Projet LS-COLIN. Quel outil de notation pour quelle analyse de la LS ?

Annelies Braffort[4], Annick Choisier[4], Christophe Collet[4], Christian Cuxac[1], Patrice Dalle[3], Ivani Fusellier[1], Rachid Gherbi[4], Guillemette Jausions[3], Gwénaëlle Jirou[1], Fanch Lejeune[2,4], Boris Lenseigne[3], Nathalie Monteillard[1], Annie Risler[1], Christophe Rolet[1], Marie-Anne Sallandre[1]

1- Sciences du Langage, Université Paris 8
2- CAMS-LaLIC, université Paris 4
3- IRIT-TCI, université Toulouse 3
4- LIMSI, CNRS, Orsay

Introduction

En 1999, le Ministère de l'Education Nationale, de la Recherche et de la Technologie lançait une action concertée incitative de recherche en sciences de la cognition, appelée "Cognitique". L'objectif était d'accompagner le développement remarquable des sciences de la cognition en favorisant des synergies entre les différentes disciplines concernées : neurosciences, psychiatrie, psychologie, linguistique, philosophie, anthropologie, informatique, mathématiques, logique, intelligence artificielle, robotique, etc… Il s'agissait en particulier de favoriser des collaborations entre, d'une part, les sciences humaines et sociales, et, d'autre part, les sciences du cerveau et / ou le secteur de l'informatique, des mathématiques et des sciences pour l'ingénieur.

C'est dans ce cadre qu'est né le projet LS-COLIN : *" Langues des signes : Analyseurs privilégiés de la faculté de langage; apports croisés d'études linguistiques, cognitives et informatiques (traitement et analyse d'image) autour de l'iconicité et de l'utilisation de l'espace* ". Il regroupe des équipes des universités de Paris 8 (Sciences du langage), Paris 4 (CAMS-LaLIC) et Toulouse3 (IRIT-TCI) et du CNRS (LIMSI). La responsabilité scientifique est assurée par Ch Cuxac, linguiste, professeur à Paris 8.

Cet article décrit les objectifs du projet, la démarche adoptée et les premiers travaux menés depuis le démarrage de l'action, à la fin de l'année 2000.

1. Présentation de l'action LS-COLIN

1.1. Hypothèses sur la Langue des Signes

L'étude des langues des signes permet d'entrevoir au plus près ce qu'est une langue, en remettant en question les frontières habituelles de ce que nous considérons comme appartenant au domaine du "linguistique".

a- Le cadre théorique

Nous proposons une analyse linguistique des langues des signes, dont le principe fondateur est la théorie de l'iconicité définie par Cuxac [Cuxac, 1996 & 2000], elle-même inscrite dans une sémiologie plus générale. Cette théorie peut s'appliquer à toutes les langues des signes (LS) : les LS standardisées comme la LSF, mais aussi les autres langues comme la LSI (Langue des signes internationale), les langues de petites communautés de Sourds et enfin les langues des signes primaires mises au point par des Sourds isolés.

Dans cette théorie, il est essentiellement question de l'iconicité dite d'image, c'est-à-dire du lien de ressemblance formelle entre le signe et ce à quoi il réfère, dans le monde réel.
Ce type d'iconicité structure le lexique des langues des signes, et rend compte de la constitution des signes dits standard. Cependant, un autre type d'iconicité, diagrammatique, intervient de façon massive dans la structure syntaxique des énoncés en langue des signes. On note une ressemblance entre l'agencement des signes dans l'espace et le rapport des éléments dans l'espace de référence. Cette iconicité, qui n'est pas spécifique aux langues signées, trouve son origine au niveau langagier des représentations sémantico-cognitives. Elle n'est cependant plus autant perceptible dans les langues vocales [Haiman, 1985] du fait de l'obligation de linéariser l'énoncé oral. Au contraire, elle se maintient selon certaines règles en langues signées du fait de leur expression dans une modalité gestuelle spatio-temporelle.

Nous savons grâce aux travaux de Goldin-Meadow, Yau, Fusellier-Souza que des personnes sourdes vivant en milieu entendant ou vivant en petites communautés ont mis en œuvre des stratégies communicationnelles fondées sur un processus d'iconicisation de l'expérience perceptivo-pratique. Celui-ci est à l'origine de toutes les langues des signes pratiquées dans le monde, quel que soit leur statut institutionnel. De ce fait, ces langues constituent aujourd'hui un terrain d'observation privilégié des faits synchroniques concernant la sémiogénèse (formation de signes ou de systèmes de signes par une communauté linguistique donnée) des langues.

Dans cette perspective sémiogénétique, la bifurcation postulée par Cuxac [Cuxac, 2000] détermine deux pôles qui coexistent et entre lesquels le va-et-vient est constant, dans tous types de discours. On a d'une part la grande iconicité, traces structurales d'une visée illustrative (ou iconicisatrice) de l'expérience vécue, qui donne à voir tout en disant, d'autre part les signes standard, sans visée illustrative. C'est par ces deux grands axes qu'une grammaire de la LS peut être établie.
Il est étonnant d'observer que les langues des signes naturelles utilisent des structures autant syntaxiques que sémantiques présentant un grand degré d'iconicité. Ainsi, on peut exprimer des idées complexes sans recourir au lexique standard. Cependant, il est intéressant d'analyser à quel moment du discours les signes standard sont nécessaires et inversement, et dans quels cas la grande iconicité est obligatoire (par exemple dans le cas d'énoncés absurdes qui ne peuvent être traduits que par un transfert personnel) ; on cite souvent l'exemple de Cuxac [1996] "le chocolat mange le garçon".
Nous tentons de démontrer que ces structures de grande iconicité, trop souvent considérées comme de la pantomime, sont bien des éléments linguistiques et constituent même la démonstration la plus subtile et la plus convaincante de ce qu'est une langue. Elles sont quasiment identiques d'une langue des signes à l'autre, ce qui permet une intercompréhension rapide entre des signeurs de nationalités différentes.
Toutefois, aucune étude systématique n'ayant encore été menée sur ces structures, des questions fondamentales demeurent.

Les grammaires cognitives apportent un cadre d'analyse pertinent à tous les niveaux d'analyse de la LS. Ces grammaires offrent effectivement des points de vue intéressants sur une langue iconique et spatiale. Elles nous permettent d'envisager l'iconicité pour les langues orales donc a fortiori pour les langues gestuelles. Elles envisagent le passage du cognitif au linguistique en trois niveaux, cognitif, langagier et de la langue. Aux deux premiers niveaux se réalisent des opérations de représentation, cognitives et langagières communes à toutes les langues. Chaque langue détermine son propre système, à travers son lexique et sa syntaxe, mais les opérations langagières sont toujours du même type. Une langue gestuelle peut s'inscrire dans un tel

modèle. Enfin, et surtout, ce modèle recours à un vocabulaire abstrait topologique et spatial pour formaliser les opérations langagières, ainsi qu'à des schématisations qui rappellent fortement l'agencement des signes dans l'espace de signation. On peut mettre en évidence en LSF une construction de l'espace du même ordre que les constructions envisagées par les grammaires cognitives : repérages, orientation de la relation prédicative, positionnement de l'espace par rapport au signeur.

b- sur l'apport du traitement d'image

Les recherches menées en France sur la LS tant en linguistique qu'en informatique, ont ceci en commun qu'elles exploitent des données réelles de terrain en s'appuyant sur des corpus d'images vidéo. Cette démarche nécessaire à l'analyse de la LS, est néanmoins délicate à mener, car elle nécessite la réalisation d'une transcription (linguistique) ou étiquetage (informatique) des séquences vidéo, en tenant compte de divers aspects :

- la quadrimensionnalité du support ;
- la prise en compte simultanée de paramètres sémiotiques tels que les gestes, le regard, la mimique faciale ainsi que les mouvements du visage et du corps ;
- la difficulté du maniement de systèmes de transcription tenant compte de tous ces effets de sens ;
- l'inadaptation des rares logiciels d'aide à la transcription (quand ils sont accessibles) des documents vidéos.

Cette transcription rend le travail d'analyse extrêmement long et fastidieux.

Un domaine de recherche particulièrement intéressant concerne l'élaboration de systèmes informatiques dédiés à la LS, dont le but est de faire de l'analyse, de la reconnaissance ou de la génération automatique. Dans le cadre de notre étude, nous utilisons des caméras vidéo, qui permettent de capter les mouvements des mains, mais aussi du buste, les expressions faciales ainsi que le regard. Les séquences d'images ainsi enregistrées servent de support aux études informatiques notamment dans le domaine du traitement d'images.

Pour les chercheurs spécialisés en traitement d'image, l'analyse de séquences vidéo de signeurs en LS présente un intérêt particulier :

- C'est un des rares cas où l'image véhicule un sens explicite, où elle est produite avec une intention de communication et où les mouvements sont produits en respectant les règles d'une grammaire (contrairement aux scènes habituellement traitées en vision par ordinateur). De ce fait, on peut exploiter cet aspect sémantique et cette base linguistique pour introduire de nouveaux mécanismes d'analyse, de représentation et de reconnaissance (choix des primitives visuelles, analyse multi-niveaux, prédiction pilotant l'analyse, etc.).
- De plus, l'étude des relations spatiales et la conception d'outils d'interaction portant sur l'organisation spatiale, sont des éléments essentiels qui peuvent être ré-exploitées dans le cadre de systèmes d'Interaction Homme-Machine (IHM), ou dans celui d'IHM de système d'interrogation d'images ou de construction de scènes.

1.2. Objectifs de l'action

Nous croyons qu'une recherche croisée de la linguistique et de l'informatique peut nous apporter des réponses intéressantes concernant le fonctionnement interne des langues des signes. La technologie informatique peut formuler une modélisation de paramètres qui pourrait valider l'hypothèse de la bifurcation fonctionnelle prédite par C. Cuxac. Concernant la recherche sur l'espace en LSF, le recours à l'analyse d'image peut s'intégrer de deux

manières : d'une part afin de mettre à l'épreuve les paramètres formels dégagés, qui ne sont pas de type phonologique mais morphémique ; d'autre part, afin d'objectiver la construction de l'espace telle qu'elle est réalisée dans un énoncé, par la mise en relief des portions d'espace pertinisées, et des relations entre les espaces topologiques.

Pour réaliser cette étude nous avons dans un premier temps confronté nos approches de l'analyse de corpus vidéo sur la LS. Nous avons ensuite défini les objectifs suivants :

- Extraire des primitives signifiantes de séquences vidéo utiles pour la reconnaissance et l'interprétation tant au niveau linguistique qu'au niveau informatique ;
- Trouver et quantifier des structures dans une séquence d'image (événements périodiques, configurations...) ;
- Mettre au point un système de transcription manuel à partir de vidéo numérique adapté aux besoins des linguistes et des informaticiens en tenant compte des différentes approches.

2. Etat de l'art et présentation de la démarche retenue

La première étape du projet a donc été de faire le point sur les différentes démarches d'analyse d'un corpus de LS et de recenser les formalismes existants et les environnement informatiques les mettant en œuvre.

De manière pragmatique, il était intéressant d'étudier la variabilité et les constantes de plusieurs transcriptions sur la base de la même séquence vidéo. Un extrait de la vidéo " Blanche neige " d'IVT a été transcrit par des linguistes et informaticiens du groupe. Après analyse, on a obtenu presque autant de transcriptions que de transcripteurs. Il a fallu donc faire un effort pour expliciter la démarche de chacun des transcripteurs. En dehors des différences au sein des linguistes et au sein des informaticiens, il est apparu clairement que la démarche était guidée selon ce que chacun cherchait : Le linguiste savait ce qu'il cherchait et tentait de le trouver (segmenter donc) dans la séquence, alors que l'informaticien avait moins de connaissance *a priori* et tentait d'extraire des évènements spatio-temporels.
Notre objectif à ce stade était de trouver un mode de transcription commun aux linguistes et aux informaticiens qui permettrait :

- de passer entre les niveaux (du formel au signifié), dans les deux sens,
- de guider, cibler voire anticiper les traitements d'images (prédiction-vérification)

Avant de décrire plus en détail notre démarche, un rapide tour d'horizon des différents systèmes de transcription et d'écriture nous permettra ensuite de nous positionner et de définir l'intérêt de travailler sur le développement d'un système de transcription adaptée à la langue des signes.

2.1. Systèmes d'écriture et de transcription

Trois types de systèmes existent : systèmes classiques d'écriture monolinéaire, systèmes dits " en partition " et éditeurs multimédias associant la vidéo à la transcription.

a- Systèmes d'écriture

Les systèmes classiques de transcription sont basés sur une décomposition linéaire des signes gestuels en différents paramètres caractéristiques (configuration, orientation, emplacement, mouvement et mimique en général) et une écriture de leur forme signifiante par l'intermédiaire de symboles. Comparé à Stokoe (1960) qui ne paraît pas exhaustif et s'inscrit davantage dans une démarche phonologique, Hamnosys (1985) semble s'imposer et cherche à décrire l'ensemble des phénomènes qui concourent à la réalisation d'un signe et s'apparente

d'avantage à un API du geste. Si de tels systèmes posent la problématique d'une écriture, ils paraissent insuffisants en terme de représentativité et ne permettent pas de transcrire de manière intéressante les structures de grande iconicité (ils visent essentiellement des énoncés standards - signes discrets -). Ils ne rendent pas compte non plus convenablement de l'utilisation pertinente de l'espace (rapports spatiaux, mémorisation des emplacements, place du signeur). Il s'agit moins d'outils de transcription autorisant une réelle démarche d'analyse. Par ailleurs, l'écriture linéaire (et se limitant à la forme signifiante) peut rendre la lisibilité opaque et pas toujours pertinente.

b- Systèmes de transcription " en partition *"*

Ils sont multilinéaires permettant à la fois une lecture temporelle d'un paramètre (axe horizontal) ou une lecture de la simultanéité (axe vertical). Ils répondent davantage aux besoins de recherche en étant aussi le support d'une analyse et d'une compréhension. Ils sont économiques offrant la transcription d'un paramètre seulement quand il paraît pertinent dans la compréhension du sens d'un énoncé, sans perdre en représentativité de phénomènes de l'espace et de l'iconicité. Par ailleurs, à la place des valeurs signifiantes des signes sous forme symbolique, on voit apparaître des notations décrites au moyen de mots de la langue écrite dominante apportant davantage de clarté (glose). Néanmoins ces infiltrations ne sont pas sans poser des problèmes (influence de la langue dominante, choix du vocabulaire) et il paraît important pour mieux comprendre ces notations de ne pas se dissocier de la nature même de la langue des signes. L'exploitation en direct de la vidéo nous semble indispensable.

c- Editeurs multimédias

Les logiciels actuellement développés associent la vidéo en complément des annotations (Signstream, 1995 ; Syncwriter, 1994). Signstream par exemple relie chaque énoncé à sa séquence visuelle et sa transcription sous forme de partition en proposant différents champs à l'utilisateur en fonction de la nature des évènements encodés. Dans de tels outils, l'intégration de la vidéo facilite le processus de segmentation et d'annotation en assurant une meilleure cohésion et une meilleure transparence des informations retenues et en rendant moins fastidieux le travail de transcription que celui qui est fait à partir d'un magnétoscope. De tels éditeurs permettent en outre de répondre à différents problèmes en terme de recueil et d'analyse de données : accessibilité, vérification, comparaison, stockage et recherche d'informations. Nous nous orientons vers un tel outil, cependant par rapport aux éditeurs existants, le nôtre doit pouvoir intégrer différents niveaux d'analyse en proposant une meilleure structuration et hiérarchisation des données. Une distinction entre ce qui relève de la description et de l'interprétation des données d'une part et une réflexion sur les concepts utilisés d'autre part sont nécessaires. Dans ce sens, la vidéo ne doit pas se limiter semble-t-il à améliorer l'ergonomie et la lisibilité des systèmes de notation mais une analyse d'image devrait à terme permettre l'interprétation des données à partir d'indices visuels.

2.2. Notre démarche

Les traitements informatiques des images, avec peu ou pas de connaissance sur la LS, s'efforcent d'extraire des évènements spatio-temporels, de les isoler, sans référer directement à la signification en langue. Les motivations de tels traitements sont décrits dans la section précédente.

Pour un travail linguistique, la transcription sert de notation pour le corpus, et pas de support à l'analyse. Elle est destinée à fournir au lecteur un moyen de se représenter la séquence gestuelle analysée. Vu les limitations des systèmes de transcription existants, il faut noter qu'aucun des linguistes du groupe ne se sert de ceux-ci, chacun a mis au point sa propre façon de transcription. La notation permet de mettre en évidence, sur l'axe temporel, des

distinctions jugées pertinentes au regard de la signification de l'énoncé. Après avoir compris le sens, les linguistes segmentent en fonction d'une grille préalable d'analyse et notent sur des partitions l'apparition des éléments sélectionnés comme étant pertinents dans l'analyse. La notation est postérieure à l'analyse, et la transcription proposée est en quelque sorte une justification du découpage opéré. Le choix des paramètres formels mentionnés, nombre de lignes de la partition, dépend des options théoriques retenues par chacun, conduisant à de nombreux éléments d'interprétation.

Les niveaux d'observation, formel ou phonologique, morphologique, syntaxique, d'interprétation ou sémantique sont souvent mal distingués, comme en témoigne l'emploi d'une nomenclature qui ne rend pas compte des différents niveaux d'analyse. Cette réflexion sur les types de transcription a permis dans un premier temps de prendre conscience de la nécessité de distinguer les niveaux d'analyse dans la transcription et du besoin de clarifier les paradigmes de nomenclature respectifs. Elle a aussi mis en lumière les possibilités d'aller plus loin ensemble, et de faire d'un outil d'écriture un véritable outil de transcription et d'analyse.

Démarche cognitivo-linguistico-informatique

Il est intéressant de rapprocher le besoin d'une transcription lisible, et les possibilités données par l'analyse d'image d'isoler des variations formelles et de les segmenter. La notation en partition, qui rend bien compte de la multiplicité des paramètres, ainsi que de la dimension temporelle, ne permet pas facilement de mettre en évidence l'utilisation de l'espace. Le recours aux images est incontournable :

- images fixes avec des indications de mouvement,
- séquence incrustée au-dessus de la partition,
- ou découpage en séquences, avec images initiales et finales.

Mais ceci nécessite de s'interroger sur la nature des images à exploiter en fonction de nos objectifs. C'est l'objet d'un travail en cours sur les conditions d'acquisition et d'enregistrement (format, angle de vue, nombre de caméras, qualité…). Il faut aussi susciter préalablement un travail en collaboration sur plusieurs plans : d'une part les linguistes doivent distinguer entre les appellations utilisées dans les différents niveaux d'analyse, recenser les critères de segmentation ainsi que les familles de paramètres (mimiques, configurations manuelles, postures) ou les combinaisons de paramètres complémentaires (regroupement en familles morphologiques) ; d'autre part les informaticiens doivent unifier leurs modèles pour une meilleure intégration des modules de traitement et des représentations de données.

Ce travail devrait faciliter l'analyse d'image par à la fois une démarche d'intégration de modules et une démarche descendante favorisant la prédiction des positions et l'anticipation des mouvements. La segmentation spatio-temporelle et ses critères discriminants sont au cœur de nos préoccupations communes : en traitement automatique des images et en délimitation des unités linguistique à différents niveaux.

Dans ce contexte, deux objectifs privilégiés sont à l'étude, la qualité des images et les critères de segmentation, pour arriver à un outil de transcription qui serve à la fois :

- en dehors du support multimédia, à fournir une notation sur un support papier,
- à partir d'une séquence d'images réelles, de viser la reconstitution du schème sémantico-cognitif spatio-temporel sous-jacent, et le mode de construction de l'espace,
- et d'espérer réaliser (dans un avenir plus lointain), une animation virtuelle d'un énoncé.

Après avoir mené ce travail de clarification sur les objectifs de l'analyse aussi bien linguistique qu'informatique, et parallèlement à la constitution d'un corpus conçu pour

permettre ces analyses, des travaux ont pu démarrer sur la réalisation d'un outil de transcription et sur les modules de traitement d'image qu'il pourrait utiliser.

3. Réalisation d'un éditeur de partition

3.1 - Spécification d'un éditeur entièrement manuel

Rôle de l'éditeur

Le rôle de l'éditeur de partition est de permettre la visualisation et la manipulation de séquences vidéo de LSF et la réalisation de transcriptions de vidéos selon des critères utiles à la fois aux linguistes et aux informaticiens.

- *Visualisation et manipulation de séquence vidéo* : L'éditeur permet la visualisation de films numérisés à la manière d'un magnétoscope : parcours avant, arrière, arrêt sur image, ralenti….

- *Transcription de vidéos* : Il permet aussi de réaliser une transcription de vidéos, c'est-à-dire associer à une ou plusieurs images des informations symboliques ou numériques. Une fois la transcription réalisée, on peut la sauvegarder afin de pouvoir la restituer lors d'un chargement ultérieur de la séquence vidéo.

Aspect graphique de l'éditeur

L'éditeur est constitué de deux parties (Figure 1). La partie supérieure contient la visualisation de la séquence vidéo, avec une série d'icônes à cliquer représentant les fonctionnalités classiques des magnétoscopes. La partie inférieure comporte la transcription associée.

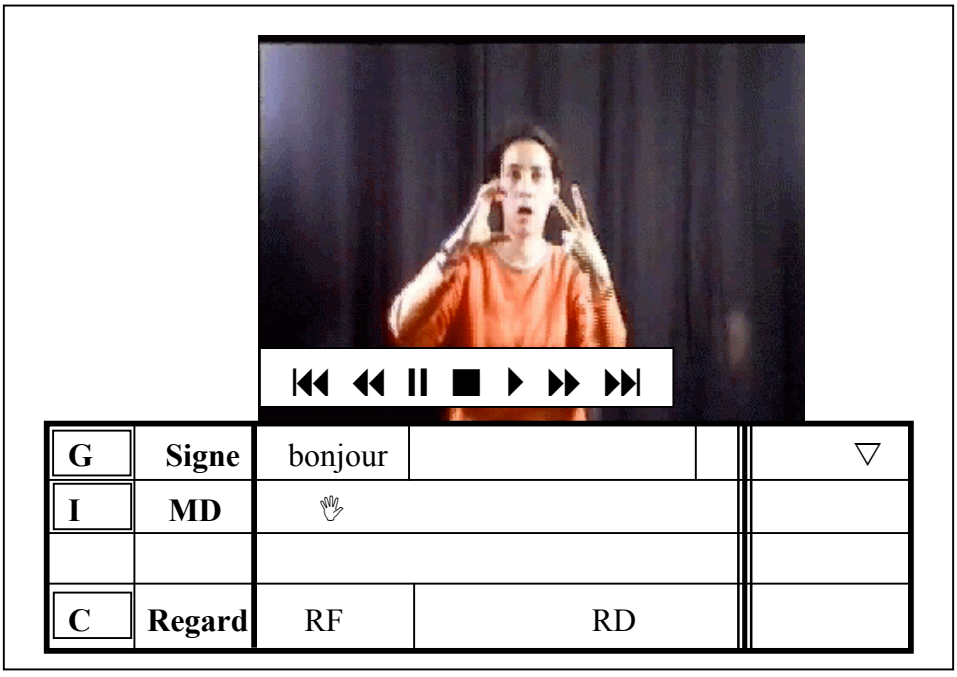

Figure 1 : aspect graphique de l'éditeur

Description détaillée de la partie transcription

La transcription est représentée sous forme d'une partition. L'axe horizontal représente le temps et l'axe vertical contient l'ensemble des paramètres sélectionnés (ex : direction du regard, mouvement des mains, fonction syntaxique, type de transfert…).

Axe horizontal :

Chaque paramètre est décrit au sein d'une bande horizontale :

- La première section à partir de la gauche contient un bouton qui permet de choisir le mode de représentation de la transcription. Ce mode de représentation peut être de plusieurs types :
  - I : icône (ex : " ✋ " pour configuration main plate),
  - C : code (ex : " RF " pour Regard Face),
  - G : description littérale (ex : " boujour "),
  - V : valeur numérique (ex : 5cm).
- La deuxième section contient le nom du paramètre (ex : " Signe ", " MD " pour Main Dominée).
- Les sections suivantes contiennent les valeurs correspondantes du paramètre pour une séquence temporelle donnée (ex : " boujour "). Lorsque c'est possible, un menu contextuel déroulant donne la liste des valeurs disponibles pour un paramètre donné.
- On peut trouver des bandes de paramètre " fermées ", pour lesquelles il y a des valeurs sur toute la bande et des bandes " ouvertes ", décrivant des événements à un moment donné de la séquence vidéo.

Il est possible par ailleurs d'établir des liaisons entre différentes parties de la transcription :

- Soit des liens auxquels on peut associer une référence appartenant à la liste des références déjà rencontrées, ou une référence " à créer ",
- Soit des références (spatiales, temporelles,…) pouvant être liées aux liens existant ou à venir.

Cela permet par exemple d'associer un classificateur au signe standard auquel il fait référence.

Axe vertical :

- On peut regrouper certains paramètres au sein d'un " groupe de paramètres " visuellement identifié lorsque cela est justifié (ex : main dominante + main dominée + deux mains).
- Une barre verticale synchronisée avec le défilement de la vidéo peut se déplacer le long de la transcription selon l'axe temporel.
- Il est possible de grouper verticalement des paramètres avec des zones de couleurs différentes ou en utilisant la transparence, afin par exemple de visualiser les zones relatives aux signes standards et celles de grande iconicité.

Une transcription vidéo est composée de plusieurs partitions différentes. En effet, un utilisateur peut vouloir étudier plusieurs extraits d'une même vidéo ou transcrire de plusieurs manières différents la même vidéo.

Généricité, flexibilité, ouverture

L'éditeur peut être utilisé à la fois par des linguistes et par des informaticiens. Il est donc prévu des facilités de personnalisation et de modification de la partie transcription.

Chaque utilisateur peut choisir :

- les paramètres qu'il souhaite étudier
- le type d'affichage pour chaque paramètre (icône, code…)
- l'ordre dans lequel ces paramètres sont affichés.

Un utilisateur peut récupérer une transcription déjà réalisée et modifier la forme de la présentation.

Chaque utilisateur peut ajouter :

- de nouveaux paramètres
- de nouvelles valeurs aux paramètres

Pour cela, des profils utilisateurs sont associés à chaque transcription. Ainsi, à la réouverture d'une transcription donnée par un utilisateur donné, celui-ci retrouve la configuration de l'éditeur telle qu'il l'avait définie précédemment.

3.2. Possibilité d'automatisation partielle par traitement d'images

Les informations reportées dans les différents niveaux de la partition proviennent d'une analyse et d'une interprétation, par le linguiste, de la séquence d'images (détermination des composantes de l'image et analyse des mouvements). Afin de l'aider dans cette analyse, et donc de faciliter l'utilisation de l'éditeur de partitions, nous avons étudié dans quelle mesure le traitement d'image permettrait d'automatiser certaines tâches.

Nous allons présenter les différents problèmes que l'on peut tenter de résoudre par traitement d'image, d'une façon générale, puis spécifier ce qui peut être réalisé dans le cadre limité du projet LS-COLIN et enfin nous présenterons les études menées et les résultats intermédiaires obtenus.

a- Fonctionnalités du traitement d'image

a1. Analyse au niveau d'une image

Les paramètres de la LS figurant dans la transcription concernent des éléments du corps, comme la main ou les sourcils ; le TI doit donc savoir retrouver leurs correspondants dans l'image. On va donc segmenter l'image en zones ayant des propriétés visuelles communes comme la couleur (zones de peau), ou localiser des configurations de pixels particulières (indices visuels). Ce traitement peut être entièrement automatisé ou au contraire mené en coopération avec l'utilisateur, celui-ci indiquant la zone approximative de l'élément à étudier ou désignant un pixel de cet élément, le système de TI déterminant ensuite les frontières précises de l'élément.

Ces éléments peuvent être composés; on doit donc savoir structurer les composants élémentaires en composants plus complexes (yeux + nez + bouche + joues + front $\Rightarrow$ visage).

Ces paramètres sont enfin caractérisés par un ensemble de valeurs intrinsèques (position de la main) ou relationnelles (main à la hauteur de la bouche). Ces valeurs peuvent être déterminées par programme, en effectuant des mesures directes dans l'image (yeux fermés),

ou en reconstruisant l'information 3D associée (direction du regard). Elles concernent les éléments de base (orientation de la main) ou des éléments composés (visage souriant)

A partir des mesures effectuées sur les entités 2D ou 3D ou à partir des indices visuels et en exploitant des connaissances a priori (modélisation des composants du corps), on peut identifier les entités (région de couleur peau + taille + forme ⇒ main), en différents niveaux d'abstraction suivant la complexité de l'information introduite dans le programme (région de couleur peau ⇒ main ⇒ main droite ⇒ main dominante ⇒ signe standard …).

a2. Analyse au niveau de la séquence d'images

La transcription concerne aussi le mouvement des mains ou de corps, éléments qu'il faut donc savoir suivre dans la séquence d'images. Les paramètres portent alors sur une caractérisation de ce mouvement (vitesse, direction, signature de trajectoire, …). On retrouve ici aussi la notion de structuration, les mouvements élémentaires pouvant être combinés en des configurations plus complexes (balancement), et la notion de mesures relatives (trajectoires parallèles).

Enfin le mouvement étant fugitif, il sera intéressant, pour faciliter son analyse par le linguiste, de pouvoir le matérialiser et de le visualiser en superposition dans l'image (techniques de " réalité augmentée "). Il s'agit donc de construire une image qui enregistre l'historique du mouvement en rendant compte à la fois de son occupation spatiale et de son déroulement temporel.
D'une façon plus générale, il serait très pertinent de pouvoir représenter l'espace de signation, d'y faire figurer les référents mis en place au cours du discours et de matérialiser les zones désignées par les pointeurs.

a3. Remarque

La reconnaissance d'un élément ou d'une configuration peut souvent être établie à partir de sa signature visuelle dans l'image ou dans une image transformée. Par contre les mesures ou les descriptions faisant référence à l'espace de signation (trajectoires, pointages, détection d'un mouvement vers l'avant, …) nécessitent une reconstruction partielle 3D. L'image seule ne suffit alors pas car des informations ont été irrémédiablement perdues lors du processus de formation de l'image (projection perspective, occultations). Il faut donc disposer d'informations supplémentaires pour pouvoir reconstruire l'information 3D ; elles peuvent être obtenues par :

- raisonnement (étude des lignes de fuite, de la variation de la taille d'un objet,…)
- ajout de connaissances supplémentaires (par exemple sur la taille réelle des composants) ou de contraintes sur la géométrie de la scène.
- ajout de données supplémentaires :
    - en utilisant une seconde image, prise depuis un point de vue différent, et en appliquant les techniques de stéréovision,
    - en utilisant plusieurs images dans le temps et en interprétant les mouvements dans les images.

Dans notre cadre de travail actuel (séquence d'images vidéo mono-caméra), les possibilités de reconstruction 3D sont très limitées.

b- L'apport du TI dans le projet LS-COLIN

b1. Les problèmes abordés

Dans le cadre du projet LS-COLIN, compte-tenu de la durée du projet et de son financement, seules les fonctions de base pourront être étudiées et expérimentées. Leur évaluation dans différents contextes (analyse de la robustesse) et l'étude des fonctionnalités évoluées débordent du cadre de ce projet.
Ces fonctionnalités de base concernent l'extraction d'entités et la reconnaissance de certaines d'entre elles, un premier jeu de mesures, l'extraction et la mise en évidence des mouvements principaux.

b2. Résultats obtenus

La première partie du projet ayant consisté à définir et à spécifier ces fonctionnalités, à partir d'une étude des besoins des linguistes, les résultats sont encore limités. De plus, en l'absence de dispositif d' acquisition de séquences d'images numériques et d'un protocole rigoureux de prise de vue, au début du projet, nous avons été contraints de traiter des images de qualité médiocre, disponibles dans le domaine public et non réalisées pour cette étude.

- Réalisation de l'interface de pilotage d'une séquence vidéo

Cette interface fournit les fonctions usuelles d'un lecteur vidéo et permet de récupérer les données associées à l'image courante, pour leur traitement. Elle constitue le cadre 1 de l'interface décrite en 3.1

- Extraction des zones d'intérêt

Deux modules ont été réalisés :

- extraction du locuteur, par comparaison avec une image de référence du fond, filtrage et traitement morphologique. Ce module fournit la silhouette du locuteur.

- extraction des entités mains et tête. Ce module utilise une signature de la couleur caractéristique de la peau, dans l'espace HSV, pour sélectionner les zones candidates, puis il identifie la tête et les mains sur des critères de taille et de forme.

- Visualisation du mouvement

Nous avons implémenté une représentation des motifs temporels, en adaptant au cas de gestes de la LS, les tMHI (Time Motion History Image) de [Davis 97]. Une séquence d'image est résumée en une seule image indiquant comment s'est fait le mouvement. L'intensité de chaque pixel est fonction de l'historique du mouvement. Cette représentation permet à la fois de visualiser le mouvement et de fournir une information globale et locale sur le mouvement, pouvant être utilisée pour la reconnaissance de configurations dynamiques.

ces réalisations ne constituent qu'une première étape. Les travaux se poursuivent dans quatre directions :

- Segmentation, traitement d'images et suivi, pour rendre ces modules plus robustes
- Mesures : taille, position, distance des entités de traitement d'image, signature de trajectoires
- Détection d'événement, indices spatio-temporels
- Visualisation de mouvements, de marqueurs

Conclusion

La collaboration d'équipes provenant de disciplines différentes, pour l'étude d'un même objet, ici la langue des Signes, renforce automatiquement l'exigence de rigueur et de précision sur les objectifs de l'étude, sur ses constituants, sur les formalismes et les modèles utilisés et sur la terminologie employée. La première phase du projet LS-COLIN a mis en évidence cette exigence. La participation de plusieurs linguistes ayant un regard focalisé sur des aspects différents de la LS impose de faire l'inventaire des différents niveaux d'analyse. L'étude d'un

outil de transcription rend nécessaire l'explicitation des constituants qui interviennent (partie du corps, paramètres du geste, …). De même la segmentation temporelle de la partition impose, pour qu'elle puisse être effectuée par le traitement d'image, de savoir la justifier en terme d'indices visuels repérables dans la séquence d'images. Ce n'est qu'à l'issue de ce travail qu'on peut déterminer les contraintes à respecter dans la définition et l'acquisition d'un corpus de vidéo de locuteurs en LS.

Nous avons décrit la démarche utilisée dans la première phase du projet LS-COLIN pour mener à bien ce travail de définition et de spécifications. La deuxième phase devrait permettre de fournir des outils élaborés sur ces bases et permettant une analyse rigoureuse de la langue, et facilitant ainsi les échanges entre les linguistes.