Avatar signeur -Synthèse de la langue des signes française à partir de texte

Sylvie Gibet

Avatar signeur – Synthèse de la langue des signes française à partir de texte

Sylvie Gibet

IRISA, Université Bretagne Sud
Campus de Tohannic, rue Yves Mainguy
F-56017 Vannes cedex
sylvie.gibet@univ-ubs.fr

RÉSUMÉ. *Nous présentons dans cet article un système de synthèse de mouvement multimodal qui produit de la langue des signes française (LSF) à partir d'énoncés textuels au moyen d'un personnage virtuel 3D appelé avatar signeur. Notre système SignCom s'appuie sur un principe de composition multicanale, chaque canal d'information étant associé à une information linguistique (depuis le niveau phonologique vers les niveaux lexical, syntaxique ou sémantique) ou articulatoire. La composition permet un agencement spatio-temporel de ces éléments qui s'exécutent en parallèle, et donne la possibilité d'éditer et de générer des phrases en LSF. Les nouveaux modules de synthèse qui enrichissent le système initial sont décrits. Ils incluent la synthèse de mouvement corporel et des mains ainsi que la synthèse faciale, et mettent en œuvre des dynamiques grammaticales propres à la LSF, en s'appuyant sur les concepts fondamentaux de spatialité et d'iconicité. Enfin, nous présentons les principaux défis technologiques qui restent à relever avant de conclure.*

MOTS-CLÉS : *langues des signes, synthèse texte-vers-LS, avatar signeur*

ABSTRACT. *In this paper, we present a multimodal synthesis system that translates text-to-LSF (French sign language) by means of a 3D virtual character, also called virtual signer. Our SignCom system is based on a multichannel composition mechanism, each channel being associated to linguistic information (from the phonological level to the lexical, syntactic or semantic levels), or to articulatory information. The composition is based on a spatio-temporal arrangement of these elements that are parallelized, and is able to edit and generate utterances in LSF. The new synthesis modules that enrich the initial system are described, including body movement synthesis, facial and hand movement synthesis. They implement grammatical dynamics specific to sign language, based on the fundamental concepts of spatiality and iconicity. Finally, we present the main technological challenges that remain before concluding.*

KEYWORDS: *Sign languages, Text-to-SL synthesis, Signing avatar*

1. Introduction

Les langues des signes repoussent les frontières habituelles des théories linguistiques associées aux langues vocales. Ceci est principalement dû au fait qu'elles utilisent l'information visuelle et gestuelle, contrairement aux langues vocales qui utilisent le canal audio-oral. Ainsi, les personnes sourdes développent avec la pratique de cette langue une dextérité dans leur gestuelle et dans leur perception visuelle, une acuité de représentation de l'espace et une capacité d'expression qui se manifestent à travers les différentes modalités[1] de communication propres aux langues gestuelles, incluant les mouvements des mains, les mouvements corporels non manuels, les mimiques faciales, la direction du regard et la labialisation éventuellement associée au son. C'est pourquoi ces langues peuvent être qualifiées de multimodales.

Cette spécificité de la gestualité s'accompagne de mécanismes iconiques et spatiaux omniprésents dans les langues des signes. L'iconicité met en jeu des processus par lesquels le locuteur décrit l'expérience vécue, imaginée ou exécutée en s'inspirant de représentations imagées ou mimétiques (Cuxac, 2000). Elle est caractérisée par le lien de ressemblance plus ou moins étroit entre les entités du monde réel, le référent et le signe qui s'y rapporte. Cuxac propose ainsi une théorie de l'iconicité dans laquelle plusieurs structures linguistiques se combinent lors d'activités discursives : les structures dites de grande iconicité à visée illustrative et les structures dites standard (dans leur forme de citation) sans visée illustrative, comprenant des unités lexicales, de pointage ou des unités dactylologiques (Sallandre et Garcia, 2020). Millet propose une grammaire descriptive de la langue des signes française (LSF) qui s'appuie également sur la spatialité et l'iconicité structurant à tous les niveaux (phonologique, lexical, syntaxico-sémantique) cette langue (Millet, 2019).

En nous appuyant sur ces théories linguistiques de la LSF, nous nous intéressons aux outils numériques à destination des personnes sourdes signantes qui permettent de produire automatiquement des contenus en langue des signes (LS). À l'heure actuelle, la plupart des applications disponibles reposent sur de la vidéo. Or, si la vidéo est le média le plus partagé par les sourds, elle ne permet pas de garantir l'anonymat et impose des contraintes fortes au niveau du stockage et du transport d'information. En contrepartie, la production automatique de contenus en LS et la visualisation au moyen d'avatars signeurs 3D constituent une réponse alternative appropriée et permettent à la fois la réduction des informations stockées, l'anonymisation ainsi que la manipulation des informations pour éditer, visualiser et produire à moindre coût de nouveaux énoncés.

Nous nous focalisons ici sur les systèmes de génération, de synthèse et de traduction texte-vers-LS (que nous regrouperons sous le sigle TSL[2]) et qui incorporent des avatars signeurs. Avec les avancées significatives du traitement automatique des langues parlées, ces systèmes TSL connaissent un regain d'intérêt ces dernières années. Dans les LS, la traduction peut être réalisée en deux étapes, comme illustré dans la figure 1. La première transforme le français écrit en un langage pivot intermédiaire

1. Terme employé dans le domaine des agents conversationnels animés.
2. *Text to Sign Language.*

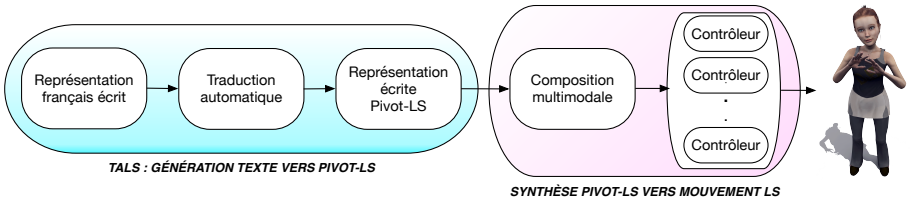

FIGURE 1. *Traduction texte-vers-LS en deux étapes : 1) Génération texte vers Pivot-LS, 2) Synthèse Pivot-LS vers mouvement en LS*

(appelé *Pivot-LS*) qui tient compte des spécificités des LS (partie *TALS* : traitement automatique des langue des signes). La seconde permet de passer de la spécification symbolique exprimée dans ce langage pivot à la production d'un flux continu de postures de l'avatar, au moyen d'un système de synthèse comprenant un procédé de composition multimodale et un ensemble de contrôleurs de mouvement.

Bien que les travaux linguistiques sur les LS aient permis de mieux appréhender les mécanismes grammaticaux en jeu (Millet, 2019), la conception de système TSL se révèle être une tâche complexe, qui soulève deux problèmes principaux : *i)* la génération TALS est loin de couvrir toute la variabilité linguistique des LS et n'est pas encore entièrement automatisée ; *ii)* la synthèse est confrontée à la nature même des signes, par essence multimodale, et est contrainte par des structures linguistiques sous-jacentes.

Nous nous concentrons dans cet article sur ces deux types de problèmes indissociables, en mettant davantage l'accent sur le second qui considère la synthèse automatique d'une description textuelle aux mouvements LS. Plus précisément, nous présentons notre système *SignCom* qui est une version étendue du système précédemment développé (Gibet *et al.*, 2011). Ce système s'appuie sur un principe de composition phonologique pour construire les signes, et d'édition d'énoncés en LSF, en combinant des modèles de synthèse basée données et de synthèse procédurale. Après avoir décrit les principaux travaux de l'état de l'art, nous présentons notre système *SignCom*, en soulignant les principales avancées technologiques réalisées ces dernières années, puis nous décrivons les principaux défis qui restent à relever pour traduire un texte en LS. Quelques exemples en LSF illustrent notre propos.

2. État de l'art

Les études sur les LS sont relativement récentes, avec des approches linguistiques très diversifiées, des modèles d'animation de personnages virtuels et des réalisations informatiques qui dépendent des avancées technologiques et des données disponibles. Dans cette section, nous explorons les principales représentations linguistiques des LS utilisées pour les systèmes TSL avec avatars signeurs ; pour une revue exhaustive sur le sujet, se référer à (Naert *et al.*, 2020) et (Núñez-Marcos *et al.*, 2023).

Représentations linguistiques des langues des signes. Les premiers travaux sur la phonologie des langues des signes ont donné lieu à différents types de représentations. Parmi celles-ci, les travaux de Stokoe (Stokoe, 1972) ont abouti à la description phonologique de l'ASL (*American Sign Language*) sous la forme d'une combinaison de paramètres constituant les signes : l'emplacement du signe, la forme de la main et son mouvement. Cette représentation paramétrique repose sur la possibilité de distinguer le sens des signes à partir de la modification d'un de ses paramètres constitutifs (notion de paire minimale). Un dictionnaire de l'ASL a été créé à partir de cette représentation. Poursuivant les travaux de Stokoe, d'autres paramètres qui participent à la formation et à la distinction des signes ont été identifiés. Ils comprennent l'orientation de la main ainsi que les éléments non manuels tels que les expressions faciales, la direction du regard, les orientations du buste et certains gestes corporels (Battison, 1978). Le système de notation HamNoSys (Prillwitz et Zentrum, 1989) reprend les paramètres précédents et transcrit de manière linéaire les signes en utilisant les symboles informatiques *Unicode*. Les travaux linguistiques remarquables sur l'ASL menés par Liddell et Johnson ont abouti à la définition d'un modèle phonétique qui s'appuie sur le schéma *Posture-Detention-Transition-Shift* (PDTS), en distinguant sur chaque composante phonologique – configuration de la main (HC), orientation (FA), placement (PL), caractéristiques non manuelles (NM) – des éléments statiques et des éléments dynamiques transitionnels.

En linguistique computationnelle, les gestes des LS ont été décrits au moyen de formalismes allant de scripts à des langages informatiques dédiés. Ainsi, le langage SiGML (Elliott *et al.*, 2008) basé sur HamNoSys a été développé pour générer les animations d'avatars 3D. Ce langage a ensuite été étendu en incorporant le modèle PDTS de Johnson et Liddell (Glauert et Elliott, 2011). Une grammaire générative de signes a également été développée à partir d'un système de composition phonologique parallélisé s'appuyant sur des cibles et des mouvements articulés entre cibles (Gibet *et al.*, 2001). Le langage formel AZee, quant à lui, intègre un formalisme symbolique qui s'appuie sur la géométrie et a pour ambition de représenter un ensemble de procédés grammaticaux de manière parallélisée et non linéaire (Filhol *et al.*, 2017).

Ces langages de script ou de spécification permettent de décrire des signes ou des énoncés de manière très analytique et précise. Cependant, la spécification de nouveaux signes peut être très fastidieuse. De plus, la plupart de ces langages intègrent au sein de leur formalisme des éléments temporels explicites, par exemple SIGML, EMBRScript (Héloir et Kipp, 2010) et AZee, où les postures clés de l'avatar sont spécifiées à des instants prédéterminés. Par contre, le modèle *de partition/constitution (P/C)* (Huenerfauth, 2006) propose un schéma de synchronisation implicite qui repose sur une représentation 2D d'un graphe syntaxique 3D. Ce modèle facilite la visualisation et la coordination d'éléments linguistiques sur des axes temporel et spatial. Peu de représentations linguistiques se sont intéressées aux flexions grammaticales des signes. Parmi celles-ci, le projet ATLAS (Lombardo *et al.*, 2010) en LS italienne incorpore des processus flexionnels impliquant l'emplacement, la configuration et le mouvement, ainsi que des spécificateurs de forme et de taille (SFT). De son côté, le système AZee-Paula (Filhol et McDonald, 2018) permet la génération d'un large panel de mécanismes flexionnels en ASL.

Systèmes d'animation d'avatars signeurs. Parmi les méthodes et technologies disponibles pour animer des avatars signeurs, on distingue trois approches principales.

La première consiste, à partir de données restreintes, à animer l'avatar en utilisant des méthodes dites de *synthèse pure*. Nous regroupons dans cette catégorie les approches à base de postures clés, qu'elles soient déterminées manuellement ou automatiquement, associées à des processus d'interpolation, et les approches dites procédurales qui automatisent le processus de génération de mouvement. Avec de telles méthodes, il est possible de synthétiser des signes isolés identifiés par une glose[3] et de construire des séquences signées à partir de procédés de concaténation et de mélange de mouvements. Cela nécessite de contrôler toute la chaîne de production dans ses moindres détails, depuis la spécification des éléments linguistiques de base, leur agencement au moyen d'un langage de description ou de spécification, jusqu'à la synthèse du mouvement. Si ces systèmes d'animation, qui couplent un langage *symbolique* à un moteur d'animation, permettent d'atteindre des objectifs de précision et de contrôle fin des postures statiques, ils aboutissent généralement à des mouvements peu naturels, voire robotisés. De plus, le processus de spécification, long et fastidieux, conduit à la création d'un nombre limité de signes, avec peu ou pas de possibilités de flexions grammaticales. Enfin, la gestion du temps reste complexe à mettre en œuvre, tant au niveau des signes (gestion de la synchronisation entre les composantes des signes) que des transitions entre signes (gestion de la coarticulation). Deux systèmes relatifs à cette approche ont été développés. Avec EMBR (Kipp *et al.*, 2011), les signes dans leur forme de citation sont générés à partir de séquences de poses spécifiées au moyen du langage EMBRScript. JASigning intègre quant à lui le moteur d'animation *AnimGen* qui permet la création de signes spécifiés à partir du langage SiGML (Kennaway *et al.*, 2007 ; Elliott *et al.*, 2008). Ces systèmes d'animation ont été utilisés pour différentes LS (Ebling *et al.*, 2016 ; Efthimiou *et al.*, 2010 ; Roelofsen *et al.*, 2021). Dans les deux cas, la flexion des signes n'est possible qu'au niveau lexical. Plus récemment, le système *Paula* développé à DePaul University génère des animations à partir de la représentation formelle AZee, en combinant des techniques à base de postures clés et d'algorithmes procéduraux qui améliorent la fluidité du mouvement et facilitent la synthèse multimodale (McDonald *et al.*, 2016 ; McDonald et Filhol, 2021).

La seconde approche consiste à développer des méthodes de *synthèse basée données*. Dans ce cas, les mouvements du signeur virtuel sont capturés par des techniques de capture de mouvement qui enregistrent simultanément les mouvements manuels, corporels, ainsi que les expressions faciales et la direction du regard. Par exemple, les projets *SignCom* (Gibet *et al.*, 2011), *Sign3D* (Gibet *et al.*, 2016) ou *Rosetta* (Dauriac *et al.*, 2022) ont permis de développer un système d'animation d'avatars signeurs en LSF à partir de mouvements capturés haute résolution. Les approches basées données facilitent la production d'animations fluides et crédibles d'avatars 3D. Elles permettent de rejouer des séquences relativement longues de LS, mais aussi de modifier des phrases préenregistrées pour créer de nouveaux énoncés. Cependant, la manipulation et l'adaptation des mouvements à de nouveaux contextes nécessitent la prise

3. Une glose est la représentation lexicale en français écrit du signe.

en compte de processus complexes afin de conserver la cohérence du contenu en LS produit, tant au niveau des animations que de l'intelligibilité des phrases en LS.

Les deux types de méthodes – synthèse basée données et synthèse pure – peuvent être combinées pour conduire à des méthodes de synthèse dite *hybride*. Il est en effet possible de remplacer des segments de mouvement par des mouvements synthétisés, ou de combiner des méthodes procédurales avec des données en s'appuyant éventuellement sur des techniques d'apprentissage automatique. Cette synthèse hybride apporte une certaine flexibilité et la possibilité d'enrichir les bases de données en générant, à partir de séquences existantes, des séquences signées avec variations synthétisées. De telles approches ont été développées, notamment en combinant mouvements capturés et méthodes procédurales pour l'étude des verbes directionnels (Huenerfauth *et al.*, 2015) ou pour générer des énoncés avec flexion spatiale (modification de la spatialisation ou du pointage) ou syntaxique (modification de l'agent, de l'objet ou du bénéficiaire) (Naert *et al.*, 2021). Le projet *Rosetta* se place également dans cette approche, en combinant dans un processus d'édition parallélisée une synthèse basée données avec des algorithmes procéduraux (Dauriac *et al.*, 2022).

Il est à noter l'émergence de technologies TSL, notamment celles développées par Keia[4] pour la LSF, et *Hand Talk App*[5] pour la LS brésilienne (Libras) et l'ASL.

L'annotation des données est au cœur des systèmes d'analyse et de synthèse des LS. En effet, c'est lors du processus de segmentation et d'étiquetage des contenus LS que l'information linguistique est incorporée au niveau des pistes du système d'annotation. Ces pistes permettent d'encoder des informations textuelles (gloses) de nature phonétique, phonologique, lexicale, syntaxique ou sémantique. Elles permettent également d'informer sur le type de mouvement (s'agit-il d'un mouvement signifiant ou d'une transition, etc.), ou bien de structurer hiérarchiquement des groupes d'articulations. L'annotation s'appuie par conséquent sur un formalisme linguistique et sur une représentation structurelle du mouvement qui conditionnent la reconnaissance ou la synthèse des LS. Les premiers systèmes d'annotation on été réalisés de manière manuelle à l'aide d'outils informatiques dédiés (Chételat-Pelé et Braffort, 2008). Plus récemment, des modèles d'annotation par apprentissage automatique ont vu le jour. Ils sont appliqués sur des données de capture de mouvement (Naert *et al.*, 2018) ou sur des données vidéo (Chaaban *et al.*, 2021).

Systèmes de traduction automatique texte-vers-LS. Avec l'avènement de l'apprentissage profond, des modèles récents à base de réseaux de neurones (NN) ont été développés avec succès pour la traduction automatique en LS, en s'inspirant du principe de la traduction du texte vers la parole ou d'une langue parlée vers une autre. Quelques états de l'art permettent de recenser ces systèmes TSL (Kahlon et Singh, 2021) ou plus largement parole/texte-vers-LS et LS-vers-texte (Farooq *et al.*, 2021 ; Núñez-Marcos *et al.*, 2023). Dans ce contexte, des NN ont été proposés pour la traduction en LS arabique (Brour et Benabbou, 2019), ou pour la traduction

4. https://www.keia.io/
5. https://www.handtalk.me/en/

de l'anglais parlé en ASL dans le cadre applicatif de la diffusion de bulletins météo. Des réseaux générateurs de type GAN, associés à des graphes de mouvement, ont également été proposés pour produire des vidéos de LS à partir de phrases en langue parlée (Stoll *et al.*, 2020). Les résultats sont prometteurs mais encore insuffisants.

3. Caractéristiques des langues des signes

Nous partons de l'hypothèse que la formation des signes et des énoncés en LS est déterminée par la réalisation simultanée de formes de main, d'orientations, d'emplacements, de mouvements, qui constituent les unités minimales, dites phonologiques des signes (Stokoe, 1972 ; Battison, 1978 ; Liddell et Johnson, 1989). La composition parallèle de ces unités phonologiques en nombre restreint permet de construire un ensemble structuré et codifié qui définit les bases du lexique, avec une économie de représentation qui est à rapprocher de la structure phonétique des langues vocales. De plus, les LS s'appuient sur deux dynamiques essentielles de l'expression gestuelle, à savoir la spatialité et l'iconicité. L'ensemble des règles qui sont structurées relativement à ces dynamiques gestuelles fonde la grammaire de la LSF. Une autre spécificité des LS concerne la difficulté de séparer les différents niveaux linguistiques – phonologique, lexical, syntaxique et discursif – qui sont propres aux langues vocales. Dans cette section, nous évoquons de manière non exhaustive quelques mécanismes linguistiques de la formation des signes isolés et des énoncés en LSF. Nous nous focalisons plus particulièrement sur les procédés incorporés à notre système de synthèse. La terminologie est empruntée au modèle linguistique de Millet (Millet, 2019).

3.1. *Formation des signes isolés*

La formation des signes isolés dans leur forme de citation nécessite la combinaison spatio-temporelle des composantes phonologiques décrites précédemment. Ces signes sont toujours exécutés de la même manière à un emplacement spécifique de l'espace du signeur, qui peut être soit une zone neutre de l'espace qui l'entoure, soit un emplacement sur son corps. Ces signes n'étant pas soumis aux processus de flexion grammaticale, leurs variations proviennent essentiellement de la façon dont les mouvements sont exécutés (occupation de l'espace, cinématique des mouvements). Leur réalisation requiert toutefois une grande précision, la modification d'une composante phonologique engendrant un sens différent, comme le signe [CHOCOLAT] qui devient le signe [BRICOLER][6] en modifiant le mouvement de la main dominante (vitesse et amplitude), l'emplacement et les configurations manuelles étant inchangés. Du point de vue temporel, les règles de synchronisation entre les éléments composant le signe doivent être respectées.

6. http://www.sematos.eu/lsf.html

3.2. *Formation des énoncés à visée illustrative*

Nous explorons ci-après quelques mécanismes grammaticaux en LSF, en organisant notre propos suivant les concepts de spatialité et d'iconicité, les éléments de spatialité étant implicitement reliés aux dynamiques iconiques. Il s'agit d'une description succincte, schématisée à des fins de modélisation pour la synthèse présentée dans la section 4. La flexion grammaticale, qui se traduit par la modification d'une ou de plusieurs composantes phonologiques, conduit à modifier le sens de l'énoncé.

Espace de signation, Locus. L'espace de signation est défini comme étant l'espace dans lequel le discours du locuteur va se déployer. Le signeur utilise cet espace en positionnant des entités, animées ou non, présentes dans son discours. Pour ce faire, il définit des zones symboliques discrètes, parfois présémantisées (par exemple pour représenter la 1^{re} ou la 3^{e} personne). Les emplacements abstraits (*Locus*) associés à ces zones de l'espace de signation, permettent d'identifier et de rappeler ces entités, apportant ainsi une cohérence sémantique à la phrase.

Ancrage et spatialisation. Au niveau lexical, les signes sont réalisés à un emplacement neutre (juste devant le signeur) ou sur son corps. Certains signes à ancrage neutre peuvent être repositionnés à d'autres emplacements de l'espace de signation (spatialisation), ce qui modifie ainsi leur rôle syntaxique/sémantique dans l'énoncé. Par exemple, pour décrire le placement d'un livre sur une étagère, le signe [LIVRE] est d'abord signé dans son ancrage lexical, puis la forme de la main en *proforme* [LIVRE] est positionnée à un endroit cible sur l'étagère.

Pointage. Le pointage a soulevé de nombreuses questions dans la communauté internationale (Garcia *et al.*, 2011 ; Blondel *et al.*, 2004). Il peut prendre différentes formes, soit en étant porteur d'une signification propre (valeur de pronom par exemple), soit en ayant l'objectif de montrer une entité. Dans ce dernier cas, l'entité pointée constitue l'information signifiante, le mouvement de pointage réalisant le déplacement de la main vers cet élément pointé. La désignation d'un emplacement (ou locus) se fait souvent par pointage de l'index ou autre configuration manuelle.

Formes de main. La configuration manuelle (HC) comporte une forte dimension iconique. Du point de vue lexical, elle est l'une des composantes de formation des signes. Par exemple le signe [ESCARGOT] devient [LIMACE] en modifiant la HC (Y devient H), les mouvements étant inchangés (Naert *et al.*, 2021).

Spécificateurs de forme et de taille (SFT). Les SFT sont des procédés qui permettent de décrire la forme ou la taille des entités signées. Si l'on prend l'exemple des signes lexicalisés [BOL] et [VERRE] en LSF, ils peuvent être fléchis de manière à leur adjoindre une valeur adjectivale, au niveau de la forme pour devenir [VERRE-A-EAU] ou [VERRE-DE-CHAMPAGNE], ou bien au niveau de la taille ([GRAND-BOL], [PETIT-BOL]) (Gibet *et al.*, 2011).

Verbes directionnels ou à trajectoire. Ces verbes s'exécutent au moyen d'un mouvement allant d'un locus à l'autre et mettent en jeu plusieurs actants (agent, objet, bénéficiaire). Ainsi, il est possible de fléchir le verbe [DONNER] suivant différentes

trajectoires allant d'un locus agent à un locus bénéficiaire, ces actants pouvant être des pronoms positionnés dans l'espace de signation. Par exemple, la phrase en français « Je te donne » peut être traduite en LSF par un mouvement de la main d'une zone neutre près du buste (personne 1) vers une zone devant le signeur (personne 2), alors que la phrase « Tu me donnes » est traduite par une trajectoire inversée. Outre leur flexion suivant la trajectoire, certains verbes directionnels peuvent être fléchis suivant l'actant objet. Dans ce cas, la forme de la main représentant l'objet est modifiée. Ainsi, dans les exemples en français « Je te donne un verre » ou « Je te donne un livre », les objets sont représentés par des HC différentes, représentant soit un verre, soit un livre.

Proformes manuelles statiques. Les proformes statiques sont représentées par des HC qui référencient, lorsqu'elles sont maintenues, des éléments lexicaux. Elles peuvent représenter des entités animées (personnes, véhicules). Elles assurent également une fonction pronominale (substitution à une entité lexicale), ou anaphorique. Ce mécanisme favorise ainsi le rappel d'une entité dans le discours. Par exemple, la proforme [PERSONNE] (index pointé vers le haut) peut être rapidement positionnée dans l'espace de signation. De plus, la personne peut être représentée dans différentes positions (debout, assise ou allongée), associées à des HC (et orientations) différentes. Par extension, on peut représenter facilement plusieurs personnes dans un espace (autour d'une table par exemple) ou dans une salle de conférence. Les proformes permettent également de positionner des entités les unes par rapport aux autres. Par exemple, la phrase « Le stylo est dans le verre » peut se traduire en LSF par la forme de main en proforme [STYLO] qui vient se positionner dans la proforme représentant le signe [VERRE]. Les proformes constituent ainsi des procédés iconiques efficaces qui assurent la cohérence syntaxique du discours.

Proformes manuelles dynamiques. Certains comportements ou démarches peuvent être représentés par des proformes dynamiques. C'est le cas par exemple lorsque l'on veut décrire la démarche d'un humain ou d'un animal (un oiseau, un ours, un lion, etc.). La forme de la main représente la forme de la patte de l'animal, et le mouvement des mains indique quant à lui la qualité de la démarche (souplesse/raideur, légèreté/lourdeur) (Naert *et al.*, 2021).

Expressions faciales. Les expressions faciales (FE), par essence iconiques, sont fondamentales dans les LS car elles véhiculent trois types d'informations d'ordre lexico-syntaxique (Millet, 2019) : des informations relatives à la modalité de la phrase, des informations expressives liées à la nature émotionnelle du discours, ou des informations de nature adverbiale ou adjectivale.

– *Modalité de la phrase.* Trois modalités principales ayant un rôle syntaxique s'expriment à travers des FE appropriées : la modalité assertive, correspondant à une phrase affirmative, s'exprime le plus souvent par un visage neutre ; la modalité interrogative se traduit par une FE exprimant l'interrogation ; la modalité impérative est employée lorsque le locuteur donne un ordre. La négation quant à elle peut s'exprimer par le signe [NON], par une mimique faciale, ou les deux combinés. Enfin, une mimique faciale, associée à un mouvement du buste et de la tête peut aussi exprimer la condition (« S'il pleut, je reste à la maison »).

– *Affect.* Au-delà des FE primaires identifiées par (Ekman et Friesen, 1978) (joie, colère, peur, etc.), il existe de très nombreuses FE en LSF (inquiétude, admiration, réflexion, etc.) (Cuxac, 2000). D'autres FE expriment les états affectifs caractérisant l'attitude du locuteur vis-à-vis du contenu de l'énoncé. Les principales concernent l'expression exclamative (surprise) ou dubitative (doute) (Millet, 2019).

– *Valeur adverbiale ou adjectivale.* Enfin certaines mimiques faciales, dans des situations de dialogue ou de récit, ont des fonctions adverbiales (gonflement des joues qui accompagne la phrase « Le vent souffle fort »), ou adjectivales (joues rentrées ou rebondies dans « un homme mince ou gros »).

Ces FE varient en fonction du style du locuteur, de son état affectif et de la morphologie de son visage. De plus, les FE peuvent être combinées entre elles en fonction du contexte du discours.

Autres mouvements – tête, buste, épaules. Les mouvements impliquant d'autres parties du corps (tête, buste, épaules) sont aussi porteurs d'information. Ainsi, en LSF, le torse légèrement penché en avant indique une action réalisée dans le futur. L'orientation du buste est également utilisée pour passer d'un personnage à l'autre dans une narration. Enfin, pour des formes 2D ou 3D, il est possible de décrire les différents niveaux de détail de l'objet en inclinant en avant le buste.

De nombreux autres procédés grammaticaux existent en LSF (proformes non manuelles, prises de rôle, exploitation du regard, etc.). Ils sont décrits précisément dans l'ouvrage de Millet (Millet, 2019).

Dans la suite de cet article, nous nous intéressons à la modélisation, la spécification et l'implémentation de ces procédés linguistiques, en mettant l'accent sur les avancées techniques du système TSL *SignCom*.

4. Notre système de synthèse *SignCom* texte-vers-LSF

Les langues des signes requièrent une grande précision et un haut niveau de réalisme dans l'exécution des mouvements corporels, manuels et faciaux, afin qu'ils soient compris et acceptés par les sourds. La capture du mouvement (MoCap) associée à l'animation d'un avatar 3D permet de répondre à ces exigences. Cependant, le coût de production de ces données MoCap reste très élevé, et il est pertinent d'enrichir les bases de données existantes au moyen de processus d'édition. C'est cette motivation qui nous a conduits à développer un système d'édition et de synthèse multimodale produisant simultanément les mouvements corporels, manuels, parallèlement aux expressions faciales et à la direction du regard. Ainsi, dans le système *SignCom* (Gibet *et al.*, 2011), l'édition a permis de construire de nouvelles phrases, en coupant/collant/transformant/mixant les segments de mouvement sélectionnés, et de produire automatiquement l'animation d'un avatar signeur en 3D. Différents processus d'édition on été explorés, i) par remplacement de signes ou de groupes de signes, ii) par instanciation de schémas syntaxiques prédéfinis, ou, iii) par remplacement de composantes phonologiques des signes – configurations des mains, mouvements ma-

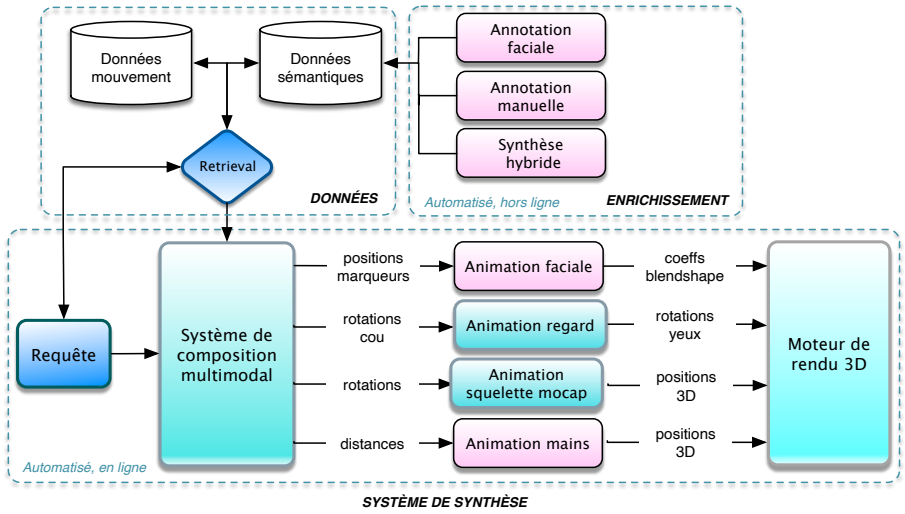

FIGURE 2. *Système de synthèse texte-vers-LSF (SignCom) avec avatar signeur ; en bleu : système de base (Gibet* et al.*, 2011) ; en rose : avancées technologiques*

nuels, du torse et de la tête, et expressions faciales (Gibet, 2018). Cependant, si le système *SignCom* dans sa première version était capable de manipuler des contenus LS aux niveaux phonologique, lexical et discursif, en s'appuyant sur des processus d'annotation manuelle, il ne permettait pas de prendre en compte des mécanismes grammaticaux plus complexes tels que ceux évoqués dans la section 3.

Nous avons développé une extension du système *SignCom* qui intègre certains procédés flexionnels de la grammaire de la LSF se rapportant à la spatialité et à l'iconicité propres à cette langue. Ainsi, la possibilité de fléchir des composantes des signes – notamment les composantes manuelles ou faciales –, permet de produire des variations grammaticales des phrases du corpus initial. Ces processus flexionnels s'expriment à partir des nouveaux modules de synthèse présentés dans cette section. La figure 2 illustre le système *SignCom* dans son ensemble, en intégrant les modules fonctionnels de base et ceux constituant les avancées technologiques. Après avoir rappelé le principe de la synthèse concaténative (section 4.1), nous décrivons ci-après les modules d'annotation automatique (section 4.2) qui visent à réduire le temps d'annotation manuelle, puis nous présentons les modules de synthèse hybride corporelle (section 4.3), de synthèse faciale (section 4.4) et de synthèse du mouvement des mains (section 4.5) qui enrichissent notre système initial.

4.1. *Principe de la synthèse concaténative*

Notre système de synthèse concaténative repose sur des données de mouvement préalablement capturées. Ces données sont caractérisées par des séquences de postures du squelette du signeur. Elles sont enregistrées au moyen d'un système de capture de mouvement à base de marqueurs passifs et de caméras infrarouges, qui permettent de déterminer de manière très précise la position des marqueurs placés sur le corps, les mains et le visage de l'acteur (fréquence d'acquisition de 200 Hz). Il en résulte la constitution d'un ensemble de séquences de mouvement représentant les énoncés du corpus en LSF. Ces données sont ensuite annotées en suivant des schémas multicanaux pour lesquels nous distinguons i) les canaux linguistiques qui respectent la description phonologique des signes (Johnson et Liddell, 2011) et leur classe grammaticale (Johnston, 1998), et ii) les canaux physiques qui correspondent à des groupes d'articulations (main, bras, etc.).

Le système de synthèse est divisé en deux parties : un processus hors ligne de stockage des données et un processus en ligne d'extraction des données et de contrôle de l'animation. La nécessité d'encoder dans les énoncés signés, d'une part les informations linguistiques traduisant la structure multilinéaire des LS et d'autre part les flux de mouvement, nécessite de construire au préalable deux bases de données hétérogènes couplées, l'une contenant les données sémantiques issues de l'annotation, l'autre les données de mouvement (positions ou angles aux articulations du squelette). La base de données sémantique réalise le couplage entre les données symboliques suivant notre schéma multicanal d'annotation et une liste de mouvements indexés au moyen d'un nom, de marqueurs temporels, et d'une séquence d'articulations (au sens biomécanique) impliquées dans le mouvement. Il s'agit d'un couplage *un-vers-plusieurs* pour tenir compte de l'existence de plusieurs instances d'un même signe ou partie de signe dans le corpus. Un langage de requêtes multiconditions permet, à partir de la spécification d'éléments propres à cette annotation, d'extraire automatiquement un flux continu de postures du mouvement qui lui correspondent.

Le système d'animation à proprement parler s'appuie sur un processus de composition multicanale du mouvement qui reçoit en entrée des flux de données associées à des segments corporels qui sont, soit des groupes d'articulations du squelette appelés *effecteurs* (corps, bas du corps, torse, colonne vertébrale, bras gauche/droit, main gauche/droite), soit des données propres aux expressions faciales ou à la direction du regard. La composition est réalisée à la fois spatialement, avec des niveaux de priorité appliqués aux effecteurs, en suivant l'organisation hiérarchique structurelle du squelette, et temporellement en déclenchant au moment approprié le contrôleur de synthèse spécifique à l'effecteur considéré. Pour chaque élément du squelette, un contrôleur paramétré permet de rejouer le mouvement préenregistré avec la possibilité de lui appliquer des traitements spécifiques tels que la répétition, l'inversion, etc. La synthèse du mouvement du regard est réalisée par un modèle de cinématique inverse guidé par des positions pointées en 3D couplées aux mouvements de la tête. Puis une technique de mélange de mouvement est appliquée hiérarchiquement aux données de sortie des contrôleurs, afin de fluidifier le mouvement produit. Ce système permet,

par agencement de mouvements préenregistrés, de construire de nouveaux énoncés en substituant des signes ou des portions de phrase par d'autres, ou en modifiant des éléments linguistiques sur un ou plusieurs canaux phonologiques. Dans l'exemple de la phrase en français « J'aime les jus de fruits », transformée en « Je n'aime pas le jus d'orange », le mouvement du buste ainsi que celui du bas du corps et du bras gauche sont conservés. Par contre, les mouvements de la tête et du bras droit, ainsi que l'expression faciale sont modifiés, de façon à préserver la cohérence sémantique de la phrase (Duarte, 2012).

4.2. *Annotation par apprentissage automatique*

Nous avons développé un système d'annotation couplé au système de synthèse concaténative, qui s'appuie sur des algorithmes d'apprentissage automatique (Naert *et al.*, 2018). En effet, la précision de l'annotation, à la fois spatiale (nature et structure des pistes d'information) et temporelle (marqueurs temporels associés aux segments) conditionne la finesse d'édition du mouvement, la cohérence grammaticale du résultat et la qualité de l'animation produite. Cependant, annoter manuellement un corpus LSF constitue une tâche chronophage qui nous a conduits à opter pour un processus d'annotation automatique pour les HC et FE. L'annotation des HC est réalisée à travers une chaîne de traitements qui permet d'extraire les principaux descripteurs manuels, de segmenter les phrases à partir d'une détection de seuil sur des distances moyennes variationnelles, puis d'étiqueter ces phrases au moyen de méthodes d'apprentissage automatique (ML). Une évaluation quantitative a été réalisée sur la base d'un corpus contenant 32 classes de configurations manuelles. Un sous-ensemble de 29 distances a permis de classifier de manière optimale les HC. Les différentes méthodes utilisées ont donné des scores de bonne classification de 87 % pour la régression logistique, 89 % pour les k plus proches voisins (kNN) avec k = 3, et de 93 % pour la méthode *Support Vector Machine* (SVM). L'annotation des FE a été réalisée de manière analogue. Le processus de segmentation s'appuie sur la détection des *maxima* des dérivées première et seconde des descripteurs de *blendshape* (section 4.4). Les résultats de ML donnent une précision de 91 % pour les forêts aléatoires, de 86 % pour SVM et de 72 % pour kNN (k = 3). À la fois pour les HC et FE, l'étiquetage automatique est ensuite effectué sur la base de la classe prédominante sur chaque segment.

4.3. *Enrichissement par synthèse hybride du mouvement corporel*

La synthèse corporelle dite *hybride* a pour objectif d'enrichir la base de données initiale en adjoignant à ces données des données synthétisées avec variations flexionnelles, facilitant ainsi la création de nouveaux énoncés en LSF. En effet, la capture d'un ensemble volumineux de données de mouvements MoCap s'avère longue et fastidieuse, et il existe encore peu de corpus couvrant la grande variabilité des LS. Nous décrivons ci-après quelques techniques de synthèse qui permettent d'augmenter les données enregistrées tout en respectant les mécanismes de spatialité et d'iconicité

de la LSF. Certaines de ces techniques ont été implémentées et évaluées (pointage, modifications lexicales et syntaxiques à partir de la manipulation des HC) (Naert *et al.*, 2021) ; d'autres procédés sont spécifiés (spatialisation, verbes directionnels, SFT) en vue d'une intégration dans notre système *SignCom*.

Pointages dans l'espace de signation. Nous avons vu que l'espace de signation définit des zones spatiales discrétisées qui peuvent être exploitées dans le discours en LSF. En particulier, les gestes de pointage définissent des procédés syntaxiques de type locus/pointage qui permettent d'assurer la fonction pronominale et de référencer des entités du discours. Cependant, ces gestes, présents dans les données capturées, correspondent à un ensemble limité de zones pointées. Or, par synthèse, il est possible de produire un nombre illimité de gestes de pointage qui couvrent l'espace de signation. À cette fin, nous avons défini un modèle d'inversion cinématique (IK) qui, à partir de la seule spécification d'un ensemble de positions cibles référencées (locus), génère le mouvement de pointage vers ces cibles.

Spatialisation. Ce même procédé peut être utilisé pour répondre à la problématique de spatialisation des signes. En effet, disposant des signes réalisés dans la zone de leur ancrage lexical, la technique de l'IK permet de déplacer la main vers la zone souhaitée et de signer l'entité lexicale à cette position spécifique.

Verbes directionnels. Les verbes directionnels sont caractérisés par une trajectoire qui définit la trace du mouvement d'un locus à l'autre et permet ainsi de distribuer les rôles actanciels agent/bénéficiaire/objet dans la phrase. Notre système permet de produire de telles phrases en spécifiant les locus correspondant aux pronoms souhaités, puis par IK en synthétisant le mouvement de la main. Par exemple, la phrase en français « Je te donne un livre » peut être modélisée par l'expression paramétrée [DONNER]([PRO-1],[PRO-2],[LIVRE]) dans laquelle le verbe [DONNER] s'exécute par un mouvement de la main, depuis la localisation du pronom « je » [PRO-1] vers celle du pronom « tu » [PRO-2] avec une configuration manuelle qui est celle du signe [LIVRE]. Pour générer la phrase « Tu me donnes un livre », il suffit d'inverser le mouvement de [PRO-2] à [PRO-1]. Pour générer la phrase « Je lui donne un livre », nous spécifions le nouveau locus correspondant à la 3^e personne [PRO-3], et synthétisons par IK le mouvement de [PRO-1] à [PRO-3] tout en conservant la configuration manuelle. Pour les verbes directionnels comportant un actant objet, la phrase est signée avec une configuration manuelle qui correspond à l'objet. Ce procédé de synthèse peut être généralisé à la plupart des phrases comportant des verbes directionnels.

Spécificateurs de forme et de taille (SFT). Nous distinguons les SFT qui agissent sur les mouvements, configurations manuelles ou orientations des mains, ou sur une combinaison de ces composantes des signes, et nous proposons ci-dessous une liste non exhaustive de flexions des signes rencontrées en LSF. Ces SFT ont été modélisés et spécifiés. Ils impliquent des procédés de synthèse très différents. Seules les trois premières situations ont été implémentées.

– La taille de la trajectoire du mouvement peut être modifiée en spécifiant une trajectoire rectiligne dans le plan transversal du signeur et en synthétisant le mouvement

par une technique d'IK ou d'interpolation. C'est le cas du signe [TABLE], dans lequel les mains plates s'écartent plus ou moins en fonction de la taille de la table.

– Il est relativement simple de spécifier et de remplacer la configuration manuelle statique sur tout le signe, comme dans les signes [VERRE-FIN], [GROS-VERRE], où seule la forme de la main change (*C* plus ou moins ouvert).

– Modifier dynamiquement la configuration manuelle au cours d'un signe nécessite d'employer un modèle de cinématique directe (FK) entre deux ou plusieurs formes clés de la main. Ainsi, à partir du signe [VERRE], il est possible de générer le signe [COUPE-DE-CHAMPAGNE], dans lequel la forme de la main s'évase vers le haut.

– D'autres SFT impliquent de modifier simultanément la trajectoire et l'orientation de la main, la configuration manuelle étant statique. C'est le cas du signe [BOL] dont la taille peut varier. De la même manière, on peut modifier les trajectoires et configurations manuelles simultanément, l'orientation étant statique, comme dans le signe [BANANE]. Les techniques de synthèse supposent de générer des trajectoires plus ou moins complexes et de synthétiser par IK le mouvement des articulations des bras, et de manière simultanée de synthétiser par FK les configurations ou orientations manuelles dynamiques.

Évaluations du système global. Deux évaluations perceptuelles ont été réalisées, dans lesquelles nous avons comparé les résultats de synthèse avec les données de rejeu MoCap qui constituent la vérité terrain (Naert *et al.*, 2021). La première évaluation concerne l'étude de la spatialisation et du pointage. 57 participants de bon niveau en LSF et au-delà (très bon, natif et interprète) ont répondu à des questionnaires sur le web avec des consignes vidéo, les réponses étant fournies au moyen de textes (menus déroulants), d'images ou de vidéos. Pour la tâche consistant à reconnaître l'emplacement du signe [BOL] à travers la visualisation de 8 vidéos (5 vidéos de synthèse et 3 de rejeu), le taux de reconnaissance est de 86 % pour les signes synthétisés contre 63 % pour les signes rejoués, la différence s'expliquant par une plus grande variabilité des signes réels. L'évaluation du réalisme de ces mêmes signes a donné un score moyen de 3,6 pour les énoncés de synthèse contre 3,8 pour les énoncés de rejeu, sur une échelle de Likert de 5 points, montrant qu'il n'y a pas de différence significative entre les séquences synthétisées et les séquences de rejeu (*p-value* de 0,031). La seconde étude a permis d'évaluer le réalisme des gestes de pointage. 9 vidéos ont été présentées aux mêmes participants (6 vidéos de synthèse et 3 de rejeu). Les scores sont respectivement de 3,15 pour les gestes de synthèse et de 3,45 pour les gestes de rejeu. Là également, nous avons conclu qu'il n'y a pas de différence marquée entre les gestes réels et de synthèse (*p-value* de 0,081).

La seconde évaluation concerne les processus relatifs à la manipulation de configurations manuelles (HC), statiques ou dynamiques. Pour cette étude, 39 participants ont visualisé 20 vidéos présentées dans un ordre aléatoire (13 de synthèse et 7 de rejeu), représentant des signes issus de la dactylologie tels que [LSF] ou [OK], ou différentes démarches d'animaux (par exemple celles du coq et du chat). Pour la tâche consistant à évaluer le remplacement des HC par d'autres (5 vidéos), les résultats ont montré qu'il n'y avait pas de différence significative entre les énoncés de rejeu ou de synthèse, ces

derniers étant même jugés plus réalistes. Pour la tâche qui consiste à reconnaître les signes issus de la dactylologie (15 vidéos), les taux de reconnaissance obtenus pour la synthèse sont de 95 % pour les signes synthétisés et de 91 % pour les signes de rejeu. De plus, aucune différence notable n'a été observée en ce qui concerne le réalisme des signes synthétisés (score de 3,08/5) et rejoués (score de 3,03/5).

4.4. *Synthèse faciale*

Motivation et approche. La synthèse de mimiques faciales associées aux trois fonctions de la LSF – expression de la modalité de la phrase, expression adverbiale ou expression affective – nécessite de couvrir un large spectre de mimiques expressives, avec toute la richesse et les nuances spécifiques aux LS. Nous nous sommes intéressés plus spécifiquement aux mimiques faciales affectives dans le cadre des émotions de base (Ekman et Friesen, 1978), ainsi qu'aux mimiques exprimant les modalités négative, interrogative, exclamative ou injonctive de la phrase. Ces expressions faciales (FE) peuvent se combiner entre elles avec différents degrés d'intensité. De plus, afin de préserver la cohérence grammaticale des phrases produites, nous avons opté pour une technique de synthèse faciale basée données capturées, ce qui permet d'atteindre une grande précision, tant spatiale que temporelle (fréquence d'acquisition > 200 Hz) ainsi que la synchronie avec les mouvements corporels et manuels. Par ailleurs, nous avons adopté une représentation paramétrique unifiée qui s'appuie sur des formes clés (*blendshapes*) associées à des coefficients de pondération (dits coefficients de *blendshape*). Ces coefficients sont calculés automatiquement à partir des données de MoCap grâce à une chaîne de synthèse que nous détaillons ci-après.

Synthèse à base de formes clés (*blendshapes*). L'animation basée *blendshapes* permet de construire une expression faciale $\boldsymbol{v}$ (vecteur des N positions du maillage de l'avatar) à partir de la combinaison linéaire de N formes de base $\boldsymbol{b}_i$, chacune étant pondérée par un coefficient c_i, la forme $\boldsymbol{b}_0$ correspondant à l'expression neutre :

$$\boldsymbol{v} = \boldsymbol{b}_0 + \sum_{i=1}^{i=N} c_i \boldsymbol{b}_i \qquad [1]$$

Les formes de base représentent des expressions unitaires impliquant une petite partie du visage (sourcil, bouche, menton, etc.). La figure 3 (partie droite) illustre le principe de la synthèse par blenshapes. L'équation 1 peut se réécrire : $\boldsymbol{v} = B.\boldsymbol{c}$, où B représente la matrice des formes de base et $\boldsymbol{c}$ le vecteur des coefficients de *blendshapes*. Ce type de modèle présente plusieurs avantages. D'une part, il fournit un niveau d'abstraction permettant le transfert d'une animation d'un modèle d'avatar 3D à un autre. D'autre part, il s'agit d'un modèle linéaire très simple, ce qui conduit à des temps de calcul autorisant les applications temps réel. Enfin, cette représentation stable et régulière permet la segmentation et l'étiquetage des séquences d'expressions faciales. Une étude préalable (Reverdy *et al.*, 2015) a permis de confirmer l'hypothèse selon laquelle l'utilisation de *blendshapes* pour animer le visage permet non seulement

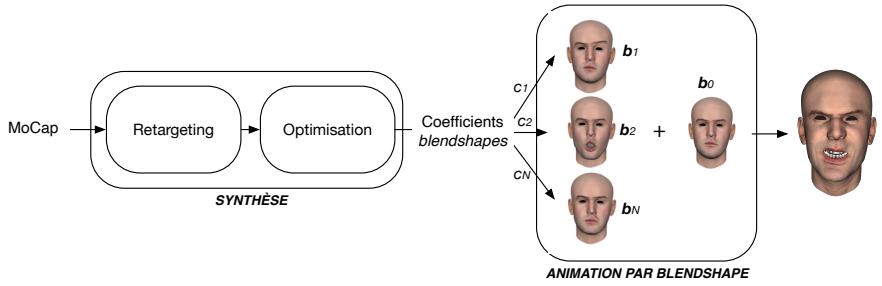

FIGURE 3. *Chaîne de synthèse faciale à partir des données MoCap*

de réduire les coûts de calcul, mais aussi de produire une animation faciale convaincante. Par la suite, afin de synthétiser des expressions précises et subtiles, nous avons choisi un grand nombre de formes de base (soit 51). Cette étude a également permis de comparer et d'évaluer plusieurs jeux de marqueurs (nombre et positionnement) grâce à une méthode de *clustering*, conduisant ainsi à un jeu optimal de 41 marqueurs.

Chaîne de synthèse. Notre méthode LSTS de synthèse faciale permet de transformer automatiquement des données 3D de MoCap faciale en coefficients de *blendshape* utilisés ensuite pour l'animation d'un avatar virtuel (Reverdy, 2019). Cette méthode comporte deux étapes principales (figure 3). La première, dite de *retargeting*, a pour objectif d'adapter morphologiquement les trajectoires enregistrées sur l'acteur afin qu'elles correspondent à la morphologie de l'avatar ciblé. Après un post-traitement qui consiste à supprimer les transformations rigides (translation et rotation) des données 3D et à aligner les données dynamiques de l'acteur sur celles de l'avatar, une méthode de régression de type *RBF* (*Radial Basis Function*) permet de résoudre ce problème d'adaptation. La deuxième étape exploite une méthode d'optimisation pour synthétiser automatiquement les coefficients de *blendshape* de l'avatar à partir des positions de marqueurs. Le problème revient à minimiser l'erreur quadratique entre les positions variationnelles des P marqueurs MoCap $\delta\boldsymbol{m} = \hat{\boldsymbol{m}} - \hat{\boldsymbol{m}}_{\mathbf{0}}$ obtenues après *retargeting* et celles des P points du maillage de l'avatar correspondant, calculées à partir du modèle de *blendshapes* $\delta\boldsymbol{a} = B_P.\boldsymbol{c}$, où B_P correspond à la projection de la matrice B sur les P points du maillage alignés sur les marqueurs MoCap :

$$\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}} ||\delta\boldsymbol{m} - \delta\boldsymbol{a}|| + \alpha_b.E_b + \alpha_t.E_t \qquad [2]$$

E_b et E_t étant deux énergies de régularisation qui pénalisent l'espace des solutions de ce problème d'optimisation. L'énergie de normalisation E_b permet d'éviter que les coefficients de *blendshape* ne sortent de l'intervalle $[0, 1]$. L'énergie Laplacienne de déformation E_t permet de minimiser la déformation de la structure du maillage par rapport au maillage original dans son expression neutre. Les paramètres de pondération α_b et α_t sont tels que $\alpha_b + \alpha_t$ = 1

Résultats. Nous avons travaillé principalement sur les expressions affectives en choisissant les six classes d'émotions primaires – la colère (C), le dégoût (D), la peur (P), la joie (J), la tristesse (T), la surprise (S) – auxquelles nous avons ajouté le neutre (N), et sur les modalités syntaxiques assertive, interrogative, exclamative et négative (Reverdy, 2019). Un corpus a permis de valider notre méthode de synthèse *LSTS* et d'animation. Il a été enregistré pour un seul signeur de niveau B2 au moyen de deux dispositifs de capture employés simultanément : le système MoCap précédent et une caméra RGB-D. Ce corpus est constitué de séquences alternant les expressions neutres et expressives pour les 6 émotions, en tenant compte de 3 degrés d'intensité différente (N – C1 – N – C2 – N – C3 – N – J1 – N – J2 – ... avec 1 : faible ; 2 : marqué ; 3 : exagéré) et de phrases expressives avec différentes modalités syntaxiques, soit au total environ 30 minutes d'enregistrement. Une première étude perceptuelle a été réalisée sur la base de ce corpus. 27 personnes (17 hommes et 10 femmes, âgés de 17 à 31 ans) ont participé en répondant à des questionnaires en ligne. 54 vidéos de synthèse réalisées en variant les facteurs de pondération des deux énergies de régularisation leur ont été présentées dans un ordre aléatoire. Tout d'abord, le choix d'un paramétrage optimal de la méthode *LSTS* a été établi en analysant subjectivement les facteurs de reconnaissance de l'émotion, d'identification de l'intensité perçue et de réalisme des animations. Les résultats ont donné pour le meilleur modèle de synthèse un taux de reconnaissance moyen de 62,5 %, les émotions les mieux reconnues étant la joie, la surprise et la colère (88 %), un score de reconnaissance de l'intensité de 5,09 (sur une échelle de Likert de 1 à 7) et un score de réalisme de 5,2/7.

Une seconde étude perceptuelle a permis de valider le modèle *LSTS* par comparaison avec le modèle *FS* proposé par *faceshift* [7] qui est une référence au niveau de l'état de l'art, en utilisant les mêmes facteurs d'évaluation auxquels on a rajouté le critère de fidélité par rapport aux vidéos originales. 41 personnes séparées en 2 groupes ont participé à l'étude. Les méthodes *FS* et *LSTS* ont donné des résultats similaires, avec des taux de reconnaissance respectifs de 62,4 pour *LSTS* et 62,6 pour *FS*, et un réalisme de 5,1/7 pour *LSTS* et de 4,9/7 pour *FS*. La fidélité des vidéos de synthèse par rapport aux vidéos réelles a donné des scores de 5,1/7 pour *LSTS* et de 4,9/7 pour *FS*.

4.5. *Synthèse du mouvement des mains*

En raison de la rapidité et de la précision des gestes de la LSF, la reconstruction des données manuelles est particulièrement longue et fastidieuse. Parmi les différentes méthodes exploitables (captation des positions des marqueurs, par gants équipés d'accéléromètres/gyroscopes, captation basée vision), l'utilisation de la MoCap à base de marqueurs présente les résultats les plus aboutis et en adéquation avec la qualité attendue pour les LS, grâce à la précision spatiale et temporelle de ces marqueurs. Elle souffre néanmoins d'une phase coûteuse de post-processing pour corriger les problèmes d'occultation, très nombreux en LS.

7. *faceshift* : `http://www.faceshift.com/product/,2012`

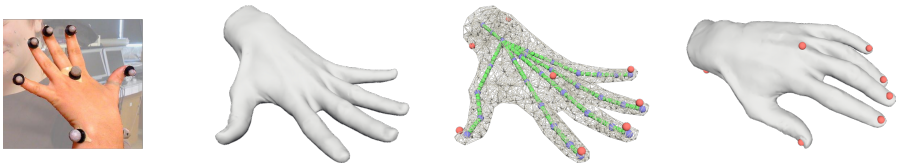

FIGURE 4. *Synthèse du mouvement des mains en 4 étapes : 1. Mains avec marqueurs réfléchissants ; 2. Maillage de référence ; 3. Maillage volumétrique intégrant le squelette et les positions de marqueurs ; 4. Posture générée par notre système*

Nous proposons une chaîne d'animation complète pour la synthèse du mouvement des mains à partir d'un jeu de marqueurs simplifié (figure 4.1). Notre méthode se fonde sur une technique de déformation Laplacienne qui intègre dans une structure de contrôle volumétrique le maillage haute résolution, le squelette, ainsi que les emplacements des marqueurs pertinents (figure 4.3). Une méthode d'optimisation itérative, qui préserve à la fois les caractéristiques géométriques, les longueurs des segments et les butées articulaires, est appliquée à cette structure. En suivant cette approche présentée dans Le Naour *et al.* (2019), nous avons montré la capacité du modèle d'optimisation à animer et à déformer interactivement des modèles de main en haute définition à partir d'un faible nombre de contraintes de position, tout en conservant tous les détails du mouvement. Ce modèle a été appliqué avec succès aux mouvements des mains et des doigts en LSF, avec la possibilité d'éditer et d'adapter les données morphologiques de façon à ce que les contraintes de contact et de non-interpénétration des doigts soient bien respectées. Une évaluation quantitative a été réalisée sur la base des erreurs *root mean square* entre la séquence synthétisée et la vérité terrain (erreur < 5 %).

5. Problèmes non résolus

Ces problèmes sont multiples. Nous présentons ci-après de manière non exhaustive ceux qui nous paraissent les plus importants au regard des TSL.

Modélisation des mécanismes flexionnels des LS. Les systèmes actuels TSL sont loin d'être complètement automatisés et d'intégrer l'ensemble des mécanismes flexionnels des LS. Cette constatation concerne à la fois les représentations linguistiques des LS et les techniques de synthèse employées. Cela peut s'expliquer par le fait que les LS sont des langues à modalité gestuelle, dont le fonctionnement et les structures linguistiques sont encore peu formalisés au sens des langages formels en informatique. Quelques procédés linguistiques ont été modélisés et traduits dans des langages informatiques dédiés (section 2). Cependant, ces langages sont encore loin de couvrir la multitude de phénomènes répertoriés, de manière générique et paramétrisée. De plus, il est nécessaire d'aller vers l'automatisation des processus, depuis la traduction des phrases en langage Pivot-LS vers la spécification de l'animation (figure 1).

Conjointement, les modèles d'animation classiques d'avatars ne peuvent s'appliquer directement aux LS car ils ne s'adaptent pas aux variations flexionnelles mentionnées. Ces deux objectifs doivent être envisagés simultanément. Cela implique de maîtriser à la fois le traitement automatique des langues signées (TALS) et les méthodes d'animation par ordinateur.

Coordination multi-effecteur et synchronisation temporelle. Du point de vue linguistique *computationnelle*, l'un des enjeux majeurs concerne la prise en compte du temps dans le formalisme de synchronisation entre les canaux d'information. Le type de langage utilisé détermine la façon dont cette synchronisation est gérée. En particulier, les langages de script ou impératifs gèrent le temps de manière explicite, ce qui peut s'avérer fastidieux pour traduire les phrases dans le formalisme choisi. Des langages réactifs permettraient d'intégrer ces éléments de synchronisation, mais ils sont en général plus difficiles à manipuler. Du point de vue de la synthèse du mouvement, l'un des défis principaux réside dans la coordination multi-effecteur et la synchronisation des mouvements générés sur chacune des pistes – soit par extraction dans la base de données, soit par synthèse –, de façon à respecter les schémas spatio-temporels des signes. En effet, pour que les signes générés paraissent plausibles, il est nécessaire que les mouvements relatifs à des groupes d'articulations respectent des schémas de synchronisation propres au contrôle moteur : par exemple la configuration manuelle doit toujours être atteinte avant que la main ait atteint son objectif cible (Duarte, 2012). De plus, si l'on mélange plusieurs styles de mouvements, la recherche d'une cohérence de style peut imposer de compresser ou de dilater des portions de mouvement de façon à ce que les règles de synchronisation soient vérifiées (Héloir et Gibet, 2007). Cette synchronisation se répercute également aux mouvements secondaires apparaissant dans certains signes. Enfin, il est primordial de gérer la coarticulation, et ceci aux niveaux intrasigne et intersigne de façon à tenir compte du contexte passé et futur de chaque signe dans la séquence générée.

Adaptation morphologique et prise en compte des contacts. La plupart des avatars signeurs utilisent les données signées d'un seul locuteur. L'adaptation morphologique à d'autres avatars signeurs passe par des processus d'adaptation morphologique (*retargeting*) permettant par exemple de transférer les animations vers d'autres personnages (homme, femme, enfant, voire animal). De plus, les contraintes des LS sont liées au contenu des signes et des énoncés. En effet, de nombreux signes impliquent des contacts entre les deux mains, ou entre chacune des mains et une partie du corps. Ces contraintes spatiales doivent être spécifiées précisément dans l'espace du signeur (par exemple « les mains doivent rester au-dessus de la table ») ou exprimées de manière qualitative (par exemple « l'index doit toucher la paume de la main »). Les algorithmes proposés pour la synthèse des mouvements manuels (section 4.5) sont capables de traiter un ensemble de ces contraintes numériques par optimisation, par exemple en ajoutant une connaissance liée à l'environnement ou en permettant le relâchement de certains degrés de liberté. Il serait intéressant d'intégrer ces contraintes dans le système de synthèse grâce à un langage utilisateur de haut niveau.

Modélisation physique. La physique des mouvements (au sens des forces mises en œuvre) fait partie intégrante des dynamiques gestuelles impliquées dans les LS. Ainsi, la façon dont les contacts sont exécutés (de manière effleurée ou frappée) modifie le sens des signes. Dans un futur proche, il paraît inévitable que les systèmes de synthèse d'avatars signeurs soient modélisés et simulés physiquement.

Rareté des données. Les données sont au cœur des technologies et méthodes employées pour les avatars signeurs. Plusieurs types de données sont disponibles : la vidéo, la capture de mouvement, les images et les textes (français écrit, LSF-glosée, annotations, etc.). Plusieurs questions se posent pour la définition du corpus. La première concerne le compromis entre étendue et profondeur du corpus. Si l'objectif est de disposer d'un lexique qui couvre un large domaine, comprenant plusieurs thématiques, un corpus étendu sera privilégié. Si, au contraire, l'objectif est d'avoir un vocabulaire limité et de le réutiliser dans différents énoncés avec flexions grammaticales, alors on choisira l'approche en profondeur. Dans ce cas, de nombreuses instances des mêmes signes avec variations doivent être considérées dans le vocabulaire prédéfini. La deuxième question concerne la nature des variations elles-mêmes qui doivent être incluses dans le corpus pour l'édition et la synthèse. Pour pallier ces difficultés, on peut à court terme remplacer les données de MoCap par des données vidéo, plus faciles à acquérir. Cela permet de disposer de gros volumes de données et d'envisager la traduction TSL en exploitant des méthodes performantes d'apprentissage profond développées dans le domaine de la vision et de l'animation par ordinateur. Enfin, une préoccupation essentielle relative à la construction du corpus est la qualité actée ou spontanée des mouvements produits par les locuteurs signants.

Traduction automatique texte-vers-LS et LS-vers-texte. Les systèmes actuels de traduction automatique d'une langue vocale vers une autre laissent entrevoir la possibilité de traduire automatiquement une langue parlée/écrite vers une LS. Les approches de *deep learning* (DL) devraient faciliter cette étape. Cependant, l'absence de système d'écriture communément accepté pour les LS ne permet pas de disposer de suffisamment de données mettant en correspondance un texte dans une langue vocale et sa transcription écrite en LS. Avec le peu de corpus parallèles disponibles, seuls des systèmes de traduction à base de règles, ou ceux portant sur un vocabulaire restreint sont actuellement en cours de développement. De nouvelles méthodes d'apprentissage frugal reposant sur des connaissances préalables devraient être capables d'aider les modèles DL à intégrer de nouveaux concepts à partir de peu d'exemples.

Par ailleurs, les approches TSL neuronales basées vidéos s'appuient sur la transformation entre données vidéo 2D et séquences de postures en 3D. Cependant, si les architectures de réseaux neuronaux conduisent à des séquences de postures relativement précises spatialement, elles n'intègrent toujours pas les configurations manuelles, ou du moins pas de manière précise. De plus, parmi les approches développées, très peu s'intéressent aujourd'hui à la qualité du mouvement produit. Or, la précision des configurations manuelles en LS, et la manière dont les séquences de LS se déroulent dans le temps, constituent des enjeux majeurs pour le développement des systèmes de traduction automatique TSL.

Notons que la reconstruction de séquences de postures squelettiques 3D à partir de vidéo 2D est un moyen de constituer des bases de données conséquentes associant vidéo et MoCap. Celles-ci peuvent être exploitées pour la reconnaissance de signes à partir de vidéos ou pour la synthèse de mouvements à partir de texte. Il serait certainement intéressant de s'affranchir du passage par le squelette 3D, et de produire directement du texte en français à partir de vidéos LS (pour la reconnaissance), ou des animations LS à partir de texte en français (pour la synthèse).

Enfin, il subsiste la question de l'alignement vidéo/texte qui n'est pas résolue. Il est nécessaire de l'aborder dans le langage pivot ou à défaut dans la langue vocale.

6. Conclusion et perspectives

Dans cet article, nous avons abordé les questions principales qui se posent pour la traduction et la synthèse en LS à partir d'énoncés textuels, et décrit les avancées principales de notre système TSL *SignCom*. Ce système intègre au niveau de sa conception les bases permettant de produire des contenus en LSF en respectant les procédés grammaticaux répertoriés. En particulier, il s'appuie sur la composition multimodale de segments de mouvements attachés d'une part à des composantes linguistiques de la LSF, et d'autre part à des contrôleurs de synthèse spécifiques. La possibilité d'éditer les énoncés, depuis le niveau phonologique jusqu'aux niveaux lexical, syntaxique et discursif permet ainsi de construire de nouveaux énoncés en LSF. Plusieurs modules de synthèse ont été intégrés au système initial. Tout d'abord, il devient possible de synthétiser des mécanismes flexionnels de la grammaire de la LSF qui s'appuient sur la spatialité et les dynamiques iconiques de cette langue. Notamment notre système permet de générer automatiquement des processus de type (locus/pointage/spatialisation) des signes, facilitant ainsi l'agencement des référents dans l'espace de signation. Il permet également de générer de manière flexible certains procédés propres à l'iconicité, comme les proformes manuelles lexicales ou syntaxiques. Un module de synthèse des expressions faciales a également été implémenté et évalué. Ce module génère automatiquement, à partir de données capturées, des données d'animation faciale qui peuvent être utilisées dans le système d'édition de manière similaire aux données de mouvement. Compte tenu du corpus enregistré, il est capable de synthétiser les qualités affectives et modales des mimiques faciales. De plus, un module spécifique de synthèse du mouvement des mains a été développé pour s'adapter aux exigences de précision et de rapidité des LS. La prise en compte d'un maillage volumétrique des mains, incorporant un modèle de squelette permet d'atteindre un niveau de précision jusqu'à présent inégalé, tout en évitant les interpénétrations des mains entre elles et avec les autres parties du corps. La synthèse du mouvement des mains ainsi que l'animation faciale ont donné des résultats très satisfaisants. Enfin, un procédé d'annotation automatique, appliqué aux configurations manuelles et expressions faciales permet d'accélérer la constitution des bases de données annotées pour la synthèse concaténative.

Si les avancée des systèmes TSL avec avatars signeurs sont très prometteuses, de nombreuses questions de recherche restent ouvertes. En particulier, l'un des enjeux majeurs consiste à mieux intégrer les travaux sur les formalismes de représentation des LS et les systèmes d'animation. De plus, la grande variabilité de ces langues visuo-gestuelles et la complexité des mécanismes de flexion qu'elles sous-tendent, nécessitent la mise en œuvre de processus de modélisation dédiés, d'un point de vue linguistique et animation, ce qui ouvre des voies de recherche encore peu explorées. Dans un futur proche, la possibilité de capturer de grands volumes de données et le développement des méthodes d'apprentissage automatique profond vont conduire à des systèmes automatiques de synthèse texte-vers-LS ou LS-vers-texte, dans la mesure où ils intègrent une connaissance linguistique des LS. Plus largement, cela ouvre des perspectives vers des systèmes de traduction automatique des langues vocales vers les LS ou vice-versa, ou d'une LS vers une autre.

7. Bibliographie

Battison R., *Lexical borrowing in American sign language*, ERIC, 1978.

Blondel M., Tuller L., Lecourt I., « Les pointés et l'acquisition de la morphosyntaxe en LSF », *La linguistique de la LSF : recherches actuelles. Silexicales 4*, In Berthonneau, A-M. and DAL, G. (eds), Univ. Lille 3, p. 17-32, 2004.

Brour M., Benabbou A., « ATLASLang MTS 1 : Arabic Text Language into Arabic Sign Language Machine Translation System », *Proc. Computer Science*, vol. 148, p. 236-245, 2019.

Chaaban H., le Gouiffès M., Braffort A., « Automatic Annotation and Segmentation of Sign Language Videos : Base-level Features and Lexical Signs Classification », *Int. Conf. VISIGRAPP 2021, Vol. 5*, p. 484-491, 2021.

Chételat-Pelé E., Braffort B., « Sign Language Corpus Annotation : toward a new Methodology », *LREC 2008, Marrakech, Morocco*, ELRA, 2008.

Cuxac C., *La langue des signes française (LSF) : les voies de l'iconocité (French) [French Sign Language : the iconicity ways]*, Faits de langues, Ophrys, 2000.

Dauriac B., Braffort A., Bertin-Lemée E., « Example-based Multilinear Sign Language Generation from a Hierarchical Representation », *Sign Language Translation and Avatar Technology ; Junction of the Visual and the Textual ; Challenges and Perspectives*, p. 21-28, 2022.

Duarte K., Motion Capture and avatars as Portals for Analyzing the Linguistic Structure of Sign Languages, PhD thesis, Université Bretagne Sud, 2012.

Ebling S., Glauert J. R., Kennaway J., Marshall I., Safar E., « Building a Swiss German Sign Language avatar with JASigning and Evaluating it among the Deaf community », *Universal Access in the Information Society*, vol. 15, p. 577-587, 2016.

Efthimiou E., Fontinea S., Hanke T., Glauert J., Bowden R., Braffort A., Collet C., Maragos P., Goudenove F., « Dicta-sign–sign language recognition, generation and modelling : a research effort with applications in deaf communication », *Workshop on the Representation and Processing of Sign Languages, LREC 2010*, p. 80-83, 2010.

Ekman P., Friesen W., *Facial Action Coding System : A Technique for the Measurement of Facial Movement.*, Consulting Psychologists Press, 1978.

Elliott R., Glauert J. R., Kennaway J., Marshall I., Safar E., « Linguistic modelling and language-processing technologies for Avatar-based sign language presentation », *Universal Access in the Information Society*, vol. 6, n° 4, p. 375-391, 2008.

Farooq U., Rahim M., Sabir N., Hussain A., Abid A., « Advances in machine translation for sign language : approaches, limitations, and challenges », *Neural Computing and Applications*, 11, 2021.

Filhol M., McDonald J., « Extending the AZee-Paula shortcuts to enable natural proform synthesis », *Workshop on the Representation and Processing of Sign Languages, LREC 2018*, Japan, 2018.

Filhol M., McDonald J., Wolfe R., « Synthesizing Sign Language by connecting linguistically structured descriptions to a multi-track animation system », *Int. Conf. on Universal Access in Human-Computer Interaction*, Springer, p. 27-40, 2017.

Garcia B., Sallandre M.-A., Schoder C., L'Huillier M.-T., « Typologie des pointages en Langue des Signes Française (LSF) et problématiques de leur annotation », *in* Boutora, L. et Braffort, A (eds), *TALN 2011*, Montpellier, France, p. 107-119, 2011.

Gibet S., « Building French Sign Language Motion Capture Corpora for Signing Avatars », *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018.

Gibet S., Courty N., Duarte K., Le Naour T., « The SignCom System for Data-driven Animation of Interactive Virtual Signers : Methodology and Evaluation », *ACM Transactions on Interactive Intelligent Systems*, vol. 1, 2011.

Gibet S., Lebourque T., Marteau P., « High level Specification and Animation of Communicative Gestures », *Journal of Visual Languages and Computing*, vol. 12, p. 657-687, 2001.

Gibet S., Lefebvre-Albaret F., Hamon L., Brun R., Turki A., « Interactive editing in French Sign Language dedicated to virtual signers : requirements and challenges », *Universal Access in the Information Society*, vol. 15, n° 4, p. 525-539, 2016.

Glauert J., Elliott R., « Extending the SiGML notation : a progress report », *Int. Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, vol. 23, 2011.

Héloir A., Gibet S., « A Qualitative and Quantitative Characterization of Style in Sign Language Gestures », *Gesture Workshop*, 2007.

Héloir A., Kipp M., « Real-time animation of interactive agents : Specification and realization », *Applied Artificial Intelligence*, vol. 24, n° 6, p. 510-529, 2010.

Huenerfauth M., Generating American Sign Language classifier predicates for English-to-ASL machine translation, PhD thesis, University of Pennsylvania, 2006.

Huenerfauth M., Lu P., Kacorri H., « Synthesizing and Evaluating Animations of American Sign Language Verbs Modeled from Motion-Capture Data », *SLPAT@Interspeech*, 2015.

Johnson R., Liddell S., « A segmental framework for representing signs phonetically », *Sign Language Studies*, vol. 11, n° 3, p. 408-463, 2011.

Johnston T., « The Lexical Database of AUSLAN (Australian Sign Language) », *Proceedings of the First Intersign Workshop : Lexical Databases*, Hamburg, 1998.

Kahlon N., Singh W., « Machine translation from text to sign language : a systematic review », *Universal Access in the Information Society*, 07, 2021.

Kennaway R., Glauert J. R., Zwitserlood I., « Providing signed content on the Internet by synthesized animation », *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 14, n° 3, p. 15, 2007.

Kipp M., Héloir A., Nguyen Q., « Sign Language Avatars : Animation and Comprehensibility », *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, 2011.

Liddell S., Johnson R., « American Sign Language : The phonological base », *Sign Language Studies*, vol. 64, n° 6, p. 195-278, 1989.

Lombardo V., Nunnari F., Damiano R., « A virtual interpreter for the Italian sign language », *International Conference on Intelligent Virtual Agents*, Springer, p. 201-207, 2010.

McDonald J., Filhol M., « Natural synthesis of productive forms from structured descriptions of sign language », *Machine Translation*, vol. 35, n° 3, p. 363-386, 2021.

McDonald J., Wolfe R., Schnepp J., Hochgesang J., Jamrozik D., Stumbo M., Berke L., Bialek M., Thomas F., « An automated technique for real-time production of lifelike animations of American Sign Language », *Universal Access in the Information Society*, vol. 15, n° 4, p. 551-566, 2016.

Millet A., *Grammaire descriptive de la langue des signes française : dynamiques iconiques et linguistique générale*, UGA Editions, 2019.

Naert L., Larboulette C., Gibet S., « A survey on the animation of signing avatars : From sign representation to utterance synthesis », *Comput. Graph.*, vol. 92, p. 76-98, 2020.

Naert L., Larboulette C., Gibet S., « Motion synthesis and editing for the generation of new sign language content », *Machine Translation*, vol. 35, n° 3, p. 405-430, 2021.

Naert L., Reverdy C., Larboulette C., Gibet S., « Per channel automatic annotation of sign language motion capture data », *Workshop on the Representation and Processing of Sign Languages : Involving the Language Community, LREC 2018*, 2018.

Núñez-Marcos A., de Viñaspre O. P., Labaka G., « A survey on Sign Language machine translation », *Expert Systems with Applications*, vol. 213, p. 118993, 2023.

Prillwitz S., Zentrum H., *HamNoSys : Version 2.0 ; Hamburg Notation System for Sign Languages ; An Introductory Guide*, Signum-Verlag, 1989.

Reverdy C., Data-driven annotation and synthesis of facial expressions in French sign language, PhD thesis, Université Bretagne Sud, 2019.

Reverdy C., Gibet S., Larboulette C., « Optimal marker set for motion capture of dynamical facial expressions », *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, ACM, p. 31-36, 2015.

Roelofsen F., Esselink L., Mende-Gillings S., Smeijers A., « Sign Language Translation in a Healthcare Setting », *In Translation and Interpreting Technology Online (TRITON)*, p. 110-124, 2021.

Sallandre M.-A., Garcia B., « Semiological Approach to Sign Languages and "gloss-based notations" : Issues related to SL sub-units annotation », *Hesperia : Anuario de Filología Hispánica*, vol. 22, p. 57-79, 2020.

Stokoe W. C., *Semiotics and Human Sign Language*, Walter de Gruyter Inc., 1972.

Stoll S., Camgoz N. C., Hadfield S., Bowden R., « Text2Sign : towards sign language production using neural machine translation and generative adversarial networks », *International Journal of Computer Vision*, vol. 128, p. 891-908, 2020.