t_w

We propose a first and straightforward counting function that consists in positively counting a unit $u_d \in U_D$ if and only if there exists a unit of the same class in U_G, within a certain margin t_w – respectively a unit $u_g \in U_G$ is counted positively if and only if there exists a unit of the same class in U_D, within a certain margin t_w.

More precisely, the time gap between the middle of u_d and the middle of the closest unit of the same class in U_G is compared to t_w, in order to decide whether u_d is a correct detection – respectively, the time gap between the middle of u_g and the middle of the closest unit of the same class in U_D is compared to t_w, in order to decide whether u_g is correctly detected.

In this configuration, precision, recall and F1-score are named $P^\star_w(t_w)$, $R^\star_w(t_w)$ and $\text{F1}^\star_w(t_w)$.

5.2.2.2 Counting units with thresholds $\bar{t}_p$ and $\bar{t}_r$ on their normalized temporal intersection

Wolf et al. [2014] have proposed and applied a similar but refined set of metrics, adapted for human action recognition and localization, both in space and time:

> *This measure is designed to penalize information loss, which occurs if actions or (spatial or temporal) parts of actions have not been detected, and it should penalize information clutter, i.e. false alarms or detections which are (spatially or temporally) larger than necessary.*

Because our data are only labeled in time, we set aside the space metrics, although they would definitely be useful with adapted annotations. The metrics derived below are thus slightly modified with respect to [Wolf et al., 2014].

In this setting $P^\star$ and $R^\star$ are calculated by finding the *best matching units*. Precision matches each unit of the detected list to one of the units in the ground truth list, whereas recall matches each unit of the ground truth to one of the units in the detection list, as for $P^\star_w(t_w)$ and $R^\star_w(t_w)$.

For each unit u_d in the list U_D, one can define the best match unit in U_G as the one maximizing the normalized temporal overlap between units (and a symmetric formula for the best match of a unit u_g in U_D):

$$\text{BestMatch}\left(u_d, U_G\right) = \underset{u_g \in U_G}{\text{argmax}} \begin{cases} \dfrac{2 \ \#\text{ of frames}\left(u_g \cap u_d\right)}{\#\text{ of frames}\left(u_g\right) + \#\text{ of frames}\left(u_d\right)} & \text{if Class}(u_g) = \text{Class}(u_d) \\ 0 & \text{otherwise} \end{cases} \tag{5.13}$$

$$\text{BestMatch}\left(u_g, U_D\right) = \underset{u_d \in U_D}{\text{argmax}} \begin{cases} \dfrac{2 \ \#\text{ of frames}\left(u_g \cap u_d\right)}{\#\text{ of frames}\left(u_g\right) + \#\text{ of frames}\left(u_d\right)} & \text{if Class}(u_g) = \text{Class}(u_d) \\ 0 & \text{otherwise.} \end{cases} \tag{5.14}$$

$P^\star_{pr}(\bar{t}_p, \bar{t}_r)$ and $R^\star_{pr}(\bar{t}_p, \bar{t}_r)$ are then expressed as:

$$P^\star_{pr}(\bar{t}_p, \bar{t}_r) = \frac{1}{|U_D|} \sum_{u_d \in U_D} \text{IsMatched}\left(\text{BestMatch}\left(u_d, U_G\right), u_d, \bar{t}_p, \bar{t}_r\right) \tag{5.15}$$

$$R^{\star}_{pr}(\bar{t}_p, \bar{t}_r) = \frac{1}{|U_G|} \sum_{u_g \in U_G} \text{IsMatched}\left(u_g, \text{BestMatch}\left(u_g, U_D\right), \bar{t}_p, \bar{t}_r\right). \tag{5.16}$$

IsMatched decides whether two units are sufficiently similar, with two criteria (not counting the fact that the two units must obviously belong to the same class):

- The number of frames which are part of both units is sufficiently large with respect to the number of frames in the detected set, *i.e.* the detected excess duration is sufficiently small.
- The number of frames which are part of both units is sufficiently large with respect to the number frames in the ground truth set, *i.e.* a sufficiently long duration of the unit has been found.

This can be written down as follows:

$$\text{IsMatched}\left(u_g, u_d, \bar{t}_p, \bar{t}_r\right) = \begin{cases} 1 & \text{if } \begin{cases} \dfrac{\#\text{ of frames}\left(u_g \cap u_d\right)}{\#\text{ of frames}\left(u_d\right)} > \bar{t}_p \\ \dfrac{\#\text{ of frames}\left(u_g \cap u_d\right)}{\#\text{ of frames}\left(u_g\right)} > \bar{t}_r \\ \text{Class}(u_g) = \text{Class}(u_d) \end{cases} \\ 0 & \text{otherwise.} \end{cases} \tag{5.17}$$

In the end, the F1-score can be derived similarly to Equation 5.8:

$$\text{F1}^{\star}_{pr}(\bar{t}_p, \bar{t}_r) = 2\left(P^{\star}_{pr}(\bar{t}_p, \bar{t}_r)^{-1} + R^{\star}_{pr}(\bar{t}_p, \bar{t}_r)^{-1}\right)^{-1}. \tag{5.18}$$

Equations 5.15, 5.16 and 5.18 enable to plot curves for variable thresholds $\bar{t}_p$ and $\bar{t}_r$. Integrated metrics can finally be defined as follows:

$$I_p = \int_0^1 \text{F1}^{\star}(\bar{t}_p, 0)\, \mathrm{d}\bar{t}_p \tag{5.19}$$

$$I_r = \int_0^1 \text{F1}^{\star}(0, \bar{t}_r)\, \mathrm{d}\bar{t}_r \tag{5.20}$$

which correspond to areas under curves of $\text{F1}^{\star}$, and a final average measure:

$$I_{pr} = \frac{1}{2}\left(I_p + I_r\right). \tag{5.21}$$

Other interesting values include $P^{\star}_{pr}(0,0)$, $R^{\star}_{pr}(0,0)$ and $\text{F1}^{\star}_{pr}(0,0)$, that correspond to counting matches as units with at least one intersecting frame.

5.3 Training error function

5.3.1 Categorical cross-entropy

When training Neural Networks (NNs), global metrics like the ones proposed in Section 5.2 are not sufficient. A continuously differentiable function – the training *loss* – is required, in order to use gradient descent algorithms. For the m^{th} categorical descriptor (including the case of binary descriptors), the categorical cross-entropy $\mathscr{L}^m$ is an adapted metric. It estimates the discrepancy between ground

truth values d_t^m and predicted values $\widehat{d_t^m}$ – encoded with a *one-hot* vector format:

$$\mathscr{L}^m = -\sum_{t=1}^{T}\sum_{c=1}^{C} d_t^m(c)\log\left(\widehat{d_t^m}(c)\right). \tag{5.22}$$

This is used to update the network weights and prevent overfitting, but can not be easily interpreted to estimate model performance.

5.3.2 Weighted error

Because of class imbalance, especially in the case of the detection of rare events, one may penalize very frequent classes with class weights α_c. These weights are usually computed as:

$$\alpha_c = \left(\frac{\#\text{ training frames with label } c}{\#\text{ training frames}}\right)^{-1}. \tag{5.23}$$

The weighted training loss can then be expressed as:

$$\mathscr{L}^m = -\sum_{t=1}^{T}\sum_{c=1}^{C} \alpha_c d_t^m(c)\log\left(\widehat{d_t^m}(c)\right). \tag{5.24}$$

Last, we can look for a compromise between unweighted loss 5.22 and 5.24 by introducing a global weight correction factor $\beta \in [0,1]$ such that different levels of weight correction can be used:

$$\mathscr{L}^m = -\sum_{t=1}^{T}\sum_{c=1}^{C} ((1-\beta)+\beta\alpha_c) d_t^m(c)\log\left(\widehat{d_t^m}(c)\right). \tag{5.25}$$

Unweighted loss then corresponds to $\beta = 0$ and fully weighted loss to $\beta = 1$.

Conclusion

In this chapter, we have introduced the second contribution of this thesis, that is a broader definition for Continuous Sign Language Recognition along with some proposals in terms of performance metrics. It consists of stacking many linguistic descriptors together, in an effort to describe SL discourse as precisely as possible, according to available linguistic annotations. Performance metrics can be computed at the frame level or unit-wise, which is most likely more meaningful. Last, we also discussed training error functions.

In the next chapter, we will introduce an adapted framework that is both generalizable and compact.

Chapter 6

A generalizable and compact framework

In Chapter 3, two main frameworks for Sign Language Recognition (SLR) were discussed. On the one hand, signer representation features can be built independently from the learning model, as a first step. On the other hand, some end-to-end frameworks only have one learning phase, in which the signer representation and SLR part are simultaneously learned.

Both approaches have benefits and drawbacks. The end-to-end frameworks – using Convolutional Neural Networks (CNNs) for instance – are easier to set up and do not require any prior knowledge on the signer representation. However, they require more data and may not be easily generalizable, for instance they may be sensible to changes in scaling appearance, lighting *etc.* When signer representation and learning model are decoupled, the generalizability with respect to new types of videos is increased, and one does not need to retrain the whole network in case new linguistic descriptors are added to the model. The reduced demand on training data is also an important benefit of such models, as annotated SL corpora are not that large, and allows for much faster training. Also, the *black box* architecture of end-to-end models does not enable one to get a straightforward feedback on which signer features are linguistically relevant for recognition.

For all these reasons, especially the fact that available training data are limited in quantity, we have chosen to resort to a separated approach. Section 6.1 details our proposal for a relevant, light and generalizable signer representation, then Section 6.2 outlines how such a signer representation can be coupled to a Recurrent Neural Network (RNN) for general Continuous Sign Language Recognition (CSLR).

6.1 Signer representation

Since available training data are limited in quantity, we have decided to partly rely on pre-trained models for signer representation. Therefore, it made sense to handle upper body, face and hands separately – which are usually dealt with in very specific ways, whether in SL-specific or non SL-specific models.

6.1.1 Upper body processing

While ten years ago, most SLR methods were usually based on optical flow and skin color detection [Gonzalez Preciado, 2012; Cooper et al., 2011; Lefebvre-Albaret, 2010], CNNs have emerged as a very

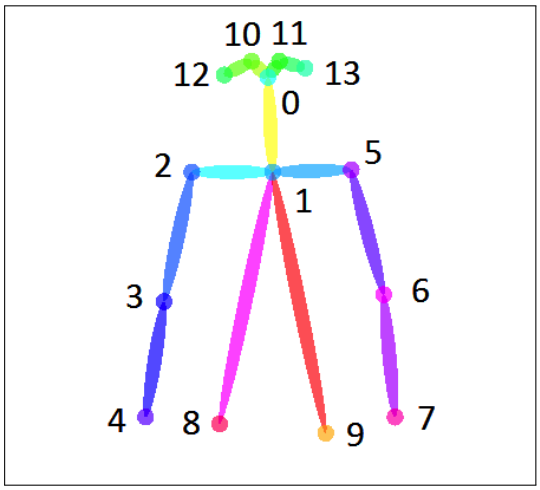


Figure 6.1: 2D upper body keypoints from OpenPose

effective tool to get relevant features from images. OpenPose (OP) [Cao et al., 2017; Wei et al., 2016] is a powerful open source library, with real-time capability for estimating 2D body pose.

In this section, we discuss the benefits of OP and variants. We then introduce more advanced 3D pose estimation modules, that should enable to leverage more information than plain 2D. Indeed, as discussed above, SLs are 3-dimensional in space, by nature.

6.1.1.1 2D pose (`Image` → `2D`)

Widely used in the gesture recognition community, the 2D body pose estimation module of OP has multiple benefits. It is fast – close to real-time for 25 frames per second (fps) videos on a modestly powerful desktop computer –, easy to use and works well even when part of the body is missing from the image. This is a great benefit, as most SLs videos only show the upper body. Many other pose estimation models we have experimented do not offer this feature.

As a matter of fact, we only keep the 14 upper body keypoints and leave out the leg keypoints. An illustration is given in Figure 6.1.

As pointed out by the work of Dilsizian et al. [2016, 2018], and since SLs are 3-dimensional, 3D pose might be a good aid for SLR. Therefore, we developed a 3D pose estimation model, presented below.

6.1.1.2 3D pose (`Image` → `3D`)

Although `Image` → `3D` models do exist [Xiang et al., 2018; Pavlakos et al., 2017; Yang et al., 2017; Rogez et al., 2017], we have not been able to find one fitting our requirements. Indeed, as for `Image` → `2D` models, prediction usually fails when part of the body is missing from the image, or when the person is not centered with respect to the image. Another type of issue is related to the training data of these models. As they have not been trained with SL data, our experience is that they do not perform well when fed with SL images.

6.1.1.3 3D pose (`Image` → `2D` → `3D`)

Fortunately, the 2D OP estimates have proven robust even on SL videos. Therefore, we decided to rely on OP in order to get good 2D estimates (see Section 6.1.1.1), then train a `2D` → `3D` Deep Neural Network (DNN), reproducing the architecture from [Zhao et al., 2016] (see Figure 6.2).

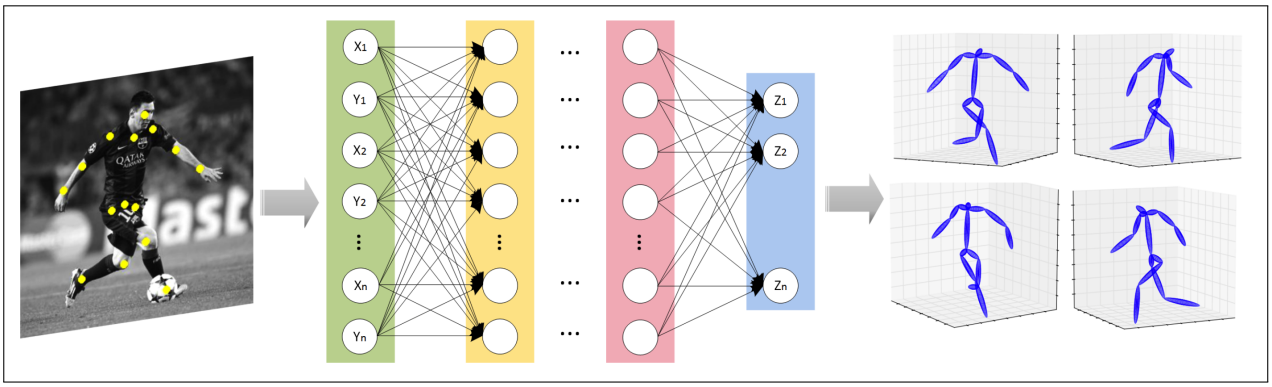


Figure 6.2: Deep Neural Network architecture for 2D → 3D estimation [Zhao et al., 2016]

Formalization Such a model aims to learn the function f that estimates the third coordinate for each landmark of the signer in frame t, that is, with n as the number of landmarks ($n = 14$ in our case):

$$f : \begin{array}{rcl} \mathbb{R}^{2n} & \longrightarrow & \mathbb{R}^{n} \\ [(\widehat{x}_{1t}, \widehat{y}_{1t}), (\widehat{x}_{2t}, \widehat{y}_{2t}), \ldots, (\widehat{x}_{nt}, \widehat{y}_{nt})] & \longmapsto & [\widehat{z}_{1t}, \widehat{z}_{2t}, \ldots, \widehat{z}_{nt}] \end{array} \tag{6.1}$$

where $\widehat{x}_{it}$, $\widehat{y}_{it}$ and $\widehat{z}_{it}$ are standardized version of the original coordinates x_{it}, y_{it} and z_{it}, with respect to the whole training dataset. Standardization if applied so that the process in invariant to translations and scaling:

$$\begin{cases} \widehat{x}_{it} &= \dfrac{x_{it} - \bar{x}_i}{\Sigma} \\ \widehat{y}_{it} &= \dfrac{y_{it} - \bar{y}_i}{\Sigma} \\ \widehat{z}_{it} &= \dfrac{z_{it} - \bar{z}_i}{\Sigma} \end{cases} \tag{6.2}$$

$$\Sigma = \frac{1}{2}\left(\sigma_{\text{stdev}}(x_i) + \sigma_{\text{stdev}}(y_i)\right) \tag{6.3}$$

with average values $\bar{x}_i$, $\bar{y}_i$ and $\bar{z}_i$ and standard deviations $\sigma_{\text{stdev}}(x_i)$ and $\sigma_{\text{stdev}}(y_i)$ computed on the whole training dataset. Note that $\sigma_{\text{stdev}}(z_i)$ is not used. The reason is that Equations 6.1 and 6.2 must be invertible if non-normalized predictions are to be computed on a test set, where obviously z_i is unknown.

The training loss is defined as the Euclidean distance between predictions and ground-truth data.

Choice of training data The training data we decided to use for training consisted in motion capture data from the French Sign Language (LSF) corpus MOCAP1 [LIMSI and CIAMS, 2017]. This data has been particularly valuable since it contains high precision 3D landmarks recording of LSF, from four different signers.

Data preprocessing Because the position of sensors in MOCAP1 did not exactly match that of OP landmarks – as it can be seen in Figure 6.3 – preprocessing motion capture data mostly consisted in establishing simple relations between the latter and the former. In detail, each OP landmark position was estimated as a linear combination of some motion capture sensors position – for instance, OP

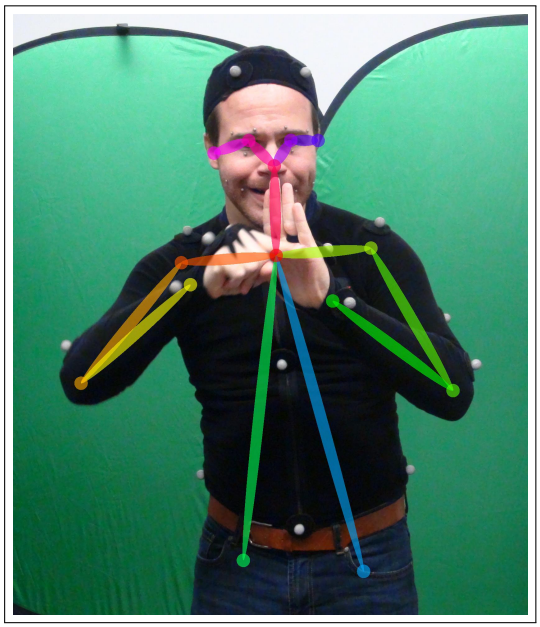


Figure 6.3: OpenPose estimate on a frame from MOCAP1 [LIMSI and CIAMS, 2017]. The motion capture sensors and OpenPose keypoints position do not match perfectly.

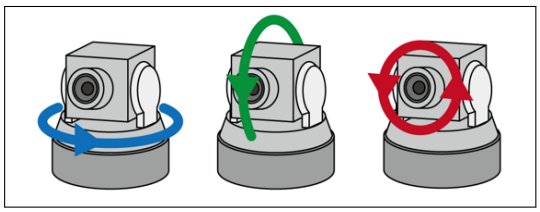


Figure 6.4: Camera angles: pan, tilt, roll

wrist position is calculated as the average of the position of two sensors that were placed around the wrist during motion capture.

Data augmentation In order to increase model generalizability, data augmentation techniques are used during training. In detail, the 3D data from MOCAP1 are randomly rotated at each training epoch, with added pan $\Delta\theta_p \in [-45°, +45°]$, added tilt $\Delta\theta_t \in [-20°, +20°]$ and added roll $\Delta\theta_r \in [-5°, +5°]$ (see angles definition on Figure 6.4). With this technique, the generalizability of the trained model is drastically increased.

Implementation details The proposed DNN is implemented with Keras [Chollet et al., 2015] on top of Tensorflow [Abadi et al., 2016]. All hidden layers use Rectified Linear Unit activation [Nair and Hinton, 2010], with Dropout to prevent overfitting [Srivastava et al., 2014]. RMSProp is used as the gradient optimizer [Tieleman and Hinton, 2012]. Six neuron layers are stacked, with sizes $[28, 28, 28, 28, 28, 14]$.

The proposed `Image` $\rightarrow$ `2D` $\rightarrow$ `3D` pipeline for processing the 3D upper body pose from signers in Red-Green-Blue (RGB) frames is shown in Figure 6.5.

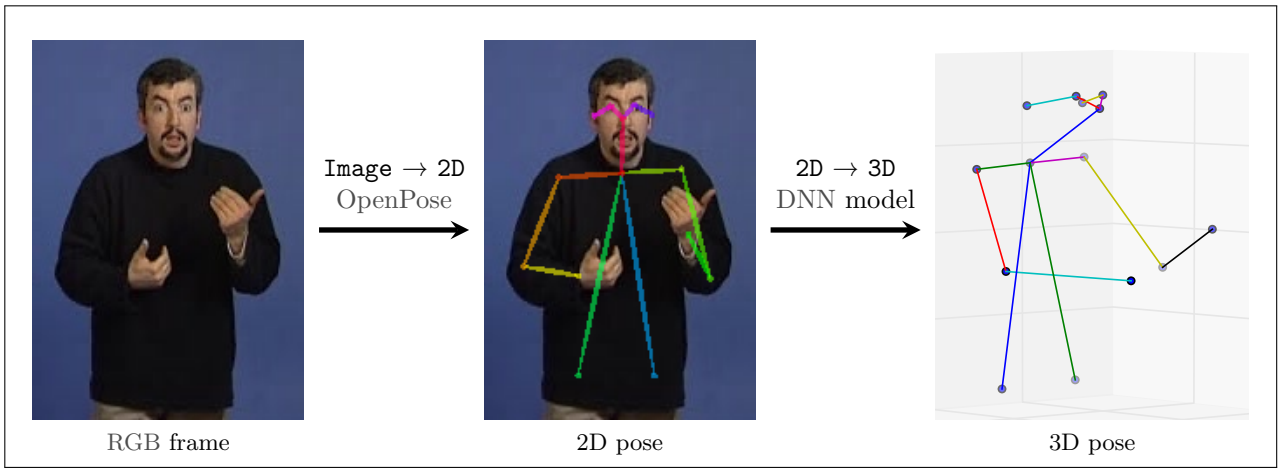


Figure 6.5: Proposed `Image` → `2D` → `3D` pipeline for the upper body pose, applied to a random frame from the French Sign Language corpus LS-COLIN [Braffort et al., 2001]. OpenPose enables to get 2D estimates, then a DNN model is used to estimate the missing third coordinate of each landmark.

6.1.2 Hand processing

Hands are obviously one of the main articulators in SL. Although linguists do not all share a common ground for the description and linguistic role of sub-units for the hands, three important parameters have been identified. More specifically – at least from an articulatory point of view – the location, shape and orientation of both hands are known to be critical, along with the dynamics of these three variables, that is: hand trajectory, shape deformation and hand rotation.

In this section, we present different possible ways of processing hands.

6.1.2.1 2D pose (`Image` → `2D`)

In addition to the body pose feature, the OP library also includes a 2D hand pose estimation module, with RGB images as input [Simon et al., 2017], illustrated on Figure 6.6.

From our experience, this module is quite sensitive to the image resolution, and even more to the video frame rate. Indeed, very poor results are obtained on blurred images, which is often the case for the hands with 25-30 fps videos in standard resolution. This is mostly due to the fact that hands can move fast in SL production, causing motion blur arounds the hands and forearms.

Moreover, the hand shapes used in SL can be very sophisticated, and somehow never used in the daily life of non-signing people. Therefore, in all likelihood, the data that were used to train the OP models did not include such hand configurations, which sometimes makes predictions unreliable.

That being said, the OP hand module can still be seen as a good and light proxy for hand representation.

6.1.2.2 3D pose (`Image` → `3D`)

Ideally, one would greatly benefit from a frame-wise 3D hand pose estimate on RGB images. Hands are indeed an extremely fine means of carrying information in SL discourse with strong iconic properties in a natively continuous space, possible interactions between dominant and weak hands, conventional hand shapes to represent proforms, *etc.*

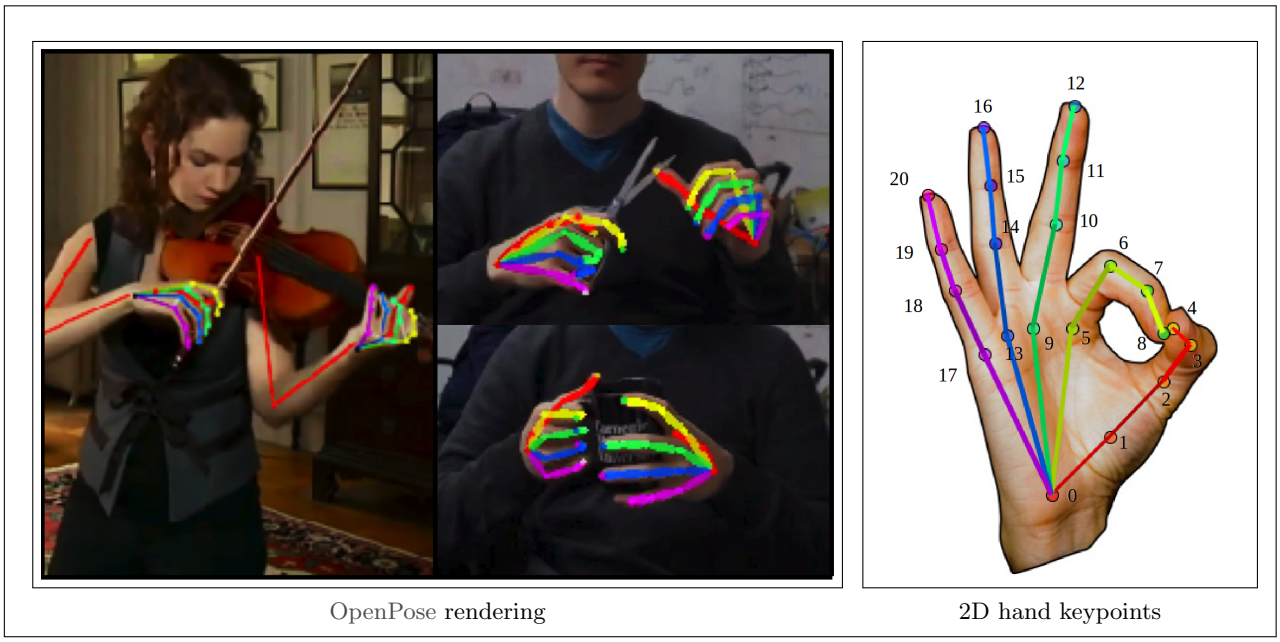


Figure 6.6: 2D hand pose renderings from OpenPose [Simon et al., 2017], with keypoints numbering.

Although such algorithms have been developed [Xiang et al., 2018; Iqbal et al., 2018; Spurr et al., 2018; Zimmermann and Brox, 2017; Mueller et al., 2017; Panteleris et al., 2018], we have not found any that was able to provide a reliable estimate on hand pose on real-life 25 fps SL videos. Indeed, the same issues as described in Section 6.1.2.1 are encountered. First, the motion blur issue described earlier makes the frame-wise estimation of hand pose very difficult. Also, most of these models were trained with still images of high resolution, which is usually not the case in SL videos. Furthermore, specific SL hand shapes are almost never reconstructed correctly, as they were not seen during the training phase of these sophisticated models.

6.1.2.3 Hand shape estimates

While 3D hand pose estimation models – and, to a lesser extent, 2D models – have not appeared to be reliable to this day when applied to real-life RGB videos, another possible direction is to extract more global features from hand crops.

As written earlier, hand parameters include location, shape and orientation. Focusing on hand shape – thus setting aside location and orientation –, a SL-specific model was developed in [Koller et al., 2016a]. This CNN model classifies cropped hand images into 61 predefined hand shapes classes.

Training data and generalizability The model was trained on more than a million frames, including motion blurred images. Three SL corpora of different types were compiled, with different proportions: isolated lexical signs from Danish Sign Language (DTS) – 12% – and New Zealand Sign Language (NZSL) – 23%; Continuous Sign Language (CSL) in German Sign Language (DGS) – 65%. The final hand shape distribution can be seen on Figure 6.7.

Even though the hand shapes frequency of occurrence is very likely to vary between different SLs, we have made the assumption that SLs other than DTS, NZSL and DGS could still be dealt with without retraining the model. Indeed, a lot of hand shapes are obviously shared across most SLs, since they are used to depict salient and/or primary forms (flat, round, square, *etc.*)

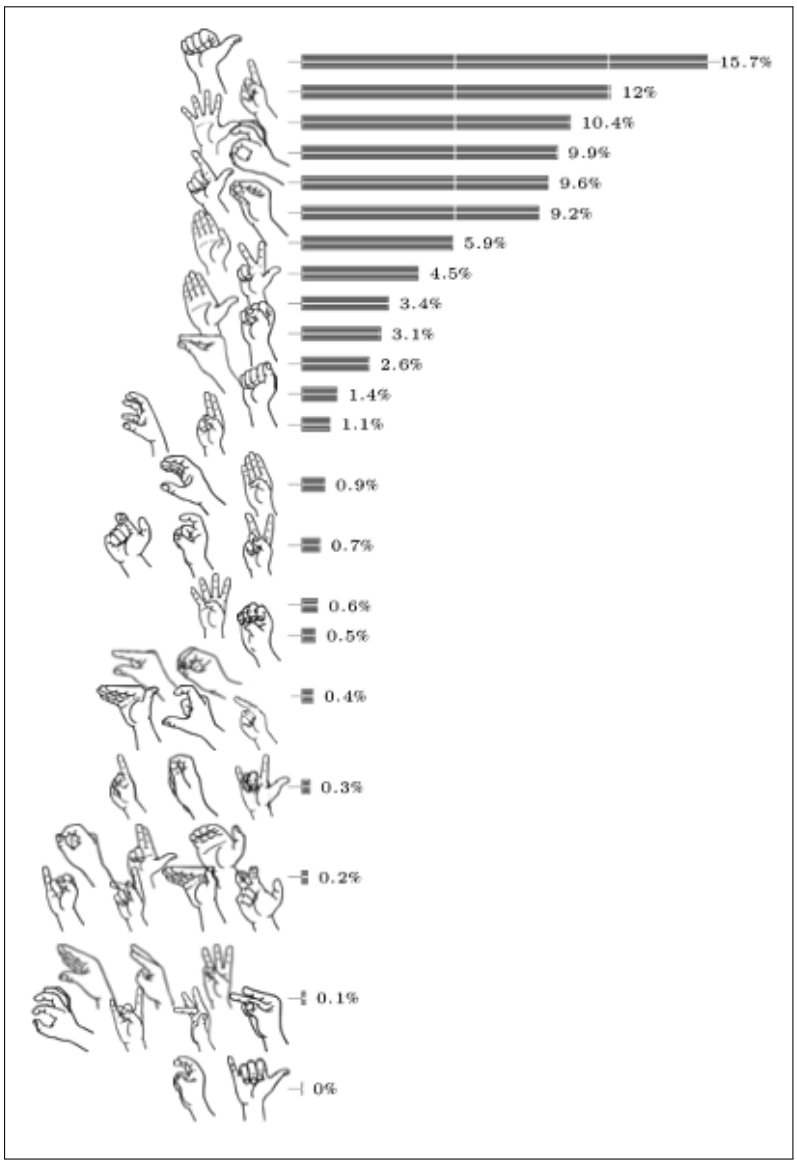


Figure 6.7: Hand shapes distribution in the dataset used by Koller et al. [2016a]

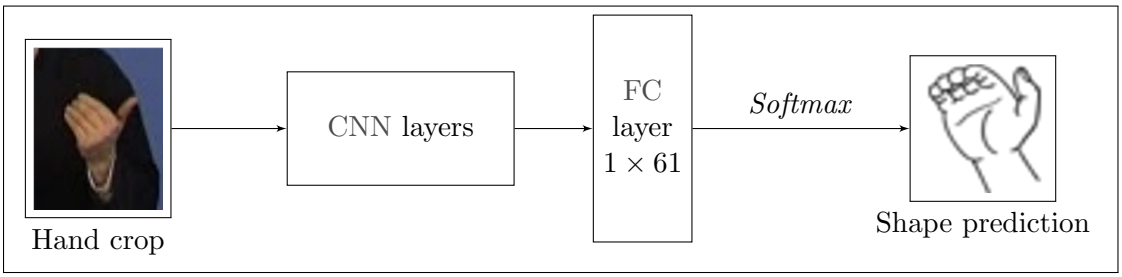
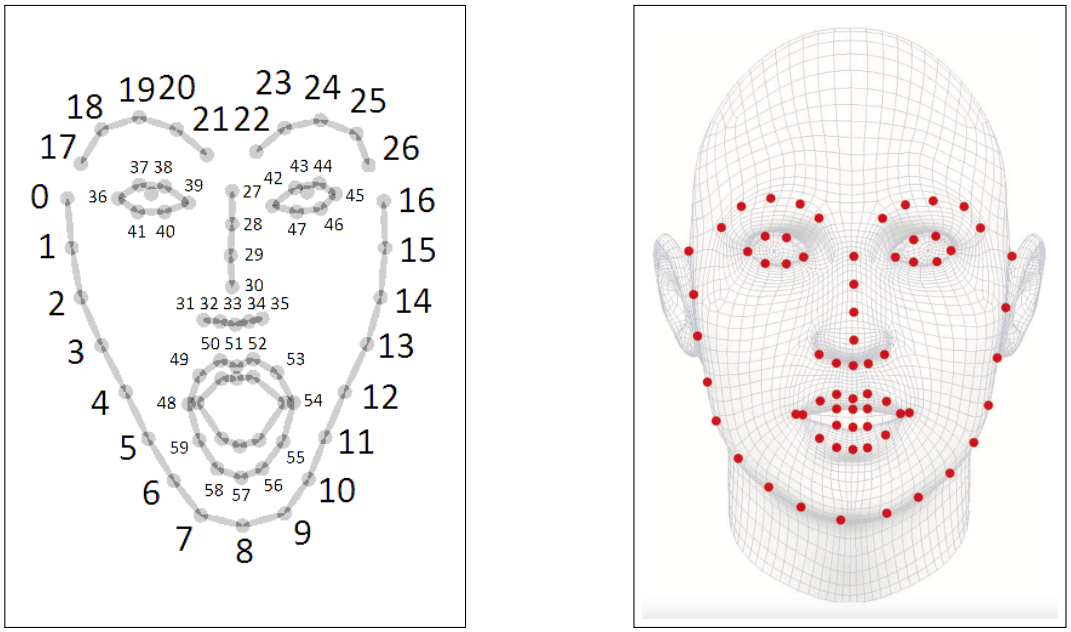


Figure 6.8: Synoptic architecture for the `1-miohands-v2` model from [Koller et al., 2016a]. Hand crop images are processed by several Convolutional Neural Network (CNN) layers, then a final Fully Connected (FC) layer enables to estimate probabilities for each of the 61 classes, with a softmax operation.



(a) 2D (OpenPose) (b) 3D [Bulat and Tzimiropoulos, 2017]

Figure 6.9: 2D (a) and 3D (b) face keypoints

Model usage and details The trained prediction model `1-miohands-v2` is publicly available[1], under the Caffe architecture [Jia et al., 2014]. A simplified scheme is presented on Figure 6.8. The input of the model is a cropped hand image, that is processed by several CNN layers. The final layer is of Fully Connected (FC) type, with 61 neurons, one for each class. The model outputs the most probable class with a softmax operation. However, one can also choose to extract the output of the last fully-connected layer and thus get a representation vector of size 61, for each hand, instead of the unique value corresponding to the most probable class.

Because the model takes centered hand crops as input, one should note that the position of each hand must be estimated first. To this end, we use the upper body estimation from OP – Koller et al. [2016a] originally used a hand tracker based on dynamic programming. With E_r and W_r as the right elbow and right wrist positions, the position of the center of the right hand H_r is estimated as

$$\overrightarrow{E_r H_r} \simeq 1.2\overrightarrow{E_r W_r} \tag{6.4}$$

that is

$$\begin{cases} x_{H_r} & \simeq & 1.2x_{W_r} - 0.2x_{E_r} \\ y_{H_r} & \simeq & 1.2y_{W_r} - 0.2y_{E_r} \end{cases} \tag{6.5}$$

where x_{W_r}, y_{W_r}, x_{E_r} and y_{E_r} are provided by the OP upper body estimate. The equations for the left hand position (x_{H_l}, y_{H_l}) are identical.

6.1.3 Face and head pose processing

6.1.3.1 2D pose (`Image` → `2D`)

Similarly to body pose – Section 6.1.1.1 – and to hand pose – Section 6.1.2.1 –, the OP library makes it possible to get an estimate on the 2D face pose. The 70 keypoints are shown on Figure 6.9a.

6.1.3.2 3D pose (`Image` → `3D`)

Although OP outputs a 2D estimate on the facial pose, a reliable 3D estimate is directly obtained from video frames thanks to a CNN model [Bulat and Tzimiropoulos, 2017] trained on 230,000 images.

The 68 landmarks are identical to that of the 2D model of Section 6.1.3.1, minus the two pupils. An illustration is given on Figure 6.9b.

6.1.4 Final signer representation: from raw data to relevant features

With X as video frames, the final signer representation $x = \mathscr{R}(X)$ that we propose is simply a combination of the previously introduced raw data, and/or manufactured features that are known or assumed to be relevant for SLR.

6.1.4.1 Hand shapes

Based on the Deep Hand model [Koller et al., 2016a], the x^{hand}_{shapes} vector consists of hand shapes probabilities for both hands, with size 122 (2×61 scalars per hand).

6.1.4.2 Raw pose data

Based on Sections 6.1.1, 6.1.2 and 6.1.3, a *raw* feature vector could be set up as a simple combination of some of the following raw body part features:

- $x^{b}_{\mathbf{raw2D}}$ 2D raw body pose, size 28 (14 2D landmarks)
- $x^{b}_{\mathbf{raw3D}}$ 3D raw body pose, size 42 (14 3D landmarks)
- $x^{fh}_{\mathbf{raw2D}}$ 2D raw face/head pose, size 140 (70 2D landmarks)
- $x^{fh}_{\mathbf{raw3D}}$ 3D raw face/head pose, size 204 (68 3D landmarks)
- $x^{hand}_{\mathbf{raw2D}}$ 2D raw hand pose, size 126 (2×[21 2D landmarks plus 21 confidence scores])

However, raw data might not be optimal for training SLR systems. Indeed, raw values are highly correlated, with a lot of redundancy, they can be difficult to interpret and are not always meaningful for SLR. Thus, we propose to derive a more relevant feature vector for body, face and head pose, after normalizing raw data with distance between shoulders, in order to increase generalizability.

6.1.4.3 Relevant 3D feature vector $x^{bfh}_{\mathbf{feat3D}}$

Drawing inspiration from previous work in gesture recognition [Granger and el Yacoubi, 2017; Wu et al., 2016; Neverova et al., 2014], a relevant feature vector x^{bfh}_{feat3D} should include pairwise positions and distances, as well as joint angles and orientations:

[1] https://www-i6.informatik.rwth-aachen.de/~koller/1miohands/

- Relative position (3D vector) and Euclidean distance of each hand with respect to parent elbow, plus first and second order derivatives,
- Relative position (3D vector) and Euclidean distance of each elbow with respect to parent shoulder, plus first and second order derivatives,
- Relative position (3D vector) and Euclidean distance of each shoulder with respect to point 1 in Figure 6.1, plus first and second order derivatives,
- Cosinus and orthonormal vector of the elbow and shoulder angles, plus first and second order derivatives.

In order to reduce the dimensionality of the face/head feature vector, the following components are computed and included in x^{bfh}_{feat3D}:

- 3 Euler angles for the rotation of the head, plus first and second order derivatives,
- Mouth size (horizontal and vertical distances),
- Relative motion of each eyebrow to parent eye center,
- Position of nose landmark with respect to point 1 in Figure 6.1.

In SL, the location of hands with respect to specific parts of the body and head is known to be related to families of concepts – which is a strong argument for the iconic origin of many signs [Östling et al., 2018]. Relatedly, the detection of contacts between hands and specific locations of the body is known to increase recognition accuracy [Dilsizian et al., 2018]. Therefore, the feature vector also includes the relative position between each wrist and the nose, plus first and second order derivatives.

Moreover, because SLs make intensive use of hands, their relative arrangement is crucial. Battison [1974], for instance, establishes a typology of bimanual signs based on the relative motion and shapes of hands. Therefore, we also include the relative position and distance of one wrist to the other to x^{bfh}_{feat3D}, plus first and second order derivatives.

Finally, the 3D feature vector is of size 176.

6.1.4.4 Relevant 2D feature vector $x^{bfh}_{\textbf{feat2D}}$

Because the advantage of 3D data over 2D is not proven, and because one may want to further reduce the feature vector size, we also derived a relevant 2D feature vector, in the same manner as the 3D one. In this case, positions, distances and angles are actually projected positions, distances and angles on the 2D plane.

With the exact same features as in Section 6.1.4.3, x^{bfh}_{feat2D} is of size 96.

6.2 Learning framework: a convolutional and recurrent architecture

With a signer representation $x = \mathscr{R}(X)$ as a concatenation of some of the vectors introduced in Section 6.1.4, setting up a learning and prediction model consists in defining $\mathscr{M}$ such that:

$$\begin{cases} \mathscr{M}(x) & = & \widehat{Y}_{\text{CSLR}} \\ \widehat{Y}_{\text{CSLR}} & \simeq & Y. \end{cases} \tag{6.6}$$

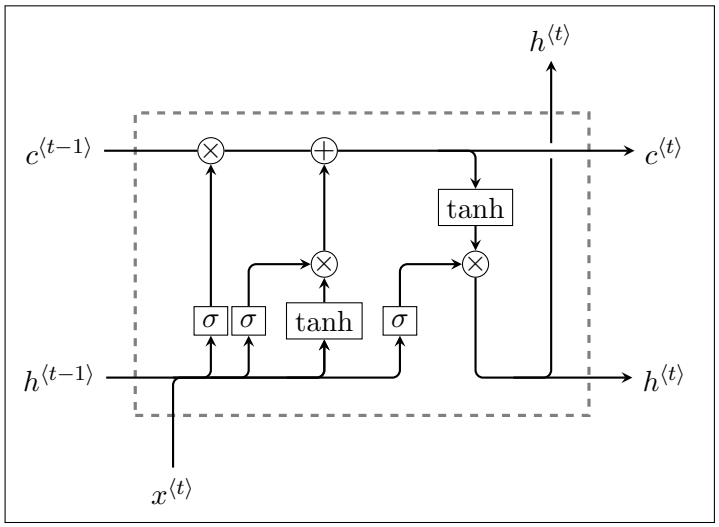


Figure 6.10: Diagram illustrating the internal function of a Long Short-Term Memory (LSTM) unit, with x as input, h as hidden state and c as cell state.

In Section 3.3.3, we presented different types of learning frameworks taking one-dimensional time-series as input, which is our case. The most effective architectures have used Hidden Markov Models (HMMs), Conditional Random Fields (CRFs) and RNNs.

HMMs are generative models, with the assumption that the system is a Markov process, with hidden states. One of the strong assumptions of HMMs is that state transitions only depend on the current state, not on anything in the past. Also, the number of states must be pre-defined. CRFs are discriminative models that can accommodate context information. The training phase of CRFs is somehow tricky, and does not allow for easy retraining when new data are available.

On the other hand, RNNs are discriminative models that use a form of memory to learn temporal dependencies. Conversely to HMMs and CRFs, they are very good at learning complex hidden features from data. In this sense, they are sometimes considered as feature extractors and integrated into RNN-HMM frameworks, for instance.

In this section, we present the compact architecture that we have built for training CSLR models, compatible with the signer representation detailed in Section 6.1. It is mainly designed around recurrent layers (Section 6.2.1), which are complemented by a first convolutional layer (Section 6.2.2). The final setup is summarized in Section 6.2.3.

6.2.1 Recurrent layers

In our experiments, we have chosen to use RNNs, mainly for the following reasons: they are good to build complex features from not always meaningful input, they are very modular and straightforward to train, and they exhibit the best results in the field of Gesture Recognition (GR) and SLR.

One issue with traditional RNNs is that they undergo vanishing gradient issues, making them impractical for learning long-time dependencies – or with high-frequency data. To overcome this issue, a specific type of RNN, referred to as Long Short-Term Memory (LSTM) was introduced [Gers et al., 2000].

6.2.1.1 Long Short-Term Memory units

LSTMs include a cell state c in addition to the usual hidden state h of RNNs. A schematic for the internal function of a LSTM unit is shown on Figure 6.10, with input x, cell state c and hidden state h. The equations of any LSTM cell can be written as:

$$\begin{cases} i^{\langle t \rangle} &= \sigma\left(W^{xi}x^{\langle t \rangle} + W^{hi}h^{\langle t-1 \rangle} + b^i\right) \\ f^{\langle t \rangle} &= \sigma\left(W^{xf}x^{\langle t \rangle} + W^{hf}h^{\langle t-1 \rangle} + b^f\right) \\ c^{\langle t \rangle} &= f^{\langle t \rangle}c^{\langle t-1 \rangle} + i^{\langle t \rangle}\tanh\left(W^{hc}h^{\langle t-1 \rangle} + W^{cx}x^{\langle t \rangle} + b^c\right) \\ o^{\langle t \rangle} &= \sigma\left(W^{ho}h^{\langle t-1 \rangle} + W^{xo}x^{\langle t \rangle} + b^o\right) \\ h^{\langle t \rangle} &= o^{\langle t \rangle}\tanh\left(c^{\langle t \rangle}\right) \end{cases} \tag{6.7}$$

where W are matrices and b are vectors which values are updated during training, and σ is the sigmoid or logistic function:

$$\sigma : \left| \begin{array}{ccc} \mathbb{R} & \longrightarrow &]0,1[\\ x & \longmapsto & \dfrac{1}{1+\exp(-x)}. \end{array} \right. \tag{6.8}$$

We have chosen LSTMs as the base unit for our learning networks, which can easily be stacked in several LSTM layers.

6.2.1.2 Stacked Bidirectional LSTM layers

When real-time predictions are not needed, forward and backward LSTM units can be paired to form Bidirectional LSTMs (BLSTMs). Several layers of BLSTMs can then be stacked, as shown on Figure 6.11 detailing a two-layer BLSTM. Our experiments usually include one to four BLSTM layers.

6.2.2 Adding temporal convolutions

An interesting addition for helping the network build relevant features is to set up the first layer as a one-dimensional temporal convolution. Temporal convolutions can help with noisy high-frequency data like ours, and are good to learn temporal dependencies [Pigou et al., 2018]. A convolution layer with a kernel width of three frames is included on Figure 6.11.

6.2.3 Final setup and list of parameters

Using the Keras library [Chollet et al., 2015] on top of Tensorflow [Abadi et al., 2016], we have then built a modular architecture[2] with the following options and parameters:

- A convolutional layer
  - Number of filters
  - Kernel size
- One or multiple LSTM or BLSTM layers

[2] https://github.com/vbelissen/cslr_limsi/

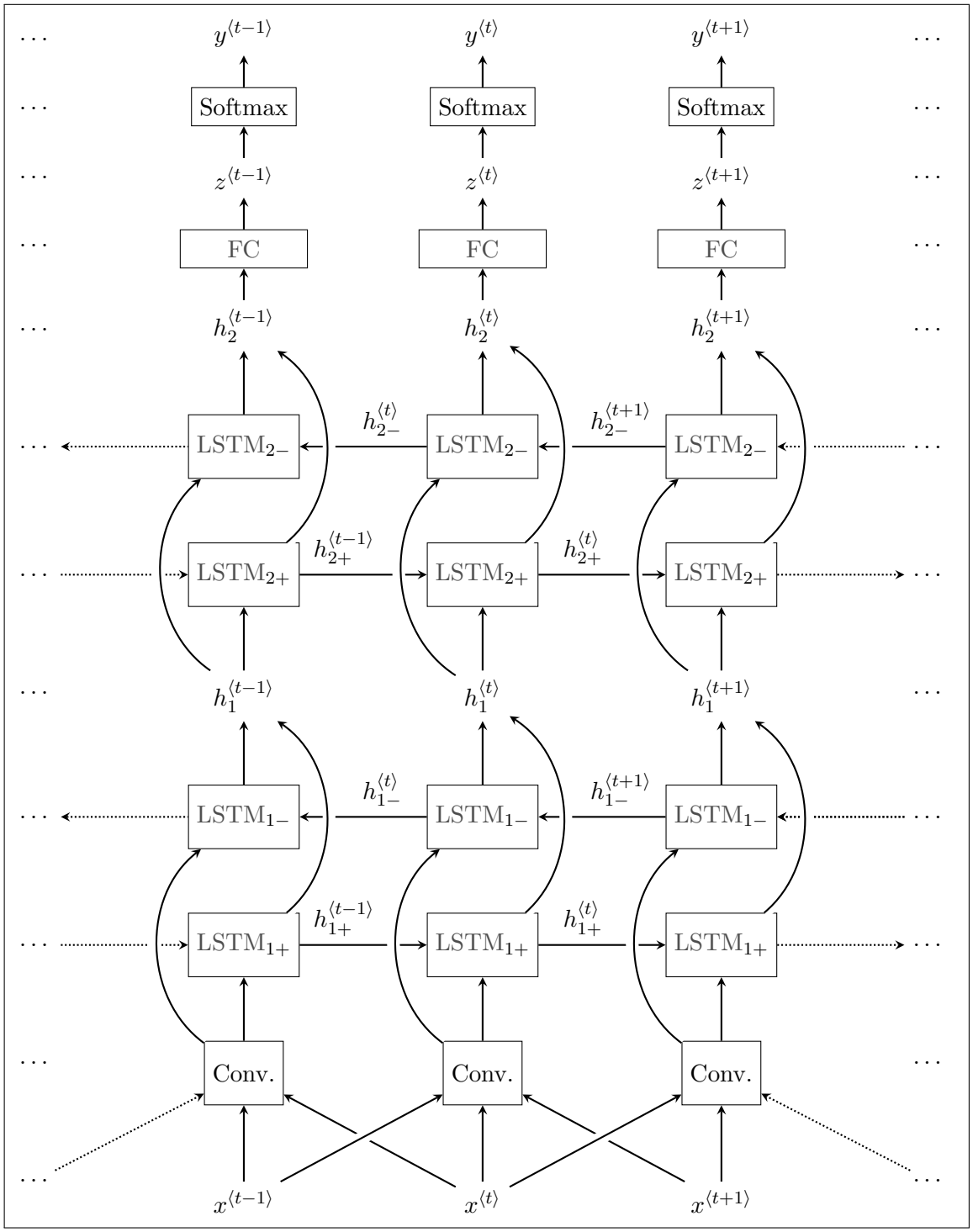


Figure 6.11: Unrolled representation of a two-layer Bidirectional LSTM (BLSTM) network for temporal classification, with input x and output y. The cell state c – visible on Figure 6.10 – is omitted, for sake of clarity. Upstream of the LSTM layers, the input is first convolved, with a convolution kernel width of three frames on this scheme.

  - Number of units

- A FC layer
- A final softmax operation for classification.

Finally, the training phase is associated to many parameters as well:

- Learning rate and optimizer
- Batch size
- Sequence length
- Dropout rate
- Data imbalance correction (*cf.* Equation 5.25)
- Choice of metric.

Conclusion

This chapter has presented the chosen strategies and architecture for the data representation and the CSLR model. This has led to an open access implementation, which is the third contribution of this thesis.

A new model requiring validation

In Part II of this thesis, we have sought to open the way to a broader acceptation of Continuous Sign Language Recognition (CSLR) and proposed three contributions to the field.

In Chapter 4, we have introduced some linguistic-driven SL corpora. Although the recorded SL is very natural, they are not directly usable for CSLR purposes, primarily because the annotation data are not consistent. Thus, we have developed a remake of a French Sign Language (LSF) dataset, Dicta-Sign–LSF–v2, that is a very natural dialogue corpus with fine and consistent annotation.

Then, we have proposed in Chapter 5 a newer formulation of the CSLR problem, consisting in the simultaneous recognition of many linguistic descriptors, instead of focusing on very conventionalized signs, also referred to as Fully Lexical Signs (FLSs). This recognition of descriptors captures a much more important part of SL discourse, for instance Depicting Signs (DSs) or Pointing Signs (PTSs), that are crucial for Sign Language Understanding (SLU). Along with it, we have suggested the use of appropriate performance metrics, especially at the level of manual units, that can be used to assess both time-wise and space-wise reliability of prediction models with respect to expert annotations.

Finally, an adapted framework was developed in Chapter 6. A compact and generalizable representation of signers in videos was laid out, which can be used as input to a large spectrum of learning models. In fact, we propose to use a convolutional Recurrent Neural Network (RNN) architecture, with Long Short-Term Memory (LSTM) as base units, which are known to be effective for learning long-time dependencies.

In the next and last part of this thesis, we aim to validate this work. In Chapter 7, we focus on the signer representation and learning architecture, in a quantitative way. Then, in Chapter 8, we analyze prediction results, more qualitatively, in order to validate the point of a wider acceptation of CSLR. Last, we lay out perspectives for the future of CSLR.

Part III

Validation, results and perspectives

Chapter 7

Model validation: quantitative results

The purpose of the current chapter is to validate different choices and proposals presented in Chapters 5 and 6. First, we introduce the general settings and choices for this chapter (Section 7.1). Then, we analyze the global performance of the proposed model and signer representation in Section 7.2, using a unique standard configuration. Subsequently, we analyze the sensitivity of the recognition performance with respect to the network parameters and training settings (Section 7.3). Last, we compare the performance of different signer representation options (Section 7.4).

7.1 General settings

In this section, we discuss the linguistic descriptors that we have decided to focus on for the model validation (Section 7.1.1), the metrics used (Section 7.1.2) and the common training settings for all experiments (Section 7.1.3).

7.1.1 Model outputs

In this chapter, we have decided to focus on the – binary – recognition of four manual unit types: Fully Lexical Signs (FLSs), Depicting Signs (DSs), Pointing Signs (PTSs) and Fragment Buoys (FBuoys). These categories are representative of the variety of SL linguistic structures, with conventional and illustrative units, as well as elements highly used within the diagrammatic iconicity of SL. Other types show a too small number of instances for the results to be significant, or even for the network to converge (see detail in Table 4.3 and Figure 4.5).

7.1.2 Metrics

Following the discussion of Section 5.2, the chosen performance metrics for validation include frame-wise and unit-wise measures, with detail below:

Frame-wise metrics:

- Accuracy
- F1-score

Unit-wise metrics:

- Margin-based F1-score $\text{F1}^\star_w(t_w)$, with margin $t_w = 12$ frames (half a second)
- Normalized intersection-based F1-score $\text{F1}^\star_{pr}(\bar{t}_p, \bar{t}_r)$, with $\bar{t}_p = 0, \bar{t}_r = 0$ (counting positive recognition for units with at least one intersecting frame)
- Associated integral value I_{pr}.

Although we have included it for the sake of completeness, the frame-wise accuracy is not really informative (see Section 5.2). Therefore, we mostly rely on unit-wise metrics and frame-wise F1-score to look for the best network settings, training hyperparameters and signer representation.

7.1.3 Common training settings

All training sessions, unless otherwise specified, are conducted with the following common settings:

- The training loss is the weighted binary/categorical cross-entropy of Equation 5.25.
- The gradient descent optimizer is RMSProp [Tieleman and Hinton, 2012].
- A common cross-validation split of the data is realized in a signer-independent fashion, with 12 signers in the training set, 2 in the validation set and 2 in the test set.
- Each run consists of 150 epochs. Only the best model is retained, in terms of performance on the validation set. During training, only the frame-wise F1-score is used to make this decision. Although we have argued that unit-wise metrics are better fitted for performance assessment, the frame-wise F1-score is actually strongly correlated with unit-wise metrics and is much faster to compute, which is why we focus on this measure during training.

7.2 Baseline performance of a standard configuration

As a first baseline, we develop in this section the results of a standard configuration. This reference point can then be compared to variations in the network architecture (Section 7.3) or in the signer representation (Section 7.4).

7.2.1 A standard configuration ($\mathcal{S}$)

The standard configuration is defined with the following settings, that are set to reach a sort of compromise, so that the model converges quickly while showing a good average performance:

In terms of network architecture:

- Network parameters:
  - One Bidirectional LSTM (BLSTM) layer;
  - 50 units in each LSTM cell;
  - 200 convolutional filters as a first neural layer, with a kernel width of size 3.
- Training hyperparameters:
  - A batch size of 200 sequences;
  - A dropout rate of 0.5;

		Frame-wise		Unit-wise		
		Acc	F1 $_{(P/R)}$	$F1^{\star}_{w}(t_w = 12)$ $_{(P/R)}$	$F1^{\star}_{pr}(0,0)$ $_{(P/R)}$	I_{pr}
FLS	μ	0.83	0.64 (0.56/0.74)	0.86 (0.76/0.98)	0.78 (0.65/0.98)	0.52
	σ	0.01	0.01 (0.02/0.02)	0.02 (0.02/0.03)	0.04 (0.05/0.01)	0.03
DS	μ	0.95	0.40 (0.35/0.49)	0.48 (0.36/0.74)	0.44 (0.32/0.72)	0.31
	σ	0.01	0.04 (0.07/0.08)	0.06 (0.07/0.06)	0.06 (0.07/0.06)	0.04
PTS	μ	0.97	0.31 (0.41/0.26)	0.46 (0.40/0.56)	0.45 (0.39/0.55)	0.33
	σ	0.01	0.02 (0.07/0.05)	0.04 (0.06/0.10)	0.05 (0.06/0.11)	0.03
FBuoy	μ	0.98	0.14 (0.25/0.10)	0.13 (0.12/0.15)	0.19 (0.22/0.18)	0.11
	σ	0.01	0.04 (0.07/0.04)	0.03 (0.02/0.05)	0.04 (0.05/0.05)	0.03

Table 7.1: Average (μ) and standard deviation (σ) values from 7 identical simulations for the binary recognition of Fully Lexical Signs, Depicting Signs, Pointing Signs and Fragment Buoys, for the standard configuration defined in Section 7.2, on the validation set of Dicta-Sign–LSF–v2.
Metrics displayed are frame-wise accuracy and F1-score, as well as unit-wise margin-based F1-score $F1^{\star}_{w}(t_w)$, with margin $t_w = 12$ frames (half a second), and normalized intersection-based F1-score $F1^{\star}_{pr}(0,0)$ (counting positive recognition for units with at least one intersecting frame).

- No weight penalty in the learning loss ($\beta = 0$ in Equation 5.25);
- Samples arranged with a sequence length of 100 frames.

In terms of signer representation:

The chosen signer representation corresponds to configuration #16 in Table 7.4, that is: 3D body and face preprocessed data, along with both OpenPose and Deep Hand estimates as hand features, for a total feature vector size of 424.

7.2.2 Results

The results are summarized in Table 7.1, in which we report average values and standard deviation after seven identical simulations, for the binary recognition of FLSs, DSs, PTSs and FBuoys.

From this table, it appears that the best results are obtained for the recognition of FLSs, with a 64% frame-wise F1-score and a 52% I_{pr}. DSs and PTSs get comparable performance values, while FBuoys are not very well recognized – 14% frame-wise F1-score and 11% I_{pr}. Except for FBuoys, one can note that the recall is usually higher than the precision, which means that there are more false positives than false negatives.

These differences in terms of performance can be explained first by the discrepancy with respect to the number of training instances: as can be seen in Table 4.3, FLSs account for about 75% of the manual units, while this drops to 11% for DSs and for PTSs. Only 589 FBuoy instances are annotated in Dicta-Sign–LSF–v2, that is about 2% of the total number of manual units.

However, other reasons could explain these differences. DSs are a very broad category of units – many

sub-categories can be listed, as detailed in Section 4.2.4.3 – with a lot of inner variability. Also, the role of eye gaze is known to be crucial in DSs, however our signer representations include no gaze information. PTSs are very short, sometimes they last only one or two frames in 25 frames per second (fps) videos. As for FBuoys, they correspond a to a maintained hand shape at the end of a bimanual sign, when it bears a linguistic function, which is not easy to detect (sometimes the hand shape is held for other reasons, and is not annotated as a FBuoy).

Now we have established baseline results, the next sections will focus on evaluating the influence of network parameterization and signer representation on the performance.

7.3 Validation of the network architecture

The focus of this section is on the validation of the Recurrent Neural Network (RNN) architecture presented in Section 6.2. We aim at analyzing the sensitivity of the performance with respect to the network parameters, as well as find the best hyperparameters for training.

7.3.1 Validation setup

Signer representation

In this series of experiments, the signer representation is set the same for all runs and corresponds to configuration #16 in Table 7.4, that is: 3D body and face preprocessed data, along with both OpenPose and Deep Hand estimates as hand features, for a total feature vector size of 424.

Network configurations

The 18 network configurations are all defined as variations of $\mathcal{S}$, the standard configuration, by changing the value of one parameter at a time:

- Network parameters:
  - 1 ($\mathcal{S}$), 2 or 3 BLSTM layers;
  - 10, 50 ($\mathcal{S}$) or 90 units in each LSTM cell;
  - 50, 200 ($\mathcal{S}$) or 350 convolutional filters as a first neural layer, with a kernel width of size 3, or no convolution layer at all (0 filters);
- Training hyperparameters:
  - A batch size of 50, 100, 200 ($\mathcal{S}$) or 400 sequences;
  - A dropout rate of 0, 0.25, 0.5 ($\mathcal{S}$) or 0.75;
  - No weight penalty ($\mathcal{S}$) in the learning loss – $\beta = 0$ in Equation 5.25 –, or a weight penalty of 0.5 or 1;
  - Samples arranged with a sequence length of 50, 100 ($\mathcal{S}$) or 200 frames.

Then, for FLSs, DSs, PTSs and FBuoys, the performance of this standard configuration can be compared to that of the same configuration, with only one setting changed.

7.3.2 Results

Results on the validation set are presented in Tables 7.2 for FLSs and DSs, and 7.3 for PTSs and FBuoys.

In these two tables, each line corresponds to a certain network configuration, either the standard one, or a modified version (one setting changed). For each metric (except accuracy), the best setting is in bold. Not all metrics yield the same conclusion with respect to the best settings, although the agreement is generally strong. In case of disagreement, we have used the integrated unit-wise metric I_{pr} as decision rule, which is highlighted in the two tables.

For instance, let us focus on Depicting Signs (Table 7.2). With a batch size of 200 sequences, the standard configuration yields a I_{pr} of 0.31. With a value of I_{pr} of 0.44, a batch size of 50 sequences is thus preferable, although other metrics like $\mathrm{F1}^{\star}_{w}(t_w = 12)$ and $\mathrm{F1}^{\star}_{pr}(0, 0)$ would indicate a best batch size of 100 sequences.

Some general conclusions can then be noted[1]:

- It appears that annotations with the fewest instances – typically FBuoys – are more sensitive to network parameters, which is expected (see for instance the variation of I_{pr} with respect to the number of convolution filters). Conversely, categories with the most training instances like FLSs show very little sensitivity to the hyperparameters.
- From the results, it seems that the first convolution layer is always beneficial to the performance, especially when few training instances are available.
- In any case, there is no point in stacking more than two BLSTM layers. One is generally enough, which is also beneficial to training time.
- The standard configuration is generally very close to the best performing configuration and thus can be used for further experiments on all considered linguistic descriptors.

[1]When drawing conclusions, one should be careful of the inner variability of the results, which is visible through the standard deviation of the results for the standard configuration, in Table 7.1. Differences are not necessarily significant.

				Frame-wise		Unit-wise		
				Acc	F1 (P/R)	$\mathrm{F1}^{\star}_{w}(t_w = 12)$ (P/R)	$\mathrm{F1}^{\star}_{pr}(0,0)$ (P/R)	I_{pr}
Fully Lexical Signs	BLSTM layers	1	$(\mathcal{S})$	0.83	**0.64** (0.56/0.74)	0.86 (0.76/0.98)	0.78 (0.65/0.98)	0.52
		2		0.88	0.62 (0.58/0.66)	**0.88** (0.79/0.98)	0.81 (0.69/0.98)	**0.58**
		3		0.82	0.61 (0.58/0.65)	**0.88** (0.80/0.97)	**0.83** (0.74/0.95)	0.56
	LSTM units	10		0.88	**0.67** (0.57/0.80)	**0.88** (0.80/0.98)	0.81 (0.69/0.98)	**0.56**
		50	$(\mathcal{S})$	0.83	0.64 (0.56/0.74)	0.86 (0.76/0.98)	0.78 (0.65/0.98)	0.52
		90		0.88	0.62 (0.60/0.64)	0.87 (0.82/0.92)	**0.84** (0.76/0.93)	**0.56**
	Conv. filters	0		0.83	0.60 (0.47/0.83)	0.73 (0.64/0.86)	0.75 (0.60/0.99)	0.47
		50		0.83	**0.64** (0.56/0.75)	0.87 (0.77/1.00)	**0.79** (0.65/0.98)	**0.52**
		200	$(\mathcal{S})$	0.83	**0.64** (0.56/0.74)	0.86 (0.76/0.98)	0.78 (0.65/0.98)	**0.52**
		350		0.84	**0.64** (0.60/0.69)	**0.88** (0.79/1.00)	0.77 (0.63/0.98)	**0.52**
	Batch size	50		0.86	**0.67** (0.66/0.68)	**0.90** (0.83/0.98)	**0.83** (0.74/0.93)	**0.58**
		100		0.88	**0.67** (0.59/0.77)	0.86 (0.76/0.98)	0.78 (0.66/0.96)	0.55
		200	$(\mathcal{S})$	0.83	0.64 (0.56/0.74)	0.86 (0.76/0.98)	0.78 (0.65/0.98)	0.52
		400		0.84	0.64 (0.54/0.80)	0.82 (0.69/0.99)	0.73 (0.57/1.00)	0.46
	Dropout	0		0.87	0.60 (0.55/0.65)	0.85 (0.78/0.92)	0.78 (0.69/0.90)	0.51
		0.25		0.83	0.63 (0.57/0.71)	0.84 (0.73/0.98)	**0.81** (0.69/0.97)	**0.56**
		0.5	$(\mathcal{S})$	0.83	0.64 (0.56/0.74)	**0.86** (0.76/0.98)	0.78 (0.65/0.98)	0.52
		0.75		0.82	**0.65** (0.53/0.84)	0.82 (0.71/0.98)	0.75 (0.61/0.99)	0.49
	Weight balance	0	$(\mathcal{S})$	0.83	0.64 (0.56/0.74)	**0.86** (0.76/0.98)	0.78 (0.65/0.98)	0.52
		0.5		0.82	**0.67** (0.54/0.88)	0.81 (0.72/0.91)	0.84 (0.73/1.00)	**0.57**
		1		0.80	0.66 (0.50/0.97)	0.70 (0.75/0.66)	**0.85** (0.74/1.00)	0.56
	Seq. length	50		0.89	0.65 (0.60/0.71)	**0.88** (0.79/0.99)	0.80 (0.67/0.98)	0.54
		100	$(\mathcal{S})$	0.83	0.64 (0.56/0.74)	0.86 (0.76/0.98)	0.78 (0.65/0.98)	0.52
		200		0.86	**0.66** (0.58/0.77)	0.86 (0.77/0.97)	**0.82** (0.71/0.97)	**0.55**
Depicting Signs	BLSTM layers	1	$(\mathcal{S})$	0.95	0.40 (0.35/0.49)	0.48 (0.36/0.74)	0.44 (0.32/0.72)	0.31
		2		0.97	**0.55** (0.54/0.57)	**0.51** (0.37/0.84)	**0.59** (0.45/0.85)	**0.34**
		3		0.95	0.39 (0.49/0.32)	0.45 (0.45/0.46)	0.49 (0.49/0.48)	0.31
	LSTM units	10		0.98	**0.52** (0.45/0.61)	0.30 (0.20/0.60)	0.32 (0.22/0.60)	0.26
		50	$(\mathcal{S})$	0.95	0.40 (0.35/0.49)	0.48 (0.36/0.74)	0.44 (0.32/0.72)	0.31
		90		0.98	0.41 (0.72/0.29)	**0.50** (0.64/0.41)	**0.49** (0.65/0.39)	**0.36**
	Conv. filters	0		0.97	0.26 (0.56/0.17)	0.43 (0.45/0.41)	0.43 (0.44/0.41)	0.28
		50		0.95	**0.41** (0.37/0.46)	**0.53** (0.42/0.71)	**0.53** (0.42/0.71)	**0.39**
		200	$(\mathcal{S})$	0.95	0.40 (0.35/0.49)	0.48 (0.36/0.74)	0.44 (0.32/0.72)	0.31
		350		0.93	0.31 (0.24/0.44)	0.31 (0.20/0.72)	0.31 (0.20/0.72)	0.20
	Batch size	50		0.97	**0.64** (0.65/0.63)	0.52 (0.43/0.67)	0.54 (0.44/0.68)	**0.44**
		100		0.98	0.54 (0.63/0.47)	**0.49** (0.39/0.65)	**0.62** (0.58/0.67)	0.42
		200	$(\mathcal{S})$	0.95	0.40 (0.35/0.49)	0.48 (0.36/0.74)	0.44 (0.32/0.72)	0.31
		400		0.92	0.28 (0.26/0.31)	0.46 (0.34/0.72)	0.47 (0.35/0.74)	0.30
	Dropout	0		0.98	**0.61** (0.66/0.58)	0.51 (0.39/0.72)	**0.64** (0.54/0.78)	**0.42**
		0.25		0.96	0.49 (0.46/0.53)	**0.56** (0.43/0.80)	0.56 (0.43/0.80)	0.39
		0.5	$(\mathcal{S})$	0.95	0.40 (0.35/0.49)	0.48 (0.36/0.74)	0.44 (0.32/0.72)	0.31
		0.75		0.95	0.21 (0.21/0.21)	0.41 (0.30/0.62)	0.41 (0.30/0.62)	0.16
	Weight balance	0	$(\mathcal{S})$	0.95	**0.40** (0.35/0.49)	0.48 (0.36/0.74)	0.44 (0.32/0.72)	0.31
		0.5		0.95	0.39 (0.34/0.45)	**0.49** (0.37/0.71)	**0.49** (0.37/0.71)	**0.34**
		1		0.91	0.39 (0.26/0.82)	0.36 (0.23/0.81)	0.39 (0.25/0.81)	0.28
	Seq. length	50		0.98	**0.54** (0.61/0.49)	0.40 (0.31/0.57)	0.40 (0.31/0.55)	**0.33**
		100	$(\mathcal{S})$	0.95	0.40 (0.35/0.49)	**0.48** (0.36/0.74)	**0.44** (0.32/0.72)	0.31
		200		0.91	0.20 (0.20/0.20)	0.41 (0.31/0.62)	0.40 (0.29/0.62)	0.25

Table 7.2: Best validation performance on Dicta-Sign–LSF–v2, with different network and training settings, for FLSs and DSs. Each line corresponds to a particular configuration of the network, either the standard configuration ($\mathcal{S}$), or with only one setting changed. Bold values correspond to the best value for each setting category. In the end, I_{pr} is used to decide the best settings.

				Frame-wise		Unit-wise		
				Acc	F1 (P/R)	$\mathrm{F1}^{\star}_{w}(t_w = 12)$ (P/R)	$\mathrm{F1}^{\star}_{pr}(0,0)$ (P/R)	I_{pr}
Pointing Signs	BLSTM layers	1	$(\mathcal{S})$	0.97	**0.31** (0.41/0.26)	**0.46** (0.40/0.56)	**0.45** (0.39/0.55)	**0.33**
		2		0.97	0.22 (0.25/0.20)	0.34 (0.27/0.45)	0.25 (0.18/0.42)	0.19
		3		0.97	0.12 (0.41/0.07)	0.34 (0.39/0.31)	0.33 (0.38/0.30)	0.23
	LSTM units	10		0.97	0.16 (0.14/0.18)	0.43 (0.36/0.54)	0.27 (0.21/0.39)	0.19
		50	$(\mathcal{S})$	0.97	**0.31** (0.41/0.26)	**0.46** (0.40/0.56)	**0.45** (0.39/0.55)	**0.33**
		90		0.97	0.18 (0.34/0.12)	0.33 (0.35/0.30)	0.33 (0.39/0.28)	0.25
	Conv. filters	0		0.98	0.04 (0.17/0.02)	0.15 (0.27/0.11)	0.07 (0.12/0.05)	0.05
		50		0.95	0.23 (0.18/0.31)	0.32 (0.21/0.65)	0.26 (0.17/0.62)	0.20
		200	$(\mathcal{S})$	0.97	**0.31** (0.41/0.26)	**0.46** (0.40/0.56)	**0.45** (0.39/0.55)	**0.33**
		350		0.97	0.30 (0.31/0.30)	**0.46** (0.37/0.60)	**0.45** (0.35/0.60)	0.32
	Batch size	50		0.97	0.10 (0.14/0.07)	0.18 (0.19/0.17)	0.17 (0.18/0.16)	0.13
		100		0.98	0.03 (0.19/0.02)	0.18 (0.46/0.11)	0.12 (0.31/0.08)	0.08
		200	$(\mathcal{S})$	0.97	**0.31** (0.41/0.26)	**0.46** (0.40/0.56)	**0.45** (0.39/0.55)	**0.33**
		400		0.95	0.09 (0.18/0.06)	0.35 (0.30/0.41)	0.25 (0.20/0.33)	0.15
	Dropout	0		0.97	0.21 (0.18/0.24)	0.38 (0.30/0.51)	0.34 (0.26/0.49)	0.24
		0.25		0.97	0.29 (0.33/0.26)	0.45 (0.35/0.64)	**0.45** (0.35/0.64)	**0.34**
		0.5	$(\mathcal{S})$	0.97	**0.31** (0.41/0.26)	0.46 (0.40/0.56)	**0.45** (0.39/0.55)	0.33
		0.75		0.96	**0.31** (0.26/0.39)	**0.50** (0.37/0.77)	**0.45** (0.32/0.77)	0.32
	Weight balance	0	$(\mathcal{S})$	0.97	**0.31** (0.41/0.26)	**0.46** (0.40/0.56)	**0.45** (0.39/0.55)	**0.33**
		0.5		0.95	0.26 (0.19/0.40)	0.40 (0.27/0.77)	0.35 (0.24/0.68)	0.27
		1		0.95	0.27 (0.21/0.39)	0.43 (0.31/0.73)	0.41 (0.28/0.73)	0.29
	Seq. length	50		0.98	0.02 (0.10/0.01)	0.12 (0.49/0.07)	0.07 (0.25/0.04)	0.05
		100	$(\mathcal{S})$	0.97	**0.31** (0.41/0.26)	**0.46** (0.40/0.56)	**0.45** (0.39/0.55)	**0.33**
		200		0.96	0.09 (0.31/0.05)	0.50 (0.55/0.46)	0.38 (0.43/0.34)	0.24
Fragment Buoys	BLSTM layers	1	$(\mathcal{S})$	0.98	**0.14** (0.25/0.10)	**0.13** (0.12/0.15)	**0.19** (0.22/0.18)	**0.11**
		2		0.98	0.06 (0.22/0.03)	0.06 (0.12/0.04)	0.06 (0.15/0.04)	0.04
		3		0.98	0.10 (0.24/0.06)	0.06 (0.07/0.06)	0.12 (0.23/0.08)	0.06
	LSTM units	10		0.97	**0.19** (0.20/0.17)	**0.15** (0.10/0.27)	**0.21** (0.17/0.31)	**0.13**
		50	$(\mathcal{S})$	0.98	0.14 (0.25/0.10)	0.13 (0.12/0.15)	0.19 (0.22/0.18)	0.11
		90		0.98	0.09 (0.22/0.06)	0.07 (0.08/0.06)	0.13 (0.23/0.09)	0.08
	Conv. filters	0		0.98	0.01 (0.30/0.01)	0.03 (0.11/0.02)	0.01 (0.05/0.01)	0.01
		50		0.98	0.11 (0.29/0.07)	0.11 (0.12/0.10)	0.15 (0.20/0.12)	0.09
		200	$(\mathcal{S})$	0.98	0.14 (0.25/0.10)	0.13 (0.12/0.15)	0.19 (0.22/0.18)	0.11
		350		0.97	**0.21** (0.23/0.20)	**0.15** (0.10/0.27)	**0.22** (0.17/0.30)	**0.13**
	Batch size	50		0.98	0.01 (0.18/0.01)	0.03 (0.09/0.02)	0.03 (0.12/0.02)	0.02
		100		0.97	0.10 (0.16/0.07)	0.12 (0.11/0.13)	0.17 (0.21/0.15)	0.10
		200	$(\mathcal{S})$	0.98	0.14 (0.25/0.10)	**0.13** (0.12/0.15)	**0.19** (0.22/0.18)	**0.11**
		400		0.98	**0.17** (0.28/0.13)	0.10 (0.10/0.10)	0.15 (0.19/0.12)	0.09
	Dropout	0		0.98	0.08 (0.19/0.05)	0.11 (0.11/0.11)	0.16 (0.18/0.15)	0.09
		0.25		0.98	**0.20** (0.39/0.13)	**0.15** (0.16/0.14)	0.20 (0.32/0.15)	0.12
		0.5	$(\mathcal{S})$	0.98	0.14 (0.25/0.10)	0.13 (0.12/0.15)	0.19 (0.22/0.18)	0.11
		0.75		0.97	**0.20** (0.19/0.20)	**0.15** (0.10/0.31)	**0.23** (0.17/0.35)	**0.14**
	Weight balance	0	$(\mathcal{S})$	0.98	0.14 (0.25/0.10)	**0.13** (0.12/0.15)	**0.19** (0.22/0.18)	**0.11**
		0.5		0.92	**0.18** (0.11/0.43)	0.09 (0.05/0.44)	0.14 (0.08/0.53)	0.08
		1		0.71	0.10 (0.05/0.77)	0.05 (0.03/0.33)	0.10 (0.06/0.84)	0.06
	Seq. length	50		0.98	0.02 (0.20/0.01)	0.06 (0.17/0.04)	0.04 (0.14/0.02)	0.03
		100	$(\mathcal{S})$	0.98	**0.14** (0.25/0.10)	**0.13** (0.12/0.15)	**0.19** (0.22/0.18)	**0.11**
		200		0.98	0.07 (0.33/0.04)	0.06 (0.11/0.04)	0.08 (0.30/0.05)	0.05

Table 7.3: Best validation performance on Dicta-Sign–LSF–v2, with different network and training settings, for PTSs and FBuoys. Each line corresponds to a particular configuration of the network, either the standard configuration ($\mathcal{S}$), or with only one setting changed. Bold values correspond to the best value for each setting category. In the end, I_{pr} is used to decide the best settings.

7.4 Validation of the signer representation

Besides validating the network architecture and parametrization, different options in terms of signer representation must be evaluated, which is the purpose of the current section.

7.4.1 Validation setup

In order to adopt the best possible signer representation, we have defined 16 combinations of the feature vectors presented in Section 6.1.4. The detail of these configurations is given in Table 7.4, in which we also indicate the final representation vector size (for each frame), ranging from 218 for combination 5 to 494 for combination 9. Body and face data are either 2D or 3D, raw or made of preprocessed features, while hand data are made of OpenPose estimates, Deep Hand predictions or both.

Combination #16 corresponds to the signer representation in the standard configuration ($\mathcal{S}$, Section 7.2.1).

Then, for FLSs, DSs, PTSs and FBuoys, the performance of each combination was computed, with the network parameters and training settings defined as those of the standard configuration of Section 7.2.1. Results are presented below.

7.4.2 Results

Tables 7.5 and 7.6 present the model performance metrics on the validation set, for each of the 16 combinations and each of the four different annotation types.

In each table, one line corresponds to a particular combination, *i.e.* a certain signer representation. As for Section 7.3, for each metric (except accuracy), the best setting is in bold. Not all metrics yield the same conclusion with respect to the best settings: in case of disagreement, we have used the integrated unit-wise metric I_{pr} as decision rule, which is highlighted in the two tables.

For instance, for the binary recognition of Fully Lexical Signs, the best combination – with an I_{pr} of 0.60 – is #15: 3D features, with hand shapes from the Deep Hand model. For Depicting Signs, best performance is reached by 2D features, with both OpenPose and hand shape data. Pointing Signs are better recognized with 3D features and both OpenPose and hand shape data. Last, Fragment Buoys should be recognized with 2D or 3D features, with OpenPose data alone.

A few general insights can be drawn from these results:

- From the results, it is clear that using preprocessed data instead of raw values is always beneficial to the model performance, whatever the linguistic category. For linguistic annotations with few training instances like PTSs or FBuoys, the model is not even able to converge with raw data.
- In the end, it appears that 3D estimates do not always improve the model performance, compared to 2D data. FLSs and FBuoys are better recognized when using 3D, while DSs and PTSs should be predicted using 2D data.

  However, this surprising result might stem from the limited quality of the 3D estimates that we used. True 3D data (instead of estimates trained on motion capture recordings) might indeed be more reliable thus beneficial in any case.
- In terms of hand representation, it appears that the Deep Hand model is beneficial when recognizing FLSs, while OpenPose estimates alone correspond to the best choice – or very close to

#	Configuration				Corresponding feature vectors and size (Section 6.1.4)								Total size
	Body and face		Hands		x^{b}_{raw2D}	x^{b}_{raw3D}	x^{fh}_{raw2D}	x^{fh}_{raw3D}	x^{hand}_{raw2D}	x^{hand}_{shapes}	x^{bfh}_{feat2D}	x^{bfh}_{feat3D}	
			OP	HS	28	42	140	204	126	122	96	176	
1	2D	Raw			✓		✓						168
2			✓		✓		✓		✓				294
3				✓	✓		✓			✓			290
4			✓	✓	✓		✓		✓	✓			416
5		Features									✓		96
6			✓						✓		✓		222
7				✓						✓	✓		218
8			✓	✓					✓	✓	✓		344
9	3D	Raw				✓		✓					246
10			✓			✓		✓	✓				372
11				✓		✓		✓		✓			368
12			✓	✓		✓		✓	✓	✓			494
13		Features										✓	176
14			✓						✓			✓	302
15				✓						✓		✓	298
16 ($\mathcal{S}$)			✓	✓					✓	✓		✓	424

Table 7.4: Detail of the 16 signer representations that are compared in Section 7.4. Configuration #16 corresponds to the standard configuration ($\mathcal{S}$).

it – for the other linguistic categories. The fact that Deep Hand alone performs quite well for the recognition of FLSs and not for the other types of units could be explained that FLSs use a large variety of hand shapes, whereas other units like DSs use few hand shapes, but are rather very determined by the hand orientation, that is not captured by Deep Hand. In other words, it is likely that DSs give a more balanced importance to all hand parameters than FLSs.

		Body and face	Hands OP	Hands HS		Frame-wise Acc	Frame-wise F1 (P/R)	Unit-wise $F1^{\star}_{w}(t_w = 12)$ (P/R)	Unit-wise $F1^{\star}_{pr}(0,0)$ (P/R)	I_{pr}
Fully Lexical Signs	2D	Raw				0.80	0.58 (0.51/0.68)	0.57 (0.75/0.46)	0.75 (0.76/0.73)	0.42
			✓			0.79	0.16 (0.44/0.10)	0.45 (0.87/0.30)	0.34 (0.68/0.23)	0.19
				✓		0.82	0.55 (0.56/0.55)	0.83 (0.85/0.80)	0.75 (0.76/0.74)	0.45
			✓	✓		0.80	0.19 (0.52/0.12)	0.60 (0.86/0.46)	0.42 (0.61/0.32)	0.26
		Features				0.86	0.68 (0.65/0.71)	0.88 (0.81/0.97)	0.78 (0.66/0.94)	0.56
			✓			0.85	0.66 (0.61/0.73)	0.85 (0.75/0.99)	0.79 (0.67/0.98)	0.57
				✓		0.84	0.63 (0.60/0.66)	0.87 (0.78/0.99)	0.75 (0.61/0.96)	0.49
			✓	✓		0.85	**0.69** (0.60/0.82)	0.87 (0.78/0.98)	0.81 (0.69/0.98)	0.59
	3D	Raw				0.81	0.42 (0.56/0.34)	0.68 (0.81/0.59)	0.57 (0.64/0.52)	0.32
			✓			0.82	0.47 (0.59/0.40)	0.82 (0.82/0.82)	0.64 (0.65/0.64)	0.40
				✓		0.83	0.45 (0.64/0.34)	0.73 (0.87/0.62)	0.62 (0.76/0.52)	0.38
			✓	✓		0.80	0.37 (0.51/0.29)	0.73 (0.83/0.66)	0.58 (0.66/0.51)	0.34
		Features				0.83	0.65 (0.55/0.78)	0.84 (0.73/0.99)	0.73 (0.59/0.97)	0.51
			✓			0.87	**0.69** (0.66/0.73)	0.89 (0.81/0.98)	0.80 (0.69/0.95)	0.57
				✓		0.86	**0.69** (0.64/0.75)	**0.90** (0.82/0.99)	**0.83** (0.73/0.97)	**0.60**
			✓	✓	($\mathcal{S}$)	0.83	0.64 (0.56/0.74)	0.86 (0.76/0.98)	0.78 (0.65/0.98)	0.52
Depicting Signs	2D	Raw				0.92	0.24 (0.18/0.36)	0.20 (0.14/0.34)	0.22 (0.16/0.37)	0.15
			✓			0.94	0.30 (0.24/0.40)	0.35 (0.24/0.62)	0.34 (0.23/0.59)	0.21
				✓		0.92	0.10 (0.08/0.13)	0.28 (0.19/0.56)	0.28 (0.18/0.55)	0.16
			✓	✓		0.93	0.36 (0.27/0.53)	0.40 (0.27/0.77)	0.42 (0.28/0.81)	0.27
		Features				0.95	0.41 (0.37/0.46)	0.33 (0.25/0.48)	0.32 (0.24/0.47)	0.25
			✓			0.97	0.55 (0.54/0.56)	**0.67** (0.58/0.78)	**0.68** (0.60/0.78)	0.44
				✓		0.95	0.43 (0.37/0.52)	0.40 (0.29/0.64)	0.37 (0.26/0.62)	0.26
			✓	✓		0.97	**0.59** (0.53/0.66)	0.61 (0.50/0.81)	0.64 (0.53/0.81)	**0.46**
	3D	Raw				0.97	0.24 (0.68/0.14)	0.32 (0.66/0.21)	0.32 (0.65/0.21)	0.23
			✓			0.90	0.28 (0.18/0.60)	0.39 (0.26/0.75)	0.41 (0.27/0.80)	0.24
				✓		0.94	0.14 (0.13/0.16)	0.31 (0.25/0.43)	0.24 (0.18/0.34)	0.15
			✓	✓		0.88	0.25 (0.16/0.62)	0.32 (0.20/0.80)	0.25 (0.15/0.81)	0.17
		Features				0.93	0.36 (0.27/0.53)	0.25 (0.16/0.55)	0.22 (0.14/0.50)	0.17
			✓			0.97	0.50 (0.52/0.49)	0.52 (0.44/0.63)	0.55 (0.46/0.70)	0.37
				✓		0.92	0.34 (0.24/0.57)	0.29 (0.18/0.72)	0.25 (0.16/0.60)	0.17
			✓	✓	($\mathcal{S}$)	0.95	0.40 (0.35/0.49)	0.48 (0.36/0.74)	0.44 (0.32/0.72)	0.31

Table 7.5: Performance assessment on the validation set of Dicta-Sign–LSF–v2, for different signer representations, applied to the recognition of FLSs and DSs. Each line corresponds to a particular signer representation, see Table 7.4. Bold values correspond to the best value for each setting category. In the end, I_{pr} is used to decide the best representation.

	Body and face		Hands			Frame-wise		Unit-wise		
			OP	HS		Acc	F1 (P/R)	$\mathrm{F1}^{\star}_{w}(t_w = 12)$ (P/R)	$\mathrm{F1}^{\star}_{pr}(0,0)$ (P/R)	I_{pr}
Pointing Signs	2D	Raw				——		Not converged	——	
			✓			0.97	0.23 (0.25/0.21)	0.28 (0.24/0.34)	0.24 (0.19/0.33)	0.21
				✓		——		Not converged	——	
			✓	✓		0.97	0.14 (0.22/0.10)	0.22 (0.26/0.18)	0.22 (0.26/0.18)	0.18
		Features				0.97	0.14 (0.27/0.10)	0.31 (0.35/0.29)	0.21 (0.22/0.20)	0.12
			✓			0.97	0.30 (0.40/0.24)	0.46 (0.38/0.59)	**0.45** (0.37/0.59)	0.29
				✓		0.97	0.14 (0.17/0.12)	**0.51** (0.41/0.67)	0.28 (0.21/0.41)	0.17
			✓	✓		0.97	**0.32** (0.41/0.27)	0.48 (0.42/0.58)	0.43 (0.35/0.58)	0.32
	3D	Raw				——		Not converged	——	
			✓			0.97	0.04 (0.10/0.03)	0.17 (0.24/0.13)	0.17 (0.24/0.13)	0.12
				✓		——		Not converged	——	
			✓	✓		0.96	0.15 (0.16/0.13)	0.30 (0.32/0.28)	0.22 (0.22/0.21)	0.13
		Features				0.96	0.11 (0.13/0.10)	0.37 (0.28/0.54)	0.24 (0.19/0.35)	0.12
			✓			0.96	0.27 (0.26/0.28)	0.48 (0.35/0.72)	0.44 (0.32/0.67)	0.31
				✓		0.97	0.09 (0.13/0.06)	0.43 (0.42/0.43)	0.24 (0.20/0.28)	0.12
			✓	✓	$(\mathcal{S})$	0.97	0.31 (0.41/0.26)	0.46 (0.40/0.56)	**0.45** (0.39/0.55)	**0.33**
Fragment Buoys	2D	Raw	⋆	⋆		——		Not converged	——	
		Features				0.97	0.24 (0.26/0.23)	0.16 (0.12/0.24)	0.24 (0.21/0.29)	0.15
			✓			0.98	**0.32** (0.43/0.26)	0.17 (0.15/0.20)	0.25 (0.26/0.23)	**0.16**
				✓		0.98	0.23 (0.40/0.16)	0.14 (0.16/0.13)	0.22 (0.32/0.17)	0.15
			✓	✓		0.96	0.30 (0.24/0.39)	**0.19** (0.12/0.40)	0.24 (0.16/0.46)	0.15
	3D	Raw	⋆	⋆		——		Not converged	——	
		Features				0.98	0.13 (0.31/0.08)	0.15 (0.20/0.12)	0.20 (0.35/0.14)	0.12
			✓			0.98	0.31 (0.43/0.25)	**0.19** (0.15/0.26)	**0.26** (0.24/0.30)	**0.16**
				✓		0.98	0.12 (0.35/0.07)	0.16 (0.25/0.12)	0.21 (0.41/0.14)	0.13
			✓	✓	$(\mathcal{S})$	0.98	0.14 (0.25/0.10)	0.13 (0.12/0.15)	0.19 (0.22/0.18)	0.11

Table 7.6: Performance assessment on the validation set of Dicta-Sign–LSF–v2, for different signer representations, applied to the recognition of PTSs and FBuoys. Each line corresponds to a particular signer representation, see Table 7.4. Bold values correspond to the best value for each setting category. In the end, I_{pr} is used to decide the best representation.

Conclusion

In this chapter, we have quantitatively assessed the performance of the recognition model presented in the previous chapter. Using adapted metrics, we have sought the appropriate values in terms of training hyperparameters and learning settings. Because we have only varied one parameter at a time in terms of network configuration, there remains room for performance improvement.

Then, we have used this model to look for the best signer representation, based on many options developed earlier on. In the end, this analysis provides interesting insights on the influence of signer representation on the recognition capabilities of a model such as the one we have developed. These first results suggest that hand representation is crucial, and should be built in conjunction with the linguistic descriptors. Furthermore, preprocessed features are very effective and should be analyzed further.

More generally, this chapter has proven that very interesting Continuous Sign Language Recognition (CSLR) performance results could be met by the compact model and generalizable signer representation that we developed earlier on.

In the next chapter, we will use the trained models and analyze their results on the test set, mostly in a more qualitative way.

Chapter 8

Recognition results and qualitative analysis

After analyzing the framework and signer representation options in Chapter 7, we intend to develop in this chapter a more qualitative analysis.

We have decided to focus on the binary recognition of four manual unit types – Fully Lexical Signs (FLSs), Depicting Signs (DSs), Pointing Signs (PTSs) and Fragment Buoys (FBuoys) – that show a sufficient number of instances for the prediction model to converge during training.

First, we will analyze the impact of signer-independence and task-independence on the prediction results (Section 8.1), then we will present a certain number of test sequences, along with prediction results.

8.1 Signer-independence, task-independence

8.1.1 Setup

In this first set of experiments, we set the network parameters and training settings as those of the standard configuration of Section 7.3.1. As for Chapter 7, we focus on the binary recognition of four manual unit types: FLSs, DSs, PTSs and FBuoys. Also, the metrics used for performance assessment are those defined in Section 7.1.2.

Because we are considering both the problem of signer-independence and that of task-independence, four cases are to be analyzed. Task 9 is excluded from all of them, as it corresponds to a very different task from the others (more detail in Section 4.2).

Signer-dependent and task-dependent (SD-TD)

In this case, we randomly pick 60% of the videos for training, 20% for validation and 20% for testing. Some signers and tasks are then shared across the three sets.

Signer-independent and task-dependent (SI-TD)

In this case, we randomly pick 10 signers for training, 3 signers for validation and 3 signers for testing. All tasks are then shared across the three sets.

Signer-dependent and task-independent (SD-TI)

In this case, we randomly pick 5 tasks for training, 2 tasks for validation and 1 task for testing. All signers are then shared across the three sets.

Signer-independent and task-independent (SI-TI)

In this last case, we randomly pick:

- 8 signers for training, 4 signers for validation and 4 signers for testing;
- 3 tasks for training, 3 tasks for validation and 2 tasks for testing.

This roughly[1] corresponds to a 55%-27%-18% training-validation-testing split in terms of video count.

Notably in this setting, a fraction of the videos has to be left out – videos that correspond to signers in the training set, and tasks in the other sets, *etc.* In the end, it is thus expected that the amount of training data is more likely to be a limiting factor than for the three previously described configurations.

8.1.2 Results

The four different settings correspond to different sets of videos. For the results to be as comparable as possible, we have actually repeated seven times the random video split and training of each model. In the end, the presented results are actually averaged out values from seven repeats.

These results are summarized in Table 8.1, using the same performance metrics as in Chapter 7.

Surprisingly, it appears that results for the configurations SD-TD, SI-TD and SD-TI perform relatively close, which supports the idea that the proposed signer representation and learning framework are good at generalizing to unseen signers and unseen tasks. The fact that performance is much lower in the SI-TI configuration thus suggests that the amount of training data is indeed a limiting factor in our case.

[1] Some particular tasks are missing for some signers.

			Frame-wise		Unit-wise		
	SI	TI	Acc	F1 (P/R)	$\mathrm{F1}^{\star}_{w}(t_w = 12)$ (P/R)	$\mathrm{F1}^{\star}_{pr}(0,0)$ (P/R)	I_{pr}
FLS			0.78	0.57 (0.48/0.71)	0.77 (0.69/0.88)	0.72 (0.61/0.90)	0.47
	✓		0.78	0.54 (0.46/0.67)	0.79 (0.72/0.88)	0.72 (0.62/0.86)	0.46
		✓	0.79	0.56 (0.52/0.62)	0.83 (0.77/0.91)	0.73 (0.65/0.85)	0.48
	✓	✓	0.65	0.45 (0.34/0.72)	0.67 (0.60/0.80)	0.65 (0.53/0.89)	0.39
DS			0.94	0.26 (0.41/0.20)	0.30 (0.35/0.28)	0.31 (0.39/0.28)	0.20
	✓		0.92	0.30 (0.43/0.26)	0.33 (0.39/0.30)	0.35 (0.43/0.33)	0.22
		✓	0.92	0.24 (0.38/0.21)	0.33 (0.37/0.38)	0.32 (0.37/0.35)	0.19
	✓	✓	0.92	0.11 (0.22/0.08)	0.19 (0.23/0.18)	0.18 (0.25/0.16)	0.11
PTS			0.96	0.20 (0.19/0.25)	0.35 (0.28/0.52)	0.26 (0.21/0.41)	0.18
	✓		0.97	0.15 (0.28/0.12)	0.30 (0.37/0.30)	0.23 (0.29/0.23)	0.15
		✓	0.96	0.20 (0.26/0.19)	0.40 (0.38/0.42)	0.30 (0.29/0.33)	0.20
	✓	✓	0.94	0.07 (0.09/0.11)	0.20 (0.21/0.31)	0.11 (0.11/0.20)	0.07
FBuoy			0.97	0.19 (0.22/0.20)	0.12 (0.11/0.26)	0.21 (0.18/0.36)	0.12
	✓		0.94	0.10 (0.20/0.07)	0.11 (0.15/0.19)	0.11 (0.15/0.14)	0.08
		✓	0.93	0.07 (0.07/0.07)	0.06 (0.05/0.09)	0.08 (0.06/0.10)	0.05
	✓	✓	0.98	0.01 (0.01/0.01)	0.02 (0.01/0.09)	0.02 (0.01/0.09)	0.01

Table 8.1: Performance assessment with respect to signer-independence (SI) and task-independence (TI) on the test set of Dicta-Sign–LSF–v2, for the binary recognition of four linguistic descriptors (FLSs, DSs, PTSs and FBuoys).

8.2 Example-based analysis

Although performance metrics provide interesting insights on the results of the proposed model, a more qualitative analysis is needed. In this section, we analyze the prediction results of the model developed in Chapter 6, on six test sequences of Dicta-Sign–LSF–v2. The network parameters and signer representation are decided from the optimization tables from Chapter 6.2.3. The chosen setup is signer-independent and task-dependent (SI-TD). The test signers are then unknown both from the training and validation sets. These results complement preliminary analyses focused on DSs and developed in [Belissen et al., 2020b].

In this analysis, we have trained four binary descriptors, corresponding to FLSs, DSs, PTSs and FBuoys. In the following figures (8.1 to 8.6), we show, from top to bottom: a few key thumbnails, a proposed English translation, expert annotations for FLSs and Partially Lexical Signs (PLSs) – each on three tracks, corresponding to right-handed, two-handed or left-handed units – and model predictions. Because all descriptors are binary, a positive prediction is equivalent to a probability greater than 0.5.

For each sequence, the quantitative performance is summarized in Table 8.2. This includes frame-wise – Accuracy, F1-score – and unit-wise metrics – margin-based F1-score $\mathrm{F1}^{\star}_{w}(t_w)$, with margin $t_w = 12$ frames (half a second), normalized intersection-based F1-score $\mathrm{F1}^{\star}_{pr}(\bar{t}_p, \bar{t}_r)$, with $\bar{t}_p = 0, \bar{t}_r = 0$ (counting positive recognition for units with at least one intersecting frame), and associated integral value I_{pr}.

It is important to note that the values in Table 8.2 are not really comparable to those of the tables in Chapter 7. Indeed, the videos of the current chapter belong to the test set and are then different than the validation videos from Chapter 7. Furthermore, in Chapter 7 the performance was computed on the whole length of the videos. Because Dicta-Sign–LSF–v2 is a dialogue corpus with continuous recording for each signer, about half of each video has almost no annotations, which in the end results in a higher rate of false positives in terms of predictions, thus in a lower precision.

Video `S3_T1_B0`, frames 7340-7375 (Figure 8.1)

We have translated this very short sequence as *"From my experience, here is what I would advise."*. Its syntactic structure is quite simple, making use of two FLSs and one PTS, according to the annotation, in a sequential way. The PTS is a self reference, playing the role of a subject. No iconic structure is observed in this sequence.

The model predictions appear to be very good for FLSs, with the two signs segmented close to what is annotated. This is confirmed in Table 8.2, with $I_{pr} = 0.89$. The probabilities for DS and FBuoy remain close to zero, which is consistent with the annotation. This can only be evaluated through frame-wise accuracy (equal to 100%), since all other metrics are undefined due to the absence of positive predictions and annotations.

As regards PTSs, two units are detected, although only one is annotated. However, looking at thumbnail 6, it seems that a very short – only one frame – but real PTS is realized (pointing to the first person), very similar to the one that was annotated. One should thus be aware that annotators sometimes miss some units, especially very brief ones. In the end, the I_{pr} for PTSs is only 0.44 on this simple sequence.

Figure 8.1: LSF sequence from Dicta-Sign–LSF–v2 (video reference: `S3_T1_B0`, see Chapter 4). Expert annotations for right-handed (RH), two-handed (2H) and left-handed (LH) Fully Lexical Signs (FLSs) and Partially Lexical Signs (PLSs) are given.

Video S7_T2_A10, frames 660-790 (Figure 8.2)

This is a longer and much more complex sequence with all four types of annotations – ten FLSs, two DSs, three PTSs and one FBuoy. We propose the following translation: *"For you to decide between those two [touristic] options, I will present them one after the other, then we will also discuss prices."*.

This sequence makes extensive use of space at the syntactic level. Indeed, the lexical sign HESITATE is iconically reactivated, with one hand corresponding to an option A and the other hand to another option B. Using pointing signs and a visible tilt in the upper body, as well as localized signs like EXPLAIN, the two options are sequentially referred to in a very spatial and visual way.

FLSs are detected quite correctly, with an I_{pr} of 0.60. Two pointing signs are detected, while one is missed. The two successive DSs are correctly detected, even though they are not segmented like the annotations. In the end of the sequence – and to a lesser extent the beginning – DSs are predicted by the model although they are not annotated. However, they do include a form of iconicity – as mentioned earlier, it is spatial iconicity used at the syntactic level.

The unique FBuoy is not detected, resulting in $I_{pr} = 0$.

Video S7_T2_A10, frames 885-990 (Figure 8.3)

This sequence is rather sequential and includes an illustrative structure around frame 920, with the left hand of the lexical sign PARIS iconically reactivated into a FBuoy, while the right hand performs a DS-Size&Shape (DS-SS). We propose the simple translation *"You will need some time to explore Paris!"*.

All FLSs are detected correctly, but two *false positives* are observed in the vicinity of the illustrative structure. The unique PTS is perfectly recognized. The DS unit is very well detected too, while the simultaneous FBuoy is detected but much shorter than it is annotated.

Interestingly, the FLS VISIT is also detected as a DS. This makes some sense as it is produced in quite an iconic way, in a form of Transfer of Persons (T-P), emphasized by the gaze moving away from the addressee and the crinkled eyes.

Video S7_T2_A10, frames 1710-1820 (Figure 8.4)

This sequence, translated as *"I advise you to climb up the Eiffel Tower, you will then get a very nice panoramic view."*, includes six FLSs, one DS, one PTS and one FBuoy. A form of T-P can be observed – thumbnails 3 and 9-10 – although Ts-P were not annotated in Dicta-Sign–LSF–v2

Most FLSs are accurately detected, although the last one is detected as two separate units and a false positive can be seen during the illustrative structure. The unique DS is detected with a temporal shift. Simultaneously, the FBuoy is also detected, but much shorter than the annotation. The FLS VISIT is detected as a DS, probably for the same reason as explained in the previous sequence. The unique annotated PTS is missed by the prediction model – though one will notice that the recognition probability peaks at the right time, with a maximum value of about 0.3.

Video S7_T2_A10, frames 3398-3485 (Figure 8.5)

Translated as *"At the very center of this area, there is a large building surrounded by restaurants."*, this sequence is particularly interesting in that it delivers the message in a straightforward spatial way, fully exploiting the signing space. Only two FLSs are annotated, while four DS units, two PTSs

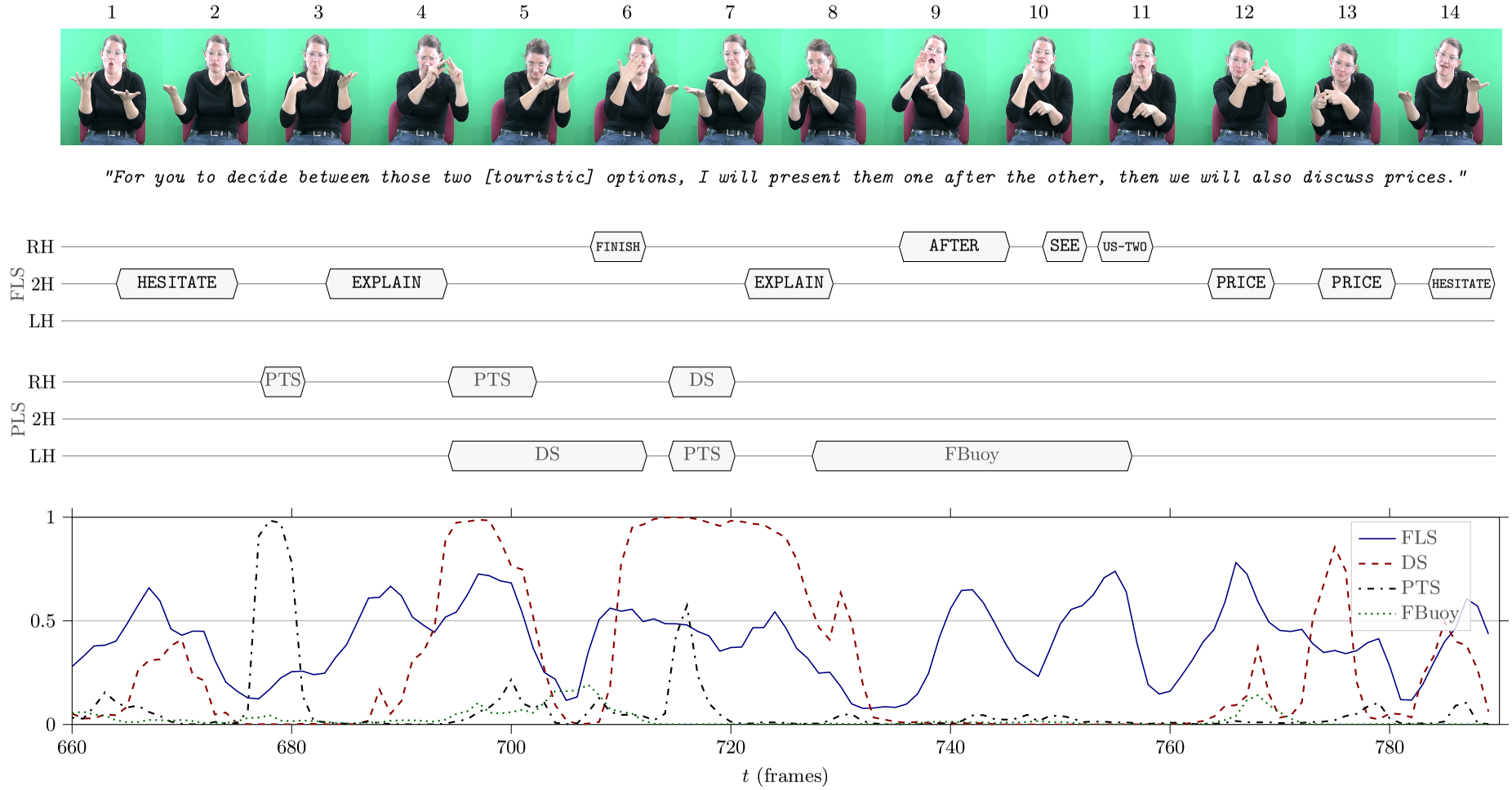


Figure 8.2: LSF sequence from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`, see Chapter 4). Expert annotations for right-handed (RH), two-handed (2H) and left-handed (LH) Fully Lexical Signs (FLSs) and Partially Lexical Signs (PLSs) are given.

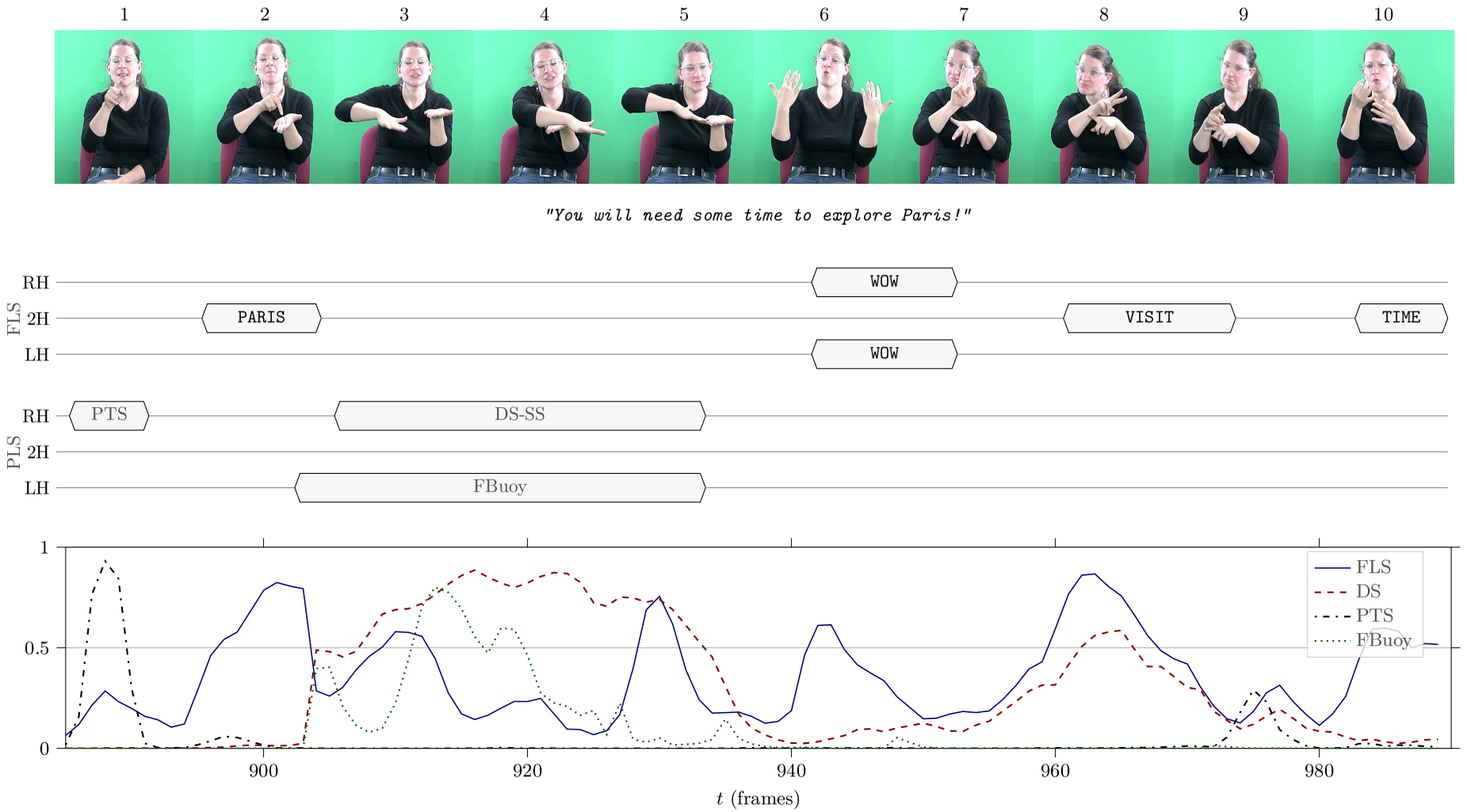


Figure 8.3: LSF sequence from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`, see Chapter 4). Expert annotations for right-handed (RH), two-handed (2H) and left-handed (LH) Fully Lexical Signs (FLSs) and Partially Lexical Signs (PLSs) are given.

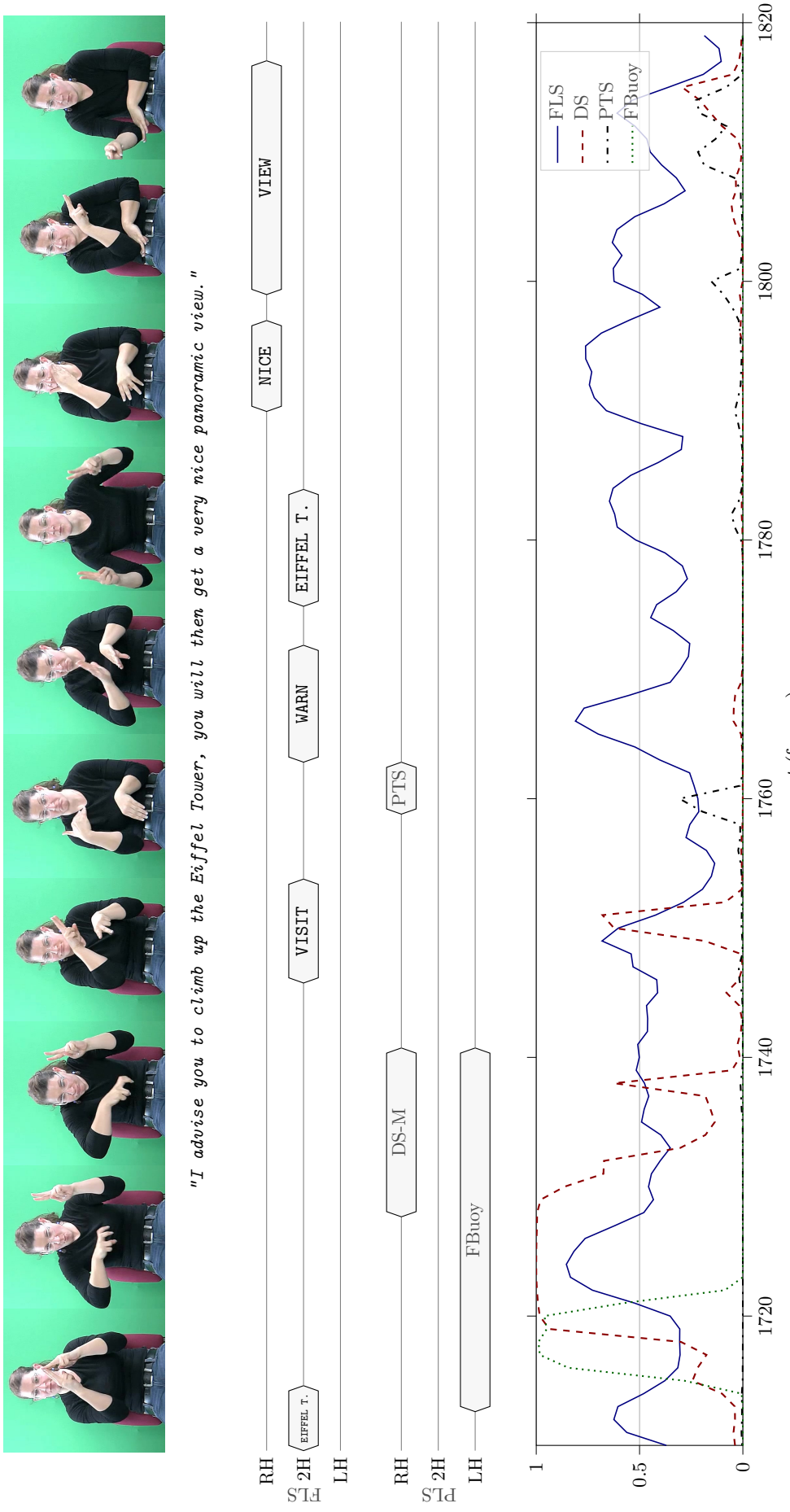


Figure 8.4: LSF sequence from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`, see Chapter 4). Expert annotations for right-handed (RH), two-handed (2H) and left-handed (LH) Fully Lexical Signs (FLSs) and Partially Lexical Signs (PLSs) are given.

and one long FBuoy are noted. An area is first introduced, then a large building is introduced in the middle and restaurants are placed around it using a proform.

Predictions for FLSs are not very satisfying, with an I_{pr} of 0.30, due to many false positives. However, it seems that some of these could actually have been annotated, in some way, as FLSs. For instance, the first false positive (around frame 3400) strongly resembles the lexical sign `AREA`, which presents an inner degenerated iconicity. Similarly, around frame 3435 the DS for building very much looks like the lexical sign `BUILDING`, although it is performed in a very iconic way, hence the DS annotation. The three DS units are well detected, as well as the unique FBuoy, although for a shorter time than the annotation. The two PTSs are accurately detected, and a false PTS positive can be seen around frame 3445, which may stem from the particular hand shape used for the lexical sign `RESTAURANT` (see thumbnail 6). A false FBuoy positive is detected at the beginning of the sequence, which actually makes sense since the left-hand DS-Ground (DS-G) is highly similar to a FBuoy, both in form and function.

Video `S7_T2_A10`, frames 5285-5385 (Figure 8.6)

This sixth and last sequence is quite illustrative. We have proposed the following translation: *"You should definitely go see this place where birds fly all around buildings."*.

The annotation includes four FLSs and three DSs. The first annotated unit, `BUILDING`, is annotated as a FLS although it is produced very iconically, and thus could have been annotated as DS. This unit is not recognized as FLS. Then, the illustrative structure around frame 5300 is recognized as such and as a lexical sign too. The lexical sign `BIRD` is very well detected as a FLS, and is indeed performed in a very standard way. The subsequent DS is well detected too and includes proforms for birds on both hands. The final two FLSs are accurately detected. One can note that the lexical unit `GO` is also recognized as a DS, which makes sense as it is clearly signed in the form of a T-P.

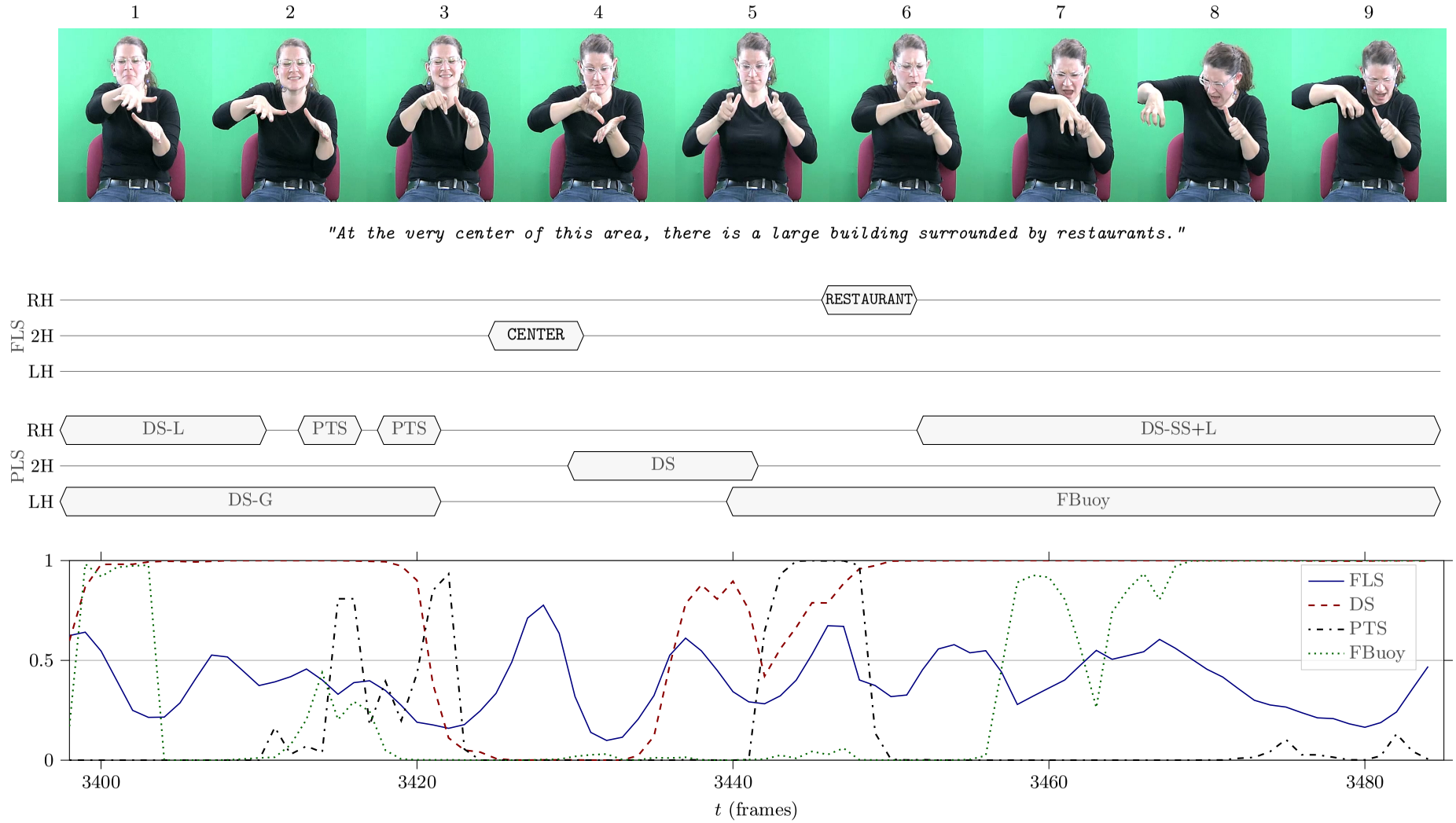


Figure 8.5: LSF sequence from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`, see Chapter 4). Expert annotations for right-handed (RH), two-handed (2H) and left-handed (LH) Fully Lexical Signs (FLSs) and Partially Lexical Signs (PLSs) are given.

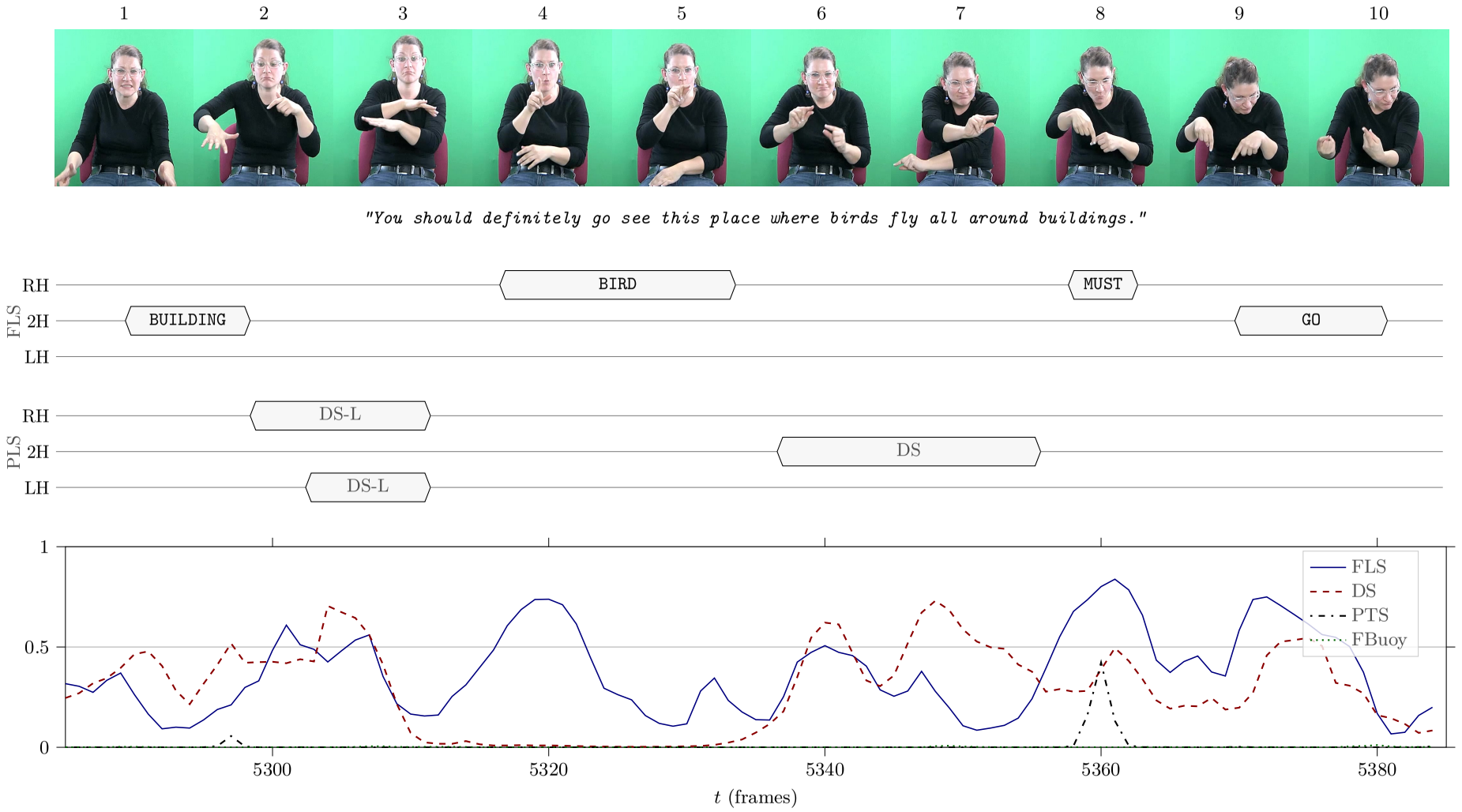


Figure 8.6: LSF sequence from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`, see Chapter 4). Expert annotations for right-handed (RH), two-handed (2H) and left-handed (LH) Fully Lexical Signs (FLSs) and Partially Lexical Signs (PLSs) are given.

		Frame-wise		Unit-wise		
		Acc	F1 (P/R)	$\mathrm{F1}^{\star}_{w}(t_w = 12)$ (P/R)	$\mathrm{F1}^{\star}_{pr}(0,0)$ (P/R)	I_{pr}
Figure 8.1	FLS	0.92	0.89 (1.00/0.80)	1.00 (1.00/1.00)	1.00 (1.00/1.00)	0.89
	DS	1.00	-	-	-	-
	PTS	0.92	0.57 (0.50/0.67)	0.67 (0.50/1.00)	0.67 (0.50/1.00)	0.44
	FBuoy	1.00	-	-	-	-
Figure 8.2	FLS	0.64	0.53 (0.71/0.43)	1.00 (1.00/1.00)	0.85 (0.80/0.90)	0.60
	DS	0.82	0.56 (0.47/0.68)	0.67 (0.50/1.00)	0.67 (0.50/1.00)	0.44
	PTS	0.91	0.40 (0.80/0.27)	0.80 (1.00/0.67)	0.80 (1.00/0.67)	0.59
	FBuoy	0.77	0.00 (- /0.00)	0.00 (- /0.00)	0.00 (- /0.00)	0.00
Figure 8.3	FLS	0.77	0.65 (0.73/0.58)	0.92 (0.86/1.00)	0.83 (0.71/1.00)	0.61
	DS	0.93	0.88 (0.83/0.93)	0.67 (0.50/1.00)	0.67 (0.50/1.00)	0.63
	PTS	0.93	0.75 (1.00/0.60)	1.00 (1.00/1.00)	1.00 (1.00/1.00)	0.80
	FBuoy	0.78	0.38 (1.00/0.23)	1.00 (1.00/1.00)	1.00 (1.00/1.00)	0.57
Figure 8.4	FLS	0.68	0.63 (0.67/0.60)	0.94 (0.89/1.00)	0.88 (0.78/1.00)	0.64
	DS	0.85	0.41 (0.35/0.50)	0.80 (0.67/1.00)	0.80 (0.67/1.00)	0.26
	PTS	0.97	0.00 (- /0.00)	0.00 (- /0.00)	0.00 (- /0.00)	0.00
	FBuoy	0.81	0.36 (1.00/0.22)	1.00 (1.00/1.00)	1.00 (1.00/1.00)	0.61
Figure 8.5	FLS	0.72	0.29 (0.20/0.50)	0.73 (0.57/1.00)	0.44 (0.29/1.00)	0.30
	DS	0.81	0.88 (0.86/0.90)	1.00 (1.00/1.00)	1.00 (1.00/1.00)	0.83
	PTS	0.84	0.22 (0.18/0.29)	0.80 (0.67/1.00)	0.40 (0.33/0.50)	0.30
	FBuoy	0.73	0.69 (0.84/0.59)	0.50 (0.33/1.00)	0.80 (0.67/1.00)	0.53
Figure 8.6	FLS	0.73	0.58 (0.70/0.50)	0.80 (0.67/1.00)	0.60 (0.50/0.75)	0.48
	DS	0.78	0.54 (0.72/0.43)	0.89 (0.80/1.00)	0.75 (0.60/1.00)	0.49
	PTS	1.00	-	-	-	-
	FBuoy	1.00	-	-	-	-

Table 8.2: Frame-wise accuracy, F1-score and integrated unit-wise metric I_{pr} for six sequence examples, illustrating the recognition of FLSs, DSs, PTSs and FBuoys. In case of no unit in the annotation (resp. the predictions), recall (resp. precision) can not be computed.

Conclusion

In this chapter, we have evaluated the proposed approach on the test videos of Dicta-Sign–LSF–v2, for the binary recognition of four linguistic descriptors: Fully Lexical Signs, Depicting Signs, Pointing Signs and Fragment Buoys.

First, we have shown that the original model and signer representation that we have built are good at generalizing to unseen signers or unseen tasks, with consistent results in Signer-Independent and Task Independent settings. This is especially interesting for extensions of this work, as eventually, all Sign Language Recognition (SLR) systems are to be replicable to new signers and tasks. We have also shown that the amount of training data currently appears to be a limiting factor for the model performance. A challenge for the future is thus to get many more hours of finely annotated SL corpora, or develop unsupervised or partially supervised frameworks.

In a more qualitative sequence-base analysis of the predictions, we have then highlighted the merits of our approach. The predictions of the four descriptors are generally well in line with the annotations, and could be used to describe a much broader part of SL discourse than the pure Continuous Lexical Sign Recognition (CLexSR) approach. Moreover, many of the observed discrepancies can actually be explained by the subjectivity in the annotation, some annotation mistakes or even the unclear boundary between certain categories, in terms of linguistic definition. In particular, the degenerated or *dormant* iconicity in many lexical signs is such that these units can be signed in a more or less iconic fashion, so that the FLS *versus* DS opposition may not always make sense. Consequently, it may have been more appropriate to allow for both unit types to be positively annotated at the same time in the original corpus. More generally, the predictions of the proposed model could help question the exclusivity and relevance of certain linguistic categories. This will however require an even more thorough analysis of the results in order to ensure that no erroneous conclusions are drawn due to shortcomings in the signer representation or learning model.

Conclusions and perspectives

Now is the time to bring this thesis to a conclusion. In these last pages, we start by going back to the original motivations and the related problem statement. We then summarize and discuss our main contributions and findings. Finally, we consider future perspectives of this work and Continuous Sign Language Recognition (CSLR) in general.

Back to the problem statement

In Chapters 1 and 2, we have developed on the complexity and specificities of SLs. Each SL has its own lexicon, made of conventionalized units called lexical signs. However, with a visual-gestural modality, all SLs include context-dependent illustrative or iconic structures and exploit the signing space to organize discourse, while simultaneously making use of multiple language articulators.

This visual syntax of SLs as well as these complex illustrative structures, referred to as transfers or classifier constructions, are generally ignored from the field of CSLR. Instead, CSLR has focused on the development of models for the recognition of lexical signs in the form of sequences within SL videos – hence the more appropriate naming Continuous Lexical Sign Recognition (CLexSR) –, which is thoroughly developed in Chapter 3. These models can be used to understand the general meaning of simple utterances, however they are bound to fail in the case of more illustrative or spatial discourse, that is the very natural features of SLs. Relatedly, available corpora for SLR are very specific. They are often elicited from simple and artificial sentences, even though the popular RWTH-Phœnix-Weather corpus, made of interpreted German Sign Language (DGS), is somehow more natural. In any case, the annotation of these corpora only include lexical signs. These works are in no way useless, and they have definitely contributed to push SLR forward. We can only regret that the authors almost never acknowledge the complexity of SL linguistics and the limited scope of their work in this regard.

In light of this state of the art, we have then asked ourselves the following question: what should and could be envisioned so that CSLR can effectively be seen as a stepping stone towards – in the medium term – Sign Language Understanding (SLU) and – in the long run – Sign Language Translation (SLT)?

Our main contributions and findings

Our first contribution focuses on improving the input data for CSLR systems. Corpora made by linguists are very relevant in terms of language quality and representativeness, yet the annotation is generally inconsistent and incomplete, thus not usable as such. Therefore, we have redesigned a French Sign Language (LSF) corpus previously made by linguists, Dicta-Sign–LSF–v2. In particular, we have

ensured quality and consistency in the annotation for 11 hours of recordings, which we have also made public. This is presented in Chapter 4, with detailed explanations on the annotation categories and statistics.

We then developed in Chapter 5 a formal and general description of the problem of CSLR, understood as the parallel recognition of linguistic descriptors, which should be as relevant as possible. Accordingly, we have discussed and introduced adapted performance metrics, going past the commonly used measures in the case of CLexSR. This represents our second contribution.

Our third contribution is an open-access proposal and implementation for an original combination of signer representation and learning model, which is detailed in Chapter 6. Using a mix of publicly available and self-developed models, we were indeed able to derive a generalizable and compact representation of signers in videos, with a separate processing of upper body, hands and face and the manufacturing of linguistically relevant features. This signer data can then be used as input to a convolutional and recurrent neural network, which allows fast training while being known to be an effective architecture for gesture recognition. This decoupled framework presents a few advantages compared to end-to-end approaches, like the reduced needs in terms of training data since only the learning model has to be trained, or the feedback it can give on the relative importance of different signer representation features for the recognition performance.

Finally, we have conducted a thorough analysis of the recognition performance of the proposed model for four very different linguistic descriptors – Fully Lexical Signs (FLSs), Depicting Signs (DSs), Pointing Signs (PTSs) and Fragment Buoys (FBuoys) – on Dicta-Sign–LSF–v2. Although the amount of training data can be a limiting factor, quantitative results on the validation set – Chapter 7 – show that with the right choice of signer representation and learning parameters, very promising performance values are met. Subsequently, a qualitative analysis on the test set demonstrates the merits of the proposed approach. The model is shown to be resilient to signer-independence and task-independence, which is a relief as the generalizability is often an overlooked issue in SLR. A detailed analysis on a few test sequences shows that there is generally good agreement between annotations and predictions. Furthermore, many cases of disagreement can be explained by errors, subjectivity or ambiguity in the annotation process.

Perspectives

As a first attempt of generalizing CSLR on a linguistically relevant SL corpus, our work has proven effective while leaving room for improvement.

In terms of signer representation and learning framework, we have tried to build a compact solution requiring limited data while ensuring generalizability. As there is an important overlap with the very active research field of human activity recognition, this part of our work could easily be improved in the near future. As regards the signer representation, the hand area is notably critical and we have to acknowledge that the solution we could come up with is not optimal. Three-dimensional hand modeling, for instance, would enable to capture more of the signed information. Furthermore, we have shown that preprocessed signer features performed better than raw data. This lead should definitely be explored further, and will imply the work of linguists to determine features that are as relevant as possible. More refined state-of-the-art learning architectures developed for Natural Language Processing (NLP) are also a potential lead although generally they can not be used as is for frame-wise predictions.

Another direction for improvement is related to the availability of data. As we have pointed out, the amount of training data appears to be a bottleneck for the performance of CSLR models: this is thus a major challenge for the evolution of CSLR. Similarly to the work we have conducted when redesigning Dicta-Sign to Dicta-Sign–LSF–v2, a corpus made for SLR purposes, adapting corpora made by linguists seems both realistic and highly cost-efficient. However, because ensuring consistency in the annotation is highly time-consuming, another direction worth exploring is the adoption of weakly supervised or even unsupervised learning methods for these partially or inconsistently annotated corpora. Exploiting methods to help annotate videos much faster using visual and dynamic cues like [Chaaban et al., 2019] could also be very useful.

On a close topic, the content of data itself and the recognition objectives of CSLR models, which are obviously related, could and should definitely be extended in order to capture much more of the information conveyed in SL. In particular, SL corpora are always – to the best of our knowledge – annotated along a temporal axis alone, which makes it especially unnatural to capture the spatial characteristics of signed utterances. As we have detailed in this thesis, the diagrammatic iconicity of SLs indeed consists in organizing discourse spatially, which temporal annotation alone can not describe. Research in this direction is crucial and should allow for better integration of the way information and discourse are represented in SLs, which eventually will lead to much better SLU.

Finally, and in contrast to the usual approach of using linguistic knowledge to support NLP applications, models such as the ones we have developed in this thesis could be of great help to linguists. Primarily, for linguists who aim at improving the description of SLs, the ability or inability of a prediction model to effectively discriminate between different types of units, for instance, could indeed be used as a measure of the very relevance of such categories. Incidentally, our research may also be used to analyze other oral languages and specifically the role of co-verbal gestures with respect to cognitive operations, as suggested by Cuxac [2001] (*SLs are ideal analyzers for the language faculty*). In this way, we hope that this work will have started to bridge the gap between the communities of linguists and computer scientists, for the benefit of both.

Appendix A

Performance metrics for temporal data: illustration

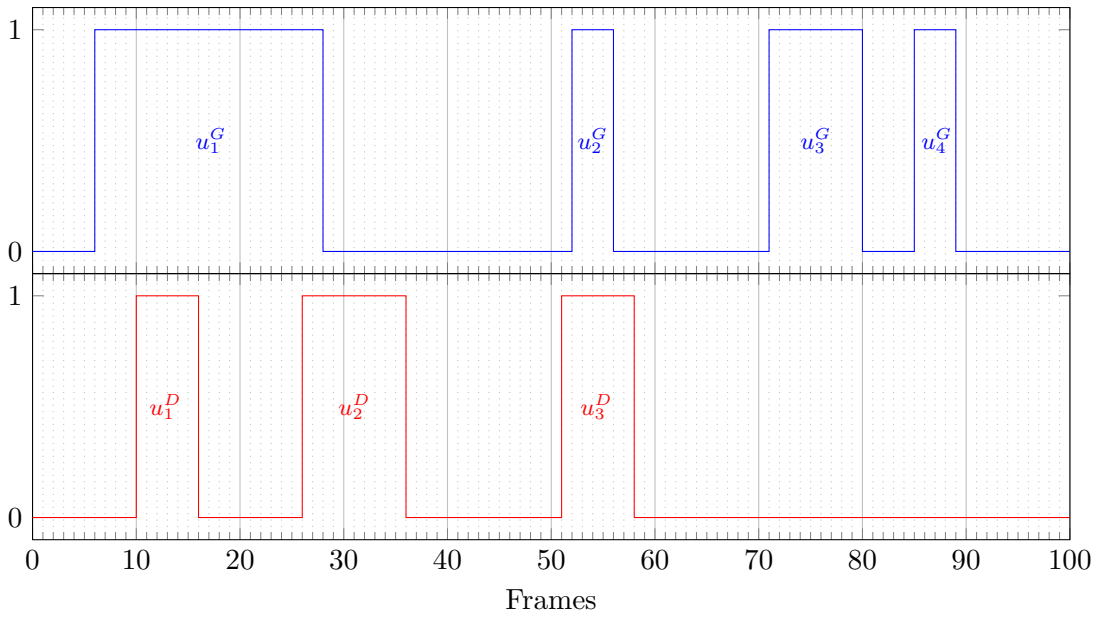


Figure A.1: Annotated (top, blue) and predicted (bottom, red) data in a dummy binary classification problem. Four units are annotated, while three are detected.

In this appendix, we illustrate the performance metrics presented in Section 5.2. We choose the case of binary classification, with a dummy sequence for which fictitious annotated and predicted data are given in Figure A.1.

A.1 Frame-wise metrics

The frame-wise metrics presented in Section 5.2.1 are easily computed. First, the accuracy is the rate of correctly predicted frames, including class 0:

$$\text{Acc} = 0.61.$$

Frame-wise precision and recall are computed from the count of true positives, false positives and false negatives frames (see Equations 5.6 and 5.7):

$$P = \frac{15}{15+11} \simeq 0.58$$
$$R = \frac{15}{15+28} \simeq 0.35$$

which yield:

$$\text{F1} \simeq 0.44.$$

A.2 Unit-wise metrics

A.2.1 $\mathbf{P^\star_w}$, $\mathbf{R^\star_w}$, $\mathbf{F1^\star_w}$

Let us first note that:

- The closest unit from u_1^D is unit u_1^G, with 4 frames of shift between their respective centers.
- The closest unit from u_2^D is unit u_1^G, with 14 frames of shift between their respective centers.
- The closest unit from u_3^D is unit u_2^G, with 0.5 frame of shift between their respective centers.

Also:

- The closest unit from u_1^G is unit u_1^D, with 4 frames of shift between their respective centers.
- The closest unit from u_2^G is unit u_3^D, with 0.5 frame of shift between their respective centers.
- The closest unit from u_3^G is unit u_3^D, with 21 frames of shift between their respective centers.
- The closest unit from u_4^G is unit u_3^D, with 32.5 frames of shift between their respective centers.

From Equations 5.11 and 5.12, unit-wise precision and recall as a function of a margin t_w can be written as:

$$P_w^\star(t_w) = \frac{1}{3}\left(\mathbb{1}_{t_w>4} + \mathbb{1}_{t_w>14} + \mathbb{1}_{t_w>0.5}\right)$$
$$R_w^\star(t_w) = \frac{1}{4}\left(\mathbb{1}_{t_w>4} + \mathbb{1}_{t_w>0.5} + \mathbb{1}_{t_w>21} + \mathbb{1}_{t_w>32.5}\right).$$

With margins of half a second (12 frames) or one second (25 frames):

$$P_w^\star(12) \simeq 0.67$$
$$R_w^\star(12) = 0.5$$
$$\text{F1}_w^\star(12) \simeq 0.57$$

and

$$P_w^\star(25) = 1$$
$$R_w^\star(25) = 0.75$$
$$\text{F1}_w^\star(25) \simeq 0.86.$$

A.2.2 $\mathbf{P^\star_{pr}}$, $\mathbf{R^\star_{pr}}$, $\mathbf{F1^\star_{pr}}$

From Equations 5.13 and 5.14, one can note that:

- The best match for unit u_1^D is unit u_1^G, with 7 intersecting frames over the original 7 frames of u_1^D.
- The best match for unit u_2^D is unit u_1^G, with 3 intersecting frames over the original 11 frames of u_2^D.
- The best match for unit u_3^D is unit u_2^G, with 5 intersecting frames over the original 8 frames of u_3^D.

Also:

- The best match for unit u_1^G is unit u_1^D, with 10 intersecting frames over the original 23 frames of u_1^G.
- The best match for unit u_2^G is unit u_3^D, with 5 intersecting frames over the original 5 frames of u_2^G.
- The best match for unit u_3^G is any unit u_i^D, because there is no intersection.
- The best match for unit u_4^G is any unit u_i^D, because there is no intersection.

Then, with IM standing for IsMatch, $P^\star_{pr}$ and $R^\star_{pr}$ of Equations 5.15 and 5.16 can be simply expressed as:

$$\begin{aligned}
P^\star_{pr}(\bar{t}_p, \bar{t}_r) &= \frac{1}{3}\left(\mathrm{IM}(u_1^D, u_1^G, \bar{t}_p, \bar{t}_r) + \mathrm{IM}(u_2^D, u_1^G, \bar{t}_p, \bar{t}_r) + \mathrm{IM}(u_3^D, u_2^G, \bar{t}_p, \bar{t}_r)\right) \\
&= \frac{1}{3}\left(\mathbb{1}_{\{\frac{7}{7} > \bar{t}_p, \frac{7}{23} > \bar{t}_r\}} + \mathbb{1}_{\{\frac{3}{11} > \bar{t}_p, \frac{3}{23} > \bar{t}_r\}} + \mathbb{1}_{\{\frac{5}{8} > \bar{t}_p, \frac{5}{5} > \bar{t}_r\}}\right) \\
R^\star_{pr}(\bar{t}_p, \bar{t}_r) &= \frac{1}{4}\left(\mathrm{IM}(u_1^D, u_1^G, \bar{t}_p, \bar{t}_r) + \mathrm{IM}(u_3^D, u_2^G, \bar{t}_p, \bar{t}_r) + \mathrm{IM}(u_i^D, u_3^G, \bar{t}_p, \bar{t}_r) + \mathrm{IM}(u_i^D, u_4^G, \bar{t}_p, \bar{t}_r)\right) \\
&= \frac{1}{4}\left(\mathbb{1}_{\{\frac{7}{7} > \bar{t}_p, \frac{7}{23} > \bar{t}_r\}} + \mathbb{1}_{\{\frac{5}{8} > \bar{t}_p, \frac{5}{5} > \bar{t}_r\}} + 0 + 0\right).
\end{aligned}$$

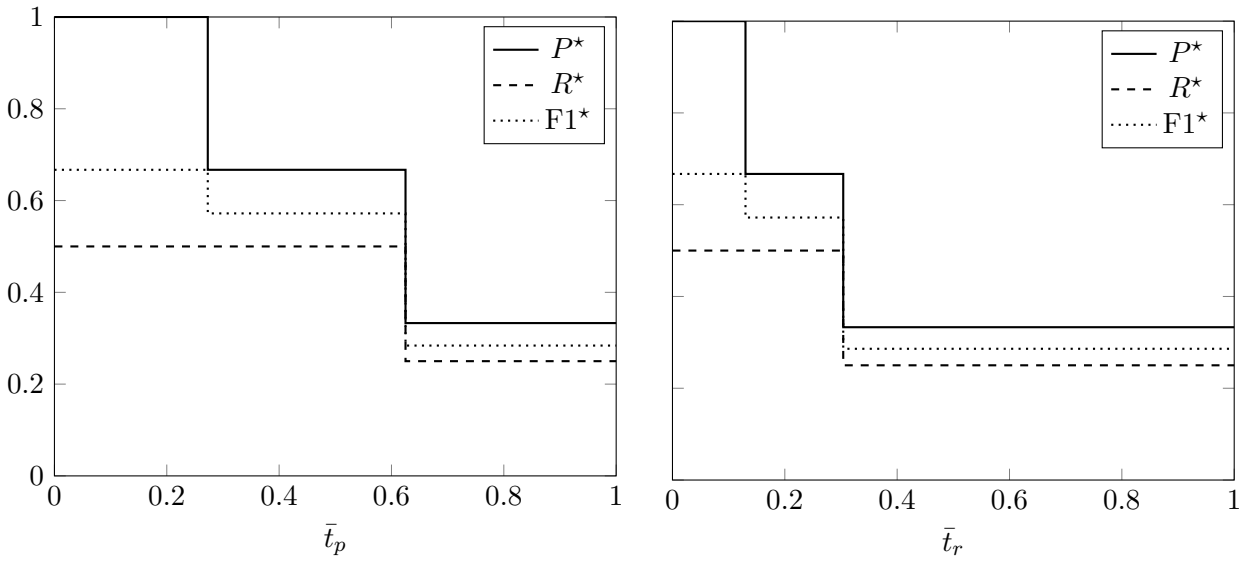
These formula make it possible to draw curves for $P^\star_{pr}$, $R^\star_{pr}$ and $\mathrm{F1}^\star_{pr}$, either with fixed $\bar{t}_r = 0$ or fixed $\bar{t}_p = 0$. This is shown in Figure A.2.

The calculation of area under curves (Equations 5.19, 5.20 and 5.21) then yields:

$$I_p \simeq 0.490$$
$$I_r \simeq 0.385$$

and finally:

$$\boxed{I_{pr} \simeq 0.438}.$$



(a) $\bar{t}_r = 0$ (b) $\bar{t}_p = 0$

Figure A.2: Unit-wise $P^\star_{pr}$, $R^\star_{pr}$ and F1$^\star_{pr}$ values, in the case of the dummy sequences of Figure A.1, as a function of $\bar{t}_p$ ($\bar{t}_r = 0$), or as a function of $\bar{t}_r$ ($\bar{t}_p = 0$).

Peer-reviewed publications during the Ph.D.

Journal papers

Valentin Belissen, Annelies Braffort, and Michèle Gouiffès. Experimenting the Automatic Recognition of Non-Conventionalized Units in Sign Language. *Algorithms*, 13(12):310, 2020.

Conference papers

Valentin Belissen, Michèle Gouiffès, and Annelies Braffort. 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 12th International Conference on Language Resources and Evaluation (LREC 2020)*, ELRA, 2020.

Valentin Belissen, Michèle Gouiffès, and Annelies Braffort. Improving and Extending Continuous Sign Language Recognition: Taking Iconicity and Spatial Language into account. In *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. Satellite Workshop to the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, ELRA, 2020.

Valentin Belissen, Annelies Braffort, and Michèle Gouiffès. Towards Continuous Recognition of Illustrative and Spatial Structures in Sign Language. In *Sign Language Recognition, Translation and Production (SLRTP)*, ECCV Workshops, vol. 4, 2020.

Doctoral consortia

Valentin Belissen. Sign Language Video Analysis For Automatic Recognition and Detection, *14th IEEE International Conference on Automatic Face and Gesture Recognition*, Lille, France, 2019.

Valentin Belissen. Sign Language Video Analysis For Automatic Recognition and Detection, *20th International ACM SIGACCESS Conference on Computers and Accessibility*, Galway, Ireland, 2018.

Bibliography

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, volume 16, pages 265–283, 2016.
(cited p. 110, 118)

Charlotte Baker and Dennis Cokely. American Sign Language. *A Teacher's Resource Text on Grammar and Culture. Silver Spring, MD: TJ Publ*, 1980.
(cited p. 86)

Robbin Battison. Phonological Deletion in American Sign Language. *Sign Language Studies*, 5(1): 1–19, 1974.
(cited p. 39, 116)

Auguste Bébian. *Mimographie, ou Essai d'écriture mimique propre à régulariser le langage des sourds-muets*. L. Colas, 1825.
(cited p. 39)

Valentin Belissen, Michèle Gouiffès, and Annelies Braffort. 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 12th International Conference on Language Resources and Evaluation (LREC 2020), ELRA*, 2020a.
(cited p. 23, 50, 87)

Valentin Belissen, Michèle Gouiffès, and Annelies Braffort. Improving and Extending Continuous Sign Language Recognition: Taking Iconicity and Spatial Language into account. In *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. Satellite Workshop to the 12th International Conference on Language Resources and Evaluation (LREC 2020), ELRA*, 2020b.
(cited p. 142)

Dominique Boutet, Claudia S Bianchini, Patrick Doan, Léa Chèvrefils, Chloé Thomas, Morgane

Rébulard, Adrien Contesse, Claire Danet, Jean-François Dauphin, and Mathieu Réguer. Réflexions sur la formalisation, en tant que système, d'une transcription des formes des Langues des Signes: l'approche Typannot. In *SHS Web of Conferences*, 2020.
(cited p. 46)

Annelies Braffort. *Reconnaissance et compréhension de gestes, application à la langue des signes.* PhD thesis, Université de Paris XI - Orsay, 1996.
(cited p. 53, 71, 77, 79)

Annelies Braffort, Annick Choisier, Christophe Collet, Christian Cuxac, Patrice Dalle, Ivani Fusellier, Rachid Gherbi, Guillemette Jausions, Gwenaelle Jirou, Fanch Lejeune, et al. Projet LS-COLIN. Quel outil de notation pour quelle analyse de la LS. *Journées Recherches sur la langue des signes. UTM, Le Mirail, Toulouse*, 2001.
(cited p. 111)

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
(cited p. 114, 115)

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
(cited p. 72, 74)

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
(cited p. 15, 68, 73, 75, 76, 77)

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
(cited p. 13, 15, 74, 76, 77)

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.
(cited p. 19, 55, 86, 108)

Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? a New Model and the Kinetics Dataset. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.
(cited p. 63)

Hussein Chaaban, Michèle Gouiffès, and Annelies Braffort. Towards an Automatic Annotation of French Sign Language Videos: Detection of Lexical Signs. In *International Conference on Computer Analysis of Images and Patterns*, pages 402–412. Springer, 2019.
(cited p. 155)

Xiujuan Chai, Hanjie Wang, and Xilin Chen. The DEVISIGN Large Vocabulary of Chinese Sign Language Database and Baseline Evaluations. Technical report, Institute of Computing Technology, 2014.
(cited p. 23, 58, 59, 60)

François Chollet et al. Keras. https://keras.io, 2015.
(cited p. 110, 118)

Helen Cooper, Brian Holt, and Richard Bowden. Sign Language Recognition. In *Visual Analysis of Humans*, pages 539–562. Springer, 2011.
(cited p. 60, 78, 107)

Onno Crasborn, Inge Zwitserlood, and Johan Ros. Corpus NGT. *An open access digital corpus of movies with annotations of Sign Language of the Netherlands (Video corpus). Centre for Language Studies, Radboud University Nijmegen*, 2008.
(cited p. 23, 86)

Onno A Crasborn. The Sign Linguistics Corpora Network: towards standards for signed language resources. *Workshop 3 (annotation)*, 2010.
(cited p. 89)

Onno A Crasborn and Inge Zwitserlood. The Corpus NGT: an Online Corpus for Professionals and Laymen. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Satellite Workshop to the 6th International Conference on Language Resources and Evaluation (LREC 2008), ELRA*, pages 44–49. Paris: ELRA, 2008.
(cited p. 23, 86)

Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7361–7369, 2017.
(cited p. 72, 74)

Christian Cuxac. French Sign Language: Proposition of a Structural Explanation by Iconicity. In *Proceedings of the 1999 International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 165–184. Springer, 1999.
(cited p. 13, 40, 41, 42, 49, 50)

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

(cited p. 22, 40, 41, 43)

Christian Cuxac. Les langues des signes: analyseurs de la faculté de langage. *Acquisition et interaction en langue étrangère*, (15):11–36, 2001.
(cited p. 155)

Christian Cuxac and Marie-Anne Sallandre. Iconicity and arbitrariness in French Sign Language: Highly iconic structures, degenerated iconicity and diagrammatic iconicity. *Empirical Approaches to Language Typology*, 36:13, 2007.
(cited p. 40)

Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition. In *International Conference on Artificial Neural Networks*, pages 646–657. Springer, 2019.
(cited p. 57, 64, 65)

Ferdinand De Saussure. *Course in General Linguistics*. Columbia University Press, 1916.
(cited p. 39)

Asa DeMatteo. Visual imagery and visual analogues in American Sign Language. *On the other hand: New perspectives on American Sign Language*, pages 109–136, 1977.
(cited p. 40)

Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, and Dimitris N Metaxas. A New Framework for Sign Language Recognition based on 3D Handshape Identification and Linguistic Modeling. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, 2014.
(cited p. 61, 72)

Mark Dilsizian, Zhiqiang Tang, Dimitris Metaxas, Matt Huenerfauth, and Carol Neidle. The Importance of 3D Motion Trajectories for Computer-based Sign Recognition. *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining. Satellite Workshop to the 10th International Conference on Language Resources and Evaluation (LREC 2016), ELRA*, 2016.
(cited p. 61, 108)

Mark Dilsizian, Dimitris Metaxas, and Carol Neidle. Linguistically-driven Framework for Computationally Efficient and Scalable Sign Recognition. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
(cited p. 13, 15, 62, 63, 64, 65, 108, 116)

Alistair DN Edwards. Progress in Sign Language Recognition. In *Proceedings of the 1997 International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 13–21. Springer, 1997.
(cited p. 79)

Gareth J Edwards, Christopher J Taylor, and Timothy F Cootes. Interpreting Face Images using Active Appearance Models. In *Proceedings of the 1998 3rd IEEE International Conference on Automatic Face & Gesture Recognition*, pages 300–305. IEEE, 1998.
(cited p. 60, 72)

Eleni Efthimiou, Stavroula-Evita Fontinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Goudenove. Dicta-Sign: Sign Language Recognition, Generation, and Modelling with Applications in Deaf Communication. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. Satellite Workshop to the 7th International Conference on Language Resources and Evaluation (LREC 2010), ELRA*, pages 80–83, 2010.
(cited p. 23, 87)

Gaolin Fang, Wen Gao, Xilin Chen, Chunli Wang, and Jiyong Ma. Signer-Independent Continuous Sign Language Recognition Based on SRN/HMM. In *Proceedings of the 2001 International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 76–85. Springer, 2001.
(cited p. 71)

Benoit Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, et al. Automatic human utility evaluation of ASR systems: Does WER really predict performance? In *INTERSPEECH*, pages 3463–3467, 2013.
(cited p. 67)

Michael Filhol and Annelies Braffort. What constraints for representing multilinearity in Sign language? In *Proceedings of the 7th International Workshop on Constraint Solving and Language Processing (CSLP12)*, page 106. Citeseer, 2012.
(cited p. 44)

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, volume 9, pages 3785–3789, 2012.
(cited p. 24, 68, 70)

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1911–1916, 2014.
(cited p. 24, 68, 70)

Nancy Frishberg. Arbitrariness and iconicity: historical change in American Sign Language. *Language*, pages 696–719, 1975.
(cited p. 41)

Wen Gao, Gaolin Fang, Debin Zhao, and Yiqiang Chen. Transition Movement Models for Large Vocabulary Continuous Sign Language Recognition. In *Proceedings of the 2004 6th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 553–558. IEEE, 2004.
(cited p. 72)

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000.
(cited p. 117)

Matilde Gonzalez Preciado. *Computer Vision Methods for Unconstrained Gesture Recognition in the Context of Sign Language Annotation.* PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2012.
(cited p. 107)

Nicolas Granger and Mounîm A el Yacoubi. Comparing Hybrid NN-HMM and RNN for Temporal Modeling in Gesture Recognition. In *Proceeding of Advances in Neural Information Processing Systems (NIPS 2017)*, pages 147–156. Springer, 2017.
(cited p. 115)

Pierre Guitteny. *Le passif en langue des signes.* PhD thesis, Université Michel de Montaigne - Bordeaux III, 2006.
(cited p. 34)

Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. Hierarchical LSTM for Sign Language Translation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.
(cited p. 72, 74)

Dan Guo, Shengeng Tang, and Meng Wang. Connectionist Temporal Modeling of Video and Language: a Joint Model for Translation and Sign Labeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 751–757. AAAI Press, 2019a.
(cited p. 72, 73, 74)

Dan Guo, Shuo Wang, Qi Tian, and Meng Wang. Dense Temporal Convolution Network for Sign Language Translation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 744–750. AAAI Press, 2019b.
(cited p. 72, 73, 74)

Thomas Hanke. HamNoSys –Representing Sign Language Data in Language Resources and Language Processing Contexts. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, volume 4, pages 1–6, 2004.
(cited p. 21, 46, 47, 87)

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based Sign Language Recognition without Temporal Segmentation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

(cited p. 23, 71, 72, 74)

Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand Pose Estimation via Latent 2.5 D Heatmap Regression. In *Proceedings of the 2018 European Conference on Computer Vision (ECCV)*, 2018.
(cited p. 112)

Elena Jahn, Reiner Konrad, Gabriele Langer, Sven Wagner, and Thomas Hanke. Publishing dgs corpus data: Different formats for different needs. In *Proceedings of the 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. Satellite Workshop to the 11th International Conference on Language Resources and Evaluation (LREC 2018), ELRA*, pages 83–90, 2018.
(cited p. 86)

Terry Janzen. *Topics in signed language interpreting: Theory and practice*, volume 63. John Benjamins Publishing, 2005.
(cited p. 45)

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.
(cited p. 114)

Trevor Johnston. Creating a corpus of Auslan within an Australian National Corpus. In *Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, pages 87–95, 2009.
(cited p. 23, 49, 83)

Trevor Johnston and L De Beuzeville. Auslan Corpus Annotation Guidelines. *Centre for Language Sciences, Department of Linguistics, Macquarie University*, 2016.
(cited p. 46, 49, 50, 80, 90, 91, 93, 96)

Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. In *Proceedings of the 2018 British Machine Vision Conference (BMVC)*, 2018.
(cited p. 15, 23, 57, 58, 60, 63, 64, 65)

Mohammed Waleed Kadous et al. Machine Recognition of Auslan Signs using PowerGloves: Towards Large-Lexicon Recognition of Sign Language. In *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, volume 165, 1996.
(cited p. 53)

Oscar Koller, Jens Forster, and Hermann Ney. Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.

(cited p. 68, 72, 74)

Oscar Koller, Hermann Ney, and Richard Bowden. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data is Continuous and Weakly Labelled. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3793–3802, 2016a.
(cited p. 14, 72, 73, 74, 112, 113, 114, 115)

Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *Proceedings of the 2016 British Machine Vision Conference (BMVC)*, 2016b.
(cited p. 72, 74)

Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.
(cited p. 68, 72, 74, 75)

Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *International Journal of Computer Vision,*, 126(12):1311–1325, 2018.
(cited p. 74)

Oscar Koller, Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2019.
(cited p. 13, 73, 74)

François Lefebvre-Albaret. *Traitement automatique de vidéos en LSF Modélisation et exploitation des contraintes phonologiques du mouvement.* PhD thesis, Université Paul Sabatier-Toulouse III, 2010.
(cited p. 107)

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In *Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1469, 2020.
(cited p. 24, 57)

Yanqiu Liao, Pengwen Xiong, Weidong Min, Weiqiong Min, and Jiahao Lu. Dynamic Sign Language Recognition Based on Video Sequence with BLSTM-3D Residual Networks. *IEEE Access*, 7:38044–38054, 2019.
(cited p. 63, 65)

Scott K Liddell. Grounded Blends, Gestures, and Conceptual Shifts, 1998.
(cited p. 40)

Scott K Liddell. *Grammar, Gesture, and Meaning in American Sign Language.* Cambridge University

Press, 2003.
(cited p. 50, 90)

Kian Ming Lim, Alan WC Tan, and Shing Chiang Tan. Block-based Histogram of Optical Flow for Isolated Sign Language Recognition. *Journal of Visual Communication and Image Representation*, 40:538–545, 2016.
(cited p. 61, 64, 65)

LIMSI and CIAMS. MOCAP1. https://hdl.handle.net/11403/mocap1/v1, 2017. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.
(cited p. 109, 110)

LIMSI and IRIT. Dicta-Sign-LSF-v2. https://hdl.handle.net/11403/dicta-sign-lsf-v2, 2019. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.
(cited p. 23, 87)

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL https://www.aclweb.org/anthology/D15-1166.
(cited p. 72)

Elizabeth Macken, John Perry, and Cathy Haas. Richly grounding symbols in ASL. *Sign Language Studies*, 81(1):375–394, 1993.
(cited p. 40, 41)

Aleix M Martínez, Ronnie B Wilbur, Robin Shay, and Avinash C Kak. Purdue RVL-SLLL ASL Database for Automatic Recognition of American Sign Language. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, pages 167–172. IEEE, 2002.
(cited p. 23, 61)

Richard P Meier. Elicited imitation of verb agreement in American Sign Language: iconically or morphologically determined? *Journal of Memory and Language*, 26(3):362–376, 1987.
(cited p. 13, 43)

Dimitris N Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael, and Carol Neidle. Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2414–2420, 2012.
(cited p. 77)

Melanie Metzger. *Sign language interpreting: Deconstructing the myth of neutrality*. Gallaudet University Press, 1999.
(cited p. 45)

Laurence Meurant, Aurélie Sinte, and Eric Bernagou. The French Belgian Sign Language Corpus A User-Friendly Searchable Online Corpus. In *Proceedings of the 7th workshop on the Representation and Processing of Sign Languages: Corpus Mining. Satellite Workshop to the 10th International Conference on Language Resources and Evaluation (LREC 2016), ELRA*, pages 167–174, 2016.
(cited p. 23, 86)

Agnès Millet and Jean-Marc Colletta. Présentation: Des mouvements corporels à la syntaxe des langues gestuelles et de la communication parlée. *LIDIL: Revue de linguistique et de didactique des langues*, (26):7–26, 2002.
(cited p. 40)

Agnès Millet, Nathalie Niederberger, and Marion Blondel. French Sign Language. In Julie Bakken Jepsen, Goedele De Clerck, Sam Lutalo-Kiingi, and William B McGregor, editors, *Sign languages of the world: A comparative handbook*, chapter 10, page 273. Walter de Gruyter GmbH & Co KG, 2015.
(cited p. 44)

Bill Moody, Agnès Vourc'h, Michel Girod, and Anne-Catherine Dufour. *La langue des signes, dictionnaire bilingue LSF/Français*. Editions IVT, 1997.
(cited p. 48)

Bernard Mottez and Harry Markowicz. Intégration ou droit à la différence. *Paris: Centre d'Etudes des Mouvements Sociaux*, 1979.
(cited p. 34)

Bernard Mottez, Harry Markowicz, and David Armstrong. Deaf Identity. *Sign Language Studies*, 68 (1):195–216, 1990.
(cited p. 34)

Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. GANerated Hands for Real-time 3D Hand Tracking from Monocular RGB. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
(cited p. 112)

Kouichi Murakami and Hitomi Taguchi. Gesture Recognition using Recurrent Neural Networks. In *Proceedings of the 1991 SIGCHI Conference on Human factors in Computing Systems*, pages 237–242, 1991.
(cited p. 53)

Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML10)*, pages 807–814, 2010.
(cited p. 110)

Carol Neidle and Christian Vogler. A new web interface to facilitate access to corpora: Development of the ASLLRP data access interface (DAI). In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Satellite Workshop to the 8th International Conference on Language Resources and Evaluation (LREC 2012), ELRA*. Citeseer, 2012. URL http://www.bu.edu/asllrp/ncslgr.html.
(cited p. 23, 77, 86)

Carol Neidle, Ashwin Thangali, and Stan Sclaroff. Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Satellite Workshop to the 8th International Conference on Proceedings of the Language Resources and Evaluation (LREC 2012), ELRA*. Citeseer, 2012.
(cited p. 23, 55, 56, 60, 61, 63)

François-Xavier Nève. *Essai de grammaire de la langue des signes française*, volume 271. Librairie Droz, 1996.
(cited p. 40)

Natalia Neverova, Christian Wolf, Graham W Taylor, and Florian Nebout. Multi-scale Deep Learning for Gesture Detection and Localization. In *Proceedings of the 2014 European Conference on Computer Vision (ECCV)*, pages 474–490. Springer, 2014.
(cited p. 115)

Sylvie CW Ong and Surendra Ranganath. Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):873–891, 2005.
(cited p. 60)

Gerardo Ortega. Iconicity and Sign Lexical Acquisition: A Review. *Frontiers in Psychology*, 8:1280, 2017.
(cited p. 13, 41, 42)

Robert Östling, Carl Börstell, and Servane Courtaux. Visual Iconicity Across Sign Languages: Large-Scale Automated Video Analysis of Iconic Articulators and Locations. *Frontiers in psychology*, 9: 725, 2018.
(cited p. 41, 116)

Carol Padden. Verbs and Role-shifting in American Sign Language. In *Proceedings of the fourth national symposium on sign language research and teaching*, volume 44, page 57. National Association of the Deaf Silver Spring, MD, 1986.
(cited p. 43)

Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a Single RGB Frame for Real Time 3D Hand Pose Estimation in the wild. In *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018.

(cited p. 112)

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
(cited p. 76, 77)

Georgios Pavlakos, Xiaowei Zhou, Konstantinos Derpanis, and Kostas Daniilidis. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272. IEEE, 2017.
(cited p. 108)

Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. Sign Language Recognition Using Convolutional Neural Networks. In *Proceedings of the 2014 European Conference on Computer Vision (ECCV)*, pages 572–578. Springer, 2014.
(cited p. 63)

Lionel Pigou, Aäron Van Den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. *International Journal of Computer Vision,*, 126(2-4):430–439, 2018.
(cited p. 63, 118)

Elena Antinoro Pizzuto, Paola Pietrandrea, and Raffaele Simone. *Verbal and Sign Languages. Comparing Structures, Constructs, Methodologies.* Mouton De Gruyter, 2007.
(cited p. 40)

Siegmund Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, and Arvid Schwarz. DGS Corpus Project-Development of a Corpus Based Electronic Dictionary German Sign Language / German. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Satellite Workshop to the 6th International Conference on Language Resources and Evaluation (LREC 2008), ELRA*, 2008.
(cited p. 23, 86)

Junfu Pu, Wengang Zhou, Jihai Zhang, and Houqiang Li. Sign Language Recognition Based on Trajectory Modeling with HMMs. In *Proceedings of the 2016 International Conference on Multimedia Modeling (ICMM)*, pages 686–697. Springer, 2016.
(cited p. 15, 23, 59, 60, 61, 65)

Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative Alignment Network for Continuous Sign Language Recognition. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4165–4174, 2019.
(cited p. 72, 74)

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object

Detection with Region Proposal Networks. In *Proceeding of Advances in Neural Information Processing Systems (NIPS 2015)*, pages 91–99, 2015.
(cited p. 63)

Annie Risler. Les classes lexicales en LSF envisagées à partir de la fonction adjectivale. *Silexicales, Université de Lille*, 2007.
(cited p. 44)

Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net: Localization-Classification-Regression for Human Pose. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, United States, 2017.
(cited p. 108)

M Sallandre. Iconicity and Space in French Sign Language. *Typological studies in language*, 66:239, 2006.
(cited p. 41)

Marie-Anne Sallandre. *Les unités du discours en Langue des Signes Française. Tentative de catégorisation dans le cadre d'une grammaire de l'iconicité.* PhD thesis, Sciences Du Langage (SDL) - Université Paris 8, 2003.
(cited p. 45)

Marie-Anne Sallandre and Christian Cuxac. Iconicity in Sign Language: a theoretical and methodological point of view. *Lecture Notes in Computer Science*, pages 173–180, 2002.
(cited p. 41)

Marie-Anne Sallandre, Antonio Balvet, Geoffrey Besnard, and Brigitte Garcia. Étude exploratoire de la fréquence des catégories linguistiques dans quatre genres discursifs en LSF. *Lidil. Revue de linguistique et de didactique des langues*, (60), 2019.
(cited p. 93)

Adam Schembri. British Sign Language Corpus Project: Open Access Archives and the Observer's Paradox. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Satellite Workshop to the 6th International Conference on Language Resources and Evaluation (LREC 2008), ELRA*, pages 165–169, 2008.
(cited p. 23, 86)

Adam Schembri, Kearsy Cormier, Trevor Johnston, David McKee, Rachel McKee, and Bencie Woll. Sociolinguistic variation in British, Australian and New Zealand Sign Languages. In *Sign Languages*, pages 476–498. Cambridge University Press, 2010.
(cited p. 35)

Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *Proceedings of the 2017 IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), 2017.
(cited p. 111, 112)

Dan I Slobin, Nini Hoiting, Marlon Kuntze, Reyna Lindert, Amy Weinberg, Jennie Pyers, Michelle Anthony, Yael Biederman, and Helen Thumann. A Cognitive/Functional Perspective On The Acquisition Of "Classifiers". *Perspectives on classifier constructions in sign languages*, 2:271–296, 2003.
(cited p. 41)

C Smith, EM Lentz, and K Mikos. Vista American Sign Language series: Signing naturally, 1988.
(cited p. 86)

Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal Deep Variational Hand Pose Estimation. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
(cited p. 112)

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2627435.2670313.
(cited p. 110)

William C Stokoe. Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *Studies in Linguistics*, 8, 1960.
(cited p. 39)

William C Stokoe. Classification and Description of Sign Languages. *Current trends in linguistics*, 12: 345–371, 1972.
(cited p. 39, 46, 60)

William C Stokoe. Semantic Phonology. *Sign Language Studies*, 71(1):107–114, 1991.
(cited p. 44)

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *Proceedings of the 2018 British Machine Vision Conference (BMVC)*. British Machine Vision Association, 2018.
(cited p. 73, 75)

Ted Supalla. The Classifier System in American Sign Language. In *Noun Classification and Categorization*, volume 7, pages 181–214. Colette Grinevald Craig, 1986.
(cited p. 40, 41)

Valerie Sutton. *Lessons in sign writing*. SignWriting, 1995.
(cited p. 46, 72)

Sarah F Taub. *Language from the body: Iconicity and metaphor in American Sign Language*. Cambridge University Press, 2001.
(cited p. 41)

T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.
(cited p. 110, 126)

Els Van der Kooij. *Phonological Categories in Sign Language of the Netherlands: The Role of Phonetic Implementation and Iconicity*. Netherlands Graduate School of Linguistics, 2002.
(cited p. 44)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceeding of Advances in Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008, 2017.
(cited p. 76)

M. Vermeerbergen, L. Leeson, and O.A. Crasborn. *Simultaneity in Signed Languages: Form and Function*. Amsterdam studies in the theory and history of linguistic science. John Benjamins, 2007. ISBN 9789027247964.
(cited p. 44, 101)

Christian Vogler and Dimitris Metaxas. Adapting Hidden Markov Models for ASL Recognition by Using Three-dimensional Computer Vision Methods. In *Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 1, pages 156–161. IEEE, 1997.
(cited p. 71)

Ulrich Von Agris and Karl-Friedrich Kraiss. Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition. *Proceedings of the 2007 International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, 2007.
(cited p. 24, 59, 60, 61, 68, 69, 78)

Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. The Significance of Facial Features for Automatic Sign Language Recognition. In *Proceedings of the 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008.
(cited p. 15, 60, 61, 65, 74, 78)

Hanjie Wang, Xiujuan Chai, Xiaopeng Hong, Guoying Zhao, and Xilin Chen. Isolated Sign Language Recognition with Grassmann Covariance Matrices. *Proceedings of the 2016 ACM Transactions on Accessible Computing (TACCESS)*, 8(4):14, 2016.
(cited p. 63, 65)

Hanjie Wang, Xiujuan Chai, and Xilin Chen. A Novel Sign Language Recognition Framework Using Hierarchical Grassmann Covariance Matrix. *IEEE Transactions on Multimedia*, 21(11):2806–2814,

2019.
(cited p. 63, 65)

Jennifer Wehrmeyer. Eye-tracking Deaf and hearing viewing of sign language interpreted news broadcasts. *Journal of Eye Movement Research*, 2014.
(cited p. 45)

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016.
(cited p. 62, 108)

Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, et al. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14–30, 2014.
(cited p. 104)

Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez. Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 38(8):1583–1597, 2016.
(cited p. 63, 115)

Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
(cited p. 108, 112)

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
(cited p. 63)

Hee-Deok Yang and Seong-Whan Lee. Robust Sign Language Recognition by Combining Manual and Non-Manual Features based on Conditional Random Field and Support Vector Machine. *Pattern Recognition Letters*, 34(16):2051–2056, 2013.
(cited p. 78)

Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning Feature Pyramids for Human Pose Estimation. In *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCV)*, 2017.
(cited p. 108)

Zhaoyang Yang, Zhenmei Shi, Xiaoyong Shen, and Yu-Wing Tai. SF-Net: Structured Feature Network

for Continuous Sign Language Recognition. *arXiv preprint arXiv:1908.01341*, 2019.
(cited p. 72, 73, 74)

Polina Yanovich, Carol Neidle, and Dimitris N Metaxas. Detection of Major ASL Sign Types in Continuous Signing For ASL Recognition. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.
(cited p. 78)

Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition. In *Proceedings of the 27th DAGM Joint Pattern Recognition Symposium*, pages 401–408. Springer, 2005.
(cited p. 23, 61)

Ruiqi Zhao, Yan Wang, C Fabian Benitez-Quiroz, Yaojie Liu, and Aleix M Martinez. Fast and Precise Face Alignment and 3D Shape Reconstruction from a Single 2D Image. In *Proceedings of the 2016 European Conference on Computer Vision (ECCV)*, pages 590–603. Springer, 2016.
(cited p. 108, 109)

Hao Zhou, Wengang Zhou, and Houqiang Li. Dynamic Pseudo Label Decoding for Continuous Sign Language Recognition. In *Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1282–1287. IEEE, 2019.
(cited p. 72, 73, 74)

Christian Zimmermann and Thomas Brox. Learning to Estimate 3D Hand Pose from Single RGB Images. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. URL https://lmb.informatik.uni-freiburg.de/projects/hand3d/.
(cited p. 112)

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : De la reconnaissance de signes à la compréhension automatique de langue des signes : une prise en compte des unités non conventionnalisées

Mots clés : Reconnaissance de langue des signes, Langue des signes continue, Iconicité, Linguistique des langues des signes, Représentation du signeur, Réseaux de neurones récurrents

Résumé : Les langues des signes (LS) se sont développées naturellement au sein des communautés de Sourds. Elles intègrent des signes lexicaux, c'est-à-dire des unités conventionnalisées du langage, mais aussi des structures iconiques, i.e. où forme et sens du message sont liés. La plupart des travaux de recherche en reconnaissance automatique de LS se sont pourtant attelés à reconnaitre les signes lexicaux, utilisant des corpus artificiels, parfois en LS interprétée.

Dans cette thèse, nous souhaitons élargir cette perspective pour envisager la reconnaissance d'éléments utilisés pour la construction du discours ou au sein de structures illustratives.

Les corpus de linguistes sont pour cela intéressants car la langue y est naturelle et les annotations détaillées, cependant pas toujours cohérentes. Nous proposons donc la refonte d'un corpus de dialogue en langue des signes française, Dicta-Sign-LSF-v2, annoté de manière détaillée et cohérente. Nous redéfinissons alors le problème de la reconnaissance automatique de LS comme la reconnaissance de divers descripteurs linguistiques, avec des métriques adaptées. Nous développons par ailleurs une représentation compacte et généralisable des signeurs dans les vidéos par un traitement parallèle des mains, du visage et du haut du corps, puis une architecture d'apprentissage adaptée consistant en un réseau de neurones récurrent et convolutionnel.

Nous montrons enfin l'effectivité du modèle proposé, d'abord via une analyse approfondie du paramétrage du modèle et de la représentation des signeurs, puis par l'étude détaillée de prédictions pour la reconnaissance de quatre descripteurs linguistiques. Cette étude, avec des résultats très encourageants, montre le bien-fondé de l'approche proposée et ouvre la voie à une acception plus large de la reconnaissance continue de langue des signes.

Title: From Sign Recognition to Automatic Sign Language Understanding: Addressing the Non-Conventionalized Units

Keywords: Sign Language Recognition, Continuous Sign Language, Iconicity, Sign Language Linguistics, Signer Representation, Recurrent Neural Networks

Abstract: Sign Languages (SLs) have developed naturally in Deaf communities. They make use of lexical signs, i.e. conventionalized units of language, but also iconic structures, i.e. when the form of an utterance and the meaning it carries are related. Most research in automatic Sign Language Recognition (SLR) has however focused on recognizing lexical signs, using corpora that are artificial, sometimes made of interpreted SL.

In this thesis, we wish to broaden this perspective and consider the recognition of elements used for the construction of discourse or within illustrative structures.

To this end, corpora developed by linguists are valuable as the language is natural and the annotations are detailed, however not necessarily complete or coherent. We thus propose the redesign of a French Sign Language dialogue corpus, Dicta-Sign-LSF-v2, finely and consistently annotated. We then redefine the problem of automatic SLR as the recognition of various linguistic descriptors, with adapted performance metrics. Moreover, we develop a compact and generalizable representation of signers in videos by parallel processing of the hands, face and upper body, then an adapted learning architecture based on a recurrent and convolutional neural network.

Finally, we show the effectiveness of the proposed model, first through an in-depth analysis of the parameterization of the learning model and the representation of the signers, then by studying in detail predictions for the recognition of four linguistic descriptors. This study, with very encouraging results, shows the soundness of the proposed approach and paves the way for a wider understanding of continuous Sign Language Recognition.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France