Création semi-automatique d'un lexique bilingue langue des signes française (LSF) / français pour l'annotation de vidéos de LSF

Julie Lascar, Annelies Braffort, Michèle Gouiffès

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Création semi-automatique d'un lexique bilingue langue des signes française (LSF) / français pour l'annotation de vidéos de LSF

Julie Lascar Annelies Braffort Michèle Gouiffès
Université Paris-Saclay, CNRS, LISN
Campus Universitaire Bat 507, rue du Belvédère, 91405 Orsay, France
prenom.nom@lisn.upsaclay.fr

RÉSUMÉ

Cet article présente nos contributions sur la constitution de ressources et le traitement automatique au service de l'analyse de vidéos de langue des signes française (LSF) et plus particulièrement de l'annotation automatique des unités lexicales. À ce jour, nous avons pu constituer de manière semi-automatique un lexique bilingue comportant 88 entrées et utilisé pour élaborer un classificateur qui a permis d'annoter automatiquement des unités lexicales sur le corpus Mediapi-rgb.

ABSTRACT

Semi-automatic creation of a bilingual French sign language (LSF) / French lexicon for annotating LSF videos

This article presents our contributions to the constitution of resources and automatic processing for the analysis of French sign language (LSF) videos, and more specifically the automatic annotation of lexical units. To date, we have been able to build a bilingual lexicon containing 88 entries in a weakly supervised manner. This lexicon has been used to build a classifier, used to automaticaly annotated lexical units on the Mediapi-rgb corpus.

MOTS-CLÉS : Langue des signes française, lexique bilingue, annotation automatique.

KEYWORDS: French sign language, bilingual lexicon, automatic annotation.

1 Introduction

Les Langues des Signes (LS) sont des langues naturelles pratiquées au sein des communautés de Sourds et la Langue des Signes Française (LSF) est celle pratiquée en France. À ce jour, les LS sont encore très peu dotées et les recherches sur ces langues sont assez récentes, en particulier dans le domaine du traitement automatique. Nos projets actuels ont pour objectif de contribuer à la constitution de ressources et de traitements automatiques au service de l'analyse de vidéos de LSF. Cet article présente un point d'étape sur nos contributions autour de l'annotation automatique.

FIGURE 1 – Capture d'écran du site Média'Pi !

2 Constitution d'un lexique bilingue

Il n'existe pas à l'heure actuelle de lexique bilingue en contexte exploitable pour le traitement automatique de la LSF. D'une manière générale, les ressources de LSF utilisables pour le traitement automatique sont peu nombreuses (Braffort, 2022). Une précédente étude (Bull, 2023) a permis de constituer un jeu de données utilisable pour le traitement automatique. Il a été constitué à partir du corpus Mediapi-rgb, comportant 86h de vidéos en LSF produites par des journalistes ou présentateurs sourds du média bilingue en ligne Média'Pi ! [1], accompagnées de sous-titrage en français (figure 1).

À partir de ce jeu de données, nous avons constitué un premier lexique bilingue avec une approche faiblement supervisée :

- nous avons sélectionné une liste de mots présents dans les sous-titres de telle sorte que : 1) à part quelques exceptions, ils possèdent peu, ou pas de synonymes et 2) leur équivalent en LSF varie peu en fonction du contexte. Nous avons ainsi opté pour les jours, les mois, certaines villes et pays ainsi que du vocabulaire lié à l'actualité de l'époque (masque, chômage, gilet jaune...) ;
- pour chaque mot, nous avons sélectionné toutes les vidéos pour lesquelles le sous-titre contient ce mot, puis nous avons calculé la similarité entre paires de vidéos (méthode de la similarité cosinus) ;
- dans chaque vidéo, on retient les numéros d'images pour lesquelles la similarité dépasse un certain seuil (empiriquement fixé à 0.6) sur un nombre consécutif d'au moins 4 images.

1. https ://www.media-pi.fr/

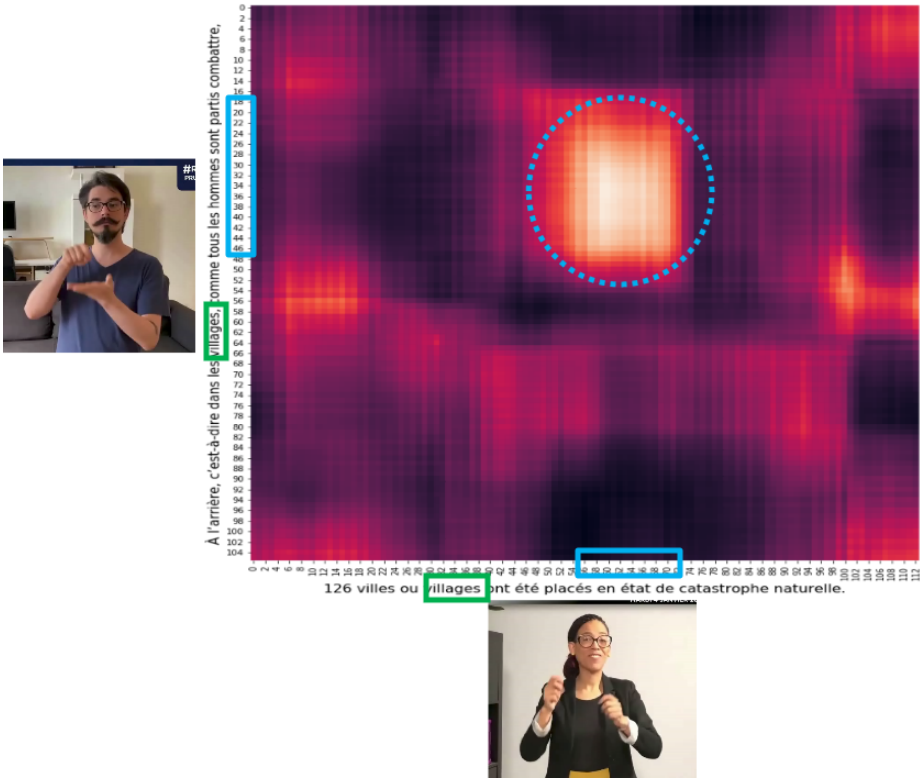

FIGURE 2 – Exemple de matrice de similarité entre deux vidéos pour le mot « village ».

La figure 2 illustre une matrice de similarité entre deux vidéos dont les sous-titres comportent le mot « village ». La zone plus claire est celle pour laquelle la similarité est la plus forte.

Certains signes peuvent présenter des variantes de forme qui dépendent du locuteur. Par exemple quelques signes représentant les mois sont réalisés par des locuteurs avec une main et par d'autres avec les deux mains. Pour distinguer ces variantes, nous avons procédé à un regroupement automatique des vidéos par locuteur. Par ailleurs, certains mots peuvent présenter des variantes de sens, comme par exemple le mot « place » qui est signé différemment selon son sens (ex : mettre en place, quatrième place, place de la Concorde). Nous avons utilisé un modèle de language Bert (Devlin *et al.*, 2019) pour regrouper les mots selon leur sens dans le contexte de la phrase.

Une dernière étape a consisté à regrouper les vidéos capturées pour chaque mot en utilisant une méthode de partitionnement des données (algorithme des k-moyennes). Cela a permis d'une part, d'éliminer les erreurs de détection et d'autre part, de détecter certaines variantes, comme illustré sur la figure 3.

Dans ces figures, après réduction de la dimension des données par analyse par composantes principales (ACP), on a projeté sur les deux axes principaux les séquences obtenues pour les signes ITALIE et NOVEMBRE. Dans la figure de gauche, on obtient deux groupes : le plus grand regroupe les vidéos correspondant effectivement au signe ITALIE, le plus petit regroupe les erreurs de détection. Dans la figure de droite, on obtient aussi deux groupes

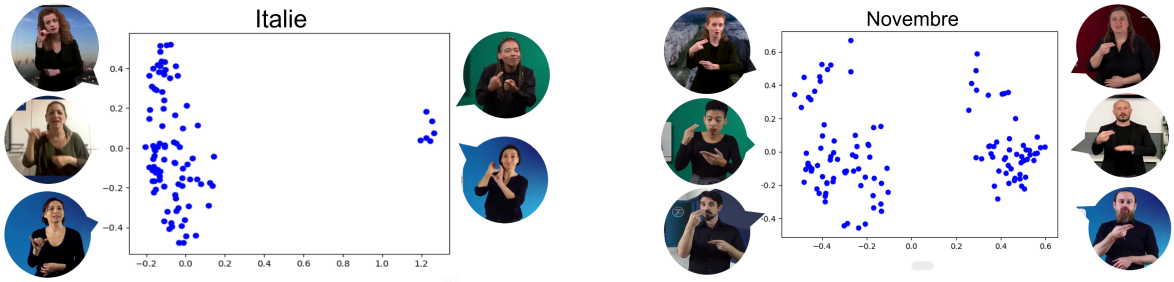

FIGURE 3 – Groupes obtenus pour les signes ITALIE et NOVEMBRE (après réduction de la dimension des données par ACP).

qui permettent de différencier deux variantes pour le signe NOVEMBRE : celui de gauche regroupe les signes effectués à deux mains et celui de droite regroupe les signes effectués à une seule main.

Nous obtenons ainsi des classes associant forme et sens distinctes et présentant peu d'erreurs. Une dernière phase de contrôle réalisée par des experts est en cours et va permettre de ne garder que des occurences correctes et bien segmentées, c'est-à-dire pour lesquelles la séquence d'images comporte l'ensemble du signe et aucun élément de coarticulation avec les signes précédent et suivant.

3 Constitution d'un classifieur basé sur ce lexique

La constitution de ce lexique bilingue a permis d'annoter une partie des données d'entraînement du corpus Mediapi-rgb. Ces données annotées ont ensuite été utilisées pour entraîner un classifieur, qui prend en entrée les vidéos sous forme de séquences d'images et produit en sortie des séquences d'entiers, chaque entier identifiant pour chaque image la classe correspondante. L'architecture du système, illustrée figure 4, comporte deux modèles :

- Le premier a pour rôle d'encoder les vidéos sous forme de plongements.
- Le second est le classifieur qui prend en entrée ces plongements .

Des séries d'expériences ont été menées en faisant varier les modèles utilisés. Pour l'étape de calcul des plongements, trois modèles simples ont été comparés, tous adaptés au fait que les données sont de nature temporelle : un modèle de type *transformers* entraîné sur des données d'actions humaines (*Video Swin Transformer* (Liu *et al.*, 2022)), un modèle à base de réseaux de convolution 3D (I3D) réentraîné sur un corpus de BSL (langue des signes britanique) (Renz *et al.*, 2021) et de nouveau le modèle *Video Swin Transformer* mais cette fois-ci réentrainé sur des vidéos de BSL (Prajwal *et al.*, 2022). Le troisième modèle s'avère le plus performant et c'est celui que nous avons conservé pour la suite.

Pour l'étape de classification plusieurs modèles ont été comparés (des perceptrons et des LSTM, à une ou deux couches). Nous avons optimisé les paramètres en utilisant le score F1, qui est la moyenne harmonique entre la précision et le rappel. Plus précisément, chaque

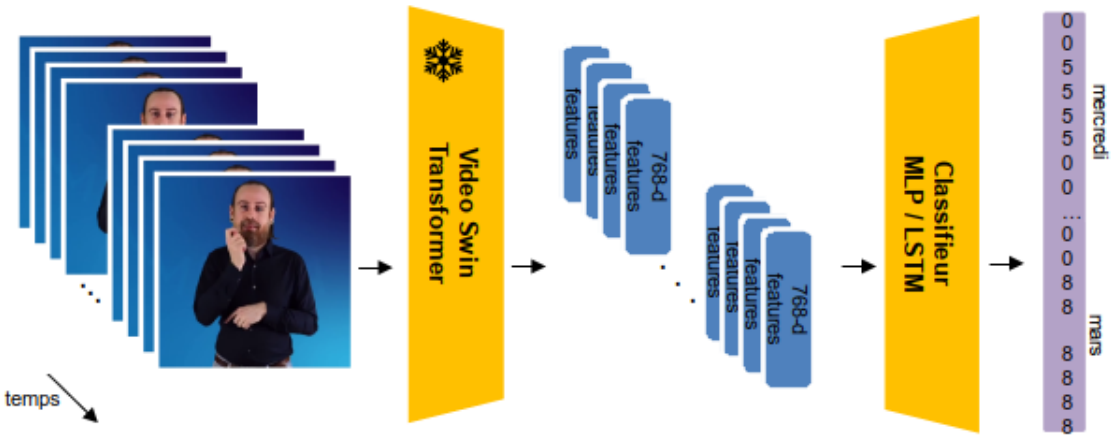

FIGURE 4 – Architecture du classifieur.

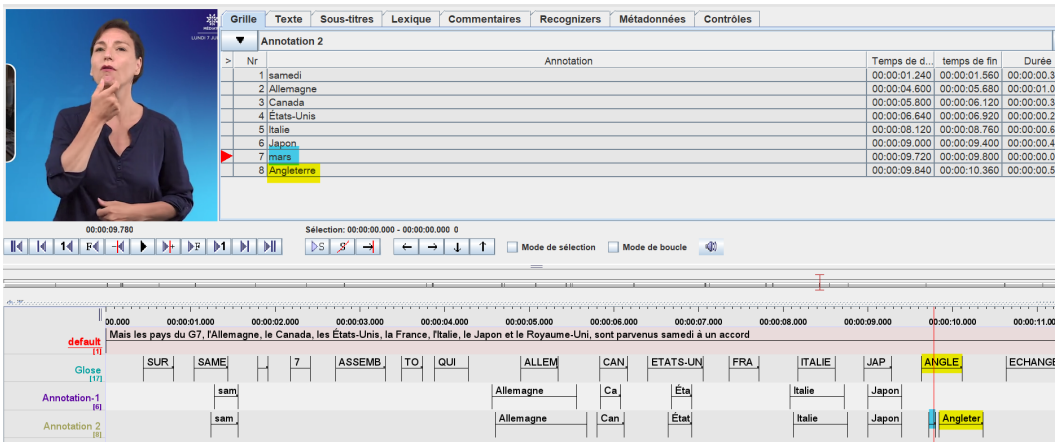

FIGURE 5 – Annotation par des experts avec Elan.

classe se voit attribuer un score F1 (voir l'annexe pour le détail du calcul). Le score final correspond à la moyenne des scores F1 de chaque classe. Nous avons retenu une architecture LSTM à une couche qui donne le meilleur score (score F1 : 0.79).

Ce classifieur a ensuite été utilisé pour annoter de nouveau les données d'entraînement. Ainsi, 5 876 vidéos et 7 259 signes ont été annotés automatiquement. Le score F1 est insuffisant pour évaluer la qualité des annotations, en particulier sur la qualité de la segmentation temporelle. Pour cette étape aussi, une analyse fine des résultats par des experts est en cours. La figure 5 montre un exemple d'annotation à l'aide du logiciel Elan (Crasborn & Sloetjes, 2008) des unités lexicales sur une vidéo en LSF accompagnée du sous-titrage « Mais les pays du G7, l'Allemagne, le Canada, les Etats-Unis, la France, l'italie, le Japon et le Royaume-Uni, sont parvenus samedi à un accord ».

L'annotation comporte trois pistes : *Glose* est renseignée manuellement par un expert, tandis que *Annotation-1* et *Annotation-2* sont renseignées automatiquement, respectivement suite à l'étape 1 basée sur le calcul de similarité et à l'étape 2 à l'aide du classifieur.

Sur cet exemple, l'ensemble des signes présents dans le lexique ont été détectés à l'exception du signe FRANCE et du signe ANGLETERRE dans l'étape 1. L'étape 2 améliore les résultats car le signe ANGLETERRE a été détecté. Cependant un segment glosé MARS a été inséré. Il se trouve que les signes MARS et ANGLETERRE ont des formes proches (configuration et emplacement de la main). Enfin, globalement, les annotations automatiques ont bien positionné les segments dans le flux temporel mais ne sont pas très précis. L'expertise en cours sur la qualité des occurrences du lexique bilingue (construites lors de l'étape 1) montre qu'un grand nombre d'occurrences sont incomplètes ou incluent des parties des coarticulations avec les signes précédent et suivant. L'amélioration de la qualité du lexique bilingue devrait significativement améliorer les résultats du classifieur, y compris sur l'aspect repérage temporel.

4 Bilan et perspectives

Il s'agit des tous premiers pas vers un système d'annotation automatique d'unités lexicales dans des vidéos de LSF et beaucoup reste à faire sur l'exploitation des sous-titres, la représentation des vidéos, l'architecture du classifieur et l'évaluation des performances du système. Ces différents axes sont actuellement en chantier, ainsi que l'augmentation du vocabulaire, qui dépasse maintenant les 1000 classes. Nous envisageons aussi d'exploiter les annotations fines réalisées par des experts sur le corpus Dicta-Sign-LSF-V2 (Belissen *et al.*, 2020), afin de récupérer des occurrences de bonne qualité et ajouter de nouvelles classes. Nous envisageons aussi d'évaluer le classifieur sur ce corpus.

Par la suite, ce travail doit être étendu à d'autres unités gestuelles telles que les structures illustratives très fréquentes dans les LS et qui ne peuvent pas être listées dans un dictionnaire, ce qui implique de concevoir une approche différente.

Références

BELISSEN V., BRAFFORT A. & GOUIFFÈS M. (2020). Dicta-Sign-LSF-v2 : Remake of a continuous French Sign Language dialogue corpus and a first baseline for automatic sign language processing. In *Language Resources and Evaluation Conference*, p. 6040–6048, Marseille, FR : ELRA.

BRAFFORT A. (2022). Langue des Signes Française : Etat des lieux des ressources linguistiques et des traitements automatiques. In *Journées Jointes des GRD LIFT et TAL*, p. 131–138, Marseille, FR : CNRS.

BULL H. (2023). *Learning sign language from subtitles*. Thèse de doctorat. Université Paris-Saclay.

CRASBORN O. & SLOETJES H. (2008). Enhanced elan functionality for sign language corpora. In *Work. on the Representation and Processing of Sign Languages @ LREC Conf.*, p. 39–43, Marrakech, Morocco.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Conf. of the NAACL association : Human Language Technologies*, p. 4171–4186 : ACL.

LIU Z., NING J., CAO Y., WEI Y., ZHANG Z., LIN S. & HU H. (2022). Video swin transformer. In *Conf. on Computer Vision and Pattern Recognition*, p. 3202–3211, New Orleans, USA : IEEE.

PRAJWAL K. R., BULL H., MOMENI L., ALBANIE S., VAROL G. & ZISSERMAN A. (2022). Weakly-supervised fingerspelling recognition in british sign language videos. In *British Machine Vision Conference*, London, UK : BMVA.

RENZ K., STACHE N. C., ALBANIE S. & VAROL G. (2021). Sign language segmentation with temporal convolutional networks. In *Conf. on Acoustics, Speech and Signal Processing*, p. 2135–2139, Toronto, Canada : IEEE.

Annexe : calcul du score F1 d'une classe

On obtient le score F1 d'une classe c de la manière suivante :

a. On calcule le score F1 pour chaque vidéo dans laquelle la classe c est présente en comparant les prédictions du modèle (Pred) aux annotations effectuées lors de l'étape 1 (GT).

$$\text{Vidéo 1} \rightarrow \begin{cases} \text{GT : 000111111000...} \\ \text{Pred : 011111000000...} \end{cases} \qquad \text{F1} = 0.55$$

$$\text{Vidéo 2} \rightarrow \begin{cases} \text{GT : 000111111000...} \\ \text{Pred : 000001110000...} \end{cases} \qquad \text{F1} = 0.75$$

...

b. On calcule la moyenne des scores F1 de chaque vidéo.