Corpus bilingue sous-titrage et Langue des Signes Française : la problématique de l'alignement automatique des données

Julie Halbout[1] Diandra Fabre[2]
(1) Univ. Paris-Saclay, CNRS, LISN, Orsay, France
(2) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France
(1) firstname.lastname@lisn.fr
(2) firstname.lastname@univ-grenoble-alpes.fr

RÉSUMÉ

Dans cet article, nous présentons une étude sur la problématique de l'alignement automatique des données dans un corpus constitué de discours en français parlé, sous-titrés en français écrit et interprétés en langue des signes française (LSF). Après une introduction précisant le processus bien particulier de l'interprétation en langue des signes, nous dressons un tour d'horizon des ensembles de données existants pour la LSF ainsi que les spécificités du corpus Matignon-LSF, constitué à partir des comptes-rendus vidéos hebdomadaires du conseil des ministres. Nous montrons ensuite sur quelques exemples certains des phénomènes observés sur la problématique de l'alignement temporel entre les sous-titres synchronisés avec l'audio, et la LSF interprétée qui subit un décalage temporel. Nous en concluons que le niveau d'alignement ne peut pas être celui des phrases en français écrit et proposons quelques pistes pour la suite.

ABSTRACT

Bilingual subtitling and French sign language corpus : the problem of automatic data alignment

In this article, we present a study on the problem of automatic data alignment in a corpus consisting of spoken French discourse, subtitled in written French and interpreted into French sign language (LSF). After an introduction outlining the specific process of sign language interpreting, we provide an overview of existing datasets for LSF and the specific features of the Matignon-LSF corpus, based on the weekly video reports of the Council of Ministers. We then use a few examples to illustrate some of the phenomena observed with regard to the problem of temporal alignment between subtitles synchronized with the audio, and the interpreted LSF which is subject to a time lag. We conclude that the level of alignment cannot be this of French written sentences, and propose a few avenues for further research.

MOTS-CLÉS : Langue de Signes Française, LSF, corpus, interprétation, alignement.

KEYWORDS: French Sign Language, LSF, dataset, interpretation, alignment.

1 Introduction

Bien que le traitement automatique des langues des signes (LS) soit un domaine en pleine expansion, la grande majorité de ces langues sont encore peu dotées en termes de corpus disponibles pour la

recherche. La Langue des Signes Française (LSF) n'y fait pas exception. Une source potentielle de données en LS est la télévision (Koller *et al.*, 2015; Albanie *et al.*, 2021), où le nombre d'émissions interprétées a connu une augmentation significative ces dernières années. Cependant, l'accès à ces données n'est généralement pas aisé, soit pour des raisons légales soit à cause de simples considérations techniques. En France, les compte-rendus hebdomadaires du Conseil des Ministres produisent des vidéos qui sont systématiquement interprétées en LSF (fig. 1), en licence ouverte [1].

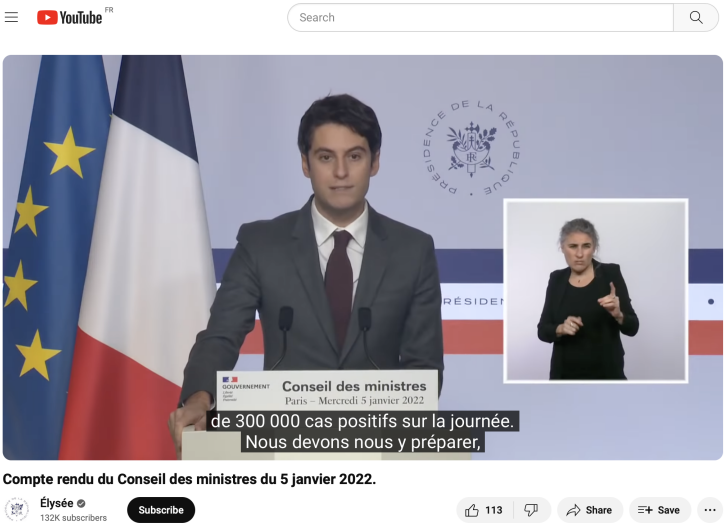


FIGURE 1 – Capture d'écran d'une vidéo montrant un compte-rendu télévisés du Conseil des Ministres du gouvernement français interprétés en LSF.

La principale modalité linguistique des programmes télévisés est la parole. Celle-ci peut être sous-titrée, soit automatiquement avec toutes les erreurs potentielles que cela comporte, soit dans un studio de sous-titrage en direct avec des contraintes de temps et de format, ou en temps différé (Buet, 2022). Le discours oral peut également être traduit ou interpreté en LS. Dans ce cas, la situation est particulière et diffère selon qu'il s'agit de traduction ou d'interprétation.

La traduction consiste à traduire un texte d'une langue source vers la langue maternelle du traducteur. Le traducteur dispose de toute la latitude et le temps nécessaire pour faire des recherches et revenir sur son travail. Dans le cas d'une traduction vers les LS, celle-ci est le plus souvent réalisée par des traducteurs sourds dont la LS est la langue première.

L'interprétation consiste à transposer en simultané (ou presque) un discours oral (parlé ou signé) d'une langue source à une langue cible. Il peut s'agir d'une interprétation d'une LS source vers une LS cible et, dans ce cas, les interprètes sont généralement des personnes sourdes dont la LS cible est la langue première. Dans le cas de la traduction d'une langue parlée vers une langue signée, il est nécessaire que l'interprète entende l'orateur. Par conséquent, à moins que l'interprète ne soit un CODA (*Child Of Deaf Adult* c'est-à-dire enfant d'adulte sourd), il/elle n'interprète pas vers sa langue maternelle [2]. Il a été noté qu'il existe des différences entre la production des interprètes entendants

1. https ://www.etalab.gouv.fr/licence-ouverte-open-licence/
2. A noter que depuis quelques années, des co-interprétations sont mises en place : il s'agit d'un binôme constitué d'un

et celle des interprètes sourds (Stone & Russell, 2011). D'une manière générale, les personnes qui signent couramment peuvent faire la différence entre un discours interprété et un discours direct, non interprété, ainsi qu'entre un discours signé par un sourd de naissance et un autre signé par un entendant. Notons enfin qu'en raison du processus d'interprétation, la langue source peut interférer avec la langue cible. Ainsi, le discours interprété peut s'avérer très éloigné de celui qu'aurait directement produit un signeur (dans une version originale), à commencer par le fait que la LS interprétée tend à suivre de près la structure grammaticale de la langue parlée (Dayter, 2019). Cependant, peu de travaux ont été consacrés à la description ou à la quantification de ces différences.

Cette étude porte sur un jeu de données nommé *Matignon-LSF* [3], constitué à partir des vidéos des comptes-rendus du Conseil des Ministres. Il s'agit donc de discours en LSF interprétée, comportant tous les biais notés plus haut. Cependant il peut constituer une source de données très intéressante pour la recherche en linguistique sur l'étude du discours interprété en LSF, mais aussi pour le traitement automatique des LS. Il présente également l'avantage d'être en données ouvertes. Il s'agit cependant de données langagières de nature très différentes qui, de plus, ne sont pas synchronisées a priori. La question abordée dans cet article concerne le problème, spécifique à ce type de corpus, de l'alignement temporel entre les données de langue parlée et de langue signée.

Ainsi, après un bref aperçu des corpus actuellement disponibles en particulier en LSF (section 2), les sections 3 et 4 reprennent succintement la présentation du jeu de données, la collecte et le traitement des données, et les sections suivantes présentent la problématique de l'alignement entre les contenus en langue parlée et en LS interprétée (section 5), les premières expérimentations menées sur ce sujet (section 6), puis une discussion sur les observations réalisées (section 7).

2 Jeux de données en LSF

Dans le cadre du récent projet européen Easier [4], le rapport (Kopf *et al.*, 2023) a permis de dresser une vue d'ensemble des jeux de données existants pour les Langues des Signes européennes. Ces ensembles de données ont été divisés en deux catégories : les corpus réalisés en laboratoire et les données de radiodiffusion. Les premiers offrent des données de haute qualité avec des transcriptions et des annotations riches, tandis que les secondes sont disponibles en grandes quantités, mais généralement sans métadonnées ni expertise linguistique. Depuis la publication de ce rapport, d'autres jeux de données ont été publiés, tels que BSL-1K (Albanie *et al.*, 2020) et plus récemment BOBSL en langue des signes britannique (BSL) (Albanie *et al.*, 2021), qui représente un changement d'échelle en termes de quantité, fournissant aux chercheurs plus de 1200 heures de LS interprétée à partir d'émissions de la BBC. Dans le même ordre d'idées, les données YouTube-ASL (Uthus *et al.*, 2024) en LS américaine (ASL) totalise près de 1000 heures de vidéos extraites du web. Toujours en ASL, le corpus How2Sign, publié en 2023, est particulièrement intéressant, car il s'agit du plus grand corpus de laboratoire de LS originales (non interprétées ou traduites). Il a déjà fait l'objet de plusieurs travaux (Duarte *et al.*, 2021).

La LSF a fait l'objet de plusieurs corpus durant les 10 dernières années (Braffort, 2022). La plupart d'entre eux ont été construits en laboratoire, souvent à des fins d'analyse linguistique et sont de très

interprète entendant interprétant depuis la parole vers une LS et d'un interprète sourd réinterprétant depuis la LS de l'interprète entendant vers une LS plus naturelle.

3. *Matignon* désigne la résidence officielle du Premier ministre français et s'étend au gouvernement français.
4. https ://www.project-easier.eu/

petite taille (moins de 4 heures). Les corpus plus conséquents comme Creagest (Balvet *et al.*, 2010) ne sont que partiellement annotés.

Le corpus DictaSign-LSF-v2, contenant 8 heures de dialogues (Belissen *et al.*, 2020a) partiellement annotés, s'avère pertinent pour la reconnaissance de signes en contexte, que ce soient des signes lexicaux (Ouakrim *et al.*, 2023) ou des structures linguistiques (Belissen *et al.*, 2020b). Pour pallier le manque de données, deux corpus en LSF ont récemment été rendus disponibles : Mediapi-Skel (Bull *et al.*, 2020a) et Mediapi-RGB (Ouakrim *et al.*, 2024). Ce dernier est constitué de vidéos issues du media bilingue Médiapi [5] produites en LSF par des journalistes et des présentateurs sourds et accompagnées de sous-titres en français. Ces corpus ont été préparés pour l'analyse automatique : les sous-titres sont alignés avec les vidéos en LSF, la traduction en français ayant été produite à partir des vidéos en LSF et non l'inverse. En raison du modèle économique de ce média, seules 86h de vidéos correspondant à des contenus dits « froids »sont mises à disposition.

Le jeu de données Matignon-LSF a été constitué dans l'objectif de collecter un ensemble de données de LSF qui soit à la fois large et ouvert. Il est décrit synthétiquement dans les sections suivantes.

3 Présentation succinte du corpus Matignon-LSF

Les Conseils des ministres [6] du gouvernement français ont lieu une fois par semaine à l'Elysée. Ils sont suivis par un compte-rendu prononcé par le porte-parole du gouvernement qui est filmé, sous-titré et, depuis juillet 2020, interprété en LSF. Le jeu de données Matignon-LSF est basé sur les interprétations et sous-titres en LSF de ces compte-rendus. % Il n'est pas possible d'obtenir d'informations sur le processus de travail des interprètes, tenus au secret. Toutefois, nous supposons qu'ils disposent de certains éléments pour préparer leur interprétation. À ce jour, le corpus comprend 67 vidéos. La figure 2 montre les 20 noms les plus fréquents de l'ensemble de données, révèlent ainsi un discours centré sur la politique française (cinq premiers mots : *ministre*, *question*, *mesure*, *français* et *président*).

59 vidéos se composent du discours du porte-parole du gouvernement (qui varie de 4 minutes à 20 minutes, avec une moyenne d'environ 12 minutes), suivi d'une session de questions-réponses avec les journalistes. Cette partie peut varier en fonction des sujets et du nombre de journalistes présents dans la salle de presse (de 8 minutes à près d'une heure, avec une moyenne de 23 minutes). Dans cinq autres vidéos, les ministres sont invités à présenter leur point de vue après l'intervention du porte-parole, et des questions leur sont posées en plus de celles du porte-parole. Dans les trois vidéos restantes, la conférence de presse se déroule sans porte-parole et les ministres prononcent directement leur discours, avec une session de questions-réponses plus courte. Les 67 vidéos diffusées ont une durée totale de 39 heures, avec une durée moyenne de 36 minutes. La répartition de la durée des vidéos est présentée dans la Figure 3(a).

Les sous-titres (en français écrit) de l'ensemble de données sont composés d'un total de 447k tokens [7] pour un vocabulaire de 10k. À partir des sous-titres, 18k phrases ont été extraites, comme décrit dans la section 4. Matignon-LSF comprend 15 interprètes. Les caractéristiques du jeu de données sont résumées dans le tableau 3. À ce jour, le corpus Matignon-LSF se situe entre Mediapi-Skel et Mediapi-RGB en termes de taille.

5. https ://www.media-pi.fr/
6. https ://www.gouvernement.fr/conseil-des-ministres
7. tokenisation réalisée avec SpaCy `https://spacy.io/`

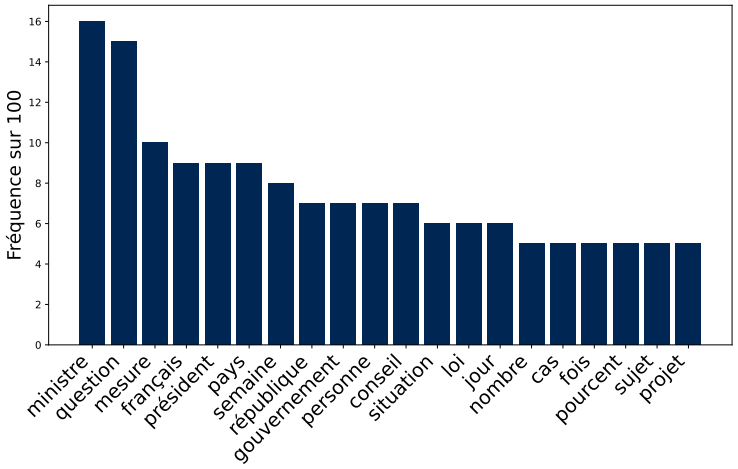


FIGURE 2 – Les 20 noms les plus fréquents dans les sous-titres de Matignon-LSF.

4 Collecte et traitement des données

Chaque semaine, le compte-rendu du Conseil des Ministre est filmé et déposé sur Youtube [8] et/ou Dailymotion [9] et accompagné d'un ensemble correspondant de sous-titres écrits en français alignés sur l'audio. Les vidéos originales ont une résolution de 1080 px et une fréquence d'acquisition de 30 fps. Les différents processus conçus pour automatiser la collecte et la mise en forme des vidéos et des sous-titres sont documentés dans un dépôt GitHub organisé comme une boîte à outils pour permettre la reproduction et l'expansion du corpus, car de nouveaux communiqués de presse ont lieu une fois par semaine. Nous présentons ces processus succinctement.

Vidéos. Les vidéos sont recadrées de manière à ne conserver que l'encadré contenant l'interprète, normalisé aux dimensions 494×494 px. Ces vidéos en LSF sont accompagnées de l'audio et des sous-titres français associés.

Prétraitements des vidéos. Certaines méthodes (Tarrés *et al.*, 2023; Renz *et al.*, 2021) s'appuient sur des représentations denses générées par apprentissage, tel que le modèle I3D (Carreira & Zisserman, 2017) pour représenter les données vidéos. C'est cette dernière architecture qui a été retenue pour extraire des caractéristiques des vidéos, plus précisément le modèle spécialisé fourni par Varol *et al.* (2021) et entrainé sur un corpus de LS britannique (BSL).

Prétraitement des sous-titres. Les sous-titres étant limités en longueur pour des raisons d'affichage, ils ne forment pas nécessairement des phrases. Cependant, les tâches de traduction opèrent

8. https ://www.youtube.com/@ELYSEE
9. https ://www.dailymotion.com/gouvernementFR

Durée totale (h)	39
#vidéos	67
#sous-titres	51131
#phrases	18000
#mots français	10000
#signeurs	15
#porte-paroles	3*
Résolution des vidéos (px)	494×494
Fréquence d'acquisition (fps)	30

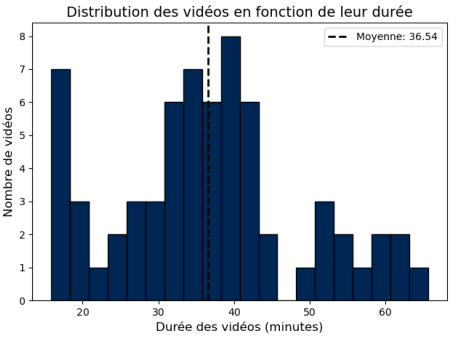
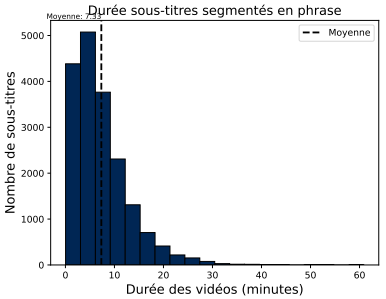
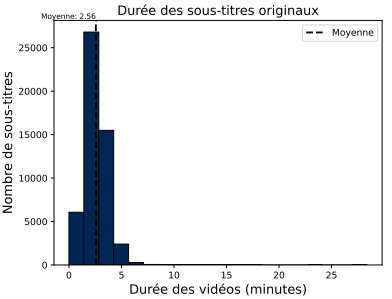


FIGURE 3 – Présentation du corpus (Les journalistes et les ministres ne sont pas inclus) et distribution des durées des vidéos.



(a)



(b)

FIGURE 4 – Distribution des durées des sous-titres (a) avant extraction des phrases et (b) après.

souvent au niveau de la phrase. C'est pourquoi une segmentation au niveau de la phrase a été produite à partir des sous-titres, en se basant sur l'approche de Albanie *et al.* (2021). Cela nous permet également d'avoir des données comparables au corpus BOBSL pour nos expériences sur l'alignement 6.

La disparité de la durée des sous-titres entre les sous-titres originaux et les sous-titres segmentés en phrases est illustrée dans les figures 4. La durée moyenne passe ainsi de 2,56 à 7,33 secondes.

5 La problématique de l'alignement temporel sous-titres/LSF interprétée

Le corpus Matignon-LSF est destiné à la recherche en traitement automatique des langues des signes, en particulier à la traduction. Ce type de tâche nécessite des prétraitements de segmentation des sous-titres et des vidéos et la mise en correspondance temporelle de ces segments afin d'obtenir des données bilingues alignées. Les segments peuvent correspondre à des signes (unités lexicales par exemples), à des portions d'énoncés, ou à des discours complets. Cependant, produire un tel alignement à partir d'un ensemble de données en LS interprétées est un véritable défi. En effet, les sous-titres sont alignés sur l'audio (la parole), alors que l'interprétation en LS est effectuée avec une

latence qui varie dans le temps, mais aussi d'un/une interprète à l'autre. Ainsi l'utilisation de ce type de corpus vient avec de nouvelles questions de recherche sur l'alignement entre données de parole, sous forme audio ou textuelles et données de LS, sous forme de vidéo. C'est sur cette problématique que porte notre étude. Nous présentons ici nos premières observations et expérimentations sur ce sujet. Les exemples présentés dans cette section sont extraits d'une même vidéo [10] et sont signés par la même interprète, mais les observations relevées ici s'appliquent à l'ensemble de données. Dans cette étude nous avons sélectionné quelques phrases dans les sous-titres et observé les positionnements temporels des segments correspondants aux vidéos associées.

La figure 5 illustre le phénomène de décalage temporel, avec un premier exemple composé de deux phrases consécutives : « Un cap pour contrôler l'épidémie (en vert). », « Un cap pour relancer notre pays (en gris). ». Le segment temporel correspondant à l'affichage des sous-titres dans la vidéo est noté *Sous-titres*. Le segment temporel correspondant à la LSF interprétée correspondante, repéré manuellement par un expert, est noté *LSF interprétée*. On observe d'une part, un décalage temporel des débuts de segment de 3.6 secondes et d'autre part, que la durée des deux énoncés signées (4,64 secondes) est plus longue que celle des deux énoncés parlées (3,9 secondes). Par conséquent, nous pouvons déjà noter qu'un simple décalage du segment des Sous-titres de manière à aligner son début avec celui de la LSF interprétée n'est pas suffisant car la fin des segments n'est pas alignée : segment *Décalage automatique* dans la figure.

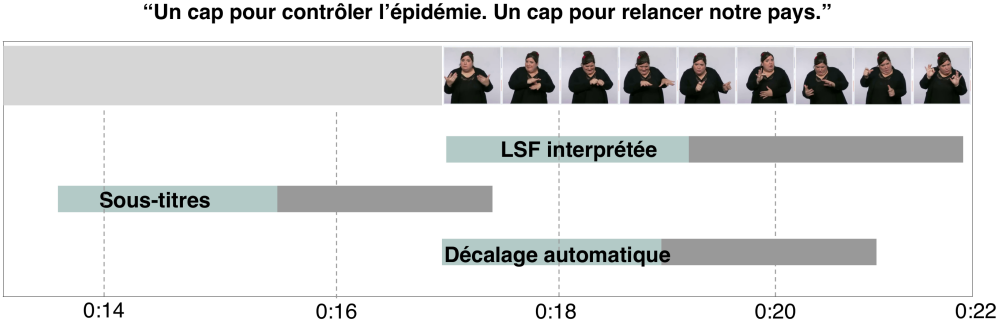


FIGURE 5 – Exemple 1 : Deux énoncés consécutifs, en vert et en gris. Segments temporels de la LSF interprétée, des sous-titres, du décalage des sous-titres pour aligner le début avec la LSF.

L'alignement manuel nécessite un temps considérable, comme l'explique (Bull *et al.*, 2020b). Il faut environ 10 à 15 heures à un expert en LS pour synchroniser les sous-titres avec une heure de vidéo continue en LS. Nous avons donc choisi d'explorer une méthode d'alignement automatique, décrite dans la section suivante.

6 Expérience d'alignement automatique

La tâche d'alignement consiste à trouver pour un segment dans une langue source un segment de sens équivalent dans la langue cible. Elle se distingue des tâches de reconnaissance de signes en continu et

10. ID : 0IViLqFmgIY

de localisation temporelle de séquences de signes comme expliqué dans (Bull *et al.*, 2021). Entre autres, la reconnaissance de signes continus présuppose que les contenus sont déjà segmentés, et la localisation temporelle de signes ne se fait pas à l'échelle d'une unité de sous-titre. Dans notre cas, nous cherchons à aligner des segments de texte (les sous-titres) avec des segments de vidéos (la LSF interprétée). Les sous-titres sont segmentés en « unités de sous-titres »à partir desquels nous pouvons facilement reconstituer des phrases. Par contre, la vidéo n'est pas segmentée, ce qui rend ce travail particulièrement complexe. C'est une tâche très peu explorée en LS, et à notre connaissance une seule étude, appliquée à la LS britannique (BSL), a été menée sur ce sujet (Bull *et al.*, 2021). Récemment reprise en modifiant le pré-traitement des données textuelles et en intégrant une *perte d'alignement négative* dans (Jang *et al.*, 2025), cette approche a été étendue avec une stratégie de d'entraînement auto-supervisé pour l'alignement. Ces deux méthodes ont d'abord été pré-entraînées sur des tâches de *sign spotting*, ce qui suggère l'existence de données annotées.

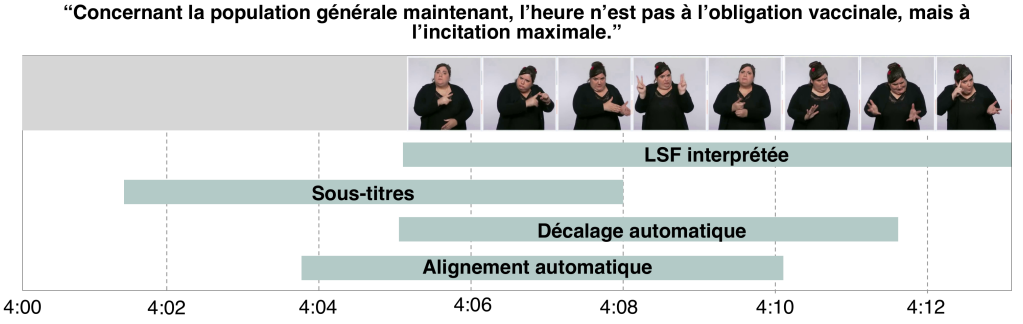
Nous avons mené des expériences d'alignement automatique sur Matignon-LSF en partant du modèle de (Bull *et al.*, 2021). Ce modèle combine des *Transformers* (pour la mise en correspondance des segments de LS avec les sous-titres associés) avec du Dynamic Time Wrapping pour l'alignement long-terme à proprement parler. La limite principale est de nécessiter un *a priori* sur le décalage moyen entre la vidéo et le texte. C'est cependant un bon point de départ pour faire une base de référence pour la LSF.

Nous avons cherché à ajuster le modèle d'origine en utilisant le corpus bilingue LSF/français Mediapi-RGB (Ouakrim *et al.*, 2024). Étant donné que Mediapi-RGB est parfaitement aligné, nous avons forcé le désalignement, en introduisant du bruit dans les repères temporels (timecodes) des sous-titres (entre 0 et 3,5 secondes). Cela nous a permis d'avoir un corpus avec une vérité terrain pour l'ajuster. À ce jour, cette méthode ne nous a pas permis d'améliorer les performances par rapport à ce que nous obtenions sans ajustage, performances qui n'étaient pas elles-mêmes pas encore suffisamment satisfaisantes pour justifier une évaluation quantitative utile.

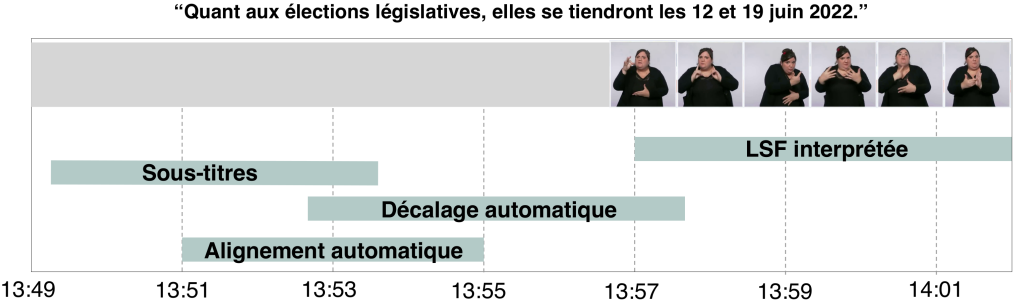
Cependant, afin d'illustrer de manière qualitative les résultats, nous montrons sur deux autres exemples extraits du même discours les segments temporels obtenus en appliquant un décalage systématique du début du segment des sous-titres de 3,5s et un alignement calculé automatiquement. Dans la Figure 6 nous retrouvons les segments de type LSF interprétée, Sous-titres et Décalage automatique complétés par des segments notés « Alignement automatique » qui correspondent à l'alignement fourni par le meilleur modèle pour la correction automatique d'alignement de sous-titres.

Nous observons dans ces exemples que les segments Alignement automatique n'ont été que légèrement décalés par rapport aux segments Sous-titres, moins que les segments de Décalage automatique fixés à 3.5 secondes. Par contre, ils sont légèrement plus long, mais pas de manière notable. L'exemple 2 de la Figure 6 correspond à la phrase « Concernant la population générale maintenant, l'heure n'est pas à l'obligation vaccinale, mais à l'incitation maximale ». Ici, *population générale* est opposée à *personnel soignant* mentionné dans des phrases précédentes. Le contexte, à savoir le fait que la vaccination est obligatoire pour le personnel médical mais seulement encouragé dans la population générale, est rappelé en LS mais pas à l'oral. Cela induit un rallongement significatif de la version LSF interprétée.
L'exemple 3 de la Figure 6 illustre une autre différence entre langue parlée et signée. L'audio original

qui précède cet énoncé est particulièrement rapide et riche en informations temporelles et spatiales : « En Guadeloupe, en Martinique, en Guyane, à Saint-Pierre-et-Miquelon, à Saint-Barthélémy, à Saint-Martin et en Polynésie française, les électeurs voteront le samedi pour tenir compte du décalage horaire ». Dans ce cas, la traduction en LSF nécessite davantage de développement, et entraîne un décalage d'autant plus important entre le Français et la LSF. Les segments de l'alignement comme du décalage automatique sont très en avance par rapport à la LSF interprétée. L'alignement automatique se termine bien avant le début de la LSF interprétée, alors que le décalage automatique termine pendant que la LSF interprétée commence.



Exemple 2



Exemple 3

FIGURE 6 – Segments temporels pour deux exemples. LSF interprétée, sous-titres, décalage automatique des sous-titres de 3,5s et alignement automatique produit par le meilleur des modèles d'alignement testés.

7 Discussion

Nous avons observé empiriquement qu'un décalage manuel de 3,5 secondes pouvait parfois suffire à aligner le début des segments temporels des sous-titres et de la LSF. Cependant, comme on peut le voir dans l'exemple 3 (figure 6), ce n'est pas toujours le cas. Les personnes politiques ont tendance à formuler des énoncés longs et complexes. Dans le corpus Matignon-LSF, un énoncé peut durer jusqu'à

50 secondes. En Français, les énoncés peuvent commencer par des mots qui n'apportent aucune information. Même lorsqu'ils sont prêts, les interprètes doivent attendre pour obtenir suffisament d'éléments d'information avant de commencer l'interprétation.

Parfois, ces longs énoncés sont redondants. Dans ce contexte, pour éviter de répéter deux fois la même information, les interprètes peuvent fusionner trois énoncés parlés en deux énoncés en LS. Cette situation peut créer une perturbation dans l'alignement, puisque nous ne pouvons pas trouver la paire correspondante pour chacun des trois énoncés parlés. Par exemple, toujours dans la même vidéo, nous avons cet énoncé dans les premières secondes : « Hier, le président de la République a pris la parole devant les Français. Il a donné un cap, un cap qu'il nous a répété ce matin en conseil des ministres. Un cap franc et clair. », qui se traduit en langue des signes par quelque chose comme « Hier, le président de la République a annoncé un cap clair. Ce matin, en conseil des ministres, pareil, un cap clair ». Ce type de stratégie permet de raccourcir le délai d'interprétation. Par conséquent, les interprètes ne sont pas systématiquement en retard, et peuvent s'approcher d'une interprétation en temps-réel lorsqu'il y a des redondances dans le discours. C'est également le cas dans d'autres situations, par exemple lorsque des données, du texte ou des informations sont affichés devant eux, ou lorsqu'ils peuvent préparer le discours avant la présentation. C'est le cas dans l'ensemble de données BOBSL : les données doivent être informatives et compréhensibles pour le public, le contenu du discours est écrit avant le début du journal télévisé. Dans ce scénario, les interprètes peuvent s'en servir comme support. Ils savent quand une idée se termine, et peuvent anticiper le contenu de la phrase.

Au travers de ces exemples, nous avons illustré quelques-uns des principaux défis de l'alignement entre la parole et la LSF interprétée. Les décalages temporels entre les deux langues sont variables et la durée d'énoncés de contenu équivalents varie aussi entre les deux langues. Faire des paires d'énoncés LSF/Français ne peut pas être systématique car il n'y a pas toujours de correspondance. Selon le contexte, le Français peut être plus rapide, mais la LSF interprétée peut aussi simplifier des énoncés, fusionner plusieurs énoncés pour n'en faire qu'un seul. Enfin, le besoin de rappeler ou préciser le contexte dépend de la langue utilisée, des besoins de repréciser le contexte en fonction des contraintes linguistiques et apporte une complexité supplémentaire à l'alignement entre Français et LSF.

La traduction automatique impliquant des LS a jusqu'alors été considérée comme une tâche au niveau de la phrase : les contenus vidéos sont découpés et présentés au modèle sous forme de clips isolés associés à des sous-titres. Une étude récente (Tanzer *et al.*, 2024) a montré que ce niveau de granularité n'était pas adéquat du fait de dépendances linguistiques à longue portée, qui rendent une grande proportion de ces clips incompréhensibles s'ils sont considérés isolément. Nous pouvons par exemple citer l'utilisation du pointage en LS qui permet de représenter de manière spatiale des entités spécifiées plus tôt dans le discours et placées dans l'espace de signation afin de pouvoir y faire référence plus tard. La présence d'un pointage dans un clip vidéo nécessite d'avoir une connaissance de la manière dont l'espace de signation a été structuré précédemment. Cela pourrait expliquer au moins en partie le fait que beaucoup d'informations contextuelles sont présentes dans les énoncés interprétés en LSF dans notre corpus, ce qui a pour effet d'allonger la durée des segments.

Nous en concluons que le niveau d'alignement ne peut pas être celui des phrases en français écrit. Une piste pourrait être de s'intéresser d'abord au niveau lexical, puis tenter de repérer des structures de plus haut niveau. Cela nécessite en premier lieu d'annoter le corpus Matignon-LSF, une tâche sur laquelle nous travaillons actuellement. Pour cela, nous nous basons sur l'approche proposée par (Lascar *et al.*, 2024). Cette méthode comprend plusieurs étapes. La première consiste à construire un

lexique bilingue (incluant les variantes potentielles d'une unité lexicale donnée) de manière faiblement supervisée. Le lexique ainsi obtenu est ensuite contrôlé par des experts en LSF, afin en particulier de vérifier la qualité de la segmentation et de supprimer toutes les occurrences mal identifiées ou mal segmentées. Ces données servent ensuite à entraîner un classifieur supervisé pour l'annotation automatique des unités lexicales.

8 Conclusion

Dans cet article, nous avons présenté Matignon-LSF ainsi que les études que nous menons dessus concernant l'alignement entre les sous-titres et la LSF. En effet, Matignon-LSF est un corpus de LSF interprétée, ce qui induit un décalage temporel entre les deux langues.

Nous soulignons la complexité d'aligner les sous-titres (issus de la parole) avec les segments vidéos correspondants en LS interprétée. En ce qui concerne l'alignement automatique, nous avons testé deux approches qui ne donnent pas encore de résultats probants. Les travaux futurs devraient se concentrer sur la conception d'algorithmes permettant l'alignement à différents niveau de granularité, d'abord au niveau lexical, puis à des niveaux plus hauts.

Références

ALBANIE S., VAROL G., MOMENI L., AFOURAS T., CHUNG J. S., FOX N. & ZISSERMAN A. (2020). Bsl-1k : Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020 : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, p. 35–53 : Springer.

ALBANIE S., VAROL G., MOMENI L., BULL H., AFOURAS T., CHOWDHURY H., FOX N., WOLL B., COOPER R., MCPARLAND A. *et al.* (2021). BOBSL : BBC-Oxford British Sign Language Dataset. In *ArXiv preprint*.

BALVET A., COURTIN C., BOUTET D., CUXAC C., FUSELLIER-SOUZA I., GARCIA B., L'HUILLIER M.-T. & SALLANDRE M. A. (2010). The creagest project : a digitized and annotated corpus for french sign language (lsf) and natural gestural languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, p. 469–475.

BELISSEN V., BRAFFORT A. & GOUIFFÈS M. (2020a). Dicta-Sign-LSF-v2 : Remake of a Continuous French Sign Language Dialogue Corpus and a First Baseline for Automatic Sign Language Processing. In ELRA, Éd., *LREC 2020, 12th Conference on Language Resources and Evaluation*, Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France. HAL : hal-02541792.

BELISSEN V., BRAFFORT A. & GOUIFFÈS M. (2020b). Experimenting the automatic recognition of non-conventionalized units in sign language. *Algorithms*, **13**(12), 310–336.

BRAFFORT A. (2022). Langue des signes française : Etat des lieux des ressources linguistiques et des traitements automatiques. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, p. 131–138 : CNRS.

BUET F. (2022). *Modèles neuronaux pour la simplification de parole, application au sous-titrage*. Theses, Université Paris-Saclay. HAL : tel-03920729.

BULL H., AFOURAS T., VAROL G., ALBANIE S., MOMENI L. & ZISSERMAN A. (2021). Aligning subtitles in sign language videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, p. 11552–11561.

BULL H., BRAFFORT A. & GOUIFFÈS M. (2020a). MEDIAPI-SKEL -A 2D-Skeleton Video Database of French Sign Language With Aligned French Subtitles. In *12th Conference on Language Resources and Evaluation (LREC 2020)*, p. 6063–6068, Marseille, France. HAL : hal-02952340.

BULL H., BRAFFORT A. & GOUIFFÈS M. (2020b). MEDIAPI-SKEL -A 2D-Skeleton Video Database of French Sign Language With Aligned French Subtitles. In *12th Conference on Language Resources and Evaluation (LREC 2020)*, p. 6063–6068, Marseille, France. HAL : hal-02952340.

CARREIRA J. & ZISSERMAN A. (2017). Quo vadis, action recognition ? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 6299–6308.

DAYTER D. (2019). *Collocations in non-interpreted and simultaneously interpreted English : a corpus study*. Routledge.

DUARTE A., PALASKAR S., VENTURA L., GHADIYARAM D., DEHAAN K., METZE F., TORRES J. & GIRO-I NIETO X. (2021). How2Sign : A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2735–2744.

JANG Y., CHOI J., AHN J. & CHUNG J. S. (2025). Deep understanding of sign language for sign to subtitle alignment.

KOLLER O., FORSTER J. & NEY H. (2015). Continuous sign language recognition : Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, **141**, 108–125.

KOPF M., SCHULDER M. & HANKE T. (2023). The sign language dataset compendium. DOI : 10.25592/uhhfdm.12017.

LASCAR J., GOUIFFÈS M., BRAFFORT A. & DANET C. (2024). Annotation of lsf subtitled videos without a pre-existing dictionary. In *Workshop on the Representation and Processing of Sign Languages at the International Conference on Language Resources and Evaluation (sign-lang@LREC)*.

OUAKRIM Y., BEAUTEMPS D., GOUIFFÈS M., HUEBER T., BERTHOMMIER F. & BRAFFORT A. (2023). A multistream model for continuous recognition of lexical unit in french sign language. In *29° Colloque sur le traitement du signal et des images"*, volume 2023-1182, p. 461–464 : GRETSI - Groupe de Recherche en Traitement du Signal et des Images.

OUAKRIM Y., BULL H., GOUIFFÈS M., BEAUTEMPS D., HUEBER T. & BRAFFORT A. (2024). Mediapi-RGB : Enabling technological breakthroughs in french sign language (LSF) research through an extensive Video-Text corpus. *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, **2**, 139–148.

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

STONE C. & RUSSELL D. (2011). Interpreting in international sign : decisions of deaf and non-deaf interpreters. In *Proceedings of World Association of Sign Language Interpreters Conference*.

TANZER G., SHENGELIA M., HARRENSTIEN K. & UTHUS D. (2024). Reconsidering sentence-level sign language translation. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 6262–6287,

Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.360.

TARRÉS L., GÁLLEGO G. I., DUARTE A., TORRES J. & GIRÓ-I NIETO X. (2023). Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 5624–5634.

UTHUS D., TANZER G. & GEORG M. (2024). Youtube-asl : A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems*, **36**.

VAROL G., MOMENI L., ALBANIE S., AFOURAS T. & ZISSERMAN A. (2021). Read and attend : Temporal localisation in sign language videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 16857–16866.