UNIVERSITÉ DE LORRAINE | BIBLIOTHÈQUES UNIVERSITAIRES

# AVERTISSEMENT
# DROIT D'AUTEUR

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document. Toute contrefaçon, plagiat ou reproduction illicite encourt une poursuite pénale.

Contact bibliothèque : ddoc-theses-contact@univ-lorraine.fr
(Cette adresse ne permet pas de contacter les auteurs)

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10

UNIVERSITÉ DE LORRAINE
34 COURS LÉOPOLD — CS 25233
54052 NANCY CEDEX
TÉL : 03 54 50 54 00
CONTACT@UNIV-LORRAINE.FR
WWW.UNIV-LORRAINE.FR

UNIVERSITÉ DE LORRAINE

**École doctorale IAEM Lorraine**

# Vers une approche interdisciplinaire fondée sur les données pour l'analyse multimodale et la modélisation computationnelle du geste co-verbal

## THESE

Présentée et soutenue publiquement le 27 avril 2026

pour l'obtention du

**Doctorat de l'Université de Lorraine**
**(mention informatique)**

par

Mickaëlla Grondin Verdon

**Composition du jury**

| | | | | |
|---|---|---|---|---|
| *Président :* | M. | **Bernard GIRAU** | Professeur des Universités | UMR 7503 |
| *Encadrants :* | M. | **Slim OUNI** | Professeur des Universités | UMR 7503 |
| | Mme. | **Domitille CAILLAT** | Maitresse de conférences | UMR 5267 |
| *Rapporteurs :* | Mme. | **Annelies BRAFFORT** | Directrice de recherche CNRS, HDR | UMR 9015 |
| | M. | **Christophe PARISSE** | Chargé de recherche INSERM, HDR | UMR 7114 |
| *Examinateurs :* | Mme. | **Gaëlle FERRE** | Professeure des Universités | UMR 15076 |
| | Mme. | **Magalie OCHS** | Professeure des Universités | UMR 7020 |

**Laboratoire Lorraine de Recherche en Informatique et ses Applications** UMR 7503
**Praxiling** UMR 5267

# Remerciements

Je tiens en premier lieu à remercier Slim Ouni, mon directeur de thèse, ainsi que Fabrice Hirsch, pour m'avoir donné l'opportunité de travailler sur ce projet doctoral et pour la confiance et la bienveillance dont ils ont fait preuve à mon égard. Je remercie également Domitille Caillat ma co-encadrante de thèse, sans qui cette thèse n'aurait pu atteindre un tel aboutissement. Je lui suis profondément reconnaissante pour son soutien constant, son écoute attentive et son engagement sans faille dans ce projet. Merci à mes deux encadrants pour leurs retours constructifs, leur disponibilité et la relecture attentive de ce manuscrit. Enfin, je remercie sincèrement les équipes Multispeech et Praxiling pour leur accueil. Je tiens tout particulièrement à remercier Nasser-Eddine Monir pour sa présence, nos nombreuses discussions, et les bons conseils qu'il a pu m'apporter, ainsi que Sam Bigeard, dont l'enthousiasme constant à chacun de mes passages au laboratoire a rendu ces moments particulièrement agréables.

De très nombreuses heures ont été consacrées à ce travail de thèse, et ce aux côtés de Noah Hinschberger. Merci infiniment à lui pour toutes ces heures passées presque tous les jours, week-ends compris, à se motiver mutuellement sur nos projets doctoraux respectifs, et pour avoir été présent pour me rassurer et m'aider dès qu'il le pouvait.

Je me sens très chanceuse d'avoir été si bien entourée tout au long de cette aventure, et je remercie chaleureusement tous ceux avec qui j'ai partagé des moments inoubliables en dehors de la thèse, et qui m'ont accompagnée jusqu'à son aboutissement. Je pense notamment à Pierre, Florian, Cylia, Manyl, Anna, et Noah, pour tous ces week-ends à Strasbourg, et soirées sur Discord ; à Antoine Thill, pour avoir été présent à la fois à Montpellier et à Nancy, et tous ces moments où nous avons dansé ensemble ; à Félix Poiroux pour tous les envois d'Halloween ; à Alexis et Pauline, pour ces weekends rénovation que j'espère réitérer, les sessions crafts, et nos échanges de manhwa ; et enfin, à Frédérique Sueur, toujours présente pour m'écouter dès que nécessaire et pour nos sessions Mario Kart.

Je remercie du fond du cœur le Dr Juillé Emmanuelle et le Dr Christelle Quenot, pour leur accompagnement médical sans faille, ce depuis plusieurs années, leur écoute, et pour avoir toujours fait en sorte que je puisse poursuivre mon doctorat malgré les difficultés.

Enfin, je remercie ma famille pour son soutien, et plus particulièrement ma maman, Angélique Verdon. Merci à toi de m'avoir convaincue d'aller en lycée général, pour avoir cru en moi et en la possibilité que je poursuive ma scolarité malgré les obstacles. Je remercie également la famille Le Verge–Sérandour pour leur présence et leur soutien, et tout particulièrement Yohan Le Verge–Sérandour, envers qui je n'ai plus les mots pour exprimer toute ma reconnaissance, pour toutes ces années partagées et tous ces moments précieux, notamment ceux avec Méiose et Robert, et pour tous ceux que nous continuerons à partager.

Je dédie cette thèse à
Lise-Véronne et Hornet.

**Garama !**

# SOMMAIRE

**INTRODUCTION** - 11 -

**PARTIE I FONDEMENTS THEORIQUES ET ENJEUX DES ETUDES DE LA GESTUALITE** - 25 -

CHAPITRE 1 L'ETUDE DU GESTE DANS LE CHAMP DES *GESTURES STUDIES* - 27 -
CHAPITRE 2 L'ETUDE DE LA GESTUALITE EN SCIENCES COMPUTATIONNELLES - 39 -
BILAN DE PARTIE - 57 -

**PARTIE II RESSOURCES ET DONNEES MULTIMODALES** - 61 -

CHAPITRE 3 CORPUS MULTIMODAUX POUR L'ETUDE DE LA GESTUALITE : ENJEUX ET NECESSITE - 63 -
CHAPITRE 4 ANALYSE CRITIQUE DE L'ENRICHISSEMENT D'UN CORPUS MULTIMODAL - 81 -
CHAPITRE 5 PROTOCOLE D'ENRICHISSEMENT DE CORPUS DE GESTES - 95 -
BILAN DE PARTIE - 109 -

**PARTIE III APPORTS ET PERSPECTIVES DE L'EXPERTISE DES SCIENCES DU LANGAGE** - 113 -

CHAPITRE 6 L'ANNOTATION MULTIMODALE COMME OUTIL DE MODELISATION DU GESTE CO-VERBAL - 115 -
CHAPITRE 7 L'ANNOTATION MULTIMODALE COMME OUTIL D'EVALUATION DES GESTES SYNTHETIQUES - 147 -
BILAN DE PARTIE - 165 -

**PARTIE IV APPORTS ET PERSPECTIVES DE L'EXPERTISE DES SCIENCES COMPUTATIONNELLES** - 169 -

CHAPITRE 8 ANNOTATION AUTOMATIQUE DES ASPECTS FORMELS DU GESTE MANUEL CO-VERBAL - 171 -
CHAPITRE 9 SEGMENTATION SEMI-AUTOMATIQUE DU GESTE SOUS CONTRAINTE DE DONNEES D'ENTRAINEMENT LIMITEES - 195 -
BILAN DE PARTIE - 225 -

**CONCLUSION : LES GAINS D'UNE APPROCHE INTERDISCIPLINAIRE** - 229 -

**ANNEXES** - 235 -

**BIBLIOGRAPHIE** - 243 -

**RESUME** - 259 -

**ABSTRACT** - 261 -

**TABLE DES MATIERES** - 263 -

**TABLE DES FIGURES** - 271 -

**TABLE DES TABLEAUX** - 275 -

# GLOSSAIRE

| | |
|---|---|
| **AD** | Durée moyenne, *average duration* |
| **Ann** | Annotatrice |
| **AS** | Geste bimanuel asymétrique |
| **ASR** | Reconnaissance automatique de la parole, *automatic speech recognition* |
| **B** | Taille du batch |
| **Ba** | *Baseline* |
| **BCE loss** | *Binary Cross-Entropy loss* |
| **BEAT** | *Body Expression Audio Text* |
| **BF** | Aller-retour, *back and forth* |
| **BH** | Geste bimanuel, *bimanual hands* |
| **biRNN** | RNN bidirectionnel, *bidirectionnal recurrent neural network* |
| **BML** | *Behavior Markup Language* |
| **CH** | Main fermée, *closed hand* |
| **CID** | *Corpus of International Data* |
| **CNN** | Réseau de neurones convolutifs, *convolutional neural network* |
| **COSMOS** | *CO-speech gesture Segmentation on Motion Sequences* |
| **D** | Paume orientée vers le bas, *down* |
| **D1** | Dimension primaire |
| **D2** | Dimension secondaire |
| **E** | Paume orientée sur la tranche, *edge* |
| **ECA** | Agent conversationnel animé, *embodied conversational agents* |
| **Ext.** | Outlier extrême |
| **Fe** | *Features* |
| **F1-frame** | F1-score appliqué aux frames |
| **F1-segment** | F1-score appliqué aux segments |
| **FN** | Faux négatif |
| **Fps** | *Frame per second* |
| **Fr** | Paume vers l'interlocuteur, *front* |
| **GESRes** | *Gesture Exploration and Synthesis Resource* |
| **GNN** | Réseau de neurones graphiques, *graph neural network* |
| **GRU** | *Gated Recurrent Unit* |
| **GRUb** | GRU bidirectionnel |
| **GUI** | *Gesture User Interface* |
| **H** | Position horizontale |
| **I** | Position intermédiaire |
| **IE** | Index engagé |
| **IEMOCAP** | *Interactive Emotional Dyadic Motion Capture* |
| **IoU** | *Intersection over union* |
| **IPU** | Unités inter-pausales, *inter pausal units* |
| **IQR** | Ecart interquartiles, *interquartile range* |
| **LH** | Main gauche, left hand |
| **Loss** | Fonction de perte |
| **LSTM** | *Long Short Term Memory* |
| **MCH** | Main partiellement fermée, *mid closed hand* |
| **MD** | Paume partiellement vers le bas, *mid-down* |

| | |
|---|---|
| **MDu** | Durée médiane, *median duration* |
| **MED** | Médiane |
| **MFA** | *Montreal Forced Aligner* |
| **MFCC** | *Mel-frequency cepstral coefficients* |
| **MH** | Mains en miroir, *mirror hands* |
| **MLP** | Perceptron multicouches, *multilayer perceptron* |
| **Mocap** | Capture de mouvement, *motion capture* |
| **MOY** | Moyenne |
| **MP** | Matrice de poids |
| **MU** | Paume partiellement vers le haut, *mid-up* |
| **Nat.** | Données naturelles |
| **NLP** | Traitement automatique du langage naturel, *natural language processing* |
| **OH** | Main ouverte, *open hand* |
| **Out.** | Valeur aberrante, *outlier* |
| **PCA** | Analyse en composantes principales, *principal component analysis* |
| **PH** | Geste bimanuel avec les mains parallèles, *parallel hands* |
| **PL** | Boucle progressive, *progressive loop* |
| **PS** | Progression scandée, *progressive step* |
| **Q1** | Premier quartile |
| **Q3** | Troisième quartile |
| **Quat** | Quaternions |
| **R1** | Référentiel initial de BEAT |
| **R2** | Référentiel centré sur le locuteur |
| **RB** | Bloc résiduel, *residual bloc* |
| **RH** | Main droite, *right hand* |
| **RIL** | Moulinet, *repetitive inversed loop* |
| **RNN** | Réseau de neurones récurrents, *recurrent neural network* |
| **PReLu** | ReLu paramétrique |
| **SD** | Ecart-type, *standard deviation* |
| **SEI** | Index partiellement engagé, *semi-engaged index* |
| **Sig** | Fonction sigmoïde |
| **SO** | Paume orientée vers le locuteur, *self-oriented* |
| **SOH** | Main partiellement ouverte, *semi-open hand* |
| **ST-GCN** | Réseaux convolutifs graphiques spatio-temporels, *spatio-temporal graph convolutional networks* |
| **Synt.** | Données de synthèse |
| **T** | Nombre de *frames, time* |
| **TE** | Pouce engagé, *thumb engaged* |
| **TIE** | Pouce et index engagés, *thumb index engaged* |
| **TMIE** | Pouce, majeur et index engagés, *thumb middle index engaged* |
| **TN** | Vrai négatif, *true negative* |
| **Tot** | Total |
| **TP** | Vrai positif, *true positive* |
| **Tps** | Temps |
| **t-SNE** | *t-distributed Stochastic Neighbor Embedding* |
| **U** | Paume vers le haut, *up* |
| **V** | Position verticale |

# Introduction

## 1 La gestualité, un champ d'étude en soi

### 1.1 Geste et *Gesture studies*

#### 1.1.1 Un objet à la croisée de plusieurs disciplines

L'analyse du discours bénéficie d'une tradition de recherche ancienne et riche. L'étude des gestes remonte à l'Antiquité, où ils étaient d'abord perçus comme des gestes oratoires, majoritairement centrés sur les mouvements des mains. Des auteurs tels que Cicéron (106-43 av. J.-C) ou Quintilien (35-96 apr. J.-C) ont été parmi les premiers à les commenter et à en proposer des classifications rudimentaires. Ce n'est toutefois qu'à partir du XVI^e siècle que l'intérêt scientifique pour les gestes commence à se structurer de manière plus systématique. Dans les premières approches, les gestes étaient généralement interprétés comme des manifestations visibles de caractéristiques physiques ou d'intentions internes, servant à rendre perceptibles les émotions ou les pensées. Cette conception analogique se prolonge dans le domaine des arts visuels : des figures comme Leon Battista Alberti (1404-1472) ou Léonard de Vinci (1452-1519) ont ainsi cherché à capturer la valeur expressive des gestes dans leurs œuvres, considérant le corps humain comme un vecteur de sens artistique. Durant le siècle des Lumières, l'étude des gestes s'oriente vers des questionnements sur l'origine du langage. Le geste y est vu comme une forme corporelle primitive de communication, notamment chez l'enfant. Denis Diderot (1713-1784) reconnaît explicitement son rôle central dans la communication des personnes sourdes, anticipant ainsi certaines réflexions modernes sur les langues signées. À cette époque, le geste est encore souvent pensé comme un substitut ou un prédécesseur de la parole.

Le XIX^e siècle marque une rupture : les gestes commencent à être étudiés comme un système autonome, doté de propriétés sémiotiques propres. Divers travaux décrivent alors les gestes dans différentes communautés, en les classifiant selon leur forme ou leur signification. Cette période voit également l'émergence des premières réflexions structurées sur les dimensions sémantiques, syntaxiques et pragmatiques du geste. Wilhelm Wundt propose ainsi l'une des premières analyses [1] du lien entre gestes et parole, en s'intéressant particulièrement aux langues des signes (envisagées comme langues gestuelles) qu'il considère comme des systèmes linguistiques à part entière. La fin du XIX^e siècle connaît cependant un essoufflement de cet intérêt. Ce n'est qu'au début du XX^e siècle qu'un nouveau tournant s'opère : les chercheurs cessent de considérer le geste uniquement comme une trace des origines du langage, pour l'intégrer pleinement à la communication verbale. Le geste n'est plus conçu comme universel ou archaïque, mais comme profondément dépendant des normes culturelles et sociales. Ce déplacement marque une transition majeure : la communication est désormais pensée comme un phénomène fondamentalement multimodal. Cette approche se consolide, grâce aux travaux

de chercheurs majeurs tels que Ogston (XIX$^{e}$ siècle), Birdwhistell (1918-1994), McNeill (1933-), Ekman (1934-2025) et Friesen (XX$^{e}$ siècle), Kendon (1934-2022), Kita (1963-), Condon (XX$^{e}$ siècle) ou de Ruiter (XXI$^{e}$ siècle). Tous contribuent à redéfinir le geste non plus comme un simple accompagnement de la parole, mais comme un constituant essentiel de l'interaction verbale. Ils insistent sur l'intégration dynamique et fonctionnelle des gestes au discours oral, soulignant leur rôle dans la structuration, l'expressivité et la mise en relation intersubjective.

Ainsi, le champ de la gestualité s'affirme progressivement comme un domaine à la croisée de la linguistique, de la psychologie, de l'anthropologie et des sciences cognitives. Depuis la seconde moitié du XX$^{e}$ siècle, l'analyse du langage se développe dans une perspective résolument multimodale. L'objectif est désormais de mieux comprendre les mécanismes profonds qui sous-tendent nos gestes, et d'éclairer la nature complexe et subtile des relations entre parole et geste au sein de la communication interpersonnelle.

### 1.1.2 Le geste comme acte communicatif

Le concept de communication a été largement étudié dans de nombreuses disciplines scientifiques, offrant des perspectives variées sur sa nature multifacette [2]. Les recherches en psycholinguistique définissent la communication comme la transmission intentionnelle d'un contenu propositionnel [3]. Les perspectives sociologiques insistent sur l'influence des normes culturelles dans les processus de communication et d'interaction [4], en soulignant notamment l'importance des indices non verbaux dans la construction des relations et des dynamiques sociales [5]. Par ailleurs, l'approche écosystémique envisage la communication comme un phénomène intégrant des éléments verbaux et non verbaux, porteurs de sens dans un contexte donné [6]. Ces différentes approches enrichissent la compréhension de la communication humaine en mettant en lumière son caractère inévitable et dynamique. C'est un processus complexe reposant sur de multiples canaux permettant de transmettre du sens et de faciliter l'interaction. Au-delà de la simple transmission d'informations, la parole constitue un moyen de créer du lien, de favoriser la coopération et de naviguer dans les interactions sociales.

Les composantes de la communication recouvrent un ensemble large de phénomènes, distinguant par exemple les signes verbaux des signes non verbaux. Les signes non verbaux peuvent à leur tour être caractérisés selon leur nature statique ou dynamique, leur caractère intentionnel ou non, ainsi que leur implication, ou non, dans l'exercice de la communication. Dans ce continuum, seuls certains mouvements du corps participent réellement à la communication et sont produits intentionnellement par les locuteurs : les gestes. Kendon [7, 8] définit ainsi le « geste » comme une action visible délibérément porteuse de sens. McNeill [9, 10], quant à lui, restreint cette définition aux mouvements spontanés des mains et des doigts qui accompagnent la parole. Dans la mesure où de nombreuses études se concentrent principalement sur les gestes manuels, la définition proposée par McNeill est aujourd'hui l'une des plus couramment adoptées. Toutefois, le geste peut aussi être considéré plus largement, comme le propose Calbris [11] (définition qui rejoint celle de Kendon), en tant que tout mouvement visible d'une partie du corps effectué dans une visée communicative. Dans ce cadre,

les mains, les pieds, le regard, les mouvements de la tête, ou encore les expressions faciales peuvent constituer des objets d'étude de la gestualité.

## 1.2 Geste et traitement informatique

### 1.2.1 L'essor des approches outillées

Les gestes ont longtemps été étudiés principalement par observation directe, souvent dans le cadre d'analyses qualitatives portant sur des extraits restreints et choisis. L'émergence des nouvelles technologies a toutefois marqué une rupture méthodologique, en permettant d'enregistrer et de traiter ces données de manière informatisée. Ces outils ont ouvert la possibilité d'analyser les gestes sur des durées et quantités bien plus conséquentes et dans des contextes expérimentaux variés. Ils offrent également l'opportunité d'intégrer différents formats complémentaires, tels que l'audio, le texte, et surtout la vidéo, devenue centrale pour l'étude de la gestualité [12]. Dans ce cadre, l'élaboration de corpus audiovisuels s'est imposée comme une étape essentielle à l'ère de l'analyse de données de parole multimodale depuis la fin du XX^e^ siècle. L'accès à des ensembles de données plus larges et plus diversifiés est supposé garantir des études plus objectives, en s'éloignant du « *cherry picking* » des analyses qualitatives classiques et en faisant de la quantité de données un gage de robustesse [13]. Entre les années 80 et le début du XXI^e^, les vidéos, enregistrées au format `VHS`, devaient être visionnées sur un magnétoscope. Le chercheur ralentissait fréquemment la vitesse de lecture pour examiner avec précision les nuances gestuelles et leur synchronisation avec le discours. La transcription des propos, ainsi que l'annotation des gestes et d'autres comportements, s'effectuait alors à l'aide d'un logiciel de traitement de texte [14], de tableurs ou encore par recopie des trajectoires gestuelles par transparence directement sur l'écran.

Dans le domaine des sciences du langage, l'exploitation efficace des corpus a ensuite été facilitée par le développement de logiciels spécialisés pour l'annotation et l'enrichissement des données, tels qu'ELAN [15] ou PRAAT [16]. Ces outils permettent de systématiser l'annotation, de synchroniser plusieurs modalités (audio, vidéo, transcription) et d'ouvrir la voie à des analyses statistiques et comparatives sur de larges ensembles. Ces nouvelles approches ont ainsi permis d'améliorer la précision temporelle des analyses, jusqu'au niveau de la milliseconde, tout en offrant la possibilité d'effectuer des extractions de données de manière plus rapide.

Du point de vue du domaine de l'informatique, la dynamique des grands corpus émergents a encouragé l'essor d'analyses à plus grande échelle, quantitatives et descriptives, rendues possibles par l'augmentation des capacités de stockage et de partage, ainsi que par la montée en puissance du calcul informatique. Au fil du temps, l'informatique a apporté de nouveaux paradigmes de modélisation, de nouvelles méthodes d'expérimentation et d'évaluation, et de nouvelles perspectives pour formaliser et analyser les données. Les corpus numériques ne constituent donc pas seulement une ressource : ils façonnent le paysage contemporain de l'analyse gestuelle en permettant la combinaison de nouvelles approches qualitatives et quantitatives, et en offrant d'autres perspectives de reproductibilité des résultats et de cumuler d'avantage de données. L'approche computationnelle de la gestualité propose dès lors un cadre

unifié reliant l'observation (analyses à très grande échelle de données de parole multimodales), la modélisation (développement de modèles statistiques ou neuronaux et représentations numériques automatiques) et la prédiction (identification automatique de phénomènes communicatifs, tels que les gestes). Ainsi, des outils tels que les systèmes de transcription automatique, comme le *speech-to-text* ou l'ASR, ou la capacité à modéliser des dépendances non linéaires complexes, à apprendre des représentations latentes et à formaliser mathématiquement des concepts, facilitent l'intégration et l'exploitation des données multimodales. L'un des avantages majeurs des approches outillées réside dans la capacité à exploiter pleinement différentes méthodologies complémentaires de traitement et d'analyse de données, compétences encore largement sous-exploitées dans les sciences du langage.

### 1.2.2 Approche computationnelle fondée sur les modèles d'apprentissage

Au fil du temps et au gré des approches, le geste a été étudié à travers une variété de supports et de formats, depuis des dessins manuels jusqu'aux dispositifs plus récents tels que la capture de mouvement, qu'elle soit basée sur l'estimation 3D à partir de données 2D ou sur la reconstruction 3D articulée à partir de capteurs corporels. Les progrès matériels – caméras à haute résolution, capteurs de profondeur plus précis, systèmes de capture multicaméras et dispositifs de *tracking* avancés – ont considérablement amélioré la qualité des enregistrements, permettant de collecter des données de mouvements plus détaillées et plus fiables. Parallèlement, les outils informatiques d'analyse de données ont ouvert de nouvelles possibilités de représentation et de traitement des gestes. L'avènement de l'animation 3D d'avatars a ainsi marqué une étape importante, en offrant la possibilité de modéliser la communication gestuelle à travers des ensembles de règles guidant l'association de certains mouvements à la parole, comme dans le cas du système GRETA [17] (un agent conversationnel animé ECA [18, 19]). Avec l'ère du *big data*, l'essor du *machine learning* puis du *deep learning* a favorisé l'émergence d'approches dites *data-driven*, où l'analyse computationnelle repose directement sur l'exploitation de vastes ensembles de données pour entraîner des modèles destinés à produire et à analyser des gestes.

Aujourd'hui, les modèles de génération de gestes (systèmes capables de créer automatiquement des séquences de gestes) constituent le cœur de l'étude de la gestualité en informatique. Ils visent principalement à produire des mouvements gestuels synchronisés avec un discours, à partir de la parole audio, du texte ou d'une combinaison de ces modalités, et trouvent des applications dans de nombreux domaines tels que les agents conversationnels et assistants virtuels, les avatars en réalité virtuelle ou augmentée, les jeux vidéo et autres environnements interactifs, ou encore la robotique humanoïde et sociale. Leur objectif est de générer des gestes qui s'insèrent de manière cohérente dans le déroulement temporel de la parole, tout en respectant les contraintes cinématiques du corps humain afin de produire des mouvements plausibles.

### 1.2.3 Limites des approches outillées

L'émergence des grands corpus constitue à la fois un atout et une limite pour l'étude de la gestualité. Si le traitement automatique de volumes massifs de données permet d'accélérer certaines analyses, par l'accès qu'ils offrent notamment à un large panel de cas, il ne peut se substituer à l'expertise humaine. Svartik [20] souligne ainsi les limites d'une étude de la parole fondée uniquement sur la taille des corpus : une distance excessive peut se créer entre l'utilisateur final et le matériel original. Le contenu du corpus risque en effet d'être traité comme un objet canonique et dépourvu de contexte, ce qui peut biaiser les analyses qui en sont tirées et limiter leur validité. Or, cette expertise humaine reste justement difficile à mobiliser lorsque la quantité de données à analyser est importante. On se trouve alors confronté à un compromis entre quantité et qualité de données : l'idée selon laquelle « plus grand est toujours meilleur » peut conduire à négliger l'adéquation réelle du corpus. Un volume conséquent de données faiblement structurées, mal annotées ou traitées de manière approximative, ne saurait en effet compenser l'absence d'une description qualitative rigoureuse.

Dans le domaine informatique, la tendance à privilégier la quantité au détriment de la qualité est fréquente, en particulier dans une logique de *machine learning* où de larges volumes de données, nécessaires au fonctionnement de ces modèles, sont perçus comme un gage de performance, et alors même que les grands corpus de gestes généralement utilisés ne sont majoritairement pas annotés. Un tel déséquilibre, bien que séduisant par la possibilité de produire des indicateurs quantitatifs, des statistiques ou des visualisations, pose des problèmes méthodologiques majeurs et limite la portée interprétative des analyses et des modèles qui en découlent. En effet, tout système automatique nécessite un cadre et des consignes précises, ce qui implique un minimum d'analyse préalable pour déterminer ce qui est pertinent à extraire ou à mesurer. Les approches auto-supervisées permettent certes d'exploiter des corpus peu annotés, mais elles dépendent toujours d'un apprentissage non supervisé sur de grands ensembles, et peinent encore à capturer la sémantique fine du lien geste-parole. Il semble alors nécessaire d'intégrer ces précisions sémantiques, notamment au moyen d'annotations expertes destinées à guider l'apprentissage des modèles. Ainsi, au-delà d'un certain seuil, l'augmentation massive de données peut générer des coûts disproportionnés sans bénéfices clairs, notamment si la qualité des annotations n'est pas garantie. Sans contrôle rigoureux de la pertinence, de la cohérence et de la précision de ces annotations, la valeur scientifique et opérationnelle des corpus s'en trouve donc prévisiblement fortement diminuée.

Dans le domaine de l'étude du geste, cette critique prend une importance particulière. En dehors de la perte de contexte inhérente à l'utilisation des grands corpus – qui peut faire oublier la manière dont le geste se construit et se comprend dans une interaction donnée –, il convient de souligner que les performances en termes de classifications (ex. : système d'annotation fonctionnelle automatique) ou de génération de gestes ne sont pas encore à la hauteur des connaissances établies par les *gesture studies* et ne répondent donc pas aux besoins actuels pour faire avancer notre compréhension de la gestualité co-verbale. Idéalement, en effet, les gestes générés automatiquement devraient être pertinents du point de vue communicatif, en

étant adaptés au contenu sémantique et pragmatique du discours qu'ils accompagnent. Mais en pratique, cet objectif reste largement hors de portée des modèles actuels, qui peinent à dépasser une génération gestuelle principalement rythmique [21], faiblement ancrée au contenu sémantique du discours et encore moins à ses dimensions pragmatiques. D'un point de vue technique, les fonctions de perte (*loss functions*) sont généralement orientées vers l'analyse géométrique et les régularités statistiques des données : elles mesurent la continuité, la régularité et la proximité par rapport aux données d'entraînement, souvent en réalisant une moyenne de toutes les dimensions du geste. Aussi les dimensions sémantiques ou autres caractéristiques spécifiques de certains gestes ne sont-elles pas prises en compte, l'apprentissage n'étant pas enrichi du sens des mouvements à traiter.

L'une des raisons de ce manque est le coût très élevé de l'annotation manuelle, qui requiert une expertise spécialisée. À ce jour, aucun outil automatique ou semi-automatique ne permet d'alléger de manière significative la tâche d'enrichissement des corpus pour les annotateurs. Les initiatives existantes visent surtout la segmentation et l'extraction de caractéristiques motrices, laissant de côté les aspects les plus complexes de la gestualité, qui ont trait à leur interprétation. Par ailleurs, les études portant sur la génération de gestes restent peu concluantes en termes de profondeur de sens [21]. Le recours à des questionnaires perceptifs pour évaluer les gestes produits par des modèles génératifs semble parfois relever davantage d'une logique performative que véritablement informative [22], dans la mesure où ces questionnaires paraissent parfois plus déterminants pour l'acceptation d'une publication que pour l'évaluation effective des gestes. De plus, ces gestes de synthèse ne sont presque jamais évalués en termes de fonctionnalité ou de forme, telle que définie et décrite par les gestualistes, ce qui limite fortement la portée et la pertinence de ces évaluations. Enfin, si l'analyse des représentations internes (*latent spaces*) fait l'objet de nombreux travaux en intelligence artificielle, elle reste encore peu exploitée pour l'étude et la validation d'hypothèses linguistiques sur la gestualité, en particulier dans les modèles dédiés à la génération ou à la reconnaissance.

# 2 Problématique et questions de recherche

Les projets véritablement pluridisciplinaires sur les gestes restent aujourd'hui l'exception. Le manque de temps, de moyens, mais aussi parfois de reconnaissance mutuelle entre disciplines, ne permettent pas de mobiliser ces expertises précieuses pour porter un regard qualitatif et contextualisé sur les gestes étudiés. Les enjeux liés à la modélisation de la gestualité, notamment dans les contextes de l'animation, de la synthèse de comportements humains ou de l'interaction humain-machine, soulignent pourtant l'importance cruciale d'un dialogue interdisciplinaire. Ainsi, la pluridisciplinarité ne constitue pas un simple cadre contextuel, mais bien une condition indispensable pour conduire une étude rigoureuse et complète des phénomènes gestuels, particulièrement dans des recherches visant à articuler observation empirique, annotation fine et modélisation automatique. Le principal obstacle réside donc dans le manque concret de collaboration réellement pluridisciplinaire entre les chercheurs et la perméabilité des approches. Les systèmes de génération ou de classification sont encore trop rarement conçus et évalués en tenant compte des fondements théoriques en gestualité et des finalités fonctionnelles du geste.

Cette lacune limite la portée des conclusions qui peuvent en être tirées de ces modèles, non seulement en termes de performances quantitatives, mais surtout quant à leur capacité à rendre compte des fonctions communicatives du geste, de leur adéquation aux cadres théoriques existants et de leur transférabilité à d'autres contextes discursifs ou interactionnels. Dans cette perspective, le développement de modèles véritablement intégrés, cohérents vis à vis des connaissances théoriques existantes et rigoureusement validées, constitue un enjeu central pour approfondir la compréhension des mécanismes gestuels tout en proposant des outils robustes pour l'analyse, la prédiction et la génération de gestes.

La problématique centrale de cette thèse touche ainsi à la conceptualisation d'une **méthodologie d'analyse qualitative et outillée du geste**, objet d'étude hautement complexe. Le postulat de cette recherche est que **la modélisation de la gestualité manuelle co-verbale ne peut être pleinement menée qu'en articulant les approches et perspectives issues à la fois des sciences du langage et des sciences computationnelles**, sans privilégier un domaine au détriment de l'autre. D'une part, en effet, les modèles informatiques peuvent tirer profit des acquis théoriques et méthodologiques développés en sciences du langage afin d'enrichir ses modèles, notamment en intégrant des descriptions fonctionnelles et des cadres théoriques établis dans les domaines de l'analyse du geste. D'autre part, les études menées en sciences du langage peuvent bénéficier des approches, des paradigmes et des outils proposés par les sciences computationnelles pour la constitution et le traitement de large ensemble de données, la conduite d'analyses quantitatives, la visualisation des comportements étudiés et la modélisation de phénomènes complexes difficilement appréhendables par des méthodes exclusivement manuelles et qualitatives.

C'est dans ce cadre que se pose la question de recherche principale de ces travaux : *En quoi une méthodologie d'analyse pluridisciplinaire peut-elle faire progresser l'analyse et la modélisation du geste manuel co-verbal ?* Pour répondre à cette question, cette thèse poursuivra deux objectifs principaux. Il s'agira d'une part de montrer l'intérêt d'une approche informatique fondée sur les données, plutôt que strictement sur les modèles, dans une perspective où les modèles computationnels seraient nourris d'informations tirées des cadres théoriques et descriptifs issus des sciences du langage, et non réduits à un simple traitement statistique de données de mouvements. Il s'agira d'autre part de démontrer le rôle central de l'annotation, en particulier de l'annotation multimodale du geste, comme point d'articulation entre l'observation qualitative et la formalisation automatique. Ainsi, dans une approche résolument centrée sur des données enrichies, l'annotation gestuelle apparaît à la fois comme un instrument d'analyse et comme un point d'entrée privilégié pour la modélisation computationnelle.

Organisée selon ces objectifs, cette thèse progressera autour de quatre questions de recherche :

*(Q1) Quels cadres théoriques structurent aujourd'hui l'étude de la gestualité dans le champ des gesture studies et dans les approches computationnelles ?*

*(Q2) Quels sont les éléments constitutifs d'un corpus multimodal suffisamment fiable et adapté à une analyse qualitative du geste co-verbal ?*

*(Q3) En quoi les pratiques d'annotation multimodale issues des sciences du langage peuvent-elles constituer un levier méthodologique structurant pour les approches computationnelles de la modélisation du geste co-verbal ?*

*(Q4) Dans quelle mesure les paradigmes d'analyse et de traitement de données issues de l'approche computationnelle peuvent-ils constituer un levier méthodologique pour l'analyse pluridisciplinaire et l'annotation multimodale fiable du geste co-verbal ?*

# 3 Contexte de recherche

Ce projet de recherche a été mené au sein de deux unités mixtes de recherche : le Laboratoire lorrain de Recherche en Informatique et ses Applications (LORIA, UMR 7503) et Praxiling (UMR 5253). Chacun de ces laboratoires possède une spécialisation propre. Praxiling est un laboratoire de sciences du langage, qui réunit une équipe transdisciplinaire composée de linguistes, de phonéticiens et de neuropsychologues. Son objectif est d'explorer les liens entre langage, cognition et pratiques sociales, en analysant à la fois les usages concrets du langage et les mécanismes cognitifs qui les sous-tendent. C'est afin d'acquérir une expertise approfondie en sciences du langage que j'ai réalisé ma première année de doctorat au sein de ce laboratoire, à Montpellier. Lors de ma deuxième année, j'ai intégré l'équipe Multispeech du laboratoire en informatique LORIA. Les recherches de cette équipe portent sur l'étude de la communication parlée et couvrent un large spectre, allant de la perception humaine des indices acoustiques à l'analyse et à la reconnaissance automatique de la parole. Ces travaux s'inscrivent dans une perspective appliquée, visant à développer des interfaces vocales et visuelles performantes pour des systèmes embarqués et multimodaux. C'est dans ce cadre que la dimension informatique de ma thèse a été particulièrement approfondie.

Cette recherche s'inscrit initialement dans le projet Syncogest, récemment soutenu par l'ANR dans le cadre de l'appel à projets 2024, et dédié à l'étude du geste manuel co-verbal. Ce projet fédère trois équipes aux expertises complémentaires en informatique (Multispeech), linguistique (Praxiling) et sciences du mouvement (EuroMov DHM). Les recherches menées dans le cadre de cette thèse ont bénéficié d'un financement *MITI Prime80* du CNRS, destiné à soutenir les projets à forte dimension interdisciplinaire.

La synergie entre les deux équipes de recherche en informatique et en sciences du langage a permis d'articuler des approches théoriques, méthodologiques et techniques complémentaires : définitions fonctionnelles et segmentations fines du geste issues des sciences du langage d'un côté, outils automatiques de gestion et paradigmes d'analyse de données à grande échelle proposés par l'informatique de l'autre. La présente thèse s'inscrit ainsi pleinement dans ce cadre d'interactions entre disciplines. Cette complémentarité, si elle constitue une richesse indéniable, s'est toutefois accompagnée de défis notables : différences de vocabulaire scientifique, contraintes propres à chaque domaine, attentes de résultats divergentes, ou encore méthodes de travail distinctes. Ces obstacles ont nécessité un effort constant de coordination, de dialogue

et d'adaptation dès les premiers instants du projet, reflétant la complexité réelle de l'intégration pluridisciplinaire. Un tel rapprochement ouvre néanmoins un espace particulièrement fertile pour l'élaboration de nouvelles méthodologies, la mutualisation des ressources et la formation de jeunes chercheurs capables d'évoluer à l'interface de plusieurs champs scientifiques. Cette collaboration durable entre disciplines favorise la conception d'outils innovants, plus performants et mieux adaptés, répondant aux besoins spécifiques aussi bien des sciences humaines que de l'informatique, tout en valorisant et respectant les expertises propres à chaque domaine. À plus long terme, une telle approche vise à renforcer la cohérence des recherches multimodales et à encourager l'émergence de projets ambitieux, tant à l'échelle nationale qu'internationale.

# 4 Organisation de la thèse

Ce manuscrit se structure en quatre parties. **La première partie** de cette thèse, intitulée « Fondements théoriques et enjeux des études de la gestualité », pose le cadre théorique sur lequel se fonde cette recherche. Il vise à identifier les fondements conceptuels et méthodologiques qui guident l'étude des gestes dans les domaines des sciences du langage et de l'informatique, et s'organise dès lors en deux chapitres :

**Chapitre 1 -- L'étude du geste dans le champ des *Gestures* studi*es.*** Ce premier chapitre donne un aperçu des définitions théoriques et des cadres méthodologiques issus des sciences du langage pour l'étude multimodale du geste co-verbal.

**Chapitre 2 -- L'étude de la gestualité en sciences computationnelles.** Ce deuxième chapitre présente les principales méthodes et paradigmes mobilisés pour la modélisation du geste co-verbal dans une perspective computationnelle.

**La deuxième partie**, intitulée « Ressources et données multimodales », s'intéresse aux méthodes de constitution des corpus, d'enrichissement des données et d'évaluation des annotations multimodales. Cette partie vise à identifier et formaliser les critères qui garantissent qu'un corpus multimodal est réellement exploitable et suffisamment fiable pour l'analyse du geste co-verbal. Elle se structure en trois chapitres.

**Chapitre 3 -- Corpus multimodaux pour l'étude de la gestualité : enjeux et nécessité**. Lorsqu'il n'est pas possible de constituer son propre corpus, il est nécessaire d'examiner les ressources existantes et d'évaluer leur adéquation avec le projet de recherche. Or, dans le cas spécifique de l'étude de la gestualité co-verbale, une première question s'impose : *Quels corpus sont actuellement disponibles ?* Ces corpus, même rattachés à un même domaine de spécialité, peuvent différer considérablement par leur taille, leurs modalités ou leurs contextes de production. De ce fait, ils peuvent être plus ou moins adaptés à certains usages et présenter des limites lorsqu'il s'agit d'atteindre des objectifs de modélisation, notamment computationnelle.

**Chapitre 4 -- Analyse critique de l'enrichissement d'un corpus multimodal**. Lorsque le choix d'un corpus repose sur sa disponibilité et les données qui le compose,

une seconde question émerge : *Quelles méthodes permettent de s'assurer de la qualité et de la réexploitabilité d'un corpus existant ?* En effet, il est indispensable de vérifier les critères garantissant la fiabilité et la cohérence des données, notamment en termes d'enrichissement, avant de fonder l'ensemble d'un projet sur des ressources. Ces considérations invitent également à réfléchir aux pratiques d'élaboration de corpus et à la transparence méthodologique, éléments essentiels pour maximiser la réutilisabilité des corpus au sein de la communauté scientifique.

**Chapitre 5 -- Protocole d'enrichissement de corpus de gestes**. Ce cinquième chapitre présente le protocole d'enrichissement établi dans le cadre de cette thèse pour obtenir un ensemble de données fiables et qualitatives destiné à tester l'approche interdisciplinaire défendue dans cette thèse. Ce protocole, pensé pour produire des données structurées, cohérentes et interprétables, est ainsi à la fois ancré dans les usages et pratiques d'annotation descriptives et fonctionnelles issus des sciences du langage et pensé pour une exploitation efficace par des modèles informatiques

**La troisième partie**, intitulée « Apports et perspectives de l'expertise des sciences du langage », met en évidence la manière dont l'expertise des sciences du langage constitue à la fois un cadre méthodologique et un outil opérationnel pour modéliser les gestes co-verbaux, l'annotation multimodale experte servant de première approche computationnelle. Cette partie s'organise en deux chapitres.

**Chapitre 6 -- L'annotation multimodale comme outil de modélisation du geste co-verbal**. Ce chapitre supporte l'idée que, bien que coûteuse et chronophage, l'annotation multimodale fournit déjà un premier matériau pour une analyse fine et précise du geste. Il s'agira alors ici d'identifier les types de caractérisations effectivement capturés par les annotations, et de montrer en quoi ceux-ci peuvent constituer un levier central pour le développement de nouveaux outils computationnels dédiés à la modélisation du geste.

**Chapitre 7 -- L'annotation multimodale comme outil d'évaluation des gestes synthétiques.** L'objectif de ce chapitre est de démontrer comment les annotations expertes peuvent permettre de dépasser les limites des évaluations perceptives, en fournissant des critères pertinents pour mesurer plus objectivement et plus efficacement la qualité et la diversité de gestes générés automatiquement, répondant ainsi à un besoin explicitement formulé dans le domaine, tout en permettant de caractériser précisément les limites actuelles des gestes de synthèse et, par contraste, de mettre en évidence les spécificités des gestes naturels.

**La quatrième et dernière partie**, intitulée « Apports et perspectives de l'expertise des sciences computationnelles », met en évidence les contributions de cette thèse reposant sur l'analyse automatique de mouvement et l'apprentissage automatique supervisé, tant pour l'enrichissement semi-automatique de corpus que pour le développement de nouvelles méthodologies pour l'analyse automatique des gestes. Cette partie s'organise en deux chapitres.

**Chapitre 8 -- Annotation automatique des aspects formels du geste manuel co-verbal.** Sur la base d'un repérage préalable des gestes, les outils issus des technologies informatiques offrent des moyens objectifs et précis pour annoter les différents aspects formels d'un geste. Ils permettent ainsi de s'affranchir en partie de la subjectivité inhérente aux annotations manuelles, d'alléger le travail d'annotation, et ainsi de rendre possible le traitement de volumes de données plus importants. Ce chapitre défendra dès lors une méthodologie permettant une annotation des caractéristiques formelles qui soit cohérente et qualitative et présentera les possibilités de leur automatisation.

**Chapitre 9 -- Segmentation semi-automatique du geste sous contrainte de données d'entrainement limitées.** Sur la base d'un entraînement sur un jeu de données restreint dans lequel les gestes ont été préalablement segmentés – manuellement ou automatiquement – il devient possible de guider un modèle vers le repérage automatisé des gestes. Ce chapitre montrera alors comment des annotations multimodales peuvent contribuer à mettre en place un modèle de segmentation de gestes, et ce même dans un contexte de données limitées. Un modèle capable de discriminer les gestes des mouvements non communicatifs signifierait l'existence d'une représentation latente structurée, suffisamment informative pour encoder les propriétés distinctives du geste. Une telle représentation permettrait alors de focaliser les apprentissages sur le matériau réellement pertinent lors d'analyses automatiques multimodales.

Enfin, la thèse se conclut sur les apports de l'approche interdisciplinaire défendue, par la synthèse de l'ensemble des travaux présentés et la mise en lumière des différents questionnements constituant le fil rouge de cette recherche.

# 5 Valorisations de recherche

## 5.1 Contributions scientifiques

Cette thèse apporte plusieurs contributions significatives au domaine de l'étude de la gestualité, tant sur le plan méthodologique qu'analytique. Elle propose notamment de premiers outils permettant de vérifier la qualité minimale de l'enrichissement d'un corpus multimodal, en utilisant des statistiques descriptives destinées à identifier les incohérences, lacunes ou biais éventuels dans les données secondaires. Elle introduit également une première version d'un protocole d'enrichissement visant à produire une description précise et fiable des gestes, intégrant des catégories pertinentes pour une approche computationnelle et permettant une caractérisation systématique du geste dans ses différentes dimensions, formelles ou fonctionnelles. Nous introduisons le corpus *reBeat-14* (13 minutes), dont les fichiers sont issus du corpus BEAT et ont été enrichis selon le protocole établi dans le cadre de cette thèse. L'annotation multimodale experte est également mise en œuvre comme méthodologie d'évaluation des gestes générés, offrant un cadre de référence fiable pour comparer gestes naturels et synthétiques, et ainsi mieux cerner les limites des résultats des modèles de génération automatique.

Parallèlement, la thèse développe une méthodologie d'annotation automatique des aspects formels du geste, fondée sur une analyse géométrique et cinématique permettant de standardiser et d'objectiver l'annotation des aspects formels du geste. Cette méthodologie est complétée par une interface de visualisation (*Gesture User Interface,* GUI) en développement, destinée à faciliter l'exploration, la vérification et l'exploitation des annotations. Enfin, cette recherche introduit une première version exploratoire d'un modèle d'apprentissage supervisé (*Co-speech gesture segmentation on motion sequences*, COSMOS) pour la segmentation du geste dans des contextes d'annotations limitées. L'objectif de cette démarche n'est pas de proposer un modèle d'état de l'art entraîné sur des centaines d'heures de données, capable de résoudre la tâche de segmentation de manière absolue. Notre contribution réside plutôt dans la preuve de concept d'une approche *human-in-the-loop* en régime de données extrêmement limitées. Il s'agit de montrer qu'il est possible d'entraîner un « modèle compagnon » sur une très petite quantité de données expertes (environ 13 minutes) afin d'accélérer et de pré-annoter la segmentation pour un locuteur spécifique, tout en laissant le contrôle final à l'annotateur humain.

Ces travaux visent ainsi à attester que l'étude de la gestualité est intrinsèquement pluridisciplinaire, nécessitant la collaboration entre sciences du langage et sciences computationnelles pour produire des corpus fiables, des modèles robustes et des analyses fiables, pertinentes et approfondies. L'ensemble de ces contributions ouvre ainsi de nouvelles perspectives pour la recherche, la modélisation et l'évaluation des gestes co-verbaux dans des contextes multimodaux.

## 5.2 Publications et médiations scientifiques

### 5.2.1 Articles de conférence associés à une communication en congrès international

**GRONDIN-VERDON M.** [1], CAILLAT D.[1], ABEL L., OUNI S. **2025**. Evaluating Automatic Hand-Gesture Generation Using Multimodal Corpus Annotations: The Benefits of a Multidisciplinary Approach. In *In Proceedings of the International Workshop on Generation and Evaluation of Non-verbal Behaviour for Embodied Agents* (GENEA '25), October 27–28, 2025, Dublin, Ireland. ACM, New York, NY, USA, 9 pages. 10.1145/3746268.3759430 ⟨hal-05330689⟩

**GRONDIN-VERDON M.**, CAILLAT D., OUNI S. **2024**. Qualitative study of gesture annotation corpus: Challenges and perspectives. In *International Conference on Multimodal Interaction* (ICMI Companion '24), November 04–08, 2024, San Jose, Costa Rica. ACM, New York, NY, USA, 9 pages. 10.1145/3686215.3688820 ⟨hal-04767088⟩

**GRONDIN-VERDON M.**, CAILLAT D., OUNI S. **2026**. A Conceptual *Framework* for Co-Speech Manual Gesture Formal Description and Automatic Annotation. LREC 2026 (in review)

[1] Les auteur·rice·s ont contribué·e·s de manière équivalente à ce travail.

### 5.2.2 Communication sur résumé

**GRONDIN-VERDON M.**, CAILLAT D., OUNI S. **2025**. Encoding the Spatial Features of Co-Verbal Manual Gestures: a *Framework* for Automated Annotation. *12th International Conference on Multimodality* (ICOM), University of Groningen, Oct 2025, Groningen, Netherlands. ⟨hal-05343752⟩

CAILLAT D., **GRONDIN-VERDON M.**, OUNI S. **2026**. Exploring Form–Meaning Mapping in Co-Verbal Gestures: A Methodological Approach to the Semantic Value of Formal Features. University of Trier, Trier, Germany DGFS (accepted)

CAILLAT D., **GRONDIN-VERDON M.**, OUNI S. **2025**. Iconic Co-Verbal Gestures: Study of Some Facets of Iconicity. *At Dimensions of Iconicity in the Visual Modality*, February 13–14, 2025, University of Göttingen. ⟨hal-04962406⟩

CAILLAT D., **GRONDIN-VERDON M.**, OUNI S. **2025**. Exploring Gesture Formalization: Encoding Features and Automation Strategies. At International Society of *Gesture studies* (ISGS) 10th, *Motion to Meaning: Innovations in Multimodal Language Research*, July 9–11, 2025, Radboud University, Nijmegen. ⟨hal-05180973⟩

CAILLAT D., MARIN L., DODANE. C, HIRSCH F., OUNI S., SLANGEN P., GUYOT P., COLOTTE V., MORGENSTEN A., ABEL L., **GRONDIN VERDON M.**, LOZANO GOUPIL J. **2022**. Synchronization of speech and gestures in an interactional context (SyncoGest Project). ISGS 2022 - *9th Conference of the International Society for Gesture studies*, Jul 2022, Chicago, United States. ⟨hal-03875218⟩

### 5.2.3 Poster

**GRONDIN-VERDON M.**, CAILLAT D., OUNI S., **2025**. COSMOS: Semi-automatic tool development for co-speech gesture segmentation based on expert *frame* annotations. GENEA 2025 - *Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Workshop*, Oct 2025, Dublin, Ireland ⟨hal-05280497⟩

### 5.2.4 Médiation scientifique grand public

**GRONDIN-VERDON M.**, CAILLAT D., HIRSCH F., OUNI S. **2022**. La gestualité, Roll-on du stand de Praxiling, Fête de la science de Montpellier. Université Paul-Valéry Montpellier.

# Partie I
# Fondements théoriques et enjeux des études de la gestualité

Cette partie présente un état de l'art de l'étude de la gestualité. Elle examine d'une part les définitions issues des sciences du langage, au fil de la présentation des théories sur la coordination parole-gestes, des méthodes de descriptions, et des typologies fonctionnelles et formelles. D'autre part, elle explore les approches computationnelles de la gestualité, en détaillant les principes et méthodes de segmentation, de classification et de génération des gestes, ainsi que les enjeux liés à l'interprétabilité et à l'évaluation des modèles.

L'examen de ces contributions issues des sciences du langage et des sciences computationnelles, de leurs apports, leurs limitations et obstacles méthodologiques actuels respectifs, vise à poser les fondations de l'approche pluridisciplinaire défendue dans cette recherche de thèse. L'objectif de cette synthèse préliminaire est ainsi de poser les bases de la complémentarité des approches envisagées, en introduisant les éléments jugés les plus pertinents et utiles pour la modélisation et l'analyse de la gestualité co-verbale.

# Chapitre 1
# L'étude du geste dans le champ des *Gestures studies*

## 1 Définition du geste

Le corps humain est capable de produire une grande diversité de signaux sociaux mobilisant différentes parties du corps, telles que le visage (incluant les mouvements du cou et de la tête, le regard, les sourcils, la bouche et les expressions faciales), les membres supérieurs (comprenant le torse, les épaules, les bras, les avant-bras, le coude, les mains et les doigts) et les membres inférieurs (comprenant les jambes, les pieds, et les genoux). On parle alors de *mouvement* lorsqu'il y a une altération ou un déplacement physique du corps ou d'une de ces parties. Les informations véhiculées par ces mouvements, de nature variée, peuvent être perçues, interprétées et comprises par un interlocuteur, bien qu'ils ne soient pas nécessairement produits de manière intentionnelle par le locuteur. Un tremblement, par exemple, peut informer de l'état émotionnel d'un locuteur sans qu'il ne soit délibérément produits. De tels mouvements se distinguent donc de ceux intentionnellement réalisés par un locuteur.

Au sein des mouvements intentionnels, tous ne participent pas directement à la communication. Les mouvements intentionnels non communicatifs recouvrent ainsi les actions réalisées au cours d'une interaction, comme attraper un verre pour boire ou manipuler un objet (mouvements qui visent à atteindre un objectif spécifique), ou les gestes dits *de confort*, qui consistent par exemple à se gratter la tête ou à croiser les jambes. Les gestes réellement tenus pour être co-verbaux, donc associés à la parole, se distinguent en ce qu'ils sont produits intentionnellement et qu'ils ont pour fonction de contribuer à la transmission d'un message. Kendon [7, 8], entre autres, définit ainsi les gestes co-verbaux comme des actions visibles délibérément porteuses de sens. Ils constituent ainsi une composante à partie entière du processus de parole et entretiennent à ce titre un lien étroit avec le discours, par leur synchronisation avec la production verbale et prosodique, et par les rôles multiples qu'ils jouent dans la transmission du message. Les gestes co-verbaux sont donc fondamentalement ancrés au

discours, avec lequel ils fonctionnent en synergie sur le plan de la construction du sens. D'un point de vue formel, ces gestes sont le plus souvent définis comme des mouvements réalisés avec les bras et les mains [9, 10], cette définition peut être élargie à tout signal multimodal délibérément communicatif (ex. : mimiques, postures, etc.).

Dans le cadre de cette thèse, nous nous concentrerons exclusivement sur l'étude des gestes co-verbaux manuels. À ce titre, le terme « geste » sera spécifiquement réservé dans cette recherche à la désignation (1) de mouvements intentionnels et communicatifs (2) produits avec les mains dans un contexte d'élaboration verbale. Le terme « mouvement », quant à lui, sera affecté aux références à un déplacement physique en tant que tel, indépendamment de toute visée communicative.

# 2 La production gestuelle en contexte

## 2.1 Coordination entre geste et parole

Les gestes occupent une place centrale dans le processus de la parole, où ils impactent aussi bien la production et la compréhension [23, 24] que la dynamique globale d'une interaction. Les gestes soutiennent de fait l'élaboration verbale en apportant un appui kinesthésique [25] et contribuent à la conceptualisation [26] ainsi qu'à la formulation d'idées complexes [27], permettant par exemple d'exprimer plus clairement des notions abstraites [28]. Leur rôle est particulièrement visible dans la verbalisation d'expériences subjectives difficiles à décrire, comme l'expression de la douleur [29], et dans la mise en récit d'informations complexes, qu'ils contribuent à structurer et à enrichir [30]. Du côté de la réception, les gestes jouent également un rôle dans la perception du message. Ils contribuent par exemple à lever les ambiguïtés discursives [31], notamment en fournissant des indices visuels complémentaires [32] ou en soulignant l'organisation d'énoncés syntaxiquement ambigus [33]. La congruence entre gestes et parole accélère dès lors le traitement de l'information et révèle une influence réciproque entre ces deux modalités [23]. Cette congruence peut même aller jusqu'à modifier la perception des sons [34]. Les gestes contribuent ainsi à la compréhension globale d'un message, de même qu'ils participent à la régulation d'une interaction. Ils facilitent par exemple la gestion des tours de parole, pouvant servir d'indicateurs de fin de tour [35, 36], tandis que l'orientation du corps et du regard jouent un rôle essentiel dans la gestion de l'interaction [37, 38].

## 2.2 Ancrage culturel et situationnel

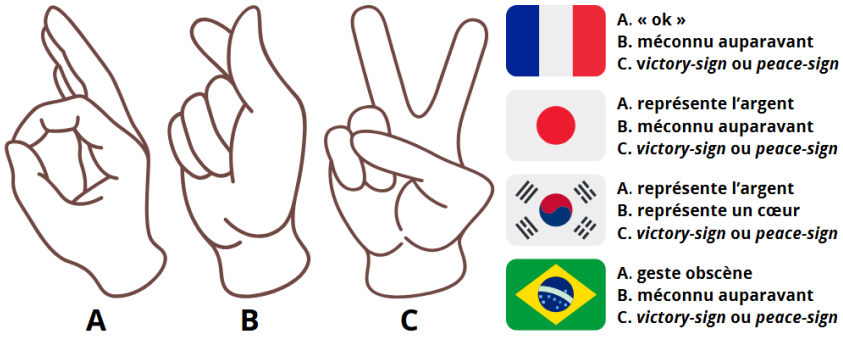



*Figure 1. Illustration des gestes pouvant être propres à chaque communauté*

Si les gestes ne peuvent que très difficilement être interprétés indépendamment du contexte discursif dans lequel ils s'inscrivent, tant ils sont étroitement liés au discours auquel ils participent, certains d'entre eux relèvent de variations interculturelles relativement codées et dépendent ainsi d'une culture donnée (Figure 1). Aussi, en l'absence du cadre culturel et interactionnel dans lequel ces gestes prennent sens, leur interprétation peut devenir ambiguë, voire impossible. Ainsi, un geste considéré comme poli, emphatique ou neutre dans une culture peut prendre une tout autre valeur dans une autre, illustrant le rôle structurant des conventions socioculturelles dans l'interprétation gestuelle. De nombreuses études ont été publiées concernant ces gestes dits « conventionnels » ou « emblématiques », dont le sens est univoque et répond à une « convention » propre à un (ou plusieurs) pays [39, 40, 41, 42, 43, 44, 45, 46, 47].

Au-delà des aspects culturels, la situation interactionnelle elle-même influence fortement la forme, la fréquence et la fonction des gestes [48]. Le cadre de communication (formel ou informel), le degré d'engagement des interlocuteurs, la nature de la tâche, le type de discours ou encore la relation interpersonnelle modulent en effet la manière dont les gestes sont produits et interprétés. Un même locuteur adoptera ainsi des stratégies gestuelles différentes selon qu'il raconte une histoire, argumente un point de vue, donne des instructions ou participe à une conversation spontanée. De la même manière, le contexte matériel et spatial joue également un rôle essentiel : la disponibilité de l'espace gestuel, la présence d'objets physiques, la distance entre les interlocuteurs ou encore la configuration de la pièce déterminent la possibilité de produire certains gestes et influencent leur amplitude et leur orientation. Les gestes ne sont donc jamais aléatoires, mais émergent en coordination avec le discours, l'environnement, la tâche et les participants.

L'ancrage des gestes constitue ainsi un cadre interprétatif indispensable pour comprendre leur variabilité ou leur stabilité. Il rappelle que tout geste est produit à la fois comme un acte individuel et comme un acte social, façonné à la fois par les normes collectives, les circonstances locales et les dynamiques interactionnelles. Ce double ancrage culturel et situationnel est à la base de toute analyse rigoureuse de la gestualité et représente un point clé pour la modélisation et l'interprétation des données multimodales.

# 3 Eléments de description de la gestualité

L'étude des gestes co-verbaux s'appuie sur un ensemble de différentes caractéristiques telles que leur composition interne (phases de réalisation), leurs propriétés physiques (forme, emplacement, etc.), et leur contribution à l'activité communicative (lien avec le discours). Ces descripteurs, largement partagés au sein de la communauté des gestualistes, constituent un cadre d'analyse commun, bien que non réellement standardisé dans le détail, pour rendre compte de la diversité des productions gestuelles. Ils permettent ainsi de décrire la façon dont le geste se déploie dans le temps, les caractéristiques corporelles, spatiales et dynamiques qui structurent son mouvement, ainsi que les dimensions fonctionnelles qui éclairent son rôle dans la production et l'interprétation du discours.

## 3.1 Segmentation du geste

### 3.1.1 Les unités gestuelles selon Kendon

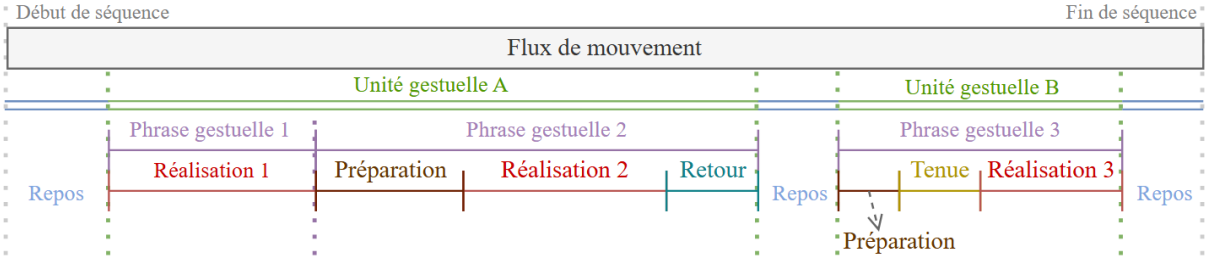



*Figure 2. Représentation de la segmentation du geste dans le temps, basé sur les travaux de Kendon [7, 8]*

Le repérage des gestes au sein du flux des mouvements d'un locuteur (mouvements intentionnels ou non, et communicatif ou non) consiste à discrétiser un signal quasi continu en plaçant les frontières temporelles délimitant les occurrences de geste. La finesse de cette segmentation dépend alors de la granularité souhaitée. Introduite par Kendon [7, 8], la notion d' « unités gestuelles » (*gesture units*) renvoie à la plus grande segmentation observable, comme illustré dans la Figure 2. Ces unités sont en effet définies prototypiquement par la transition d'une position de repos (non-mobilisation des membres impliqués dans la réalisation des gestes) à une autre position de repos (qui peut différer de la position de repos initiale). Pendant cet intervalle de temps, un locuteur peut réaliser un ou plusieurs gestes consécutifs. Une unité gestuelle peut donc contenir une ou plusieurs « phrases gestuelles » (*gesture phrases*), qui elles, renvoient chacune à l'élaboration d'un geste individuel seulement. Une phrase gestuelle, soit l'élaboration d'un geste individuel, se décompose elle-même en plusieurs phases.

### 3.1.2 Définition des phases selon Kendon

Dans sa réalisation la plus complète, selon Kendon [7, 8], une phrase gestuelle comprendra toutes les phases suivantes : préparation, repos, réalisation (*stroke*), tenue, rétraction. Pour certains auteurs, le geste réside dans la seule phase de *stroke*, là ou d'autres le définiront comme le résultat de l'ensemble des phases décrites par Kendon. L'absence de définitions et de pratiques réellement standardisées sur ce point notamment (mais sur d'autres également)

conduit ainsi à une diversité d'usages selon les chercheurs et les projets, qui complique le rapprochement des études et la comparaison des résultats. Les phases gestuelles sont toutefois traditionnellement entendues comme telles :

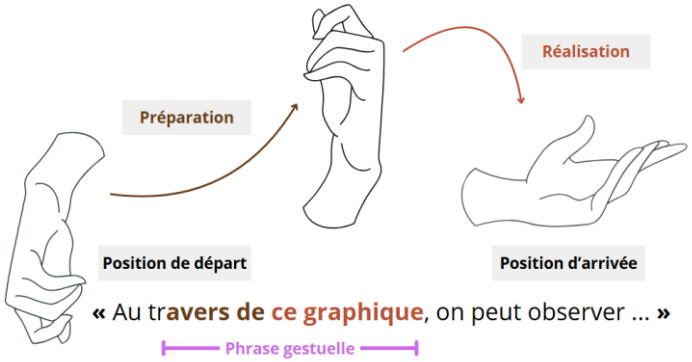



*Figure 3. Exemple de phrase gestuelle incluant une préparation et une réalisation*

**Préparation** : La phase de préparation (Figure 3) débute par le mouvement des articulateurs engagés dans la réalisation du geste, depuis la position de repos vers une position dans l'espace où le *stroke* du geste peut commencer. Cette phase est optionnelle : si les articulateurs sont déjà dans la position nécessaire à la réalisation du geste (par exemple parce que ce geste succède immédiatement à un autre), il ne nécessitera pas de phase de préparation.

**Réalisation** : La phase de réalisation (Figure 3) constitue le cœur du geste. Chaque geste ne comporte qu'une seule phase de réalisation. C'est au cours de cette phase que le geste prend son sens [7, 8, 9], et que les dynamiques de mouvement liées à l'« effort » et à la « forme » se manifestent avec le plus de clarté. Du point de vue du signal physique, le mouvement produit pendant cette phase est ainsi généralement caractérisé par un pic de vitesse et d'énergie. Dans le cadre de cette thèse, comme nous le détaillerons plus loin, nous nous concentrerons uniquement sur les phases de *stroke*.

**Tenue** : La phase de tenue (*hold*) correspond à un arrêt temporaire des articulateurs du geste et au maintien de la dernière position de la phase précédente. Cette phase facultative peut être observée avant la phase de réalisation du geste (*pre-stroke hold* [49]) ou après (*post-stroke hold* [49]). Pendant la phase de tenue, les articulateurs restent dans une position spécifique destinée à installer ou à prolonger le geste [49] ; la vitesse du mouvement est alors proche de zéro, et la tension musculaire reste soutenue.

**Rétractation** : La phase de rétraction renvoie au processus de retour à la position de repos des articulateurs impliqués dans le geste. La rétraction peut être partielle (due par exemple à l'enchaînement sur un autre geste), complète (retour à la position de repos), ou totalement absente (transition directe vers un autres gestes sans retour, ni partiel ni complet, à la position de repos intermédiaire).

## 3.2 Approches spatiales et articulaires

### 3.2.1 La description de l'espace gestuel

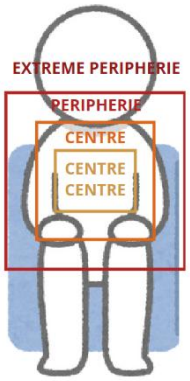



*Figure 4. Illustration de l'espace gestuel décrit par McNeill [9, 10]*

L'espace gestuel selon McNeill [9, 10] est un concept central dans l'étude des gestes co-verbaux. Il désigne la zone autour du corps où se déploient les gestes manuels accompagnant la parole, comme illustré dans la Figure 4. Cet « espace de l'expression gestuelle » n'est pas seulement un espace physique, mais aussi un espace signifiant où les gestes prennent place dans leur association au discours. McNeill souligne que cet espace est organisé de manière dynamique selon la position du locuteur, la nature de la communication, et les intentions expressives. Il comprend généralement les zones proches du tronc, principalement devant et sur les côtés du corps. L'espace gestuel est également lié aux contraintes perceptives et ergonomiques qui influencent la forme et la fonction des gestes produits.

On notera que dans la description de l'espace gestuel proposée par McNeill, la dimension de profondeur n'est pas prise en compte : le découpage spatial repose sur un référentiel bidimensionnel latéralité / hauteur (plan frontal), adopté du point de vue d'un locuteur assis face à son interlocuteur. Dans le cadre d'une modélisation computationnelle, l'espace dans lequel se produit le geste est représenté en 3D, typiquement à partir des positions x (latéralité), y (hauteur) et z (profondeur) issues de la capture de mouvement. Cette approche oblige ainsi à repenser les zones du geste en tenant compte de l'ensemble du volume disponible, et donc à intégrer l'espace compris devant le locuteur (axe Z). La gestion simultanée de ces trois dimensions reste cependant très largement peu explorée dans les approches classiques des *gesture studies*. Certaines études proposent toutefois de combiner aux deux dimensions de l'espace gestuel de McNeill la dimension de profondeur, alors définie comme la distance du geste par rapport au corps [50].

### 3.2.2 La description articulaire

L'approche kinésiologique de Boutet propose une description formelle et articulée du geste [51], centrée sur les mouvements des articulateurs (mains, bras, avant-bras, doigts) et leurs degrés de liberté (ddl). Chaque articulateur est défini dans son contexte anatomique, avec des limites et contraintes liées aux articulations voisines. Les gestes sont décrits dans un espace tridimensionnel, utilisant soit un référentiel axial (directions générales et plans) soit polaire (angles articulaires centrés sur chaque articulation). Les mouvements sont caractérisés selon les ddl : flexion/extension, abduction/adduction, rotations interne/externe, pronation/supination, circumduction. Les amplitudes maximales et les combinaisons possibles sont spécifiées pour chaque articulation (doigts, pouce, main, avant-bras, bras). Les articulations étant interconnectées, certains mouvements limitent ou influencent d'autres mouvements, rendant certains ddl interdépendants. Cette taxonomie permet de décrire précisément la configuration et la dynamique des gestes, adaptée aux analyses informatiques et à la modélisation formelle,

Toutefois, ce type de description, centré sur l'exécution articulatoire, s'avère plus difficile à mobiliser dans le cadre d'annotations multimodales fondées sur la perception du geste.

## 3.3 Descriptions formelles

Différentes typologies formelles ont été proposées pour décrire les gestes, et leur définition dépend étroitement de l'articulateur considéré, chaque partie du corps présentant des caractéristiques et contraintes propres. Pour la description de la main, Kendon [8], regroupe les gestes en quatre familles basées soit sur la forme de la main, soit sur l'orientation de la paume. La *G-family* (*grip*) regroupe les gestes où les bouts des doigts sont rapprochés. La *R-family* (*ring*) concerne les gestes dans lesquels au moins deux doigts (le pouce et l'index) forment un cercle. Les familles *Open Hand Prone* et *Open Hand Supine* se basent sur l'orientation de la paume : dans la première, la paume est tournée vers le bas, dans la seconde, elle est tournée vers le haut, correspondant à la famille *Palm Up Open Hand* décrite par Muller [52]. Il est à noter que cette typologie n'est pas strictement formelle : Kendon associe en effet aux différentes familles de gestes des fonctions et des intentions spécifiques. Par exemple, il décrit les familles G et R comme étant respectivement liées à la négation ou à l'interruption, et au fait de donner ou d'offrir.

Indépendamment de ces associations fonctionnelles, de nombreux travaux se concentrent sur des dimensions formelles plus détaillées et plus systématiquement observables. Parmi les dimensions les plus couramment étudiées figure la manualité, soit le fait qu'une seule main ou les deux mains soient mobilisées. Lorsque les deux mains sont engagées, il est alors possible de distinguer des configurations symétriques ou asymétriques. D'autres paramètres directement liés à la forme et à l'orientation de la main sont également souvent pris en compte. L'orientation de la paume [53] constitue ainsi un élément fréquemment décrit, bien que ses modalités de notation puissent varier selon les approches. La configuration de la main elle-même peut s'appuyer sur des inventaires issus des langues des signes, tels que HamNoSys [54] ou l'inventaire de langue des signes américaine de Friedman [55] utilisé par McNeill, ou sur des catalogues descriptifs plus généraux [56, 57, 58].

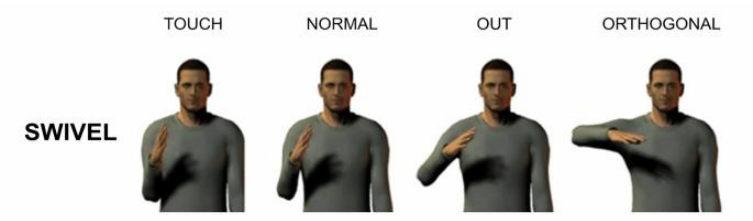



*Figure 5. Catégories d'encodage de l'arm swivel [59]*

Au-delà des propriétés morphologiques de la main, l'analyse gestuelle formelle porte également certaines caractéristiques dynamiques du geste. Celles-ci incluent notamment la direction [53, 60] ou la forme de la trajectoire [58, 61, 59], la taille du geste [62], ainsi que la vitesse de déplacement [53, 62]. Des mouvements plus globaux peuvent également être décrits, tels que le pivotement du bras (*arm swivel*, Figure 5) [59, 62] ou la distance entre les mains

[59]. Ces différents paramètres permettent de caractériser le geste comme une unité dynamique inscrite dans le temps et l'espace.

L'ensemble de ces dimensions constitue un cadre potentiellement riche pour la description des gestes, sans pour autant être systématiquement exploité dans son intégralité selon les études. De même, lorsqu'un même paramètre est pris en compte, ses modalités de catégorisation, son niveau de granularité et ses modes de combinaison varient fortement d'un système de notation à l'autre et selon les objectifs de recherche.

## 3.4 Approches interprétatives

Traditionnellement, le geste est également fréquemment interprété dans sa dimension fonctionnelle, par l'attribution d'une ou de plusieurs étiquettes qui rendent compte de son ou ses rôles et de sa ou ses relations avec la parole. Plusieurs typologies fonctionnelles existent alors, répertoriant différentes fonctions que peuvent remplir les gestes. Bien que cette section ne présente pas de manière exhaustive toutes ces typologies, qui se recoupent en grande partie, on peut mentionner entre autres celles de Wundt [1], Efron [40], Ekman et Friesen [41, 63], McNeill [9, 10], Kendon [7, 8], ou encore Cosnier [64], qui ont fortement façonné l'approche interprétative tout au long du XX$^{e}$ siècle. De nos jours, le choix de la typologie dépend souvent du projet de recherche et peut potentiellement inclure des classifications développées en interne.

Dès 1969, Ekman et Friesen [41, 63] ont proposé une perspective fonctionnelle en distinguant cinq grandes catégories de gestes : les *emblèmes*, codifiés culturellement et interprétables indépendamment de la parole ; les *illustrateurs*, qui accompagnent le discours pour le compléter ou le préciser ; les *régulateurs*, chargés de structurer l'échange conversationnel ; les gestes *expressifs*, traduisant des états émotionnels ; et les *adaptateurs*, souvent inconscients, orientés vers soi ou liés à la manipulation d'objets. Cette première classification a posé les bases d'une description systématique des gestes co-verbaux et a largement influencé les recherches ultérieures.

McNeill [9, 10] , dans les années 1990, complète cette approche en insistant sur le lien entre geste et discours, en distinguant parmi les gestes *illustrateurs* de Ekman et Friesen, les gestes *déictiques*, utilisés pour désigner des objets ou des points dans l'espace avec une directionnalité signifiante ; les gestes *iconiques*, qui représentent visuellement certaines caractéristiques des référents évoqués dans le discours ; les gestes *métaphoriques*, traduisant des concepts abstraits ; et les gestes de *battement*, qui accompagnent rythmiquement la parole. Cette classification, qui offre donc une première distinction des manières dont le geste s'articule au contenu propositionnel, est de nos jours la plus fréquemment utilisée.

Kendon [7, 8] complètera cette classification par la catégorie des gestes *pragmatiques*, définis comme illustrant des aspects du discours non directement liés au contenu référentiel ou propositionnel de l'énoncé. Il identifie ainsi trois catégories fonctionnelles au sein de ces gestes pragmatiques : les *modaux*, qui contribuent à indiquer comment un énoncé doit être interprété (ex. : appréciation, insistance, etc.) ; les *performatifs*, qui signalent la nature de l'acte de parole

dans lequel s'engage le locuteur ; et les *parsing*, qui contribuent au marquage de différents aspects de la structure du discours.

Cosnier [64] subdivisera quant à lui les gestes qu'il appelle « conversationnels » en gestes quasi-linguistiques (les emblèmes de Ekman et Friesen), porteurs d'un contenu sémantique explicite et parfois substitutifs à la parole, en gestes dits *synchronisateurs*, dont le rôle est de faciliter la coordination des tours de parole et la gestion temporelle de la conversation, et en gestes co-verbaux, produits simultanément à la parole. Les gestes co-verbaux, qui regroupent notamment les catégories de McNeill, sont à leur tour subdivisés en trois types :

- les gestes *référentiels*, qui comprennent les gestes *déictiques* mais aussi les gestes dits *illustratifs*, qui eux regroupent les gestes *iconiques* (spatiographiques, pictographiques et kinémimiques, selon qu'ils dépeignent un espace, une image ou un mouvement) et les *idéographiques* (les métaphoriques de McNeill),
- les gestes *expressifs*, véhiculant émotions et connotations discursives,
- et les gestes *paraverbaux*, qui organisent le rythme et la structure de l'énonciation.

Si les catégories de McNeill constituent la référence la plus répandue à l'heure actuelle, l'évolution de ces classifications et la nécessité parfois d'utiliser d'autres catégories témoignent d'une approche encore en réflexion, sans consensus strict.

On notera pour finir que la caractérisation fonctionnelle des gestes peut, le cas échéant, être complétée par le repérage d'une affiliation directe au plan verbal. On peut ainsi annoter parfois l'*affilié lexical* d'un geste, soit le mot (ou groupe de mots) dont le sens est le plus étroitement lié au sens du geste [65]. On peut également repérer des *affiliés prosodiques*, qui correspondent aux traits prosodiques saillants avec lequel un geste semble se synchroniser [66].

## 3.5 Complémentarité des approches

Les approches formelles et interprétatives offrent des perspectives distinctes mais profondément complémentaires pour l'analyse des gestes co-verbaux. La première centrée sur la description du mouvement, de l'espace d'exécution et des contraintes articulaires, fournie un cadre rigoureux permettant de caractériser les gestes en termes de structure, de dynamique, d'amplitude et d'organisation spatiale. Elles s'attachent à décomposer le geste en paramètres mesurables – positions, trajectoires, ddl, etc. – qui rendent possible une modélisation fine et une analyse quantitative du mouvement. Les approches interprétatives, quant à elles, mettent l'accent sur la fonction, la valeur sémantique et la relation du geste au discours, en s'intéressant aux intentions communicationnelles, aux dimensions pragmatiques et aux processus énonciatifs. Plutôt que d'opposer ces perspectives, leur articulation permet d'appréhender le geste comme un phénomène à la fois moteur et signifiant : les propriétés physiques du mouvement conditionnent les formes que peut prendre le geste, tandis que ses fonctions discursives orientent le choix, la structuration et la synchronisation de ces formes. L'intégration de ces deux dimensions (biomécanique et sémiotique) constitue ainsi un cadre d'analyse particulièrement

riche pour comprendre la complexité des gestes co-verbaux et leurs multiples rôles dans la communication multimodale.

# 4 Conclusion de chapitre

L'analyse gestuelle en science du langage repose historiquement sur des méthodes d'observation, de transcription et d'annotation manuelle particulièrement exigeantes. Chaque séquence gestuelle doit être segmentée, décrite et interprétée dans son contexte verbal, prosodique et interactionnel. Cette approche, fondée sur une expertise humaine, garantit une finesse d'analyse et une sensibilité au sens, mais elle implique un investissement temporel considérable. Annoter quelques minutes d'interaction peut nécessiter plusieurs heures, voire plusieurs jours de travail, selon la complexité du comportement à décrire et le niveau de détail recherché.

Les descriptions fines, telles que celles portant sur les phases gestuelles (préparation, *stroke*, tenue, rétraction) ou sur les fonctions sémiotiques du geste (iconique, déictique, rythmique, métaphorique, etc.), mobilisent des critères parfois ambigus, dont la mise en œuvre demande de nombreux allers-retours entre les annotateurs et le corpus. Ce caractère chronophage limite la taille des corpus analysables et rend difficile la généralisation des résultats. Il explique également la rareté des bases de données gestuelles annotées de manière exhaustive et standardisée, freinant la reproductibilité et la comparaison entre études.

Outre la charge de travail, les approches descriptives en sciences du langage se heurtent à des difficultés méthodologiques profondes. La première tient à la subjectivité et la difficulté inhérente à l'interprétation du geste : même avec des consignes détaillées, les annotateurs peuvent diverger dans leurs critères de segmentation, leur choix de catégorisation fonctionnelle ou leur manière d'interpréter la relation geste-parole. Puisqu'il n'existe pas de consensus strict sur la manière d'annoter certains gestes, il n'est pas garanti que l'annotation soit comparable à celle produite par d'autres équipes de recherche. La granularité de l'analyse, par exemple, peut ainsi poser un problème : certaines études privilégient une description très fine, *frame* par *frame* ou par micro-segment, tandis que d'autres opèrent à un niveau plus global, en considérant le geste comme une unité fonctionnelle dans le discours. Ce choix influence directement les résultats, rendant difficile la comparaison entre travaux ou la création de standards communs. Néanmoins, des initiatives telles que des sessions collaboratives impliquant plusieurs chercheurs issus de différentes équipes permettent de renforcer la cohérence et de favoriser une harmonisation des pratiques.

La reproductibilité demeure donc un enjeu méthodologique central. La dépendance à des annotations humaines, souvent contextualisées et partiellement intuitives, limite la possibilité de reproduire exactement une analyse donnée. Dans un contexte où la recherche tend vers des approches quantitatives et comparatives à grande échelle, cette variabilité méthodologique devient un obstacle majeur. Loin de remettre en cause la valeur des approches descriptives, ces limites soulignent la nécessité d'une collaboration interdisciplinaire, combinant l'annotation

humaine, le *gold standard* pour interpréter et décrire les gestes, avec la rigueur et la scalabilité des approches computationnelles.

# Chapitre 2
# L'étude de la gestualité en sciences computationnelles

## 1 Le geste comme donnée numérique

La gestualité, longtemps étudiée dans une perspective linguistique, psychologique ou anthropologique, fait aujourd'hui l'objet d'un intérêt croissant de la part des sciences informatiques. Récemment, le geste a été investi non plus pour être compris, comme en sciences du langage, mais pour répondre à des objectifs technologiques et industriels : développer des systèmes capables de détecter, d'interpréter ou de reproduire automatiquement les gestes. L'essor des outils de capture de mouvement, des modèles d'apprentissage automatique et des méthodes de représentation multimodale ouvre ainsi la voie à une conception du geste non seulement comme phénomène lié à l'expression orale, mais aussi comme une donnée mesurable, analysable et modélisable. Cette transformation n'est pas seulement technique : elle s'accompagne d'un changement de paradigme, où le geste est progressivement envisagé à travers le prisme de sa structure dynamique, de sa coordination multimodale très précise allant jusqu'aux millisecondes (avec la parole, la prosodie, le regard, etc.), et de sa formalisation géométrique et temporelle.

### 1.1 Distinction entre gestes et mouvement

L'une des premières difficultés dans l'étude computationnelle de la gestualité réside dans la définition même du geste, qui ne recouvre pas les mêmes réalités que celles précédemment décrites. Dans les *gesture studies*, le geste est envisagé comme un signe expressif inscrit dans une situation de communication, tandis qu'en sciences computationnelles, la distinction entre geste et mouvement n'est pas toujours opérée. Les approches informatiques tendent ainsi à considérer le geste avant tout comme un événement moteur mesurable, décrit à travers des trajectoires, des vitesses ou des rotations dans un espace de coordonnées (le plus souvent

cartésien) plutôt que comme une unité communicative contextualisée. Le mouvement constitue alors la donnée de base, indépendamment de toute valeur sémiotique. Dans ce cadre, la dimension fonctionnelle du geste est rarement prise en compte : les modèles apprennent principalement à détecter des régularités cinématiques ou temporelles, souvent de nature rythmique, sans accès direct, notamment via des annotations gestuelles explicites, aux fonctions communicatives ou interprétatives du geste. Cette différence de niveau d'abstraction entre les deux approches est fondamentale. Là où les sciences du langage visent à interpréter ce qu'est un geste (en analysant ses apports interactionnels et ses corrélations de sens avec les formes réalisées en contexte), les sciences computationnelles cherchent à représenter le mouvement, le modéliser et le segmenter en tant que donnée numérique. Cette opération de formalisation n'est toutefois pas une fin en soi : elle s'inscrit généralement dans l'objectif de mettre en relation le mouvement avec le discours, en cherchant à aligner les différentes modalités dans lesquelles il intervient, afin de rendre ces relations exploitables par des modèles automatiques.

## 1.2 Représentation computationnelle du geste

Le passage du geste observé au geste modélisé repose sur la formalisation de sa représentation. En sciences computationnelles, cette représentation doit être à la fois suffisamment riche pour capturer la complexité du mouvement, et suffisamment structurée pour être traitée automatiquement. Deux grandes formes de représentation se distinguent et se complètent : les représentations basées sur le signal, et les représentations symboliques ou fonctionnelles.

Les représentations basées sur le signal décrivent le geste à partir de ses caractéristiques physiques et cinématiques. Le mouvement est alors représenté sous forme de trajectoires, de positions dans l'espace tridimensionnel ou de rotations, généralement obtenues par capture de mouvement (*motion capture*) ou par estimation de pose 3D à partir de la vidéo. Chaque *frame* correspond à un ensemble de coordonnées (x, y ou x, y, z) associées à différents points du corps (épaules, coudes, poignets, etc.). Si ces positions offrent une description simple du mouvement, elles présentent toutefois une limite majeure : elles ne capturent pas directement les informations de rotation essentielles pour décrire précisément les extrémités articulées. C'est pourquoi les représentations rotationnelles constituent un complément de représentation. Les rotations en trois dimensions peuvent être exprimées selon plusieurs formalismes, chacun avec ses avantages et ses contraintes. Les angles d'Euler, fondés sur trois rotations successives autour d'axes fixes, sont intuitifs mais sujets au *gimbal lock*, qui entraîne une perte de degré de liberté. Les matrices de rotation, exemptes de singularités, fournissent une description complète mais redondante, puisqu'elles mobilisent neuf paramètres pour trois degrés de liberté. Les quaternions offrent une solution plus compacte et numériquement stable, évitent le *gimbal lock* et permettent des interpolations fluides, ce qui explique leur large adoption en robotique, en *motion capture* et en animation, malgré leur manque d'intuitivité. Les *exponential maps* (ou représentations axe-angle [67]) décrivent une rotation par un vecteur dont la direction encode l'axe et la norme encode l'angle. Elles sont particulièrement adaptées à l'apprentissage automatique, bien que des ambiguïtés puissent apparaître pour des rotations proches de $\pi$.

Enfin, la rotation d'un articulateur peut être représentée par la représentation 6D *continous Rotation* [68]. Composée de deux vecteurs 3D, elle fournit un signal spatialement continu, contrairement aux quaternions, *exponential maps* et des angles d'Eulers, qui peuvent présenter des discontinuités pour des rotations proches. La représentation 6D est donc moins sujette aux biais d'approximation, et favorise une descente de gradient plus stable. La représentation du signal offre une description fine, continue et objectivable du mouvement, adaptée à l'analyse quantitative et à la modélisation par apprentissage automatique.

Les représentations discrètes, à l'inverse, traduisent le mouvement en catégories ou en unités manipulables par des algorithmes. En sciences computationnelles, ces catégories prennent fréquemment la forme de labels binaires (présence/absence), d'états ou de classes, qui visent à rendre le geste exploitable dans des cadres formels ou statistiques. Comparées aux descriptions fines et multidimensionnelles présentées dans le chapitre précédent, ces représentations impliquent souvent une réduction de la richesse descriptive, susceptible d'entraîner une perte de granularité spatio-temporelle. Dans le cas de la représentation du geste, l'adoption d'une discrétisation à grain relativement large conduit fréquemment à lisser ou à abstraire des dimensions fines du mouvement, au profit d'une catégorisation plus globale.

Ces approches restent limitées dans leur portée interprétative : sans annotation ou information contextuelle, elles ne permettent pas de distinguer les gestes des mouvements non communicatifs, qu'ils soient intentionnels ou non. Des travaux récents en modélisation commencent toutefois à intégrer, encore de manière limitée, des représentations discrètes issues des *gesture studies*, en particulier sous la forme de schémas hiérarchiques et multimodaux [69]. Dans ces approches, le niveau bas encode les caractéristiques différentes modalités (mouvement, audio, etc.) tandis que le niveau haut modélise la structure temporelle et fonctionnelle du geste, en capturant ses phases, ses types ou son alignement avec la parole. Cette structuration hiérarchique favorise une intégration plus fine des dimensions du geste, et elle ne se limite plus à la représentation simple du mouvement. Cette formalisation ouvre la voie à une compréhension plus globale de la communication humaine, où le geste est appréhendé non seulement comme un signal mesurable, mais comme une composante dynamique et signifiante du langage incarné. Toutefois, les travaux intégrant des dimensions descriptives ou interprétatives plus fines, telles que définies en sciences du langage, restent encore minoritaires.

# 2 Types de modélisation du geste

## 2.1 Reconnaissance et segmentation

En traitement automatique, diverses méthodes sont mobilisées afin de détecter un geste, donc de le distinguer du reste des gesticulations d'un locuteur, ou de reconnaître une phase de sa réalisation : certaines utilisent des seuils définis sur des *features* (vitesse, orientation, distance), d'autres s'appuient sur des modèles probabilistes ou statistiques [70, 71], ou sur des approches récentes d'apprentissage profond [72, 73, 74], capables de capturer les dynamiques temporelles complexes du signal gestuel. L'un des principaux défis reste la constitution de données annotées : réaliser une segmentation automatisée fine des phases peut s'avérer

extrêmement chronophage. Cette contrainte a conduit de nombreuses études à se limiter à des problèmes plus restreints, comme la détection d'une phase particulière (souvent le *stroke* [72, 75, 76]) ou l'identification binaire de la présence ou l'absence d'un geste [77, 78, 79, 80, 71, 69], notamment au travers de la détection de leurs limites temporelles [81]. Si quelques travaux récents se sont aventurés vers la reconnaissance multi-phases [70, 73], ces approches demeurent encore peu répandues et souffrent souvent de problèmes de reproductibilité ainsi que d'une forte dépendance aux données d'entraînement. Parallèlement, la détection du geste peut être abordée directement au sein de la tâche de classification.

## 2.2 Classification

La classification gestuelle a pour objectif d'attribuer une ou plusieurs étiquettes à des gestes temporellement identifiés, selon leur fonction communicative ou leur forme. On distingue ainsi généralement les classifications formelles [69, 82, 83] (ex. : main droite en hauteur, paume orientée vers l'intérieur, geste circulaire, forme de main globale) et fonctionnelles [69, 72] (ex. : déictique, iconique, métaphorique, battement) lorsqu'elles ne sont pas binaires. Selon le cadre méthodologique, l'apprentissage du modèle de classification peut être supervisé, à partir d'un corpus annoté manuellement sur la forme ou la fonction des gestes, ou non supervisée, lorsque les catégories émergent des données elles-mêmes. Les principaux défis résident dans la forte variabilité intra- et inter-individuelle des gestes, la rareté des données enrichies, et la difficulté à représenter des fonctions communicatives souvent implicites. La tâche de classification peut par ailleurs être utilisée pour guider la génération gestuelle [84, 85, 70], notamment avec l'apprentissage de styles [86] ou encore de représentations fonctionnelles précises afin de les reproduire [87, 88, 89].

## 2.3 Génération

La génération de gestes à partir de données multimodales vise à produire des mouvements naturels, expressifs et cohérents, généralement synchronisés avec une séquence de parole ou un contenu textuel comme donnée d'entrée. Elle nécessite donc une modélisation conjointe de l'intention communicative (aspects sémantiques et pragmatiques du contenu du discours) et de la réalisation motrice (propriétés cinématiques et temporelles des mouvements réalisés en parallèle). Selon les approches, les modèles s'appuient sur des données variées, envisagées ensemble ou séparément : texte écrit [90, 91], signal vocal [84], représentations sémantiques ou prosodiques [86], scénarios narratifs structurés, ou encore séquences gestuelles préexistantes [92], dans une logique de continuation d'animation (suite à un mouvement réel) ou d'animation complète. Plusieurs défis demeurent centraux : la synchronisation multimodale (alignement parole-gestes), l'expressivité (transmission fluide des intentions et affects), la variabilité (éviter la répétitivité et les gestes stéréotypés), ainsi que l'adaptation contextuelle (prise en compte du style du locuteur [86], du contenu du discours, ou de la situation interactionnelle). Les systèmes récents intègrent plusieurs modalités en entrée, comme texte + audio [85], texte + pose corporelle [93] ou encore texte + audio + pose [94]. Ces travaux trouvent des applications dans les agents conversationnels, les avatars virtuels, les dispositifs d'accessibilité (langue des signes,

communication augmentée), la formation, la simulation ou encore les industries créatives. Les recherches actuelles s'orientent vers des architectures multimodales et hiérarchiques capables de générer des comportements complets – gestes manuels, postures, regards, prosodie, expressions faciales –, avec un objectif de fluidité, d'expressivité et de cohérence contextuelle.

# 3 Outils pour la modélisation

## 3.1 Modélisation par règles

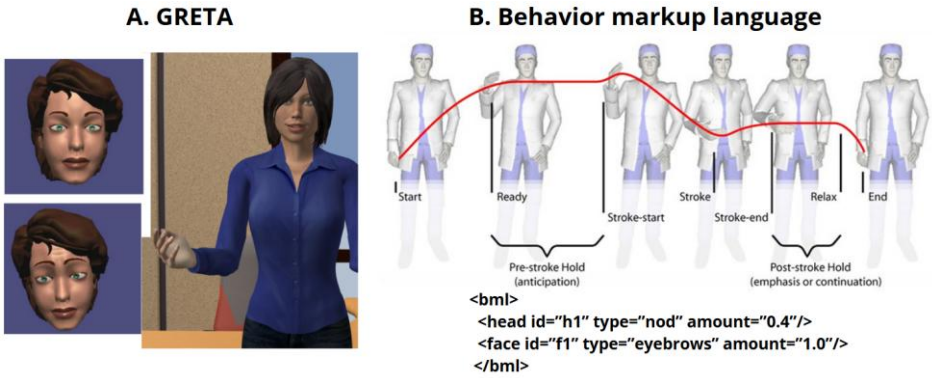



*Figure 6. (A) Visuels initiaux de l'agent GRETA [17] et l'une de ses mises à jour. (B) Illustration de l'utilisation du Behavior Markup Language (BML [95]) pour la génération de séquence gestuelle*

Les premières modélisations gestuelles à destination des agents conversationnels animés (ECA) reposaient principalement sur des systèmes à règles, où des comportements prédéfinis étaient déclenchés par des conditions linguistiques ou intentionnelles. Parmi les outils fondateurs figurent BEAT [96], NVBG [97] et Cerebella [98], ainsi que des extensions centrées sur des schémas conceptuels ou imagés [98, 99]. Ces approches, fondées sur des langages spécifiques au domaine souvent incompatibles entre elles, ont mené à la proposition du *Behavior Markup Language* [95] (Figure 6), désormais standard pour spécifier des comportements multimodaux. Malgré leur intérêt méthodologique, les systèmes à règles restent limités : ils reposent sur des animations fixes, peu variées, et produisent des enchaînements souvent peu naturels, comme l'illustre GRETA [17] (Figure 6), incapable de générer des gestes nouveaux ou adaptés au contexte dans ses versions initiales. Leur dépendance à des hypothèses figées empêche la modélisation des processus gestuels sous-jacents et limite leur capacité d'adaptation.

Ces limites ont favorisé l'émergence d'approches fondées sur les données, en particulier l'apprentissage automatique puis profond. Grâce aux avancées issues des études sur la parole, le texte et la vision, les modèles neuronaux permettent d'apprendre directement la variabilité et la synchronie multimodale sans passer par des règles explicites. Ils offrent la possibilité d'identifier des motifs latents dans les données et d'explorer de nouveaux cadres descriptifs. Selon les besoins – prise en compte des relations spatiales, des dépendances temporelles ou de l'alignement parole-geste – différents types de réseaux sont aujourd'hui employés pour modéliser la synchronie multimodale dans l'état de l'art. Les modèles les plus simples, tels que les réseaux à perceptrons multicouches (MLP), servent souvent de *baselines* ou de composants internes pour transformer des représentations, mais ils ne permettent pas de capturer directement la

dynamique temporelle. Ce type d'outils a notamment été utilisé dans le cadre de développement du *Gesticulator* [85]. Cependant, cette architecture est incapable de modéliser les relations spatiales et temporelles présentes dans les données, ce qui l'empêche de tirer pleinement parti de la richesse de l'entrée. Pour modéliser la structure séquentielle et les variations du mouvement, les approches contemporaines s'appuient plutôt sur des architectures conçues pour traiter des données temporelles capables de gérer à la fois la continuité du signal gestuel, ses dépendances à long terme et son alignement fin avec la parole.

## 3.2 Modélisation dynamique et temporelle

### 3.2.1 Réseaux de neurones récurrents (RNN)

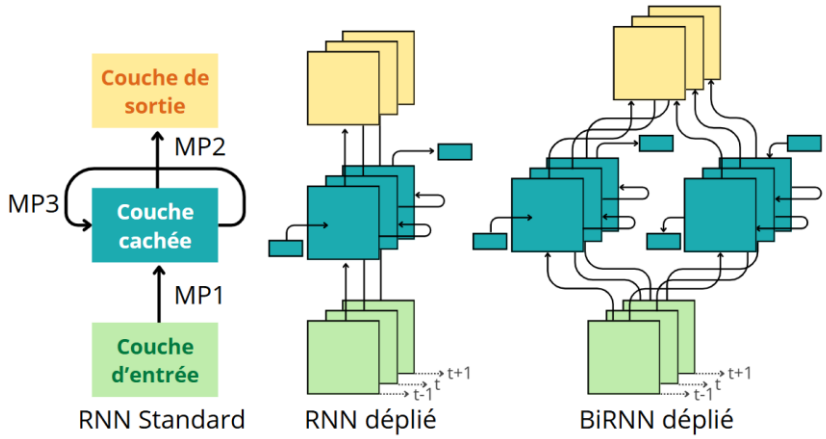



*Figure 7. Architecture de RNN standard et déplié, et de RNN bidirectionnel (flèches : matrices de poids MP)*

Contrairement aux MLP, les RNNs [100] introduisent la notion de réutiliser l'état précédent pour calculer l'état suivant. À l'instant t, la sortie est calculée à partir de l'entrée à t et de l'état caché à t-1, ce qui crée une « mémoire » des entrées passées (Figure 7) et cet état sera transmis à t+1. Cela permet d'exploiter efficacement la corrélation temporelle des données et d'apprendre des motifs temporels. Une variante de ces modèles, les réseaux bidirectionnels, repose sur deux RNN (biRNN) [101] (Figure 7) qui traitent la séquence dans des sens opposés : l'un du passé vers le futur et l'autre du futur vers le passé. Cette configuration permet au modèle d'exploiter simultanément le contexte antérieur et le contexte ultérieur, dépassant ainsi la limitation d'un traitement strictement linéaire et unidirectionnel des données. Cependant, les RNN et biRNN standards rencontrent en pratique des difficultés liées à la stabilité de l'apprentissage. En effet, la rétropropagation à travers le temps implique de faire remonter les gradients sur de longues chaînes d'opérations successives. Ce processus tend soit à réduire les gradients jusqu'à les rendre négligeables, soit à les amplifier de manière incontrôlée. Pour pallier ces limitations, plusieurs variantes de RNN ont été proposées : le LSTM (*Long Short-Term Memory*) [102], qui étend le RNN classique en introduisant des cellules de mémoire et des portes (entrée, oubli, sortie) permettant de réguler le flux d'information sur de longues séquences, et le GRU (*Gated Recurrent Unit*) [103], une version simplifiée des LSTM dotée de deux portes seulement (réinitialisation et mise à jour), plus rapide à entraîner tout en conservant la capacité à modéliser des dépendances temporelles étendues. Cet outil a été employé notamment pour la

génération de gestes [84, 104]. À titre d'exemple, Woo et ses collaborateurs [105] exploitent ce principe pour construire des représentations multimodales – combinant visage, tête et audio – tant pour un agent que pour un humain, dans le but de générer les gestes de l'agent.

Les LSTM et GRU sont conçus pour gérer des séquences temporelles qui se déroulent sur plusieurs *frames*, un atout pour modéliser les gestes. Leur mémoire récurrente permet de suivre les transitions et les motifs de mouvement sur des durées plus étendues que les CNNs (présentés plus loin). Ces modèles peuvent alors analyser la séquence pas à pas, en s'appuyant sur les éléments déjà produits, ce qui permet de préserver la dynamique naturelle du geste, contrairement à une prédiction *frame* par *frame* indépendante. Néanmoins, ces modèles présentent certaines limites : ils peuvent avoir du mal à capturer les relations entre gestes distants, sont sensibles aux hyperparamètres (taille des couches, longueur des séquences), et leur traitement séquentiel est moins parallélisable, ce qui peut limiter la longueur de séquence analysable.

### 3.2.2 Transformers

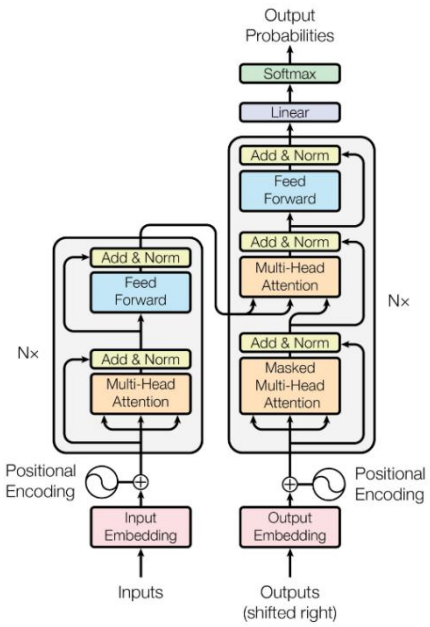



*Figure 8. Architecture d'un transformer [102]*

Les modèles de type RNN, LSTM ou GRU ont démontré leur efficacité pour traiter des séquences temporelles, mais leur traitement strictement séquentiel limite à la fois la vitesse d'apprentissage et la modélisation des dépendances longues ou complexes. Pour dépasser ces contraintes, les *Transformers* ont été introduits [106] (Figure 8) comme une architecture entièrement fondée sur des mécanismes d'attention. À la différence des RNN, les *Transformers* traitent l'ensemble de la séquence en parallèle, ce qui permet de capturer directement les interactions entre tous les éléments de la séquence, quelle que soit leur distance. Chaque couche combine un module d'auto-attention, qui met en relation chaque élément avec l'ensemble de la séquence, et un réseau de neurones entièrement connecté pour transformer ces représentations. Grâce à ce mécanisme, les *Transformers* sont particulièrement intéressants pour la modélisation de la gestualité co-verbale : l'auto-attention permet de capturer les dépendances non locales et asynchrones qui caractérisent le lien entre parole et geste (ex. : relier des traits prosodiques saillants et un stroke, même éloigné temporellement). Bien que très performants pour modéliser les dépendances à long terme (comme le lien sémantique entre un geste et un mot distant), les *Transformers* sont dépourvus de biais inductif local. Ils nécessitent, par conséquent, des volumes de données importantes pour apprendre efficacement ces régularités, ce qui les rend peu adaptés aux configurations reposant sur des corpus d'annotations expertes de taille réduite. Ce type de modèle a notamment été utilisé dans Text2gestures [107], afin de produire une séquence de mouvements depuis un texte.

L'empilement de plusieurs couches constitue l'encodeur, tandis que le décodeur génère les sorties à partir des représentations produites par l'encodeur, en intégrant éventuellement les informations des sorties précédentes pour guider la prédiction. Dans le cas de Text2Gestures, le modèle encode les *embeddings* de mots ainsi que des informations additionnelles (ex. : étiquette émotionnelle, genre du locuteur, ou manualité) et les corrélations entre ces différentes informations servent à orienter la génération des gestes. Bien que les *Transformers* soient bien plus gourmands en mémoire et en ressources de calcul, ils ont établi des performances de référence dans de nombreuses tâches, notamment avec des modèles tels que BERT [108] pour la représentation vectorielle de mots.

Ces modèles présentent un intérêt certain grâce à leur capacité à paralléliser le traitement des séquences, qui permet de gérer efficacement de longues séries de données gestuelles, et pour leur efficacité à produire des représentations de données pertinentes. Cependant, ils restent très sensibles aux choix de paramétrage et aux contraintes matérielles. La mémoire disponible impose une limite sur la taille de la fenêtre de *tokens* traitable, ce qui peut entraîner une troncation ou des approximations dans la représentation des données. De plus, les *Transformers* nécessitent de très grandes quantités de données pour réduire les biais auxquels ils sont sensibles et apprendre des régularités fiables, ce qui les rend peu adaptés aux corpus de gestes de petite taille.

## 3.3 Modélisation structurelle et explicite du corps

### 3.3.1 Réseaux de neurones convolutifs (CNN)

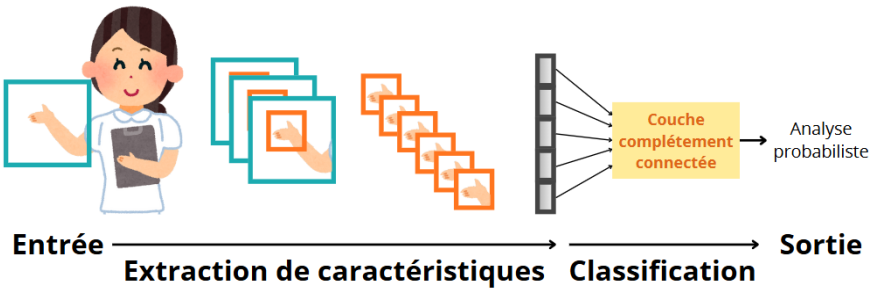



*Figure 9. Illustration de l'extraction de caractéristiques par un CNN, sur une image 2D, pour la classification*

Les outils présentés précédemment ne permettent pas de capturer efficacement la dimension spatiale. C'est là que les CNN (*Convolutional Neural Networks*) [109] interviennent (Figure 9). Les convolutions sont des opérations mathématiques appliqués à un champ réceptif destinées à extraire de nouvelles caractéristiques, dont les poids sont appris automatiquement par le réseau. En empilant plusieurs couches de convolution, le champ réceptif s'élargit et les caractéristiques extraites deviennent de plus en plus abstraites. Les CNN sont particulièrement adaptés aux données structurées comme les images 2D, mais peuvent aussi traiter efficacement des signaux 1D, comme la parole. Les CNN sont caractérisés par plusieurs paramètres qui influencent directement leur champ réceptif, c'est-à-dire la portion d'entrée observée par chaque neurone dans une couche donnée. Parmi ces paramètres, on distingue notamment la taille du

noyau (*kernel size*), qui détermine l'étendue locale du traitement, le pas de convolution (*stride*), qui définit le décalage du noyau entre deux positions successives, et le *padding*, qui permet de contrôler la dimension des cartes de caractéristiques intermédiaires. Cette flexibilité rend les CNN particulièrement adaptés à l'extraction hiérarchique de caractéristiques dans des données structurées, qu'il s'agisse d'images, de séquences temporelles ou de représentations gestuelles spatiales. Le biais inductif de localité spatio-temporelle propre aux CNN est particulièrement adapté au signal gestuel : il permet de capter les variations brusques de vélocité typiques des phases de préparation et de *stroke*, indépendamment de leur position exacte dans la séquence. Dans de nombreux travaux, CNN et RNN sont combinés pour extraire des représentations abstraites et les analyser dans le temps. Par exemple, Tuyen et ses collaborateurs [110] utilisent des CNN pour transformer progressivement des représentations d'entrée en représentations plus élevées, puis revenir à des caractéristiques de sortie, ici pour générer des gestes contrôlant un robot Pepper.

Il est à noter que les opérations de *pooling* (réduction des dimensions des cartes de caractéristiques extraites) peuvent entraîner une perte d'information spatiale, nuisant à la localisation exacte des motifs gestuels saillants. Ces modèles restent également vulnérables au décalage de domaine (*domain shift*), c'est-à-dire lorsque la distribution des données de test diffère de celle des données d'entraînement. Dans le domaine de l'étude des gestes, cela peut concerner des variations de locuteur, d'angle de caméra, de vitesse ou de style de mouvement, susceptibles de réduire significativement les performances si le modèle n'a pas été exposé à une diversité suffisante de conditions. Malgré ces limites, les CNNs présentent l'avantage de capturer les motifs locaux au sein des séquences, ce qui est particulièrement pertinent pour l'analyse des gestes et l'examen précis de segments locaux de mouvement. Grâce au partage des poids, un même filtre est appliqué sur l'ensemble de l'entrée, ce qui réduit le nombre de paramètres du modèle. C'est notamment pour cette raison qu'ils restent adaptés lorsque les données disponibles sont limitées, les rendant attractifs pour des configurations expérimentales pratiques.

### 3.3.2 Réseaux de neurones graphiques (GNN)

Les réseaux de neurones graphiques (GNN) [111] permettent de traiter des données structurées, comme les graphes, où les voisins ne sont pas uniformément organisés. Dans les CNN classiques, le noyau de convolution peut agréger des nœuds sans lien de voisinage direct, ce qui perturbe la structure spatiale de l'entrée. Les GNN contournent ce problème en utilisant une matrice d'adjacence qui guide l'agrégation des nœuds lors de la convolution, préservant ainsi la structure du graphe tout en conservant l'efficacité de traitement des CNN. Cette capacité à exploiter explicitement la topologie des données est particulièrement utile pour les squelettes humains, où les articulations et segments anatomiques forment un graphe naturel. Des GNN spatiotemporels (STGCN) ont ainsi été employés pour modéliser simultanément les dimensions spatiale et temporelle des données squelettiques, tant pour la reconnaissance d'actions [112, 113] que pour la génération de gestes [94]. Ces architectures peuvent également être combinées avec des RNN, LSTM ou *Transformers* pour capturer les dépendances temporelles tout en conservant la structure spatiale. La modélisation de gestes à l'aide de GNN,

en combinaison avec des architectures temporelles, permet de capturer à la fois la structure spatiale et les dynamiques temporelles des mouvements.

Cependant, la complexité croissante de ces modèles soulève des questions cruciales sur la compréhension de leur fonctionnement et la justification de leurs prédictions. En effet, les GNNs sont limités par la complexité des structures qu'ils doivent représenter. Plus la structure de données (nombre de nœuds et nombre de *features* par nœud) est complexe, plus le modèle peut être sujet à de l'*over-smoothing* (les représentations des nœuds deviennent indiscernables avec l'empilement des couches) ou à de l'*over-squashing* (une trop grande quantité d'informations lointaines est compressée dans un vecteur de taille fixe, limitant leur propagation). Les gestes, pouvant présenter des structures complexes, par exemple en combinant plusieurs articulations, chacune étant liée à de multiples paramètres (positions dans l'espace, rotations, ou autres composantes saillantes du geste), bénéficient de la capacité des GNNs à capturer cette topologie ; toutefois, cette même complexité peut devenir problématique si elle n'est pas correctement gérée par le modèle, en raison des risques d'*over-smoothing* ou d'*over-squashing* mentionnés précédemment.

### 3.3.3 Modèles de diffusion

Les modèles de diffusion sont des modèles génératifs, initialement développés pour la génération d'images [114], et désormais étendus à d'autres types de signaux. Leur principe repose sur deux processus itératifs : une diffusion vers l'avant et une diffusion inverse. Le premier consiste à corrompre progressivement une donnée issue du jeu d'entraînement en y ajoutant du bruit aléatoire (généralement modélisé par une distribution gaussienne), tandis que le second correspond à l'apprentissage du processus inverse, visant à restaurer un échantillon cohérent à partir de ce bruit. L'apprentissage du modèle consiste ainsi non seulement à débruiter les données, mais aussi à apprendre implicitement leur structure sous-jacente à travers les relations entre les versions progressivement bruitées et les versions reconstruites. Une fois entraîné, le modèle peut alors échantillonner une distribution de bruit et la transformer en une sortie de haute qualité.

Dans le cadre de la génération de gestes, ces modèles sont de plus en plus mobilisés. Par exemple, lors du GENEA 2023 [115], la majorité des auteurs ont utilisé des modèles de diffusion pour le *gesture generation challenge.* Le modèle de Deichler et ses collaborateurs [116] présentent de très bons résultats d'évaluation perceptive, en comparaison avec les autres auteurs participant du *challenge.* On peut notamment citer DiffGesture [117], qui souligne plusieurs avantages des modèles de diffusion, tels qu'une meilleure qualité des mouvements générés et une diversité plus importante par rapport aux approches précédentes, comme celle des *Transformers.* D'après les auteurs, ces modèles permettraient en effet de mieux capturer les associations intermodales, telles que les correspondances entre audio et mouvement, tout en maintenant la cohérence temporelle. Les modèles de diffusion semblent donc particulièrement prometteurs pour la génération de gestes, grâce à leur capacité à capturer la structure temporelle et intermodale du mouvement tout en produisant des séquences réalistes et cohérentes.

Malgré leurs performances remarquables en génération, les modèles de diffusion présentent plusieurs limites. Leur processus itératif implique des temps d'inférence élevés et des coûts computationnels importants, rendant leur usage délicat dans des contextes temps réel (tels que les ECAs). Ils requièrent en outre de larges volumes de données pour apprendre des distributions riches et diversifiées, et sont très sensibles aux biais induits par les données. La complexité de leurs architectures et la sensibilité aux choix d'hyperparamètres compliquent également leur déploiement, la compréhension des processus internes au modèle et leur reproductibilité, la sortie du modèle étant souvent difficile à contrôler. Enfin, l'intégration de contraintes explicites propres à certains types de données peut représenter un défi supplémentaire lors de la mise en place de ces modèles.

# 4 Interprétabilité et explicabilité des modèles

## 4.1 Définition

Interprétabilité et explicabilité sont deux notions distinctes en lien avec l'analyse des modèles. L'interprétabilité désigne le degré de compréhensibilité qu'un humain peut avoir du fonctionnement d'un modèle ou de ses décisions [118]. L'explicabilité, quant à elle, se rapporte à la capacité d'un modèle à produire des représentations de son fonctionnement interne qui soient interprétables. Elle suppose la mise en évidence de relations directes – et non de simples corrélations – entre les différentes étapes ou états du modèle, relations généralement exprimées en termes de causes ou de raisons. Dans le cas d'un modèle algorithmique, l'explicabilité porte à la fois sur l'algorithme lui-même, sur ses étapes et articulations internes, ainsi que sur les caractéristiques d'entrée et de sortie. Expliquer un modèle revient donc à fournir, d'une part, une interprétation de ses entrées et sorties, et d'autre part, une interprétation des relations causales plus ou moins directes qui les relient. Autrement dit, l'interprétabilité renvoie à la compréhension du fonctionnement interne et de la logique du modèle, tandis que l'explicabilité concerne la justification de ses décisions en sortie à partir des facteurs ou variables mobilisés.

À l'heure actuelle, très peu de modèles sont étudiés en détail pour comprendre précisément leur fonctionnement, et ce pour plusieurs raisons. On distingue généralement trois types de contextes en termes d'interprétabilité : les *boîtes blanches*, les *boîtes grises* et les *boîtes noires*, allant du plus favorable au moins favorable en matière de transparence. Dans le cas d'une *boîte blanche*, le fonctionnement interne du modèle est entièrement connu et interprétable : chaque étape du traitement est identifiable, traçable et explicable. À l'inverse, dans le cas d'une *boîte noire*, seules l'entrée et la sortie sont connues, tandis que le processus interne reste obscur – soit parce qu'il n'a pas encore été analysé (par manque de temps ou de moyens), soit parce que l'architecture du modèle elle-même ne permet pas un suivi explicite des données, rendant l'interprétation difficile, voire impossible. Entre ces deux extrêmes se trouve le cas de la *boîte grise*, dans lequel une partie du fonctionnement est comprise ou documentée, mais sans que le modèle soit totalement transparent ou explicable dans son ensemble.

## 4.2 Hyperparamètres et fonction de perte

Les paramètres du modèle sont organisés par couche sous la forme de matrices ou, plus généralement, de tenseurs de poids. Leur observation peut fournir certaines informations partielles, telles que la distribution des poids, la sparsité ou certaines structures locales. Cependant, elles ne permettent pas de comprendre directement la manière dont le réseau encode les informations : deux réseaux ayant des poids différents peuvent apprendre des représentations internes très similaires. En revanche, l'étude de l'impact des paramètres d'apprentissage (tels que le taux d'apprentissage, la régularisation ou certains choix architecturaux) sur les performances permet d'analyser plus indirectement la dynamique d'apprentissage du modèle. Observer l'évolution des performances lorsque ces hyperparamètres varient peut fournir des indications sur la stabilité, la capacité de généralisation ou la sensibilité du modèle, même si cela ne donne pas accès de manière explicite aux représentations internes apprises.

La fonction de perte (*loss*) quantifie la performance du modèle localement (sur un *batch* ou un ensemble de données) en mesurant l'écart entre les prédictions du modèle et les valeurs attendues. Elle est utile pour suivre l'apprentissage, détecter l'*overfitting* (sur-apprentissage des données d'entrainement) ou comparer différents modèles, et constitue ainsi une première indication sur la qualité du modèle, même si elle ne renseigne pas directement sur la structure ou les représentations internes. Le type de modélisation a un impact direct sur le choix de fonction de perte. Plusieurs options existent, selon le type de tâche : pour la classification, on utilise généralement la *binary cross-entropy* (binaire) ou la *categorical cross-entropy* (multi-classes) ; pour la régression, on privilégie le *mean squared error* (MSE) ou le *mean absolute error* (MAE) ; pour des tâches temporelles ou séquentielles, ces fonctions peuvent être adaptées pour traiter des séquences de sorties. Si la *loss* est mal adaptée ou mal optimisée, le réseau peut produire des prédictions imprécises ou incohérentes, reflétant mal la structure réelle des données. Une mauvaise convergence se traduit par des activations internes peu représentatives.

## 4.3 Outils d'analyse des représentations internes

Lorsque l'architecture du modèle le permet, il est possible d'examiner ce que le réseau apprend réellement, au-delà des simples mesures de performance globale. Plusieurs outils permettent d'analyser les représentations internes des modèles. Ces analyses offrent la possibilité d'identifier quelles unités s'activent pour certaines entrées, de comprendre les transformations appliquées aux données ou encore de mettre en évidence la structure de celles-ci dans l'espace latent.

Les méthodes d'analyse des activations et de projection de l'espace latent permettent d'explorer la structure interne des données telles qu'elles sont représentées au sein des modèles. Ces approches visent à comprendre comment le modèle organise l'information dans ses couches intermédiaires et à mettre en évidence des patterns ou des regroupements caractéristiques. Parmi les techniques couramment utilisées, on peut citer l'analyse en composante principale (PCA), qui effectue une réduction de dimension linéaire pour visualiser les tendances principales des activations, et le t-SNE *t-distributed Stochastic Neighbor Embedding* (t-SNE [119]), qui

propose une réduction non linéaire favorisant la mise en évidence des clusters et de la structure locale des données. Une alternative plus récente, *Uniform Manifold Approximation and Projection* (UMAP [120]), offre une réduction non linéaire plus rapide tout en conservant à la fois la structure globale et locale.

Pour comparer les représentations internes entre couches ou entre modèles, des méthodes comme le *Centered Kernel Alignment* (CKA [121]) permettent de quantifier la similarité des représentations, tandis que la *Singular Vector Canonical Correlation Analysis* (SVCCA [122]) met en relation les sous-espaces des activations pour identifier des correspondances structurelles.

Les méthodes d'attribution de l'importance des *features* permettent de comprendre quelles entrées influencent le plus les prédictions d'un modèle. Parmi les approches les plus utilisées, les *SHapley Additive exPlanations* (SHAP [123]) et les Integrated *Gradients* [124] quantifient la contribution de chaque *feature* à une prédiction donnée, offrant ainsi une mesure rigoureuse et théoriquement fondée de l'importance des entrées. *Local Interpretable Model-agnostic Explanations* (LIME [125]) adopte une approche différente, en approximant localement un modèle complexe par un modèle simple, interprétable, afin d'expliquer la prédiction pour un exemple spécifique.

Dans le cas des *Transformers*, l'interprétation des mécanismes d'attention constitue un outil précieux pour comprendre comment le modèle relie les différentes parties d'une séquence. La méthode *Attention Roll-out* [126] permet de calculer la contribution cumulée des poids d'attention à travers les différentes couches. Cette approche permet d'identifier quelles parties de la séquence sont réellement prises en compte par le modèle pour produire ses prédictions, et de mieux comprendre les dépendances non locales et asynchrones capturées par les mécanismes d'auto-attention.

Dans le contexte de l'étude des gestes, ces outils offrent un avantage certain : ils permettent de visualiser et de comprendre comment le modèle encode les informations apprises, et d'identifier d'éventuelles erreurs ou biais dans la représentation des données. Ils complètent ainsi les évaluations classiques basées sur la performance globale, en offrant une lecture plus fine de ce que le modèle « voit » et apprend à partir des données d'entraînement. Cependant, ces méthodes n'expliquent pas directement les concepts sémantiques sous-jacents aux gestes : elles fournissent principalement une interprétation technique des processus internes du modèle, mais ne suffisent pas à réaliser une analyse fonctionnelle automatique des gestes ou à capturer leur signification communicative. Cependant certains travaux récents tentent d'avancer dans cette direction en concevant des architectures spécifiques, capables de produire des représentations gestuelles interprétables par des spécialistes [89]. Le modèle est alors conçu pour apprendre des représentations robustes des gestes, facilitant ainsi l'analyse et l'interprétation de son apprentissage.

# 5 Evaluation des modèles

## 5.1 Métriques quantitatives

La capacité du modèle à effectuer de bonnes prédictions est évaluée à l'aide de différentes métriques quantitatives [127]. Ces calculs reposent sur les étiquettes attribuées aux prédictions du modèle : dans le cas d'une prédiction *frame* par *frame par exemple*, chaque *frame* prédite est catégorisée selon sa correspondance avec la vérité terrain. On distingue alors quatre types de prédictions : les vrais positifs (TP, prédiction correcte de la présence d'une caractéristique), les faux positifs (FP, prédiction à tort de la présence d'une caractéristique), les vrais négatifs (TN, prédiction correcte de l'absence d'une caractéristique) et les faux négatifs (FN, prédiction à tort de l'absence d'une caractéristique).

Parmi les mesures les plus couramment utilisées, *l'accuracy* ($\frac{TP+TN}{TP+TN+FP+FN}$) correspond à la proportion de prédictions correctes parmi l'ensemble des observations, combinant à la fois les vrais positifs et les vrais négatifs. La précision ($\frac{TP}{TP+TP}$) mesure la fiabilité des prédictions positives, c'est-à-dire la proportion de vrais positifs parmi toutes les prédictions positives du modèle. Le rappel ($\frac{TP}{TP+TN}$) indique la capacité du modèle à détecter toutes les occurrences réelles d'un événement, en évaluant la proportion de vrais positifs parmi l'ensemble des événements réellement présents. Enfin, le F1-score ($\frac{Precision \times Rappel}{Précision+Rappel} \times 2$) constitue une mesure combinée de précision et de rappel, offrant un compromis entre ces deux critères et permettant de résumer la performance globale du modèle sur les prédictions positives.

Il convient de noter que dans le contexte de la segmentation de séries temporelles, exiger une exactitude stricte à la *frame* près (*frame-wise*) peut s'avérer trop rigide, les frontières du geste humain (ex. : début de la préparation, fin de la rétractation) étant intrinsèquement floues. C'est pourquoi, dans les approches computationnelles de détection d'action, les métriques *frame-wise* sont généralement complétées par des métriques *segment-wise* intégrant une marge de tolérance temporelle (telles que l'*Intersection over Union*), permettant de considérer une prédiction comme correcte dès lors qu'elle recouvre la majorité du segment expert. L'*Intersection over Union* (IoU) correspond au ratio entre l'intersection (la durée où le segment prédit et le segment réel se chevauchent) et l'union (la durée totale couverte par les deux segments). Une IoU élevée indique que le segment prédit correspond étroitement au segment réel, tandis qu'une IoU faible signale un décalage ou un chevauchement partiel.

Dans le cadre de l'évaluation de la génération de mouvements, plusieurs métriques objectives peuvent être calculées pour quantifier la qualité des séquences générées. Les métriques classiques comparent les positions ou les rotations des joints entre le mouvement de référence et le mouvement généré. La *Mean Absolute Error* (MAE) calcule l'erreur absolue moyenne des coordonnées des joints sur l'ensemble des *frames*. Chaque écart contribue de manière équivalente, sans pénalisation plus forte pour les grandes erreurs, et la métrique fournit une mesure directe de la distance moyenne, exprimée dans les mêmes unités que les données. La *Mean Squared Error* (MSE) mesure l'écart quadratique moyen entre les positions prédites

et réelles des joints. En élevant les écarts au carré, elle accentue l'impact des erreurs importantes et est particulièrement sensible aux prédictions très éloignées de la vérité. La *Root Mean Squared Error* (RMSE) correspond à la racine carrée de la MSE, permettant de revenir à la même unité que les données et de mieux interpréter l'amplitude moyenne des erreurs. Enfin, la *Percentage of Correct Keypoints* (PCK) calcule le pourcentage de joints 3D prédits se trouvant à moins d'un seuil fixé (par exemple 15–50 mm) par rapport aux positions de référence

Il est également possible de comparer les distributions statistiques des mouvements réels de ceux produits par un modèle de génération. La *Fréchet Gesture Distance* (FGD) [128] permet d'évaluer la similarité globale, en encodant dans un espace latent, à l'aide d'*embeddings,* ces mouvements, puis compare les distributions gaussiennes de ces représentations. Elle mesure à la fois la distance entre les moyennes et les différences de covariance de ces distributions. Une valeur faible de FGD indique que les gestes générés reproduisent de manière réaliste la variabilité et les corrélations observées dans les gestes naturels. Bien que la valeur de cette métrique ne soit pas directement interprétable en termes physiques, elle semble corrélée de manière plus fiable aux évaluations perceptives (présentées dans la section suivante) des mouvements [129], en comparaison avec d'autres métriques telles que l'accélération moyenne, la vitesse et des analyses statistiques de corrélations entre deux ensembles de données.

## 5.2 Évaluations perceptives

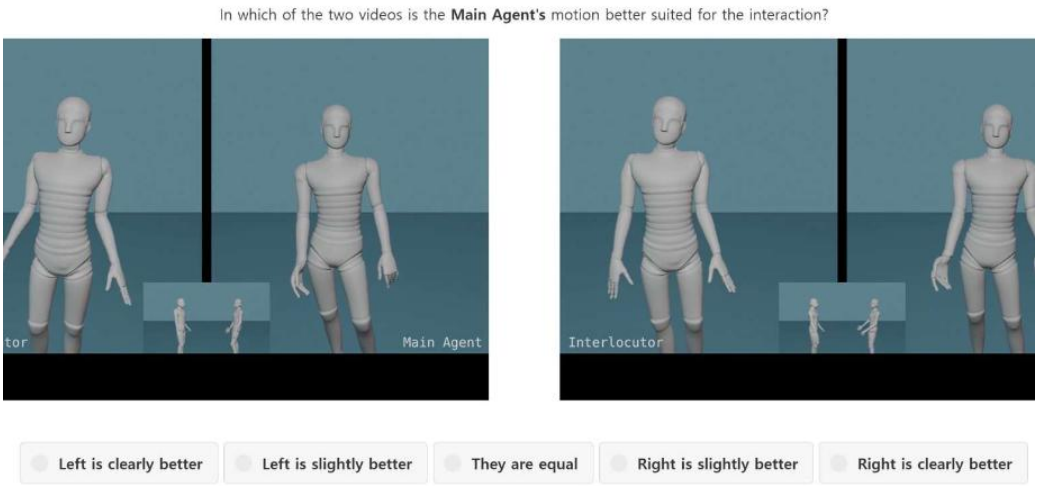



*Figure 10. Capture d'écran de l'interface utilisée lors de l'évaluation perceptive [115, 130]*

Dans le cadre de modèles de génération gestuelle, comme ceux utilisés pour animer des agents conversationnels virtuels, les méthodes d'évaluation traditionnelles, fondées sur des métriques objectives ou quantitatives, se révèlent souvent insuffisantes. En effet, les gestes sont des phénomènes hautement perceptifs, expressifs et contextuellement dépendants. Leur pertinence ou leur réalisme ne peut pas toujours être jugé à l'aide de simples comparaisons numériques avec des données de référence. C'est pourquoi il est désormais courant de recourir à des évaluations subjectives, impliquant des participants humains dans le processus d'analyse qualitative des productions, comme le propose le *GENEA Challenge* [131, 115] (Figure 10), initiative de référence dans ce domaine. Il s'agit généralement d'évaluation à grande échelle (via

plateformes de *crowdsourcing*) des gestes générés automatiquement par différents modèles, dans l'objectif de tenter d'évaluer et de classer leurs performances sur la base d'un protocole d'évaluation identique.

Trois dimensions principales sont prises en compte pendant ces évaluations. La première concerne le caractère humain des mouvements : les participants doivent juger si les gestes semblent naturels et réalistes (*Human-likeness*), sans prendre en compte le contenu du discours ou la présence d'un interlocuteur. Cette première évaluation permet d'estimer la fluidité, la biomécanique et la crédibilité du mouvement de manière isolée. La deuxième dimension vise à mesurer l'adéquation des gestes au discours de l'agent (*Appropriateness for agent speech*). Ici, les évaluateurs analysent si les gestes produits accompagnent efficacement et pertinemment le contenu verbal, tout en maintenant une certaine qualité gestuelle globale. Cette évaluation permet d'examiner la synchronisation et la complémentarité entre les modalités verbales et gestuelles. Enfin, la troisième dimension, dite dyadique, est plus complexe car elle implique un interlocuteur virtuel : les participants évaluent l'adéquation des gestes de l'agent au comportement de son interlocuteur (*Appropriateness for interlocutor behaviour*), qu'il s'agisse de la parole ou des gestes de celui-ci. Cette tâche est particulièrement pertinente dans le contexte de la co-construction du dialogue et de la synchronisation interactive, puisqu'elle interroge la capacité du modèle à produire des gestes ajustés, cohérents et dynamiques dans une interaction simulée.

Cependant, plusieurs travaux récents [21], soulignent des limites importantes de ces évaluations menées de manière individuelle. Tout d'abord, les protocoles utilisés sont rarement standardisés : les questions posées, les types de mouvements évalués et les méthodes de réponse varient d'une étude à l'autre, ce qui rend difficile la comparaison des résultats et l'interprétation des scores. Ensuite, ces évaluations tendent à être plus performatives qu'informatives [22], c'est-à-dire qu'elles sont souvent conçues et réalisées dans une perspective de publication (en tant qu'étape obligatoire pour l'acceptation d'un papier), plutôt que dans l'objectif de fournir une analyse rigoureuse des performances du modèle et donc de pouvoir les comparer entre eux. Enfin, de manière générale, les sorties des modèles de gestes synthétiques sont jugées peu crédibles par les évaluateurs [21], quelle que soit la dimension considérée, ce malgré les bons scores en évaluation perceptive. Cela met en évidence le besoin de méthodes d'évaluation plus robustes, plus détaillées et plus objectives, qui pourraient dès lors servir à guider le développement de modèles gestuels plus fiables en ciblant avec précision les aspects actuellement insatisfaisants des gestes de synthèse, ce que ne fournissent pas à l'heure actuelle ces évaluations. Dans tous les cas, ces études lorsqu'elles sont informatives s'avèrent essentielles, car elles permettent de rendre compte de la qualité formelle/dynamique (*human-likeness*), expressive (*speech appropriateness*) et interactive (*behavior appropriateness*).

# 6 Conclusion de chapitre

Les approches computationnelles appliquées à l'étude de la gestualité reposent sur la modélisation et le traitement automatique de données hautement dimensionnelles, souvent issues de la capture de mouvement ou de l'estimation de poses à partir de la vidéo. Ces données,

bien qu'abondantes et précises, imposent des contraintes fortes en termes de calcul, de traitement et de représentation. Le volume des signaux à manipuler – parfois plusieurs centaines de milliers de coordonnées articulaires – nécessite des architectures de modèles capables de préserver la cohérence temporelle tout en réduisant la complexité spatiale.

L'apprentissage automatique, qu'il soit supervisé ou non supervisé, requiert généralement par ailleurs de grandes quantités de données. Or, la majorité de ces grands corpus ne sont pas annotés, et les gestes ne se distinguent pas nécessairement du flux continu de mouvements. En conséquence, les modèles sont souvent entraînés sur des représentations incomplètes voire inexistantes du geste, privilégiant la dimension cinématique, sans distinguer ce qui est réellement communicatif de ce qui ne l'est pas, et au détriment des aspects fonctionnels des gestes et de leurs liens étroits avec la parole. Enfin, les choix techniques – type d'encodeur, granularité temporelle, représentation du mouvement (angles, quaternions, coordonnées cartésiennes) – influencent directement la qualité de la modélisation. Ces contraintes computationnelles, si elles garantissent une formalisation rigoureuse du mouvement, tendent parfois à réduire la richesse sémiotique du geste à une suite de vecteurs numériques, sans pouvoir le distinguer des mouvements non communicatifs.

Si les approches computationnelles ont permis de franchir un cap dans la modélisation automatique du mouvement, elles demeurent souvent éloignées des préoccupations théoriques et méthodologiques des sciences du langage. Les modèles entraînés à détecter ou classifier des gestes s'appuient rarement sur une conceptualisation fine des catégories gestuelles ou des fonctions communicatives, pourtant estimées centrales dans les *gestures studies*. Cette déconnexion conduit à une représentation partielle du geste : les modèles identifient des régularités spatiales et temporelles sans nécessairement relier ces régularités à de possibles rôles interactionnel ou discursif. Autrement dit, ils décrivent « comment » le corps bouge, mais rarement « pourquoi ». De plus, les critères d'évaluation des modèles (précision, rappel, F1-score) privilégient la performance quantitative au détriment de la validité interprétative. Or, une détection exacte d'un geste n'implique pas nécessairement une compréhension correcte de sa fonction dans le discours.

En dépit de leurs progrès techniques, les approches computationnelles restent limitées dans leur capacité à relier la modélisation du mouvement à l'interprétation fonctionnelle et formelle des gestes.

# Bilan de partie

Les *gesture studies* ont apporté plusieurs contributions majeures à l'étude du geste en modifiant profondément la manière dont celui-ci est conceptualisé dans les sciences du langage. Le geste n'est plus considéré comme un simple substitut de la parole ni comme un accompagnement secondaire du discours verbal, mais comme une composante à part entière de la communication multimodale, participant activement à la construction du sens en interaction étroite avec la parole. Dans cette perspective, le geste est appréhendé comme un objet d'étude autonome, doté de propriétés propres tout en étant intrinsèquement lié aux autres modalités auxquelles il se combine. Les méthodologies développées dans ce champ d'étude visent ainsi à rendre compte de la complexité des gestes à travers des descriptions fines et systématiques. Elles permettent notamment d'analyser leurs fonctions communicatives, leurs caractéristiques formelles et spatiales (trajectoire, configuration de la main, l'espace gestuel mobilisé...), ainsi que leurs relations temporelles et sémantiques avec le discours. Pour mener de telles analyses, les travaux en *gesture studies* s'appuient majoritairement sur des procédures d'annotation multimodale détaillées, souvent réalisées manuellement par des experts à l'aide de logiciels dédiés. Si cette approche garantit une grande richesse descriptive et une forte validité interprétative, elle requiert néanmoins un investissement humain conséquent, ce qui limite son applicabilité à des corpus de très grande taille, et limitent de fait fortement la portée des analyses et leur généralisabilité.

Plus récemment, les approches computationnelles de la gestualité ont donné accès à de nouveaux outils et de paradigmes pour l'étude et la modélisation du geste. Là où les sciences du langage privilégient une analyse qualitative, contextualisée et interprétative, les méthodes issues de l'apprentissage automatique et du traitement du signal reposent sur une formalisation numérique du mouvement, dans l'objectif de représenter les gestes sous forme de séries temporelles, de vecteurs de caractéristiques ou de représentations latentes exploitables par des algorithmes. Ce décalage méthodologique ne traduit pas une opposition entre disciplines, mais plutôt une différence de granularité et de finalité dans la manière de rendre compte du phénomène gestuel : les premières visent principalement la compréhension approfondie du geste, ancrée dans la situation communicative et dans son articulation avec le discours, tandis que les

secondes cherchent à modéliser automatiquement ses régularités, à prédire son occurrence ou à en générer des réalisations réalistes.

Plusieurs limites ont pu être observées dans ces deux domaines. En sciences du langage, l'enrichissement de corpus sur lequel se fonde une analyse repose majoritairement sur des procédures manuelles. Bien que celles-ci produisent des données souvent considérées comme des *gold standards*, elles demeurent particulièrement coûteuses et complexes : le nombre d'experts disponibles est restreint, le temps nécessaire à l'annotation est considérable, et certaines caractéristiques du geste restent difficiles à appréhender de manière fiable, notamment lorsque l'analyse s'appuie principalement sur de la vidéo 2D et sur des jugements perceptifs. Par ailleurs, les typologies mobilisées ne sont pas toujours standardisées (ex. : la description formelle peut fortement varier) et les schémas d'annotation varient fortement d'un projet à l'autre, ce qui limite la comparabilité entre corpus. Du côté des approches computationnelles, l'étude du geste s'inscrit principalement dans une logique de traitement de données en vue de l'entraînement de modèles automatiques. Or, ces modèles, quelle que soit la tâche à laquelle ils sont dédiés, demeurent largement perfectibles. Les gestes générés présentent souvent une faible accroche au discours et une richesse sémantique limitée, tandis que la détection automatique des phases gestuelles comme la classification fonctionnelle du geste reste encore largement sous-performantes. En outre, l'explicabilité des modèles appliqués à la gestualité est rarement interrogée, ce qui ne permet donc pas d'analyser les logiques d'associations motivant la reconnaissance ou la génération de gestes par ces modèles. Si certaines études analysent les données mobilisées lors de l'apprentissage, elles renseignent peu sur la manière dont les gestes sont effectivement interprétés par le modèle dans l'ensemble des dimensions qui les caractérise. Pourtant, dans le cas qui nous occupe, comprendre comment un réseau de neurones profond parvient à représenter la synchronisation entre les gestes et la parole – ou, plus généralement, à reconnaître des patterns complexes dans des données multimodales – est essentiel pour progresser sur cette tâche spécifique et mieux appréhender les mécanismes d'apprentissage sous-jacents. Ces représentations permettent de mieux comprendre comment et pourquoi le modèle regroupe certains gestes dans l'espace latent, ouvrant ainsi de nouvelles perspectives pour l'étude conjointe des gestes et de la parole. Enfin, l'évaluation des gestes générés repose majoritairement sur des études perceptives et ne permet pas de déterminer si les gestes produits sont effectivement porteurs de sens.

Pour dépasser ces limites, cette thèse défend le recours à des approches interdisciplinaires comme une voie privilégiée pour renouveler les méthodologies d'étude des gestes co-verbaux. L'étude de la gestualité en sciences computationnelles gagnerait en effet à une intégration fine des connaissances issues des sciences du langage aux paradigmes issues de l'intelligence artificielle ou méthodes d'automatisation, afin de produire des modèles à la fois crédibles sur le plan comportemental et robuste sur le plan computationnel. Du point de vue des sciences du langage, les approches computationnelles peuvent être envisagées comme des paradigmes d'exploration et de structuration des données, au service d'une compréhension approfondie des relations entre gestes et parole au travers des analyses automatiques qu'elle propose. C'est précisément dans cet espace commun que s'inscrit la motivation de la présente recherche, qui

vise à développer une méthodologie permettant une modélisation cohérente et rigoureuse de la gestualité grâce à des outils computationnels adaptés, qui reste exploitable, qualitative, cohérente et efficace dans le cadre des projets de recherche appliqués à l'étude du geste.

Malgré l'existence d'un ensemble croissant d'outils pour le traitement automatique de la parole, de la vidéo ou du mouvement, ces ressources restent peu mobilisées dans les études du geste en sciences du langage. Cette situation tient à plusieurs facteurs : d'une part une méconnaissance des méthodes disponibles, mais aussi des barrières pratiques liées à l'implémentation, à l'infrastructure nécessaire pour mettre en place de telles technologies. En conséquence, de nombreuses analyses continuent de reposer quasi exclusivement sur des approches manuelles, sans tirer pleinement parti des possibilités offertes par les méthodes computationnelles pour explorer de grands volumes de données, objectiver certaines régularités ou tester systématiquement des hypothèses à large échelle. Une telle couverture empirique permettrait de mieux cerner l'étendue du champ des associations possibles entre gestes et parole, au-delà des régularités déjà bien documentées dans la littérature. Ainsi, les paradigmes d'analyse issus des sciences computationnelles, notamment via l'apprentissage profond, décuplent les possibilités : ces outils ne se limitent plus à traiter ou visualiser des données, mais permettent d'aller plus loin, en révélant des régularités, relations complexes et associations dont la finesse, la variété ou l'abondance rendraient l'appréhension humaine impossible, tout en offrant une formalisation mathématique des phénomènes observés.

De manière réciproque, les performances des modèles dépendent étroitement du corpus utilisé, et par conséquent des données disponibles, du niveau de granularité de ces données, et de leur qualité. Or, les approches computationnelles intègrent encore rarement les descriptions fines et les connaissances produites par les sciences du langage. Les modèles gagneraient pourtant à tirer parti de cette expertise, qui pourrait guider l'apprentissage dans sa prise en compte des liens entre gestes et discours, enrichir la représentation des gestes et améliorer la pertinence des prédictions. En l'absence de cette articulation, les jeux de données restent souvent peu représentatifs de la complexité du geste co-verbal, limitant à la fois la fidélité des gestes générés et la capacité des modèles à saisir leurs dimensions formelles et fonctionnelles.

Bien que l'interdisciplinarité constitue une piste prometteuse, plusieurs difficultés peuvent néanmoins être anticipées. Par exemple, lorsqu'un geste assure simultanément plusieurs fonctions, la classification fonctionnelle automatique du geste ne peut être considérée comme une tâche de reconnaissance de classe unidimensionnelle par entité analysée. Cette réalité pose un défi majeur : dans une approche multilabel, le modèle doit apprendre à distinguer des dimensions fonctionnelles partiellement chevauchantes, ce qui complique à la fois l'apprentissage et la prédiction par rapport à une classification traditionnelle multiclasse. Par ailleurs, même dans ce dernier cadre, des difficultés subsistent : certaines catégories fonctionnelles sont sous-représentées dans les corpus, et certaines séquences ou types de gestes sont rares ou très courts. Ce déséquilibre statistique peut également biaiser l'apprentissage et limiter la performance des modèles pour la reconnaissance de types fonctionnels.

De même, la segmentation des gestes – qui consiste à discrétiser un signal fondamentalement continu, en particulier le mouvement, afin d'identifier et de délimiter les unités gestuelles au sein d'un flux de mouvement plus large – implique de déterminer des frontières temporelles dans un signal qui évolue progressivement et sans ruptures nettes. Or, cette continuité rend l'exercice intrinsèquement délicat, tant pour les annotateurs humains que pour les modèles automatiques. Les transitions entre phases sont souvent graduelles, potentiellement ambiguës ou dépendantes du contexte discursif, ce qui complique l'identification de limites temporelles précises. Pour les modèles d'apprentissage automatique, cette absence de frontières objectives et stables constitue un défi majeur : les prédictions peuvent être décalées de quelques *frames*, fragmenter un même geste en plusieurs segments ou, au contraire, fusionner plusieurs gestes distincts. La segmentation automatique du geste ne relève ainsi pas d'une simple détection de ruptures dans le signal, mais d'une interprétation fine de dynamiques motrices étroitement liées au discours et à la situation communicative.

L'approche interdisciplinaire défendue ici vise donc à faire émerger une compréhension partagée et approfondie de la gestualité co-verbale, à la fois sur les plans théorique, méthodologique et computationnel. Malgré l'existence d'un socle commun entre sciences du langage et approches computationnelles, la modélisation outillée de la gestualité co-verbale ne peut se réduire à une simple juxtaposition de disciplines. Il est crucial d'envisager l'interdisciplinarité dès le point de départ, c'est-à-dire lors de la constitution, de la collecte et de l'enrichissement des corpus, plutôt que de se concentrer uniquement sur le développement de modèles ou le traitement automatique de données non représentatives de ce qu'est réellement un geste. Une approche fondée sur les données permet non seulement d'assurer la qualité, la représentativité et la pertinence des annotations, mais aussi de mieux guider l'apprentissage des modèles en intégrant les dimensions formelles et fonctionnelles du geste. Elle favorise ainsi une exploitation plus fidèle et plus riche des corpus, en plaçant la connaissance experte et l'observation du geste au cœur de la modélisation, plutôt que de réduire le geste à un simple signal exploitable par un algorithme, tout en usant de la puissance et de l'innovation des approches computationnelles pour faire progresser notre compréhension de la gestualité co-verbale humaine.

# Partie II
# Ressources et données multimodales

Cette partie s'intéresse à la construction, à l'enrichissement et à la réutilisation de corpus dédiés à l'étude de la gestualité. Elle vise à répondre à la première question de recherche de ces travaux, en dégageant les éléments constitutifs d'un corpus multimodal qui soit suffisamment fiable et adapté pour l'analyse outillée du geste co-verbal.

Elle commence par un panorama des principaux corpus existants, mettant en évidence leurs caractéristiques et leurs limites. Elle présente ensuite une étude de cas portant sur l'analyse de la qualité de l'enrichissement d'un corpus donné, proposant plusieurs méthodes et outils mobilisables pour réaliser une première vérification de surface. Enfin, elle décrit la conception et l'annotation du corpus d'étude de cette recherche, en détaillant les choix méthodologiques et les stratégies adoptées pour produire des données riches et exploitables.

Le défi central de cette partie consiste à montrer comment concilier une approche réaliste, attentive aux limites pratiques, répondant aux exigences requises pour produire un enrichissement de corpus adapté à l'analyse fine du geste et exploitable dans des cadres computationnels.

# Chapitre 3
# Corpus multimodaux pour l'étude de la gestualité : enjeux et nécessité

## 1 Introduction

L'étude approfondie de la gestualité humaine repose sur l'analyse de données diversifiées, riches et contextualisées. Ces données jouent un rôle central dans la compréhension de la communication humaine pour plusieurs raisons essentielles. Tout d'abord, elles permettent d'explorer, ce au long terme, la variété des gestes dans des contextes sociaux, culturels et linguistiques différentes. En examinant ces occurrences de gestes, les chercheurs peuvent déchiffrer des significations fines, des nuances expressives et des règles implicites qui régissent leur production et leur interprétation. Cette diversité rend possible l'identification de gestes universels, culturellement spécifiques ou encore individualisés, et permet d'analyser la stabilité ou l'évolution des gestes au fil du temps. Ensuite, ces données constituent un appui essentiel pour valider, nuancer ou ajuster les théories existantes sur la gestualité, en offrant des preuves empiriques solides et systématiquement documentées.

L'objectif de ce chapitre est alors de proposer un état des lieux, non exhaustif mais représentatif, des ressources existantes pour l'analyse du geste co-verbal. Dans un premier temps, nous présenterons les principes méthodologiques fondamentaux régissant la constitution d'un corpus, en mettant l'accent sur les spécificités liées à l'étude du geste : choix du dispositif d'enregistrement, modalité de capture, procédures d'enrichissement et exigences en matière de documentation. Dans un second temps, nous dresserons un panorama des principaux corpus multimodaux à disposition, en considérant leur taille, les modalités qu'ils intègrent (vidéo, audio, motion capture, transcriptions alignées, annotations multimodales, etc.) ainsi que les contextes d'enregistrement qui influencent leurs usages possibles. Cette présentation permettra de mettre en lumière la diversité des corpus constitués et les différents paradigmes méthodologiques qui

les sous-tendent. Enfin, nous examinerons de manière critique les usages, les apports et les limites de ces corpus pour les études de la gestualité, mais aussi pour leur exploitation dans des approches computationnelles telles que l'apprentissage automatique et la segmentation automatisée de gestes. Cette discussion soulignera à la fois la richesse de ces ressources et les défis persistants en matière de qualité, de standardisation et de disponibilité des données. En articulant ces trois perspectives, ce chapitre vise à clarifier l'état actuel des ressources pour l'étude du geste co-verbal et à identifier les enjeux méthodologiques et technologiques qui orienteront les développements futurs.

# 2 Modes de constitution d'un corpus de gestes

Un corpus de gestes désigne un ensemble structuré de données multimodales (vidéo, audio, *motion capture*, transcription, etc.) recueillies dans un contexte défini, permettant l'observation, la description et l'analyse des gestes d'un ou de plusieurs locuteurs. Il s'agit d'une ressource organisée qui documente à la fois le geste (mouvement, fonction, forme, etc.) et son environnement communicatif (parole, interaction, situation). L'objectif d'un tel corpus multimodal est de fournir une base empirique pour étudier la production et l'interprétation des gestes : décrire leurs formes et leurs fonctions, ou encore comparer leurs usages selon différents contextes sociaux, culturels ou linguistiques, tels que l'acquisition du langage, la pédagogie ou diverses situations émotionnelles. Pour atteindre ces objectifs, il est indispensable de disposer d'un volume suffisant de gestes, produits en conséquence du protocole défini pour la collecte des données. La quantité de gestes disponibles conditionne en effet la robustesse des analyses empiriques et la pertinence des conclusions, et dépend pour beaucoup du dispositif mis en place, notamment du type de tâche ou de stimulus, des conditions d'enregistrement, ou encore des instructions données aux participants. Ainsi, la constitution d'un corpus exploitable repose à la fois sur un protocole de collecte capable de susciter une gestualité riche et variée, ou au contraire ciblée sur un type de geste donné, et sur des enregistrements suffisamment nombreux pour permettre des analyses fiables et généralisables.

## 2.1 Situations de communication suscitées par le protocole

Les situations de communication, déterminées par le contexte, les intentions communicatives et les participants impliqués, façonnent d'abord la production verbale, dont les contraintes et les dynamiques orientent ensuite la gestualité co-verbale associée. Ainsi les corpus multimodaux peuvent-ils être constitués à partir de tâches très diverses proposées aux participants, qui façonnent la nature, la fréquence et la richesse des gestes produits. L'analyse de la communication, et en particulier de la gestualité co-verbale, doit donc impérativement prendre en compte ces variations contextuelles, afin d'interpréter correctement les comportements observés et d'éviter des conclusions biaisées par le protocole ou par la situation de collecte.

### 2.1.1 Monologue et dialogue

Un monologue correspond à l'intervention d'un seul locuteur généralement en l'absence d'un interlocuteur (ex. : un monologue théâtral) ou dans un cadre où l'interlocuteur ne doit pas répondre (ex. : conférence) ou ne peut pas intervenir (ex. : allocution télévisée). Si cette intervention est préparée, la gestualité produite peut être travaillée en amont, et potentiellement donner à voir une plus grande proportion de gestes destinés à structurer le discours plutôt que d'autres types de gestes (ex. : gestes servant à réguler une interaction). À l'inverse, un dialogue désigne un échange impliquant au moins deux interlocuteurs, qui alternent les rôles de locuteur et d'auditeur dans une dynamique d'action et de réaction (ex. : conversations informelles, entretiens, débats, etc.). Dans ce contexte, les gestes sont par définition plus spontanés (donc plus variés) et adaptatifs. En comparaison avec la situation de monologue, les gestes d'une interaction sont potentiellement moins liés à la structuration du discours et davantage délégués à la régulation de l'interaction en cours.

### 2.1.2 Spontanée, dirigé, acté

Les situations spontanées correspondent à des situations de discours naturelles, lors desquelles théoriquement la gestuelle est elle-même spontanée, c'est-à-dire qu'elle est produite sans l'influence d'indications préalables sur la manière dont les participants devraient se mouvoir ou s'exprimer. Les locuteurs évoluent librement dans l'environnement, selon leurs habitudes et leur style personnel, ce qui favorise l'émergence de gestes authentiques. Les situations dirigées, en revanche, s'inscrivent dans un cadre contrôlé : des consignes précises sont données aux participants pour orienter leurs productions verbales ou gestuelles, tout en leur laissant une certaine marge de liberté. Par exemple, on peut leur demander d'exprimer une émotion spécifique, de réaliser une tâche spécifique, de rester assis sur une chaise ou de ne pas dépasser une zone définie pour la captation, etc. Ces contraintes permettent de standardiser certaines conditions tout en conservant une part de naturalité dans les paroles et gestes observés, tant qu'aucune consigne n'est donnée les concernant spécifiquement. Enfin, les situations actées constituent une forme particulière de situation dirigée, où l'énonciation du locuteur est volontairement mise en scène, comme au théâtre ou au cinéma. Dans le cadre d'un corpus acté multimodal, il peut être demandé aux locuteurs d'amplifier leurs gestes, de les styliser, ou encore de les répéter. Le locuteur ne produit plus un contenu spontané, mais joue un rôle dans lequel sa production verbale et sa gestuelle sont orchestrées selon des objectifs précis. Cette distinction entre situations spontanée, dirigée et actée est essentielle, car elle conditionne donc le degré d'authenticité des gestes et influence la manière dont ils peuvent être interprétés et utilisés dans des analyses linguistiques ou multimodales.

### 2.1.3 Activité libre, élaboration discursive, coopération

L'activité libre correspond à un monologue ou une interaction menés sans consignes particulières, dans un cadre aussi spontané que possible. Le ou les participants orientent eux-mêmes leur production discursive, ce qui permet, dans le cadre d'un corps multimodal, de recueillir des gestes variés et naturels, reflétant les habitudes et styles individuels. Cependant,

ce type de tâche comporte certains risques, notamment le partage d'informations personnelles jugées sensibles (orientation sexuelle, croyances religieuses, etc.). De plus, ce type d'activité libre peut également ne pas générer les éléments ciblés en nombre suffisant, ce qui limite alors parfois la quantité de données exploitables pour certaines analyses. Dans le cas de l'élaboration discursive, les participants effectuent cette fois une tâche verbale relativement cadrée, telle que raconter une histoire, décrire une image ou reformuler une information. Dans une visée d'étude multimodale, ce type de situation favorise la production de gestes étroitement liés à un type de discours donné, tout en assurant une meilleure comparabilité des données entre participants, ce qui est particulièrement utile pour l'analyse systématique de la gestualité. Enfin, les tâches de coopération, qui constituent une forme particulière d'élaboration discursive, impliquent la participation conjointe de deux personnes ou plus dans des activités collaboratives comme la résolution de problèmes, des jeux collectifs ou la construction d'objets. Dans une perspective multimodale, ces situations favorisent l'apparition de gestes interactifs, orientés vers la coordination, la négociation et la régulation sociale, et permettent d'observer des comportements gestuels émergents dans des contextes d'échange et de co-construction. Il est possible de combiner plusieurs de ces paramètres lors d'une captation afin de diversifier les situations de production verbale et gestuelle, et d'ainsi constituer un corpus à la fois représentatif et varié.

## 2.2 Types de captation

Diverses méthodes d'acquisition peuvent être envisagées, chacune présentant là encore des caractéristiques propres qui influencent de manière significative la qualité des données, mais surtout la manière de les traiter. Elles impactent notamment la clarté de la parole, la précision des mouvements captés, le niveau de bruit dans les enregistrements et plus largement le degré de contrôle et d'influence exercé sur la situation de communication.

### 2.2.1 Extraction automatique

L'extraction automatique consiste à collecter automatiquement des données à partir de contenus généralement disponibles en ligne. Elle présente généralement l'avantage de permettre l'accès à de grands volumes de données, pouvant aller au-delà de 100 heures de parole dans le cas de corpus linguistiques oraux. Toutefois, cette abondance s'accompagne de certaines limitations. Les enregistrements collectés, par exemple sur des plateformes de diffusion ou dans des archives médiatiques, peuvent être de qualité variable et ne reflètent pas toujours des situations de communication authentiques. De plus, le montage vidéo peut altérer la continuité des interactions. Ces facteurs peuvent affecter la fiabilité des analyses, en particulier lorsque les outils utilisés sont sensibles à la qualité des données. En définitive, le principal inconvénient de l'extraction automatique réside dans le manque de contrôle sur les conditions de production des données, susceptibles d'influencer de manière significative le temps nécessaire de prétraitement des données, mais surtout la représentativité et la robustesse des résultats obtenus.

### 2.2.2 Environnement contrôlé

Les ensembles de données collectés dans des environnements « contrôlés » sont produits et enregistrés en laboratoires ou dans des salles spécialisées. Cette condition d'enregistrement implique généralement des équipements et matériels sophistiqués : systèmes de capture de mouvement, microphones haute qualité pour la parole, caméras à haute résolution pour enregistrer tous les aspects de la situation, etc. Cependant, la production de données en environnement contrôlé reste limitée par le temps et les ressources : chaque mise en place de salle, recrutement de participants, enregistrement et traitement postérieur est chronophage, ce qui peut rendre difficile la constitution de très grands corpus. Un autre facteur important est le biais introduit par le cadre contrôlé : être dans un laboratoire, mais plus encore le simple fait de se savoir observé et enregistré peut engendrer du stress ou de l'anxiété, ou encore accentuer certains traits comme la timidité, rendant les participants plus conscients de leurs gestes et de leur parole, et donc moins naturels. Cet effet peut en outre être particulièrement amplifié lorsque les locuteurs et les auditeurs ne se connaissent pas.

### 2.2.3 Situation authentique et naturelle

Cette approche désigne un scénario dans lequel les participants sont observés dans leur environnement naturel, au cours d'une activité libre, sans consignes ni directives particulières. L'objectif de ce cadre de captation est de pouvoir observer des comportements du quotidien, sans altérer les routines individuelles. Cette méthode vise ainsi à collecter des données reflétant fidèlement les usages et expériences authentiques des individus, en minimisant les biais liés à un environnement artificiel ou contrôlé. Cependant, certaines contraintes matérielles et logistiques limitent souvent la mise en place de ce type d'observation. Ainsi, la qualité technique de ces collectes peut s'avérer insuffisante, ne serait-ce qu'en raison de problèmes de cadrage, de prise de son ou de conditions d'enregistrement. De plus, là encore, la conscience de la présence d'un matériel d'enregistrement, tout comme le fait de se savoir observé, peut tout de même influencer les comportements des participants.

# 3 Mise en œuvre du protocole d'enrichissement

L'enrichissement d'un corpus consiste à ajouter des données supplémentaires (parfois nommées « données secondaires ») aux données collectées (« données primaires »). Il peut s'agir d'éléments écrits (ex. : le texte d'une séquence audio énoncée par le locuteur, des descriptions ou des catégorisations), joints dans des fichiers de différents formats (ex. : au format `.txt`, `.eaf` ou `.TextGrid`). Dans le cadre d'une étude multimodale, l'enrichissement de corpus implique souvent la réalisation de transcriptions verbales (transposition du signal audio en écrit) et d'annotations variées (catégorisations verbales [formes lemmatiques, relations de dépendances...] ou gestuelles [description, fonctions...]). Ces enrichissements peuvent être réalisés manuellement ou de manière automatisée.

## 3.1 Types d'annotateurs

Selon les corpus, les enrichissements réalisés manuellement sont le fruit d'annotateurs plus ou moins experts. Lorsqu'on évoque un annotateur « expert », on fait généralement référence à une personne issue du champ d'études concerné (en lien avec la nature des spécificités à annoter), familiarisée avec les définitions mobilisées, les conventions d'annotation et les outils spécifiques au domaine (comme les logiciels ELAN [15], Anvil [132]ou Praat [16] dans le cas de corpus multimodaux). Ce profil est particulièrement recherché, car il garantit un haut niveau de fiabilité des annotations réalisées, et de là, de l'analyse des phénomènes annotés. Dans le cas de corpus multimodaux dédiés à l'analyse gestuelle, cette fiabilité a notamment trait à la segmentation, la classification et l'interprétation fonctionnelle ou formelle des gestes annotés.

Cependant, dans la réalité des projets académiques, les annotateurs experts sont rares, car ils impliquent une expertise particulière, et coûteux, car la tâche est chronophage (elle peut aller au-delà d'une heure d'annotation par minute de corpus, selon le type d'enrichissement requis). Dès lors, le manque de temps, la complexité des tâches à effectuer et les contraintes budgétaires rendent difficile leur mobilisation à grande échelle. Lorsqu'un financement le permet, il peut être envisagé d'employer une personne spécifiquement responsable de cette tâche, mais il reste rare qu'un projet mobilise plus d'un ou deux annotateurs qui y soient dédiés. Dans tous les cas, une phase de formation reste indispensable pour assurer une compréhension précise des objectifs, des méthodes et des outils utilisés dans le cadre du projet de recherche. Il convient également de souligner que, malgré cette préparation, occuper exclusivement un poste d'annotateur n'est pas toujours confortable sur le long terme, en raison de la nature exigeante et relativement répétitive de cette tâche.

Face à ces difficultés, de plus en plus de recherches s'orientent vers des annotateurs dits « non-experts », c'est-à-dire des personnes extérieures au domaine, souvent recrutées sur des plateformes participatives comme *Prolific* ou *Amazon Mechanical Turk*. Ces plateformes permettent un recrutement rapide, à coûts restreints, et offrent une certaine souplesse dans la gestion des données à annoter. La rémunération constitue une motivation immédiate pour les participants, et les chercheurs peuvent ainsi récolter un volume important de données en peu de temps. Cependant, cette approche soulève plusieurs limites. De fait, il n'est pas garanti que les participants comprennent les consignes, même lorsqu'elles sont détaillées, ce qui est d'autant plus problématique dès lors qu'il s'agit de réaliser des tâches relativement complexes, comme l'est typiquement l'annotation des gestes, qui nécessite diverses compétences (observation, catégorisation, cohérence temporelle) difficiles à mobiliser sans réelle formation ou expérience antérieure. En l'absence de synchronisation ou de calibration entre annotateurs, les résultats peuvent en outre s'avérer hétérogènes, voire inutilisables sans une phase de vérification systématique par des personnes plus expertes. Cela pose donc la question de l'équilibre entre quantité (accès facilité à de grandes masses de données produites) et qualité (fiabilité et cohérence des annotations), équilibre particulièrement délicat à atteindre dans des contextes à ressources limitées.

## 3.2 Protocole d'enrichissement

Pour assurer la cohérence d'un jeu d'annotations, le protocole d'enrichissement d'un corpus peut reposer sur plusieurs conditions [14].

### 3.2.1 La rédaction d'un manuel de codage

Un manuel de codage poursuit deux objectifs principaux. D'une part, il sert à documenter les pratiques pour les annotateurs eux-mêmes, afin de leur fournir des directives précises sur la manière de procéder. D'autre part, il vise à fournir à la communauté scientifique ou aux futurs utilisateurs du corpus une description détaillée des pratiques employées, permettant de comprendre et de réutiliser les données de manière informée. Il formalise ainsi les objectifs de l'enrichissement, les unités d'analyse pertinentes (ex. : *frame*, segment, tour de parole, etc.), les catégories d'annotation et les listes d'étiquettes utilisées, les critères de décision, les cas limites ou ambigus, ainsi que les outils techniques mobilisés. Ce manuel n'est pas figé : il peut évoluer au fur et à mesure de l'avancement du projet, en fonction des difficultés rencontrées, des cas non prévus initialement ou des ajustements nécessaires pour renforcer la cohérence inter-annotateurs. Il est conçu comme un outil dynamique, mis à jour régulièrement et partagé avec l'ensemble des annotateurs. Cette démarche est couramment adoptée dans la plupart des projets de recherche en sciences du langage.

### 3.2.2 La phase de synchronisation (calibration) des annotateurs

L'étape de synchronisation consiste à former les annotateurs, qu'ils soient experts ou non, et à les faire travailler sur un sous-ensemble commun du corpus lors de sessions de travail partagées. L'objectif est de s'assurer qu'ils partagent une compréhension uniforme des consignes et appliquent les critères d'annotation de manière cohérente. Cette phase permet également d'identifier de potentielles ambiguïtés du protocole, d'ajuster les définitions et d'anticiper les difficultés méthodologiques.

### 3.2.3 Validation des annotations

Une fois les annotations réalisées, l'étape de validation permet d'évaluer leur fiabilité et leur cohérence. Cette validation peut prendre plusieurs formes : comparaison entre annotateurs (mesure de l'accord inter-annotateur avec le kappa de Cohen par exemple), vérification par un annotateur indépendant (sur la base du manuel de codage), ou encore utilisation de protocoles statistiques pour détecter les anomalies ou incohérences. L'objectif de cette phase est de s'assurer que les annotations respectent bien les critères définis lors de la phase de calibration, et qu'elles peuvent servir de base solide pour des analyses ultérieures. Dans le cas de corpus destinés à l'apprentissage automatique ou à la modélisation gestuelle, cette étape permet également d'identifier des éléments problématiques au sein des données (ex. : problèmes de format de données ou segments ayant une moindre qualité) qui pourraient nuire aux performances des modèles.

Manuel de codage, phase de calibration et validation constituent donc les trois piliers méthodologiques garantissant la valeur scientifique d'un enrichissement de corpus et sa fiabilité pour une exploitation par la communauté.

# 4 Variétés de formats d'enrichissement d'un corpus multimodal

Les fichiers générés lors de l'enrichissement d'un corpus multimodal peuvent adopter différents formats, tels que `.txt`, `.eaf`, `.TextGrid`, `.csv`, etc. afin d'être exploités par la suite. Le choix du format dépend de la nature de l'information ajoutée, de la tâche d'enrichissement réalisée, ainsi que de la nature des traitements envisagés.

## 4.1 Transcription et annotation de la parole

La transcription de la parole consiste à convertir le signal audio en une représentation écrite du discours. Elle constitue l'une des étapes fondamentales de l'enrichissement de corpus multimodaux, notamment lorsqu'il s'agit d'étudier l'articulation entre le verbal et d'autres modalités, telles que la prosodie ou la gestualité. La forme la plus courante est la transcription orthographique, qui vise à représenter de manière standardisée les mots prononcés par les locuteurs, en respectant toutefois souvent, dans la mesure du possible, leur élocution réelle (hésitations, reformulations, marques d'oralité, etc.). Le degré de détail de la transcription peut alors varier selon les objectifs de recherche. Dans certains cas, une transcription des mots est suffisante ; dans d'autres, la transcription peut être enrichie par des annotations prosodiques (intonation, accentuation, pauses), ou encore une segmentation fine des unités syntaxiques, lexicales ou pragmatiques, etc. Il devient alors possible d'étudier le discours à différents niveaux de granularité : la séquence entière, la phrase, le mot, la syllabe ou encore le phonème, chaque unité annotée offrant un angle d'analyse spécifique. Par exemple, l'analyse morphosyntaxique repose sur une segmentation au niveau du mot, tandis que des études prosodiques peuvent nécessiter une distinction syllabique ou phonémique. La transcription de la parole peut être alignée au flux audio, et dès lors être mobilisée comme point d'ancrage temporel pour l'annotation des autres modalités. Le geste peut ainsi être temporellement localisé au sein du flux verbal et dès lors être mis en relation avec le discours concomitant (affiliation temporelle de l'exécution du geste). Sur cette base, il est alors possible de caractériser, et donc d'analyser, les différents modes de synchronicité à l'œuvre dans le flux gestuel et discursif produit par le locuteur.

La qualité de la transcription verbale impacte donc directement la précision, la fiabilité et la pertinence des enrichissements apportés au corpus, en particulier lorsque ces enrichissements reposent sur des alignements temporels fins. La qualité de la transcription repose également fortement sur le respect rigoureux de l'orthographe, dont l'exactitude est indispensable aux analyses lexicales et sémantiques. En effet, dans le cadre d'analyses automatiques outillées portant sur de relativement grands ensembles de données, la standardisation des graphies constitue un prérequis méthodologique : à défaut, les outils ne peuvent reconnaître de manière fiable les unités lexicales.

## 4.2 Segmentation des gestes

La segmentation des gestes repose souvent sur un découpage en différentes phases, selon le modèle proposé par Kendon [7, 8] présenté au chapitre 1. En pratique, il est relativement rare d'annoter l'ensemble des phases gestuelles, en raison de la charge temporelle de travail que cela implique. Le choix d'annoter une ou plusieurs d'entre elles dépend des objectifs spécifiques de l'étude. Un compromis fréquent consiste à annoter uniquement l'unité gestuelle complète (segment compris entre deux positions de repos et regroupant possiblement plusieurs gestes successifs), ou la phrase gestuelle (ensemble de l'élaboration d'un geste sans distinction des phases qui le composent), ou le *stroke* seulement (phase « noyau » d'un geste). Ce dernier est l'élément le plus couramment annoté ; en se concentrant sur cette phase, les annotateurs peuvent en effet capturer l'essentiel du contenu gestuel tout en optimisant le temps d'annotation.

La segmentation gestuelle peut être réalisée selon différents procédés, en fonction du support utilisé. Pour une segmentation à partir d'une vidéo RGB, la tâche repose sur l'observation attentive des mouvements corporels de la séquence vidéo afin de repérer les indices délimitant le début ou la fin des gestes, tels que des changements de posture, des accélérations, des pauses ou des variations dans la direction du mouvement [133, 134]. Ces repères sont interprétés comme des indices perceptibles signalant une transition d'une phase gestuelle à une autre, ou la limite d'une unité gestuelle plus large selon le type d'annotation choisi. Cette méthode reste cependant sujette à variabilité, notamment dans les cas ambigus en termes de mouvement, par exemple lorsque les gestes sont très fluides. Elle requiert donc une bonne familiarité avec les critères de segmentation et, idéalement, un consensus préétabli de ces critères dans le cadre d'une compréhension commune du protocole de la part des annotateurs.

Pour les données tridimensionnelles de type *mocap*, le principe de segmentation reste similaire à celui appliqué à la vidéo RGB, bien que les indices perceptibles diffèrent sensiblement. À partir des positions enregistrées, il est possible de reconstruire une séquence sous la forme d'un avatar animé ou de générer une vidéo de synthèse (de face ou selon un angle choisi). L'annotateur dispose alors d'un point de vue contrôlable (orientation, zoom sans perte de résolution), ce qui peut faciliter l'identification des transitions gestuelles reconstituées numériquement, en particulier pour des gestes rapides ou subtils. Cependant, l'absence de certaines informations d'origine, comme les expressions faciales (pas forcément reconstituées numériquement) ou la corpulence du locuteur (effacée de la reconstruction numérique) peut limiter certaines interprétations. Comme pour les vidéos, la qualité de l'enregistrement reste dans tous les cas un facteur déterminant : une captation bruitée (variations indésirables ou erronés des données qui ne reflètent pas le mouvement réel), partielle ou imprécise peut rendre difficile l'identification fiable des phases gestuelles.

## 4.3 Catégorisations formelles et fonctionnelles des gestes

Catégoriser des gestes consiste à leur attribuer des étiquettes catégorielles et discrètes à partir de la segmentation réalisée. Cette catégorisation peut être fonctionnelle, et dès lors renvoyer au rôle communicatif des gestes, en lien avec leur affiliation lexicale ou prosodique, ou

formelle, et détailler leurs caractéristiques visuelles et motrices. S'agissant de l'annotation fonctionnelle, l'annotateur interprète le geste dans son contexte de production afin de déterminer son intention communicative, en se basant sur les étiquettes et définitions incluses dans le manuel de codage. Ces annotations exigent une bonne maîtrise du protocole et des critères de catégorisations et requièrent généralement des ajustements lors de la phase de calibration des annotateurs. S'agissant de l'annotation formelle, les gestes sont étiquetés selon leurs dimensions motrices et posturales telles que la configuration de la main, la direction et la trajectoire du mouvement réalisé, l'amplitude, ou la vitesse. Ces informations fonctionnelles et formelles permettent par la suite de mener une analyse descriptive des gestes et constituent une base pour des études quantitatives ou pour l'alimentation de modèles de détection automatique.

La qualité des annotations dépend ainsi à la fois de la précision de la segmentation et de la transcription, de la clarté des catégories et des instructions données, ainsi que de la compréhension du protocole et des spécificités de l'objet à annoter par les annotateurs. Elle requiert également une cohérence intermodale et temporelle, notamment par l'alignement des gestes avec le discours et, le cas échéant, avec d'autres modalités comme le regard ou l'expression faciale. De nombreux facteurs peuvent affecter cette qualité, ce qui plaide pour une rigueur méthodologique pour produire des corpus fiables et exploitables, que ce soit pour l'analyse descriptive ou la modélisation computationnelle, tout en s'appuyant sur des outils d'enrichissement maîtrisés.

# 5 Outils d'enrichissement de corpus

En sciences du langage, l'étude de la communication multimodale s'appuie sur des interfaces spécialisées pour l'enrichissement manuel des corpus. On retrouve ainsi une certaine diversité d'outils et de fonctionnalités au sein de logiciels tels qu'ELAN [15], Praat [16], ANVIL [132], CLAN [135] ou encore SPPAS [136] ; la liste de fonctions présentée dans cette section n'a toutefois pas vocation à être exhaustive, l'objectif étant plutôt de présenter certaines fonctionnalités particulièrement utiles dans le cadre de l'enrichissement de corpus multimodaux.

- La ligne d'annotation temporelle (*tiers* ou *acteurs)* est un outil qui permet de délimiter dans le flux de parole plusieurs segments temporels ciblés correspondant à des éléments relevant d'une même catégorie (mots, syllabes, présence d'un geste, fonction d'un geste, forme de la main, etc.). Cette fonctionnalité, présente dans la plupart des logiciels d'annotation, est particulièrement utile car elle permet de partitionner le discours en différents niveaux d'informations (verbales, prosodiques, gestuelles), tout en révélant leur articulation. Ainsi la délimitation temporelle des mots énoncés par un locuteur, reportée par exemple sur un *tier* « words », est-elle reliée à la délimitation temporelle des gestes réalisés par le locuteur, dont la présence peut être consignée sur un *tier* « gesture ». Ces *tiers* peuvent donc être multiples et être hiérarchisés selon un schéma d'annotation particulier (relation de dépendance « parent-enfant » entre certaines lignes d'annotation dans le cas d'ELAN). Le contenu d'une ligne d'annotation peut provenir d'un enrichissement manuel ou automatique, selon la nature des éléments qui y sont annotés.

- La création de lexiques (le *vocabulaire contrôlé* dans ELAN par exemple) permet de définir des familles d'étiquettes. Ces lexiques peuvent être associés à un *tier* spécifique et présentent plusieurs avantages : ils limitent les choix, au sein d'un *tier*, aux seules catégories prédéfinies (par exemple, une liste de fonctions possibles des gestes, ou une liste de phénomènes prosodiques observable sur une syllabe, etc.). Ils évitent donc la saisie manuelle de catégorisations, et par conséquent les erreurs typographiques. Enfin, les lexiques permettent de réduire le temps nécessaire à l'annotation et la vérification des saisies. Ils contribuent également à la structuration et à la cohérence du schéma d'annotation.

- La visualisation multimodale des données à traiter constitue une fonctionnalité commune à la majorité de ces logiciels d'annotations, bien que les types de modalités proposées diffèrent selon les outils : Praat, par exemple, permet de visualiser l'enrichissement réalisé dans des *tiers* de manière synchronisée avec une représentation spectrographique du signal audio, modalités auquel ELAN couple également la vidéo, et ANVIL, la visualisation des données tridimensionnelles.

- Les fonctions d'import et d'export de fichiers, relativement répandues, permettent d'exploiter des formats spécifiques à certains logiciels dans d'autres outils (ex. : l'import de fichiers `.TextGrid` dans des logiciels autres que Praat) ou d'exporter les enrichissements réalisés vers des formats plus génériques (ex. : `.eaf` vers `.csv` dans le cas d'ELAN), ce qui facilite l'interopérabilité entre différents outils.

Ensemble, ces différents outils constituent un socle technique essentiel pour la description, l'analyse et la modélisation de la communication multimodale. Leur usage reste cependant majoritairement manuel, rendant l'enrichissement des corpus particulièrement chronophage, surtout dans le cadre de schémas d'annotation fins, donc complexes. Des fonctionnalités automatiques d'enrichissement et d'analyse peuvent être mobilisées, soit directement au sein des logiciels dédiés aux sciences du langage, soit à l'aide d'outils informatiques externes, qui ne sont pas nécessairement intégrés aux environnements classiques d'annotation.

- Les systèmes de reconnaissance automatique de la parole (ASR) facilitent la transcription orthographique. Bien qu'ils ne soient pas directement intégrés aux logiciels classiques d'annotation en sciences du langage, ils peuvent être mobilisés comme outils complémentaires pour générer automatiquement une transcription à partir d'un signal audio.

- La segmentation et l'alignement temporel semi-automatiques du contenu verbal peuvent être réalisés à l'aide de modules tels qu'EasyAlign (dans Praat), SPPAS, ou encore *Montreal Forced Aligner* (MFA) [137], ce dernier étant largement utilisé en informatique. Ces outils permettent de segmenter automatiquement le flux audio en différentes unités textuelles, à des niveaux de granularité variables (du mot au phonème) tout en les alignant temporellement avec le signal audio. Ces procédures sont qualifiées de semi-automatiques dans la mesure où elles nécessitent généralement une vérification et parfois une correction humaine.

- Différentes analyses automatiques de données sont aussi disponibles, la plupart offrant des statistiques descriptives classiques (nombre d'occurrences, fréquences, durées, etc.) et d'autres proposant des fonctionnalités plus spécifiques, comme l'analyse du signal acoustique (intensité, hauteur de la voix, etc. dans Praat) ou encore l'analyse morphosyntaxique (détection automatique des *tokens* de mots dans CLAN). Les modèles de langage et bibliothèques informatiques issus du traitement automatique du langage naturel (NLP) permettent d'analyser de grands corpus textuels, comme *spaCy* [138], qui offre une annotation verbale automatique avec catégorisation grammaticale (*part-of-speech*) et analyse des dépendances syntaxiques.

Bien que ces outils semi-automatiques offrent un gain de temps significatif par rapport à l'enrichissement manuel, leur utilisation requiert souvent des compétences techniques spécifiques. C'est particulièrement le cas des outils issus du domaine informatique. Les langages de programmation permettent de concevoir des *pipelines* d'analyse et d'extraction de données personnalisés, tout en facilitant la réutilisation d'outils, de bibliothèques et de modules existants. Les compétences et les ressources matérielles requises peuvent limiter leur mobilisation dans les études en sciences du langage, où ces outils restent peu accessibles aux linguistes ou aux annotateurs.

# 6 Partage et science ouverte dans le domaine de la gestualité

## 6.1 Disponibilité des données et risques

La diffusion des corpus multimodaux s'effectue généralement sur des plateformes spécialisées dans le partage de ressources linguistiques et multimodales, comme *Ortolang* [139] pour les corpus en français. Ces plateformes offrent des infrastructures permettant de gérer l'accès aux données, de documenter les métadonnées et d'assurer la traçabilité des usages. Les corpus peuvent également être directement mis à disposition par les laboratoires et équipes de recherche responsables de leur constitution sur des espaces dédiés. La diffusion de corpus multimodaux soulève des enjeux particuliers liés à la sensibilité des données : les enregistrements peuvent contenir des informations personnelles ou identifiables, typiquement des visages, ou divers éléments contextuels révélant l'identité des participants. L'accès est donc souvent limité aux chercheurs agréés, soumis à des conventions de confidentialité, ou les données sont anonymisées conformément aux exigences du RGPD en Europe et aux règles éthiques en vigueur. Cette gestion des données sensibles impose un compromis entre ouverture scientifique et protection des participants, limitant parfois la disponibilité complète des corpus.

## 6.2 Corpus disponibles

La liste des corpus disponibles présentée dans cette section n'est pas exhaustive : seuls certains corpus, sélectionnés pour leur pertinence dans une perspective d'exploitation interdisciplinaire, seront présentés afin d'illustrer les types d'ensembles de données disponibles dans le cadre d'une analyse des gestes. L'attention a été portée sur l'ensemble des données

primaires collectées (vidéo, audio, ou *mocap*) et des données secondaires d'enrichissement (transcriptions et annotations verbales ou gestuelles) fournies par les auteurs de ces corpus, en ce qui concerne à la fois leur richesse et leur qualité. Les corpus présentés ont été examinés attentivement, de manière à tirer un enseignement de leurs avantages et de leurs limites.

### 6.2.1 PACO

Le corpus PACO [140] est constitué de données conversationnelles spontanées issues de 15 interactions dyadiques en face à face, totalisant plus de 5 heures d'enregistrement (chaque interaction durant environ 20 minutes). Il a été créé pour étudier l'impact du manque de terrain d'entente personnel sur la collaboration des participants et, plus particulièrement, sur leur sourire lors des transitions thématiques. Les participants sont 30 étudiants francophones âgés de 20 à 35 ans. Chaque locuteur était équipé d'un micro-casque (Sennheiser HSP4 EW, 48 kHz) et filmé par deux caméras frontales (Canon XF105). Une transcription orthographique enrichie (TOE) a été réalisée manuellement sur 10 des 15 dialogues, incluant les phénomènes oraux typiques (pauses remplies, faux départs, etc.). Le corpus se distingue par un protocole d'annotation semi-automatique du sourire, mais il ne comporte pas d'annotations gestuelles détaillées ni de données de capture de mouvement.

### 6.2.2 CID

Le *Corpus of Interactional Data* (CID) [141] est un corpus audiovisuel français de conversations dyadiques. Ce corpus comprend 8 heures de données, réparties sur 8 interactions distinctes, enregistrées en deux séries, en 2003 et 2005. Il rassemble 16 participants francophones, natifs, âgés d'environ 25 à 40 ans. Un trait distinctif et essentiel du CID est que les participants étaient familiers les uns des autres. Les interactions sont toutes non mixtes (5 dyades Femme-Femme ; 3 dyades Homme-Homme). Le protocole d'enregistrement, effectué en chambre anéchoïque au LPL à Aix-en-Provence, reposait sur un protocole dans lequel les participants étaient invités à se raconter soit des conflits professionnels (pour la première série), soit des histoires insolites (pour la seconde série). Le CID est hautement enrichi, intégrant un découpage automatique en Unités Inter-Pausales (IPU) et une transcription orthographique enrichie du signal de parole. Il bénéficie d'annotations nombreuses et variées couvrant plusieurs niveaux linguistiques (phonétique, morphosyntaxique, prosodique, discursif, gestuel) et conversationnels (*feedbacks*, disfluences, rires, contours intonatifs) qui ont généré de nombreuses publications. Malheureusement, ce corpus ne contient pas de données de *mocap*, et la qualité des vidéos RGB est insuffisante (images et mouvements souvent flous du fait de la faible résolution).

### 6.2.3 TSG

Le corpus *Trinity Speech-Gesture* (TSG) [142] se déclinent en deux versions : TSG-I et TSG-II. Ces ensembles de données, diffusés sous licence, se concentrent sur les mouvements d'un seul individu, un acteur anglophone natif. Le protocole expérimental est similaire dans les deux cas : le locuteur s'exprime librement sur une variété de sujets, en se déplaçant dans un

espace délimité, mais faiblement contraint. L'interaction est minimale, l'interlocuteur en arrière-plan se limitant à fournir un retour visuel par des hochements de tête. La première version comprend environ 4 heures d'enregistrements réparties en 23 séquences, avec des données audio (44 kHz) et de la *mocap* (≈ 60 fps, 53 marqueurs). Toutefois, bien que les doigts sont inclus dans la capture, la qualité des données les concernant est jugée insuffisante, car trop bruitée, et rend leur exploitation quasiment impossible [143, 131]. Dans la seconde version, les auteurs proposent une captation améliorée des mouvements des doigts, avec l'ajout d'environ 6 heures de données supplémentaires. Cependant, aucune annotation gestuelle n'est proposée dans ces corpus.

### 6.2.4 SaGA

Le corpus audiovisuel *Bielefeld Speech and Gesture Alignment Corpus* (SaGA) [144] a été conçu pour l'étude de l'interaction entre la parole et les gestes co-verbaux, dans une perspective interdisciplinaire, combinant des objectifs de recherche linguistique théorique et de modélisation computationnelle, notamment pour la simulation du comportement communicatif multimodal chez des agents virtuels ou des robots. L'étude a utilisé un environnement de réalité virtuelle (*SaGA town*) comme contexte expérimental. Vingt-cinq dyades ont participé à l'expérience, chacune comprenant un *Route-Giver* (donneur d'itinéraire) qui décrivait un parcours et les points d'intérêt associés à un *Follower* (suiveur). Le corpus comprend 280 minutes de matériel vidéo et contient environ 6000 gestes et 39 435 mots. L'enrichissement du corpus, documenté dans un manuel de codage, est effectué de manière systématique et précise, couvrant à la fois les éléments verbaux et les gestes. Seule la phase de *stroke* a été segmentée par les annotateurs. La catégorisation fonctionnelle des gestes se concentre seulement sur les gestes iconiques et les classe selon huit modes de représentation distincts (tels que *Indexing*, *Shaping* ou *Drawing*). Elle inclut également une annotation formelle fine des gestes (forme de la main, position du poignet, orientation de la paume et du dos de la main, et trajectoire du mouvement). Ce corpus présente un intérêt certain pour l'étude des interactions et des gestes, mais il ne couvre qu'un sous-ensemble limité de catégories fonctionnelles de gestes. De plus, il se distingue par le fait d'être entièrement en allemand.

### 6.2.5 MPIEMBM

Le jeu de données *MPI Emotional Body Motion* (MPIEMBM) [145] est un corpus axé sur la reconnaissance des émotions corporelles. Les locuteurs, des acteurs anglophones, produisent des discours actés à partir de scripts, auxquels ils ajoutent librement les nuances émotionnelles qu'ils souhaitent. La quantité de données est limitée, avec une durée totale d'environ 1,5 heure pour 8 participants, chaque locuteur étant enregistré entre 10 et 15 minutes. Le corpus fournit les fichiers de *mocap* (120 fps) avec 22 marqueurs et les transcriptions verbales, mais les audios associés ne sont pas disponibles. En outre, si ce corpus offre un enrichissement centré sur les émotions et une segmentation par séquence de mouvement, il ne propose pas de segmentation gestuelle au sens strict.

### 6.2.6 IEMOCAP

La base de données *Interactive Emotional dyadic MOtion CAPture* (IEMOCAP) [146] est un ensemble de 151 dialogues enregistrés privilégiant l'analyse de l'expression émotionnelle. Ce corpus comprend environ 12 heures de données pour 10 locuteurs, en anglais. Il se compose de sessions dyadiques au cours desquelles des acteurs réalisent des improvisations ou jouent des scénarios pré-écrits, spécifiquement conçus pour susciter des expressions émotionnelles. Chaque dyade réalise deux passations, lors desquelles seul l'un des deux participants (puis l'autre) est équipé et enregistré en *mocap*. La *motion capture* a été réalisée à l'aide de caméras *Vicon*, principalement centrées sur le visage et les mains. Cinquante-trois marqueurs étaient fixés sur le visage, tandis que les mains n'étaient équipées que de trois marqueurs chacune, un choix cohérent avec un corpus conçu prioritairement pour l'analyse des expressions émotionnelles, et non pour la modélisation gestuelle (manuelle). L'audio a été enregistré simultanément, pour les deux locuteurs, à l'aide de deux microphones de haute qualité (48 kHz), et la captation vidéo assurée par deux caméras numériques à haute résolution. Le corpus est présenté comme multimodal et enrichi. Sur le plan verbal, il inclut des transcriptions complètes des dialogues, ainsi que des alignements au niveau du mot, de la syllabe et du phonème. Les sessions sont segmentées manuellement en unités d'énoncés, chacune étant annotée par au moins trois annotateurs humains. Le corpus propose des annotations émotionnelles à la fois catégorielles (colère, joie, excitation, tristesse, frustration, peur, surprise, autre, et état neutre) et dimensionnelles (valence, activation, dominance), permettant ainsi une analyse fine des variations expressives. En complément, des informations liées aux mouvements et angles de la tête sont également fournies, ce qui renforce la richesse multimodale du corpus. Toutefois, comme mentionné précédemment, la *motion capture* reste centrée sur le visage et limitée à trois points par main, ce qui restreint les possibilités d'analyse de la gestualité co-verbale manuelle. De manière corrélative, aucune annotation gestuelle n'est proposée.

### 6.2.7 Aoyama Gakuin

Le corpus présenté *Aoyama Gakuin* [147] a été conçu avec un focus spécifique sur certains types de gestes, en particulier les gestes iconiques et métaphoriques. Les participants devaient expliquer des sujets préparés, ce qui leur permettait d'adapter mentalement leurs gestes pour rendre le discours plus compréhensible. Le corpus comprend 5 heures d'enregistrements de 2 participants, réalisés en *mocap* avec 51 marqueurs (dont 32 sur les mains) à 100 fps. Il inclut des données tridimensionnelles et de l'audio, et les auteurs mentionnent un enrichissement gestuel. Cependant, ces annotations ne semblent pas accessibles ou disponibles publiquement.

### 6.2.8 NoXi

Le corpus Noxi [148] est un corpus accessible sur demande, dédié à l'étude de l'interaction sociale et des comportements adaptatifs face à des situations imprévues, telles que les interruptions. Il repose sur des dyades expert-novice engagées dans des discussions sur des sujets d'intérêt commun, médiées par écran, et vise à capturer des interactions spontanées. Le corpus comprend environ 25 heures d'enregistrements impliquant 87 participants, avec une

prédominance de données en français, allemand et anglais, ainsi que quelques occurrences dans d'autres langues. Les participants sont installés dans des pièces distinctes, placés devant un grand écran de télévision leur permettant de se voir de la tête aux pieds. Ils peuvent se déplacer librement dans l'espace, tant qu'ils demeurent face à l'écran. Les enregistrements audio (48 kHz) sont réalisés à l'aide de microphones-casques, tandis que les données vidéo (25 fps) et de motion capture sont capturées via un capteur Kinect 2 monté sur l'écran. Selon les auteurs, le corpus est enrichi d'annotations couvrant les signaux sociaux de bas niveau (gestes, mimiques, expressions faciales), les dimensions fonctionnelles (tours de parole, actes dialogiques) ainsi que les aspects interactionnels (engagement, intérêt, fluidité). Toutefois, ces annotations ne sont pas disponibles publiquement et n'ont pas pu être retrouvées auprès des auteurs. Par ailleurs, l'usage de la Kinect 2 introduit un bruit important dans les signaux de capture, limite la résolution vidéo, et s'accompagne d'une grande variabilité interindividuelle liée à la diversité des participants et des thématiques abordées.

### 6.2.9 GESRes

Le corpus *Gesure Exploration and Synthesis Resource* (GESRes, 2024) [149] est constitué d'un ensemble de données audiovisuelles richement annotées, spécialement conçu pour l'analyse approfondie des gestes co-verbaux de la main dans la communication humaine. Il comprend 2373 gestes annotés provenant de 9 locuteurs anglophones américains, répartis en trois contextes communicationnels distincts : des conférenciers universitaires, des politiciens, et des psychothérapeutes (cliniciens). Ces enregistrements, totalisant 3 heures et 46 minutes de séquences vidéo annotées, représentent des intentions variées, telles que persuader, augmenter la compréhension, ou créer un espace invitant. GESRes se distingue par la profondeur de ses annotations qui couvrent les sept catégories fonctionnelles de gestes (iconiques, métaphoriques, beat, emblématiques, déictiques, et adapteurs), ainsi que 11 propriétés physiques détaillées comme la hauteur, la distance au corps, l'orientation de la paume, la forme de la main, et la vitesse du mouvement. Il inclut également des données de suivi de pose 3D (33 points de repère de pose et 21 de la main). De plus, le corpus introduit un lexique gestuel étendu avec 149 descripteurs de forme gestuelle (lexèmes), dont 97 n'étaient pas documentés auparavant. Bien que ce corpus ait été enrichi par des experts, la segmentation des gestes est effectuée au niveau des phrases gestuelles, sans distinction des phases du geste. Même dans ce cadre, certaines phrases contiennent plusieurs gestes (le nombre étant indiqué sur un tier), alors que, idéalement, une phrase gestuelle ne devrait correspondre qu'à un seul geste selon la définition de Kendon [7, 8].

### 6.2.10 BEAT

Le corpus *Body Expression Audio Text* (BEAT, 2022) [150] est un jeu de données multimodal en accès libre. Il a été conçu pour l'étude des gestes conversationnels, et intègre également une dimension émotionnelle (annotation 8 classes d'émotion) ainsi qu'une dimension verbale (transcription et annotation des affiliés lexicaux des gestes). Les enregistrements comprennent à la fois des situations de monologue et des interactions dyadiques. Dans ces

dernières, seul un locuteur est enregistré, l'interlocuteur étant un instructeur dialoguant à distance. Lors des monologues, les participants devaient réaliser une tâche discursive actée et dirigée, avec une orientation émotionnelle fournie par les instructeurs. Le corpus totalise 61 heures en anglais, 16 heures en chinois, 2 heures en espagnol et 2 heures en japonais, provenant de 30 locuteurs. Les données ont été enregistrées en motion capture (120 fps, 73 articulations, dont 48 pour les mains) et incluent l'audio ainsi que les transcriptions verbales au niveau du mot et du phonème. L'enrichissement se veut particulièrement complet, couvrant la segmentation gestuelle, l'annotation fonctionnelle et l'affiliation lexicale des gestes, et offrant une variété d'informations bien supérieure à celle de la majorité des autres corpus disponibles.

L'ensemble de cet enrichissement, réalisé par un nombre important d'annotateurs recrutés via une plateforme participative, souffrent toutefois d'un manque de cohérence rendant les annotations difficilement exploitables. Malgré les limites notables de son enrichissement, qui seront détaillées dans le chapitre suivant, le choix de ce corpus pour fonder la démonstration de notre approche trouve sa motivation dans le fait que le corpus BEAT s'est très tôt imposé comme un standard de fait (*benchmark*) au sein de la communauté de l'apprentissage profond pour la génération de gestes, notamment dans le cadre du GENEA *Challenge*.

# 7 Conclusion de chapitre

Il existe aujourd'hui une diversité de corpus ciblant – ne serait-ce qu'en partie – la gestualité, qui se distinguent par leur taille, le type de situation enregistrée (dialogue ou monologue), la nature des tâches proposées, la langue d'usage, les modalités disponibles (audio, vidéo ou *mocap*), la qualité et la quantité de données, la présence ou non d'annotations verbales ou gestuelles, le protocole suivi, ou encore la documentation associée. De manière générale, tout corpus audiovisuel peut contenir des occurrences de gestes, mais leur seule présence ne garantit pas qu'un corpus soit adapté à l'étude de la gestualité : même les corpus conçus à cet effet ne conviennent pas forcément à tous les approches ou objectifs de recherche. Si en théorie, il est toujours possible d'annoter la gestualité dans un corpus qui contient des gestes non annotés, en pratique, plusieurs obstacles limitent cette possibilité : la qualité et le type de données enregistrées, le format et la structuration des fichiers, ou encore la faible adéquation entre les gestes présents et les objectifs spécifiques de l'étude. La pertinence d'un corpus dépend donc non seulement de sa richesse apparente, mais aussi de la manière dont ses données sont capturées, organisées et compatibles avec les méthodes d'analyse envisagées. De plus, il convient de souligner que si de nombreux corpus existent, la plupart ne sont pas librement accessibles. Comme évoqué précédemment, la protection des données demeure essentielle, et cette vigilance est d'autant plus cruciale à une époque où l'intelligence artificielle générative et d'autres applications facilitent une exploitation massive des données. Même lorsque l'anonymisation semble complète, certains traits très spécifiques, comme ceux de la voix, peuvent permettre une identification indirecte ou une réutilisation malveillante, exposant ainsi les participants à des risques. Cette limite de diffusion pose un obstacle sur la possibilité de réutiliser ou de reproduire les analyses.

Pour toutes ces raisons, la création d'un corpus multimodal de gestes constitue une entreprise complexe et chronophage, bien au-delà des seules contraintes financières. Il ne s'agit pas seulement de collecter des données audio, vidéo ou *mocap*, mais également de concevoir un protocole expérimental pertinent, de recruter des participants adaptés, et de garantir la qualité et la diversité des situations enregistrées. Chaque étape demande une attention rigoureuse pour assurer la fiabilité des mesures et la représentativité des interactions. La captation des gestes, notamment en *mocap*, implique un équipement spécifique, des contraintes techniques et une préparation minutieuse pour éviter les artefacts ou les données bruitées. De manière complémentaire, la transcription verbale, l'annotation fonctionnelle et la segmentation des gestes représentent une charge supplémentaire, nécessitant du temps, une formation des annotateurs et des outils adaptés. Toute dégradation à ce stade peut entraîner des répercussions importantes sur les étapes ultérieures, et fausser les analyses. Sur le plan computationnel, un bruit excessif ou des artefacts dans les données peuvent biaiser l'apprentissage des modèles, dégrader leurs performances de prédiction ou de génération, limiter leur capacité de généralisation à d'autres contextes d'acquisition, et compromettre la fiabilité des annotations automatiques produites ensuite. Ces erreurs peuvent également affecter l'analyse des représentations internes des modèles, en mettant en avant des dimensions liées aux contraintes techniques plutôt qu'à des propriétés gestuelles pertinentes. L'ensemble de ces facteurs fait de la constitution et de l'enrichissement d'un corpus une démarche longue et méthodologiquement exigeante, tant sur le plan théorique que pratique. Enfin, produire systématiquement des corpus toujours plus spécifiques, sans disposer d'une base commune pour l'étude de la gestualité en différents contextes (sans effacer les spécificités nécessaires aux projets de recherche) et sans normes établies, empêche toute comparaison fiable entre travaux et rend difficile la réutilisation des données. En l'absence d'approche structurée, le problème de fragmentation des ressources persiste, et les avancées méthodologiques restent limitées.

La réutilisation de corpus existants permet de réduire significativement le temps d'une recherche, en épargnant aux chercheurs une nouvelle collecte et l'élaboration d'un protocole, mais elle n'élimine pas tous les défis méthodologiques. Il peut être nécessaire de compléter les annotations existantes, de vérifier la qualité des transcriptions ou de recalibrer certaines données, notamment pour aligner précisément les modalités lorsqu'une désynchronisation entre fichiers est observée dans un corpus. De plus, la variabilité des participants, des conditions d'enregistrement et des dispositifs techniques peut introduire des biais ou des limitations pour certaines analyses. Ainsi, bien que réutiliser un corpus permette un gain de temps appréciable, cette approche implique toujours une adaptation méthodologique et des vérifications rigoureuses pour garantir la fiabilité et la pertinence d'une étude. Entre la présentation d'un corpus et les attentes que l'on peut en avoir, il existe en effet toujours un écart potentiel qui peut réserver des surprises. Les chercheurs peuvent se heurter à un manque de documentation, à des ambiguïtés dans les annotations, à des interprétations différentes des critères méthodologiques ou encore à des variations techniques imprévues. Ces éléments peuvent influencer la qualité des données et leur adéquation aux objectifs de l'étude, notamment lorsqu'il s'agit d'analyser le geste de manière fine et rigoureuse.

# Chapitre 4
# Analyse critique de l'enrichissement d'un corpus multimodal

## 1 Introduction

Au-delà de la nature même des données collectées, la qualité de l'enrichissement d'un corpus conditionne directement la portée des analyses gestuelles et de là, la robustesse des modèles capables d'en tirer parti. Le présent chapitre propose ainsi une réflexion approfondie sur les critères permettant d'évaluer l'enrichissement d'un corpus, et sur les pratiques nécessaires pour en garantir la fiabilité, l'interprétabilité et de là, la réutilisabilité dans des contextes variés. Cette réflexion a été menée à partir d'un jeu de données sélectionné au sein des corpus présentés au chapitre précédent. L'analyse des gestes telle qu'envisagée dans cette thèse requérait en effet de pouvoir travailler sur un corpus offrant à la fois des données *mocap* pour l'analyse informatique du mouvement, des données audios de qualité pour l'analyse verbale, et un minimum d'enrichissements verbal et gestuel suffisamment riches et fiables, notamment en termes de segmentation des gestes et d'annotation fonctionnelle structurée. En l'absence de ces éléments, ni l'analyse qualitative ni l'apprentissage automatique ne pouvaient être menés de manière rigoureuse, et le gain de l'approche interdisciplinaire défendue de cette recherche ne pouvait être démontré. Le corpus GESRes n'étant pas accessible au moment de ce choix, c'est le corpus BEAT qui est apparu comme le plus pertinent et le plus complet au regard des critères requis, du fait de l'ampleur et de la diversité des annotations multimodales proposées et de son positionnement comme ressource ambitieuse pour la recherche sur les comportements humains, notamment en sciences computationnelles. Comme pour les autres corpus listés dans le chapitre précédent toutefois, BEAT souffre de certains défauts qui lui sont propres, et c'est pour en mesurer leur ampleur que nous avons notamment réalisé une vérification préalable des

annotations proposées. Cette étape intermédiaire, essentielle pour garantir la réexploitabilité de l'ensemble du corpus, a été l'occasion de mener, au travers du cas d'étude de BEAT, une première analyse critique de la qualité d'un corpus multimodal de la gestualité fondée sur l'examen de son enrichissement, et de proposer des méthodes permettant d'évaluer sa fiabilité et sa réutilisabilité.

Dans cette perspective, le corpus BEAT est examiné en trois volets complémentaires : (A) la présentation générale du corpus et de son enrichissement, (B) la description de la méthodologie d'analyse quantitative adoptée pour évaluer l'enrichissement proposé, et les résultats associés, et (C) une brève analyse de la qualité des annotations fonctionnelles. Ce chapitre conclut sur l'état de l'enrichissement proposé par les auteurs, en mettant en évidence ses apports et ses limites au regard des usages envisagés.

# 2 Le corpus BEAT

Les analyses menées dans cette recherche se concentrent spécifiquement sur la composante anglophone (environ 62h et 30 locuteurs) du corpus BEAT [150]. Ces données sont équilibrées entre sessions de conversation de 10 minutes et sessions de monologue d'une minute. Les conversations (48,38% des enregistrements anglophones), couvrant 20 sujets prédéfinis (33 % de débats, 67 % de descriptions) ont été capturées de manière naturelle et spontanée, sans consignes spécifiques. Les sessions de *self-talk* (monologues correspondant à 51,62% des enregistrements anglophones), quant à elles, consistent en des enregistrements scénarisés (protocole acté et dirigé) dans lesquels les participants jouent des réponses préparées sur différents sujets du quotidien, sans consigne concernant la gestuelle qu'ils doivent adopter. Parmi les questions posées à l'origine, 64 visaient à susciter des émotions neutres, les autres étant réparties de manière équilibrée entre sept autres catégories émotionnelles (tristesse, peur, colère, joie, dégoût, surprise, et mépris). Les participants devaient visionner 2 à 10 minutes de vidéos émotionnellement stimulantes en amont des enregistrements des monologues afin de susciter le type d'émotion ciblé. Ils devaient ensuite rejouer un contenu verbal pré-écrit en l'accompagnant de leurs propres gestes.

## 2.1 Captation et format de données

Les enregistrements ont été réalisés en environnement contrôlé, au sein d'une salle dédiée à la capture de mouvement. La *mocap* a été effectuée à l'aide de 16 caméras synchronisées fonctionnant à 120 Hz. Les participants portaient une combinaison *Vicon* équipée de 77 marqueurs réfléchissants, permettant une capture complète du corps en trois dimensions (fichiers `.bvh`). La capture des mouvements faciaux repose sur *ARKit* associé à une caméra de profondeur, permettant d'extraire 52 coefficients de *blendshape* à 60 Hz, définis selon le *Facial Action Coding System* (FACS) [151]. La base de données inclut également des enregistrements audios au format `.wav` (45 Hz). Le corpus ne contenant pas de vidéo RGB, les vidéos disponibles dans le corpus sont des reconstructions numériques des positions tridimensionnelles.

## 2.2 Enrichissements disponibles

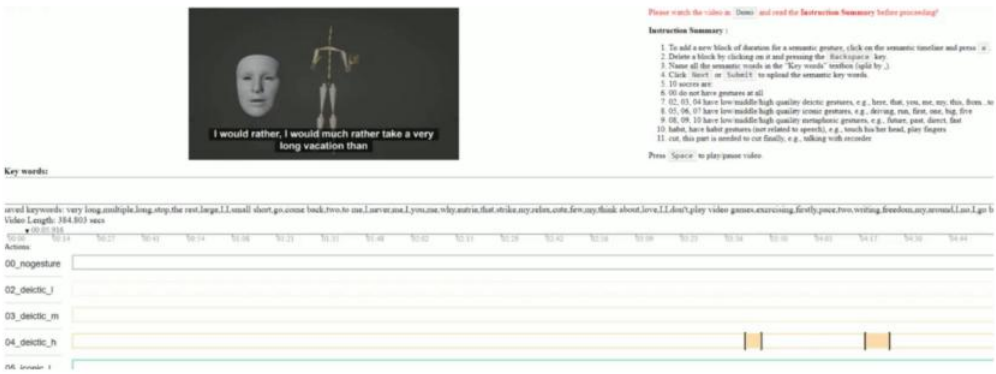



*Figure 11. Interface utilisée par les auteurs de BEAT pour l'annotation des gestes [150]*

Pour enrichir le corpus, les auteurs ont fait appel à des annotateurs non experts recrutés sur la plateforme *Amazon Mechanical Turk*. Il est important de noter qu'aucun manuel de codage n'a à priori été fourni aux annotateurs, et que la procédure d'annotation n'est décrite que de manière très succincte dans l'article de présentation du corpus. Les annotateurs recevaient apparemment uniquement une courte introduction aux consignes et une vidéo de démonstration avant d'accéder à l'interface d'annotation (Figure 11). Celle-ci permettait aux annotateurs d'écouter le signal audio et de visualiser simultanément le squelette issu de la *mocap*, les expressions faciales reconstituées sur un visage numérique placé sur le côté du squelette, et la transcription verbale.

| Score | Catégorie associée |
|---|---|
| 0 | Absence de geste |
| 1 | Geste de battement |
| 2 à 4 | Geste déictique |
| 5 à 7 | Geste iconique |
| 8 à 10 | Geste métaphorique |
| Habit | Geste de confort |

*Tableau 1. Instructions d'annotation sémantiques pour les annotateurs de BEAT*

Via cette interface, les annotateurs devaient indiquer, *frame* par *frame*, la présence d'un geste, et qualifier la nature et le degré de son sémantisme (Tableau 1), à l'aide d'un score compris entre 0 et 10 (plus le score est haut, plus son sémantisme est considéré comme fort). Ces scores correspondaient, selon les auteurs, à la classification de McNeill [9, 10] (1 pour les gestes de *battement*, 2 à 4 pour *déictiques*, 5 à 7 pour les gestes *iconiques*, et 8 à 10 pour les gestes *métaphoriques*). À ces scores liés à des catégories fonctionnelles a été ajouté le label « Habit », dont la définition n'est pas clairement établie par les auteurs ; il semble correspondre à des gestes de confort, illustrés par les exemples suivants dans l'interface d'annotation : se toucher la tête ou jouer avec ses doigts. Les annotateurs ne pouvaient annoter qu'un seul score par geste. Dans les consignes affichées sur l'interface, ces scores étaient exemplifiés par des mots-clés jugés fréquents afin de guider l'annotation : par exemple, *here*, *that* ou *this* pour

illustrer les types de mots fréquemment associés à l'occurrence de gestes *déictiques* ; *driving, run, one* pour les gestes *iconiques* ; et *future*, *past* ou *direct* pour les gestes *métaphoriques*.

Sur les 600 annotateurs initialement recrutés, seuls 118 ont été retenus après un test de qualification, dont les critères d'évaluation ne sont pas renseignés par les auteurs du corpus. La procédure d'annotation elle-même n'est pas clairement documentée, et les documents annexes fournissent peu d'informations sur les tâches assignées aux annotateurs selon les vidéos annotées, la répartition des segments annotés ou encore sur le score d'accord inter-annotateurs.

Le corpus propose différents fichiers d'enrichissement, du niveau verbal et du niveau gestuel. Au niveau verbal, les transcriptions sont fournies au format `.TextGrid`. Ces fichiers contiennent une transcription par mots et par phonèmes. La transcription orthographique a été obtenue à l'aide d'un modèle de reconnaissance vocale automatique, développé en interne, non décrit dans l'article de présentation du corpus. L'alignement temporel de la transcription (mots et phonèmes), quant à elle, a été réalisé à l'aide de MFA [137], basé sur Kaldi [152]. La procédure de vérification de ces transcriptions n'est ni mentionnée ni détaillée par les auteurs. Au niveau gestuel, chaque fichier d'enregistrement est associé à un fichier `.txt` contenant plusieurs lignes renseignant différents aspects des gestes identifiés par les annotateurs. Chaque geste est ainsi caractérisé par son type fonctionnel (tel que défini par le score présenté précédemment), son temps de début et de fin, sa durée (en secondes) et, lorsqu'il a pu être déterminé, son affilié lexical.

# 3 Statistiques descriptives de l'enrichissement du corpus

## 3.1 Analyses statistiques des annotations fonctionnelles des gestes

### 3.1.1 Annotations prises en compte dans le corpus initial

L'objectif de l'analyse quantitative des annotations gestuelles du corpus a été d'obtenir un aperçu général de leurs caractéristiques, et d'ainsi évaluer leur cohérence par rapport aux connaissances issues de l'état de l'art. Pour cela, nous avons mobilisé à la fois des outils classiques de gestion de données et des outils spécialisés dédiés à l'annotation (tel ELAN ou Praat). Cette analyse a porté sur la segmentation des gestes proposée par les auteurs, leur étiquetage fonctionnel ou encore l'annotation de leurs affiliés, en appui sur les mouvements reconstruits à partir des données de *motion capture* et sur l'audio synchronisé fourni. Les données d'enrichissement liées à l'annotation fonctionnelle des gestes ont été retraitées afin d'exploiter non pas les scores sémantiques proposés par les auteurs, mais leur correspondance directe avec la catégorisation de McNeill [9, 10], comme présenté précédemment dans le Tableau 1. La segmentation des gestes n'étant pas explicitement décrite comme correspondant à une unité gestuelle, une phrase gestuelle ou une phase en particulier, nous avons considéré par défaut que chaque segment de mouvement annoté devait correspondre à une séquence comprenant au moins un – mais seulement un – *stroke*.

Au total, 40 732 segments identifiés comme gestes ont été analysés, avec autant d'annotations fonctionnelles (une par geste) et 19 790 annotations d'affiliés.

### 3.1.2 Nombre d'annotations dans le corpus

| Catégorie | N | % |
|---|---|---|
| Battement | 20 914 | 51,34 |
| Iconique | 8162 | 20,03 |
| Métaphorique | 6933 | 17,02 |
| Déictique | 4468 | 11,05 |
| Habit | 255 | 0,62 |
| **Total** | **40 732** | **100 %** |

*Tableau 2. Statistiques descriptives globales des annotations fonctionnelles de BEAT [153]*

L'analyse des annotations du corpus original révèle une répartition fortement hétérogène entre les types de gestes annotés (Tableau 2). Les gestes de type battement dominent en effet nettement, représentant plus de la moitié des cas (51,34 % des 40 732 annotations recensées), tandis que les gestes *iconiques* (20 %), *métaphorique*s (17 %) et *déictiques* (11 %) sont moins fréquents, et que les gestes dits « habit » (0,62 %).

Cette prédominance des battements peut paraître surprenante dans un corpus visant une annotation exhaustive des gestes co-verbaux, fondée sur les catégories de McNeill et couvrant des contextes discursifs variés. On s'attendrait en effet à observer une représentation plus équilibrée des différents types gestuels. La forte concentration sur une seule catégorie suggère de potentielles retombées d'un manque de clarté concernant les critères de catégorisation et les pratiques d'annotation, dont aurait pu résulter la tendance à mobiliser les battements comme catégorie par défaut. On peut en effet supposer que ce déséquilibre marqué reflète une mauvaise compréhension des catégories fonctionnelles par les annotateurs.

### 3.1.3 Recouvrement temporel des annotations

| Catégorie | Durée (h :m :s) | % |
|---|---|---|
| Battement | 54 :42 :2.775 | 88,27 |
| Iconique | 2 :51 :13.471 | 4,61 |
| Métaphorique | 2 :21 :10.560 | 3,80 |
| Déictique | 1 :32 :60.669 | 2,49 |
| Habit | 0 :12 :53.661 | 0,35 |
| Absence de geste | 0 :17 :50.349 | 0,48 |
| **Total** | **61 h 1m 1.485 s** | **100 %** |

*Tableau 3. Recouvrement temporel des annotations dans le corpus [153]*

La répartition temporelle des gestes (Tableau 3) dans le corpus apparaît également aussi étrange qu'elle est hétérogène : les annotations indiquent que les gestes interviennent quasiment tout le temps, avec seulement 0,48 % du temps du corpus qui ne comporterait pas de geste.

Ainsi les gestes annotés comme « battements », qui représentent donc plus de la moitié des gestes annotés, couvrent-ils 88,28 % de la durée totale du corpus, les autres types de gestes ne couvrant quant à eux qu'une proportion proportionnellement très réduite d'environ 11% du corpus (*iconiques* : 4,61 %, *métaphorique*s : 3,80 %, *déictiques* : 2,49 %, et *habits* : 0,63 %). La présence continue de gestes paraît très surprenante, dans la mesure où elle implique que le locuteur gesticulerait en permanence, même lors des pauses intervenant entre et dans ses énoncés. À titre de comparaison, l'étude de Seyfeddinipur [154] fait état d'environ 55 % du temps de parole accompagné de gestes. Ce constat est d'autant plus étonnant que près de la moitié de ces gestes sont interprétés comme des *battements*, qui sont pourtant par définition des gestes brefs, calqués sur la prosodie ou le rythme ponctuellement marqué du discours. Un tel enchaînement continu de gestes semble dès lors très peu probable, ce qui interroge inévitablement les pratiques de segmentation et la fiabilité globale des annotations réalisées, et leur capacité à refléter correctement la gestualité réellement produite par les locuteurs.

## 3.2 Analyses statistiques de la durée des gestes

L'objectif de l'analyse de la durée des gestes visait à mieux comprendre la structure et la qualité des segmentations gestuelles fournies. Nous avons réalisé différentes opérations de décompte par catégorie fonctionnelle de gestes, ainsi que des mesures de durée moyenne (Moy), durée médiane (Med), et de l'écart type (SD). Ces analyses ont été choisies afin de mettre en évidence des tendances générales dans les segmentations opérées, et d'identifier de potentiels biais ou incohérences les concernant.

### 3.2.1 Durée des gestes par catégorie fonctionnelle

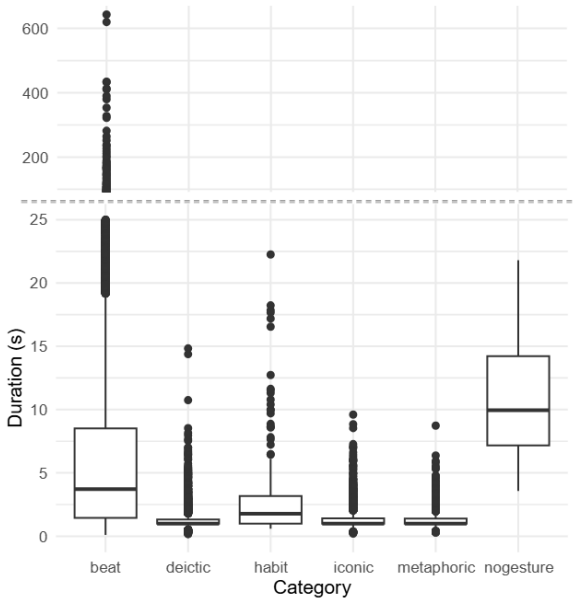



*Figure 12. Représentation visuelle de la distribution des durées de gestes dans le corpus [153]*

| Catégorie | Moy | Med | SD | Min | Max |
|---|---|---|---|---|---|
| Battement | 9,42 | 4,34 | 17,54 | 0,10 | 643,00 |
| Iconique | 1,26 | 1,00 | 0,59 | 0,21 | 9,60 |
| Métaphorique | 1,22 | 1,00 | 0,51 | 0,29 | 8,73 |
| Déictique | 1,24 | 1,00 | 0,81 | 0,16 | 14,83 |
| *Habit* | 3,03 | 1,78 | 3,62 | 0,59 | 26,82 |
| **Total** | **5,47** | **1,49** | **13,22** | **0,10** | **643,00** |

*Tableau 4. Analyse des durées de gestes (secondes) dans le corpus BEAT [153]*

L'analyse des durées des gestes annotés dans le corpus BEAT (Tableau 4, Figure 12) révèle une forte variabilité : la durée moyenne est de 5,47 secondes, tandis que la médiane n'atteint que 1,5 seconde. Cette dispersion s'explique par la coexistence d'annotations très courtes et de gestes exceptionnellement longs.

Dans le détail, on observe également des profils temporels très contrastés selon les types de gestes annotés. Les gestes catégorisés comme *battements* présentent ainsi une durée moyenne particulièrement élevée (9,42 s), avec une médiane de 4,34 secondes et un écart-type très important (17,54 s), témoignant d'une variabilité extrême au sein de cette catégorie. Leur étendue varie ainsi de 0,10 à 643 secondes (soit 10 min 43 s), révélant la présence de valeurs exceptionnellement longues – d'autant plus pour cette catégorie de gestes pourtant par définition très brefs – et confirmant une distribution fortement asymétrique où ces quelques gestes particulièrement étendus tirent la moyenne vers le haut (Figure 12, Tableau 4). En revanche, les gestes *iconiques*, *métaphoriques* et *déictiques* sont nettement plus courts et homogènes : leurs durées moyennes sont proches de 1,2 seconde (1,26 s pour les *iconiques*, 1,22 s pour les *métaphoriques*, 1,24 s pour les *déictiques*), avec des médianes proches de 1 seconde et des écarts-types inférieurs à 1 seconde également. La catégorie *habit* occupe une position intermédiaire (moy = 3,03 s, SD = 3,62 s) avec des durées allant de 0,59 à 26,82 secondes.

### 3.2.2 Durées aberrantes

Les valeurs aberrantes (*outlier,* Out.), au sein des durées de gestes, ont été identifiées par l'écart interquartile (IQR), calculé comme la différence entre le troisième ($Q_3$) et le premier quartile ($Q_1$). Une valeur a été considérée comme un *outlier* si elle se situait en dehors de l'intervalle $[Q_1 - 1{,}5 \times IQR\,, Q_3 + 1{,}5 \times IQR]$, et comme un *extreme outlier* (Ext.) si elle se situait en dehors de l'intervalle $[Q_1 - 3 \times IQR\,, Q_3 + 3 \times IQR]$. Un *outlier* correspondait donc à une observation située en dehors de l'intervalle attendu des valeurs normales d'un ensemble de données, pouvant résulter d'une erreur de mesure, d'une occurrence rare, mais réelle, ou d'autres phénomènes inhabituels. Tandis qu'un *outlier extrême* correspondait à une valeur relevant très probablement d'erreurs ou d'artefacts, et n'était donc pas interprétée comme une potentielle occurrence rare, mais valide.

| Catégorie | Outlier | | Extrême outlier | | Total | |
|---|---|---|---|---|---|---|
| | *N* | % | *N* | % | *N* | % |
| *Battement* | 1005 | 47,1 | 797 | 53,9 | 1802 | 49,9 |
| *Iconique* | 389 | 18,2 | 215 | 14,5 | 604 | 16,7 |
| *Métaphorique* | 338 | 15,8 | 159 | 10,8 | 497 | 13,8 |
| *Déictique* | 389 | 18,2 | 290 | 19,6 | 679 | 18,8 |
| *Habit* | 9 | 0,4 | 15 | 1,0 | 24 | 0,7 |
| **Total** | **2133** | **59,1** | **1479** | **40,9** | **3612** | **8,9** |

*Tableau 5. Analyse des valeurs aberrantes de durées de gestes par catégorie fonctionnelle [153]*

Sur la totalité des gestes annotés, environ 8,9 % des durées sont identifiées comme des *outliers* (Tableau 5), parmi lesquels environ 41 % correspondent à des *outliers extrêmes*. Ces valeurs concernent majoritairement les durées des *battements* (Out. 47 %, Ext. 53,9 %), lesquelles représentent à elles seules près de la moitié de l'ensemble des durées considérées comme aberrantes. Les autres catégories de gestes (*iconiques*, *métaphoriques* et *déictiques*) sont nettement moins représentées parmi ces valeurs, chacune présentant des proportions de *outliers* et *outliers extrêmes* comparables, comprises entre 14 % et 19 %, tandis que les gestes de type *habit* restent très marginalement représentés ($<$ 1 %).

Cette analyse, qui révèle avant tout l'hétérogénéité et l'incongruité des durées des *battements* annotés, complète donc l'analyse du recouvrement temporel des annotations effectuée précédemment. Effectivement, ce recouvrement du corpus par les annotations de gestes de *battement* s'explique notamment par les durées particulièrement longues des gestes de ce type tels qu'annotés dans BEAT. Si leurs durées sont surprenantes étant donné leur nature ordinairement brève et rapide [9, 10], à l'inverse, les gestes *iconiques*, *métaphoriques* et *déictiques* restent courts et homogènes, contrastant fortement avec la variabilité extrême des *battements*. Cependant, de manière générale, les durées moyennes de toutes les catégories de gestes annotés dans le corpus semblent bien plus longues que celles rapportées dans la littérature pour un *stroke*, estimé à une durée moyenne entre 0,38 s (SD = 0,14 s [155]) et 0,58 s (med. = 0,42 s [73]). De plus, ces durées moyennes ne semblent pas correspondre à une segmentation en *gesture phrase*. En effet, les *gestures phrases* sont estimées dans la littérature entre 3,1s (sd = 1,48s [156]) et 3,7s (sd = 3,27s [157]), tandis que dans BEAT, les durées moyennes des *battements* sont bien plus élevées, et celles des gestes *iconiques*, *métaphoriques* et *déictiques* sont plus faibles. Par ailleurs, au vu de la proportion élevée de battements parmi les valeurs aberrantes et les *outliers extrêmes*, l'hypothèse d'occurrences rares mais valides peut être écartée, ce qui renforce l'interprétation selon laquelle ces durées résultent plutôt de problèmes d'annotation ou de segmentation. Ces observations confirment les interrogations précédemment soulevées quant à la qualité de ces annotations, mettant là encore en évidence les apparentes incompréhensions concernant ce que constitue réellement un geste.

## 3.3 Analyses statistiques de l'affiliation lexicale des gestes

L'analyse statistique des affiliés lexicaux annotés a été menée en envisageant à la fois leur nombre, leur composition (nombre de mots par affiliation annotée) et leur justesse orthographique. Dans un premier temps, notre analyse s'est concentrée sur l'identification des erreurs orthographiques présentes dans le corpus BEAT, ces erreurs pouvant affecter l'analyse sémantique des associations entre gestes et parole en altérant le sens des affiliés, voire de la phrase, donc celui du contexte dans lequel s'inscrit le geste. Pour évaluer la potentielle ampleur des erreurs de graphies, nous avons utilisé la bibliothèque de traitement automatique du langage naturel *spaCy* [138], afin de vérifier si chaque mot annoté correspondait bien à un mot anglais valide reconnu par le modèle linguistique (modèle utilisé : `en_core_web_lg`). Chaque mot (*token*) d'un affilié a donc été analysé isolément.

| **Catégorie** | **Affilié** | | **Nombre de mots** (par affilié) | | **Erreur orthographique** | |
|---|---|---|---|---|---|---|
| | *N* | *% cat.* | *Moy* | *Med* | *N* | *% affilié.* |
| *Battement* | -- | -- | -- | -- | -- | -- |
| *Iconique* | 8145 | 99.7 | 3,8 | 2 | 69 | 0,8 |
| *Métaphorique* | 6922 | 99.8 | 3,7 | 2 | 42 | 0,6 |
| *Déictique* | 4468 | 100 | 3,7 | 2 | 47 | 1,1 |
| *Habit* | 255 | 100 | 2,2 | 2 | 3 | 1,2 |
| **Total** | **19 790** | **48,6** | **2,33** | **1** | **161** | **0,8** |

*Tableau 6. Analyse descriptive de l'affiliation lexicale dans le corpus et des détections de spaCy [153] (N = nombre d'occurrences ; moy = moyenne ; med = médiane)*

Au total, 19 790 affiliés (pouvant comprendre un ou plusieurs mots) ont été recensés dans le corpus et analysés (Tableau 6). Environ 48,6 % des gestes du corpus présentent une affiliation, ce pourcentage s'expliquant par le fait que les *battements*, qui représentent la moitié des annotations de gestes du corpus, n'ont aucune affiliation lexicale renseignée, tandis que les gestes des autres catégories en possèdent presque tous une. Il est toutefois surprenant que les battements n'aient aucune affiliation lexicale, alors même que, par définition, ils sont étroitement liés à l'énonciation et la prosodie du discours [158, 159], ce qui rend d'ordinaire très facile l'identification de leur association à un terme particulier.

Un affilié peut comprendre un ou plusieurs mots. Dans le corpus BEAT, l'ensemble des affiliés, toutes catégories de gestes confondues, correspond en moyenne à 2,33 mots (médiane = 1). En considérant les différentes catégories de gestes prises séparément, on constate que les affiliés des gestes *iconiques*, *métaphoriques* et *déictiques* comptent en moyenne environ 4 mots (médiane = 2), et les gestes *habit* 2,2 mots. Dès qu'au moins un mot de l'affilié n'est pas reconnu, l'ensemble de l'affilié est considéré comme incorrect, selon la règle « une erreur = un affilié erroné ». À titre d'exemple, dans le fichier 1_1 du locuteur Zhang, un geste *métaphorique* a pour affilié « *enbironnment* ». Ce mot, invalide en raison d'une coquille graphique, est comptabilisé comme une erreur. Une autre source d'erreurs provient de l'absence d'espaces, comme dans le fichier 0_11 du locuteur Li, où l'affilié d'un geste iconique est « *fashion magazines and the inspirationbooks* » : l'absence d'espace entre « *inspiration* » et « *books* » rend

le *token* invalide. Enfin, dans le fichier 0_6 de la locutrice Carla, un geste déictique est associé à l'affilié « *I.didn't* » : la présence d'un point à la place d'un espace crée là encore un *token* non reconnu par *spaCy* [138].

L'analyse des affiliés lexicaux des gestes annotés dans le corpus BEAT met en évidence des taux d'erreur globalement faibles, bien que légèrement variables selon les catégories gestuelles (Tableau 6). Sur l'ensemble des affiliés, 161 erreurs ont été détectées par *spaCy*, soit un taux d'erreur global de 0,8 %. Le taux d'erreur par catégorie de gestes varie entre 0,6 % et 1,2 % : 0,8 % des affiliés lexicaux des gestes *iconiques* présentent une erreur, 0,6 % pour ceux des *métaphoriques*, 1,1 % pour ceux des *déictiques*, et 1,2 % pour ceux des *habits*.

Globalement, la majorité des termes des affiliés lexicaux sont correctement reconnus, ce qui témoigne de la bonne performance de l'ASR. Certaines erreurs peuvent par ailleurs provenir de termes absents du lexique de *spaCy*, mais valides dans le contexte du discours (ex. : nom propre spécifique). Seule une relecture humaine systématique permet de garantir la justesse des transcriptions et, par conséquent, leur exploitabilité pour une analyse sémantique ou syntaxique, cependant dans le cas du corpus BEAT, les auteurs ne précisent pas si cette transcription a été révisée manuellement. Le véritable enjeu de l'enrichissement lexical de BEAT est toutefois moins lié à la précision orthographique des mots détectés qu'au choix des mots qui ont été associés aux gestes par les annotateurs. Comme nous le verrons dans les exemples qualitatifs, même lorsque le mot est correctement orthographié, l'affiliation sémantique réalisée par un annotateur non-expert (ou son absence, notamment pour les gestes de *battements*) peut s'avérer incorrecte ou peu pertinente pour appréhender la fonction du geste.

# 4 Analyse fonctionnelle : exemple de cas

## 4.1 Méthodologie de réannotation ponctuelle

Dans l'objectif d'évaluer la fiabilité des annotations des types fonctionnels des gestes du corpus, nous avons procédé à une réannotation ponctuelle d'une séquence très limitée, sélectionnée spécifiquement parce qu'elle présente deux gestes saillants (présentés dans l'analyse ci-dessous), dont l'interprétation de l'un d'entre eux a immédiatement semblé problématique et particulièrement illustrative de la mise en doute de la fiabilité de l'enrichissement original de BEAT. Cette brève réannotation, indépendante de celle qui sera détaillée plus loin, a porté sur un segment de 5 secondes extrait du fichier 0_53 (13 s 500 à 18 s 500) du locuteur Wayne (id. : n° 1). Elle a été menée sur la base des catégories fonctionnelles utilisées par les annotateurs dans le cadre de l'enrichissement de BEAT, donc celles de McNeill : gestes *métaphoriques*, *déictiques*, *iconiques*, *battements*.

La procédure consistait, sans se baser sur les annotations d'origine, à proposer une nouvelle segmentation des occurrences de gestes identifiables dans cette séquence, à les catégoriser sur le plan fonctionnel, et à renseigner leur affiliation lexicale lorsque cela s'avérait possible et pertinent. L'objectif de cette réannotation était d'évaluer de manière fine et qualitative, par comparaison, la cohérence et la pertinence des annotations originales, et d'ainsi

mettre en évidence d'éventuelles divergences ou limites du protocole d'annotation d'origine de BEAT. Bien que très restreinte, cette investigation ponctuelle nous a semblé essentielle pour compléter l'analyse quantitative, compte tenu des anomalies observées concernant les durées et la répartition des gestes.

## 4.2 Analyse comparative de l'enrichissement d'une courte séquence

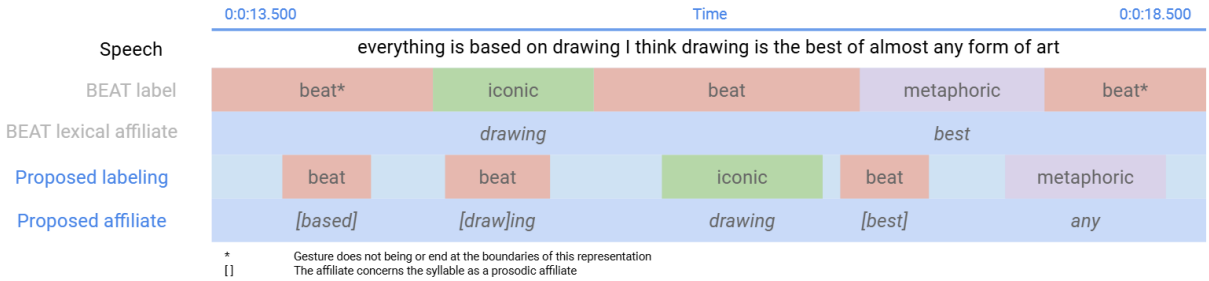



*Figure 13. Comparaison entre l'enrichissement fonctionnel original et celle proposée dans cette analyse [153]*

La Figure 13 met en parallèle les annotations originales du segment extrait du corpus avec les nouvelles annotations réalisées sur celui-ci. Pendant cette séquence Wayne prononce : « *everything is based on drawing I think drawing is the best of almost any form of art* ». Dans l'enrichissement original de BEAT, cette séquence est annotée comme une série de gestes continus (pas de pause) comprenant trois *battements*, entrecoupés d'un geste *iconique* et d'un geste *métaphorique*. Il est à noter que le premier et le dernier battement couvrent tous deux une durée allant au-delà des limites de la séquence annotée. Du point de vue de l'affiliation lexicale, le geste iconique est associé au mot « *drawing* » et le geste *métaphorique* au mot « *best* ».

Notre propre interprétation de cette séquence révèle plusieurs différences. Si les mêmes catégories de gestes ont été annotées, et en même nombre, leur ordre d'apparition diffère, tout comme leurs durées, donc avec un ancrage temporel au discours différent. Les affiliations lexicales diffèrent également : dans les nouvelles annotations, les battements sont affiliés à des syllabes – « *based* », « *best* », et « *draw* » (issu de « *drawing* ») – alors que dans BEAT, aucun affilié n'est associé aux battements. S'agissant du geste *iconique*, si son affiliation lexicale au mot « *drawing* » dans la réannotation parait identique à celle des annotateurs de BEAT, elle est en réalité plus précise. Dans la séquence d'origine en effet, il n'est pas possible de déterminer à quelle occurrence du mot « *drawing* » le geste *iconique* est affilié (« *based on drawing / drawing is* »), là où nous avons proposé une correspondance numérique explicite (non représentée dans la Figure 13) entre l'identifiant d'un geste et le ou les termes de son affilié lexical, de manière à éviter toute confusion entre occurrences identiques.

Si cette indétermination n'affecte pas la nature de l'affiliation lexicale elle-même, le terme étant identique, elle pose toutefois un problème pour l'analyse fine de la synchronie geste-parole. Elle peut en effet affecter en particulier l'étude des relations temporelles entre la réalisation du geste et l'énonciation de son affilié lexical, cette synchronisation ne pouvant plus être correctement déterminées.

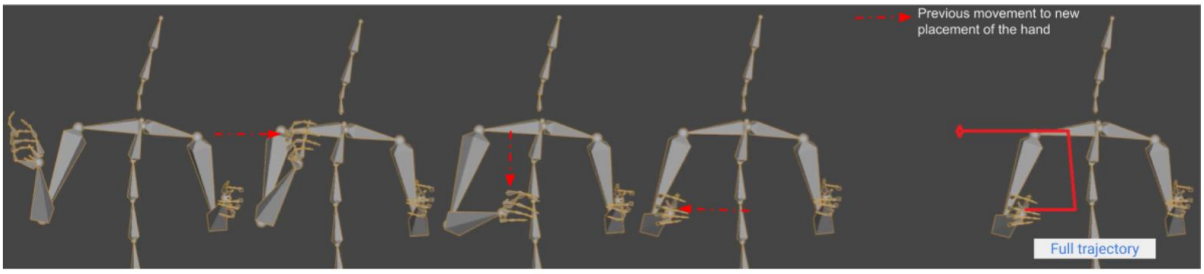



*Figure 14. Geste iconique réalisé dans la séquence réannotée [153]*

La Figure 14 illustre le déroulé du geste *iconique* tel qu'identifié dans la nouvelle annotation, soit le segment compris entre 00 :15 :504 et 00 :16 :436. Alors que dans les annotations originales de BEAT cette séquence est interprétée comme un geste de *battement*, dont l'étendue dépasse d'ailleurs le découpage proposé ici, notre analyse, fondée à la fois sur la forme du geste et le mot co-occurrent, interprète donc ce geste comme un geste iconique, en tant qu'il sert à représenter les contours de ce qui peut être compris, dans ce contexte, comme une feuille de dessin.

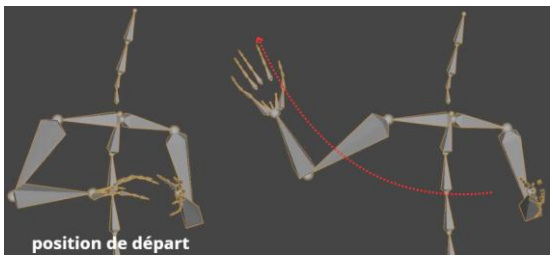



*Figure 15. Geste métaphorique réalisé dans la séquence réannotée*

Le geste *métaphorique* (Figure 15) réannoté entre 00:17.537 et 00:18.144, contre 00:16.712 à 00:17.712 dans l'annotation d'origine, suit une logique interprétative similaire. Dans ce passage, Wayne tient d'abord ses deux mains proches de son corps, puis lève brusquement la main droite vers le haut, paume ouverte vers l'avant. Du fait de la forme du geste et du discours associé, ce geste a pu être interprété comme étant affilié au mot « *any* », en tant qu'il représente de manière abstraite une portée ou une extension. Dans les annotations originales, ce segment chevauche deux annotations : la fin d'un geste interprété comme un *métaphorique*, débuté un peu plus tôt dans la séquence et associé au mot « *best* », et le début d'un battement, sans affilié, perdurant jusqu'au bout de l'extrait – interprétation dont nous n'avons pas pu reconstruire la logique.

# 5 Conclusion de chapitre

Les analyses quantitatives et fonctionnelles de l'enrichissement du corpus BEAT ont ainsi mis en évidence plusieurs caractéristiques saillantes et pour le moins étonnantes de la gestualité telle qu'annotée à l'origine. La répartition des gestes selon les catégories et les durées montre une hétérogénéité importante et certaines grosses anomalies, telles que la surreprésentation des *battements*, ou la présence de durées extrêmement aberrantes, suggérant que le protocole d'enrichissement présente de grandes faiblesses. En effet, les critères d'annotation n'ont peut-

être pas été définis de manière suffisamment claire et précise pour permettre une annotation réellement qualitative. De mêmes, les erreurs orthographiques dans les affiliés lexicaux, bien que d'un taux global faible, peuvent néanmoins affecter certaines analyses. Enfin, sur le plan fonctionnel, la brève analyse comparative menée tend à démontrer que les annotations d'origine semblent arbitraires, ou à minima très difficiles à interpréter, ce qui peut introduire des biais dans l'analyse quantitative des types de gestes, limiter la comparabilité des données entre corpus et, à long terme, compliquer la réutilisation de cet enrichissement de corpus. L'absence de manuel de codage détaillé et l'absence d'informations détaillées sur les procédures de vérifications compliquent également la reproductibilité et l'interprétation des annotations.

Toutes ces observations confirment la nécessité d'une vigilance particulière à adopter non seulement lors de la réalisation de l'enrichissement d'un corpus, mais aussi lorsqu'il s'agit de réexploiter ces données secondaires. Dans le cas de BEAT, en effet, l'enrichissement proposé ne semble pas correspondre à une *phrase gestuelle* ou un *stroke* tel que conceptualisé dans les *gestures studies*, ce qui limite la fiabilité des annotations fonctionnelles associées et leur exploitabilité pour des usages computationnels. Une telle représentation partielle ou imprécise des gestes risque, en effet, d'induire des biais qui compromettent la capacité des modèles à apprendre des régularités gestuelles robustes, avec à la clé des performances trompeuses, une généralisation limitée et des résultats difficilement interprétables.

Ce chapitre met ainsi en évidence plusieurs critères permettant d'évaluer rapidement la qualité d'un corpus à partir d'analyses statistiques mises au regard des connaissances de l'état de l'art. L'analyse des durées des annotations permet de détecter des valeurs extrêmes ou une distribution incohérente des gestes, signalant une segmentation potentiellement problématique. La vérification des *tokens* de mots associés aux gestes, par exemple via des outils de traitement automatique du langage, aide à repérer des erreurs orthographiques, des mots non lexicaux ou des problèmes de segmentation lexicale, qui sont autant de freins à une exploitation fiable des données verbales. Enfin, la comparaison de séquences d'interprétation, nécessitant toutefois de pouvoir confronter les annotations existantes à une réinterprétation contextualisée, permet de mettre en évidence des divergences dans la segmentation, le *labeling* ou l'affiliation lexicale, et d'ainsi évaluer la cohérence globale des annotations. Ces critères constituent un cadre pratique pour juger de la fiabilité et de l'exploitabilité d'un corpus, sans nécessiter une analyse plus détaillée ou exhaustive.

Si le développement de nouvelles méthodologies d'annotation constitue un progrès pour l'étude de la gestualité, il est tout aussi crucial de standardiser et de partager les pratiques existantes. En sciences du langage, la production de manuels de codage détaillés pour l'annotation gestuelle est déjà courante, mais ces ressources restent souvent circonscrites à des projets spécifiques. En sciences computationnelles, lorsqu'un enrichissement gestuel est produit, il est rare qu'il soit accompagné d'un manuel de codage détaillé, et la réalisation de ce type de document semble encore limitée. Dans l'idéal, un enrichissement de corpus destiné à l'étude du geste devrait s'appuyer sur les pratiques méthodologiques reconnues dans les *gesture studies*, garantissant que les annotations reflètent fidèlement les aspects formels et fonctionnels des gestes. Dans le cas de BEAT, cet alignement semble partiellement absent : certaines annotations

ne correspondent pas pleinement aux caractéristiques fines des gestes telles qu'elles ont été conceptualisées à l'origine (définition des catégories de McNeill) ou telles qu'elles sont habituellement décrites dans la littérature (propriétés formelles, distributionnelles, etc.). Cela suggère que l'enrichissement proposé n'a pas systématiquement intégré les critères de précision, de cohérence et de pertinence que l'on attendrait de ce type de ressources.

Dans une perspective interdisciplinaire de l'analyse du geste, l'harmonisation des pratiques et des protocoles est pourtant indispensable. Sans cadre méthodologique partagé, les ressources produites risquent d'être hétérogènes. Définir des standards communs, tant pour les procédures d'annotation que pour les formats et les critères d'enrichissement, permet non seulement d'assurer la cohérence des corpus et d'en garantir la réutilisation, mais aussi de s'accorder sur les objets étudiés et sur leur description. Cette harmonisation constitue ainsi une condition essentielle au dialogue scientifique et à garantir l'exploitabilité des corpus.

# Chapitre 5
# Protocole d'enrichissement de corpus de gestes

## 1 Introduction

Le chapitre précédent a permis de réaliser une première analyse de surface de l'enrichissement du corpus BEAT [150], mettant en évidence que les premiers indices de qualité étaient relativement faibles [153]. Dans ce contexte, il a été nécessaire de repenser les usages potentiels de ce corpus dans le cadre de cette thèse. Si les données primaires de BEAT semblaient adaptées à des traitements computationnels, les limites des annotations proposées en restreignaient toutefois l'exploitation rigoureuse, tant pour des analyses linguistiques que pour une intégration fiable dans des modèles automatiques. Ces constats ont motivé la création d'un protocole de réannotation transparent, rigoureux et systématique, visant à obtenir des données plus exhaustives, standardisées et fidèles à la gestualité observée, et solidement ancré dans l'expertise développée au sein des *gesture studies*. Dès lors, nous avons produit une nouvelle annotation d'un extrait du corpus BEAT, que nous avons nommée *reBeat-14*, et dont nous allons détailler les spécificités dans ce chapitre.

| **n°** | **Durée** | **n°** | **Durée** | **n°** | **Durée** | **n°** | **Durée** |
|---|---|---|---|---|---|---|---|
| 2 | 1 min 13 s | 43 | 54 s | 69 | 1 min 4 s | 105 | 46 s |
| 14 | 1 min 13 s | 53 | 38 s | 89 | 1 min 10 s | 107 | 44 s |
| 25 | 1 min 2 s | 60 | 55 s | 93 | 35 s | **N** | **Tps total** |
| 42 | 49 s | 64 | 1 min 3 s | 100 | 57 s | **14** | **13 min 3 s** |

*Tableau 7. Liste des fichiers réannotés et de leur durée*
*(n° = numéro de fichier ; N = nombre d'occurrences ; tps = temps)*

Deux annotatrices expertes, Domitille Caillat et moi-même, ont élaboré le protocole d'enrichissement présenté ci-après, puis ont réalisé les annotations multimodales conformément

à celui-ci. Cette réannotation a été menée sur quatorze fichiers du corpus BEAT (Tableau 7), représentant au total environ treize minutes de temps de parole (38 secondes pour le plus court et 1 minute 13 pour le plus long). Tous correspondent à des enregistrements du locuteur Wayne (natif américain), et consistent en des monologues. Ce corpus a été enrichi à partir des données primaires du corpus BEAT, à savoir les fichiers audio (`.wav`) pour la transcription de la parole et les vidéos de reconstruction des mouvements du locuteur issues de la capture de mouvement (`.bvh`) pour l'annotation des gestes. L'objectif de reBeat-14 a été de constituer un jeu de données restreint, mais fiable, destiné à pallier les limites de l'enrichissement d'origine et à rendre possible l'exploration de l'approche pluridisciplinaire défendue dans cette thèse, mêlant la robustesse d'annotations expertes de gestes aux approches computationnelles.

Ce chapitre est organisé en deux volets : il présente d'abord la méthodologie de transcription et d'annotation de la parole adoptée, puis celle de l'annotation de la gestualité, qui a été envisagée sur les plans fonctionnel et formel. Ce protocole s'inscrit dans une démarche évolutive, nourrie par des phases de réflexion et d'ajustement. À terme, ce travail a vocation à aboutir à la publication d'un manuel de codage détaillé ainsi qu'à la diffusion des données enrichies associées.

# 2 Transcription et annotation de la parole

Le protocole d'enrichissement de la parole a réuni plusieurs étapes consécutives. Le logiciel Praat [16] ainsi que des outils de traitement de texte classiques ont été utilisés pour transcrire les données audios sous forme écrite, allant de la transcription standardisée (avec ponctuation) à d'autres représentations diversement enrichies et balisées. Ce processus a conduit à la production d'un fichier `.TextGrid` contenant une suite de tiers, telle que détaillée dans le Tableau 8.

| Tier | Usage |
|---|---|
| `ortho_punct` | Transcription orthographique et ponctuée |
| `ortho_bal` | Transcription orthographique, ponctuée et balisée |
| `words` | Alignement temporel des mots |
| `pause` | Alignement temporel des pauses |
| `syll` | Alignement temporel des syllabes |
| `phones` | Alignement temporel des phones |
| `lemme` | Formes canoniques (lemmatisation) des mots |
| `pos` | Catégories morphosyntaxiques des mots |
| `dep` | Dépendance syntaxique des mots |
| `tag` | Traits morphologiques détaillés des mots |

*Tableau 8. Tiers définis dans le cadre de la transcription et de l'annotation de la parole*

## 2.1 Transcription orthographique

La transcription orthographique de chaque fichier audio du corpus a été générée automatiquement par un système de reconnaissance automatique de la parole (ASR). Le modèle

a pris en entrée un fichier `.wav` et a généré en sortie un fichier `.txt` contenant une transcription en anglais standard, sans ponctuation, sans capitalisation et sans métadonnées temporelles.

Pour ce faire, nous avons utilisé un modèle préentrainé et accessible en ligne ; le modèle *NVIDIA STT En Conformer-Transducer XLarge*. Ce modèle a une architecture *Conformer-Transducer* de grande taille (environ 644 millions de paramètres), combinant des mécanismes convolutifs et attentionnels afin de modéliser conjointement les dépendances locales et globales du signal acoustique. L'encodeur est constitué de 24 couches avec une dimension de représentation de 1024, 8 têtes d'attention et un noyau convolutionnel de taille 5. Le décodage s'appuie sur une architecture de type RNN-*Transducer*, avec un prédicteur comportant deux couches récurrentes et des dimensions cachées de taille 640. Le modèle a été entraîné à l'aide du *toolkit NeMo* sur le jeu de données *NeMo ASRSET.*

## 2.2 Révision lexicale

Nous avons ensuite procédé à une révision lexicale de la transcription générée par le modèle d'ASR, réalisée indépendamment et manuellement par chaque annotatrice dans un logiciel de traitement de texte. Une version définitive commune a ensuite été établie de manière collaborative à partir de ces révisions indépendantes. Ce travail a consisté à écouter le signal audio et à vérifier le texte généré par le système d'ASR, et a été mené en suivant des critères définis en interne pour l'homogénéisation de la transcription des fichiers : normalisation orthographique (ex. : nombres en toutes lettres, abréviations, majuscules, toponymes/anthroponymes), harmonisation de certains phénomènes oraux (ex. : formes des hésitations, marques d'élision), et correction des potentielles erreurs de reconnaissance automatique. Cette étape visait à obtenir une transcription fidèle à l'élaboration verbale du locuteur, en conservant les phénomènes oraux que l'ASR peut omettre (ex. : hésitations).

## 2.3 Révision syntaxique

Par la suite, nous avons procédé à un enrichissement typographique en ajoutant manuellement la ponctuation (virgules, points, points d'interrogation, etc.) dans un éditeur de texte afin de refléter une interprétation possible de la structuration syntaxique et prosodique du discours. L'ajout de la ponctuation a d'abord été réalisé de manière indépendante, puis harmonisé collectivement par les deux annotatrices pour la version finale (tier `ortho_punct`), en s'appuyant sur les indices prosodiques tels que l'intonation et les pauses. Cette étape offre la possibilité de réaliser par la suite des analyses au niveau de la phrase syntaxique.

## 2.4 Balisage

| **Balise encadrante** | | **Usage** |
|---|---|---|
| *Ouverture* | *Fermeture* | |
| <STAND> | <STAND/> | Standardisation orthographique |
| <ERR> | <ERR/> | Erreur, reprise ou autocorrection |
| <INTER> | <INTER/> | Interjections |
| <VOL_L> | <VOL_L/> | Baisse de volume |
| <VOL_H> | <VOL_H/> | Hausse de volume |
| <LAUGH> | <LAUGH/> | Rire confondu avec la parole |
| <ELO> | <ELO/> | Élongation |
| <ACC> | <ACC/> | Accentuation |
| **Sans encadrement** | | |
| </PAUSE/> | | Pause |
| </LAUGH/> | | Rire |

*Tableau 9. Jeu de balises pour l'annotation de phénomènes oraux*

Nous avons ensuite procédé à un enrichissement de la transcription orthographique ponctuée par l'annotation de phénomènes oraux sous forme de balises, incluant différents événements para/extralinguistiques. Ces balises ont été conçues en interne afin de coder de manière explicite, au sein de la transcription, des phénomènes produits par le locuteur en certains endroits de son énonciation. Chaque événement est ainsi reporté avec une typologie précise (Tableau 9) et intégré dans le texte exactement là où il se produit, offrant ainsi la possibilité de réaliser par la suite des analyses précises et détaillées de la synchronisation potentielle entre certaines occurrences de gestes et certains de ces phénomènes oraux. Cette étape a là encore été réalisée manuellement par les deux annotatrices, toujours dans un logiciel de traitement de texte, d'abord de manière indépendante et ensuite révisée collaborativement. Le résultat de cette étape correspond au tier `ortho_bal`.

## 2.5 Segmentation et alignement des mots et des syllabes

À partir de la transcription orthographique révisée et non ponctuée, chaque mot a été automatiquement aligné sur l'audio, et segmenté en phonèmes, également alignés sur le signal audio. Cette étape a été réalisée avec le *Montreal Forced Aligner* (MFA) [137]. Le modèle a pris en entrée chaque fichier `.txt` et a généré en sortie plusieurs `.TextGrid` référençant les silences, mots et phonèmes avec leurs temps de début et de fin. L'alignement proposé par MFA a ensuite été révisé d'abord individuellement puis collaborativement sur Praat, par les deux annotatrices. Les tiers `words`, `phones` et `pause` ont ainsi été obtenus.

La segmentation et l'alignement des mots en syllabes n'ayant pas pu être effectués automatiquement, nous avons dupliqué pour chaque mot ses temps de début et de fin dans un nouveau tier `syll`. Chaque annotatrice a ensuite inséré manuellement les timestamps à l'intérieur de l'intervalle du mot, permettant de segmenter celui-ci en syllabes alignées sur le signal audio. La vérification du tier a ensuite été faite de manière collaborative par les deux annotatrices.

## 2.6 Extractions de traits linguistiques

Enfin, nous avons appliqué un *pipeline* d'enrichissement linguistique (*spaCy* [138], modèle utilisé : `en_core_web_lg` disponible en ligne) adapté à la langue des données. Cette étape, basée sur le tier `ortho_punct`, a permis d'extraire pour chaque mot composant une phrase : les lemmes (tier `lemme`), les catégories morphosyntaxiques (tier `pos`), les traits morphologiques (tier `tag`), et les dépendances syntaxiques (tier `dep`). Ces traits ont été indexés sur les composants du tier `words` et ont hérité de leur ancrage temporel, en vue de faciliter la possibilité de mener des analyses multimodales alignées sur ces différents traits linguistiques.

# 3 Annotations gestuelles

## 3.1 Procédure générale d'annotation

Les deux annotatrices (Ann.) expertes ont élaboré le protocole d'annotations gestuelles présenté ci-après, en se reposant sur leur connaissance approfondie des travaux en *gesture studies*, tout en adaptant et remaniant les principes au besoin pour les rendre compatibles avec les objectifs et le contexte de ce travail.

| **n°** | **Durée** | **Temps d'annotation** | |
|---|---|---|---|
| | | *Ann. 1* | *Ann. 2* |
| 53 | 38 s | 2,5 jours | |
| 64 | 1 min 3 s | 11 h | |
| 89 | 44 s | 1 h 23 | 2 h 5 |
| 107 | 44 s | 40 min | 1 h 14 |
| **Total** | **3 min 9 s** | **3 j 1 h 3 min** | **3 j 2 h 19 min** |

*Tableau 10. Temps d'annotation pour chaque fichier pendant la phase de synchronisation (n° = numéro de fichier ; ann = annotatrice)*

Dans un premier temps, une phase de synchronisation a été mise en place pour garantir une compréhension partagée et une application homogène du protocole, dans le but de maximiser la fiabilité des annotations. L'objectif était alors de tester et d'affiner la méthodologie de segmentation des gestes et les critères de catégorisation fonctionnelle. Quatre fichiers ont été annotés de manière collaborative par les annotatrices, et sont désignés comme « fichiers de synchronisation » (fichiers n° 53-64-89-107 pour un total d'environ 3 minutes d'enregistrement). En pratique, cette phase, qui s'est déroulée sur 4 jours en présentiel, a consisté en plusieurs sessions de travail. Dans un premier temps, une annotation commune a été réalisée sur deux fichiers pendant les trois premiers jours (Tableau 10, fichiers 53 et 64, annotés intégralement par les deux annotatrices). Elle a ensuite été suivie d'une phase d'annotation indépendante pour les deux autres fichiers, menée de manière itérative : chaque annotatrice annotait une courte séquence, puis les annotations étaient confrontées et discutées, avant de passer à la séquence suivante.

| **Fichiers** N = 10 Durée : 9 min 28 s | **Temps d'annotation** | |
|---|---|---|
| | *Ann. 1* | *Ann. 2* |
| *Durée totale* | 8 h 50 | 14 h 40 |
| *Moyenne* | 53 min | 1 h 28 |
| *Médiane* | 49 min | 1 h 31 |
| *Ecart-type* | 18 min | 29 min |
| **Tps moyen/min** | **55,1 min ±12** | **91 min ±21** |

*Tableau 11. Analyse du temps d'annotation pendant la phase d'annotation indépendante (N = nombre d'occurrences ; ann = annotatrice ; tps = temps)*

Le reste des annotations a été réalisé de manière indépendante, les deux annotatrices ayant annoté individuellement l'ensemble des fichiers restants de *reBeat-14*. La durée totale de ces 10 fichiers est d'environ 10 minutes. D'un point de vue pratique, le temps d'annotation a varié entre 9 et 15 heures selon l'annotatrice (Tableau 11). Le temps moyen passé à annoter par minute de vidéo illustre l'intensité du travail : pour l'annotatrice 1, chaque minute de vidéo a requis en moyenne 55 min d'annotation (±12 min), tandis que pour l'annotatrice 2, ce temps a été plus élevé, avec 91 min par minute de vidéo (±21 min).

Le Kappa de Cohen [160] a été calculé sur l'ensemble des fichiers annotés de manière indépendante. Chaque fichier a été converti en vecteur binaire par *frame* (0 = absence de geste, 1 = présence de geste) avec un taux de 120 fps, puis comparé *frame* par *frame*. Le Kappa obtenu est 0,6, ce qui correspond à un accord modéré à bon selon les conventions usuelles. Cela indique que les annotatrices sont généralement d'accord sur leur segmentation de gestes, bien au-delà du hasard, même si certaines différences subsistent sur certaines *frames*.

Sur la base des annotations réalisées de manière indépendantes, l'ensemble des fichiers a été discuté et révisé collectivement lors de la phase de révision commune. Cette étape a permis de résoudre les divergences entre annotatrices, d'ajuster les conventions, d'homogénéiser l'enrichissement des fichiers, et de garantir la cohérence des choix méthodologiques. Elle a également permis de revenir sur la logique de certaines pratiques d'annotation telles qu'elles avaient été initialement conçues. Bien que chronophage, cette étape a ainsi joué un rôle décisif dans la définition du schéma d'annotation : certaines catégories ont été établies ou affinées au cours de ces discussions, nécessitant parfois des retouches des annotations initiales, en fonction des difficultés rencontrées et des ajustements nécessaires pour parvenir à une annotation commune.

## 3.2 Segmentation du flux de mouvement

Le premier type d'annotation a concerné la segmentation manuelle du flux de mouvements du locuteur, sur ELAN [15]. Dans l'interface d'annotation, nous disposions de la séquence de vidéo de synthèse (reconstruction de la *motion capture*, offrant un aperçu suffisamment précis de la gestuelle), des tiers issus de la transcription et de l'annotation de la parole décrites précédemment, ainsi que des tiers destinés à l'annotation des gestes. Le tier `activity` a constitué le tier parent hiérarchique de tous les autres tiers dédiés à l'annotation

gestuelle. Initialement, ce tier était destiné à segmenter les unités gestuelles dans le corpus, mais il a finalement servi à baliser les occurrences des potentiels gestes (plus spécifiquement l'occurrence de ce qui constitueraient une phase *stroke*) au cours du flot de gesticulations du locuteur. Sur ce tier, les gestes potentiels sont appelés « activités », en attendant que leur statut de geste ne soit réellement confirmé (ou non) sur un des tiers enfants, le tier « type ». Le processus de segmentation s'est appuyé sur des critères à la fois cinématiques [134, 133] (changements perceptifs d'accélération, de direction, ou pic de vitesse) et contextuels (interprétation en lien avec le discours ou les gestes environnants). Au total, 492 activités gestuelles ont été segmentés dans le corpus.

## 3.3 Annotation fonctionnelle de la gestualité

tier parent
↳tier enfant

| Tier | Usage |
|---|---|
| `phase` | Catégorisation de la phase gestuelle |
| `type` | Type de mouvement segmenté |
| ↳`is_communicative_d1` | Type fonctionnel du geste en D1 |
| ↳certainty_score_d1 | Score de certitude en D1 |
| ↳`is_communicative_d2` | Type fonctionnel du geste en D2 |
| ↳certainty_score_d2 | Score de certitude en D1 |

*Tableau 12. Tiers d'annotation gestuelle et leur usage (D1 = dimension primaire ; D2 = dimension secondaire)*

### 3.3.1 Caractérisation fonctionnelle de l'activité gestuelle

La labellisation des activités gestuelles a été ensuite réalisée dans les tiers `phase` et `type` (Tableau 12). Le premier de ces deux tiers était dédié à l'identification de la phase de réalisation du geste. Étant donné que les balisages réalisés dans le tier `activity` correspondaient dès le départ aux limites du segment interprété comme pouvant être le cœur du geste potentiellement identifié, seul l'étiquette *stroke* a été utilisée dans le tier `phase`. Le second tier permettait de qualifier la nature du mouvement segmenté. Le label `communicatif` a alors servi à qualifier les mouvements interprétés comme remplissant bel et bien un rôle communicatif (= soit les activités ayant réellement un statut de geste), là où le label `undefined` a servi à indiquer lorsqu'un doute était partagé entre les annotatrices quant à la nature réellement communicative du mouvement segmenté sur la ligne `activity`. Parmi les 492 activités gestuelles annotées, 479 ont été identifiées sans ambiguïté quant à leur nature communicative.

### 3.3.2 Caractérisation du rôle fonctionnel du geste

Nous avons ensuite étiqueté les rôles fonctionnels de chaque geste, c'est-à-dire de chaque activité jugée communicative, sachant que certains gestes pouvaient cumuler deux rôles. La première dimension fonctionnelle (D1) correspond au rôle fonctionnel principal du geste, tandis que la seconde dimension fonctionnelle (D2) correspond au rôle fonctionnel complémentaire, jugée moins primordiale. Les tiers `is_communicative_d1` et `is_communicative_d2` ont respectivement servi à labelliser ces fonctions (Tableau 12). Tous les gestes ont à minima été

annoté selon une D1. Afin de mieux rendre compte de la diversité des fonctions gestuelles rencontrées dans le corpus, nous avons fait le choix de regrouper des labels provenant de typologies différentes, et, lorsque nécessaire, d'introduire de nouveaux labels. Cette approche visait à offrir une description fonctionnelle plus flexible et représentative des productions gestuelles observées. Les labels fonctionnels utilisés dans le cadre de cette réannotation sont les suivants :

**Iconique** – Gestes dont la forme représente directement un aspect du référent évoqué dans le discours. Cette catégorie, issue de McNeill [9, 10], se décline en quatre sous-labels, dont trois sont des sous-spécifications proposées par Cosnier et Vaysse [64] et Ekman et Friesen [41], tandis que le dernier consiste en une variante que nous avons introduite :

- `Spatiographique` [64] : représente une configuration spatiale (distances, positions d'objets).
- `Pictographique` [64] : dessine ou mime l'apparence visuelle (ou l'une des caractéristiques) d'un référent.
- `Kinémimique` [41, 64] : imite la manière dont un être animé réalise une action (ex. : action de boire de l'eau).
- `Kinétographique` : reproduit la dynamique ou la trajectoire d'un mouvement. Nous avons intégré cette catégorie afin de distinguer clairement la représentation de la trajectoire d'un mouvement (ex. : trajectoire d'un saut) d'une imitation de ce mouvement (ex : sauter), consistant en un geste kinémimique.

**Déictique** [9, 10, 64] – Gestes dont la fonction principale est de désigner une entité, un lieu, un moment ou une personne. Nous avons jugé pertinent ici de distinguer 3 cas de figure :

- `Spatial` [64] : pointe un lieu ou une direction dans l'espace gestuel.
- `Désignation` [9, 10, 64] : indique une chose, une personne (dont le locuteur lui-même : ex. pointer sa poitrine).
- `Temporel` [64] : désigne un moment métaphoriquement spatialisé.

**Métaphorique** [9, 10] – Gestes qui représentent des concepts abstraits pour illustrer un référent lexical, sans être nécessairement lié à une métaphore verbale (ex. : représenter une sorte de mouvement de rouages pour illustrer l'idée de réfléchir).

**Modal** [8] – Gestes qui expriment ou accompagnent une modalité du discours, par le renfort ou le marquage d'une attitude liée à l'énonciation. Cette catégorie regroupe ainsi les gestes renseignant la manière dont le locuteur se positionne dans et vis-à-vis de son énoncé, qu'il s'agisse d'exprimer la négation, l'affirmation, l'incertitude, l'insistance, l'atténuation ou d'autres modalités énonciatives. Dans la classification adoptée ici, les gestes modaux comprennent les gestes traditionnellement classés en gestes « de battements », tels que décrit par McNeill [9, 10], en ce que leur rôle consiste à figurer l'insistance.

**Parsing** [8] – Gestes utilisés pour souligner la structure du discours, qu'il s'agisse de le segmenter, de structurer l'information (contraste, énumération...) ou de marquer une transition. Ils agissent comme des ponctuants discursifs visuels.

**Butterworth** [161] – Gestes produits lors d'hésitations ou de difficultés de formulation (automanipulations, mouvements répétitifs des doigts, etc.). Ils signalent souvent la recherche lexicale ou un moment de planification verbale.

**Phatique** [64] – Gestes dédiés à la gestion de l'interaction, au maintien de l'engagement du partenaire, à la gestion des tours de parole, ou à la production de *feedbacks* (ex. petits hochements ou microgestes confirmatifs).

**Quasi linguistique** [41, 64] – Gestes présentant une forme conventionnelle et fonctionnant comme un signe quasi lexical (ex. : geste de « stop », « viens », hochement signifiant « oui », mains en prière pour « s'il te plaît »). Ils peuvent être communément compréhensibles sans le verbal.

**Undefined** – Gestes dont la fonction ne peut être déterminée avec certitude à partir du contexte, de la forme, ou du lien au discours. Cette catégorie, ajoutée dans le cadre de ce travail, est utilisée lorsqu'un doute subsiste entre les annotateurs : bien que le statut de geste soit clairement reconnu, son appartenance à une fonction clairement identifiable reste problématique.

### 3.3.3 Affiliation lexicale, prosodique et gestuelle

Les gestes ont ensuite été reliés, lorsque possible et pertinent, au niveau verbal et prosodique. Une éventuelle affiliation lexicale [65] a ainsi parfois été identifiée et annotée manuellement, correspondant au mot ou à l'expression du discours auquel un geste semble associé du point de vue de sa fonction. De même, certaines affiliations prosodiques [66] ont pu être annotées, par le repérage de certaines coïncidences avec des phénomènes prosodiques (intonation, accentuation, pauses) lorsque saillantes. Enfin, l'affiliation gestuelle, qui correspond à une production gestuelle contextuellement liée à un ou plusieurs autres gestes, a aussi été annotée manuellement. Cette étape a donné lieu à des annotations permettant de mettre en évidence les différents types d'associations semblant exister entre la parole et le geste.

### 3.3.4 Caractérisation de la certitude du choix des types fonctionnels

Un score de certitude a été associé à chaque type fonctionnel annoté (en D1, et le cas échéant, en D2) afin de rendre compte du degré de confiance des annotatrices quant à l'interprétation fonctionnelle du geste. Ce score, défini sur une échelle ordinale à six niveaux allant de 0 (absence totale de certitude) à 5 (certitude absolue), consiste en un score commun déterminé sur la base des scores individuels fournis par les deux annotatrices. Les niveaux intermédiaires correspondent à des degrés croissants de confiance : 1 indique une très faible certitude, 2 une faible certitude, 3 une certitude modérée et 4 une forte certitude. L'introduction de ce score vise à renseigner le degré d'incertitude interprétative inhérente à l'annotation gestuelle et à permettre des analyses tenant compte du degré de fiabilité perçue des annotations.

## 3.4 Annotation formelle de la gestualité

L'annotation formelle de chaque geste rend compte des propriétés physiques et visuelles perçues par les annotatrices. Elles incluent la configuration de la main (forme des doigts), l'orientation de la paume, la position de la main dans l'espace du locuteur, la trajectoire du mouvement réalisé et la manualité. Cette description systématique, réalisée manuellement, avait pour objectif de pouvoir élaborer une description précise et détaillée des formes gestuelles.

### 3.4.1 Catalogue de trajectoires

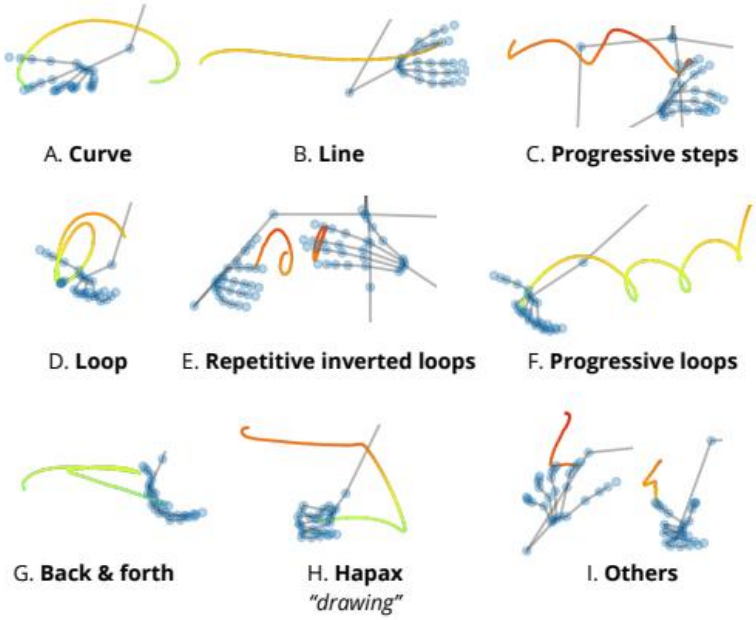



*Figure 16. Formes de trajectoires identifiées dans le corpus*

Les trajectoires des gestes co-verbaux ont été annotées selon des schèmes formels déterminés après une première observation des gestes du corpus, et sur la base de la reconstruction numérique des trajectoires des segments annotés. Elles se répartissent selon neufs catégories (Figure 16) : *curve*, *line*, *progressives steps*, *loop*, *repetitive inversed loops*, *progressive loops*, *back and forth*, *Hapax* et *autres*. Le schème *curve* (Figure 16A) désigne une trajectoire courbée, tandis que le schème *line* (Figure 16B) correspond à un déplacement rectiligne de la main dans l'espace. Le schème *progressive steps*, ou PS (Figure 16C), correspond à une progression scandée. Le schème *loop* (Figure 16D) caractérise une trajectoire circulaire ou elliptique fermée, et le schème *repetitive inversed loop*s, ou RIL (Figure 16E), décrit des boucles successives effectuées dans des directions inversées (« moulinet »). Le schème *progressive loops, ou* PL (Figure 16F), représente une série de boucles successives accompagnées d'un déplacement progressif. Le schème *back and forth, ou* BF (Figure 16G), décrit un mouvement alterné entre deux directions opposées. Enfin, l'étiquette *Hapax* (Figure 16H) désigne un mouvement unique, isolé et non répété, tandis que l'étiquette *autres* concerne des gestes qui n'ont pas de schèmes identifiés mais qui ne sont pas des mouvements isolés.

### 3.4.2 Labellisation de la manualité

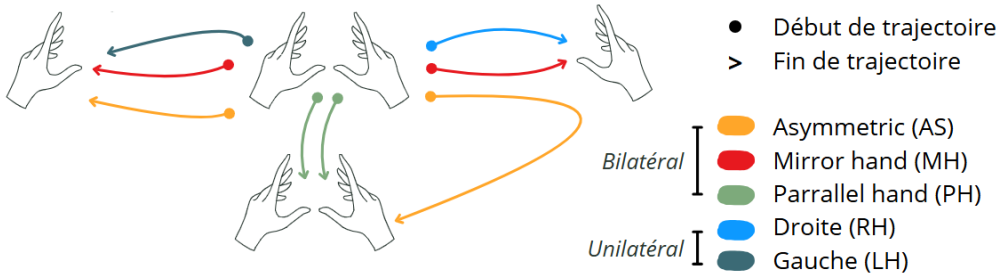



*Figure 17. Types de manualités annotées, exemples théoriques de trajectoires*

La manualité des gestes a été catégorisée en distinguant les productions bimanuelles et monomanuelles (Figure 17). Les gestes bimanuels (BH) ont été classés selon trois étiquettes : *asymmetric* (AS), lorsque les deux mains accomplissent des mouvements différents et complémentaires (par exemple représenter un « T » avec ses deux mains) ; *mirror hands* (MH), lorsque les deux mains réalisent des mouvements symétriques en miroir (ex. : dessiner un rond) ; et *parallel hands* (PH), lorsque les mouvements sont similaires et orientés dans la même direction (ex. : un glissement des deux mains vers la droite). Les gestes monomanuels, quant à eux, ont été différenciés selon la main mobilisée, droite (RH) ou gauche (LH).

### 3.4.3 Labellisation de la position de la main

La position de la main dans l'espace, basée sur l'observation de la dernière *frame* du geste segmenté, est décrite de manière simplifiée selon trois orientations principales. Ces orientations permettent de caractériser la direction dominante du geste par rapport à l'axe du corps et à l'espace environnant. La position horizontale (H) correspond à une main disposée parallèlement au sol, la position verticale (V) traduit une orientation perpendiculaire au sol, tandis que la position intermédiaire (I) recouvre les configurations situées entre ces deux extrêmes.

### 3.4.4 Labellisation de l'orientation de la paume

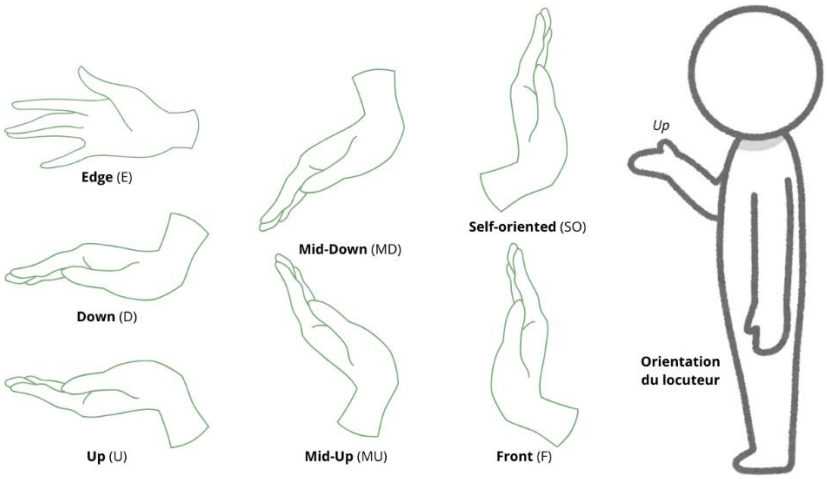



*Figure 18. Exemples illustrés d'orientations de la paume*

L'orientation de la paume de la main a également été systématiquement annotée (Figure 18) sur la base de la dernière *frame* du geste. Sept orientations ont été distinguées : *edge* (E), lorsque la paume est présentée sur la tranche ; *mid-up* (MU), orientée partiellement vers le haut ; *self-oriented* (SO), dirigée vers le corps du locuteur ; *up* (U), tournée pleinement vers le haut ; *mid-down* (MD), correspondant à une orientation intermédiaire vers le bas ; *front* (F), face à l'interlocuteur ; et *down* (D), orientée vers le bas. Cette catégorisation permet de caractériser précisément la manière dont la main est exposée dans l'espace.

### 3.4.5 Labellisation de la configuration de la main

La configuration de la main a été décrite sur la base de la disposition des doigts à la dernière *frame* d'un geste. Neuf configurations distinctes ont été annotées : *open hand* (OH), lorsque la main est complètement ouverte ; *semi-open hand* (SOH), pour une ouverture partielle ; *index engaged* (IE), lorsque l'index est mis en avant ; *mid-closed hand* (MCH), correspondant à une fermeture intermédiaire de la main ; *thumb middle index engaged* (TMIE) et *thumb engaged* (TE), impliquant spécifiquement le pouce (soit seul, soit avec le majeur et l'index) ; *closed hand* (CH), lorsque la main est fermée ; ainsi que deux formes intermédiaires, *semi-engaged index* (SEI) et *thumb index engaged* (TIE).

## 4 Conclusion de chapitre

Dans une approche résolument fondée sur les données multimodales, les choix méthodologiques opérés lors de l'annotation ont un impact direct sur la nature et la qualité des données enrichies produites. Les annotations produites lors de ce réenrichissement d'un extrait de BEAT se distinguent alors par leur granularité fine et la richesse des catégories mobilisées, qui visent un alignement étroit avec les descriptions issues de la littérature sur la gestualité co-verbale. Ce protocole s'inscrit ainsi pleinement dans l'expertise des *gesture studies*, en s'appuyant sur leurs cadres théoriques et méthodologiques bien établis. À cet égard, cet enrichissement prolonge les usages et pratiques d'annotation classiques observés dans les corpus de gestes, ce qui constitue un premier gage de la fiabilité du schéma retenu. Un protocole clairement défini, fondé sur des critères explicites et validés, garantit de fait la cohérence, la reproductibilité et la précision des annotations produites.

Parmi ces ajustements, les différentes catégories formelles et fonctionnelles ont été étendues. L'annotation fonctionnelle ne se limite pas à l'utilisation des catégories proposées par McNeill. Cette approche vise en effet à mieux caractériser la diversité des fonctions que peuvent remplir les gestes, offrant ainsi une représentation plus précise que la catégorisation fonctionnelle classique (gestes *iconiques, métaphoriques, déictiques, battements*) souvent utilisée dans les corpus de gestes, dont il peut résulter un écrasement du sémantisme des occurrences étudiées. De la même manière, les descriptions formelles ont été conçues pour être aussi exhaustives que possible, afin de couvrir l'ensemble des aspects formels pertinents du geste. Ces représentations détaillées visent à être exploitables dans le cadre de modélisations computationnelles et d'apprentissages automatiques. Un tel enrichissement s'inscrit donc résolument dans la démarche défendue dans cette thèse, visant à développer des modèles

informatiques explicitement informés de propriétés gestuelles pertinentes et précises. L'intégration d'un score de certitude quantifie le degré de confiance dans l'interprétation fonctionnelle de chaque dimension du geste, offrant ainsi plusieurs avantages : il permet de différencier les annotations plus fiables de celles plus incertaines, d'orienter les vérifications et relectures, et de pondérer les analyses ultérieures en fonction de la certitude quant à l'interprétation fonctionnelle du geste. Ce score pourra également être exploité comme un paramètre de pondération lors de l'apprentissage, en modulant l'influence relative des occurrences selon leur fiabilité estimée. Il offrira ainsi un levier de régulation du modèle, permettant d'ajuster dynamiquement l'équilibre entre robustesse et couverture des données.

Le protocole présenté ici, sur lequel se fonde les analyses des parties suivantes, reste toutefois un travail en cours, et la version exposée dans ce chapitre demeure préliminaire. À ce stade en effet, l'annotation formelle, par exemple, repose principalement sur la dernière *frame* du geste. Ce choix méthodologique, adopté pour faciliter la caractérisation de la forme du geste, ne permet toutefois pas de rendre compte pleinement de sa dynamique. La méthodologie permettant de capturer de manière cohérente la dynamique formelle des gestes reste à concevoir et à formaliser, dans le but d'obtenir une description fine des gestes tels qu'ils sont effectivement produits au cours du temps. De telles représentations spatio-temporelles du mouvement seraient d'autant plus pertinentes qu'elles pourraient potentiellement améliorer les performances des modèles d'apprentissage, dans des tâches de prédiction ou de génération automatique de gestes. Il convient en outre de souligner que ces annotations formelles, basées sur la perception de la vidéo 2D (bien que les mouvements soient issus de la *mocap* initialement), présentent certaines limites. Elles restent en effet très subjectives et peuvent être influencées par la position du locuteur dans l'espace : une variation de perspective peut en effet modifier la perception du geste par l'annotateur, et dès lors introduire un biais dans le choix des catégorisations formelles et affecter la cohérence des annotations. Cette subjectivité constitue ainsi une limite inhérente à l'annotation manuelle réalisée à partir de vidéos, qui doit être prise en compte dans l'interprétation des résultats.

L'annotation fonctionnelle du geste présente également un certain nombre de limites. Malgré la volonté de proposer un système de catégorisation intégratif, fondé sur la combinaison de plusieurs typologies fonctionnelles issues de la littérature, le dispositif retenu s'est révélé, au fil de l'annotation, insuffisamment discriminant dans certains cas. La nécessité d'introduire un label *Undefined* en constitue un indicateur méthodologique : certains gestes, bien que manifestement communicatifs, ne pouvaient être assignés de manière satisfaisante à un type fonctionnel. La difficulté rencontrée renvoie plus largement à l'absence de consensus au sein du champ des *gesture studies*, où les typologies fonctionnelles et leurs critères d'application varient sensiblement d'une étude à l'autre. Dans ce contexte, un même geste peut être catégorisé différemment selon le cadre théorique mobilisé. Il convient toutefois de nuancer ce constat : selon les objectifs analytiques poursuivis et le niveau de granularité recherché, ces catégories peuvent s'avérer pleinement satisfaisantes et opératoires. Il n'en demeure pas moins nécessaire de poursuivre la réflexion sur la définition et l'opérationnalisation des catégories fonctionnelles, afin de favoriser une annotation comparable.

Le corpus reBeat-14, qui comprend donc environ 13 minutes de données vidéo, représente toutefois en l'état un volume conséquent au regard du niveau de granularité des annotations réalisées et du temps nécessaire à leur production. Ce choix reflète un compromis assumé entre quantité et qualité : une annotation fine, et multidimensionnelle implique nécessairement une réduction de la durée totale annotée. Il convient cependant de rester conscient que ce jeu de données est relativement petit, et ne reflète donc pas la diversité des gestes possibles ou des styles propres à chaque locuteur. Certaines catégories de gestes peuvent être particulièrement sous-représentées, ce qui limite l'exhaustivité des observations et doit tempérer l'interprétation des résultats. Malgré ces limites, il est important de souligner que ce travail repose sur un corpus effectivement constitué et exploitable. Il permet une première exploration approfondie des caractéristiques des gestes, aux plans de leurs fonctions, de leurs formes, et de leurs liens avec le discours, À ce titre, il constitue un terrain expérimental pertinent, servant à la fois de base d'analyse descriptive et de point d'ancrage méthodologique pour les recherches présentées dans la suite de cette thèse, qu'il s'agisse de l'exploration des propriétés du geste, de la conception et de l'évaluation de modèles, ainsi qu'une première évaluation de la faisabilité et des limites d'approches d'annotation automatique ou semi-automatique.

# Bilan de partie

Disposer d'un corpus de gestes constitue une condition nécessaire à la production de données exploitables, sur lesquelles peuvent s'appuyer aussi bien les analyses multimodales que les modélisations computationnelles. Le corpus ne se réduit pas à un simple ensemble de données : il résulte d'un processus de conception structuré, impliquant des choix méthodologiques, techniques et théoriques à chaque étape, allant de la sélection des participants au schéma d'annotations. Les modalités retenues orientent ainsi les objets mêmes de la modélisation : elles influencent ce qui pourra être prédit, formalisé par le modèle, et les possibilités de génération. Ces décisions initiales influencent directement la qualité, la cohérence et la représentativité des données produites, et déterminent ainsi la portée des analyses possibles ainsi que la validité des modèles développés à partir de celles-ci.

Il est donc essentiel de concevoir un corpus à la fois fiable et adapté aux objectifs d'étude poursuivis. La fiabilité renvoie à la qualité intrinsèque des données : celles-ci doivent restituer avec précision les phénomènes étudiés. Dans le cas des gestes co-verbaux, cela implique une captation fidèle du mouvement, une résolution suffisante pour en analyser les paramètres pertinents, ainsi qu'une synchronisation rigoureuse avec les autres modalités, notamment la parole, tout en étant exempte d'artefacts et suffisamment riche pour refléter la variabilité naturelle des gestes dans différents contextes communicatifs. L'adéquation, quant à elle, concerne la finalité et l'exploitation du corpus. Les données doivent être structurées, annotées et organisées pour permettre leur utilisation dans des analyses qualitatives, des comparaisons quantitatives ou des modélisations computationnelles. L'enrichissement et l'annotation jouent ici un rôle central : un protocole clairement défini, aligné sur les connaissances issues des *gesture studies*, garantit la cohérence des descriptions tant sur le plan formel que fonctionnel, assure ainsi leur compatibilité avec des usages computationnels. La multidimensionnalité des descriptions (forme, fonction, contexte) permet de rendre compte de l'ensemble des aspects saillants et pertinents du geste. Par ailleurs, une documentation complète et transparente des méthodes, des métadonnées et des choix méthodologiques contribue à la fois à la fiabilité et à la réutilisabilité du corpus dans différents contextes.

Le protocole d'enrichissement de BEAT, pris ici comme cas d'étude, apparaît peu conforme aux bonnes pratiques établies dans les *gesture studies*. Les annotations produites présentent des lacunes de cohérence et de précision, et reflètent de manière limitée la réalité

des gestes co-verbaux, restreignant ainsi l'exploitation du corpus. Les auteurs, principalement orientés vers une approche computationnelle et l'animation 3D, semblent moins familiers avec les méthodologies issues des sciences du langage et les pratiques d'enrichissement, ce qui peut expliquer certaines limites observées dans la qualité et la pertinence des annotations. La stratégie adoptée, qui privilégie la quantité au détriment de la qualité en recourant à des annotateurs non experts via une plateforme de *crowdsourcing* pour traiter un corpus de grande taille, illustre la tension classique entre volume et rigueur. Plus le volume du corpus est important, plus il devient difficile de mobiliser exclusivement des annotateurs experts pour réaliser des annotations. Dans ce contexte, une approche d'enrichissement basée sur le *crowdsourcing* peut apparaître comme une solution pour traiter de grands ensembles de données. Cependant, si cette approche peut sembler pertinente sur le plan théorique, elle se révèle en pratique peu fiable et difficilement exploitable, notamment pour des tâches complexes. L'annotation multimodale des gestes co-verbaux est en effet une tâche complexe, même pour des annotateurs expérimentés : elle nécessite de segmenter correctement les gestes, d'en identifier les phases et d'interpréter leur fonction dans le contexte discursif. Pour des annotateurs non experts, ces tâches peuvent devenir particulièrement délicates et comportent un risque élevé d'erreurs, de biais ou de décisions arbitraires. Dans ce contexte, il est indispensable de prévoir un processus de vérification et de validation par des experts, afin de garantir la cohérence et la fiabilité des annotations. Or, dans le cas de BEAT, il n'est pas clairement indiqué qu'une telle validation ait été systématiquement réalisée. L'enrichissement de BEAT, dans son état actuel, présente un risque d'introduire des biais ou des erreurs plus importantes que la valeur analytique ou computationnelle qu'il pourrait apporter. En conséquence, il est préférable de ne pas s'appuyer sur cet enrichissement pour des analyses ou des modèles, afin d'éviter des conclusions potentiellement erronées.

Plutôt que de se limiter soit à des annotations multimodales entièrement réalisées par des experts ou uniquement produites par des non-experts, il est possible d'adopter une approche intermédiaire. Les tâches moins exigeantes sur le plan de l'expertise, telle que la description de mouvements, peuvent être confiées à des annotateurs non experts, tandis que les tâches complexes, comme la segmentation des gestes et l'interprétation de leur fonction, sont réalisées par des experts sur un sous-ensemble représentatif du corpus. Le reste des données peut ensuite être enrichi de manière automatique ou semi-automatique, suivi d'une vérification par des experts pour garantir la cohérence et la fiabilité des annotations. Dans ce cadre, l'échantillon initialement annoté par les experts (couvrant différents types de gestes, profils de locuteurs ou contextes communicatifs) constitue une référence fiable, utilisée pour entraîner un modèle d'apprentissage supervisé capable de générer des annotations initiales sur le reste du corpus. Les experts peuvent alors valider ou corriger ces nouvelles annotations, réduisant la charge de travail tout en maintenant la cohérence de l'enrichissement proposé dans le corpus. Cette stratégie offre un compromis efficace entre qualité et quantité, permettant de tirer parti de la richesse des données tout en garantissant la fiabilité des annotations.

C'est précisément dans ce contexte que l'approche interdisciplinaire défendue ici prend tout son sens. Dès la conception du corpus, la collaboration entre spécialistes des sciences du

langage et chercheurs en informatique permet de concilier les exigences méthodologiques liées à l'étude fine des gestes et les contraintes techniques ou computationnelles. Les experts de la gestualité apportent leur connaissance des pratiques d'annotation et d'analyse multimodale du geste, tandis que les informaticiens développent des méthodes et paradigmes qui facilitent le traitement et l'analyse des données, en formalisant des relations complexes difficilement accessibles par une observation directe. Dans le cadre de l'enrichissement automatique de corpus, cette synergie permet d'envisager des enrichissements sur de plus grand ensemble de données tout en maintenant la cohérence des annotations multimodales. Disposer de corpus étendus et rigoureusement annotés offre la possibilité de réaliser des analyses statistiques robustes, de former des modèles d'apprentissage plus performants et de développer des outils computationnels capables de capturer et de représenter la complexité des gestes et de leurs relations avec le discours. En d'autres termes, l'approche interdisciplinaire, notamment sur le plan méthodologique, ouvre la voie à une modélisation du geste à grande échelle, capable de prendre en compte ses variantes, son articulation au discours et sa complexité fonctionnelle. Une telle modélisation dépasse les limites d'une approche strictement humaine et nécessite que les systèmes d'apprentissage automatique soient intelligemment nourris par des données rigoureusement conçues et annotées.

# Partie III
# Apports et perspectives de l'expertise des sciences du langage

Cette partie examine les contributions que l'expertise en sciences du langage peut apporter à l'approche interdisciplinaire défendue dans cette thèse. Elle vise à répondre à la troisième question de recherche qui structure ce manuscrit, question portant sur la réflexion d'en quoi les pratiques d'annotation multimodales en sciences du langage peuvent constituer un levier méthodologique structurant pour les approches computationnelles de la modélisation du geste co-verbal.

Deux chapitres composent cette partie. Le premier est consacré à l'analyse du corpus reBeat-14 : il présente des statistiques descriptives sur l'enrichissement afin de proposer une première vue d'ensemble de la composition du corpus. Le second chapitre se concentre sur une analyse comparative des annotations multimodales, portant d'une part sur un fichier issu de reBeat-14 (locuteur Wayne) et, d'autre part, ce même fichier mais dont le flux de mouvement a été produit par un modèle de génération de gestes.

Le défi central de cette partie consiste à montrer que l'annotation multimodale – en particulier celle réalisée par des annotateurs dits « experts » de la gestualité – constitue déjà une première forme de modélisation, en ce qu'elle repose sur une formalisation systématique des phénomènes observés et peut ainsi servir à la caractérisation du geste, qu'il soit naturel ou de synthèse.

# Chapitre 6
# L'annotation multimodale comme outil de modélisation du geste co-verbal

## 1 Introduction

Dans le cadre d'une analyse fondée sur l'observation de données, l'impact des pratiques et méthodologies d'enrichissement du corpus d'étude est loin d'être marginal. La manière dont on annote conditionne directement ce que l'on est capable d'observer, donc de mesurer. À ce titre, et comme le soutient cette thèse, l'annotation multimodale façonne non seulement la manière dont on appréhende la description d'un geste, mais aussi les possibilités de modélisation et applications computationnelles qui en découlent. C'est sur la base de ce postulat qu'a été conçu le protocole de réannotation des fichiers de reBeat-14. Le nouvel enrichissement réalisé avait ainsi pour objectif d'identifier avec précision les gestes du corpus, et d'annoter de manière fiable et cohérente leurs propriétés formelles et fonctionnelles, afin de pouvoir en proposer une caractérisation fine et structurée.

Fondé sur l'examen des 479 gestes annotés dans re-Beat14, ce chapitre a pour objectif de mettre en lumière les types de modélisation qu'il est possible de dériver d'une annotation multimodale de qualité, tel qu'elle est pratiquée dans les *gesture studies*. Il s'organise en cinq volets d'analyse successifs : (i) statistiques générales, (ii) analyse fonctionnelle des gestes annotés, (iii) analyse formelle des gestes annotés, (iv) analyse cinématique, et, enfin (v) analyse de l'ancrage des gestes dans le discours. En dehors de méthodes spécifiques à l'étude de certains aspects (telles que les calculs de vitesse et d'amplitude), la méthodologie d'analyse des annotations du corpus repose principalement sur des mesures quantitatives descriptives classiques, telles que le décompte d'effectifs (N), les fréquences relatives, etc. Ces indicateurs permettent d'identifier des tendances dans les distributions d'étiquettes et de mener certaines comparaisons entre les gestes observés.

Cette approche, volontairement exploratoire, vise avant tout à fournir un aperçu des informations rendues accessibles par l'enrichissement multimodal du corpus. Ce chapitre vise ainsi à démontrer qu'une annotation multimodale de qualité constitue en elle-même un véritable outil de modélisation : elle encode des hypothèses théoriques, délimite précisément les aspects d'un geste et en propose une lecture contextualisée.

# 2 Statistiques générales

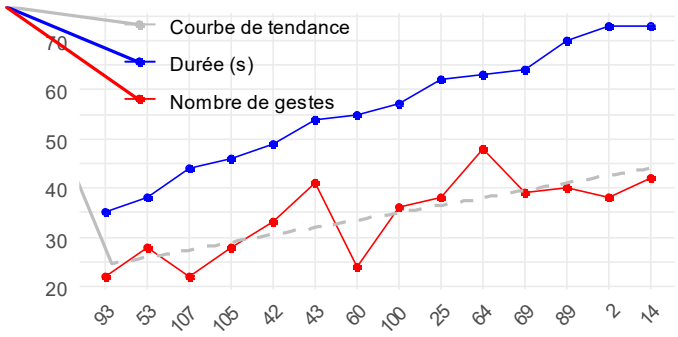



*Figure 19. Nombre de gestes par fichier et durée de fichier (secondes)*

| | Min | Max | Moy | Med | SD | N | Durée tot | % corpus |
|---|---|---|---|---|---|---|---|---|
| Nombre de gestes | 22 | 48 | 34,21 | 37 | 8,16 | 479 | 3 min 25 s | 26,18 % |
| Durée de fichier (s) | 35 | 73 | 56,46 | 57 | 12,69 | 14 | 13 min 3 s | 100 % |

*Tableau 13. Distribution du nombre de gestes par fichier (min = minimum ; max = maximum ; moy = moyenne ; med = médiane ; sd = écart-type ; N = nombre d'occurrences ; tot = total).*

Le nombre de gestes annotés par fichier varie de 22 à 48 (moyenne 34,21, médiane 37), indiquant une légère asymétrie vers les valeurs basses, avec un écart-type de 8,16 gestes (Tableau 13). Les durées des fichiers s'étendent de 35 à 73 s (moyenne 56,46, médiane 57, écart-type 12,69). La courbe de tendance (grise), calculée par régression linéaire, met en évidence que le nombre de gestes observés semble globalement corrélé à la durée des fichiers (Figure 19), le nombre de gestes étant plus élevé dans les fichiers les plus longs. Sur la durée totale des 13 min 3 s des fichiers, les gestes occupent 3 min 25 s, soit environ un quart du temps enregistré (26,2 %). Cette durée ne concerne toutefois que la sommes des durées de segments de *strokes*, les autres phases composant les phrases gestuelles n'ayant pas été annotées. De ce fait, en réalité, les participants « gesticulent » bien davantage que ne le suggère la proportion de temps calculée ici.

# 3 Analyse fonctionnelle des gestes

## 3.1 Distribution des labels fonctionnels

### 3.1.1 Types fonctionnels par dimension

| Type | Dimension 1 | | Dimension 2 | | Total | |
|---|---|---|---|---|---|---|
| | *N* | *%* | *N* | *%* | *N* | *%* |
| Butterworth | 2 | 0,42 | -- | | 2 | 0,34 |
| Déictique | 97 | 20,25 | 22 | 21,57 | 119 | 20,48 |
| Iconique | 26 | 5,43 | 4 | 3,92 | 30 | 5,16 |
| Métaphorique | 94 | 19,62 | 9 | 8,82 | 103 | 17,73 |
| Modal | 60 | 12,53 | 47 | 46,08 | 107 | 18,42 |
| Parsing | 33 | 6,89 | 7 | 6,86 | 40 | 6,88 |
| Phatique | 11 | 2,30 | 6 | 5,88 | 17 | 2,93 |
| Quasi-linguistique | 18 | 3,76 | 4 | 3,92 | 22 | 3,79 |
| Expressif | -- | | 1 | 0,98 | 1 | 0,17 |
| Undefined | 138 | 28,81 | 2 | 1,96 | 140 | 24,10 |
| **Total** | **479** | | **102** | | **581** | |

*Tableau 14. Distribution des étiquettes de types fonctionnels dans le corpus (N = nombre d'occurrences)*

Parmi les 479 gestes annotés, 345 présentent au moins une dimension fonctionnelle renseignée (72,03 % du total), tandis que les 134 gestes restants (27,98 % du total de gestes annotés) se sont vu attribuer la seule étiquette *Undefined* en tant que leur fonction, bien que perceptible, n'a pas pu être clairement déterminée. Parmi les 345 gestes dont les fonctions ont pu être renseignées, 243 comportent uniquement une D1 (70,43 % des gestes ayant au moins une dimension fonctionnelle), 102 comportent à la fois une D1 et une D2 (29,57 % des gestes). Parmi ces 102 gestes bidimensionnels, six présentent une indétermination fonctionnelle sur l'une des deux dimensions : quatre gestes ont ainsi été étiquetés *Undefined* dans leur dimension principale D1, alors que leur dimension secondaire D2 a pu, elle, être clairement identifiée, et deux autres gestes, dont la dimension D1 a pu être déterminée, se sont vus attribuer une dimension secondaire D2 sans que celle-ci n'ait pu être clairement identifiée.

Au total, ce sont donc 581 étiquettes de fonction qui ont été annotées, en tenant compte des dimensions 1 et 2 (Tableau 14). En première dimension, la catégorie fonctionnelle la plus fréquente correspond aux gestes *Déictiques* (20,25 %), suivie des gestes *Métaphoriques* (19,25 %) et, dans une moindre mesure, des gestes *Modaux* (12,53 %). Les autres catégories – *Iconiques, Parsing, Phatiques et Quasi-linguistiques* – sont nettement moins représentées, avec des proportions comprises approximativement entre 2,3 % et 6,9 %, tandis que les gestes de type *Butterworth* restent particulièrement marginaux, avec seulement deux occurrences dans l'ensemble du corpus. On remarquera également l'ampleur du nombre de gestes dont l'interprétation a été problématique en première dimension : 28,81% des gestes ayant été étiquetés *Undefined*. En seconde dimension, les gestes sont le plus fréquemment associés à une fonction *Modale* (46,08 %) et *Déictique* (21,57 %). Les autres catégories fonctionnelles (*Iconiques, Métaphoriques, Parsing, Phatiques et Quasi-linguistiques*) sont moins représentées, avec des proportions comprises entre environ 3,9 % et 8 %. Enfin, les gestes de type *Expressif* apparaissent très marginalement dans cette seconde dimension, avec seulement une occurence.

Deux occurrences de gestes annotés *Undefined* apparaissent en seconde dimension : leur fonction secondaire est attestée, mais n'a pas pu être identifiée.

Dans l'ensemble, les gestes produits par le locuteur Wayne ont majoritairement pour fonction de référer à des entités (20,25 % de gestes *Déictiques* et 21,57 % en D2), d'exprimer une modalité (12,53 % en D1 et 46,08 % en D2) ou encore d'illustrer certains concepts de manière *Métaphorique* (19,25 % en D1). Néanmoins, un pourcentage non négligeable de gestes demeure associé à une fonction indéterminée (28,81% en D1 et 2% en D2). Cette proportion met en évidence les limites inhérentes aux catégorisations fonctionnelles existantes, dont le protocole présente pourtant une version riche et mixée, notamment lorsque les indices contextuels sont insuffisants pour interpréter un geste de manière fiable ou lorsque les gestes présentent des fonctions « hybrides » difficiles à discerner. Cette proportion souligne également l'intérêt d'un système d'annotation autorisant à la fois l'incertitude des interprétations et la multidimensionnalité des gestes : il s'agit là de données cruciales tant pour l'analyse multimodale que pour la conception de modèles computationnels capables de gérer l'ambiguïté.

Précisons que pour les catégories très marginales (telles que les gestes *Butterworth* avec N=2 ou *Expressif* avec N=1), les pourcentages sont fournis à titre purement indicatif pour maintenir la cohérence de la distribution totale. Le très faible nombre d'occurrences pour ces classes ne permet évidemment pas d'en tirer des conclusions statistiques significatives, raison pour laquelle elles seront exclues de certaines analyses ultérieures.

### 3.1.2 Types fonctionnels appariés en D1 et d2

| Type | Sans d2 | | *Déictique* | Expressif | *Iconique* | *Métaphorique* | *Modal* | *Parsing* | *Phatique* | *Quasi-linguistique* | *Undefined* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | *%* | | | | | | | | | |
| Butterworth | 2 | 100 | | | | | | | | | |
| Déictique | 82 | 84,54 | 1 | | 4 | 4 | 1 | 4 | 1 | | |
| Iconique | 24 | 92,31 | 1 | | | | 1 | | | | |
| Métaphorique | 77 | 81,92 | 5 | | | 1 | 10 | 1 | | | |
| Modal | 39 | 65 | 2 | 1 | | 2 | 7 | | 4 | 4 | 1 |
| Parsing | 17 | 51,52 | 12 | | | 1 | 2 | 1 | | | |
| Phatique | 2 | 18,18 | 1 | | | | 7 | | | | 1 |
| Quasi-linguistique | -- | | | | | 1 | 17 | | | | |
| Expressif | -- | | | | | | | | | | |
| Undefined | 134 | 97,10 | | | | | 2 | 1 | 1 | | |
| **Global** | **377** | **78,71** | **22** | **1** | **4** | **9** | **47** | **7** | **6** | **4** | **2** |

*Tableau 15. Distribution croisée des types fonctionnels en dimension 1 (D1) et en dimension 2 (D2) (N = nombre d'occurrences)*

Le taux de gestes interprétés comme n'assurant qu'une fonction varie selon ces fonctions (Tableau 15). Ainsi les gestes interprétés comme *Iconiques* sont-ils presque systématiquement

étiquetés comme mono-dimensionnels, avec plus de 90 % des occurrences dans cette catégorie. Les gestes avec une première dimension *Déictique* ou *Métaphorique* sont également majoritairement mono-dimensionnels, tout en présentant toutefois un taux d'étiquetage d'une seconde dimension plus important, de l'ordre de 15 à 20 %. Les gestes ayant une fonction principale *Modale* ou de *Parsing*, quant à eux, présentent également fréquemment une deuxième dimension, respectivement à hauteur de 35 % et 49 % d'entre eux. Enfin, les gestes étiquetés comme *Phatiques* en dimension 1 constituent un cas particulier, puisqu'ils sont le plus souvent associés à une seconde dimension fonctionnelle, avec seulement 18,18 % d'occurrences sans dimension 2. Les gestes interprétés comme principalement *Quasi-linguistiques* sont quant à eux systématiquement associés à une seconde dimension. On notera pour finir le cas des gestes *Undefined*, dont la fonction principale n'a donc pas pu être déterminée avec précision : si 97% n'ont que cette seule étiquette, les 3% restants ont été annotés comme assurant une deuxième fonction.

Lorsque l'on considère les fonctions les plus fréquemment appariées sur une même occurrence, que ces fonctions soient principales ou secondaires, plusieurs régularités émergent. Les gestes ayant une fonction *Déictique* apparaissent ainsi associés à une grande variété d'autres fonctions, couvrant l'ensemble des catégories à l'exception de la fonction *Quasi-linguistique*, avec toutefois une prédominance d'association avec le rôle de *Parsing*. Les gestes de fonction *Métaphorique* et de *Parsing* sont, quant à eux, le plus souvent appariés à la fonction *Modale*. Enfin, les gestes interprétés comme *Modaux* se combinent avec l'ensemble des autres fonctions, mais présentent une association particulièrement fréquente avec la fonction *Quasi-linguistique*.

## 3.2 Durée des gestes

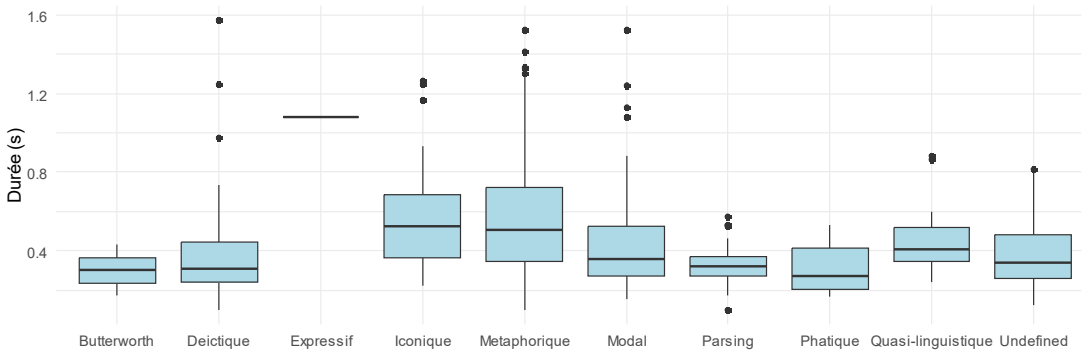



*Figure 20. Boxplots des durées de gestes dans le corpus par type fonctionnel*

| Type (D1+D2) | Min | AD | MDu | SD | Max |
|---|---|---|---|---|---|
| Butterworth | 0,17 | 0,30 | 0,30 | 0,18 | 0,43 |
| Déictique | 0,10 | 0,36 | 0,31 | 0,21 | 1,57 |
| Iconique | 0,22 | 0,56 | 0,52 | 0,29 | 1,27 |
| Métaphorique | 0,10 | 0,59 | 0,51 | 0,32 | 1,52 |
| Modal | 0,15 | 0,43 | 0,36 | 0,24 | 1,52 |
| Parsing | 0,10 | 0,33 | 0,32 | 0,10 | 0,58 |
| Phatique | 0,17 | 0,31 | 0,28 | 0,12 | 0,53 |
| Quasi-linguistique | 0,24 | 0,45 | 0,41 | 0,17 | 0,88 |
| Undefined | 0,08 | 0,38 | 0,34 | 0,16 | 0,82 |
| **Total** | **0,08** | **0,43** | **0,36** | **0,24** | **1,57** |

*Tableau 16. Statistiques descriptives des durées de gestes*
*(min = minimum ; AD = durée moyenne ; MDu = durée médiane ; SD = écart-type ; max = maximum)*

Dans l'ensemble, la durée moyenne (AD) d'un geste tel qu'annoté dans reBeat-14 (phase de *stroke* uniquement) s'élève à 0,43 seconde (SD = 0,24 s), avec une durée médiane (MDu) de 0,36 seconde (Tableau 16). L'étendue des durées de ces gestes est ainsi relativement homogène, allant de 0,10 s à 1,57 s. La médiane (0,36 s) étant proche de la moyenne (0,43 s), la distribution apparaît globalement centrée autour de cette valeur, traduisant ainsi des durées relativement équilibrées, malgré la présence de quelques gestes particulièrement courts ou longs.

La comparaison des durées en fonction des types fonctionnels (Tableau 16, Figure 20) révèle des profils relativement similaires, avec certaines légères différences toutefois. Les gestes ayant un rôle *Iconique* (AD = 0.56 s ; MDu = 0,52 s) ou *Métaphorique* (AD = 0.59 s ; MDu = 0,51 s) sont en moyenne les gestes les plus longs, bien que présentant en réalité la plus grande hétérogénéité de durées observable au sein d'un même type fonctionnel de gestes (de 0,22 s à 1,27 s). Les gestes assurant un rôle de *Quasi-linguistique* ou de *Modal* suivent, avec une durée moyenne d'environ 0,45 s (MDu entre 0,36 s et 0,41 s) pour ces étiquettes fonctionnelles. Les autres catégories (*Déictique, Phatique, Parsing)* regroupent les fonctions de gestes les plus courtes du corpus, avec une durée moyenne comprise entre 0,30 et 0,36 s et une médiane comprise entre 0,28 et 0,32 s. Enfin, les gestes dont la catégorie fonctionnelle n'a pas pu être interprétée (*Undefined*) présente des durées moyennes de 0,38 s et médianes de 0,34 s.

## 3.3 Scores de certitudes

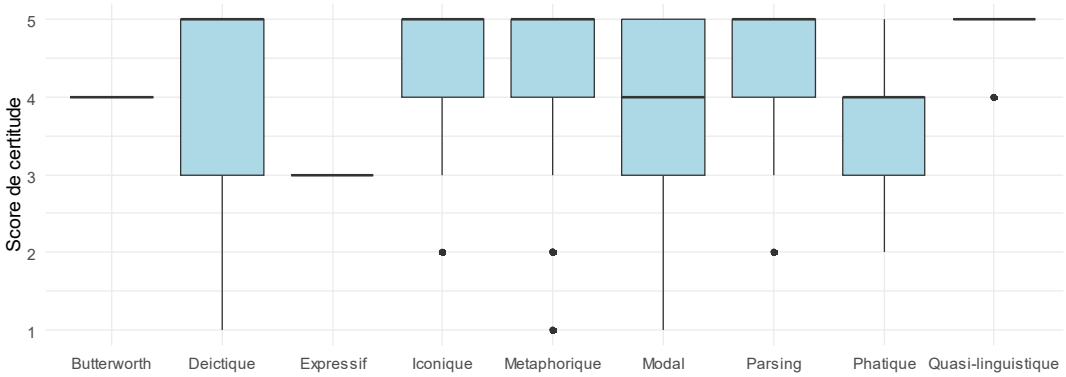



*Figure 21. Boxplots des scores de certitude par type fonctionnel*

| Type (D1+D2) | Moy | Med | SD |
|---|---|---|---|
| Butterworth | 4 | 4 | 0 |
| Déictique | 4,04 | 5 | 1,18 |
| Iconique | 4,46 | 5 | 0,89 |
| Métaphorique | 4,22 | 5 | 1,01 |
| Modal | 4,03 | 4 | 0,94 |
| Parsing | 4,45 | 5 | 0,85 |
| Phatique | 3,69 | 4 | 0,94 |
| Quasi-linguistique | 4,86 | 5 | 0,35 |
| **Total** | **4,17** | **5** | **1,02** |

*Tableau 17. Analyse des scores de certitude par type fonctionnel*
*(moy = moyenne ; med = médiane ; SD = écart-type)*

Dans l'ensemble des types fonctionnels annotés, on peut observer un score de certitude relativement élevé (Tableau 17) avec une moyenne de 4,17/5 (médiane = 5) et un écart-type de 1,02. Les scores moyens sont globalement similaires entre les différentes catégories (Figure 21), autour de 4. Certaines catégories, cependant, présentent une plus grande dispersion des scores, avec des écarts-types plus élevés, notamment s'agissant des gestes étiquetés comme *Déictiques*, *Métaphoriques*, *Modaux* et *Phatiques* (écart-type $\approx 1$). Dans l'ensemble, ces résultats suggèrent que les annotateurs étaient plutôt confiants dans l'identification des fonctions de gestes, avec des scores de certitude élevés et relativement homogènes entre les catégories. Les variations plus importantes observées pour certaines fonctions indiquent que certaines catégories ont semblés plus difficiles à interpréter, mais globalement, la fiabilité de l'annotation reste consistante.

# 4 Analyse formelle des gestes annotés

## 4.1 Manualité

### 4.1.1 Distribution générale des labels de manualité

| | **Bimanuel** | | | **Monomanuel** | |
|---|---|---|---|---|---|
| | *AS* | *MH* | *PH* | *RH* | *LH* |
| N | 23 | 228 | 18 | 183 | 27 |
| % | 8,6 | 84,8 | 6,7 | 87,1 | 12,9 |
| **Total** | **269 (56,2 %)** | | | **210 (43,8 %)** | |

*Tableau 18. Répartition des étiquettes de manualité dans le corpus*
*(AS = mains asymétriques ; MH = mains en miroir ; PH = mains parallèles ; RH = main droite ; LH = main gauche).*

Dans le corpus, on observe un équilibre relatif entre les gestes réalisés avec une seule main (monomanuels) ou avec les deux (bimanuels), avec une légère prédominance des productions bimanuelles (56,2 % contre 43,8 %, Tableau 18). Les gestes bimanuels sont majoritairement de type mains en miroir (gestes produits en symétrie axiale, *Mirror Hands* MH = 84,8 %), les gestes annotés comme asymétriques (AS) et parallèles (PH) étant minoritaires et de proportions similaires (8,6 % et 6,7 %). Les gestes monomanuels sont majoritairement effectués avec la main

droite (RH = 87,1 % ; LH = 12,9%), ce qui s'explique probablement par le fait que le locuteur est droitier.

### 4.1.2 Manualité et type fonctionnel

| Type (D1+D2) | Bimanuel | | | Monomanuel | | Total |
|---|---|---|---|---|---|---|
| | **AS** | **MH** | **PH** | **RH** | **LH** | |
| | *N = 24* | *N = 279* | *N = 21* | *N = 221* | *N = 36* | N = 581 |
| Butterworth | -- | -- | -- | 2 (100%) | -- | **2** |
| Déictique | 3 (2,52 %) | 26 (21,85 %) | 4 (3,36 %) | 72 (60,50 %) | 14 (11,76 %) | **119** |
| Iconique | 4 (13,33 %) | 8 (26,67 %) | -- | 17 (56,67 %) | 1 (3,33 %) | **30** |
| Métaphorique | 10 (9,71 %) | 56 (54,37 %) | 8 (7,77 %) | 29 (28,16 %) | -- | **103** |
| Modal | -- | 76 (71,03 %) | -- | 25 (23,36 %) | 6 (5,61 %) | **107** |
| Parsing | 4 (10%) | 11 (27,5%) | 4 (10%) | 13 (32,5%) | 8 (20%) | **40** |
| Phatique | -- | 5 (29,41 %) | -- | 11 (64,71 %) | 1 (5,88 %) | **17** |
| Quasi-linguistique | -- | 16 (72,73 %) | -- | 6 (27,27 %) | -- | **22** |
| Expressif | -- | 1 (100%) | -- | -- | -- | **1** |
| Undefined | 3 (2,14%) | 80 (57,14%) | 5 (3,57%) | 46 (32,86%) | 6 (4,29%) | **140** |

*Tableau 19. Distribution de la manualité en fonction du type fonctionnel du geste (AS = mains asymétriques ; MH = mains en miroir ; PH = mains parallèles ; RH = main droite ; LH = main gauche)*

L'analyse de la distribution des manualités selon les catégories de gestes annotées met en évidence des différences notables (Tableau 19). Les gestes assurant une fonction *Métaphorique*, *Modale* et *Quasi-linguistique* sont majoritairement réalisés à deux mains, en miroir (*Mirror Hands*, MH : de 55 % et 73 %) et, dans une moindre mesure, avec la main droite seulement (de 23% à 28 %). À l'inverse, les gestes assurant une fonction de *Déictique, d'Iconique* et de *Phatique* sont principalement produits avec la main droite seule (de 60% à 70 %), suivie de la configuration mains en miroir (environ 20 %). Les gestes avec un rôle de *Parsing* présentent quant à eux une distribution plus équilibrée, avec trois catégories de manualité observables en de mêmes proportions : main droite seule (RH = 32%), main gauche seule (LH = 20%) et mains en miroir (MH = 27,5%). Enfin, les gestes labellisés *Undefined* sur une de leur dimension fonctionnelle sont majoritairement produits avec les mains en miroir (MH, environ 57 %), sinon avec la main droite seule (environ 32 %), les autres manualités étant plus marginales. Concernant les catégories de manualités les moins fréquentes dans le corpus, les gestes réalisés selon une configuration des mains asymétrique (AS) ou parallèle (PH) sont essentiellement *Métaphorique*s (10 sur 24 manualités AS annotées et 8 sur 21 manualités PH annotées), tandis

que l'engagement de la main gauche seulement (LH) est principalement associée aux gestes *Déictiques* (14 sur 36 manualités LH annotées).

## 4.2 Forme de la trajectoire

| | **Line** | **Curve** | **Loop** | **PS** | **BF** | **PL** | **RIL** | ***Hapax*** | **Autres** | **Total** |
|---|---|---|---|---|---|---|---|---|---|---|
| N | 325 | 288 | 8 | 19 | 35 | 10 | 18 | 12 | 33 | **748** |
| % | 43,45 | 38,50 | 1,07 | 2,54 | 4,68 | 1,34 | 2,41 | 1,60 | 4,41 | **100** |

*Tableau 20. Distribution des labels de la forme de la trajectoire du geste dans le corpus (line = ligne ; curve = courbe ; PS = progression scandée ; BF = aller-retour ; PL = boucle progressive ; RIL = moulinet)*

Dans le cas de l'annotation de la forme de la trajectoire par geste (Tableau 20), on peut observer qu'une majorité de gestes a une forme rectiligne (43,45 %) ou courbée (38,50 %), les formes plus complexes représentant seulement de 1 % et 4,7 % des 748 labels utilisés pour caractériser la trajectoire de chacune des mains impliquées dans l'élaboration d'un geste.

| **Type (D1+D2)** | **BF** *N = 53* | **Curve** *N = 332* | ***Hapax*** *N = 16* | **Line** *N = 406* | **Loop** *N = 11* | **Autres** *N = 38* | **PL** *N = 10* | **PS** *N = 21* | **RIL** *N = 18* | **Total** *N = 905* |
|---|---|---|---|---|---|---|---|---|---|---|
| Butterworth | -- | 2 (100%) | -- | -- | -- | -- | -- | -- | -- | 2 |
| Déictique | 3 (1,97 %) | 78 (51,32 %) | 4 (2,63 %) | 60 (39,47 %) | 3 (1,97 %) | 2 (1,32 %) | -- | 2 (1,32 %) | -- | **152** |
| Iconique | 4 (9,52 %) | 15 (35,71 %) | 8 (19,05 %) | 10 (23,81 %) | 2 (4,76 %) | 2 (4,76 %) | -- | 1 (2,38 %) | -- | **42** |
| Métaphorique | 5 (2,83 %) | 63 (35,59 %) | 4 (2,26 %) | 49 (27,68 %) | 6 (3,39 %) | 7 (3,96 %) | 10 (5,65 %) | 17 (9,61 %) | 16 (9,04 %) | **177** |
| Modal | 23 (12,57 %) | 54 (29,51 %) | -- | 102 (55,73 %) | -- | 4 (2,19 %) | -- | -- | -- | **183** |
| Parsing | -- | 29 (49,15%) | -- | 26 (44,07%) | -- | 4 (6,78%) | -- | -- | -- | **59** |
| Phatique | -- | 4 (18,18 %) | -- | 17 (77,27 %) | -- | 1 (4,55 %) | -- | -- | -- | **22** |
| Quasi-linguistique | 15 (39,47 %) | 5 (13,16 %) | -- | 17 (44,74 %) | -- | 1 (2,63 %) | -- | -- | -- | **38** |
| Expressif | -- | 2 (100 %) | -- | -- | -- | -- | -- | -- | -- | **2** |
| Undefined | 3 (1,32%) | 80 (35,09%) | -- | 125 (54,83%) | -- | 17 (7,46%) | -- | 1 (0,44%) | 2 (0,88%) | **228** |

*Tableau 21. Distribution des labels de forme en fonction des types fonctionnels (line = ligne ; curve = courbe ; PS = progression scandée ; BF = aller-retour ; PL = boucle progressive ; RIL = moulinet)*

L'analyse des formes de trajectoires annotées dans le corpus met en évidence des différences marquées selon les types fonctionnels des gestes (Tableau 21). On observe ainsi dans un premier temps que certains types de gestes reposent majoritairement sur un nombre restreint de formes, principalement des lignes et des courbes, tandis que les autres catégories, plus complexes, restent marginales. Dans le cas des gestes assurant un rôle de *Déictique*, les trajectoires courbées sont très largement dominantes (51,32 % de l'ensemble des trajectoires annotées pour cette catégorie), suivies des lignes (39,47 %). A l'inverse, pour les gestes ayant la fonction *Métaphorique*, on observe une prédominance des lignes (35,59 %) puis des courbes (23,81 %). En revanche, pour les gestes ayant un rôle d'*Iconique*, ces proportions tendent à s'équilibrer davantage entre les catégories de formes, avec trois types particulièrement

représentés : les courbes (39,47 %), les lignes (35,71 %) et les *Hapax* (23,81 %). Les gestes assurant la fonction de *Parsing* sont quant à eux caractérisés par une équivalence de fréquence de formes entre lignes et courbes, toutes deux entre 45 et 49 %. Les gestes dont la catégorie fonctionnelle n'a pas pu être interprétée (*Undefined*) correspondent le plus souvent à des formes linéaires (54,83%) puis courbes (35,09%). La catégorie de forme « *Autres* », qui regroupe des formes de gestes dont la forme n'a pas pu être qualifiée et qui ne semble pas être un mouvement isolé comme les *Hapax,* est associée en majorité à ce groupe *Undefined.* Cette surreprésentation au sein de cette catégorie pourrait refléter le fait qu'elle est composée d'un ensemble de gestes variés, dont ni la fonction ni le sens n'ont pu être identifiés.

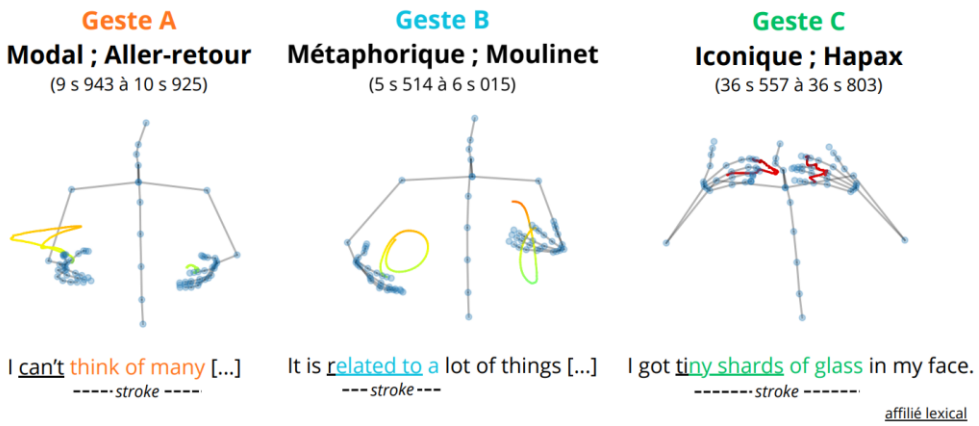



*Figure 22. Illustration de formes de gestes*

Certains types fonctionnels semblent être surreprésentés dans certaines formes minoritaires de trajectoires. Les gestes ayant une forme en aller-retour (*back and forth*, BF) sont essentiellement des gestes ayant une fonction *Modale* (23 sur 53 formes BF annotées). On peut supposer que les gestes ayant un rôle *Modal* sont fréquemment observés dans les trajectoires de type aller-retour (BF), en raison de la tendance de Wayne à produire ce schème lors de négations. Par exemple, dans le geste A (Figure 22), le locuteur dit « I can't think of many » et traduit l'impossibilité de trouver des informations par des mouvements répétés de va-et-vient horizontaux pour figurer un « non ».

Les gestes ayant un rôle *Métaphorique*, quant à eux, se distinguent en ce qu'ils sont réalisés par un plus large ensemble des catégories formelles complexes, notamment les *boucles* (6 sur 11 annotations), les *progressions scandées* (17 sur 21 annotations), les *boucles progressives* (10/10 annotations) et les *moulinets* (16 sur 18 annotations). Leur surreprésentation dans ce type de trajectoires peut potentiellement s'expliquer par la difficulté d'illustrer ce qui est dit oralement à l'aide d'un autre concept de manière visuelle, sans que celui-ci ne soit explicitement énoncé par le locuteur. Ainsi, pour éviter toute ambiguïté, ces gestes peuvent ponctuellement nécessiter une production de forme de geste plus élaborée. Le geste B (Figure 22) est un exemple de *Métaphorique* produit par Wayne pendant l'énonciation verbale « It is related to a lot of things ». La trajectoire en moulinet associée illustre visuellement les liens et le mélange, reflétant ainsi la connexion entre plusieurs éléments exprimée par le locuteur (« related to »). Cette surreprésentation de ce type fonctionnel au sein de formes en boucle (répétées ou non)

ou de progressions (bouclées ou scandées) pourrait également refléter une tendance à figurer des processus.

Enfin, la forte présence de gestes ayant une fonction d'*Iconique* dans les trajectoires « isolées » (*Hapax*), qui ne peuvent pas être regroupées avec d'autres formes similaires *Hapax*, n'est pas surprenante (8 sur 16 annotations). Ce type de gestes illustrent souvent des éléments très spécifiques du discours, ce qui se traduit par des trajectoires elles-mêmes uniques dans le corpus. Le geste C (Figure 22) en est un exemple : Wayne décrit une blessure au visage (« I got tiny shards of glass in my face ») et figure alors de petits points pour représenter le fait d'être blessé par de multiples éclats de verre sur les joues. Cette forme de trajectoire correspond précisément à ce qu'elle illustre, et aucun geste similaire n'est retrouvé ailleurs dans le corpus.

Ces résultats suggèrent que la majorité des gestes co-verbaux reposent sur des trajectoires motrices perçues comme relativement simples, principalement linéaires ou courbes, quelle que soit leur fonction. La faible association globale entre types fonctionnels et trajectoires complexes indique que ces dernières restent marginales dans le corpus, mais qu'elles tendent à se spécialiser dans certains usages fonctionnels précis.

## 4.3 Position de la main

| | **Horizontal** | **Intermédiaire** | **Vertical** | **Total** |
|---|---|---|---|---|
| N | 406 | 294 | 48 | **748** |
| % | 54,28 | 39,3 | 6,42 | **100** |

*Tableau 22. Distribution des labels de la position de la main dans le corpus*

La distribution des positions de la main par geste annoté (Tableau 22) montre une nette prédominance des configurations horizontales, qui concernent plus de la moitié des occurrences (54,28 % sur l'ensemble des positions annotées). Les positions intermédiaires constituent également une part importante des données (39,30 %), tandis que les positions verticales apparaissent très marginales dans le corpus (6,42 %).

| Type (D1+D2) | H<br>*N =462* | I<br>*N = 372* | V<br>*N = 71* | Total<br>*N = 905* |
|---|---|---|---|---|
| Butterworth | 2<br>(100%) | -- | -- | **2** |
| Déictique | 84<br>(55,26 %) | 55<br>(36,18 %) | 13<br>(8,55 %) | **152** |
| Iconique | 25<br>(59,52 %) | 11<br>(26,19 %) | 6<br>(14,29 %) | **42** |
| Métaphorique | 97<br>(54,80 %) | 69<br>(38,98 %) | 11<br>(6,21 %) | **177** |
| Modal | 54<br>(29,51 %) | 107<br>(58,47 %) | 22<br>(12,02 %) | **183** |
| Parsing | 33<br>(55,93%) | 24<br>(40,68%) | 2<br>(3,32%) | **59** |
| Phatique | 7<br>(31,82 %) | 9<br>(40,91 %) | 6<br>(27,27 %) | **22** |
| Quasi-linguistique | 13<br>(34,21 %) | 23<br>(60,57 %) | 2<br>(5,26 %) | **38** |
| Expressif | -- | -- | 2<br>(100 %) | **2** |
| Undefined | 147<br>(64,47%) | 74<br>(32,47%) | 7<br>(3,07%) | **228** |

*Tableau 23. Disposition des labels de position de la main en fonction du type fonctionnel (H = horizontal ; I = intermédiaire ; V = vertical)*

L'analyse de la répartition de ces labels de position de la main montre que l'ensemble des positions sont mobilisées quel que soit le type fonctionnel de geste (Tableau 23). On observe toutefois des différences dans les catégories dominantes selon les fonctions observées. Les gestes assurant un rôle de *Déictique, d'Iconique, de Métaphorique*, de *Parsing* ou ceux dont le type fonctionnel n'est pas défini présentent majoritairement des positions horizontales (entre 55 et 60 %), suivies des positions intermédiaires (entre 30 et 40 %), tandis que les positions verticales restent marginales. À l'inverse, pour les gestes assurant un rôle *Quasi-linguistique* et *Modal*, les deux premières catégories s'inversent : la position intermédiaire est la plus fréquente (environ 60 %), alors que la position horizontale concerne 30 % des occurrences. Enfin, les gestes *Phatiques* se distinguent par une répartition plus équilibrée entre les trois positions, avec 31 % de positions horizontales, 41 % de positions intermédiaires et 27 % de positions verticales. La fréquence plus élevée des positions verticales pour ces gestes peut potentiellement s'expliquer par la récurrence des recours aux

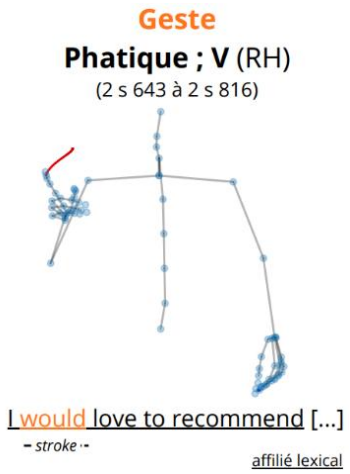



*Figure 23. Illustration de geste Phatique avec une position de main droite (RH) annotée comme verticale (V)*

gestes utilisés pour attirer l'attention de l'interlocuteur sur ce qui est dit, à l'instar du geste illustré dans la Figure 23.

## 4.4 Orientation de la paume

| | D | E | F | MD | MU | SO | U | Total |
|---|---|---|---|---|---|---|---|---|
| N | 13 | 276 | 16 | 52 | 181 | 141 | 69 | **748** |
| % | 1,74 | 36,90 | 2,14 | 6,95 | 24,20 | 18,85 | 9,22 | **100** |

*Tableau 24. Distribution des labels d'orientation de la paume dans le corpus (D = vers le bas ; E = sur la tranche ; F = vers l'avant ; MD = partiellement vers le bas ; MU = partiellement vers le haut ; SO = vers le locuteur ; U = vers le haut)*

Dans l'ensemble, les labels d'orientation de la paume (Tableau 24) révèlent que les mains du locuteur sont majoritairement orientées sur la tranche (E, 36,90 % des occurrences du corpus), sinon tournées partiellement vers le haut (MU, 25%) ou vers lui-même (SO, 20 %). Les orientations vers le haut (U) sont également présentes, avec environ 9 % des labels, tandis que l'orientation partiellement vers le bas représente environ 7 %. Les autres catégories d'orientation sont marginalement utilisées dans le corpus (D pour *Down* et F pour *Front*), avec environ 2 % chacune.

| **Type (D1+D2)** | **D** *N = 17* | **E** *N = 341* | **F** *N = 28* | **MD** *N = 73* | **MU** *N = 205* | **SO** *N = 160* | **U** *N = 81* | **Total** *N = 905* |
|---|---|---|---|---|---|---|---|---|
| Butterworth | -- | 1 (50%) | -- | -- | 1 (50%) | -- | -- | **2** |
| Déictique | -- | 59 (38 ,82%) | -- | 9 (5,92 %) | 33 (21,71 %) | 47 (30,92 %) | 4 (2,63 %) | **152** |
| Iconique | 4 (9,52 %) | 9 (21,43%) | 2 (4,76 %) | 4 (9,52 %) | 5 (11,91 %) | 11 (26,19 %) | 7 (16,67 %) | **42** |
| Métaphorique | 2 (1,13 %) | 65 (36,72%) | 2 (1,13 %) | 20 (11,30 %) | 20 (11,30 %) | 56 (31,64 %) | 12 (6,77 %) | **177** |
| Modal | 4 (2,19 %) | 79 (43,17%) | 10 (5,46 %) | 17 (9,29 %) | 37 (20,21 %) | 18 (9,84 %) | 18 (9,84 %) | **183** |
| Parsing | 3 (5,09%) | 11 (18,64%) | 2 (3,39%) | 6 (10,17%) | 26 (44,07%) | 6 (10,17%) | 5 (8,48%) | **59** |
| Phatique | -- | 19 (86,37%) | 2 (9,09 %) | 1 (4,55 %) | -- | -- | -- | **22** |
| Quasi-linguistique | 4 (10,57 %) | 13 (34,21%) | 8 (21,05 %) | 3 (7,90 %) | 3 (7,90 %) | -- | 7 (18,42 %) | **38** |
| Expressif | -- | 2 (100%) | -- | -- | -- | -- | -- | **2** |
| Undefined | -- | 83 (36,40%) | 2 (0,88%) | 13 (5,70%) | 80 (35,09%) | 22 (9,65%) | 28 (12,28%) | **228** |

*Tableau 25. Distribution des labels d'orientation de la paume en fonction du type fonctionnel (D = vers le bas ; E = sur la tranche ; F = vers l'avant ; MD = partiellement vers le bas ; MU = partiellement vers le haut ; SO = vers le locuteur ; U = vers le haut)*

L'analyse des labels d'orientation de la paume en fonction du type fonctionnel annoté met en évidence des profils distincts selon les catégories de gestes (Tableau 25). On observe que la plupart des types fonctionnels présentent une répartition des orientations de la paume couvrant l'ensemble des catégories, mais pas dans les mêmes proportions. Les gestes assurant un rôle de *Phatique* se distinguent toutefois nettement, avec une paume majoritairement placée

sur la tranche (E, 86 %), tandis que les autres orientations observées pour ces gestes (vers l'interlocuteur [F] et vers le bas [D]) sont très faiblement représentées.

La paume orientée en *Edge* constitue également la catégorie la plus représentée pour les gestes ayant un rôle de *Déictique* (38,82 %), de *Quasi-linguistique* (34,21 %) ou de *Modal* (43,17 %). Les gestes assurant une fonction de *Déictique* présentent ensuite une fréquence élevée d'orientation de la paume vers le locuteur (*self-oriented*, SO = 30,92 %), suivie d'une orientation partiellement vers le haut (*mid-up*, MU = 22 %). Pour les gestes ayant un rôle de *Modal*, l'orientation partiellement vers le haut (*mid-up* MU) arrive en deuxième position (22 %), puis trois orientations apparaissent à des niveaux comparables, autour de 10 % chacune : vers le locuteur (SO), vers le haut (U) et partiellement vers le bas (MD). Les gestes assurant un rôle de *Quasi-linguistique* présentent ensuite majoritairement une orientation de la paume vers le haut (U = 18 %) et vers l'avant (F = 21,02 %), tandis que les orientations partiellement vers le bas (MD), vers le bas (D) et partiellement vers le haut (MU) apparaissent plus marginalement, avec des proportions comprises entre 7 % et 10 %.

Plusieurs catégories fonctionnelles de gestes montrent une fréquence élevée de co-occurrence de différentes orientations de la paume. Les gestes assurant une fonction d'*Iconique* et de *Métaphorique* se caractérisent par des fréquences élevées d'orientation de la paume orientée vers le locuteur (SO) et sur la tranche (*Iconiques* : *self-oriented* SO = 26,19 % et *Edge* E = 21,43 % ; *Métaphoriques* : SO = 31,64 % et E = 36,7 %). Ces deux catégories se distinguent toutefois dans la distribution des autres orientations. Pour les gestes ayant un rôle d'*Iconique*, on observe ensuite une proportion importante de paumes orientées vers le haut (*Up*, U = 16,67 %), suivi des orientations partiellement vers le haut (MU), partiellement vers le bas (MD) et vers le bas (D), présentes dans des proportions comparables à celles observées pour les gestes assurant un rôle de *Quasi-linguistique*, autour de 10 %. Dans le cas des gestes ayant une fonction *Métaphorique*, ces mêmes orientations sont également représentées, mais dans une moindre mesure, avec des valeurs comprises entre 7 % et 11 % pour les orientations partiellement vers le bas (MD), partiellement vers le haut (MU) et vers le haut (U). Les gestes sans type fonctionnel annoté (*Undefined*) présentent également deux catégories d'orientation de la paume largement dominantes, avec l'orientation de la paume sur la tranche (*Edge*, E) et partiellement vers le haut (MU) autour de 35 %. Les orientations vers le haut (U) et vers le locuteur (SO) sont ensuite représentées plus faiblement, avec des proportions comprises entre 10 % et 12 %.

Enfin, seuls les gestes assurant un rôle de *Parsing* ne présentent pas l'orientation *Edge* comme catégorie majoritaire. Ces gestes se caractérisent principalement par une paume orientée partiellement vers le haut (*mid-up*, MU = 44 %), tandis que l'orientation sur la tranche (E) n'atteint que 18,64 % des annotations dans cette catégorie fonctionnelle. Les orientations partiellement vers le bas (MD) et vers le locuteur (SO) apparaissent ensuite dans des proportions comparables, autour de 10 % pour chacune. Pour ces gestes, la prédominance de l'orientation de la paume partiellement vers le haut (MU) pourrait s'expliquer par le fait qu'ils sont régulièrement utilisés lors des énumérations. Cette catégorie semble être une orientation de la paume récurrente pour organiser visuellement l'énonciation d'une liste, en structurant les éléments au fur et à mesure par une sorte de gestes « présentatifs » semblant répartir les contenus dans l'espace. La Figure 24 illustre ce phénomène avec un geste *Parsing* produit sur le deuxième élément de l'énumération de Wayne (« have fun »).

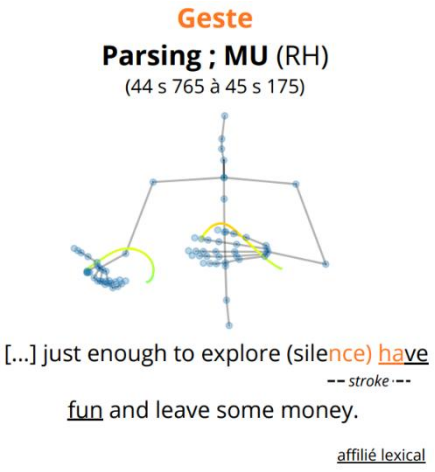



*Figure 24. Geste de Parsing avec une orientation de la paume droite (RH) partiellement vers le haut (MU)*

Si l'on analyse les catégories d'orientation les moins représentées dans l'ensemble du corpus, on constate qu'aucune catégorie fonctionnelle n'est majoritairement associée à l'orientation vers le bas (D). Pour les orientations vers l'interlocuteur (F), partiellement vers le bas (MD) et vers le haut (U), ce sont principalement les gestes assurant un rôle *Modal* qui sont associés ( F : 10 sur 28 ; MD : 17 sur 23 ; U : 18 sur 81 annotations), avec toutefois une présence notable des gestes ayant un rôle de *Quasi-linguistique* pour l'orientation vers l'interlocuteur (F : 8 sur 28 annotations) et des gestes assurant un rôle de *Métaphorique* pour les orientations partiellement vers le bas (MD : 20 sur 73 annotations), dans des proportions comparables aux gestes ayant un rôle de *Modal*.

Dans l'ensemble, ces résultats montrent que l'orientation de la paume varie selon le type fonctionnel du geste et n'est pas répartie uniformément entre les catégories. Si l'orientation *Edge* domine dans une grande partie des gestes, certaines catégories présentent des profils spécifiques sur les autres orientations annotées. Par ailleurs, certaines catégories fonctionnelles, telles que les gestes ayant un rôle *Iconique* ou *Métaphorique*, partagent des distributions proches, suggérant des mouvements comparables du point de vue de l'orientation de la paume.

## 4.5 Configuration de la main

| | CH | IE | MCH | OH | SEI | SOH | TE | TIE | TMIE | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| N | 8 | 115 | 20 | 324 | 3 | 260 | 6 | 2 | 10 | **748** |
| % | 1,07 | 15,37 | 2,67 | 43,32 | 0,40 | 34,76 | 0,80 | 0,27 | 1,34 | **100** |

*Tableau 26. Distribution des labels de configuration de la main dans le corpus (CH = main fermée ; IE = index engagé ; MCH = main partiellement fermée ; OH = main ouverte ; SEI = index partiellement engagé ; SOH = main partiellement ouverte ; TE = pouce engagé ; TIE = pouce et index engagés ; TMIE = pouce, majeur et index engagés)*

L'analyse de la répartition des labels de configuration de la main, annotés par geste et par main impliquée (Tableau 26), montre une nette dominance de la configuration *main ouverte* (OH = 43,32 %) dans le corpus. Les configurations *main partiellement ouverte* (SOH = 34,76 %) et *index engagé* (IE = 15,37 %) suivent ensuite. Les autres configurations sont très marginales, avec des fréquences comprises entre 1 % et 3 %, voire quelques occurrences très rares pour *index partiellement engagé* (SEI), *pouce engagé* (TE) et *pouce et index engagés* (TIE).

| **Type (D1+D2)** | **CH** *N = 8* | **IE** *N = 151* | **MCH** *N = 24* | **OH** *N = 409* | **SEI** *N = 3* | **SOH** *N =286* | **TE** *N = 10* | **TIE** *N = 2* | **TMIE** *N = 12* | **Total** *N = 905* |
|---|---|---|---|---|---|---|---|---|---|---|
| Butterworth | -- | -- | -- | -- | -- | 2 (100%) | -- | -- | -- | **2** |
| Déictique | -- | 36 (23,68 %) | -- | 46 (30,26 %) | 2 (1,32 %) | 63 (41,45 %) | 1 (0,66 %) | 1 (0, 6 %) | 3 (1,97 %) | **152** |
| Iconique | 6 (14,29 %) | 5 (11,91 %) | 2 (4,76 %) | 14 (33,33 %) | -- | 14 (33,33 %) | -- | -- | 1 (2,38 %) | **42** |
| Métaphorique | 2 (1,12 %) | 13 (7,35 %) | 7 (3,96 %) | 106 (59,89 %) | -- | 42 (23,73 %) | -- | 1 (0,57 %) | 6 (3,39 %) | **177** |
| Modal | -- | 25 (13,66 %) | 11 (6,01 %) | 103 (56,28 %) | -- | 39 (21,31 %) | 4 (2,19 %) | -- | 1 (0,55 %) | **183** |
| Parsing | -- | 6 (12,25%) | -- | 31 (63,27%) | -- | 21 (42,86%) | 1 (2,04%) | -- | -- | **59** |
| Phatique | -- | 21 (95,46 %) | 1 (4,55 %) | -- | -- | -- | -- | -- | -- | **22** |
| Quasi-linguistique | -- | 2 (5,26 %) | -- | 25 (65,79 %) | -- | 6 (15,79 %) | 4 (10,53 %) | -- | 1 (2,63 %) | **38** |
| Expressif | -- | -- | -- | -- | -- | 2 (100 %) | -- | -- | -- | **2** |
| Undefined | -- | 43 (18,86%) | 3 (1,32%) | 84 (36,84%) | 1 (0,44%) | 97 (42,54%) | -- | -- | -- | **228** |

*Tableau 27. Distribution des labels de configuration de la main en fonction du type fonctionnel (CH = main fermée ; IE = index engagé ; MCH = main partiellement fermée ; OH = main ouverte ; SEI = index partiellement engagé ; SOH = main partiellement ouverte ; TE = pouce engagé ; TIE = pouce et index engagés ; TMIE = pouce, majeur et index engagés)*

L'analyse des labels de la configuration des mains en fonction des types fonctionnels annotés révèle quelques différences selon les catégories de gestes (Tableau 27). Dans la majorité des cas, les gestes sont réalisés avec la main ouverte (*open hand*, OH), puis avec la main partiellement ouverte (SOH), et enfin avec l'index engagé (IE). C'est le cas pour les gestes assurant un rôle de *Métaphorique*, de *Parsing* ou de *Modal*, pour lesquels environ 60 % des configurations sont en main ouverte (OH). Les proportions de main partiellement ouverte (SOH) diffèrent cependant selon les types : elles s'élèvent à environ 22 % pour les gestes ayant une fonction de *Métaphorique* et *Modal*, tandis que pour les gestes assurant un rôle de *Parsing*, la main partiellement ouverte (SOH) concerne 42,86 % des configurations annotées pour cette fonction. Pour ces trois types fonctionnels, le label index engagé (IE) apparaît ensuite dans des proportions comprises entre 7 % et 14 %. Les gestes ayant un rôle de *Quasi-linguistique* présentent également une majorité de mains ouvertes (OH, 65,79 %), suivies de main partiellement ouvertes (SOH) dans une proportion moindre (15,79 %). Contrairement aux autres types fonctionnels mentionnés précédemment, cette catégorie fonctionnelle est ensuite principalement représentée par le pouce engagé (*thumb engaged*, TE = 10,53 %).

Certains types fonctionnels présentent un profil différent de celui décrit précédemment. Les gestes assurant un rôle de *Déictique* et ceux sans type fonctionnel défini (*Undefined*) sont majoritairement réalisés avec la main partiellement ouverte (*semi open hand*, SOH = 40 %),

suivis de la main ouverte (*open hand*, OH entre 30 % et 37 %), puis de l'index engagé (*index engaged*, IE entre 19 % et 24 %). Les gestes avec une fonction d'*Iconique*, quant à eux, présentent des proportions similaires pour les configurations de la main partiellement ouverte (SOH) et ouverte (OH) (33,33 % chacune), suivies de l'index engagé (IE = 18,86 %), représentant ainsi les configurations de main les plus fréquentes pour cette catégorie. Enfin, les gestes assurant un rôle de *Phatique* se distinguent nettement par une prédominance de la configuration index engagées (IE), qui représente 95,66 % des gestes de ce type fonctionnel. Cette prédominance peut s'expliquer par le rôle de ces gestes dans la gestion de l'interaction, qui servent tour à tour à souligner l'adressage à l'interlocuteur, à attirer son attention, à lui donner le tour de parole, etc. On voit ainsi dans la reproduction de la trajectoire du geste A (Figure 25) que Wayne produit un geste *Phatique*, avec son index engagé, lorsqu'il énonce « If you want [...] », marquant ainsi son interlocuteur comme destinataire direct de la proposition en cours.

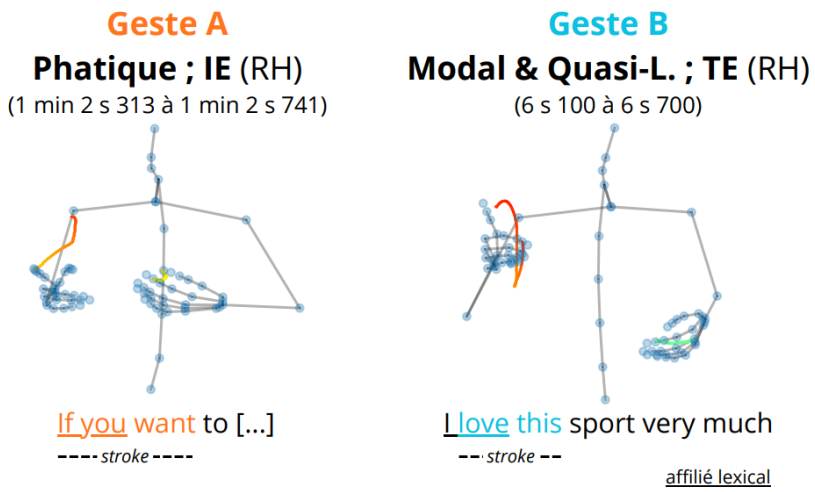



*Figure 25. Geste phatique (A) avec l'index engagé (IE) pour la main droite (RH) et geste B modal et quasi-linguistique (Quasi-L.) avec le pouce droit engagé (TE)*

Si l'on analyse les configurations de main les moins représentées dans le corpus, on observe que la configuration main fermée (CH) est majoritairement associées aux gestes qui assurent un rôle d'*Iconique* (6 sur 8 annotations), la main partiellement fermée (*mid closed hand* MCH) aux gestes ayant pour fonction l'étiquette *Modal* (11 sur 24 annotations), et le pouce engagé (*thumb engaged* TE) aux gestes assurant une fonction de *Modal* ou *Quasi-linguistique* en proportions équivalentes (4 sur 10 annotations pour chacune). La prédominance de ces deux types fonctionnels s'explique notamment par l'usage fréquent du pouce levé pour signifier « super », par exemple le geste B de la Figure 25 (Wayne utilise son pouce en l'air pour exprimer qu'il aime le sport). Dans ce contexte, cette configuration remplit à la fois une fonction modale (expression d'une appréciation) et une fonction quasi-linguistique, puisqu'elle peut être comprise indépendamment de toute parole. Enfin, la configuration avec le pouce, l'index, et le majeur engagé (*thumb, middle, index engaged* TMIE) est majoritairement représentée par des gestes assurant une fonction *Métaphorique* (6 sur 12 annotations).

# 5 Analyse cinématique des gestes

## 5.1 Méthodologies d'extractions de descripteurs du mouvement et de visualisation de trajectoires

### 5.1.1 Calcul de la vitesse

Pour chaque type fonctionnel annoté (D1 et D2), les données de mouvements de l'index droit et/ou de l'index gauche (en fonction de la manualité du geste réalisé) ont été regroupées afin d'extraire les positions à chaque *frame* et ainsi déterminer la vitesse moyenne. Pour chaque segment, les vecteurs de déplacement entre *frames* consécutives ont été calculés par différenciation successive des positions dans l'espace 3D (X, Y, Z), puis la norme de ces vecteurs a été déterminée pour obtenir la vitesse scalaire à chaque *frame*. Cette vitesse a été mise à l'échelle par le nombre de *frames* par seconde ($fps = 120$) afin d'obtenir des valeurs en cm/s. Pour pouvoir comparer des segments de longueurs différentes, une matrice a été construite pour contenir toutes les courbes de vitesse, en complétant par des valeurs `NaN` lorsque les segments étaient plus courts que la longueur maximale observée. À partir de cette matrice, des statistiques *frame-wise* ont été calculées : la moyenne (courbe rouge), la médiane (courbe verte en pointillés) et l'écart-type (zone rouge) des vitesses à chaque *frame*. Ces mesures permettent de visualiser la dynamique temporelle moyenne ainsi que la variabilité des vitesses au sein de chaque type fonctionnel de geste. En complément, la vitesse moyenne, médiane et l'écart type ont été calculés sur la globalité des segments, indépendamment de leur position dans le temps, afin de fournir une mesure globale des caractéristiques de vitesse pour chaque type fonctionnel.

### 5.1.2 Calcul de la distance parcourue et amplitude maximum

La somme des distances parcourues à chaque *frame* du geste peut refléter soit l'amplitude spatiale du geste (par exemple, des gestes en ligne particulièrement amples), soit la durée du mouvement (par exemple, des enchaînements de boucles peu amples, mais répétés). La distance parcourue par le geste a été calculée à partir des positions 3D de l'index, enregistrées *frame* par *frame*. Pour chaque geste, les différences entre positions consécutives ont été calculées afin d'obtenir les vecteurs de déplacement instantanés. La norme de ces vecteurs correspond au déplacement instantané par *frame*, et la distance totale parcourue (D) pendant le geste a été obtenue en sommant ces déplacements : $D = \sum_{i=1}^{N-1} \|p_{i+1} - p_i\|$ où $\mathrm{p}_i$ représente la position de l'index à la *frame* $i$.

L'amplitude maximale ($A_{max}$) du geste a été calculée comme la distance maximale (en cm) entre la position initiale de l'index et n'importe quelle position atteinte au cours du geste : $A_{max} = max_i \|p_{i+1} - p_i\|$ . Cette mesure reflète l'extension spatiale globale du geste, indépendamment du trajet réellement parcouru. Elle permet de distinguer les gestes très amples dans l'espace des gestes qui couvrent une grande distance cumulée en restant concentrés sur un espace restreint.

### 5.1.3 Visualisation des trajectoires

Les trajectoires (depuis les positions XYZ) des index droit et gauche ont été reconstruites dans un espace 3D pour chaque activité annotée. Seule(s) la (les) main(s) impliquée(s) a été considérée, comme renseignée dans l'annotation de la manualité. Un code couleur a été appliqué en fonction de la distance verticale (en cm) par rapport à un point de référence (hanche = 0, en bleu) jusqu'au sommet de la tête (en rouge), les points situés sous la hanche étant colorés en bleu foncé pour faciliter l'interprétation des dynamiques du mouvement sur la hauteur. Cette représentation permet aussi d'observer l'utilisation de l'axe latéral (droite-gauche) lors de l'élaboration du *stroke* par le locuteur.

## 5.2 Vitesse

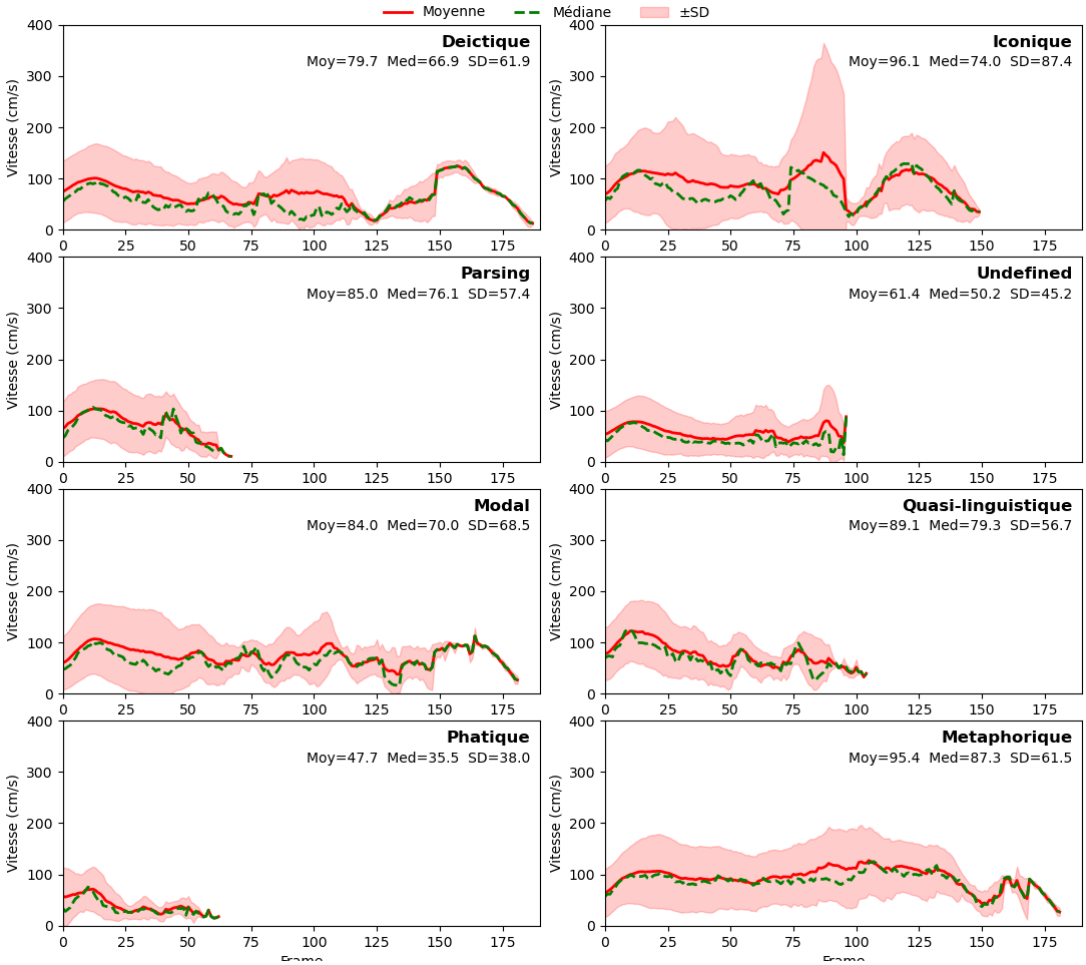



*Figure 26. Analyse temporelle de la vitesse moyenne des gestes en fonction du type fonctionnel (moy = moyenne ; med = médiane ; SD = écart-type)*

L'analyse des vitesses globales (Figure 26) révèle que les gestes présentent une vitesse moyenne de 82,3 cm/s (médiane = 70,3 cm/s, écart-type = 62,3 cm/s). En comparant les valeurs pour chaque type fonctionnel, on s'aperçoit que certaines catégories, telles que les gestes assurant une fonction *Déictique*, *Modale* ou de *Parsing*, présentent une vitesse moyenne proche de cette valeur globale, entre 80 et 85 cm/s. À l'inverse, les gestes ayant un rôle d'*Iconique*, de

*Quasi-linguistique* ou de *Métaphorique* se distinguent par des vitesses plus élevées, comprises entre 89 et 95 cm/s. Les gestes dont le type fonctionnel n'a pas été identifié montrent une vitesse plus faible, autour de 60 cm/s (médiane = 50 cm/s), tandis que les gestes assurant un rôle *Phatique* apparaissent comme les plus lents du corpus, avec une vitesse moyenne de 47 cm/s et une médiane de 35,5 cm/s.

Les durées des gestes varient effectivement entre les catégories, avec une proportion décroissante de gestes longs, ce qui réduit la précision des estimations pour les extrémités des courbes de vitesse. L'analyse des variations temporelles de la vitesse moyenne révèle cependant une tendance générale à une augmentation des vitesses pendant les premiers instants des gestes (les 20 premières *frames*), quel que soit le type fonctionnel. Par la suite, pour tous les types à l'exception des gestes assurant un rôle *Métaphorique*, on observe une diminution progressive de la vitesse au fil du temps. Pour les gestes étiquetés *Modaux*, *Quasi-linguistiques*, *Déictiques* et *Iconiques*, la vitesse présente ensuite des fluctuations après cette diminution initiale, ce qui pourrait s'expliquer par la présence de gestes dans ces catégories dont la vitesse augmente à un moment plus tardif. Ces fluctuations ne sont pas observées pour les gestes de *Parsing* ou *Phatiques*. Ces variations sont moins prononcées, mais fréquentes pour les gestes ayant une fonction *Modale* ou *Quasi-linguistique*, tandis qu'elles sont plus marquées pour les gestes assurant un rôle *Déictique* ou *Iconique*. Les gestes labélisés *Métaphorique*, en revanche, apparaissent relativement stables en termes de vitesse moyenne tout au long de leur déroulement.

## 5.3 Utilisation de l'espace

### 5.3.1 Distance parcourue

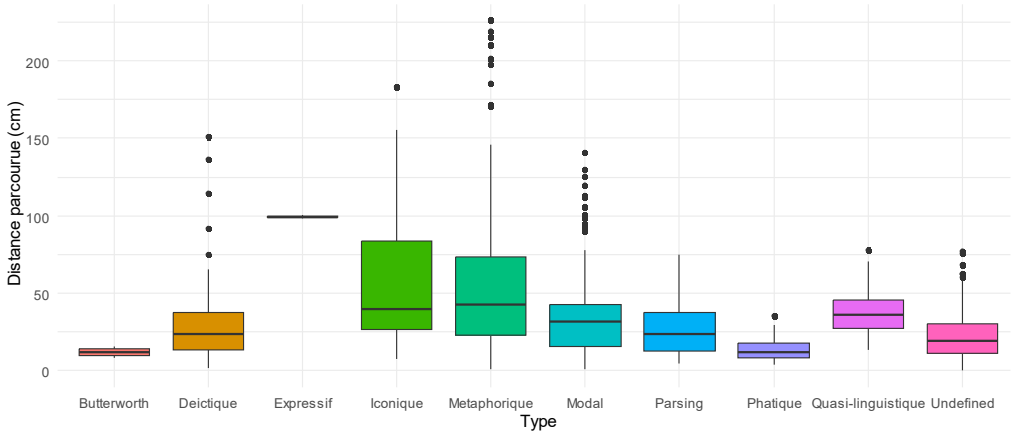



*Figure 27. Boxplot des distances parcourues (cm) en fonction du type fonctionnel annoté*

| Type | Moy | Med | SD |
|---|---|---|---|
| Butterworth | 11,85 | 11,85 | 5,65 |
| Déictique | 28,95 | 23,45 | 22,46 |
| Iconique | 56,07 | 39,80 | 43,54 |
| Métaphorique | 55,39 | 42,78 | 47,19 |
| Modal | 35,97 | 31,65 | 27,98 |
| Parsing | 26,77 | 23,58 | 16,56 |
| Phatique | 14,24 | 11,92 | 8,10 |
| Quasi-linguistique | 38,26 | 36,32 | 15,32 |
| Expressif | 99,26 | 99,26 | 1,28 |
| Undefined | 22,15 | 19,10 | 14,68 |
| **Global** | **34,94** | **27,36** | **31,82** |

*Tableau 28. Statistiques descriptives des distances parcourues (cm) par type fonctionnel (moy = moyenne ; med = médiane ; sd = écart-type)*

L'analyse de la distance parcourue (Figure 27, Tableau 28) montre que, quel que soit le type fonctionnel, l'index (droite et gauche confondues) parcourt en moyenne environ 35 cm (médiane = 27,36 cm, écart-type = 31,92 cm). Le corpus comprend à la fois des gestes très courts (minimum 0,45 cm) et des gestes très longs, atteignant jusqu'à 2,26 m, correspondant dans le corpus à des gestes circulaires particulièrement amples. Ces observations mettent en évidence une grande variabilité des distances parcourues par les mains.

Certaines catégories de gestes se rapprochent de la distance moyenne globale, avec des parcours compris entre 26 et 40 cm. C'est le cas notamment des gestes assurant un rôle *Déictique*, de *Parsing*, de *Modal*, ou de *Quasi-linguistique*. Les gestes dont l'une des dimensions n'a pas pu être identifiée sur sa fonction (*Undefined)* semblent quant à eux présenter une moindre distance parcourue en moyenne (22,5 cm). À l'inverse, les gestes ayant une fonction d'*Iconique* ou *Métaphorique* parcourent en moyenne une distance plus importante, d'environ 56 cm, et présentent une variabilité plus marquée, avec plusieurs gestes très longs. Enfin, les gestes étiquetés *Phatique* se distinguent par leur homogénéité dans la distribution des distances parcourues et leur distance relativement courte, avec une moyenne de 14 cm (médiane = 11,92 cm) et un écart-type faible en comparaison à l'écart-type des autres types de gestes.

### 5.3.2 Amplitude maximum

| Type | Moy | Med | SD |
|---|---|---|---|
| Butterworth | 6,08 | 6,08 | 2,29 |
| Déictique | 21,16 | 18,22 | 14,54 |
| Iconique | 30,41 | 24,26 | 22,69 |
| Métaphorique | 28,58 | 23,18 | 20,09 |
| Modal | 24,18 | 20,11 | 17,88 |
| Parsing | 21,12 | 18,74 | 13,41 |
| Phatique | 10,88 | 9,36 | 6,03 |
| Quasi-linguistique | 24,47 | 20,47 | 12,00 |
| Expressif | 62,12 | 62,12 | 2,33 |
| Undefined | 17,05 | 15,24 | 10,84 |
| **Global** | **22,49** | **18,67** | **16,55** |

*Tableau 29. Analyse descriptive de l'amplitude maximale en fonction du type fonctionnel (moy = moyenne ; med = médiane ; SD = écart-type)*

L'analyse de l'amplitude maximale (Tableau 29) montre que, dans l'ensemble, les gestes sont réalisés en moyenne à 23 cm de la position de départ du *stroke* (médiane = 18,67 cm, écart-type = 16,55 cm), avec des amplitudes très faibles de 0,20 cm jusqu'à des gestes très amples de 96 cm. Certaines catégories fonctionnelles présentent des amplitudes proches de la moyenne globale, notamment les gestes assurant un rôle de *Déictique*, de *Parsing*, de *Quasi-linguistique* ou de *Modal*, avec des amplitudes maximales comprises entre 21 et 24 cm en moyenne (médiane autour de 20 cm). Les gestes ayant une dimension labélisée *Undefined* se distinguent par des amplitudes légèrement plus faibles (moyenne = 17 cm, médiane = 15,25 cm), reflétant des gestes occupant moins d'espace. De même, les gestes avec la fonction de *Phatique* présentent des amplitudes limitées, avec une moyenne de 10 cm et une médiane proche de la moyenne, ce qui semble caractéristique de cette catégorie. Enfin, les gestes ayant un rôle d'*Iconique* ou *Métaphorique* se distinguent par des amplitudes plus importantes, autour de 30 cm en moyenne, et par une grande variabilité, certains gestes s'éloignant fortement de la position de départ.

### 5.3.3 Latéralité et hauteur

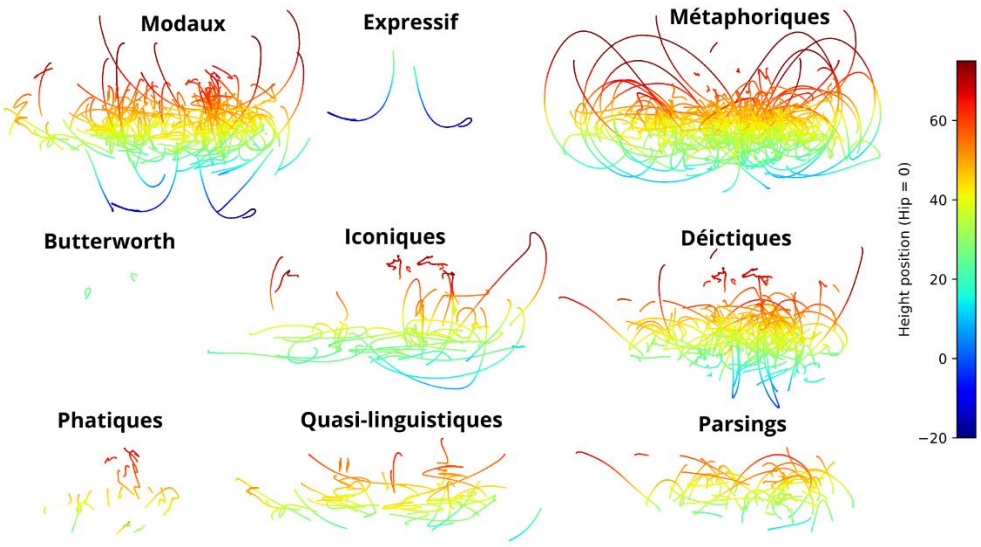



*Figure 28. Visualisation des trajectoires par type fonctionnel annoté*

On remarque que la majorité des trajectoires, quel que soit le type fonctionnel, se situe autour de 40 cm au-dessus des hanches (zones vertes à jaunes), ce qui indique que les *strokes* sont en général centrés au niveau du torse (Figure 28). Certaines catégories de gestes occupent cependant davantage l'espace en hauteur, au niveau de la tête (zones rouges) : on peut citer les gestes assurant un rôle *Modal* ou *Métaphorique*, qui dépassent fréquemment 70 cm au-dessus des hanches. Cette surreprésentation des gestes ayant un rôle *Modal*, pour cette hauteur, s'explique notamment par la présence dans le corpus de nombreux gestes exprimant l'exclamation, que le locuteur figure en levant ses deux bras de part et d'autre de son torse.

Les gestes ayant une fonction d'*Iconique* ou *Déictique* se montent également fréquemment au niveau de la tête, mais de manière plus ponctuelle. À l'inverse, très peu de gestes sont réalisés proches des hanches (zones bleues) : seuls les gestes ayant une fonction *Modale* peuvent être positionnés très bas, tandis que d'autres catégories comme *Métaphorique* et *Déictique* se situent légèrement en dessous du centre du torse (bleu clair). Cette visualisation permet également d'observer l'utilisation de l'espace sur l'axe latéral. On constate que certains types de gestes, comme les gestes ayant un rôle *Phatique*, restent relativement concentrés autour du torse sur cet axe. En revanche, les gestes assurant un rôle *Métaphorique*, *Iconique* ou *Modal* occupent une plus grande amplitude latérale, s'étendant aussi bien à gauche qu'à droite du corps. Pour les gestes remplissant un rôle *Métaphorique*, l'utilisation étendue de l'espace peut s'expliquer par le fait que le locuteur représente les concepts de « importance », de « tout » ou de « beaucoup » par une large occupation de l'espace gestuel.

# 6 Analyse de l'ancrage des gestes au discours

## 6.1 Brève description de la production verbale

Le corpus comporte 130 phrases ponctuées, pour un total de 2 182 mots. La distribution du nombre de mots par phrase met en évidence une tendance générale vers des énoncés relativement longs, avec une médiane et une moyenne autour de 16 mots et un maximum de 47 mots pour un minimum de 3. Ce profil suggère une production discursive globalement développée, caractéristique de séquences explicatives ou narratives – typiques de ces monologues – dans lesquelles les locuteurs formulent des énoncés relativement élaborés. La production verbale inclut 351 silences inter ou intra énoncés détectés automatiquement par l'ASR, soit un total d'environ 2 minutes 45 secondes sur les 13 minutes que dure le corpus, représentant ainsi environ 21 % du temps enregistré. La parole comprend un total de 388 événements annotés, correspondant aux différentes balises prosodiques et discursives insérées au sein des phrases lors de l'annotation. Les silences (`</PAUSE/>`) sont les plus représentés avec 314 occurrences. L'écart observé entre le nombre de balises `</PAUSE/>` et les silences détectés par l'ASR peut tenir, d'une part, à l'annotation des silences en début et en fin de fichier dans la transcription automatique — contrairement aux annotatrices qui n'ont relevées que les pauses perçues à l'intérieur et entre les énoncés — et, d'autre part, à un seuil de détection probablement plus fin dans le système automatique. Les phénomènes liés à l'accentuation (`<ACC><ACC/>`) apparaissent 36 fois. Les segments d'élongation (`<ELO><ELO/>`) sont plus rares, avec 11 occurrences, tandis que les erreurs (`<ERR><ERR/>`) sont très peu fréquentes, avec seulement 8 événements annotés. La prédominance des balises de silence qui ponctuent le discours, associée à la rareté des phénomènes oraux annotés, suggère que la production orale analysée s'inscrit dans un discours largement préparé.

## 6.2 Affiliation lexicale selon le type fonctionnel

Sur l'ensemble des 2 182 mots du corpus, 607 ont été identifiés comme étant associés à la réalisation d'un geste. Pour un geste bidimensionnel, l'affiliation peut être identique ou différer entre les deux dimensions. Par conséquent, l'analyse de l'affiliation selon le type fonctionnel est menée séparément pour la D1 et la D2. L'étiquette *Undefined* étant exclue de cette analyse (pas d'affiliation lexicale), 341 étiquettes fonctionnelles ont été prises en compte en D1 et 100 en D2. Sur les 345 gestes étiquetés selon au moins une dimension fonctionnelle identifiée, 230 des 243 gestes unidimensionnels (94,65 % des gestes ayant une D1 uniquement) présentent une affiliation lexicale. Sur les 102 gestes bidimensionnels (D1 et D2), 92 présentent une affiliation lexicale pour chacune des dimensions identifiée (90,20 %), tandis que 4 ne présentent aucune affiliation en D1 ou en D2 (3,92 %). Pour les six gestes bidimensionnels dont l'une des dimensions est annotée *Undefined*, un seul ne présente pas d'affiliation lexicale pour la dimension fonctionnelle identifiée.

| | Min | Moy | Med | Max |
|---|---|---|---|---|
| **D1+D2** | **1** | **1,84** | **1,50** | **7** |
| Dimension 1 | 1 | 1,85 | 2 | 7 |
| Dimension 2 | 1 | 1,82 | 1 | 5 |
| Phatique | 1 | 2,33 | 2 | 5 |
| Parsing | 1 | 2,25 | 2 | 5 |
| Iconique | 1 | 1,74 | 1 | 4 |
| Modal | 1 | 1,53 | 1 | 5 |
| Quasi-linguistique | 1 | 1,23 | 1 | 2 |

*Tableau 30. Analyse du nombre de mots par affilié*
*(D1 = dimension 1 ; D2 = dimension 2 ; min = minimum ; moy = moyenne ; med = médiane ; max = maximum)*

L'analyse du nombre de mots par affilié (Tableau 30) montre que, dans le corpus, un affilié compte au minimum 1 mot et au maximum 7 mots, avec une moyenne d'environ 1,84 mot (médiane = 1,50) toutes catégories confondues (affiliation du type fonctionnel toutes dimensions confondues). La distinction du nombre de mots des affiliés en D1 et D2 ne révèle pas de différence significative sur la moyenne, mais la médiane (1 mot en D2, contre 2 en D1) montre une légère tendance à des affiliés plus courts lorsque la fonction est annotée en D2.

L'analyse par type fonctionnel ne montre pas de profils distincts avec l'analyse du nombre de mots globale (D1+D2), sauf pour les gestes ayant une fonction de *Phatique*, de *Parsing,* d'*Iconique,* de *Modal* et de *Quasi-linguistique*. Pour les deux premières catégories, bien que la médiane reste à 2 mots, la moyenne augmente légèrement (2,25–2,33 mots/affilié), ce qui indique la présence de quelques affiliés plus longs pour ces catégories qui tirent la moyenne générale vers le haut. À l'inverse, pour les gestes assurant un rôle *Quasi-linguistique* (moy = 1,23 mot/affilié), *Iconique* (moy = 1,74 mot/affilié) ou *Modal* (moy = 1,53 mot/affilié), on observe un nombre moyen de mots légèrement plus faible, avec une médiane égale à 1, indiquant une tendance à des affiliations plus courtes en termes de nombre de mots. Cette tendance peut s'expliquer par le fait que les gestes *Iconiques* représentent souvent un objet ou une entité unique énoncée dans le discours, ou encore que les gestes *Modaux* accompagnent fréquemment des éléments grammaticaux courts, tels que des termes de négations.

| **Type** | **Dimension 1** | | **Dimension 2** | | **Total** | |
|---|---|---|---|---|---|---|
| | *N* | *%* | *N* | *%* | *D1* | *D2* |
| Butterworth | 1 | 50 | | - | 2 | - |
| Déictique | 96 | 98,97 | 21 | 95,46 | 97 | 22 |
| Iconique | 23 | 88,46 | 4 | 100 | 26 | 4 |
| Métaphorique | 94 | 100 | 9 | 100 | 94 | 9 |
| Modal | 50 | 83,33 | 43 | 91,49 | 60 | 47 |
| Parsing | 33 | 100 | 7 | 100 | 33 | 7 |
| Phatique | 9 | 81,81 | 6 | 100 | 11 | 6 |
| Quasi-linguistique | 18 | 100 | 4 | 100 | 18 | 4 |
| Expressif | | - | 1 | 100 | - | 1 |
| **Avec affiliation** | 324 | 95,01 | 95 | 95 | 341 | 100 |
| **Sans affiliation** | 17 | 4,98 | 5 | 5 | | |

*Tableau 31. Décompte de types fonctionnels avec une affiliation lexicale*
*(N = nombre d'occurrences ; D1 = dimension 1 ; D2 = dimension 2)*

Dans l'ensemble, on observe une majorité de gestes dont le type fonctionnel est associé à une affiliation lexicale (Tableau 31). En comparant D1 et D2, l'absence d'affiliation est relativement similaire, puisqu'elle concerne 4,98 % seulement des gestes avec une dimension 1 clairement identifiée et 5 % des gestes ayant une dimension 2 clairement identifiée. Certains types fonctionnels présentent presque systématiquement une affiliation lexicale (en D1 et D2), notamment les gestes assurant une fonction de *Déictique*, de *Métaphorique*, de *Parsing* ou de *Quasi-linguistique*. Cependant, en première dimension pour les gestes ayant un rôle d'*Iconique*, de *Modal* ou de *Phatique*, entre 10 et 20 % des annotations ne présentent pas d'affiliation.

Il n'est pas surprenant que certaines catégories fonctionnelles présentent des scores très élevés d'affiliation lexicale. Les gestes assurant un rôle de *Quasi-linguistique*, quand ils sont associés à du verbal, sont généralement associés à une expression plutôt fixe (ex. : secouer la main pour accompagner une salutation). De même, les gestes ayant pour fonction de *Parsing* entretiennent un lien étroit avec la parole, car ils agissent comme des ponctuants discursifs visuels. Enfin, les gestes *Iconiques, Métaphoriques* ou *Déictiques* sont également fortement liés au discours, en ce qu'ils servent à référer ou à représenter un élément du contenu verbal. Ces différentes propriétés fonctionnelles permettent ainsi de relier facilement ces gestes au terme ou à l'expression à laquelle ils se rapportent lorsqu'ils sont produits pendant l'énonciation verbale.

Il convient toutefois de nuancer les pourcentages indiquant une absence d'affiliation pour certaines catégories. Dans le cas des gestes *Modaux*, par exemple, s'il est fréquent qu'aucune affiliation lexicale ne soit mentionnée, c'est souvent parce que l'affiliation est en réalité davantage d'ordre prosodique, par exemple lors de gestes soulignant une exclamation. Dans d'autres cas, l'affiliation au discours est avant tout implicite ; c'est le cas pour quatre gestes ayant un rôle d'*Iconique* en D1 et un *Déictique* en D2 sont décomptés, dont on trouve deux exemples ci-dessous.

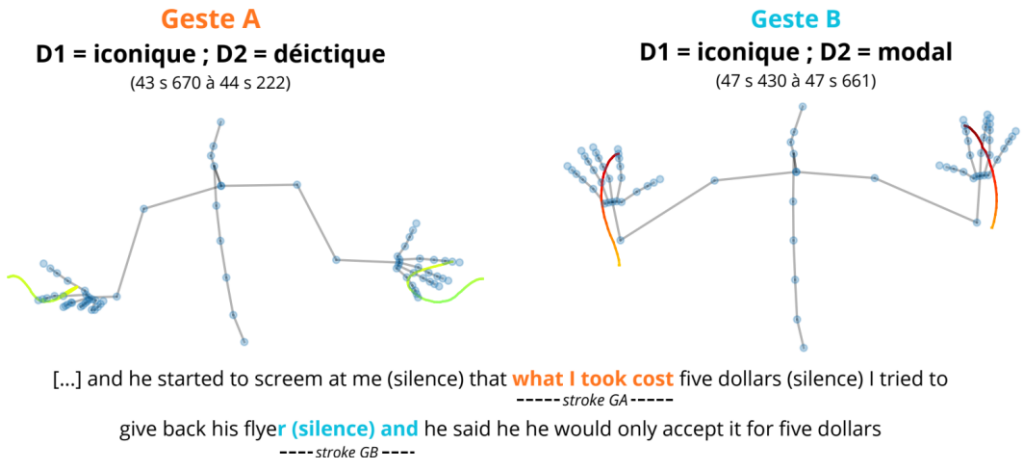



*Figure 29. Illustration de gestes sans affiliation lexicale explicite*

La Figure 29 illustre deux gestes réalisés par Wayne alors qu'il raconte une altercation qu'il a eue avec quelqu'un (« he »). Son récit porte sur le moment où l'interlocuteur lui a crié

dessus pour avoir pris un flyer, que Wayne croyait gratuit, mais que l'interlocuteur présente comme valant 5 dollars. Wayne explique ensuite qu'il a voulu le rendre, mais que l'interlocuteur a refusé de le reprendre en exigeant d'être payé 5 dollars. Tout au long de ce récit, Wayne élabore des gestes imitant les gestes produits par son interlocuteur au moment des faits.

Le geste A (Figure 29) est produit lorsque Wayne dit « what I took cost ». Pendant ce passage, il ouvre d'abord les bras, puis les avance vers l'avant tout en les rapprochant, avant de les ramener en arrière en les écartant à nouveau. Le geste est une désignation (D2) renvoyant au référent « what I took » (le flyer), et qui se faisant est un geste représentant l'action de l'interlocuteur (kinémimique, D1). L'affiliation de ce geste est donc en partie implicite, dans le sens où le geste ne peut pas être raccroché directement au « what I took » puisqu'il s'agit d'un geste rapporté, donc qui ne représente pas l'action de prendre mais une désignation par l'interlocuteur de ce qui a été pris par Wayne.

Le geste B (Figure 29) est réalisé en levant les deux mains vers le haut. Il s'agit à nouveau ici d'un geste iconique, plus précisément kinémimique (D1), dans la mesure où Wayne rejoue cette fois l'attitude de son interlocuteur, qui exprimait ainsi son refus de récupérer le flyer (modal, D2). Dans ce cadre, le geste ne peut pas être directement affilié à l'énonciation verbale « he would only accept », puisque ce qu'exprime le geste est littéralement un refus. Ainsi, l'affiliation lexicale ne peut pas être explicitement établie, bien qu'une cohérence narrative soit perceptible.

Dans tous les cas, l'analyse des affiliations lexicales annotées dans le corpus atteste d'un fort ancrage des gestes au discours, 94,78 % des gestes ayant pu être directement reliés à un segment verbal.

# 7 Conclusion

L'annotation multimodale des gestes met en évidence des tendances générales concernant les caractéristiques formelles et fonctionnelles de la production gestuelle du locuteur au sein du corpus étudié. Dans les grandes lignes, les gestes observés sont majoritairement monodimensionnels (une seule fonction assurée) et correspondent à des événements brefs, dont la durée moyenne est de 0,43 s (SD = 0,24 s). Ils sont principalement produits à deux mains (en miroir), avec des trajectoires relativement simples (ligne ou courbe). Les mains sont en général partiellement ouvertes, orientées horizontalement et les paumes tournées sur la tranche. L'analyse cinématique permet également d'estimer que les gestes sont produits à une vitesse moyenne d'environ 82 cm/s (SD = 62 cm/s), qu'ils couvrent des distances variables, parfois très courtes ou très longues, avec des amplitudes pouvant atteindre jusqu'à 30 cm par rapport au point de départ du geste. Bien que le locuteur exploite une large portion de l'espace gestuel, les productions demeurent majoritairement centrées autour du milieu du torse. Enfin, les gestes sont le plus souvent explicitement affiliés au discours, cette affiliation portant en moyenne sur deux mots.

| Légende<br>∼ : dans la moyenne<br>+ : valeur haute<br>− : valeur basse | ***Déictique (Dei.)*** | ***Iconique*** | ***Métaphorique*** | ***Modal (Mod.)*** | ***Parsing (Par.)*** | ***Phatique*** | ***Quasi-linguistique*** (Qua.) | ***Undefined*** |
|---|---|---|---|---|---|---|---|---|
| **Fréquente en D1 ou D2** | D1 | D1 | D1 | D1<br>D2 | D1 | D1 | D1 | D1 |
| **Type fonctionnel apparié fréquent** | Par. | | Mod. | Qua. | Dei. | Mod. | Mod. | - |
| **Durée** | ∼ | + | + | ∼ | − | − | ∼ | ∼ |
| **Manualité** | RH | RH | MH | MH | MH<br>RH<br>LH | RH | MH | MH |
| **Forme de trajectoire** | Courbe | Courbe<br>Ligne<br>*Hapax* | Coubre<br>Ligne | Ligne | Ligne | Ligne | Ligne<br>BF | Ligne<br>Autre |
| **Position de la main** | H<br>I | H | H<br>I | I<br>H | I | H<br>I<br>V | I | H |
| **Orientation de la paume** | E<br>SO | SO<br>E<br>U | E<br>SO | E<br>MU | MU<br>E | E | E<br>Fr | E<br>MU |
| **Configuration des mains** | SOH<br>OH | OH<br>SOH | OH | OH | OH<br>SOH | IE | OH | SOH<br>OH |
| **Vitesse, D, $A_{max}$** | ∼ | + | + | ∼ | ∼ | −− | ∼ | − |
| **Hauteur et latéralité** | + | + | ++ | ++ | ∼ | − | ∼ | − |
| **Affiliation lexicale (mots)** | 2 mots | 2 mots | 2 mots | 2 mots | 2-3 mots | 2-3 mots | 1 mot | |

*Tableau 32. Caractéristiques majoritaires des types fonctionnels déterminées lors des différentes analyses (RH = main droite ; MH = mains en miroir ; LH = main gauche ; H = horizontal ; I = intermédiaire ; V = vertical ; E = paume sur la tranche ; SO = paume vers le locuteur ; U = paume vers le haut ; MU = paume partiellement vers le haut ; Fr = paume vers l'interlocuteur ; SOH = main partiellement ouverte ; OH = main ouverte ; IE = index engagé)*

Bien que cette description moyenne du geste fournisse une cible pertinente pour guider l'apprentissage des modèles, l'objectif ne se limite pas à reproduire un comportement gestuel prototypique : il s'agit également de capturer les spécificités propres à certains types de gestes. Les analyses menées ont en effet mis en évidence l'existence de profils distincts selon les catégories fonctionnelles annotées, révélant des récurrences de variations en termes de propriétés spatiales, temporelles et cinématiques. Le Tableau 32 synthétise ainsi les caractéristiques les plus fréquemment observées pour chaque type fonctionnel (D1 et D2 confondues).

- Le **type fonctionnel *Déictique*** est fréquemment annoté en dimension principale et apparié à la fonction de *Parsing*. Les gestes assurant une fonction de déictique présentent une durée comparable à la durée moyenne des gestes du corpus. Ils sont majoritairement réalisés de la main droite (Wayne étant droitier) positionnée à l'horizontale ou sur un axe intermédiaire, avec une configuration ouverte ou partiellement ouverte. Leur trajectoire est le plus souvent courbe et leur dynamique (vitesse, distance parcourue et amplitude) se situe dans la moyenne observée pour l'ensemble des gestes. Ces gestes occupent soit le centre du torse, soit une zone

légèrement périphérique tant en hauteur qu'en latéralité. Enfin, leur affiliation lexicale est généralement composée de deux mots.

- Le **type fonctionnel *Iconique*** est également fréquemment annoté en dimension principale. Les gestes assurant cette fonction présentent généralement des durées plus longues que la durée moyenne des gestes du corpus. Ils sont eux aussi majoritairement réalisés de la main droite, mais cette fois la paume est le plus souvent orientée soit vers le locuteur, soit vers le haut, soit sur la tranche, et la configuration de la main est le plus souvent ouverte ou partiellement ouverte. Les trajectoires de ces gestes peuvent être des lignes, des courbes ou des formes variées isolées (*Hapax*). Ces gestes se distinguent par une vitesse, une distance parcourue et une amplitude supérieures aux valeurs moyennes de l'ensemble des gestes du corpus. L'espace mobilisé lors de leur production peut s'étendre légèrement au-dessus du centre du torse du locuteur ou latéralement, lorsqu'ils ne sont pas centrés. Leur affiliation lexicale est également généralement composée de deux mots.

- Le **type fonctionnel *Métaphorique***, majoritairement annoté en dimension principale, est fréquemment associé au type fonctionnel *Modal* en D2. Les gestes assurant cette fonction métaphorique présentent une durée plus longue que la durée moyenne des gestes du corpus. Ils sont majoritairement réalisés avec les mains en miroir, positionnées à l'horizontale ou sur un axe intermédiaire, avec une configuration des doigts ouverte et une paume orientée sur la tranche ou vers le locuteur. Leur trajectoire est généralement soit courbe soit linéaire, et est produite selon une vitesse et une amplitude supérieure aux moyennes observées. Ces gestes se distinguent par une mobilisation particulièrement expansive de l'espace du locuteur, plus particulièrement par le fait qu'une seule trajectoire peut mobiliser l'ensemble de cet espace (à la fois en hauteur et en largeur). Leur affiliation verbale reste le plus souvent constituée de deux mots.

- Le **type fonctionnel *Modal*** est fréquemment mobilisé à la fois en dimension principale et en dimension secondaire. Il est le plus souvent associé au type fonctionnel *Quasi-linguistique*. Les gestes assurant cette fonction de *Modal* présentent généralement une durée supérieure à la durée moyenne des gestes du corpus. Ils sont majoritairement réalisés avec les mains en miroir, placées en position intermédiaire ou horizontale et selon une configuration des doigts ouverte. La paume est généralement orientée sur la tranche ou partiellement vers le haut. Leur trajectoire est principalement linéaire, avec une vitesse, une distance parcourue et une amplitude proches de la moyenne des gestes du corpus. Comme les métaphoriques, ils se distinguent par une mobilisation expansive de l'espace du locuteur, tant en hauteur (vers le haut et vers le bas) que latéralement quand ils ne sont pas centrés sur le torse. Cependant, ils comportent moins de gestes dont la trajectoire occupe l'ensemble de l'espace du locuteur, comparativement aux gestes métaphoriques. Leur affiliation lexicale est là aussi le plus souvent constituée de deux mots.

- Le **type fonctionnel *Parsing*** est majoritairement mobilisé en dimension principale et fréquemment associé à la fonction *Déictique*. Les gestes assurant cette fonction de *Parsing* se caractérisent par une durée courte par rapport à la durée moyenne des gestes du corpus. La manualité est variable : les gestes peuvent être réalisés avec les mains en miroir ou en mobilisant une seule main (droite ou gauche). La position de la ou des mains impliquées est majoritairement intermédiaire, avec la paume partiellement orientée vers le haut ou sur la tranche, et une configuration des doigts ouverte ou partiellement ouverte. Leur trajectoire est généralement linéaire, et leur durée, leur vitesse, la distance parcourue et leur amplitude sont comparables aux valeurs moyennes des gestes du corpus. L'affiliation lexicale est parfois plus longue, étant composée de deux à trois mots en moyenne.

- Le **type fonction *Phatique*** est fréquemment mobilisé en dimension principale et souvent associé au type fonctionnel *Modal*. Les gestes assurant cette fonction de *Phatique* présentent une durée plus courte que la durée moyenne des gestes du corpus et sont majoritairement réalisés de la main droite. La position de la main est variable : elle peut être horizontale, verticale ou intermédiaire, et le plus souvent avec l'orientation de la paume sur la tranche et l'index engagé. Les trajectoires sont principalement linéaires, et adoptent une vitesse, une amplitude et une distance très faibles. L'utilisation de l'espace du locuteur est ainsi très restreinte pour ce type de gestes généralement concentrés au niveau du centre du torse. Comme pour les gestes de *Parsing*, l'affiliation lexicale des gestes *Phatiques* est composée de deux à trois mots.

- Le **type fonctionnel *Quasi-linguistique*** est majoritairement mobilisé en dimension principale et souvent associé au type fonctionnel *Modal*. Les gestes assurant cette fonction présentent une durée comparable à durée moyenne des gestes du corpus. Ils sont majoritairement réalisés avec les mains en miroir, positionnées sur un axe intermédiaire et selon une configuration des doigts ouverte. La paume est orientée soit sur la tranche, soit vers l'interlocuteur. Les trajectoires sont généralement linéaires ou en aller-retour, avec une vitesse, une amplitude et une distance parcourue proche des valeurs moyennes de l'ensemble des gestes du corpus. La production se fait principalement au centre du torse. Leur affiliation lexicale se limite généralement à un seul mot.

- Le label *Undefined*, utilisés pour les gestes dont la dimension fonctionnelle n'a pas pu être déterminée, a principalement été utilisé pour la dimension principale. Ces gestes ne présentent pas de profil particulier. Ce groupe inclut potentiellement des gestes très variés, notamment parce qu'il regroupe des gestes dont les fonctions n'ont pas pu être déterminées avec précision, ce qui empêche de dégager un profil clair à partir des récurrences, qu'elles soient formelles ou dynamiques.

Ces profils ne prétendent toutefois pas à une description exhaustive de l'ensemble des productions gestuelles, mais constituent plutôt une modélisation des tendances dominantes au

sein du corpus. Les types fonctionnels *Butterworth* et *Expressif* n'ont en revanche pas été intégrés à cette synthèse, leur faible fréquence dans le corpus ne permettant pas d'établir de profils tendanciels sur le plan statistique. À partir des quelques profils ébauchés ici toutefois, il est possible d'identifier de premières associations réciproques entre catégories de gestes. La première concerne les fonctions *Déictique* et *Parsing*, une association qui peut refléter le fait que, lors d'énumérations, le locuteur tend à désigner spatialement les différents éléments de la liste qu'il énumère. La seconde association concerne les fonctions *Modale* et *Quasi-linguistique*, ce qui peut s'expliquer par le fait que nombre de modalité peuvent s'exprimer par des gestes très codifiés qui peuvent se comprendre sans la parole (négation, appréciation, etc.).

La caractérisation des gestes, telle que décrite ici, constitue une modélisation experte qui n'est pas seulement informative, mais essentielle pour orienter le développement d'analyses multimodales et computationnelles adaptées au geste co-verbal. Elle permet notamment de déterminer quels aspects et phénomènes sont pertinents à détecter, segmenter ou classer, en vue de pouvoir, par la suite, tenter de mieux comprendre les associations entre les formes possibles d'un geste et les rôles qu'ils assurent au regard du discours. Les annotations offrent également un point d'appui pour des analyses multimodales plus fines, notamment dans une perspective de traitement automatique du langage, afin de caractériser plus précisément les types d'affiliation lexicale entre gestes et discours – qu'il s'agisse de la nature des mots concernés ou de leurs fonctions syntaxiques et sémantiques. Ces analyses n'ont pas pu être menées dans le cadre de la présente thèse, mais constituent un axe de travail envisagé à plus long terme. De plus, certaines limites techniques du corpus restreignent l'analyse fine de la synchronie entre gestes et parole. Les fichiers du corpus BEAT présentaient des désynchronisations entre les différentes modalités, nécessitant un réalignement manuel dans la mesure du possible. Malgré ces efforts, des imprécisions résiduelles subsistent, c'est pourquoi il n'a pas été possible de réaliser d'étude sur la coordination geste-parole telle que la synchronie temporelle entre le geste et son affilié. En dépit de ces contraintes, l'enrichissement proposé du corpus reBeat-14 ouvre la voie à un large éventail d'analyses multimodales, permettant d'explorer en profondeur les différents aspects du geste co-verbal.

La segmentation fine des gestes et les annotations multimodales associées dans le corpus constituent un *gold standard* précieux. Ce travail d'annotation présente toutefois des limites importantes. La mise en place du protocole, la formation des annotateurs, ainsi que la réalisation des annotations elles-mêmes se sont révélées particulièrement coûteuses en temps. Le processus a reposé majoritairement sur deux annotatrices expertes, quasiment dédiées à temps plein à cette tâche dans le cadre de la thèse. De plus, compte tenu de la taille relativement restreinte du corpus, ces résultats doivent en outre être interprétés avec prudence : s'ils permettent de mettre en évidence des tendances dans les représentations et certaines caractéristiques gestuelles, ils ne sauraient être considérés comme définitifs. Certains types fonctionnels restent en effet moins représentés que d'autres, et ces observations gagneraient à être confirmées sur des ensembles de données plus larges et plus diversifiés. La faible représentation de certaines catégories fonctionnelles limite actuellement l'entraînement de classificateurs multi-classe robustes, ce qui restreint pour l'instant les approches à des cadres

binaires ou à des regroupements de catégories. De plus, il convient également de souligner que l'annotation multimodale comporte une part de subjectivité non négligeable. En effet, les annotations restent sujettes à une certaine variabilité. Cette variabilité peut être faible entre annotateurs ayant effectué la phase de synchronisation, lorsque l'accord inter-annotateur est fort, mais elle tend à augmenter au fil du temps, à mesure que l'annotation progresse, en raison de changements dans les décisions ou dans la manière de procéder. Cette subjectivité est particulièrement prononcée dans le cadre de l'annotation formelle. Annoter les aspects formels sur la base de la perception du mouvement dans une vidéo 2D, alors que celui-ci se déroule dans un espace 3D, peut poser un problème, notamment en raison des effets de perspective qui peuvent fausser la perception de certaines positions ou orientations. Par ailleurs, certains seuils formels sont difficiles à discriminer à l'œil nu, ce qui peut limiter la précision de la catégorisation des gestes. Ces contraintes soulignent l'intérêt de développer des outils automatiques ou semi-automatiques capables d'assister l'annotation multimodale et de réduire le coût humain et temporel associé.

# Chapitre 7
# L'annotation multimodale comme outil d'évaluation des gestes synthétiques

## 1 Introduction

L'utilisation de modèles de génération automatique pour la génération de gestes co-verbaux est un domaine en pleine expansion. On peut notamment citer l'initiative du GENEA [131, 115], qui place la question de la génération et de l'évaluation des gestes synthétiques au centre des discussions lors de conférences spécialisées. Par exemple, les organisateurs du *GENEA 2023 challenge* ont invité les chercheurs du domaine de la gestualité spécialisés dans l'approche computationnelle à proposer des modèles de génération automatisée, dans l'objectif d'en évaluer les performances actuelles. Mais les rapports et travaux récents s'accordent à montrer que de nombreux défis scientifiques persistent [21]. En particulier, les modèles actuels rencontrent encore des difficultés à produire des gestes qui soient à la fois naturels et expressifs [21]. Un autre enjeu majeur concerne l'évaluation de ces gestes générés. Les tests actuellement réalisés reposent principalement sur des expériences perceptives, menées sur un échantillon plus ou moins large de personnes, où le geste est jugé selon trois dimensions : *human-likeness* (gestes naturels et réalistes), *speech appropriateness* (gestes adéquats au discours) et *behavior appropriateness* (gestes adéquats aux comportements de l'interlocuteur) [115]. Or, ces trois catégories, telles que proposées, ne s'appuient pas sur des critères fins, ce qui rend l'évaluation relativement floue. Ces catégories nécessiteraient, dans la mesure du possible, des critères plus précis en leur sein, qui pourraient être capturés par une analyse experte de la gestualité humaine.

L'objectif de ce chapitre vise alors à démontrer l'intérêt d'intégrer les caractéristiques dégagées par une analyse multimodale fondée sur des annotations expertes dans l'évaluation des productions des modèles de génération de gestes manuels co-verbaux. Le postulat est ainsi que certains traits gestuels, mis en évidence par des annotations expertes peuvent contribuer à définir des méthodes d'évaluation à la fois plus précises et plus objectives, dans un contexte où

les approches traditionnelles souffrent encore d'un manque de cohérence et de comparabilité [22]. En d'autres termes, il s'agit de s'appuyer sur les informations fournies par l'expertise d'analyse multimodale pour à la fois guider le développement des futurs modèles, en apportant un retour plus précis sur la qualité de leurs productions, mais aussi pour offrir un moyen de mieux caractériser les gestes humains en tant que tels, par la mise en évidence de certains aspects absents des gestes synthétiques.

Pour montrer comment l'expertise linguistique peut contribuer à fournir des outils d'évaluation des gestes de synthèse, nous présenterons ici une analyse comparative de deux échantillons de gestes annotés, gestes authentiques pour l'un et gestes synthétiques pour l'autre, tous deux associés à une même production verbale. Il s'agit là des résultats d'une investigation exploratoire dont l'objectif était de formaliser des critères plus objectifs pour l'évaluation des gestes générés automatiquement. Cette comparaison entend en effet démontrer la manière dont une analyse multimodale experte peut permettre de dégager des standards sur lesquels fonder une évaluation précise des gestes issus des modèles de génération, et sur lesquels baser la comparabilité de leur performance.

Pour structurer cette analyse, le chapitre débute par la présentation de la manière dont a été constitué le jeu de données de cette étude, incluant la brève présentation du modèle STARGATE [94, 162, 163] utilisé pour générer les données de synthèse évaluées dans ce chapitre, la sélection des données, et les spécificités du protocole d'annotation adopté pour annoter les données de synthèse. Il se poursuit en trois volets d'analyse, portant sur l'identification des gestes potentiels, l'interprétation fonctionnelle des gestes annotés, et enfin les aspects formels des gestes. Le chapitre conclut sur les apports d'une annotation multimodale experte pour l'évaluation des gestes de synthèse.

# 2 Méthodologie

## 2.1 Le modèle STARGATE

Les données synthétiques utilisées dans cette étude comparative ont été générées par le modèle STARGATE [94, 162] produit par L. Abel dans le cadre de son doctorat réalisé au LORIA. Ce modèle, basé sur l'apprentissage profond, a été développé à partir de l'ensemble des fichiers du locuteur Wayne du corpus BEAT [150], et conçu pour générer des gestes manuels co-verbaux à partir de l'audio et de la transcription textuelle. Il a été fondé sur une architecture auto-régressive par segments, générant plusieurs images gestuelles à chaque *frame*, lesquelles servent ensuite d'entrée pour produire le segment suivant. Cette architecture comprend trois encodeurs – audio, texte et mouvement. L'encodeur de mouvements repose sur un réseau de convolution spatio-temporelle sur graphes, permettant de projeter les données d'entrée dans un espace latent tout en préservant leur structure. Les encodeurs audio et texte, quant à eux, utilisent des réseaux convolutionnels classiques. Les espaces latents sont concaténés puis passés dans un décodeur afin de produire la séquence de mouvements. STARGATE fonctionne sur des fenêtres glissantes : une seconde de mouvement passé, et deux secondes d'audio et de texte (une seconde passée et une seconde future). Ce modèle a été choisi pour deux raisons principales :

d'une part, il a été entraîné exclusivement sur le locuteur Wayne, ce qui permettait une comparaison fiable avec les données naturelles annotées dans le cadre de cette recherche de thèse et présentées au chapitre 5, sans biais lié à une variabilité inter-individuelle ; d'autre part, il surpassait (au moment du choix du modèle) les modèles antérieurs en produisant non seulement des gestes rythmiques, mais semblait également produire des gestes sémantiques plus complexes.

## 2.2 Sélection des données

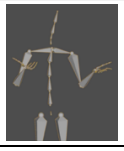
| | **Données naturelles** | **Données synthétiques** |
|---|---|---|
| **Durée de fichier** | 1 min 13 s | 1 min 11 s |
| **Origine de l'audio** | BEAT | BEAT |
| **Origine du .bvh** | BEAT | STARGATE |
| **Annoté par** | Syncogest | Syncogest |
| **FPS** | 120 | 60 |
| **Mouvement des doigts** | Oui | Non |
| **Visuel de l'agent** |  |  |

*Tableau 33. Description des fichiers naturel et synthétique [164]*

L'analyse s'est focalisée sur le Fichier 2 du locuteur Wayne. Deux versions de ce fichier ont été comparées (Tableau 33) : une version naturelle (Nat) et une version synthétique (Synt). Les deux fichiers contiennent la même production verbale issue de BEAT, mais comprennent des données de mouvement différentes : le fichier `.bvh` issu de BEAT pour Nat., et le ficher `.bvh` généré par STARGATE pour Synt. Les fichiers `.bvh` ont été utilisés pour reconstruire les mouvements sur un ECA et ainsi fournir une vidéo. Leur durée est respectivement de 1 min 13 s et 1 min 11 s. En effet, en raison d'une coupure d'une seconde au début du fichier synthétique – due à l'architecture du modèle, qui ne permet pas de générer les premières et dernières secondes de gestes à partir d'une entrée donnée –, 64 *frames* vides ont été ajoutées à la vidéo Synt pour permettre une synchronisation des deux fichiers. Les données naturelles conservent les fps d'origine du corpus BEAT (120 fps), tandis que les données synthétiques ont été générées à 60 fps. Les coordonnées des articulations dans les fichiers `.bvh` ont été recalculées en prenant la hanche comme point de référence (0, 0, 0), et exprimées en centimètres relativement à ce repère. On notera que les données naturelles incluent le mouvement détaillé des doigts, ce qui n'est pas le cas des données générées par le modèle STARGATE. Il est important de reconnaître que l'absence de la cinématique fine des doigts ainsi que la résolution temporelle réduite (60 fps) dans les données générées par STARGATE introduisent un biais intrinsèque susceptible de compliquer l'annotation. En effet, la configuration manuelle joue un rôle central dans la désambiguïsation de la fonction d'un geste. Néanmoins, même privé de la mobilité des doigts (configuration dite « en moufle »), un geste doit conserver une macro-dynamique identifiable, notamment au niveau de la trajectoire du poignet, de la phase de préparation et de la présence d'un stroke clairement distinguable, pour demeurer interprétable. L'objectif de notre analyse

est précisément d'évaluer si le modèle parvient à reproduire cette macro-dynamique du bras et de la main, indépendamment des extrémités articulaires.

La taille réduite de l'échantillon des données comparées est ici justifiée par les objectifs de cette étude : il ne s'agit pas de produire des résultats généralisables à l'ensemble des styles gestuels ni aux sorties de différents modèles de génération, mais d'explorer les principes fondamentaux et l'intérêt d'une méthodologie d'évaluation des gestes synthétiques fondée sur le recours à une annotation experte, au travers d'un protocole d'annotation dédié.

## 2.3 Protocole d'annotation et analyses

Les gestes du fichier naturel (Nat) ont été annotés dans le cadre de la constitution du corpus reBeat-14 présenté dans le Chapitre 5 ; les gestes synthétiques du fichier produit par le modèle STARGATE (Synt) a également été annoté manuellement, selon le même protocole que pour Nat, comprenant une phase d'annotation individuelle et une révision commune. Bien que les deux annotatrices étaient conscientes qu'il s'agissait de gestes de synthèse, elles ont néanmoins cherché, dans la mesure du possible, à appliquer le même protocole d'annotation que celui utilisé pour les gestes naturels.

Deux ajustements sont toutefois à noter concernant les désaccords observés lors de la révision commune concernant la segmentation et l'interprétation des gestes. En effet, dans le cas du fichier Synt, les désaccords entre les annotatrices portant sur le balisage de gestes ont été résolus en privilégiant la fenêtre temporelle la plus large, en raison de difficulté d'arbitrage que nous décrirons plus loin. Les frontières de début et de fin de gestes ont ainsi été fixées lors d'une révision commune, sur la base des temps les plus précoces et les plus tardifs annotés par les deux annotatrices. De même, les divergences concernant l'interprétation du type de phase et du type fonctionnel ont été explicitement signalées par l'ajout d'une catégorie *Unclear* dans les différents tiers concernant la phase gestuelle (tier `phase`), le type d'activité (tier `type`) et le type gestuel (tier `is_communicative_d1` ou `is_communicative_d2`). Cette catégorie, introduite lors de l'annotation des gestes de synthèse, a servi à indiquer un désaccord – que nous expliquerons plus loin – sur l'interprétation ou la fonction du geste, cas de figure non rencontré lors de l'annotation des fichiers naturels.

Les analyses ont été réalisées en comparant les deux jeux annotations obtenus, qui ont donc été analysés afin de dégager des éléments quantitatifs et qualitatifs mettant en évidence les principaux aspects dans lesquels, au sein de nos fichiers, les gestes synthétiques diffèrent des gestes naturels – apportant ainsi un éclairage sur les différents types de limites potentiellement associées au modèle étudié. La gestion des données et les analyses descriptives ont été effectuées sous RStudio, en se concentrant sur les occurrences (N), la durée moyenne (AD) en secondes et la durée médiane (MDu), elle aussi en secondes.

# 3 Identification des gestes potentiels

## 3.1 Désaccord de segmentation entre annotatrices

### 3.1.1 Modes d'analyse de l'accord inter-annotateurs sur les horodatages

Les écarts temporels entre les annotations de chaque annotatrice ont été analysés pour les gestes partageant un même identifiant d'annotation. Les repères temporels correspondent aux points de début ou de fin des gestes tels qu'annotés. Deux indicateurs sont utilisés : la durée moyenne et la durée médiane, toutes deux en secondes. Les différences absolues ont été calculées pour les temps de début et de fin des annotations, afin d'éviter tout effet compensatoire entre valeurs positives et négatives. Cette approche ne permet toutefois pas de conserver l'information de direction de la différence (c'est-à-dire si l'annotation de l'annotatrice 1 est antérieure ou postérieure à celle de l'annotatrice 2).

### 3.1.2 Mode d'affectation des identifiants des gestes annotés

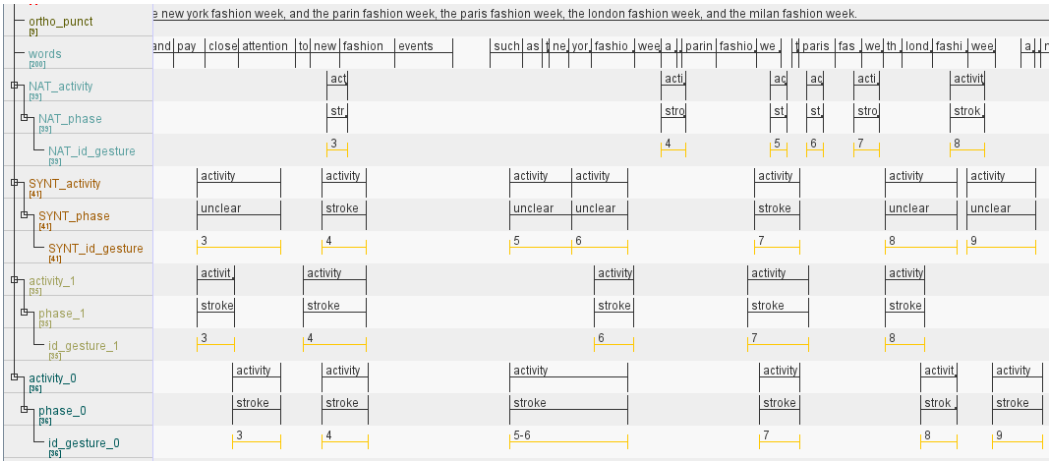



*Figure 30. Extraits des tiers d'annotation du fichier 2, avec l'annotation commune des gestes naturels (NAT), l'annotation commune des gestes de synthèse (SYNT), et les annotations individuelles des annotatrices (_0 et _1) pour les gestes de synthèse.*

La révision commune a consisté, sur la base de la comparaison des deux annotations indépendantes (Figure 30 : tiers `id_gesture_0` et `id_gesture_1`), à produire une nouvelle annotation unique (tier `SYNT_id_gesture`). Dans ce cadre, la numérotation des gestes a été recalculée sur l'ensemble des tiers, selon les principes exposés ci-dessous, ce qui a permis de renuméroter progressivement les gestes dans les deux tiers d'annotations initiales. Cette démarche garantit que l'annotation commune reflète la totalité des gestes identifiés par les deux annotatrices, en tenant compte du fait que certains gestes avaient été annotés uniquement par l'une ou l'autre, ou qu'un même geste avait été divisé en deux gestes distincts dans l'une des annotations.

- Ainsi, lorsque deux gestes annotés par les deux annotatrices (tiers avec `0` et `1` dans leur nom) se recouvraient temporellement, et que le geste validé lors de la révision commune recouvrait également les deux segmentations proposées par les annotatrices,

un même identifiant leur était attribué (par exemple, les gestes 4 et 7 en SYNT, 0 et 1).

- De même, lorsque les gestes annotés par les deux annotatrices présentaient un faible recouvrement temporel mais étaient néanmoins considérés comme relevant du même « événement gestuel » (comme les gestes 3 et 8 en SYNT, 0 et 1), un identifiant unique leur était attribué.
- Au contraire, lorsque l'une des deux annotatrices n'avait pas segmenté de geste à un endroit où l'autre en avait identifié un, l'identifiant correspondant pour ce geste identifié a été conservé dans l'ensemble des tiers, que le geste ait été retenu ou non lors de la révision commune. Cela impliquait, le cas échéant, un saut dans la numérotation au sein des tiers de l'une des annotatrices ou dans celui issu de la révision commune, afin de maintenir une continuité dans la numérotation de l'ensemble et une correspondance cohérente entre les identifiants de chaque annotation (par exemple le geste 9 en SYNT et 0, qui n'a pas été annoté par l'annotatrice 1).
- Sur le même principe, lorsqu'une annotatrice segmentait un geste unique qui correspondait, dans la révision commune ou dans l'annotation de l'autre annotatrice, à deux gestes distincts, ce segment recevait un identifiant combiné reprenant les deux identifiants concernés par le recouvrement (ex. : les gestes 5 et 6 en SYNT et le geste 5-6 en 0).

### 3.1.3 Analyse de l'accord inter-annotateur sur les repères temporels d'annotation

| **Fichier** | **Différence de repères temporels** | |
|---|---|---|
| | *AD* | *MDu* |
| Synthèse | 0,117 s | 0,195 s |
| Naturel | 0,050 s | 0,084 s |

*Tableau 34. Variabilité des repères temporels d'annotation des gestes [164]*
*(AD = durée moyenne ; MDu = durée médiane)*

Pour le corpus synthétique, la différence moyenne entre repères temporels est de 0,117 s et la médiane de 0,195 s (Tableau 34). Ces valeurs indiquent que les gestes synthétiques présentent une variabilité relativement importante dans l'annotation temporelle, suggérant que les repères sont potentiellement plus difficiles à identifier de manière précise entre les annotateurs. En comparaison, le corpus naturel présente une variabilité bien plus faible, avec une moyenne de 0,05 s et une médiane de 0,084 s. Dans les annotations des données naturelles, les activités identifiées ont ainsi été positionnées à des instants quasi identiques. Cela traduit une meilleure cohérence dans l'annotation des gestes naturels, reflétant la clarté des limites temporelles des gestes produits par le locuteur. Ainsi, cette analyse montre que les différences temporelles observées dans les annotations des gestes synthétiques reflètent les difficultés

rencontrées par les dans l'application de leur stratégie de segmentation à ces données. Contrairement aux gestes naturels – pour lesquels la phase de synchronisation entre annotateurs conduit à des annotations très proches –, les gestes synthétiques apparaissent moins clairs, ce qui entraîne des découpages plus variables et une segmentation moins consensuelle. Ce désaccord entre les annotatrices a conduit à retenir les bornes les plus larges de l'activité gestuelle lors de la révision commune, lorsque les discussions ne permettaient pas de trancher sur les bornes à adopter.

## 3.2 Gestes potentiels (activités) annotés

| | **Synthèse** | | | **Naturel** | | |
|---|---|---|---|---|---|---|
| | *N* | *AD* | *MDu* | *N* | *AD* | *MDu* |
| *Stroke* | 25 | 0,584 | 0,519 | 39 | 0,352 | 0,320 |
| *Unclear* | 16 | 0,706 | 0,669 | -- | -- | -- |
| **Activités** | **41** | **0,632** | **0,625** | **39** | **0,352** | **0,320** |

*Tableau 35. Phase des activités annotées [164]*
*(N = nombre d'occurrences ; AD = durée moyenne ; MDu = durée médiane)*

### 3.2.1 Analyse de la durée des activités

L'analyse de la durée des activités annotées met en avant un nombre comparable entre les deux fichiers du corpus (Tableau 35), avec 41 occurrences dans le fichier Synt contre 39 dans le fichier Nat. Bien que ces effectifs soient similaires, la durée moyenne des annotations est presque double dans les données synthétiques (AD = 0,632 s) comparé aux données naturelles (AD = 0,352 s). Dans les deux cas, la distribution de durées apparaît relativement homogène, avec des médianes proches des moyennes. Cette plus grande durée des gestes synthétiques est une conséquence du manque d'accord entre annotatrices sur la segmentation.

### 3.2.2 Analyse du type de phase d'activité annotées

Lorsque l'on examine plus précisément les types de phases associées à chaque activité annotée, on constate que ces phases sont exclusivement interprétées comme des *strokes* dans les données naturelles. En revanche, dans les données de synthèse, 16 des activités annotées (environ 39 %) présentent une phase annotée comme *Unclear*, indiquant un désaccord entre les annotatrices concernant les frontières de ce qui constituerait le cœur de ces gestes annotés. On notera par ailleurs que, dans le fichier de synthèse, ces segments annotés *Unclear* sont en moyenne plus longs que les segments annotés comme des *strokes* ($AD_{\text{unclear}}$ = 0,706 s ; $AD_{stroke}$ = 0,584 s). Ce sont donc principalement ces phases qualifiées d'ambiguës qui contribuent à augmenter la durée moyenne des annotations dans le fichier Synt, tandis que les activités annotées comme des phases de *strokes* présentent une différence moyenne de seulement 0,187 s entre les conditions Synt et Nat. Enfin, on observe une légère tendance à une plus grande dispersion des durées pour les *strokes* des données synthétiques (MDu = 0,519 s), avec une médiane plus éloignée de la moyenne que dans le cas des données naturelles.

Ces résultats soulignent une distinction marquée entre les deux ensembles : les gestes synthétiques sont plus longs, reflétant leur ambiguïté structurelle dans près de la moitié des cas. Cette ambiguïté s'explique par l'absence d'indices visuels distinctifs exploitables pour segmenter les gestes uniquement à partir des variations de vitesse ou des changements d'orientation des articulateurs. Elle a naturellement entraîné un manque de consensus pour déterminer ce qui constituait le *stroke* : certains segments considérés comme des phases de préparation par une annotatrice étaient annotés comme constituant des *strokes* par l'autre, et réciproquement, certains segments identifiés comme *strokes* par l'une des annotatrices étaient interprétés comme des phases de rétractation par l'autre. Ces divergences illustrent la complexité de l'annotation des gestes synthétiques et expliquent, en partie, l'ajout de la catégorie *Unclear* pour les segments dont la phase de *stroke* ne pouvait être déterminée de manière fiable.

La tendance du modèle à produire des mouvements moins dynamiques pourrait alors s'expliquer par le fait que le modèle est entraîné sur l'intégralité des séquences motrices, dans lesquelles les gestes constituent des événements ponctuels au sein d'un flux de mouvement continu. Dans ce contexte, le système pourrait apprendre à produire une trajectoire moyenne ou lissée du mouvement plutôt qu'à modéliser explicitement des caractéristiques spécifiques du geste, notamment ses variations dynamiques.

# 4 Interprétation fonctionnelle

## 4.1 Analyse des activités communicatives et non communicatives annotées

| | **Synthèse** | | | **Naturel** | | |
|---|---|---|---|---|---|---|
| | *N* | *AD* | *MDu* | *N* | *AD* | *MDu* |
| Communicatif | 31 | 0,660 | 0,625 | 38 | 0,354 | 0,322 |
| Unclear | 10 | 0,545 | 0,605 | -- | -- | -- |
| Undefined | -- | -- | -- | 1 | -- | -- |
| **Activité** | **41** | **0,632** | **0,625** | **39** | **0,352** | **0,320** |

*Tableau 36. Caractérisation des activités [164]*
*(N = nombre d'occurrences ; AD = durée moyenne ; MDu = durée médiane)*

Si l'on considère à présent le type d'activité gestuelle, on observe que les deux ensembles comportent des activités interprétées comme étant communicatives, avec 31 occurrences (76 %) en condition Synt et 38 (97,44%) occurrences en condition Nat (Tableau 36). Les activités reconnues comme communicatives sont donc non seulement plus nombreuses en condition Nat, mais surtout nettement plus longues en condition Synt (AD = 0,660 s ; MDu = 0,625s). Cette différence s'accompagne également d'une dispersion légèrement plus importante pour les données synthétiques, comme l'indique l'écart entre la moyenne et la médiane.

Concernant les activités annotées comme *Unclear*, correspondant à un désaccord inter-annotateurs sur le fait que ces segments soient réellement communicatifs, seules les données de

synthèse en présentent (N = 10 ; AD = 0,545 s, MDu = 0,605), aucun segment problématique à ce niveau n'ayant été identifié dans les données naturelles. Cette asymétrie renforce l'idée que certains gestes générés automatiquement restent difficiles à interpréter, en cohérence avec l'analyse précédente portant sur la clarté des phases gestuelles.

Enfin, une seule activité annotée comme *Undefined*, renseignant non plus un désaccord, mais un doute partagé concernant la nature communicative du segment, a été relevée dans les données naturelles, tandis qu'aucune occurrence de ce type n'apparaît dans les données de synthèse.

Dans l'ensemble, ces résultats suggèrent que les gestes synthétiques ne se distinguent pas uniquement par leur durée plus étendue, mais également par une plus grande incertitude quant à la reconnaissance de leur nature communicative, ce qui pourrait être la conséquence d'une modélisation encore imparfaite de la saillance fonctionnelle des gestes dans le processus de génération.

## 4.2 Analyse des types fonctionnels des gestes annotés

| | **Synthèse** | | | **Naturel** | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | D1 | | | **D1** | | | **D2** | | |
| | *N* | *AD* | *MDu* | *N* | *AD* | *MDu* | *N* | *AD* | *MDu* |
| Undefined | 16 | 0,671 | 0,635 | 14 | 0,343 | 0,374 | -- | -- | -- |
| Unclear | 10 | 0,612 | 0,598 | -- | -- | -- | -- | -- | -- |
| Parsing | 3 | 0,441 | 0,473 | 8 | 0,293 | 0,319 | 1 | -- | -- |
| Métaphorique | 2 | 1,135 | 1,135 | 4 | 0,691 | 0,642 | 1 | -- | -- |
| Désignation | -- | -- | -- | 5 | 0,240 | 0,217 | 2 | 0,646 | 0,646 |
| Modal | -- | -- | -- | 3 | 0,348 | 0,403 | 1 | -- | -- |
| Spatial | -- | -- | -- | 1 | -- | -- | 2 | 0,302 | 0,302 |
| Quasi-linguistique | -- | -- | -- | 1 | -- | -- | 1 | -- | -- |
| Butterworth | -- | -- | -- | 1 | -- | -- | -- | -- | -- |
| Temporel | -- | -- | -- | 1 | -- | -- | -- | -- | -- |
| **Type fonctionnel** | **31** | **0,660** | **0,625** | **38** | **0,354** | **0,322** | **8** | **0,408** | **0,336** |

*Tableau 37. Détail des types fonctionnels annotés [164]*
*(N= nombre d'occurrence ; AD = durée moyenne ; MDu = durée médiane)*

Dans les données de synthèse, deux catégories d'annotation dominent (Tableau 37) pour la D1 : *Undefined* (N=16), renseignant un doute partagé concernant le type fonctionnel du geste, et *Unclear* (N=10) renseignant sur un désaccord concernant le type fonctionnel du geste entre les annotatrices, qui totalisent plus de 80 % des annotations (durée moyenne entre 0,61 et 0,67 s). On relève quelques occurrences ponctuelles de types fonctionnels annotés en D1 : des gestes de *Parsing* (N=3) et des gestes *Métaphoriques* (N=2), avec des durées contrastées : les gestes *Métaphoriques* sont particulièrement longs (1,135 s), tandis que les gestes de type *Parsing* sont plus courts (0,441 s) en comparaison. Il est à noter qu'aucune dimension secondaire n'a été annoté pour ces données. Ces résultats mettent en évidence une grande ambiguïté et une difficulté d'interprétation des gestes synthétiques, ce qui se traduit par une diversité limitée des types de gestes reconnus dans les gestes générés automatiquement.

En contraste, 38 gestes ont été annotés dans les données naturelles, dont 8 présentant une double dimension fonctionnelle. Parmi les 38 labels annotés en D1, 24 types fonctionnels (63,16% de la totalité des labels en D1) et 14 labels *Undefined* (38,64% des D1) ont été annotés. En première dimension, on peut observer une diversité de types fonctionnels identifiés, notamment *Parsing* (N=8), *Métaphorique* (N=4), *Désignation* (N=5), *Modal* (N=3) et ponctuellement (N=1 pour chaque type) *Déictique Spatial*, *Quasi-linguistique*, *Butterworth* et *Déictique temporel*. En seconde dimension, les types fonctionnels reconnus sont également variés, avec la présence des labels *Désignation* (N=2), *Déictique Spatial* (N=2), *Métaphorique* (N=1), *Modal* (N=1) et *Quasi-linguistique* (N=1). Si l'on compare les durées des gestes à fonction *Parsing* ou *Métaphorique* en première dimension entre Synt et Nat, on observe la même dynamique que celle mise en évidence dans les analyses de durées abordées précédemment.

Le corpus synthétique présente une diversité fonctionnelle nettement plus faible que le fichier naturel, les gestes y étant majoritairement annotés comme fonctionnellement indéterminés, et sans aucune occurrence cumulant une seconde dimension fonctionnelle, malgré un nombre total de gestes comparable. À l'inverse, les gestes naturels couvrent une large variété de fonctions communicatives, et un cinquième d'entre eux sont multi-fonctionnels. Cette analyse met en évidence les difficultés d'interprétations pour les gestes de synthèse, y compris lorsque les annotatrices s'accordent sur la nature communicative des segments balisés.

## 4.3 Analyse de l'affiliation lexicale des gestes

| **Phrases** | **ID** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mots | 47 | 24 | 16 | 12 | 19 | 17 | 36 | 26 | 9 |
| | **Activités** | **8** | **4** | **5** | **2** | **2** | **2** | **8** | **6** | **2** |
| **Naturel** | Undefined | | | | | | | 1 | | |
| (24 gestes) | Activités communicatives | 8 | 4 | 5 | 2 | 2 | 2 | 7 | 6 | 2 |
| | Gestes avec affiliation | 6 | 2 | 5 | 2 | 1 | 1 | 4 | 2 | 1 |
| | **Activités** | **9** | **3** | **4** | **3** | **3** | **3** | **9** | **5** | **2** |
| **Synthèse** | Unclear | 2 | 1 | | | | 1 | 5 | 1 | |
| (5 gestes) | Activités communicatives | 7 | 2 | 4 | 3 | 3 | 2 | 4 | 4 | 2 |
| | Gestes avec affiliation | 1 | | 1 | | | 2 | | | |

*Tableau 38. Activités annotées par phrases et nombre d'affiliations lexicales [164]*

Le Tableau 38 présente la distribution des activités communicatives annotées par phrase pour les corpus naturel et synthétique. Au total, dans le fichier 2, le locuteur Wayne produit 9 phrases, comprenant entre 9 et 47 mots chacune. Sans réelle surprise, on constate que le nombre d'activités annotées tend à augmenter avec la longueur des phrases. Entre les données naturelles et synthétiques, la quantité globale d'activités annotées par phrase reste toutefois comparable.

Pour les gestes naturels, l'ensemble des gestes dont la catégorie fonctionnelle a été identifiée (N = 24, 100 %) par les annotatrices a pu être affilié au discours. Dans le cas des gestes de synthèse, parmi les cinq gestes avec un type fonctionnel identifié, quatre ont effectivement été reliés au discours (soit 80%). Il est à noter que, lors de l'annotation des gestes de synthèse, l'établissement du lien entre ces gestes et le contenu du discours s'est révélée

particulièrement difficile. Ainsi, la difficulté à identifier leur fonction communicative peut être en partie due au fait qu'ils semblent peu alignés avec le discours.

On peut supposer que cela reflète la manière dont le modèle intègre l'information textuelle lors de l'apprentissage. Il est possible que l'utilisation de vecteurs riches, dans un espace latent de grande dimension sur l'ensemble de la séquence (ou de plusieurs), rende difficile l'identification des éléments réellement pertinents pour des événements ponctuels tels que les gestes, et limite ainsi la production de gestes explicitement liés au sens lexical. On peut donc émettre l'hypothèse que, le modèle a du mal à produire des gestes clairement affiliés au contenu verbal, ce qui contribue également à la difficulté de les interpréter.

## 4.4 Analyse des scores de certitudes

| | **Naturel** | | | **Synthèse** | | |
|---|---|---|---|---|---|---|
| | *N* | *Moy* | *Med* | *N* | *Moy* | *Med* |
| D1 | 24 | 4,05 | 4 | 5 | 1,4 | 1 |
| D2 | 8 | 4,13 | 4 | -- | -- | -- |
| **Global** | **32** | **4,06** | **4** | **5** | **1,4** | **1** |

*Tableau 39. Statistiques descriptives des scores de certitude (N = nombre d'occurrence ; Moy = moyenne ; Med = médiane ; D1 = dimension 1 ; D2 = dimension 2) [164]*

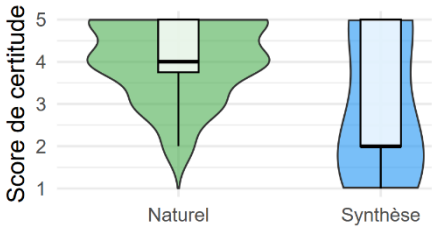



*Figure 31. Distribution des scores de certitude*

L'analyse des scores de certitude associés à l'interprétation fonctionnelle des gestes annotés reflète également une plus grande incertitude dans l'annotation des gestes synthétiques (Tableau 39, Figure 31). Dans le cas des données naturelles, on observe un score de certitude élevé, avec une moyenne autour de 4 quelle que soit la dimension analysée (médiane = 4), tandis que pour les données de synthèse cette moyenne chute à 1,4 (médiane = 1). Ces chiffres traduisent ainsi une forte confiance quant à l'interprétation fonctionnelle des gestes dans le cas des données naturelles de la part des annotateurs. Cette analyse met ainsi en évidence que, même lorsque les annotateurs s'accordent sur la nature communicative d'un segment et proposent une interprétation fonctionnelle, ils ne sont pas toujours certains de la justesse de cette interprétation dans la majorité des cas.

# 5 Analyses formelles

## 5.1 Analyse de la manualité des gestes

| | **Naturel** | | **Synthèse** | |
|---|---|---|---|---|
| | *Activité* | *Geste* | *Activité* | *Geste* |
| **Bimanuel** | 19 | 19 | 31 | 28 |
| **LH** | 3 | 3 | 1 | -- |
| **RH** | 17 | 16 | 9 | 3 |

*Tableau 40. Manualités des gestes annotés (LH = main gauche ; RH = main droite)*

L'analyse de la manualité des gestes dans les deux conditions met en évidence des différences d'utilisation des mains (Tableau 40). Dans la condition Nat, on observe une représentation relativement équilibrée entre gestes bimanuels (N = 19) et gestes monomanuels réalisés avec la main droite (RH, N = 16). En revanche, pour les gestes synthétiques, cet équilibre n'est pas observé, avec une nette surreprésentation des gestes bimanuels (N = 28) par rapport aux gestes monomanuels RH (N = 3). Les gestes impliquant uniquement la main gauche (LH) restent rares dans les deux conditions ($N_{Nat}$= 3 gestes et aucun geste en Synt). Le locuteur étant droitier, le nombre de gestes réalisés avec cette main dans les données naturelles semble cohérent, ce qui constitue également une configuration de manualité peu rencontrée par le modèle durant son apprentissage.

Cette tendance suggère que le modèle semble privilégier les gestes bimanuels, au détriment d'un équilibre entre bimanualité et monomanualité (RH) présent dans les données sur lesquels il a été entraîné. Cela pourrait s'expliquer par le fait que les mains bougent toujours légèrement. En effet, même lorsqu'un geste est monomanuel, la main non impliquée reste rarement totalement immobile, étant souvent influencée à minima par le mouvement global du corps. Il est également possible que le modèle applique un lissage sur le mouvement moyen sur l'ensemble de la séquence, y compris en dehors des gestes, ce qui empêche d'isoler les moments où le mouvement d'une main est réellement proche de zéro, et donc d'apprendre à dissocier main droite et main gauche lorsque cela est pertinent.

## 5.2 Analyse de l'utilisation de l'espace gestuel

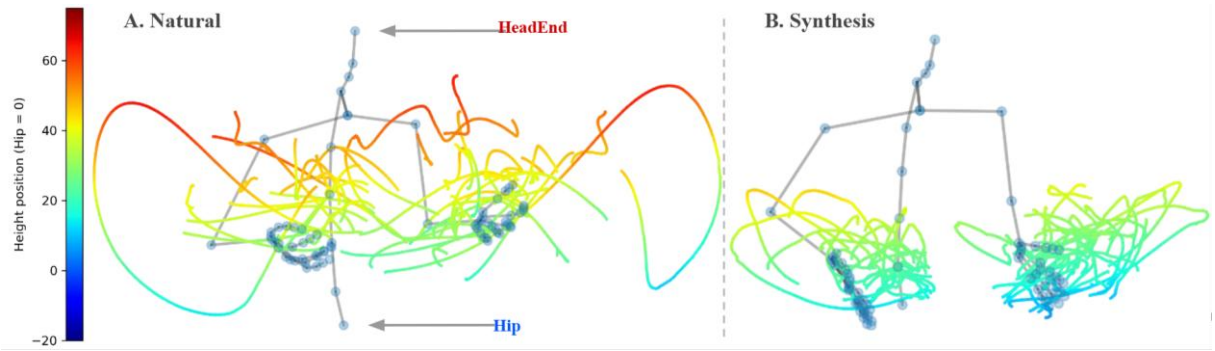



*Figure 32. Trajectoires pendant les activités annotées [164]*

La Figure 32 représente la superposition des trajectoires des mains engagées dans l'ensemble des annotations d'activités dans le fichier naturel et dans le fichier de synthèse. Pour les données naturelles (Figure 32A), on observe un gradient de couleur marqué, témoignant d'une importante variabilité spatiale des mouvements, les teintes orange et rouge indiquant la présence de gestes s'élevant à des hauteurs significatives, tandis que la dispersion latérale des trajectoires révèle un usage étendu de l'espace gestuel. Ces caractéristiques suggèrent que le locuteur mobilise un espace de gesticulation ample et étendu, modulant activement la hauteur, la direction et l'amplitude de ses gestes en fonction de son discours. Cette diversité spatiale traduit également une expressivité corporelle riche, où les mains ne se limitent pas à une zone

restreinte et évoluent librement dans l'espace gestuel, mais aussi donnant à voir des formes de trajectoires plus variées.

En revanche, les activités annotées dans les données synthétiques (Figure 32B) présentent plusieurs limitations notables. La palette colorimétrique y est cette fois considérablement réduite, concentrée autour de nuances de bleu clair et de vert, qui indique des mouvements situés à une hauteur restreinte, généralement cantonnée à la zone située au-dessus des hanches. Les gestes apparaissent globalement plus bas et moins variés, tant en hauteur qu'en amplitude horizontale, les trajectoires se superposant largement dans une même région spatiale. Ce confinement traduit une dynamique corporelle plus limitée, où les mouvements sont exécutés dans un espace gestuel restreint. Par ailleurs, les positions des mains ou l'orientation de la paume ne varient que très peu dans le cas des gestes de synthèse, avec des mouvements majoritairement symétriques, ce qui confère aux gestes un aspect plus mécanique. Ces observations suggèrent une simplification excessive des mouvements par le modèle de synthèse, qui peine à reproduire la variabilité naturelle et la dissymétrie subtile caractéristiques de la gestualité humaine.

On peut également observer une différence de posture globale entre les deux squelettes. Le locuteur naturel maintient une position relativement droite, avec un alignement cohérent du torse et de la tête, tandis que le locuteur synthétique apparaît nettement plus penché vers l'avant. Cette inclinaison du tronc, visible de manière récurrente dans les données de synthèse, suggère un biais postural introduit par le modèle de génération. Effectivement, il est également très probable qu'un biais dans les données ait joué un rôle : les locuteurs devaient souvent regarder vers le bas, ce qui a pu entraîner un biais d'apprentissage du modèle. Cette posture accentue l'impression d'artificialité des gestes synthétiques, affectant à la fois la naturalité visuelle et la cohérence spatiale globale du mouvement.

## 5.3 Analyse temporelle de la spatialité et de la vélocité

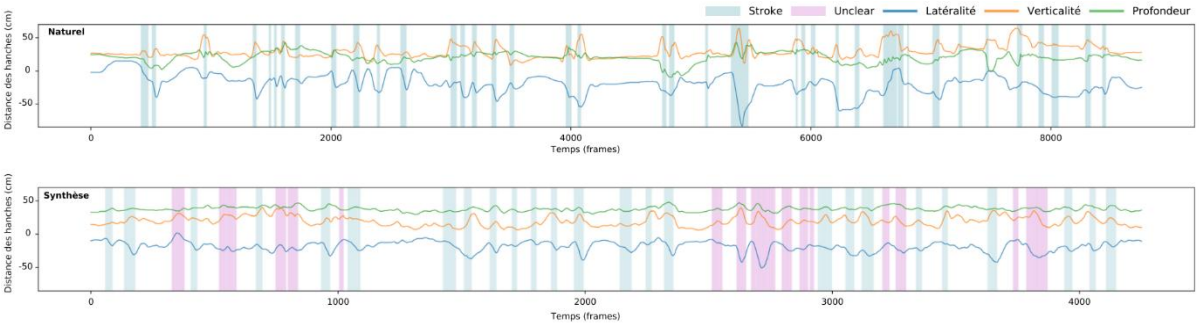



*Figure 33. Distance (cm) de l'index droit par rapport à la hanche et par frame [164]*

La Figure 33 représente la position relative de l'index droit par rapport à la hanche (référence fixée à 0), position exprimée en centimètres selon trois dimensions spatiales. La latéralité (courbe bleue) correspond à l'axe droite–gauche : les valeurs positives indiquent une position du côté droit, et les valeurs négatives du côté gauche. La verticalité (courbe jaune) renvoie à l'axe haut–bas : les valeurs positives correspondent à une position plus basse et les

négatives à une position plus haute. Enfin, la profondeur (courbe verte) correspond à l'axe avant–arrière : les valeurs positives signalent une position plus en arrière et les négatives une position plus en avant. En arrière-plan du graphique figurent également les phases annotées : les segments en bleu clair correspondent aux *strokes* annotés, tandis que les segments en rose indiquent les segments catégorisés comme *Unclear*.

L'analyse du déplacement par *frame* de l'index droit sur les trois axes montre des profils différents entre mouvement naturels et de synthèse. Dans les données naturelles (première ligne de la figure), l'index se déplace ponctuellement dans l'espace, caractérisés par des variations marquées correspondant à des pics de mouvement nettement saillants ; ces variations coïncident clairement avec les phases annotées comme *strokes*. À l'inverse, le mouvement issu de la synthèse (deuxième ligne de la figure) présente un profil cinématique plus atténué, avec des variations de position nettement moins marquées ; si certains pics de mouvement semblent ponctuellement coïncider avec les phases segmentées, cette correspondance demeure toutefois moins nette et bien moins systématique que dans les données naturelles.

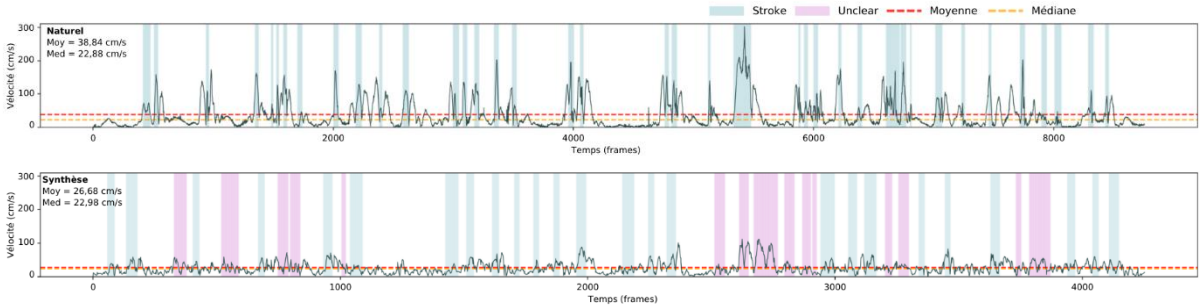



*Figure 34. Vitesse (cm/s) de l'index droit par frame [164]*

La Figure 34 représente la vitesse par *frame*, calculée à partir des positions 3D consécutives de l'index droit, pour les données naturelles et de synthèse. Comme pour la figure précédente, en arrière-plan du graphique figurent les phases annotées (*stroke* en bleu clair et *Unclear* en rose). Pour les données naturelles (première ligne de la figure), on constate que la vitesse varie entre 0 et environ 300 cm/s, avec une moyenne de 38,84 cm/s et une médiane de 22,88 cm/s. Ces valeurs indiquent que la majorité des mouvements restent relativement lents, ponctués par quelques pics de vitesse plus élevés qui tirent la moyenne vers le haut. La corrélation entre ces pics de vitesse et les phases de *stroke* annotées montre que l'identification de ces phases s'appuie fréquemment sur des variations de vitesses clairement perceptibles.

Dans le cas des données de synthèse (deuxième ligne de la figure), on peut remarquer que la vitesse par *frame* s'étend sur une plage plus restreinte de valeurs, allant de 0 à 100 cm/s. Si la médiane (22,98 cm/s) reste proche de celle des données naturelles, la différence se situe au niveau de la vitesse moyenne (26,88 cm/s) : celle-ci est proche égale de la médiane, reflétant l'absence de pics de vitesse marqués. En effet, le profil de vitesse des données synthétiques apparaît plus régulier et moins contrasté : si les variations sont fréquentes, leur amplitude reste nettement inférieure à celle des pics observés dans les données naturelles. Comme pour les

données naturelles, toutefois, on peut observer que l'annotation des phases repérées semble s'appuyer sur ces variations de vitesse, lorsqu'elles sont suffisamment perceptibles.

L'ensemble de ces analyses met en évidence que les mouvements dans les données naturelles présentent des variations spatiales et de vitesse saillante, fournissant des repères perceptibles pour l'annotation des phases. En revanche, le profil plus régulier et moins contrasté des gestes synthétiques explique la difficulté de segmentation constatée précédemment, et de manière corrélative, la complexité d'interprétation de la fonction communicative des gestes segmentés. Le « lissage » observé dans les mouvements générés par le modèle peut résulter de plusieurs facteurs techniques. D'une part, la fonction de perte peut pénaliser les grandes variations de mouvement entre *frames* consécutives, ce qui tend à réduire les pics de vitesse et d'accélération. D'autre part, des architectures reposant sur des espaces latents compacts ou des prédictions moyennées sur plusieurs *frames* peuvent produire des trajectoires « moyennes ». Les variations ponctuelles mais brèves peuvent alors ne pas être correctement représentées dans l'espace latent, limitant leur propagation et leur poids lors de l'apprentissage.

## 5.4 Exemples de gestes ambigus

### 5.4.1 Le geste de synthèse n° 5




*Figure 35. Illustration du geste n° 5 dans le fichier de synthèse*

Sur la Figure 35 est illustré le geste n° 5 du fichier de synthèse. Il se déroule de 9,703 s à 10,331 s, sur une durée d'environ 628 ms. Ce geste intervient durant la production verbale suivante : « *such as the New York Fashion Week* », plus précisément de la fin de *such* jusqu'au mot *New* (indication en orange sur la figure). Sa trajectoire procède d'abord à une remontée vers le haut, puis vers l'avant, coïncidant avec la projection du torse de l'agent. Le geste a été annoté comme une activité à phase *Unclear* et de type communicatif *Unclear*, sans affiliation lexicale attribuée. Les annotateurs n'ont pu s'accorder ni sur l'interprétation du geste, ni sur la délimitation du *stroke*. La visualisation dynamique ne permet pas d'estimer avec certitude qu'il s'agit d'une production volontaire du modèle ayant pour objectif de contribuer au message verbal : le geste est peu dynamique et peu ample, l'avancée vers l'avant relève principalement du mouvement du torse, et il est systématiquement suivi d'un autre geste, ce qui laisse planer un doute sur sa nature réelle – geste finalisé, geste avorté ou en préparation, ou simple mouvement répétitif dépourvu de sens concret, malgré un positionnement verbal approprié. Cet exemple illustre la difficulté d'effectuer des choix cohérents face à des gestes ambigus et met en

évidence les enjeux liés à l'accord entre annotateurs, en particulier lors de l'identification des repères temporels, notamment du *stroke*.

### 5.4.2 Le geste naturel n° 25

Bien que l'ambiguïté soit plus fréquente dans les gestes générés par le modèle, il est également possible de rencontrer des difficultés lors de la segmentation et de l'annotation de gestes naturels. La Figure 36 illustre le geste n° 25 du fichier naturel. Celui-ci s'étend de 49,337 s à 49,622 s, soit une durée d'environ 285 ms. Il intervient pendant la production verbale : « *I'll actually give each restaurant a score* », plus précisément sur le mot « score ». Le locuteur a l'index engagé, pointant vers le haut, la paume placée sur la tranche, et effectue une descente en ligne droite, sans changement notable de configuration ni d'orientation. Ce geste est exclusivement produit de la main droite, la main gauche ne présentant aucun mouvement clair en parallèle. Le mouvement segmenté a été identifié comme une activité de type *Undefined* et son type fonctionnel n'a pas pu être clairement identifié (*Undefined*). En effet, si le geste déictique précédent dans la séquence, associé au mot restaurant (pointage vers le haut avec l'index droit), est clairement identifiable, ce geste n° 25 semble davantage correspondre à la continuité de la rétractation du mouvement précèdent. Cette ambiguïté rend difficile de trancher entre un mouvement intentionnellement produit pour accompagner le mot « score » et un simple prolongement mécanique du geste précédent, malgré la présence de caractéristiques cinématiques pouvant évoquer un *stroke*.

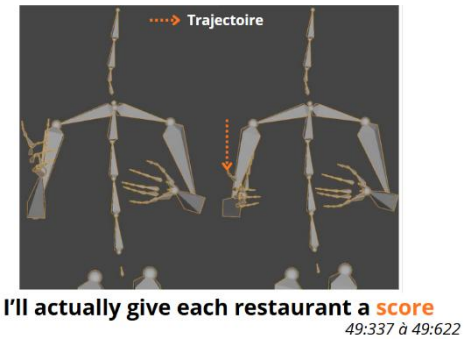



*Figure 36. Illustration du geste n° 25 dans le fichier naturel*

Ces deux exemples illustrent à quel point la frontière entre geste et mouvement peut être ténue, même pour des annotateurs expérimentés. Ils soulignent l'importance de comprendre comment l'ambiguïté inhérente à certains gestes, qu'ils soient naturels ou générés, influence directement la cohérence, la fiabilité et la reproductibilité du processus d'annotation.

## 6 Conclusion de chapitre

La comparaison, menée sur un jeu de données restreint de gestes naturels et de gestes synthétiques produits par STARGATE à partir d'un même contenu verbal permet de mettre évidence des différences marquées entre les deux extraits. S'agissant des gestes de synthèse, le faible nombre de phases clairement identifiées comme « *strokes* », la faible part de fonctions déterminées, les scores de certitude relativement bas, le peu d'ancrage au discours ainsi que les différences observées sur les aspects formels (faible diversité des trajectoires, faible occupation de l'espace, faibles pics de vitesse, peu de gestes monomanuels, etc.) révèlent que ces mouvements générés automatiquement ne correspondent pas à ce qui est observées pour les gestes naturels. Ainsi, sur la base de cet extrait d'une minute, nos observations suggèrent que le modèle STARGATE tend à produire des gestes cinématiquement lissés, dont l'interprétation fonctionnelle et l'affiliation lexicale sont dès lors plus difficilement interprétables, malgré de

bons résultats en évaluation perceptive [162]. À l'inverse, les gestes naturels présentent une plus grande diversité fonctionnelle, une affiliation lexicale clairement identifiée dans la majorité des cas et une exploitation plus riche de l'espace gestuel.

| **Fichier** | **Temps d'annotation** | | **Temps de révision** |
|---|---|---|---|
| | *Annotatrice 1* | *Annotatrice 2* | |
| Synthèse | 0 h 37 min 58 s | 1 h 20 min | 12 h |
| Naturel | 1 h 24 min 40 s | 2 h 13 min | 2 h |

*Tableau 41. Analyse du temps d'annotation et de révision des fichiers [164]*

La difficulté à annoter les gestes de synthèse, donc à les repérer et les interpréter, est très clairement reflétée par les temps d'annotation et de révision commune, présentés au Tableau 41. Si ces temps attestent d'une annotation initiale plus rapide pour les données de synthèse que pour les données naturelles (37 min et 1 h 20 contre 1 h 24 et 2 h 13 par annotatrice), elles révèlent en revanche un temps de révision nettement plus important (12 h contre 2 h). Cette asymétrie suggère que l'allongement de la phase de révision est une conséquence directe de la complexité d'analyser les gestes de synthèse, dont les caractéristiques formelles et fonctionnelles sont plus floues et ambigües que celles des gestes naturels. En effet, le manque de dynamisme et la faible accroche au discours des gestes de synthèse obligent les annotateurs à s'interroger systématiquement sur les frontières, la nature communicative et la fonction exacte de chaque segment tenu pour être un geste potentiel.

Si l'absence de données des mouvements des doigts dans le fichier de synthèse renforce ces difficultés d'interprétation – les doigts jouant un rôle clé dans la réalisation d'un geste et la désambiguïsation de sa ou ses fonctions –, cela ne suffit pas à expliquer les contrastes de l'analyse menée ici. En effet, même en se basant sur une visualisation quasi-similaire (les avatars), la distinction entre gestes naturels et synthétiques est particulièrement manifeste. Les gestes naturels mobilisent un espace plus vaste et varié, avec une amplitude verticale et latérale diversifiée, tandis que les gestes synthétiques se caractérisent par une posture plus contrainte, une symétrie accrue et une faible variabilité spatiale. Ces différences, perceptibles dès les premières *frames* et représentatives de la posture globale des locuteurs dans ces séquences, suggèrent que même une annotation menée « à l'aveugle », donc sans connaître la source des fichiers, ne pourrait empêcher de clairement distinguer des données de synthèse de données naturelles.

De manière générale, il reste difficile de déterminer avec certitude si les mouvements de synthèse interprétés comme des gestes ne sont pas simplement fortuits. Comme tout interlocuteur, les annotatrices ont naturellement cherché à relier les mouvements au discours produit en parallèle, et à ce titre, le type fonctionnel attribué à un geste synthétique peut aussi bien refléter une influence du contexte discursif que le fait que le système génère réellement un mouvement communicatif en réponse à l'entrée verbale. Par exemple, certains mouvements peuvent être interprétés comme soulignant une énumération verbale – gestes alors étiquetés de *Parsing* – en raison de leur coïncidence avec des éléments listés dans le discours, bien que cette association soit potentiellement totalement fortuite.

Notre étude constitue la première démarche de ce type utilisant une annotation experte multimodale pour comparer gestes naturels et gestes synthétiques. Bien que ces constats exploratoires ne permettent pas de généraliser sur l'ensemble des capacités du modèle, ils mettent en évidence, de manière concrète, l'intérêt d'une annotation experte pour auditer le modèle et détecter des limites qui resteraient invisibles lors d'évaluations purement perceptives. En confrontant des annotations détaillant les aspects fonctionnels et formels des gestes observés, nous avons pu identifier des caractéristiques saillantes, dès lors mal reproduites ou absentes des gestes de synthèse, telles que, donc, la diversité des trajectoires, les variations de vitesse et d'espace, ou encore l'ancrage au discours. Bien que cette étude exploratoire repose sur un seul modèle et un seul locuteur, ces résultats montrent que l'annotation multimodale constitue un outil efficace pour détecter les limites des gestes synthétiques. Elle permet ainsi de dégager des critères pouvant venir étoffer un cadre d'évaluation plus précis et plus objectif, et ouvre la possibilité de mieux comparer les résultats des différents modèles existants. L'annotation multimodale permet en effet non seulement d'expliciter l'incertitude et la variabilité dans l'interprétation des gestes, mais aussi de mettre en lumière des écarts précis par rapport aux gestes naturels, qui seraient difficiles à détecter avec une évaluation uniquement perceptive telle que réalisée à l'heure actuelle. Elle offre ainsi un moyen de réfléchir à des améliorations ciblées, en identifiant les dimensions fonctionnelles et formelles qui nécessitent un renforcement pour obtenir des gestes plus crédibles, et davantage représentatifs de la gestualité co-verbale humaine.

Une telle évaluation menée sur la base d'annotations expertes reste cependant coûteuse et chronophage, ce qui limite la possibilité de sa généralisation à grande échelle. Une première alternative consiste alors à s'appuyer sur des critères plus facilement vérifiables, permettant de contrôler systématiquement certains aspects de la cohérence et de la richesse des gestes produits. Ainsi des indicateurs tels que la variation spatiale (parcours de distance dans plusieurs dimensions de l'espace) ou encore la dynamique des mouvements (vitesse moyenne des gestes et pics de variations) par exemple pourraient ainsi constituer des métriques objectives du caractère « humain » des gestes produits automatiquement (*human-likeness*). En ce sens, cette étude pose les bases du développement de moyens d'évaluation plus objectifs, susceptibles à terme de compléter les évaluations subjectives [115, 131], encore considérées comme le standard de référence, mais limitées par leur nature peu informative sur les caractéristiques précises des gestes produits.

# Bilan de partie

L'analyse multimodale de la gestualité étudie comment les gestes s'intègrent aux autres modalités pour former un même acte de communication. Le champ des *gesture studies* offre un cadre théorique et méthodologique rigoureux pour l'étude du geste, considéré comme un phénomène ancré dans un contexte discursif spécifique et caractérisé par ses dimensions sémantiques, syntaxiques et pragmatiques. L'annotation multimodale qualitative du geste permet ainsi de le caractériser pleinement grâce à des représentations articulant simultanément sa structure temporelle, ses propriétés formelles, ses fonctions communicatives, et son lien avec le contenu du discours. Disposer de ces représentations interprétatives et structurées du geste est essentiel pour l'analyse multimodale, car elles permettent de transformer un signal de mouvement en informations analytiquement exploitables. L'analyse multimodale du corpus reBeat-14 illustre concrètement cette démarche, en permettant de caractériser systématiquement le geste. L'annotation multimodale permet ainsi de formaliser un phénomène complexe et de faire émerger des régularités et des relations pertinentes ; c'est en ce sens qu'elle constitue une première forme de modélisation du geste.

L'étude de la gestualité dans l'approche computationnelle vise à formaliser le geste via des systèmes automatiques pour l'analyser, le prédire ou le générer automatiquement. La caractérisation fine des gestes issue de l'annotation multimodale constitue un véritable levier méthodologique, en orientant la manière dont les modèles automatiques doivent être conçus. Par exemple, le fait que les gestes observés soient majoritairement de courte durée souligne l'importance d'une résolution temporelle fine pour la détection automatique et la modélisation, afin de capturer précisément ces événements au sein du flux continu de mouvement. De même, les gestes parcourant de grandes distances présentent généralement une amplitude élevée, tandis que ceux caractérisés par de faibles déplacements ont également une amplitude réduite. Toutefois, une proportion non négligeable de gestes présente des amplitudes beaucoup plus importantes et occupe un espace étendu, tant en hauteur qu'en latéralité. Ces résultats ont également des conséquences directes pour la génération automatique de gestes : les modèles ne devraient pas être contraints à une localisation spatiale trop restreinte, mais doivent être capables de produire une variabilité spatiale reflétant les comportements humains, tels qu'observés dans le corpus. Ces annotations multimodales ouvrent ainsi plusieurs perspectives pour des usages computationnels.

Les modèles d'apprentissage restent aujourd'hui insatisfaisants, rencontrant des difficultés à produire des prédictions fiables ou des gestes de synthèse naturels et cohérents. Ces limites peuvent s'expliquer par le fait que, dans certains cas, la distinction entre mouvement non communicatif et geste n'est pas nécessairement réalisée. Les modèles analysent ainsi des représentations multimodales hautement dimensionnelles (intégrant parfois des données de plusieurs modalités de tailles et de natures variées) échantillonnées à une granularité temporelle fine (allant jusqu'à la *frame*), au sein desquelles les gestes constituent des événements ponctuels. Dans cette configuration, le modèle appréhende le geste principalement comme une simple production de mouvement, sans nécessairement intégrer l'ensemble des dimensions qui le caractérisent, tel qu'elles sont explicitées dans l'annotation multimodale. Intégrer explicitement la modélisation du geste issue de l'annotation, dans l'apprentissage des modèles, permettrait de contraindre et d'orienter la représentation latente du phénomène, en les guidant vers les dimensions réellement constitutives du geste. Une telle intégration pourrait ainsi conduire à des productions plus pertinentes et plus cohérentes.

Un autre facteur peut également expliquer les limites des productions de ces systèmes automatiques. Dans le cadre de la génération de gestes, l'évaluation des productions des modèles repose majoritairement sur des expériences perceptives, qui ne permettent pas de mesurer finement les propriétés formelles et fonctionnelles des gestes. L'analyse multimodale des gestes de synthèse, telle que réalisée pour le modèle STARGATE, met en évidence que la modélisation issue de l'annotation permet de dégager des critères d'évaluation plus précis et objectifs. C'est précisément parce que l'entité « geste » est rigoureusement définie et caractérisée que l'on peut formuler des attentes spécifiques vis-à-vis des systèmes automatiques de génération automatique (et également de segmentation). Cette clarification conceptuelle constitue un préalable indispensable à l'évaluation des modèles : elle permet de dépasser des critères purement formels pour interroger la plausibilité communicative des gestes de synthèse produits. L'utilité de cette démarche apparaît de manière particulièrement saillante dans la comparaison menée entre gestes synthétiques et gestes naturels, où l'analyse multimodale rend visibles les écarts persistants dans la durée, l'occupation de l'espace, la coordination bimanuelle ou encore l'affiliation au discours. Cela illustre également le potentiel méthodologique de l'annotation multimodale pour l'approche computationnelle, en fournissant des repères précis pour guider l'évaluation des modèles. En ce sens, les pratiques d'annotation issues des *gesture studies* ne se limitent pas à un rôle descriptif, mais deviennent un véritable levier méthodologique pour structurer les approches computationnelles de la modélisation du geste co-verbal, en orientant à la fois la définition des tâches, la sélection des descripteurs pertinents et les critères d'évaluation des systèmes automatiques.

Au regard des apports méthodologiques et analytiques qu'offre l'expertise issue des *gestures studies*, cette thèse s'inscrit dans une démarche visant à étendre et à transposer les usages et pratiques de ce domaine au-delà de leur champ disciplinaire d'origine, qu'il s'agisse des protocoles d'annotation eux-mêmes ou des cadres théoriques mobilisés pour caractériser le geste co-verbal. En effet, différentes disciplines, comme les sciences du langage, les sciences du mouvement, ou encore les sciences computationnelles travaillent sur ce même objet qu'est le

geste. Dans ce cadre, disposer d'une base conceptuelle et descriptive partagée sur ce qui est observé, segmenté et analysé apparaît comme une condition essentielle pour favoriser le dialogue interdisciplinaire et, par conséquent, soutenir l'avancement conjoint de la compréhension du geste.

# Partie IV
# Apports et perspectives de l'expertise des sciences computationnelles

Cette partie examine les apports des sciences computationnelles à l'approche interdisciplinaire développée dans cette thèse. Elle porte sur l'examen de méthodologies permettant de produire des annotations multimodales et des segmentations automatiques de gestes précises et fiables.

Cette partie se compose de deux chapitres. Le premier est consacré à l'analyse automatique des aspects formels du geste et présente une méthodologie d'annotation fondée sur l'exploitation de la *motion capture*, à travers des méthodes vectorielles et des analyses de trajectoires. Le second porte sur la segmentation semi-automatique de corpus et décrit un premier modèle neuronal dédié à une tâche de classification binaire (*frame* geste/non-geste).

Le défi central de cette partie est de montrer que les approches computationnelles – telles que l'apprentissage automatique ou l'analyse des représentations tridimensionnelles du geste – peuvent contribuer à un enrichissement fiable des corpus en sciences du langage, et par conséquent, pourrait participer à l'amélioration des performances de modèles. Il s'agit à la fois de démontrer que leur raffinement méthodologique permet de pallier certaines limites de l'annotation manuelle sans altérer la qualité, et de proposer des méthodes automatiques conçues à partir des objectifs de l'annotation et des phénomènes à décrire, afin de produire des analyses interprétables et théoriquement informées.

# Chapitre 8
# Annotation automatique des aspects formels du geste manuel co-verbal

## 1 Introduction

Traditionnellement, l'enrichissement formel d'un corpus de gestes est majoritairement réalisé manuellement, en se basant sur différents aspects observables. Dans le cas de reBeat-14, plusieurs annotations ont été produites à partir des vidéos de reconstruction du mouvement afin de caractériser ces aspects formels lors de l'annotation multimodale, tels que la forme de la trajectoire ou la position de la main dans l'espace. Ces descriptions restent toutefois fortement subjectives, et la fiabilité de la caractérisation formelle en est alors limitée. En effet, l'interprétation des seuils perceptifs pour certains paramètres n'est pas toujours claire, ce qui peut entraîner des variations entre annotateurs, ou même un manque de cohérence dans la caractérisation de différents gestes. De plus, se baser sur une représentation 2D d'un mouvement intrinsèquement 3D peut entraîner une perte d'informations spatiales essentielles, limitant ainsi la précision et la fiabilité de ces descriptions. Ainsi les effets de perspective liés à l'angle de caméra ou aux déplacements du locuteur dans l'espace peuvent-ils être trompeurs, altérant la perception de certaines positions ou trajectoires.

L'exploitation des données 3D issues de la motion capture s'offre comme une solution pertinente : elle permet de se concentrer sur le mouvement réel, en conservant toutes ses dimensions et caractéristiques réelles, et non plus seulement sur la perception des mouvements via les vidéos. Cette approche constitue ainsi une alternative prometteuse face aux limites des annotations formelles manuelles. Ce chapitre explore alors le développement d'un système d'encodage pertinent à la fois pour la caractérisation des gestes (établissement de catégories formelles) et pour l'enrichissement des corpus (annotation automatique de ces caractéristiques). Nous faisons ici l'hypothèse qu'une formalisation rigoureuse des paramètres gestuels fondée sur

des critères observables, mesurables et interopérables qui n'est pas dépendante de l'annotation manuelle est (1) indispensable pour garantir précision, fiabilité et objectivité des descriptions et (2) essentielle à la modélisation computationnelle des gestes, en vue de l'amélioration des performances de modèles d'apprentissage.

On présentera donc dans ce chapitre les résultats exploratoires d'une approche méthodologique conçue pour une annotation automatique, cohérente et systématique des caractéristiques formelles du geste. Il commence par une contextualisation des pratiques d'enrichissement formel, avant de décrire l'initialisation du *framework* proposé, qui comprend le choix des descripteurs et la manière dont les différents aspects formels du geste sont conceptualisés. Sont ensuite présentés les modes de qualification formelle automatique, où l'on montre comment les paramètres quantitatifs issus de la motion capture peuvent être discrétisés et exploités pour caractériser qualitativement les aspects formels des gestes. Enfin, le chapitre introduit *Gesture User Interface* (GUI), l'interface développée dans le cadre de cette thèse, qui permet de visualiser les gestes annotés selon leurs différents aspects formels. Il convient toutefois de préciser que ce chapitre ne propose pas l'implémentation d'un nouvel outil d'annotation automatisée à proprement parler, mais présente plutôt un cadre méthodologique et exploratoire pour la formalisation de descripteurs formels en vue d'une annotation plus objective et reproductible.

Ce chapitre se clôt sur l'examen de la manière dont l'exploitation ciblée des données de *motion capture*, via des outils informatiques, constitue un levier pour améliorer la rigueur, la reproductibilité et l'objectivité de l'enrichissement, en permettant une caractérisation formelle systématique fiable.

# 2 Contextualisation

La description du mouvement peut se faire de nombreuses manières, allant de la transcription sous forme de dessins (comme dans les partitions de mouvement en danse, proposant une représentation iconique des chorégraphies [165, 166]), à des descriptions qualitatives, avec l'utilisation de termes qualifiant différents aspects du mouvement [59, 167, 168], en passant par des approches géométriques déterminant les changements de positions des membres dans l'espace [169].

La description formelle du geste s'appuie aujourd'hui sur un ensemble particulièrement riche et diversifié de paramètres. La majorité des approches classiques repose toutefois sur des descriptions qualitatives, fondées sur l'observation. Parmi les aspects fréquemment considérés, on peut citer tout d'abord la manualité, soit le fait qu'un geste soit produit avec une ou deux mains, avec des distinctions possibles entre configurations symétriques et asymétriques. La localisation spatiale du geste est également fréquemment décrite, le plus souvent en s'appuyant sur la modélisation de l'espace gestuel proposée par McNeill [9] qui structure l'espace en différentes zones selon la latéralité et la hauteur. Cette description est parfois complétée par une troisième dimension, la profondeur, envisagée comme la distance au corps [50]. La configuration de chaque main exécutant le geste constitue une autre dimension descriptive

fréquente. Elle peut être caractérisée à partir d'inventaires empruntés aux langues des signes [54, 55] ou de catalogues descriptifs élaborés spécifiquement pour la description des gestes co-verbaux [8]. Ces descriptions peuvent inclure non seulement la forme globale de la main (ouverte, fermée, en pince, pointée, etc.), mais aussi la disposition des doigts [8], ou l'orientation de la paume [53]. Les qualifications formelles peuvent également caractériser la direction ou l'orientation dans l'espace d'un geste [167, 60, 53], ou encore la forme de la trajectoire du geste.

Avec l'émergence de nouvelles technologies, et en particulier la *motion capture*, il devient possible d'analyser directement le mouvement au travers des représentations qu'elles offrent. Ainsi, à ces descripteurs qualitatifs s'ajoutent des paramètres quantitatifs, élargissant la gamme des aspects formels du geste qui peuvent être analysés. Les dimensions dynamiques sont particulièrement considérées, notamment la vitesse, dont la définition opérationnelle varie. Par exemple, dans les travaux de Ferstl [62], la vitesse est déterminée selon le pic maximal atteint au cours du geste. La vitesse du geste peut également être une mesure moyenne sur l'ensemble de la production. À ces indicateurs peuvent aussi s'ajouter des dérivés cinématiques, tels que l'accélération, qui, dans le cas des travaux de Ferstl, est évaluée à partir du premier pic de vitesse. La taille du geste est elle aussi appréhendée de différentes manières, par exemple à travers la distance totale parcourue par la main ou l'amplitude maximale atteinte dans l'espace. D'autres paramètres complémentaires, tels que la rotation du bras (*arm swivel*) [59] ou la distance entre les mains dans les gestes bimanuels [59] sont également utilisés.

Si de nombreuses études s'accordent sur le fait que le geste peut être décrit en termes de forme observable et de mesures quantitatives, les critères retenus varient fortement d'un travail à l'autre, donnant lieu à des ensembles de paramètres hétérogènes, parfois redondants ou se recouvrant partiellement. Les définitions opérationnelles de ces descripteurs diffèrent également selon les études : les choix méthodologiques varient à la fois du point de vue de la sélection des variables, de leurs modes de calcul, des unités d'analyse et des seuils d'interprétation. En outre, ces descripteurs sont généralement obtenus selon deux approches distinctes : soit par annotation manuelle, ce qui introduit une dimension subjective liée à l'évaluation des annotateurs, soit par mesures quantitatives, qui fournissent des valeurs continues (vitesse, amplitude, angles, trajectoire, etc.) mais qui ne traduisent pas directement la perception qualitative ou les caractéristiques formelles du geste.

Dans cette étude, nous considérons qu'il est possible de caractériser formellement les gestes de manière automatique et fiable à partir de données mesurables issues de la *motion capture*. Pour ce faire, un ensemble de descripteurs a été sélectionné, considéré comme cohérent et pertinent, et capable de rendre compte efficacement des différents aspects formels du geste. Pour chacun de ces descripteurs, nous proposons une conceptualisation précise des aspects qu'ils visent à décrire et expliquons la manière dont ils sont envisagés dans notre approche. Afin de qualifier de manière objective les différentes dimensions du geste, chaque paramètre est discrétisé à partir de valeurs mesurables, avec des seuils permettant de transformer les variations continues en catégories exploitables. Grâce à cette discrétisation, il devient possible de qualifier de façon reproductible les différentes variantes d'un geste au sein de chaque

dimension envisagée, en dépassant les limitations de la qualification subjective et en fournissant une base cohérente pour des analyses comparatives.

Le *framework* proposé dans ce chapitre constitue un premier pas vers une automatisation complète de l'annotation des aspects formels des gestes. Cela est rendu possible grâce à deux conditions essentielles : les gestes sont préalablement segmentés (manuellement ou automatiquement), et les données de *motion capture* correspondantes sont disponibles, permettant d'analyser précisément les mouvements au sein des segments annotés comme gestes. L'approche propose donc un cadre méthodologique combinant dimensions descriptives et computationnelles, reliant les caractéristiques formelles extraites (position, orientation, trajectoire, etc.) à une structure d'annotation exploitable par des experts humains comme par des systèmes automatiques. L'objectif est ainsi de montrer comment les technologies de capture du mouvement permettent d'obtenir des descriptions objectives, spécifiquement dédiées à la caractérisation formelle du geste, dans une démarche interdisciplinaire combinant analyse fine du mouvement et approche qualitative du geste.

# 3 Initialisation du *framework*

## 3.1 Identification des variantes de formes dans le corpus reBeat-14

Dans le cadre de l'enrichissement formel de reBeat-14, différentes formes de trajectoires ont été annotées manuellement par les annotatrices. L'analyse de ces trajectoires dans le corpus a permis d'identifier neuf formes récurrentes prédominantes : courbes, lignes, progressions scandées, boucles progressives, boucles simples, moulinets, mouvements de va-et-vient, *Hapax* (gestes associés à des significations très spécifiques), ainsi que d'autres formes plus marginales regroupées sous la catégorie « autres ».

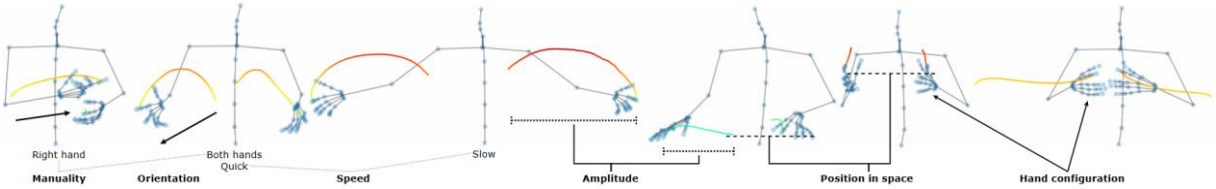



*Figure 37. Six aspects identifiés pour distinguer les variants dans chaque catégorie de trajectoire*

À partir de ces formes, six dimensions ont été identifiées comme essentielles pour distinguer les variantes au sein de chaque catégorie de trajectoire (Figure 37) : la manualité (main droite, gauche ou bilatérale), l'orientation (ex. : le geste va vers l'avant ou sur le côté), la vitesse d'exécution, l'amplitude du geste, la position du geste dans l'espace du locuteur et la configuration manuelle (ex. : la main est ouverte avec la paume vers le bas).

Plusieurs descripteurs ont alors été définis pour caractériser ces différents aspects formels, permettant de qualifier le geste et ses différentes variantes d'exécution à différents niveaux. À l'échelle du geste dans sa globalité, on considère ainsi l'orientation de sa trajectoire, sa position dans l'espace, la manualité engagée, sa vitesse et son amplitude. Pour chaque main, les mesures

portent sur leur orientation globale ainsi que sur l'orientation de la paume dans l'espace, et pour les doigts, les caractéristiques calculées ont trait à leur configuration relative entre eux et par rapport à la paume.

À l'exception de la manualité, tous ces descripteurs sont déterminés de manière objective, grâce à des calculs automatisés basés sur l'extraction des positions 3D et sur les mesures géométriques et cinématiques qui en découlent. Ce niveau de détail permet d'appréhender la finesse articulatoire des gestes selon une approche hiérarchique, décrivant le geste depuis le mouvement global jusqu'à la configuration fine des doigts. Tous les aspects pris en compte dans ce modèle ont été sélectionnés en tant qu'ils pourraient, en contexte, être porteurs de variations de sens.

## 3.2 La manualité : un cas spécifique pour la reconnaissance automatique

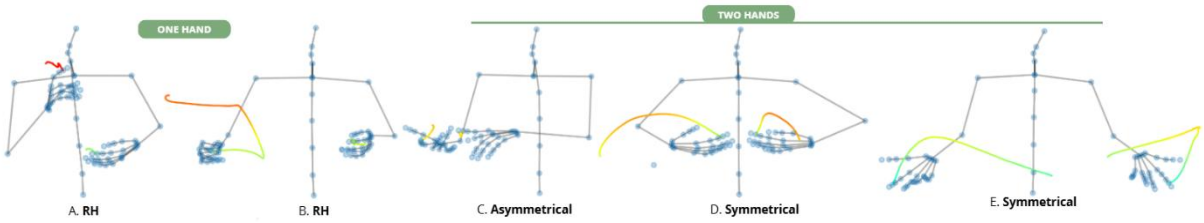



*Figure 38. Illustration de trajectoires selon la manualité*

Parmi les différents aspects formels retenus pour décrire les gestes, la manualité est le seul aspect qui ne puisse pas encore être qualifié automatiquement à partir de descripteurs géométriques ou cinématiques. En effet, même lorsque le geste est monomanuel, la main non impliquée reste rarement complètement immobile (Figure 38). Le mouvement associé peut ainsi être proche de 0 en termes de vitesse, ou au contraire être en mouvement sans pour autant participer à l'élaboration du geste. De même, des différences dans la dynamique (forme de la trajectoire, vitesse, position, etc.) des mains ne suffisent donc pas à permettre de repérer des gestes monomanuels, car il existe des gestes bimanuels asymétriques, réalisant des mouvements complémentaires du point de vue du sens véhiculé. En outre, les différences de dynamique entre les mains peuvent également être observées au sein de les gestes considérés comme « symétriques », qui souvent présentent des asymétries notables en termes de trajectoire, d'amplitude ou d'orientation [59]. Les méthodes automatiques basées uniquement sur le mouvement peinent ainsi à distinguer des mouvements non communicatifs des mouvements réellement impliqués dans la réalisation de gestes, qu'ils soient asymétriques ou imparfaitement symétriques. L'annotation de la manualité demeure donc, en l'état, un processus manuel et interprétatif tel que réalisé dans reBeat-14.

La détermination automatique de la manualité pourrait toutefois, à plus long terme, être envisagée à l'aide de méthodes d'apprentissage profond. Mais ce point dépasse le cadre du présent chapitre, qui se concentre sur les autres aspects formels, déterminés ici à partir de critères géométriques et cinématiques basés sur l'analyse de mouvements.

# 4 Qualification formelle automatique

Les calculs utilisés pour quantifier les différents descripteurs formels de ce *framework* reposent sur les données de *motion capture*, qui permettent d'extraire avec précision les informations spatio-temporelles associées aux gestes. Pour rendre ces mesures quantitatives exploitables, nous proposons des seuils de variation permettant de transformer les valeurs continues en catégories qualitatives interprétables pour la caractérisation des gestes dans ses différentes variations. Cette section présente la méthodologie des calculs automatiques proposés, organisée autour de quatre composantes principales : (1) la localisation des gestes dans l'espace du locuteur, (2) la qualification des orientations du geste, de la main et de la paume, (3) la détermination de la configuration des doigts, et (4) la qualification de la dynamique du geste, incluant la vitesse et sa taille. Chaque élément est dérivé des positions articulaires enregistrées et calculé à partir d'opérations basées sur une combinaison de mesures de distances et de vecteurs.

## 4.1 Localisations au travers des densités de position et des caractéristiques anatomiques

### 4.1.1 Détermination des positions et de l'étendue gestuelle du locuteur

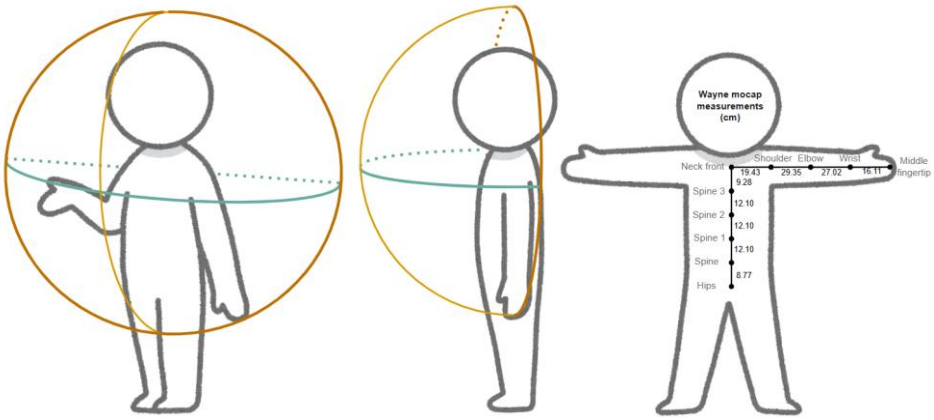



*Figure 39 .Illustration de l'étendue de l'espace gestuel et mensurations du locuteur Wayne (cm)*

La localisation spatiale des gestes consiste à déterminer leur position dans l'espace environnant du locuteur. Cette détermination s'effectue dans un espace restreint, représenté par une demi-sphère correspondant à l'amplitude maximale des bras dans toutes les directions (latérale, avant-arrière et verticale), et décomposé en trois dimensions : latéralité, hauteur et profondeur (Figure 39). Cette représentation rejoint le concept de kinésphère développé par Laban [168] dans le cadre d'analyse du mouvement en danse, qui définit l'espace personnel autour du corps dans lequel le mouvement peut se déployer.

Le périmètre de la demi-sphère peut être automatiquement déterminé à partir de la projection de la distance entre le marqueur placé à l'avant du cou et les marqueurs situés aux extrémités des index. Ces mesures de référence peuvent être capturées au début d'une session

de *motion capture* en enregistrant l'amplitude maximale des bras du locuteur (*range of motion*) dans toutes les directions possibles. Lorsque de tels enregistrements de référence ne sont pas disponibles (comme dans le corpus BEAT), la sphère peut être approximée à partir de la longueur totale du bras, en additionnant la longueur des segments morphologiques (distance entre chaque *joint* composant le squelette de l'agent virtuel) allant de l'avant du cou jusqu'aux doigts (92 cm pour Wayne). Cette approximation produit une sphère légèrement surdimensionnée, puisque le segment entre le cou et les épaules ne peut pas s'étendre complètement vers l'avant, le haut ou le bas lorsque le bras est levé dans ces directions.

L'espace gestuel est donc défini à partir des mesures corporelles réelles du locuteur, et se présente comme dynamique et adaptable : il suit les changements de posture du locuteur, étant centré sur des marqueurs situés le long de sa colonne vertébrale. Pour déterminer la localisation d'un geste dans cet espace, la ou les mains impliquées dans sa réalisation sont localisées au sein de cette demi-sphère, discrétisée en plusieurs zones. Cette discrétisation a été conceptualisée de manière à fournir un référentiel qui soit à la fois efficace et pertinent pour la classification des gestes dans l'espace. Elle a ainsi été déterminée à partir de l'observation de la distribution de certaines positions de Wayne lors de l'ensemble des gestes annotés dans reBeat-14, combinées à certaines de ses mesures anatomiques.

### 4.1.2 Mode de calcul des densités de positions des gestes

L'espace gestuel a été divisé en différentes zones fondées sur un reflet de l'occupation réelle de l'espace plutôt que déterminées de manière arbitraire. Pour ce faire, la distribution des positions des articulateurs (avant du cou, épaules, coudes, poignets et extrémités des index) a été analysée dans le corpus pendant l'élaboration de l'ensemble des gestes. Plusieurs zones ont alors pu être distinguées, des zones « typiques », où les gestes se produisent le plus fréquemment, aux zones plus périphériques.

Pour mener cette analyse de l'occupation de l'espace, les gestes annotés dans reBeat-14 ont d'abord été extraits des fichiers d'annotation et convertis en indices de *frames* (120 *fps*). Les coordonnées 3D de chaque articulation ont ensuite été extraites, pour toutes les *frames* annotées comme étant constitutives d'un segment de geste. Les marqueurs droits ou gauches ont été pris en compte en fonction de la manualité annotée (RH, LH ou bimanuel), afin de seulement extraire les positions 3D des articulateurs impliqués dans le geste.

Pour chaque axe (X : droite–gauche, Y : haut–bas, Z : avant–arrière) et chaque articulation, les données de position ont été agrégées sur l'ensemble des gestes du corpus. Les densités de probabilité ont été estimées à l'aide d'une fonction de pondération (noyau gaussien), et la valeur résultante a été évaluée sur une grille régulière de 500 points, produisant une courbe univariée représentant la probabilité d'observer des positions articulaires spécifiques le long de chaque axe. Ces densités estimées ont été visualisées avec Matplotlib sur trois sous-graphiques correspondants aux trois axes spatiaux (X, Y, Z). Les courbes de densité ont été tracées (articulateur à droite = tiretée ; articulateur à gauche = pleine), et les maxima locaux (positions les plus probables) ont été marqués par une croix rouge.

Afin de s'assurer que les positions observées ne soient pas biaisées par les changements de postures du locuteur, les données ont été alignées sur son torse en remplaçant le repère global original de BEAT (R1) par un nouveau repère orthonormé centré sur le corps du locuteur (R2). Chaque fichier `.bvh` a ainsi été traité dans Blender pour réexprimer les positions et rotations des articulations par rapport à ce nouveau repère plutôt qu'à la scène globale. Ce nouveau référentiel R2, défini par l'os `Spine2` et centré à l'avant du cou, se déplace dès lors dynamiquement avec le locuteur. Cela a permis d'obtenir des données spatiales cohérentes, rendant possible l'analyse des positions des gestes indépendamment de la localisation ou de l'orientation du corps de Wayne.

### 4.1.3 Mapping théorique de l'espace gestuel

En se basant sur la densité observée des positions des articulateurs, l'espace gestuel a été divisé en 96 zones réparties sur les 3 plans dimensionnels de la latéralité, la hauteur et la profondeur.

#### Axe droite-gauche

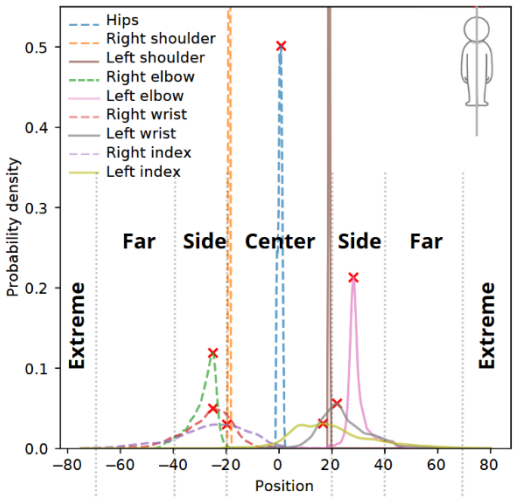



*Figure 40. Probabilités de positions des joints sur l'axe X*

Concernant la latéralité, l'observation des distributions de positions révèlent que les gestes de Wayne sont principalement produits à proximité de son corps, ceux produits de la main gauche étant centrés devant le torse gauche (pointe de l'index 16,63 cm, épaule 18,81 cm) et ceux produits de la main droite étant situés près de l'axe de l'épaule (index 19,70 cm, épaule 18,78 cm). Globalement, les deux mains couvrent également des amplitudes similaires (≈23 cm), mais dans des zones distinctes : la main droite atteint 38,61 cm vers la droite depuis la ligne médiane du corps, tandis que la main gauche reste dans un rayon de 30 cm vers la gauche depuis cette même ligne. Sur la base de ces observations, l'axe gauche-droite est alors divisé en huit zones (quatre de chaque côté de la ligne médiane du corps ; Figure 40).

- **Zones centrales** : elles s'étendent jusqu'à 20 cm de la ligne médiane, de part et d'autre de celle-ci, correspondant approximativement à la distance moyenne cou-épaule (18,80 cm). Situées devant le torse du locuteur, elles correspondent aux zones où se concentrent la plupart des gestes de la main gauche de Wayne et quelques-uns de ses gestes de la main droite.

- **Zones latérales** : elles sont situées à 20cm de la ligne médiane et s'étendent jusqu'à 40 cm de celle-ci, couvrant l'espace où sont réalisés la majorité des gestes de la main droite de Wayne et les gestes les plus externes de sa main gauche.

- **Zones éloignées** : elles s'étendent de la fin des zones latérales à 70cm de la ligne médiane, ce qui correspondant à l'amplitude maximale du bras (avant-bras et main totalisant 43,5cm) lorsque placé avec le coude sur la position moyenne observée dans le corpus (≈26cm depuis la ligne médiane).
- **Zones extrêmes** : elles couvrent l'espace allant de 70 cm de la ligne médiane à l'extension maximale des bras, atteignant ainsi les limites anatomiques de l'espace gestuel.

### Axe haut-bas

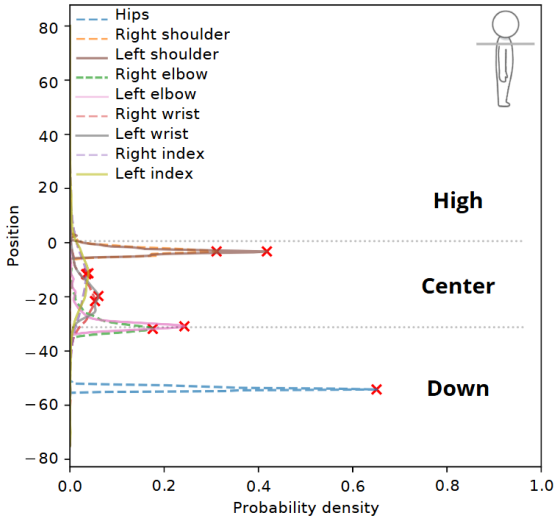



*Figure 41. Probabilités de positions sur l'axe Y*

En termes de hauteur, les densités de position présentent des valeurs relativement stables pour chaque paire d'articulateurs — ≈3 cm sous le cou pour les épaules, ≈32 cm pour les coudes, ≈20 cm pour les poignets et ≈11 cm pour les index — avec des amplitudes assez similaires des deux côtés. L'axe haut–bas est alors divisé en trois zones (Figure 41).

- La **zone centrale** s'étend du devant du cou jusqu'à 30 cm en dessous – correspondant approximativement à la hauteur moyenne du coude (≈32 cm) – englobant la majorité des positions des gestes observés.
- La **zone basse** se prolonge de 30 cm sous le devant du cou jusqu'au point le plus bas atteignable anatomiquement par le bras.
- La **zone haute** commence au niveau du devant du cou et s'étend jusqu'au point le plus haut atteignable par le bras.

### Axe intérieur-extérieur

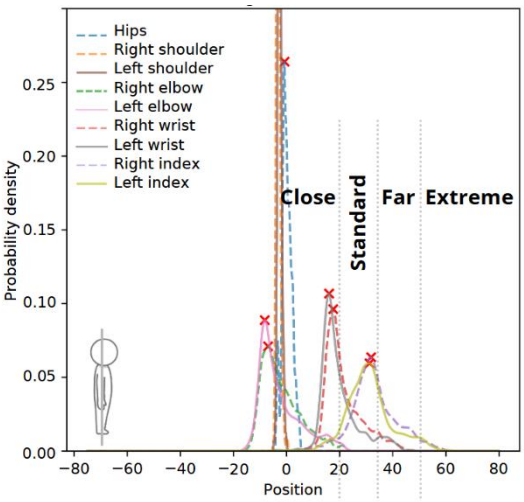



*Figure 42. Probabilités de positions sur l'axe Y*

En termes de profondeur, les positions des articulateurs du côté droit et du côté gauche sont également assez similaires, avec des distances moyennes d'environ ≈7,5 cm pour les coudes depuis le cou, 17,5 cm pour les poignets et ≈31,5 cm pour les index, présentant des amplitudes comparables des deux côtés. L'axe Z est alors divisé en quatre zones (Figure 42), comme décrit ci-dessous.

- **Zone proche** : du corps jusqu'à 20 cm en avant, correspondant approximativement à la longueur de la main (16 cm). Cette zone couvre l'espace où se produisent les gestes de contact avec

le corps, se terminant là où débutent les principales zones de concentration des gestes.

- **Zone standard** : de 20cm à l'avant du corps jusqu'à 35 cm de celui-ci, ce qui correspond aux limites atteintes par le bras (43 cm) lorsque le coude est placé dans sa position moyenne (≈8,5 cm derrière les épaules). Cette zone est celle où se concentrent quasiment tous les gestes des deux mains observées dans le corpus.
- **Zone éloignée** : depuis la fin de la zone standard jusqu'à 50 cm devant le corps, ce qui correspond cette fois à la longueur du bras (43 cm) depuis la position la plus avancée du coude observée (7 cm devant le corps), et ce qui couvre les emplacements les plus extrêmes des index observés dans le corpus.
- **Zone extrême** : depuis la fin de la zone éloignée jusqu'à l'extension anatomique maximale du bras.

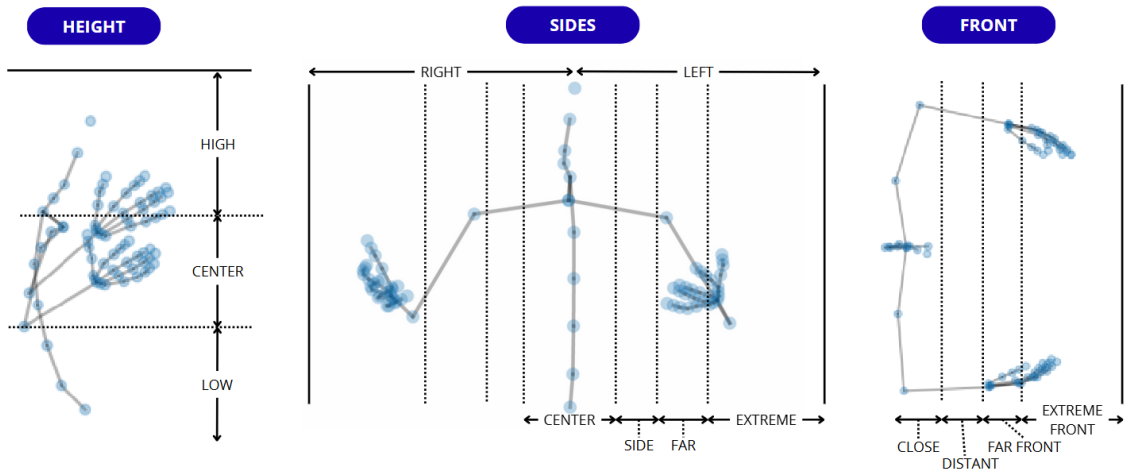



*Figure 43. Schéma de la cartographie globale, basée sur les observations de densité*

On peut observer que la cartographie de l'espace gestuel, issue de l'analyse des densités de position, (Figure 43), et selon les axes latéral et vertical, recoupe partiellement le découpage proposé par McNeill [9], apportant une justification objective à ses observations empiriques concernant la répartition des gestes dans l'espace du locuteur. Cette cartographie est également conçue pour une exploitation automatique : la localisation du geste peut alors être qualifié très précisément à partir des zones traversées par le poignet et tous les doigts tout au long de son exécution, sans risque que sa position soit perçue de façon aléatoire au sein de ses zones. Notre approche introduit en outre une nouvelle dimension gérées simultanément, la profondeur. L'espace gestuellement est alors pleinement considéré comme tridimensionnel, ouvrant la voie à des analyses plus fines et robustes des variations de localisations spatiales des gestes. De plus, cette cartographie est adaptative, sa partition s'ajuste aux mesures corporelles du locuteur (dimension des mains, des avant-bras...), et à la dynamique de son corps, puisque les repères sur lesquels elle est centré sont sur la colonne vertébrale du locuteur.

## 4.2 Orientations au travers de vecteurs

### 4.2.1 Espace tridimensionnel local discrétisé : une sphère de 64 zones

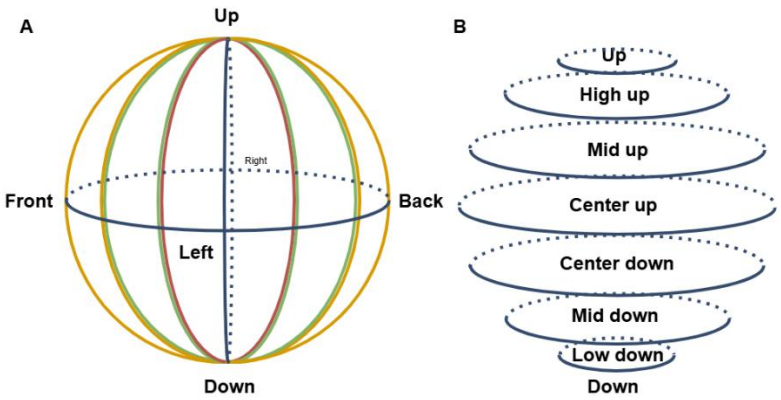



*Figure 44. Sphère 3D pour la détermination de l'orientation dans l'espace*

Pour calculer une orientation, des vecteurs (décrits ci-dessous) sont projetés dans cette sphère 3D discrétisée, centrée sur l'origine de chaque vecteur et intégrée dans l'espace gestuel décrit précédemment. Différents seuils divisent la sphère en 64 régions, qui permettent de distinguer les différentes directions de la trajectoire d'un geste, ou encore l'orientation de la main ou de la paume, et ce indépendamment de l'emplacement du geste dans l'espace du locuteur. L'azimut (az), qui correspond à la rotation dans le plan XZ (latéralité et profondeur, Figure 44A), est ainsi découpé en 8 quartiers, tout comme l'élévation (el), qui correspond à la variation le long de l'axe Y (hauteur, Figure 44B), est découpée en 8 tranches. Cette représentation en sphère rejoint le concept de *Systeme of reference* proposé par Eshkol et Harries [169] pour l'analyse du déplacement des membres dans le cadre du modèle développé pour transcrire les mouvements dans la pratique de la danse.

Dans notre modèle, chaque vecteur $\vec{v}$ est alors normalisé en un vecteur unitaire $u = \frac{\vec{v}}{|\vec{v}|}$ en divisant ses composantes par sa norme euclidienne, afin de ne conserver que la direction du vecteur, puis les coordonnées sphériques correspondantes sont déterminées en deux étapes. Ainsi, les composantes x, y, et z du vecteur unitaire sont comprises entre -1 et 1.

**Quantification des angles**

L'angle d'azimut $\theta$ est calculé, dans le plan horizontal (X, Z : droite-gauche / avant-arrière) donc autour de l'axe Y, par la fonction : $\theta = arctan2(z, x)$. Cette fonction est utilisée parce qu'elle tient compte du quadrant et renvoi un angle continu compris dans l'intervalle $(-\pi, \pi]$. Afin d'obtenir une représentation angulaire définie sur un domaine positif et compatible avec la quantification uniforme ultérieure, nous remappons cet angle dans l'intervalle $[0, 2\pi]$ en ajoutant $2\pi$ aux valeurs négatives afin d'obtenir l'intervalle en radiant sur un cercle trigonométrique complet.

L'angle d'élévation $\phi$, correspondant à l'élévation vers le haut ou le bas, est déterminé à partir de la composante z du vecteur, par la fonction : $\phi = arcsin(y)$, avec un angle en radians dans l'intervalle $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

**Qualification des angles**

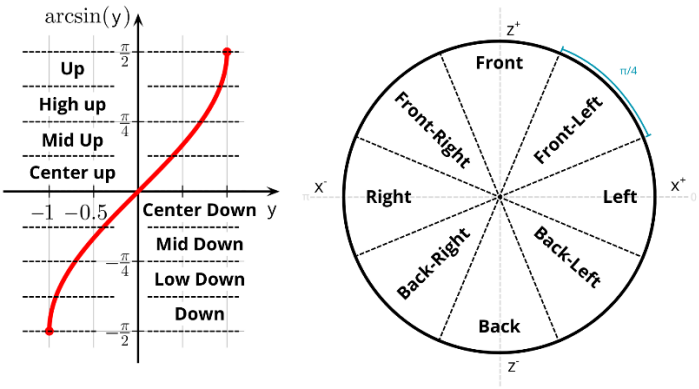



*Figure 45. Détermination des étiquettes symboliques depuis les valeurs d'angle*

Les valeurs correspondant aux bornes minimale et maximale de l'intervalle $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ pour l'élévation sont ensuite discrétisées en huit sous-intervalles uniformes. La valeur de l'angle $\phi$ est alors remplacée par l'étiquette de sa hauteur correspondante, déterminée par le sous-intervalle dans lequel elle se situe (Figure 45). Dans le cas de l'azimut, les bornes minimale et maximale de l'intervalle $[0, 2\pi]$ sont aussi discrétisée en huit sous-intervalles uniformes, répartissant ainsi le cercle azimutal en 8 bins égaux sur les axes $\frac{\pi}{8}$ $mod$ $\frac{\pi}{4}$ . La valeur de l'angle $\theta$ est alors remplacée par l'étiquette de sa position correspondante dans le plan horizontal, déterminée par le sous-intervalle dans lequel elle se situe.

La combinaison de ces deux étiquettes permet de qualifier l'orientation du vecteur dans l'espace tridimensionnel de la sphère le long des axes XYZ.

### 4.2.2 Orientation de la trajectoire du geste

Pour caractériser géométriquement une trajectoire gestuelle, quatre points (`A, B, C, D`) sont extraits (Figure 46). Les points A et B correspondent respectivement aux positions de début et de fin de la trajectoire, tandis que C est défini comme le point central du vecteur $\overrightarrow{AB}$, représentant le centre de la trajectoire du geste. Le point D correspond à l'endroit où la trajectoire s'écarte le plus de C, capturant ainsi la direction vers la plus grande amplitude du geste. Il est obtenu en calculant $\overrightarrow{CP_i}$ pour chaque position échantillonnée de l'extrémité de l'index $P_i$ et en sélectionnant le vecteur $\overrightarrow{CD}$ de plus grande norme.

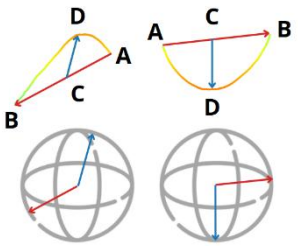



*Figure 46. Illustration des vecteurs utilisés pour la détermination de l'orientation du geste (vecteur rouge = direction principale ; vecteur bleu = direction secondaire)*

Les deux vecteurs définis fournissent une description géométrique concise du geste dans l'espace 3D. Ils sont projetés dans

l'espace de coordonnées sphériques afin de déterminer leurs orientations. La projection de $\overrightarrow{AB}$ indique ainsi la direction principale du mouvement – par exemple vers l'avant-droite ou vers le haut et la gauche – tandis que la projection de $\overrightarrow{CD}$ capture la direction secondaire, utile pour décrire les variantes de gestes courbes – par exemple des arcs vers le haut, vers l'extérieur ou latéraux (Figure 46).

### 4.2.3 Orientation globale de la main et de la paume

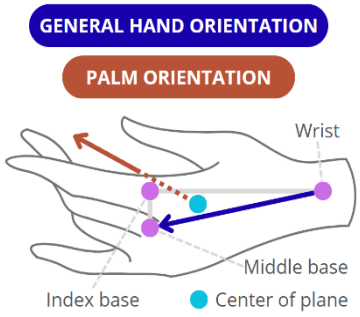



*Figure 47. Illustration des vecteurs pour la trajectoire globale de la main (vecteur bleu) et de la paume (vecteur rouge)*

Les calculs vectoriels permettent également de fournir une description objective de deux aspects clés de la configuration de la main : son orientation globale et l'orientation de sa paume. La méthodologie présentée ici ne permettant pas pour l'instant de prendre en compte la dynamique des articulateurs sur la totalité de l'élaboration du geste, ces informations sont uniquement extraites à partir de la dernière *frame* de chaque *stroke*.

Le vecteur d'orientation de la main (vecteur bleu) est défini comme le vecteur allant du poignet à la base du majeur (Figure 47). L'orientation de la paume (vecteur rouge) est quant à elle calculée comme le vecteur normal au plan défini par trois articulations de la main : le poignet, la base de l'index et la base de l'auriculaire. Afin de garantir que ce vecteur représente systématiquement une direction orientée vers l'extérieur de la paume, indépendamment de la main considérée (droite ou gauche), son orientation est corrigée à l'aide d'un vecteur de référence reliant le poignet à la base du pouce. Ce vecteur encode l'orientation anatomique de la main ; lorsque le produit scalaire entre le vecteur étudié et cette référence est négatif, cela indique qu'ils pointent dans des directions opposées. Les deux vecteurs sont ensuite projetés dans l'espace sphérique 3D, permettant de déterminer leurs directions et de caractériser l'orientation de l'élément correspondant – main ou paume.

## 4.3 Configurations des doigts au travers de calcul de distances

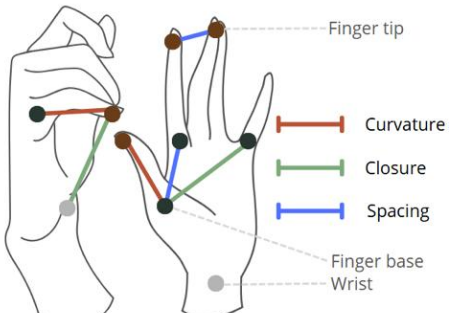



*Figure 48. Illustration des marqueurs utilisés pour le calcul de distance*

La configuration de la main implique également la description de l'arrangement des doigts. Nous avons mis en œuvre plusieurs méthodes pour quantifier la géométrie de chaque doigt sur la dernière *frame* de chaque *stroke*. Ici encore, ce calcul n'est pas dynamique sur la totalité de l'élaboration du geste. Les configurations sont définies par trois paramètres : la courbure (flexion du doigt), la fermeture (repli vers la paume) et l'écartement (distance entre les doigts). Chaque paramètre est estimé à partir des distances entre des marqueurs sélectionnés (Figure 48). Les variations

sont ensuite mises à l'échelle sur une référence allant des distances minimales aux distances maximales de chaque doigt. Ces extrêmes peuvent être automatiquement capturés au début de l'enregistrement en enregistrant l'amplitude maximale de flexion, de fermeture et d'écartement de chaque doigt lors du *range of motion*. Lorsque ces données ne sont pas disponibles (comme dans le cas du corpus BEAT), certaines mesures et seuils peuvent être approximés à partir de longueurs morphologiques de segments anatomiques (distance entre articulateurs capturés lors de la *mocap*) et de calculs alternatifs.

### 4.3.1 La courbure du doigt

La courbure est estimée à partir de la distance entre la base et l'extrémité d'un doigt, les distances les plus courtes correspondant à une plus flexion plus importante. La courbure est classée en quatre catégories (tendu, mi-tendu, mi-fléchi, fléchi), par la partition de l'espacement entre l'extension maximale et la flexion maximale en intervalles égaux. Si ces distances clés ne sont pas enregistrées, la distance maximale peut être calculée en additionnant les longueurs des segments entre la base et l'extrémité du doigt, tandis que la distance minimale peut-être fixée à une valeur théorique de 2 cm – les marqueurs de la base de l'extrémité d'un doigt ne pouvant pas se superposer lorsque les doigts sont repliés dans la paume.

### 4.3.2 La fermeture du doigt

La fermeture est calculée comme la distance entre la base du pouce et la base de l'auriculaire pour les pouces, et pour les autres doigts, entre la base de chaque doigt et le poignet ; des distances plus petites correspondent à une fermeture plus importante. La fermeture est classée en quatre variantes (ouverte, mi-ouverte, mi-fermée, fermée), par la partition de l'espacement entre l'ouverture totale et la fermeture maximale en quatre intervalles égaux. Si ces distances ne sont pas enregistrées, elles peuvent là aussi être estimées – sauf pour le pouce, pour lequel aucune alternative n'a été trouvée – en additionnant la distance de chaque articulateur entre le poignet et la première phalange pour l'ouverture maximale. La distance minimale (zéro théorique) est estimée en projetant la première phalange à 45° depuis la base du doigt vers le poignet, reflétant les limites anatomiques de flexion.

### 4.3.3 L'espacement des doigts

L'espacement des doigts est estimé en mesurant, pour les pouces, la distance entre sa première phalange et la base de l'index, et pour les autres doigts, entre les premières phalanges adjacentes, les distances plus grandes indiquant un écartement plus important. Deux ensembles de variantes reflètent les différences anatomiques. L'index, le majeur, l'annulaire et l'auriculaire, peuvent-être tour à tour qualifiés de standard (doigts à côté/repos), d'écartés, ou d'éloignés (un doigt plié vers l'intérieur). Le pouce peut quant à lui être estimé comme relevant d'une position proche (touchant l'index), standard (repos), ou écartée. Les variantes résultent là encore d'une partition uniforme de l'intervalle minimal–maximal d'écartement. Étant donné que l'écartement dépend des positions relatives de plusieurs articulateurs, aucune estimation

alternative n'est possible : les mesures de références doivent être enregistrées pour que ces qualifications puissent être réalisées automatiquement.

## 4.4 Dynamique gestuelle : la vitesse et la taille du geste

La dynamique des gestes est caractérisée grâce à différent paramètres (Tableau 42). La vitesse est calculée comme le déplacement moyen de l'extrémité de l'index entre les *frames* constituant un segment de geste, divisé par la durée d'une *frame* (120 fps). La vitesse moyenne globale du geste est ensuite calculée en cm/s, puis est ensuite catégorisée à l'aide des quartiles : rapide (supérieure au Q3), standard (autour de la médiane) et lente (inférieure au Q1).

| Variable | Q1 | Med | Q3 |
|---|---|---|---|
| Vitesse moyenne cm/s | 48,20 | 70,3 | 100,78 |
| Nombre de *frames* | 32 | 44 | 62 |
| Distance AB | 9,55 | 17,02 | 26,65 |
| Zones X traversées | 1 | 2 | 2 |
| Zones Y traversées | 1 | 1 | 2 |
| Zones Z traversées | 1 | 1 | 2 |

*Tableau 42. Propriétés dynamiques du geste dans le cas de reBeat-14*
*(Q1 = premier quartile ; Med = médiane ; Q3 = troisième quartile)*

La taille du geste est décrite au moyen de plusieurs mesures complémentaires. Tout d'abord, l'amplitude globale est calculée comme la distance euclidienne entre les points de départ et d'arrivée d'une trajectoire. Ensuite, le nombre de points échantillonnés lors d'un geste fournit une mesure de la durée du geste (elle est, par ailleurs, en lien avec la distance parcourue du geste). Enfin, la trajectoire dans l'espace gestuel du locuteur est analysée par la détermination du nombre et du type de zones traversées. Ces trois mesures sont nécessaires, car aucune ne suffit à elle seule à définir pleinement la taille d'un geste : la distance euclidienne entre départ et arrivée ne permet pas de différencier des gestes de même longueur mais de hauteurs différentes, tandis que le nombre de points de trajectoire ne distingue pas une boucle répétée compacte d'une courbe ou d'une ligne avec un nombre de *frames* similaire – par exemple, de petites boucles répétées versus une grande boucle étendue –, distinction qui n'est révélée que par le nombre et le type de zones spatiales traversées.

Pour tous les paramètres, les valeurs considérées comme « standard » sont également basées sur la médiane et les quartiles : des valeurs supérieures au Q3 indiquent des grands gestes, tandis que celles inférieures au Q1 correspondent à des petits gestes. Ces descripteurs combinés de la vitesse et de la taille offrent ainsi une caractérisation complète de la dynamique gestuelle, intégrant à la fois les propriétés cinématiques et spatiales des mouvements.

# 5 GUI : *Gesture User Interface*

L'interface utilisateur présentée dans cette section constitue l'implémentation opérationnelle du *pipeline* d'extraction formelle défini dans les sections précédentes. Elle

fonctionne comme un véritable banc d'essai, permettant de mettre à l'épreuve la modélisation mathématique du mouvement. Il est à noter que cette interface, et de là les analyses qui en découlent, reposent néanmoins sur l'existence d'une segmentation préalable des gestes (qui peut être manuelle ou automatique).

La *Gesture User Interface* (GUI) correspond à une première version expérimentale, développée dans le cadre du *framework* élaboré au sein de cette thèse (et fonctionnelle, pour l'instant, que pour les fichiers de reBeat-14). Conçue à plus long terme pour répondre aux besoins des annotateurs de corpus de gestes, elle ne fait toutefois pas l'objet d'une diffusion ou d'une publication à ce stade, l'outil nécessitant encore des développements et des validations supplémentaires. Néanmoins, dans ce chapitre, nous présentons l'apport méthodologique et scientifique de l'interface. Bien qu'encore en cours de développement, elle constitue d'ores et déjà une base fonctionnelle et opérationnelle, ayant été utilisée pour certaines des analyses présentées dans cette thèse.

> Pour une description exhaustive de l'architecture logicielle, de l'implémentation des modules (widgets) et du guide d'utilisation de l'interface, le lecteur est invité à se référer à l'Annexe C8-S5-gui.

## 5.1 Rôle stratégique de l'interface dans la démarche interdisciplinaire

Dans la section 4, les descripteurs mobilisés relèvent d'un formalisme mathématique (matrices, normes de vecteurs, angles de rotation, calculs trigonométriques, etc.) indispensable au traitement automatique du mouvement. Ils permettent de modéliser avec précision la dynamique gestuelle, d'encoder les positions articulaires, les orientations dans l'espace tridimensionnel et les variations cinématiques au fil du temps. Toutefois, sous leur forme brute, ces variables numériques demeurent difficilement compréhensibles (ou très peu informatives) pour un annotateur humain. Une matrice de rotation, une norme euclidienne ou une valeur angulaire exprimée en radians ne rendent pas immédiatement compte, d'une élévation de la main, ou d'un changement directionnel significatif. Au-delà même de leur opacité, la projection abstraite de ces variables peut également s'avérer complexe à exploiter sans un support précis permettant leur manipulation. Sans dispositif médiateur, l'annotateur se retrouve face à des grandeurs numériques ou à des qualifications formelles isolées, dont il est difficile de saisir la portée sans le contexte global du geste auquel elles s'appliquent. Les descripteurs calculés décrivent des propriétés plus ou moins locales du geste (orientation, amplitude, variation), mais leur interprétation exige une vision intégrée de ce dernier. Autrement dit, le simple passage du calcul à la visualisation ne suffit pas : encore faut-il que cette visualisation soit structurée, interactive et intelligible.

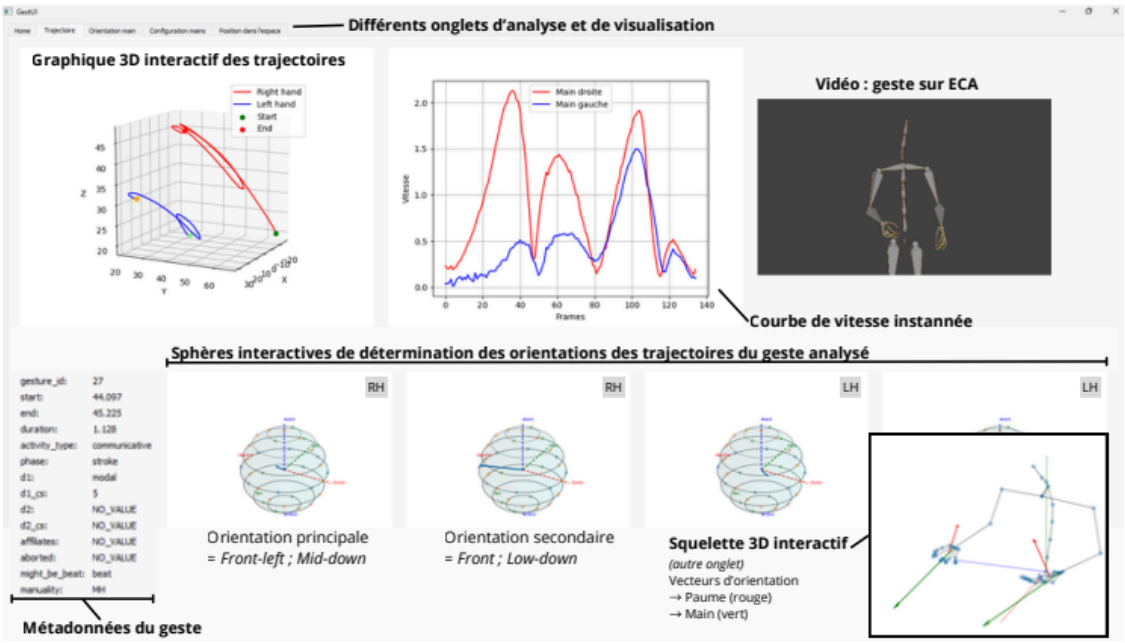



*Figure 49. Vue d'ensemble de l'interface et de certains modules qui la compose*

L'interface GUI constitue le point d'interaction principal entre l'utilisateur et le dispositif analytique conceptualisé dans ce chapitre, et ainsi intervient précisément comme un espace de médiation entre la formalisation mathématique, l'abstraction qui en découle et l'annotateur humain. Elle est composée de différents modules d'affichages et d'annotation automatiques (Figure 49), avec la vidéo du geste sur l'agent, des courbes de vitesse, d'accélérations, des graphiques 3D interactifs avec les marqueurs (donc le squelette) de la capture ou de la trajectoire, etc. Elle ne se limite pas à afficher des données : elle organise leur traduction dans un espace contextualisé, manipulable et interprétable. Les paramètres numériques deviennent des trajectoires visibles, des orientations observables, des variations dynamiques que l'expert peut explorer, ajuster et confronter à son jugement perceptif. En ce sens, la GUI joue ainsi un rôle d'interface interdisciplinaire.

## 5.2 Résolution du biais perceptif

L'annotation formelle manuelle réalisée dans un outil comme ELAN repose généralement sur l'observation d'une vidéo 2D. Or, cette représentation plane du mouvement introduit un biais perceptif structurel : elle dépend entièrement du point de vue de la caméra. L'annotateur interprète les trajectoires, les orientations et d'autres variations à partir d'une projection unique, figée, qui peut déformer la réalité spatiale du geste. Un mouvement dirigé vers l'avant (axe Z) peut ainsi sembler, selon l'angle de prise de vue, s'élever (axe Y) ou se déplacer latéralement (axe X). Ces distorsions ne relèvent pas d'une erreur de l'annotateur, mais d'une contrainte inhérente à la projection bidimensionnelle : la profondeur est écrasée, les axes sont confondus, et certaines composantes du mouvement deviennent difficilement discriminables.

Cette interface offre la possibilité de mieux appréhender les différents aspects formels des gestes en donnant accès à des représentations qui restent généralement peu accessibles lors d'enrichissements classiques de corpus. Là où l'annotation traditionnelle repose sur une

projection vidéo bidimensionnelle, l'interface rétablit l'accès à l'espace tridimensionnel dans lequel le geste évolue réellement. En restituant toutes les dimensions spatiales, elle permet d'observer de manière plus précise les trajectoires, les orientations et les configurations articulatoires. L'utilisateur peut ainsi examiner le geste (allant de la trajectoire jusqu'à la configuration d'un doigt) sous différents angles, soit par manipulation interactive de la scène (rotation autour du locuteur, changement de perspective), soit via de multiples plans de vue fixes prédéfinis (frontal, sagittal et transversal). Ces représentations permettent ainsi de s'affranchir de ce biais perceptif lors de l'annotation.

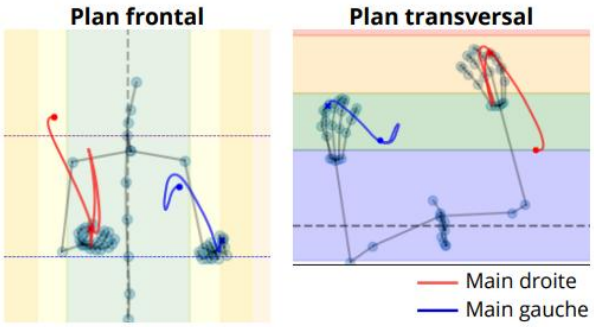



*Figure 50. Trajectoires des mains d'un geste selon les plans frontal et transversal*

Par exemple, dans le cas du geste illustré dans la Figure 50, la vue frontale (plan XY, avec en fond, les couleurs des différentes zones latérales discrétisées dans l'espace du locuteur et les traits en pointillés pour les zones en hauteur) correspond à la perspective classiquement mobilisée lors d'une annotation sur vidéo 2D. Depuis ce point de vue, l'annotateur observe, notamment pour la main droite (trajectoire rouge), un mouvement d'aller-retour vertical, interprétable comme un déplacement haut-bas. La dynamique semble principalement organisée selon l'axe Y. Cependant, la visualisation de cette même trajectoire sur le plan transversal (XZ, avec en fond les zones en termes de profondeur) modifie sensiblement la perception du mouvement. Il ne se déploie pas uniquement en hauteur, mais comporte également une composante significative selon l'axe Z, c'est-à-dire dans la profondeur. On observe en effet une avancée vers l'avant (avec un changement de zone) qui demeure invisible dans la vue frontale.

Dans le cas de la main gauche (trajectoire bleue), un biais d'interprétation peut également émerger si l'on se limite à la vue frontale (plan XY). Depuis cette perspective, la trajectoire semble se diriger vers la gauche du locuteur. Toutefois, en fonction de la posture du corps, cette déviation latérale pourrait aussi être perçue comme comportant une composante significative vers l'avant. La projection frontale ne permet pas de distinguer clairement la part respective des axes X et Z dans la trajectoire. L'annotateur peut alors surévaluer la composante de profondeur ou lui attribuer une valeur catégorielle qu'elle ne possède pas réellement. La visualisation sur le plan transversal (XZ) apporte ici un éclairage décisif. Elle montre que le mouvement comporte effectivement une légère avancée vers l'avant (axe Z), mais que cette composante reste modérée. Elle ne suffit pas à provoquer un changement de zone dans l'espace gestuel où le geste se déploie.

## 5.3 L'interface comme outil d'audit et de validation

L'interface, au-delà de sa fonction de support de visualisation, s'insère pleinement dans la boucle de développement et de validation de l'enrichissement automatique de corpus et des modèles. La contextualisation des différents aspects du geste qu'elle offre, au travers des différents modules intégrés, permet de la considérer également comme un véritable outil de diagnostic des données, des analyses computationnelles et, par conséquent, des caractérisations qui en découlent.

Les calculs automatiques reposent en effet sur les données brutes issues de la capture de mouvement. Or, ces dernières sont susceptibles de contenir du bruit, par exemple en raison de pertes temporaires de capteurs, ou des discontinuités dans la reconstruction du mouvement par le modèle biomécanique lors du post-traitement de la capture. De tels artefacts peuvent entraîner des répercussions directes sur les descripteurs calculés, avec l'apparition de valeurs aberrantes (ex. : pics de vitesse extrêmement élevés, variations angulaires, incohérences ou ruptures artificielles de la trajectoire sur quelques frames isolées, etc.). L'affichage synchronisé des courbes temporelles (vitesse, accélération), des marqueurs et du squelette 3D permet alors à l'annotateur ou à l'analyste d'inspecter précisément ces anomalies. Cette mise en correspondance directe entre signal numérique et représentation spatiale rend possible l'identification de la source du problème, et donc de déterminer s'il s'agit d'un phénomène moteur réel ou d'un artéfact technique lié à la capture.

Dans un contexte où ces données, ainsi que les descripteurs qui en sont dérivés, sont susceptibles d'être injectés dans un réseau de neurones, l'interface s'inscrit dans une logique de contrôle qualité en amont de l'analyse multimodale et de l'apprentissage automatique. Elle contribue à éviter que des artefacts techniques ne soient interprétés comme des régularités motrices par le modèle. L'outil participe ainsi pleinement à une boucle itérative de développement : calcul des descripteurs, analyse et inspection visuelle, annotation multimodale automatique, ajustement éventuel des seuils ou des modules de calcul, puis réévaluation. L'interface devient alors un espace de vérification méthodologique continue, assurant la cohérence entre données brutes, formalisation computationnelle et modélisation.

## 5.4 Perspectives : vers un écosystème interopérable pour l'annotation multimodale

Si la version actuelle de l'interface démontre la faisabilité d'une extraction et d'une visualisation objectivée des caractéristiques formelles du geste, son évolution s'inscrit dans une réflexion méthodologique plus large, articulée autour de trois défis structurants pour la communauté des études gestuelles et des sciences informatiques.

### 5.4.1 L'interopérabilité et la standardisation des annotations

A l'heure actuelle, la fragmentation des formats et des pratiques d'annotations, souvent spécifiques à un laboratoire, voire à un chercheur, limite fortement la comparaison de corpus à

grande échelle. Cette hétérogénéité freine non seulement la mutualisation des données, mais aussi la reproductibilité des analyses. Cette problématique se reflète également dans l'état actuel de l'interface développée, qui demeure pour l'instant adaptée uniquement au schéma d'annotation utilisé dans le cadre de l'enrichissement de reBeat-14, le module de chargement des données d'annotations n'étant pas encore modulable.

L'un des enjeux futurs de cet outil réside dans sa capacité à établir un pont entre les logiciels d'annotation manuelle standards, tels que ELAN (par exemple avec l'intégration d'un *plugin* GUI au sein de l'interface) et par extension les schémas d'annotation qui en sont issus, et les pipelines d'analyse et de traitement automatique intégrés à l'interface. Une telle articulation implique le développement de mécanismes de normalisation capables de prendre en charge des ontologies d'annotation hétérogènes. L'objectif est de permettre des cycles d'import et d'export de données robustes de données, garantissant la compatibilité inter-systèmes sans perte de granularité analytique.

### 5.4.2 Agnosticisme matériel

Le *framework* proposé dans ce chapitre a été développé à partir de données issues de systèmes de capture de mouvement, plus spécifiquement de fichiers au format `.bvh`, structurés selon le protocole de capture mis en place par les auteurs du corpus BEAT. Cependant, la structure interne de ces fichiers peut varier d'un corpus à l'autre, que ce soit par le nombre de capteurs utilisés ou par leur dénomination. En l'état, le *framework* ne peut donc être appliqué qu'à des fichiers conformes au même mapping que celui utilisé dans BEAT. Par ailleurs, ce format n'est pas la seule manière de représenter des données de mouvement. La recherche tend vers l'analyse de corpus écologiques, dit « in-the-wild », souvent traités via des algorithmes d'estimation de pose 2D ou 3D, tels que MediaPipe, qui reconstruit une cinématique à partir d'images vidéo.

Pour appliquer les traitements automatiques à ces différentes sources, il devient nécessaire de viser un agnosticisme au matériel d'acquisition. Cette transition implique plusieurs ajustements méthodologiques. Il est tout d'abord essentiel de s'assurer que les données nécessaires aux calculs automatiques sont disponibles, par exemple les positions 3D de la paume, indispensables pour déterminer son orientation. Selon le format initial, il peut également être nécessaire de tenir compte du bruit et de l'incertitude propres aux méthodes d'estimation de pose. L'agnosticisme matériel s'impose comme une condition nécessaire pour étendre le *framework* à des corpus hétérogènes et variés, préparant ainsi le terrain pour des analyses automatisées à grande échelle et des études comparatives inter-corpus.

### 5.4.3 Intégration de modèles prédictifs continus

Enfin, l'interface a pour vocation d'étendre ses fonctionnalités de traitement automatique. Pour la segmentation semi-automatique du geste, l'objectif à terme est d'intégrer des architectures d'apprentissage profond, telles que le modèle de segmentation COSMOS, développé dans cette thèse, afin d'enrichir les types de traitements automatiques directement

accessibles via l'interface. Cette intégration permettra de générer semi-automatiquement des pré-annotations, qui pourront ensuite être affinées par un annotateur expert, tout en facilitant l'analyse des segments produits au sein des modules d'annotation automatique formelle. Elle offrira également l'opportunité de mesurer empiriquement la robustesse des segmentations sur des ensembles de données plus volumineux. Ainsi, en combinant segmentation semi-automatique, analyse formelle automatique et visualisations variées, l'interface s'inscrit pleinement dans une boucle *human-in-the-loop*, fournissant à l'expert des supports de décision pour valider et ajuster les segmentations générées.

# 6 Conclusion de chapitre

Ce chapitre propose une méthodologie structurée pour l'annotation automatique des caractéristiques formelles du geste co-verbal, en établissant un cadre explicite, reproductible et interopérable destiné à l'enrichissement des corpus multimodaux. Les différentes dimensions descriptives du geste ont été formalisées et associées à des descripteurs calculables automatiquement à partir de critères géométriques et cinématiques extraits des données de mouvement, et suffisamment fin pour caractériser un geste dans toutes ses variantes. Cette approche permet de s'affranchir de l'annotation manuelle pour la majorité des descripteurs, tout en préservant l'information qualitative et en exploitant pleinement les caractéristiques intrinsèques du mouvement. L'organisation hiérarchique adoptée, allant des paramètres spatiaux globaux à la configuration digitale fine, a permis de structurer l'analyse selon une logique progressive et multi-échelle. Cette structuration rend le cadre modulaire et adaptable à différents objectifs analytiques, qu'il s'agisse de descriptions globales du mouvement ou d'analyses fines de la morphologie manuelle.

Le cadre méthodologique présenté a été élaboré et évalué à partir des données d'un locuteur unique, Wayne. Ce choix méthodologique a permis de développer et de stabiliser les descripteurs ainsi que les seuils de segmentation dans un environnement contrôlé, garantissant la cohérence interne du dispositif. Actuellement, la discrétisation des paramètres continus, tels que la définition des zones spatiales à 20 cm ou les quantiles de vitesse, repose sur des seuils heuristiques dérivés de la morphologie et de la distribution spécifique de ce locuteur. Bien que cette approche déterministe assure une transparence totale de l'extraction formelle, elle limite la généralisation directe à d'autres locuteurs. Il est en effet possible que certains paramètres reflètent en partie des caractéristiques propres au style gestuel de ce locuteur, notamment en ce qui concerne la dynamique, ou l'amplitude des gestes. Toutefois, les seuils définis ont été conçus comme des paramètres adaptatifs, ajustables en fonction des propriétés des données analysées. Bien que leur généralisation ne soit pas encore validée, l'étude de la variation inter-locuteur permettra d'identifier les seuils réellement distinctifs, qui se révèleraient donc potentiellement sémiotiquement significatifs chez certains locuteurs. Dans ce contexte, il reste essentiel de tester cette méthodologie sur des ensembles de données incluant plusieurs participants aux caractéristiques physiques, gestuelles et expressives variées, afin de déterminer si les seuils définis, et par conséquent la méthodologie choisie, sont suffisamment robustes ou s'ils doivent être ajustés. Cela permettra de déterminer si ces seuils sont réellement pertinents

sur le plan communicatif ou simplement idiosyncratiques. Une évolution naturelle de ce *framework* consisterait à remplacer ces seuils statiques par des méthodes de partitionnement non supervisé, telles que le clustering *K-Means* ou les Modèles de Mélange Gaussien (GMM), appliquées sur les données de calibration de chaque nouveau locuteur. Cette démarche permettrait de définir des seuils de catégorisation dynamiques, adaptés à la morphologie individuelle. L'objectif à terme est de disposer d'un cadre généralisable et adaptable, capable de fournir des annotations formelles fiables pour des corpus multimodaux étendus et diversifiés, tout en conservant la précision nécessaire pour alimenter des analyses fines de la coordination geste–parole ou l'alimentation de modèles.

Bien qu'il soit désormais possible d'automatiser une grande partie de la description formelle des gestes à partir des données de capture de mouvement, certaines dimensions restent difficiles à traiter. En particulier, la manualité demeure complexe à formaliser automatiquement, car la distinction entre geste et mouvements non communicatifs d'une main non impliquée dans le geste est trop subtile pour se passer de l'interprétation humaine. De même, la forme des trajectoires n'est pas encore entièrement automatisable. L'annotation manuelle, qui pourrait servir de référence pour l'entraînement supervisé d'outils automatiques, est elle-même sujette à un biais perceptif. Pour labelliser la forme des gestes, l'annotateur se base sur une vidéo 2D du locuteur (plan intermédiaire entre face et profil, selon les mouvements capturés), ce qui peut amener à percevoir certaines trajectoires comme linéaires ou légèrement courbes, alors que leur structure spatiale réelle peut être plus complexe. S'appuyer sur ces annotations lors d'un apprentissage automatique supervisé pourrait ainsi introduire un biais d'annotation dans les données d'entraînement et compromettre la fiabilité de la catégorisation automatique des trajectoires gestuelles. En tant qu'alternative, plusieurs tentatives ont été menées pour discriminer les trajectoires gestuelles à l'aide de méthodes d'analyse statistique et géométrique, telles que l'analyse en composantes principales et le calcul de points d'inflexion sur les trajectoires, ainsi que différentes techniques de lissage du signal. Ces approches, explorées dans le cadre de cette thèse mais non détaillées dans ce manuscrit, n'ont toutefois pas encore fourni de résultats suffisamment fiables. Cette thèse défend donc pour l'instant une approche hybride pour l'identification des formes de trajectoires, combinant la perception visuelle du mouvement (telle qu'elle serait interprétée par un interlocuteur) et l'analyse géométrique et cinématique objective du mouvement. L'objectif futur serait de mener une étude à plus grande échelle sur la perception des trajectoires, en confrontant leurs caractéristiques géométriques (nombre d'inflexions, degré de courbure, analyses angulaires, etc.) à leur interprétation perceptive. Ce type d'étude permettrait de confirmer l'intérêt d'appréhender la forme d'une trajectoire dans sa réalité tridimensionnelle, plutôt que de se limiter à l'axe de perception de l'interlocuteur (ou de l'annotateur).

L'aspect dynamique de cette méthodologie, par exemple pour la caractérisation de l'évolution de la configuration de la main au cours du geste, reste également à développer. Actuellement, les méthodologies présentées concernant cet aspect portent majoritairement sur la dernière *frame* du geste. Ce choix a été fait volontairement, afin de privilégier dans un premier temps la conceptualisation d'une méthodologie de description à partir de données statiques,

considérée comme une première étape plus accessible. Cependant, ce choix comporte des biais évidents, puisque le geste est alors réduit à sa configuration finale, alors qu'il peut évoluer de manière significative au cours du *stroke*. L'objectif futur est de progresser vers une description de plus en plus dynamique, tout en trouvant un compromis entre une annotation très fine, au niveau de chaque *frame*, et une analyse à un niveau plus global.

Malgré ces limites, l'intégration de ces modules au sein de l'interface interactive GUI constitue une avancée majeure offrant un outil d'analyse aux fonctionnalités inédites en termes de représentations et de traitements. Les descripteurs, extraits automatiquement via un ensemble de scripts et intégrés dans l'interface, permettent de caractériser chaque geste annoté selon des paramètres mesurés et clairement définis, facilitant ainsi la comparaison et la réutilisation des annotations. Cet outil constitue également un soutien précieux pour les chercheurs, en leur permettant d'examiner l'existence de corrélations récurrentes entre certaines configurations gestuelles et les significations communicatives qu'elles véhiculent. De plus, cette méthodologie d'annotation automatique comble une lacune importante des annotations manuelles, qui étaient jusqu'alors difficiles à standardiser, comparer et exploiter de manière rigoureuse, notamment dans le cadre de l'enrichissement formel de reBeat-14. Ces avancées offrent une caractérisation précise des variantes gestuelles formellement distinctive, chacune pouvant véhiculer une valeur communicative spécifique. À terme, la publication et la mise à disposition de cet outil et de ces méthodologies permettront leur appropriation par la communauté scientifique et favoriseront le développement d'analyses comparables et cumulatives sur de larges corpus multimodaux.

Au-delà des descripteurs déjà intégrés, l'architecture de l'interface GUI a été conçue pour accueillir de nouvelles analyses et offrir de nombreuses perspectives pour l'étude des gestes. Les travaux futurs sur les données du corpus pourront ainsi être directement intégrés dans l'interface, avec pour objectif de développer progressivement des modules supplémentaires, pertinents pour l'annotation, la caractérisation et l'exploitation des gestes. Cette modularité permet également d'envisager l'ajout de fonctionnalités innovantes issues d'échanges interdisciplinaires, favorisant un dialogue entre sciences computationnelles, sciences du langage, sciences du mouvement et toute autre discipline étudiant le geste. L'interface est ainsi pensée comme un outil évolutif, capable de centraliser les méthodologies les plus pertinentes tout en restant ouverte à l'extension et à l'adaptation selon les besoins des chercheurs.

Pour conclure, ce chapitre représente une première tentative structurée de convergence entre perspectives formelles, computationnelles et analyse multimodale des gestes. L'interface, ainsi que la méthodologie qu'elle intègre, constitue un outil précieux pour l'enrichissement de corpus à l'aide d'outils fiables, pour l'approfondissement des connaissances du geste et pour le développement de modèles computationnels robustes, capables de refléter avec fidélité la complexité de la communication humaine. L'approche fondée sur les paradigmes computationnels, et notamment sur le traitement de données et du signal, permet ainsi de produire des données exploitables, en particulier pour l'annotation multimodale des gestes, offrant une base fiable et reproductible pour des analyses et traitements futures.

# Chapitre 9
# Segmentation semi-automatique du geste sous contrainte de données d'entrainement limitées

## 1 Introduction

L'élaboration de modèles de reconnaissance ou de segmentation de gestes s'inscrit dans le champ de l'approche computationnelle appliquée à l'étude de la gestualité, visant à formaliser, analyser et prédire des « comportements gestuels » à partir de données variées, telles que des images RGB, des fichiers de profondeur issus de la Kinect, ou d'autres types de mesures. Plusieurs démarches en ce sens ont été référencées, notamment dans le *Chalearn challenge* [170, 171, 172, 173] pour la reconnaissance des mouvements de la main (par exemple en interaction humain-machine ou pour la reconnaissance de signes). Ces méthodes incluent l'utilisation de perceptrons multicouches [174], de machine à vecteurs de support [175, 176, 177], de modèles de Markov [178, 179, 77], ou encore le *Dynamic Time Warping* [180], parmi d'autres. Plus récemment, avec l'essor du *deep learning*, les réseaux convolutifs (CNN) [181, 82], les réseaux récurrents (RNN, LSTM) [142] ou encore les réseaux convolutifs graphiques spatio-temporels (ST-GCN) [182, 112, 183, 184], ont montré de bonnes performances pour la reconnaissance d'actions et la détection de mouvements signifiants. Cependant, la plupart de ces modèles ne sont pas spécifiquement développés pour étudier les gestes co-verbaux. Ils sont conçus pour reconnaître une large variété de mouvements qualifiés de « gestes » (par exemple, détecter une action humaine lors d'une interaction humain-machine), sans se focaliser sur ceux qui accompagnent l'élaboration du discours, tels que considérés dans le cadre de cette thèse. Le modèle proposé par Ferstl, Neff et McDonnell [142] constitue à ce jour la seule approche spécifiquement destinée à la détection de phases gestuelles telles que définies par Kendon [7, 8]. Toutefois, lorsqu'il a été testé dans le cadre de cette recherche, notamment pour l'enrichissement semi-automatique de corpus de gestes, le modèle s'est montré peu fiable pour la segmentation des *strokes*, y compris sur le corpus sur lequel il avait été initialement entraîné et dont les performances avaient été présentées comme bonnes. De plus, il n'a pas été possible de

reproduire l'exécution de ce modèle. Ainsi, bien qu'il existe des modèles visant la segmentation des gestes, aucun ne permet actuellement d'obtenir des résultats fiables et reproductibles. La segmentation des gestes reste dans la majorité des cas effectuée manuellement, comme nous l'avons fait dans reBeat-14. Bien que ces annotations soient le *gold standard* pour identifier précisément les gestes au sein du flux de mouvements accompagnant la parole, leur production reste chronophage et requiert un savoir-faire expert. Les méthodes semi-automatiques combinant traitement automatisé et intervention humaine sont alors une alternative envisagée à l'annotation purement manuelle.

Ce chapitre constitue une réflexion sur la faisabilité d'un modèle de segmentation automatique développé dans un contexte où les données d'entraînement sont limitées. Reposant sur le postulat selon lequel il n'est pas nécessaire de disposer d'un grand nombre d'annotations par locuteur pour obtenir un modèle performant, l'approche présentée ici se fonde sur le corpus reBeat-14, et plus particulièrement sur les segmentations issues de l'annotation experte et sur les données de mouvement associées. Il s'agit de s'affranchir autant que possible de la tâche chronophage d'annotation manuelle, tout en conservant l'apport de l'expertise humaine. Ce modèle de segmentation, guidé par l'expertise des annotateurs lors de son apprentissage, peut ensuite être appliqué à des données non segmentées du même locuteur, générant des prédictions que l'annotateur pourra vérifier et corriger.

L'objectif de ce chapitre n'est pas de présenter un outil prêt à l'usage, mais de proposer une méthodologie combinant modélisation computationnelle et données qualitatives issues de l'expertise en sciences du langage. Cette approche vise à permettre une segmentation semi-automatique fiable des gestes à partir d'un modèle supervisé entraîné sur la seule base d'un petit nombre de données expertes. Le chapitre s'organise en quatre volets, avec (i) l'*overview* du modèle de segmentation développé COSMOS (*CO-speech gesture Segmentation on MOtion Sequences*), (ii) les performances du modèle *baseline*, (iii) les analyses de l'impact des hyperparamètres du modèle, et (iv) les analyses de l'impact des données sur les performances du modèle *baseline*.

Le chapitre conclut sur la capacité d'un réseau de neurones à segmenter un flux de mouvement dans un cadre d'apprentissage supervisé, à apprendre des représentations gestuelles fiables et, par conséquent, à proposer des segmentations cohérentes, y compris dans un contexte où les données supervisées sont limitées.

# 2 Overview de COSMOS

## 2.1 Données utilisées

Les données constituant les différents ensembles (entraînement, validation et test) utilisés dans ce chapitre proviennent du corpus reBeat-14, présenté et analysé dans les parties II et III de ce manuscrit. Deux types de fichiers ont été exploités : les données de mouvement (fichiers originaux de BEAT `.bvh` : articulations du haut du corps), à partir desquelles les positions 3D dans l'espace ont été extraites à chaque *frame*, et les données de segmentation (`.csv` contenant

les gestes segmentés et leurs timestamps) issues de l'enrichissement de reBeat-14. Afin d'assurer une synchronisation temporelle entre ces deux sources, les annotations de segmentation ont été converties à l'échelle de la *frame* en fonction des timestamps et du *frame rate*. Ainsi, chaque fichier de mouvement a été associé à un fichier d'annotation de taille identique, avec pour chaque *frame* deux labels possibles : la classe 0 pour l'absence et la classe 1 pour la présence d'un geste. Ces labels binaires, dérivés de l'annotation experte, ont fourni une vérité terrain (données de référence considérée comme correcte et fiable) temporellement précise pour l'entraînement supervisé, sans toutefois encoder les types fonctionnels ni les formes de gestes.

Ces fichiers ont ensuite été découpés en séquences de 10 secondes (1200 *frames*). Ce choix de fenêtre temporelle a permis au modèle d'analyser plusieurs gestes au sein d'une même séquence pendant l'apprentissage. L'analyse d'un fichier complet aurait été trop volumineuse en termes de mémoire, et des séquences trop courtes auraient été susceptibles de contenir que très peu de gestes (voir aucun) dans la majorité des cas. Les séquences ont été découpées sans chevauchement afin de garantir l'indépendance des séquences utilisées lors de l'entraînement et de l'évaluation.

À l'issue de cette découpe, 71 paires de fichiers (position 3D relatives à la hanche du locuteur et label 0/1 pour chaque *frame*) ont été réparties aléatoirement en dans différents ensembles de données : l'ensemble d'entraînement (données utilisées pour l'apprentissage), l'ensemble de validation (données servant à guider l'apprentissage et au réglage des hyperparamètres) et l'ensemble de test (données utilisées uniquement après l'entraînement pour l'évaluation), selon un ratio 70/10/20. Une normalisation Min–Max a été appliquée aux séquences de l'ensemble d'entraînement afin de projeter chaque dimension dans un intervalle borné. Cette normalisation permet de préserver la dynamique relative des trajectoires articulaires, tout en facilitant la stabilité de l'optimisation et la convergence du réseau. Les paramètres de normalisation ont ensuite été appliqués aux ensembles de validation et de test.

| | **Données originales** | **Séquences après découpe de fichiers** | | |
|---|---|---|---|---|
| | | *Entraînement* | *Validation* | *Test* |
| **N fichier** | 14 | 50 × 10s | 7 × 10s | 15 × 10s |
| **N *frames*** | 93660<br>0 = 69026 ; 1 = 24934 | 60000<br>0 = 43486 ; 1 = 16514 | 8400<br>0 = 5930 ; 1 = 2470 | 18000<br>0 = 13773 ; 1 = 4227 |
| **Durée totale** | 13 min 03 s | 8 min 20 s | 1 min 10 s | 2 min 30 s |
| **N art.** | 75 | 63 | 63 | 63 |

*Tableau 43. Ensembles de données*
*(N fichiers = nombre de fichiers ; N frames = nombre de frames total ; N art. = nombre d'articulateurs)*

Au total, l'ensemble des séquences est composé de 86 500 *frames* réparties dans les différents sets de données (Tableau 43), avec 50 séquences pour l'entraînement (8 min 20 s, 72,5 % avec absence de geste), 7 séquences pour la validation (1 min 10 s, 70,6 % de *frames* sans geste) et 15 séquences (2 min 30 s, 76,4 % de *frames* sans geste). On peut remarquer un déséquilibre de classe, avec une surreprésentation de la classe 0 (sans gestes) dans les différents sets de données. Cette répartition disproportionnée, calculée d'après les annotations expertes, reflète toutefois la réalité du flux de mouvement, dans lequel les gestes sont ponctuels et

n'occupent qu'une partie limitée du signal. Ainsi, l'évaluation doit donc se concentrer sur des métriques sensibles à la classe minoritaire (rappel, précision et F1-score présentés plus loin). En effet, l'*accuracy* peut être artificiellement élevée du fait que la majorité des *frames* ne contiennent pas de geste. Enfin, chaque séquence de mouvement a une taille de [1200, 63 × 3] : 1200 *frames*, avec 63 articulateurs (du buste et des membres supérieurs) par *frame* et trois coordonnées (X, Y, Z) pour chaque articulateur. Pour les fichiers d'annotation, la taille est [1200, 1], correspondant à un label (0 ou 1) par *frame*.

## 2.2 Architecture du modèle

L'architecture du modèle, présentée dans les sections suivantes, repose sur une combinaison de CNN 1D et de GRU bidirectionnels. Bien qu'il s'agisse d'une architecture relativement classique, ce choix s'impose de manière stratégique au regard des contraintes de notre corpus (environ 13 minutes de données). Dans ce régime de *low-data*, l'architecture agit comme un régularisateur structurel naturel : l'encodeur CNN contraint le modèle à extraire des motifs locaux pertinents, tandis que le décodeur GRU bidirectionnel contrôle la modélisation des dépendances temporelles en limitant la complexité de l'historique appris. Ce compromis permet de réduire le risque de surapprentissage massif qu'aurait pu induire un modèle fondé exclusivement sur un mécanisme d'attention globale de type Transformer.

### 2.2.1 Encodeur de mouvement

#### Rôle du décodeur et pertinence des CNN

L'encodeur a pour rôle de transformer les données de mouvement en une représentation latente compacte et informative. Cette transformation vise à extraire automatiquement les *patterns* dynamiques pertinents du mouvement analysé. L'encodeur agit ainsi comme un extracteur de caractéristiques temporelles (*features*), produisant pour chaque *frame* un vecteur de représentation enrichi qui sert ensuite d'entrée au décodeur chargé de prédire la présence ou l'absence de geste.

Les réseaux convolutifs (CNN 1D) présentent plusieurs propriétés favorables à la tâche de segmentation des gestes. Les convolutions permettent de capturer efficacement des motifs locaux du mouvement, tout en utilisant les mêmes filtres, noyaux ou *kernels* qui « voient » et qui analysent chaque sous-segment, sur l'ensemble de la séquence. Cette mutualisation des filtres réduit considérablement le nombre de paramètres à apprendre, ce qui favorise une meilleure généralisation, en particulier lorsque peu de données annotées sont disponibles. De plus, les CNN 1D exploitent directement la structure temporelle des séquences de mouvement, en appliquant des filtres le long de l'axe temporel. Cela permet au modèle de capturer les dépendances locales entre *frames* successives, essentielles pour identifier le début et la fin d'un geste, tout en conservant l'information spatiale et cinématique des articulations à chaque instant. La combinaison de la nature temporelle des convolutions et du partage des filtres rend les CNN particulièrement adaptées à l'analyse des gestes co-verbaux, qui sont à la fois brefs et localisés.

### Pile de blocs résiduels pour l'encodage du mouvement

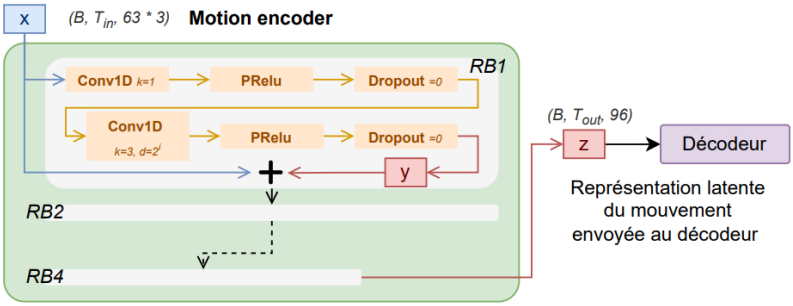



*Figure 51. Schéma de l'encodeur de mouvement*

L'encodeur (Figure 51) s'organise en 4 blocs résiduels (RB), chacun étant composé de deux couches de CNN 1D. Ces convolutions constituent le cœur du traitement de données, tandis que les connexions résiduelles assurent la transmission directe de l'information extraite entre l'entrée et la sortie du bloc.

| | Taille d'entrée | Taille de sortie |
|---|---|---|
| **RB 1** | 189 = 63 × 3 | 128 |
| **RB 2** | = sortie RB1 | 256 |
| **RB 3** | = sortie RB2 | 128 |
| **RB 4** | = sortie RB3 | 96 |

*Tableau 44. Dimensions d'entrée et de sortie des vecteurs par bloc résiduel (RB)*

Dans cette architecture, l'axe temporel correspondant aux *frames* est conservé tout au long du réseau, tandis que les *features* associées à chaque *frame* s'enrichissent progressivement à mesure que l'information circule dans les convolutions. L'entrée initiale est un tenseur de forme (B, $T_{in}$ = 1200, Fe = 63 × 3), où B est la taille du batch (sous-ensemble de données utilisé pour une seule mise à jour des poids du modèle), T le nombre de *frames* par séquence, et Fe les *features* associées aux positions 3D par articulateur où chaque *frame* représente un instant du mouvement (donc potentiellement d'un geste). Le vecteur de sortie de chaque bloc devient ensuite l'entrée du bloc suivant (Tableau 44). Ainsi, RB2 reçoit la représentation enrichie de dimensions (B, $T_{out}$, 128) produite par RB1, RB3 celle produite par RB2, et ainsi de suite jusqu'au dernier bloc, permettant au modèle de construire progressivement des représentations de plus en plus complexes. Ces connexions permettent au modèle de conserver l'information issue des couches précédentes tout en apprenant des représentations plus complexes, ce qui est particulièrement pertinent pour l'analyse de mouvements continus où certaines caractéristiques doivent être préservées sur plusieurs échelles temporelles.

### Première couche d'un bloc résiduel

Le premier CNN 1D du bloc résiduel transforme les *features* d'entrée en appliquant un filtre de taille 1 sur chaque *frame* (noyau ou *kernel* k=1). Cela signifie que chaque *frame* est traitée indépendamment, et que la couche se contente de projeter les données du vecteur d'entrée vers une représentation latente intermédiaire de dimension de sortie souhaitée pour le

bloc courant. Cette étape est suivie d'une fonction d'activation PReLU (*Parametric ReLU*), qui introduit de la non-linéarité dans le réseau. Cette non-linéarité permet au modèle de représenter des relations complexes entre les *features* plutôt que de se limiter à une simple transformation linéaire, incapable de capturer les motifs caractéristiques des données de mouvement. Ensuite, un *dropout* est appliqué ; dans notre cas, le *dropout* est fixé à 0, ce qui signifie qu'aucune unité neuronale n'est ignorée pendant l'entraînement.

**Seconde couche d'un bloc résiduel**

Le second CNN 1D prend en entrée la représentation latente produite par le premier CNN 1D, pour enrichir cette représentation. La seconde couche du bloc résiduel utilise un noyau de taille 3 (*kernel* k=3), ce qui permet de combiner l'information de la *frame* courante avec celle des *frames* voisines et de détecter ainsi des motifs temporels locaux, comme le début ou la fin d'un geste. En termes de représentation, cette couche transforme la séquence de *features* « *frame* par *frame* » en une séquence de vecteurs enrichis, qui capturent à la fois les caractéristiques de chaque *frame* et le contexte temporel immédiat. Cette représentation locale est plus informative que celle produite par la première convolution. Une dilatation est effectuée dans cette couche, avec un facteur de dilatation $d = 2^i$, où $i$ correspond au numéro du bloc résiduel dans la pile. La dilatation augmente progressivement le champ réceptif, c'est-à-dire l'étendue temporelle de la séquence que le filtre peut « voir » pour produire sa sortie, permettant au modèle de capturer à la fois des dépendances locales et des motifs à plus longue portée au fur et à mesure des itérations. Un *padding* est également configuré, ce qui signifie qu'aucune valeur artificielle n'est ajoutée aux extrémités de la séquence. Pour maintenir l'alignement temporel, un tronquage symétrique ($= \frac{(k-1)\cdot d}{2}$) est appliqué en supprimant le même nombre de *frames* au début et à la fin de la sortie afin que chaque vecteur corresponde exactement à la portion de séquence « vue » par le filtre. Enfin, une fonction d'activation PReLU est appliquée, suivie d'un *dropout* toujours fixé à 0.

**Champ réceptif total de l'encodeur**

Dans notre architecture, la première convolution ne modifie pas le champ réceptif : chaque *frame* a ses propres *features*, qui sont projetées dans un espace latent intermédiaire. La seconde convolution, en revanche, combine les *features* de la *frame* courante avec celles des *frames* voisines définies par le champ réceptif du bloc, qui est donné par $(k-1)\cdot d + 1 = 2\cdot 2^i + 1$. Ainsi, pour les quatre blocs empilés, les champs réceptifs des blocs successifs sont respectivement de 5, 9, 17 et 33 *frames*. Lorsqu'on les empile, le champ réceptif total de la séquence est obtenu en accumulant ces contributions, en soustrayant les *frames* centrales déjà comptées dans les blocs précédents, ce qui donne un champ réceptif global de 61 *frames*. Concrètement, cela signifie que, pour chaque *frame* de sortie du réseau, ses *features* finales intègrent l'information temporelle provenant d'un intervalle d'environ 61 *frames* de la séquence originale, soit approximativement 0,51 seconde à 120 fps. Cette approche permet à chaque *frame* de conserver sa position temporelle tout en enrichissant ses *features* avec un contexte local et environnant suffisant pour capturer des motifs de gestes, dont la durée est généralement inférieure à 500 ms.

### 2.2.2 Décodeur de mouvement

#### Rôle du décodeur et pertinence des GRU

Le rôle du décodeur est d'analyser les représentations latentes des séquences de mouvement produites par l'encoder et de les transformer en prédictions temporelles *frame* par *frame*, c'est-à-dire estimer la probabilité qu'un geste soit présent à chaque instant. Ainsi, l'entrée du décodeur correspond aux représentations vectorielles extraites par l'encodeur pour chaque *frame* de taille (B, $T_{out}$, 96). Chaque vecteur d'entrée représente les informations pertinentes sur la posture et le mouvement à cet instant, sans contenir encore de contexte temporel global sur la séquence.

Les GRU (*Gated Recurrent Units*) sont particulièrement efficaces pour capturer les dépendances temporelles dans les séquences. À chaque *frame*, elles filtrent les informations à retenir ou à oublier, ce qui autorise le modèle à conserver le contexte récent tout en limitant le nombre de paramètres et le coût en calcul. De plus, l'utilisation d'un GRU bidirectionnel (GRUb) permet de lire la séquence dans les deux sens, du passé vers le futur et du futur vers le passé pour chaque *frame* (analyse de la séquence vers l'avant et vers l'arrière). Cela offre au modèle la capacité de prendre en compte le contexte avant et après chaque *frame* analysée, ce qui est crucial pour détecter correctement les gestes, qui peuvent être définis non seulement par le mouvement précédent, mais aussi par celui qui suit.

#### Structure du décodeur

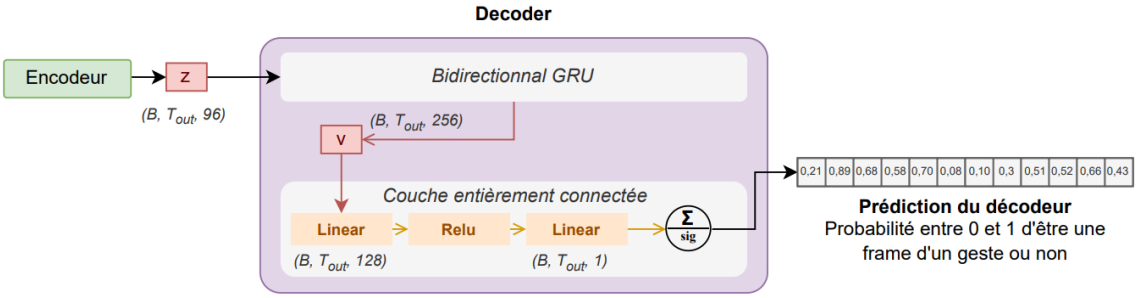



*Figure 52. Schéma de l'architecture du décodeur*

Le décodeur du modèle (Figure 52) est composé d'un GRU bidirectionnel avec une taille de couche cachée de 128, et génère deux états cachés (analyse vers l'avant et vers l'arrière), chacun étant un vecteur de H dimensions (=128) qui compacte l'information pertinente de la séquence. La sortie du GRU a une dimension (B, T, 2H), où B est la taille du batch, T le nombre de *frames* par séquence, et où 2H (=128×2) correspond à la concaténation des états cachés.

La sortie du GRU est ensuite transmise à un réseau de neurones à couches multiples (MLP), dont le rôle est de transformer la représentation de la séquence en prédictions scalaires pour chaque *frame*. Le MLP applique une transformation linéaire pour réduire la dimension des états cachés concaténés du GRU de 2H à H, en utilisant des poids et des biais optimisés lors de l'entraînement. Cette transformation permet au modèle de condenser les informations

pertinentes capturées par le GRU dans un vecteur plus compact. Il ne s'agit pas d'une simple réduction de dimension : ce vecteur résume de manière significative l'information temporelle et spatiale extraite par le GRU, afin que le MLP puisse ensuite produire des prédictions fiables. Une fonction d'activation ReLU est appliquée sur le vecteur de sortie de la couche linéaire (similaire à la fonction PReLU). Enfin, le vecteur H résultant est projeté vers une sortie scalaire par *frame* par une dernière couche linéaire, qui est ensuite convertie en probabilité de présence de geste grâce à une fonction sigmoïde (Sig). Cette fonction borne les valeurs entre 0 et 1, permettant d'interpréter directement la sortie comme une probabilité.

### 2.2.3 Classification binaire et export des annotations automatiques

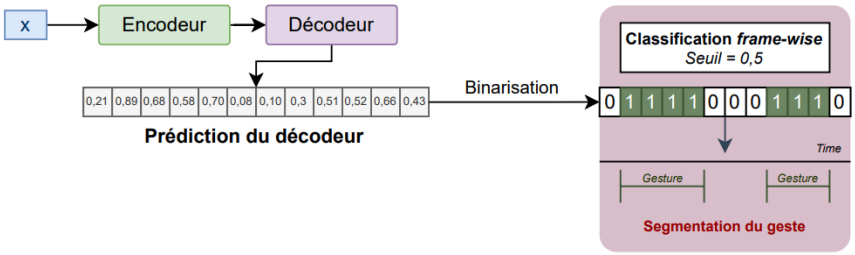



*Figure 53. Schéma de binarisation et détermination des gestes prédits*

Les étiquettes binaires sont assignées à l'aide d'un seuil fixé à 0,5 (Figure 53) : les *frames* dont la probabilité dépasse 0,5 sont considérées comme contenant un geste (label 1), tandis que celles en dessous de ce seuil sont considérées comme n'en contenant pas (label 0). Pour convertir les prédictions binaires par *frame* issues du décodeur en annotations gestuelles exploitables, une chaîne de traitement automatisée a été mise en place, permettant de générer des fichiers ELAN (`.eaf`). La fonction principale de ce *pipeline* prend en entrée : les prédictions binaires du modèle par *frame*, la fréquence d'échantillonnage (*frame rate)* du flux de mouvement, les métadonnées associées (nom du fichier, durée, etc.). Pour chaque fichier analysé, le pipeline fonctionne en plusieurs étapes :

1) Initialisation du fichier `.eaf` avec un tier dédié aux sorties du modèle.

2) Conversion des indices de *frames* en temps absolus (ms) en utilisant le *frame rate* et les métadonnées, afin d'obtenir des timestamps précis pour chaque *frame* en fonction du flux de mouvement initial.

3) Détection des prédictions positives : le pipeline identifie toutes les séquences contiguës de *frames* prédites comme positives, c'est-à-dire lorsque le label prédit est égal à 1, indiquant la présence d'un geste.

4) Annotation des segments : chaque séquence continue identifiée est considérée comme un *stroke* (le modèle ayant été entrainé sur les annotations des phases de *stroke* seulement, telles qu'annotées dans re-Beat14) et ajoutée au tier dédié du fichier `.eaf`, avec des timestamps de début et de fin exprimés en millisecondes.

5) Association des fichiers vidéo : le *pipeline* associe le fichier vidéo (.mp4) correspondant aux mouvements analysés par le modèle au fichier `.eaf` généré, permettant de visualiser la

segmentation proposée par le modèle ainsi que la reconstitution du mouvement sous format vidéo de manière synchronisée dans ELAN.

## 2.3 Configurations techniques

Le modèle a été implémenté à l'aide de PyTorch Lightning et entraîné sur un GPU unique (NVIDIA RTX A3000 Laptop) avec support CUDA, sur une machine équipée d'un CPU Intel(R) Core(TM) i9-11950H, 2,60 GHz et 32 Go de RAM. Le modèle *baseline* comporte environ 1 million de paramètres entraînables, avec une empreinte mémoire d'environ 4 Mo, et un temps d'entraînement d'environ 3 minutes.

La fonction de perte utilisée pour la mise à jour des poids du modèle lors de l'entraînement est la *Binary Cross-Entropy Loss* (BCE *Loss*), adaptée aux problèmes de classification binaire *frame* par *frame*. Elle mesure l'écart entre les probabilités prédites par le modèle et les labels binaires réels, permettant ainsi au modèle d'apprendre à produire des sorties proches de 0 ou 1 correspondant à l'absence ou à la présence d'un geste. Si $y_i$ est le label réel (0 ou 1) et $\hat{y}_i$ la probabilité prédite par le modèle pour la i-ème instance (*frame*), la BCE *Loss* moyenne sur N instances est définie par :

$$\mathrm{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)]$$

La première partie $y_i \cdot \log(\hat{y}_i)$ pénalise les faux négatifs et la seconde partie $(1 - y_i) \cdot \log(1 - \hat{y}_i)$ pénalise les faux positifs. La BCE *Loss* tend vers 0 lorsque les prédictions du modèle sont parfaitement alignées avec les labels réels.

PyTorch Lightning a été configuré pour sauvegarder automatiquement les paramètres du modèle correspondant à la meilleure valeur de la *loss* sur l'ensemble de validation, garantissant ainsi la conservation de la version du modèle la plus performante sur des données sur lesquelles il n'a pas été entraîné. L'entraînement a été effectué sur 64 époques, où une époque correspond à un passage complet sur l'ensemble des fichiers d'entraînement. Le choix de limiter l'entraînement à ce nombre d'époques a été guidé par la prévention du surapprentissage (*overfitting*). L'*overfitting* se produit lorsque le modèle devient trop spécialisé sur les données d'entraînement, perdant sa capacité à généraliser à de nouvelles séquences. En combinant la sauvegarde basée sur la *loss* de validation et un nombre d'époques contrôlé, le modèle apprend à capturer les *patterns* de mouvements pertinents tout en évitant de mémoriser des détails spécifiques aux séquences d'entraînement.

L'entraînement du modèle *baseline* a été optimisé avec l'algorithme Adam [185] en utilisant ses paramètres par défaut, un taux d'apprentissage (*learning rate*) fixé à 0,001 et un batch de 10 séquences par itération.

Pour garantir la reproductibilité des expériences, une *seed* (graine aléatoire) a été fixée lors de l'entraînement du modèle. La *seed* initialise les générateurs de nombres aléatoires utilisés pour l'initialisation des poids, le tirage aléatoire des batches et toute autre opération stochastique du processus d'entraînement. En fixant cette valeur, on s'assure d'avoir les mêmes

conditions d'entraînement, ce qui permet de comparer les performances entre différentes configurations.

# 3 Analyses des performances du modèle *baseline*

## 3.1 Métriques utilisées pour l'analyse de performance

### 3.1.1 Métriques *frame*-wise

La performance du modèle a été évaluée à l'aide de quatre métriques classiques de classification binaire, calculées à partir des vrais positifs (TP), faux positifs (FP), vrais négatifs (TN) et faux négatifs (FN). On définit ces quatre quantités à partir de la comparaison, *frame* par *frame*, entre les annotations expertes et les prédictions du modèle :

- Un vrai positif (TP, *true positive*) correspond à une *frame* d'un geste (label expert de classe 1) et qui est correctement prédite comme tel par le modèle (label prédit de classe 1).
- Un faux positif (FP, *false positive*) correspond à une *frame* qui n'est pas celle d'un geste (label expert de classe 0) mais que le modèle prédit à tort comme contenant un geste (label prédit de classe 1). Il s'agit donc d'une fausse détection.
- Un vrai négatif (TN, *true negative*) correspond à une *frame* qui n'est pas celle d'un geste (label expert = 0) et qui est correctement prédite comme tel par le modèle (label prédit = 0).
- Un faux négatif (FN, *false negative*) correspond à une *frame* d'un geste (label expert = 1) mais que le modèle ne détecte pas comme telle (label prédit = 0).

A partir de ces valeurs, ont été calculées :

$$\text{L'accuracy} = \frac{TP + TN}{TP + TN + FP + FN}\text{, la précision} = \frac{TP}{TP + FP},$$

$$\text{le rappel} = \frac{TP}{TP + FN}\text{, et le F1Score frames} = \frac{2 \cdot \text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

L'exactitude (*accuracy*) mesure la proportion globale de prédictions correctes. Elle donne une vue d'ensemble des performances du modèle, mais peut être trompeuse en cas de déséquilibre entre les classes (par exemple lorsqu'il y a, comme ici, beaucoup plus de *frames* de classe 0 que de *frames* de classe 1). La précision (*precision*) évalue la fiabilité des prédictions positives, c'est-à-dire la proportion de *frames* prédites de classe 1 qui correspondent effectivement à une *frame* d'un geste. Le rappel (*recall*) mesure la capacité du modèle à détecter les *frames* de gestes de la vérité de terrain, soit la proportion de *frames* de classe 1 correctement prédites par le modèle. Enfin, le F1-score correspond à la moyenne harmonique entre précision et rappel. Cette métrique est particulièrement pertinente dans notre contexte, du fait du déséquilibre de classe, et nécessite à la fois de limiter les fausses détections (haute précision) et de ne pas manquer de gestes réels (haut rappel). Le F1-score synthétise ces deux aspects en une valeur unique, ce qui en fait un indicateur plus robuste que l'*accuracy* seule pour évaluer

la qualité globale du modèle. La nature ponctuelle des gestes induit un fort déséquilibre de classes (72,5 % de frames sans geste). Bien que nous ayons utilisé une fonction de perte BCE standard, sans pondération des classes pénalisant davantage les faux négatifs, l'analyse approfondie via le F1-Score nous permet de garantir que le modèle ne converge pas vers la solution triviale consistant à prédire systématiquement la classe majoritaire 0.

Dans cette étude exploratoire, les performances de COSMOS sont comparées à une distribution aléatoire. Bien qu'une *baseline* heuristique (comme un simple détecteur de pics de vélocité) aurait pu être envisagée, notre objectif n'était pas de détecter des mouvements non communicatifs, mais bien de repérer des gestes (*strokes*) tels que validés par l'annotation multimodale experte. Un simple seuillage cinématique produirait un nombre massif de faux positifs (gesticulations, réajustements posturaux), ce que le modèle neuronal apprend justement à filtrer grâce à la supervision experte. Ainsi, pour calculer le F1-score « aléatoire », des séquences prédictives ont été générées aléatoirement en respectant les proportions de classes présentes dans le jeu de test (présence vs absence de geste). Ces prédictions permettent de simuler un modèle de reconnaissance « au hasard » tout en conservant la distribution réelle des labels dans les données. Le F1-score aléatoire est calculée de la même manière que présentée précédemment, au niveau des *frames*.

### 3.1.2 Métriques segment-wise

Pour évaluer la performance du modèle au niveau des segments temporels (groupe de plusieurs *frames*) prédits comme des gestes par le modèle, des métriques *segment-wise* ont été calculées. Contrairement aux métriques classiques au niveau de la *frame*, cette approche considère donc chaque segment contigu de plusieurs *frames* de classe 1 prédites par le modèle comme constituant un geste. Pour chaque geste prédit par le modèle dans les séquences de tests, le geste « de référence », tel qu'annoté lors de l'annotation experte, a été recherché, en calculant l'*Intersection over Union* (IoU). Ce calcul mesure le pourcentage de recouvrement temporel entre les annotations de gestes prédits et les annotations des gestes réels tels qu'annotés dans les mêmes fichiers. Si l'IoU dépasse un seuil défini (ici à 0,5), le geste prédit est considéré dans sa globalité comme un vrai positif (TP). Des gestes prédits non appariés avec des gestes réels sont quant à eux considérés comme faux positifs (FP), et les gestes réels non appariés avec des gestes prédits comme des faux négatifs (FN). À partir de ces comptages, la précision, le rappel et le F1-score au niveau des segments sont calculés. Étant donné que les vrais négatifs ne sont pas décomptés, le F1-score repose uniquement sur l'analyse de la proportion de segments correctement détectés (TP), des fausses détections (FP) et des gestes manqués (FN), selon la formulation suivante :

$$\text{F1Score segment} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

L'IoU moyen est également calculé uniquement sur les gestes appariés, afin d'évaluer la qualité moyenne de l'alignement temporel des gestes correctement détectés.

## 3.2 Performances sur les sets d'entraînement et de validation

### 3.2.1 Valeurs de la fonction de perte à chaque *step*

L'évolution de la fonction de perte au cours de l'entraînement (Figure 54, courbe bleue) montre une diminution progressive et régulière, caractéristique d'un processus d'optimisation stable. La valeur initiale relativement élevée de la *loss* (environ 0,55) reflète l'état d'un modèle encore peu informé, générant un nombre important d'erreurs de classification. Au cours des premières centaines d'itérations (*steps*, axe des abscisses), une décroissance rapide est observée, indiquant un ajustement efficace des paramètres du réseau en réponse au signal d'erreur. Par la suite, la courbe présente une phase de ralentissement et tend vers un plateau autour de 600 *steps*, pour atteindre une valeur proche de 0,15, suggérant que le modèle a convergé vers une solution stable sur les données d'entraînement.

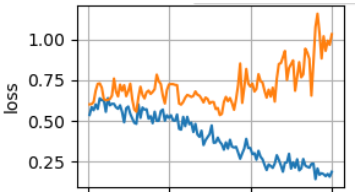



*Figure 54. Valeurs de la loss pendant l'entrainement (bleu) et la validation (orange)*

Durant les 200 premiers *steps*, la *loss* pendant la validation (courbe orange) ne présente pas de tendance nette à la convergence et oscille fortement entre 0,50 et 0,75. Cette variabilité importante indique que le modèle n'a pas encore appris une représentation stable généralisable aux données non vues : ses prédictions restent instables et sensibles aux fluctuations propres au jeu de validation, le modèle semble sous-performant pour détecter correctement ce qui est du geste de ce qui ne l'est pas. À partir d'environ 220 *steps*, une phase de convergence plus franche s'amorce, la fonction de perte diminue progressivement jusqu'à atteindre un minimum autour du *step* 324, avec une valeur de *loss* proche de 0,50. Ce point correspond au *checkpoint* retenu pour le modèle final, sauvegardé à l'époque 64, car il maximise les performances sur le jeu de validation tout en limitant le surapprentissage, que l'on peut observer après ce point avec l'augmentation progressive de la *loss*.

### 3.2.2 Métriques à chaque *step*

Pendant l'entraînement, l'*accuracy* augmente progressivement d'environ 0,75 à 0,80, ce qui indique que le modèle améliore sa capacité à classifier correctement les *frames*, correspondant à la présence ou à l'absence d'un geste (Figure 55). Autrement dit, la proportion de *frames* correctement détectées quel que soit leur classe devient plus élevée au fil des itérations. Les métriques de précision et de rappel montrent également une progression au cours de l'entraînement. La précision, qui passe d'environ 0,5 à 0,8,

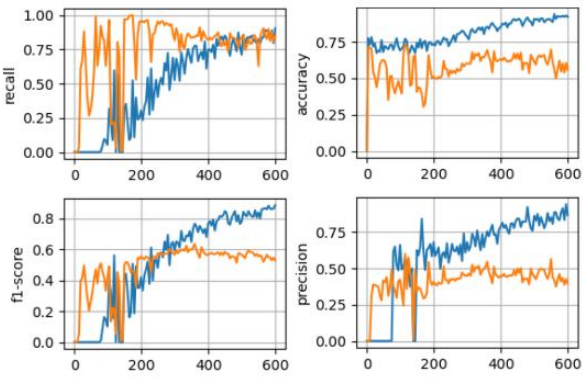



*Figure 55. Valeurs des métriques pendant l'entraînement (bleu) et la validation (orange)*

traduit une amélioration de la fiabilité des prédictions positives : le modèle réduit le nombre de faux positifs, c'est-à-dire les *frames* classées à tort comme contenant un geste. Le rappel, qui évolue de 0,10 à 0,8, reflète la capacité croissante du modèle à détecter les gestes réellement présents, en diminuant le nombre de faux négatifs. En termes de détection de gestes, ces évolutions signifient que le modèle apprend progressivement à identifier les *frames* de classe 1 de manière plus précise : au début de l'entraînement, il manque de nombreux gestes (rappel faible) et ses prédictions positives sont peu fiables (précision faible), tandis qu'à la fin, il reconnaît la majorité des gestes tout en limitant les fausses détections. L'évolution de ces deux métriques se reflète dans le F1-score, qui combine précision et rappel pour fournir une mesure unique de performance équilibrée, et qui atteint jusqu'à 0,9. Cette valeur élevée du F1-score confirme que le modèle est à la fois sensible (rappel élevé) et spécifique (précision élevée) dans la détection de *frames* positives pendant l'entraînement.

La courbe de validation (courbe orange) présente une dynamique globalement similaire à celle observée sur les données d'entraînement, mais à des niveaux légèrement inférieurs et avec davantage de fluctuations au début de l'entraînement. L'*accuracy* oscille entre 0,45 et 0,75, indiquant une capacité de généralisation correcte, mais encore perfectible sur les données non vues. Les métriques de précision, de rappel, et de F1-score sont initialement plus faibles et montrent des variations marquées durant les 200 premiers *steps,* révélant une sensibilité accrue aux échantillons du jeu de validation et une instabilité des prédictions positives. Elles se stabilisent ensuite aux alentours de 0,5 pour la précision, de 0,8 pour le rappel, et 0,6 pour le F1-score. Ces observations indiquent que le modèle, après une phase instable témoignant du sous-apprentissage pendant la validation, parvient finalement à une performance équilibrée et robuste, capable de détecter les gestes tout en limitant les fausses détections.

## 3.3 Performances sur le set de test

| Métrique | Accuracy | Précision | Rappel | F1 Score | F1 Aléatoire | IoU Moyen |
|---|---|---|---|---|---|---|
| *Frame* | 0,86 | 0,69 | 0,76 | 0,72 | 0,25 | - |
| *Segment* | - | 0,68 | 0,62 | 0,65 | - | 0,71 |

*Tableau 45. Valeurs des métriques sur le set de test (IoU = intersection over union)*

### 3.3.1 Analyse des métriques *frame-wise*

L'analyse des valeurs de métriques sur les données de test révèle des performances élevées sur l'ensemble des indicateurs (Tableau 45). L'*accuracy* de 86 % indique que la majorité des *frames*, quel que soit leur label, a été correctement classée. Bien que cette métrique puisse être surestimée en raison du déséquilibre des classes, les valeurs élevées de précision et de rappel confirment que le modèle détecte efficacement les *frames* classées comme des « gestes » tout en limitant les erreurs de classification. La précision, qui s'élève à 69 %, indique que la plupart des *frames* prédites comme des gestes correspondent réellement à cette classe selon la vérité terrain, traduisant une bonne limitation des faux positifs tout en laissant subsister une marge d'erreur résiduelle. Le rappel, de 76 %, signifie que la majorité des *frames* étiquetées « gestes » selon la

vérité de terrain ont été correctement détectées, montrant que le modèle est sensible aux *frames* de classe 1, tout en pouvant encore manquer certaines occurrences (faux négatifs). Le F1-score de 0,72, qui combine précision et rappel, reflète un bon équilibre entre la classification correcte des *frames* contenant un geste et la limitation des fausses prédictions. Comparé au score de F1 aléatoire de 0,25, cela montre que le modèle apprend significativement mieux que le hasard, confirmant sa capacité réelle à détecter les gestes sur des *frames* individuelles.

### 3.3.2 Analyse des métriques *segment-wise*

L'évaluation des performances *segment-wise* complète l'analyse *frame-wise* et permet de mieux déterminer la capacité du modèle à détecter et segmenter les gestes de manière cohérente sur des plages temporelles continues (Tableau 45). Pour le set de test, les métriques *segment-wise* révèlent un F1-score de 0,65, légèrement inférieur au F1-*frame* (0,72), ce qui est attendu puisque le calcul au niveau du geste est plus strict que pour la *frame* et nécessite qu'au moins la moitié du geste prédit recouvre le geste réel pour être considéré comme correct. Les valeurs de précision (0,68) et de rappel (0,62) confirment que le modèle détecte de manière fiable les gestes sur le jeu de test. La moyenne des IoU, de 0,71, indique que pour les gestes correctement identifiés, le recouvrement temporel avec la vérité terrain est en moyenne à 71%, signifiant que pour les gestes correctement identifiés, les gestes prédits se chevauchent en grande partie avec les segments réels, bien que de légers décalages temporels puissent subsister.

Les erreurs résiduelles au niveau des *frames*, et par extension au niveau des segments, peuvent s'expliquer par la continuité du signal gestuel, qui rend la délimitation précise des gestes difficile. Étant donné que les mouvements sont continus, le modèle peut éprouver des difficultés à déterminer exactement quelle *frame* correspond au début ou à la fin d'un geste. Cette ambiguïté temporelle aux frontières des gestes peut augmenter le nombre de faux positifs ou des faux négatifs sur certaines *frames*, même lorsque le cœur du segment est correctement détecté, reflétant de légers décalages dans le positionnement des segments. Dans le cadre d'une approche semi-automatique pour la segmentation des gestes, ce niveau de précision est néanmoins satisfaisant : il suffit que le modèle identifie correctement le centre ou la majorité du geste dans le flux de mouvement, car l'annotateur peut ensuite affiner manuellement les bordures temporelles si nécessaire. Ainsi, le modèle fournit déjà une base fiable et utilisable pour la segmentation semi-automatique des gestes du locuteur Wayne pour des fichiers avec lesquels il n'a pas été entraîné.

### 3.3.3 Analyse de la robustesse du modèle

Pour évaluer la stabilité et la fiabilité du modèle *baseline*, plusieurs évaluations ont été exécutées sur les données de test, chacune dans des conditions identiques pour garantir la reproductibilité. Cette démarche permet de confirmer que les métriques calculées sont cohérentes et robustes, et que l'inférence du modèle produit des résultats fiables et reproductibles, indépendamment d'une éventuelle variation d'exécution.

| Métrique | ***Frame*-wise** | | | **Segment-wise** | | |
|---|---|---|---|---|---|---|
| | *Moy* | *Med* | *SD* | *Moy* | *Med* | *SD* |
| F1-Score | 0,64 | 0,69 | 0,13 | 0,48 | 0,53 | 0,16 |
| Précision | 0,65 | 0,68 | 0,08 | 0,50 | 0,52 | 0,12 |
| Rappel | 0,65 | 0,69 | 0,16 | 0,48 | 0,52 | 0,17 |
| Accuracy | 0,83 | 0,86 | 0,04 | - | - | - |
| IoU moyen | - | - | - | 0,69 | 0,69 | 0,03 |

*Tableau 46. Valeurs moyennes des métriques lors d'exécutions multiples sur le set de test (Moy = moyenne ; Med = médiane ; SD = écart-type ; IoU = intersection over union)*

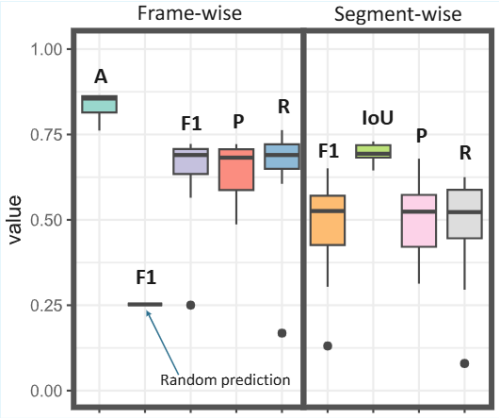



*Figure 56. Représentation graphique des distributions des valeurs moyennes de métriques (A = accuracy ; F1 = F1-Score, P = précision ; R = rappel ; IoU = intersection over union)*

L'analyse des valeurs moyennes des métriques calculées sur plusieurs exécutions du modèle (Tableau 46, Figure 56) montre des performances stables et cohérentes au niveau des *frames*, avec un F1-score, une précision et un rappel moyens aux alentours de 0,65, et des médianes proches de 0,69. Ces résultats indiquent que d'une exécution à l'autre, le modèle maintient une capacité constante à classer correctement les *frames*, confirmant sa robustesse et sa fiabilité. Les écarts entre la moyenne et la médiane suggèrent toutefois que certaines séquences sont légèrement plus difficiles à segmenter selon l'exécution, mais sans impacter significativement la cohérence globale des prédictions.

L'analyse des métriques *segment-wise* sur plusieurs exécutions répétées du modèle révèle des performances inférieures (moyennes et médianes aux alentours de 0,5 pour le F1-Score, l'accuracy et le rappel) à celles observées au niveau *frame-wise*, ainsi qu'une variabilité plus importante de ces valeurs. L'IoU est la seule métrique stable (moyenne = 0,69, médiane = 0,69, écart-type = 0,03) et montre que, lorsque les gestes sont correctement détectés, le recouvrement temporel moyen avec la vérité terrain reste élevé, quelle que soit l'exécution du modèle.

Dans l'ensemble, ces analyses confirment que le modèle *baseline* est robuste et fiable, en maintenant des performances stables et cohérentes d'une exécution à l'autre, notamment pour la détection au niveau des *frames*. Les métriques *segment-wise* semblent plus variables en comparaison. Les métriques *segment-wise* peuvent être particulièrement sensibles à l'augmentation des erreurs de détection au niveau des *frames*. En effet, si un segment prédit ne recouvre pas suffisamment le geste (seuil à 50 %) parce que moins de *frames* ont été détectées, il n'est pas comptabilisé comme vrai positif, ce qui peut réduire artificiellement la précision, le rappel et le F1-score *segment-wise*. À l'inverse, le F1 *frame-wise* reste plus stable, car il évalue chaque *frame* individuellement et est moins affecté par de petits décalages ou omissions. Dans un contexte de segmentation semi-automatique, le seuil de recouvrement utilisé pour les calculer

les métriques *segment-wise* n'est pas un critère limitant pour l'annotation. Tous les segments prédits sont proposés comme gestes potentiels, et l'annotateur peut ensuite vérifier et ajuster les frontières si nécessaire. Ainsi, l'identification des gestes corrects ne dépend pas strictement du recouvrement temporel. Dans ce contexte, nous privilégions donc le F1 *frame-wise* comme métrique principale pour évaluer les performances. Le modèle reste donc exploitable pour segmenter les gestes sur des fichiers d'un même locuteur et guider l'annotation manuelle.

### 3.3.4 Analyse d'une séquence de prédiction

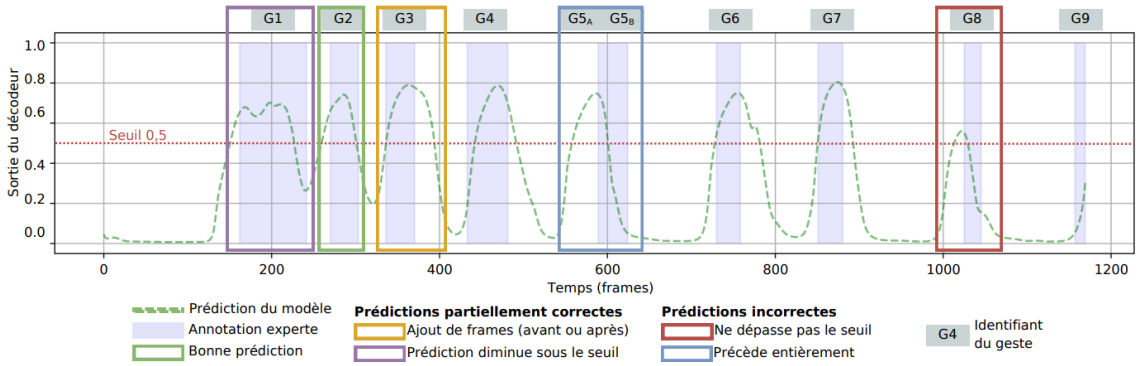



*Figure 57. Exemple de prédiction et d'annotations expertes sur une séquence*

La Figure 57 illustre une séquence issue de l'ensemble des fichiers de test, montrant à la fois les prédictions du modèle (courbe verte) et les annotations manuelles expertes (fond violet). On peut observer que, pour cette séquence, le modèle détecte un nombre de gestes similaire à celui de l'annotation experte (neuf gestes identifiés de G1 à G9), avec huit pics de valeurs en sortie du décodeur dépassant le seuil de 0,5 (donc huit gestes prédits). Le G9, en position finale dans la séquence, n'est pas détecté, ce qui pourrait s'expliquer par une troncation induite par la segmentation du fichier. Certaines prédictions peuvent être considérées comme correctes (G2 ou le G7), avec très peu de *frames* supplémentaires ou manquantes. D'autres prédictions, bien que plus ou moins proches des annotations expertes, montrent toutefois des différences pouvant être considérées comme des erreurs de prédictions.

#### Prédictions partiellement correctes

Dans cette séquence, des prédictions partiellement correctes sont observées, caractérisées par un recouvrement partiel entre les *frames* prédites et les *frames* annotées comme gestes lors de l'annotation manuelle (Figure 58). Ces cas peuvent impliquer une extension temporelle du geste prédit avec un ajout de *frames* avant le geste annoté manuellement, comme pour le G1, ou après le geste, comme pour le G3, ou au contraire, une troncation, lorsque la sortie du décodeur passe sous le seuil de 0,5 avant la fin du geste annoté manuellement (G1).

Le geste G3 annoté est un geste avec lequel le locuteur désigne dans l'espace « une petite ville » avec sa main droite. La trajectoire issue de la segmentation experte permet ainsi de voir une montée de la main (et en parallèle une descente de la main gauche). La différence principale avec la prédiction proposée par le modèle réside dans une extension temporelle de la fin du geste prédit sur la main droite. Cette extension se manifeste principalement au niveau de la main droite, où l'on observe facilement l'ajout d'une trajectoire descendante, correspondant en réalité à la phase de rétractation du geste. Pendant l'exécution du geste par la main droite, la main gauche réalise un mouvement non communicatif vers le bas, puis le moment où la main gauche devient statique coïncide avec la fin de l'exécution du geste. Ainsi, du fait de l'immobilité de la main gauche sur la portion de *frames* prédites comme faisant partie du geste, les trajectoires telles que visualisées ici semblent identiques. Dans ce contexte, la prédiction du modèle pour le G3 semble partiellement correcte, les *frames* supplémentaires devront être retirées lors de la vérification des gestes par un annotateur expert.

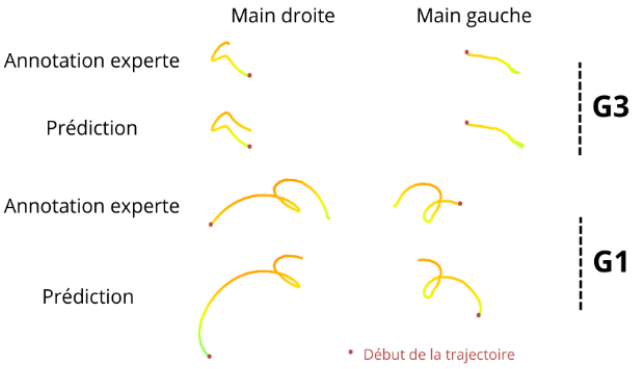



*Figure 58 Exemples de trajectoires entre prédictions et annotations expertes (G1 et G3)*

Le geste G1 est un geste de désignation réalisé avec les deux mains par le locuteur et se déroule en deux temps : une première courbe ample suivie d'une courbe plus petite. Dans le cas de la prédiction, un ajout de *frames* est également observé, au début du geste. Le modèle inclut l'ensemble de la première courbe du mouvement des mains droite et gauche, alors que l'annotation experte situe le début du geste au milieu de la première courbe. Dans le cas de ce geste, il peut être pertinent de considérer l'ensemble de la dynamique du mouvement tel que le propose le modèle. Le locuteur effectue une grande boucle préalable, ce qui peut rendre la segmentation du geste ambiguë quant à la portion à considérer comme communicative sur l'ensemble de cette boucle. D'un point de vue cinématique, il pourrait donc être pertinent de prendre en compte le mouvement dès le début de la boucle.

La prédiction du G1 présente également une troncation temporelle en fin de geste, liée à une chute de la sortie du décodeur sous le seuil de décision. Il en résulte un segment prédit dont la partie finale n'inclut pas la trajectoire descendante observée dans l'annotation experte. Cette interruption peut potentiellement s'expliquer par la dynamique du mouvement : la seconde courbe présente une amplitude plus faible que la première, et l'enchaînement des deux peut amener le modèle à détecter principalement la première. Ainsi, la première courbe pourrait exercer un poids plus important sur la sortie du décodeur, conduisant le modèle à privilégier cette portion dans sa prédiction. Cela n'est toutefois pas problématique pour l'annotation semi-automatique, car la majeure partie du geste est correctement détectée, et il suffira à l'annotateur de corriger cette fin de segment lors de la vérification.

**Prédictions incorrectes**

La séquence envisagée ici comprend également certaines prédictions incorrectes, caractérisées par un décalage temporel important entre la sortie du modèle et l'annotation experte. Par exemple, pour le geste G5, le geste prédit G5A se situe juste avant le geste annoté G5B, dont il résulte les trajectoires différentes reconstruites dans la Figure 27. Le geste annoté manuellement correspond ainsi à un stroke avec une trajectoire descendante des deux mains, tandis que les trajectoires issues de la segmentation du modèle pour le geste prédit G5A correspondent à la remontée des mains du locuteur. Il n'existe pas d'explications précises quant à la différence de segmentation constatée ici. Le décalage temporel est toutefois tel que le geste prédit et le geste annoté apparaissent comme deux segments distincts, sans superposition, ce qui est considéré comme une erreur de prédiction dans le cadre d'une comparaison stricte avec l'annotation experte.

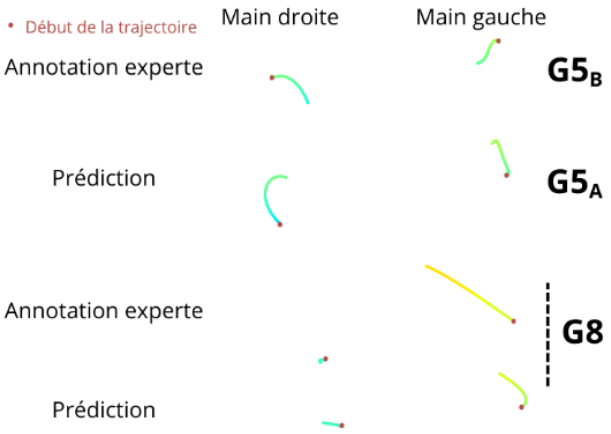



*Figure 59. Exemples de trajectoires entre prédictions et annotations expertes (G5 et G8)*

D'après l'annotation experte, le geste G8 correspond à un geste de désignation, consistant en un pointage en ligne droite vers le haut réalisé avec la main gauche. L'erreur de prédiction du modèle pour ce geste s'explique par une sortie du décodeur qui dépasse très peu le seuil de décision, et ce, sur une très courte durée. En conséquence, la trajectoire prédite pour G8 est beaucoup plus courte que le segment annoté : la prédiction inclue quelques *frames* supplémentaires, puis elle ne reprend qu'une toute petite partie du début de la ligne droite observée dans la segmentation experte. Cette faible valeur en sortie du décodeur entraîne un recouvrement temporel très limité entre la prédiction et le geste effectivement annoté, et conduit à une prédiction incorrecte du geste. Dans ce cas précis, cette prédiction semble être une erreur de la part du modèle.

Ainsi, le modèle peut produire des erreurs ponctuelles (partielles ou totales), justifiant la nécessité d'une révision humaine. Toutefois, ses bonnes performances globales en font un outil particulièrement utile, permettant un gain de temps significatif dans le processus de segmentation. Grâce à l'entraînement supervisé avec l'annotation experte, le modèle encode les mouvements en extrayant les *features* les plus discriminantes pour la reconnaissance des *frames* composant les gestes. Bien que la segmentation experte ne repose pas uniquement sur le mouvement, la dynamique constitue un critère majeur pour les identifier. Enfin, le modèle peut proposer des segmentations différentes de la référence pour certains gestes, mais qui restent plausibles d'un point de vue cinématique ou dynamique. Il peut donc être intéressant de tenter d'interpréter la logique des prédictions du modèle plutôt que de les rejeter systématiquement :

ces différences peuvent refléter une logique interne du modèle pour la reconnaissance des gestes, et potentiellement une autre manière valide de segmenter le geste.

# 4 Analyses de l'impact des hyperparamètres sur les performances du modèle *baseline*

## 4.1 Nombre de canaux et nombre de neurones

### 4.1.1 Architectures d'encodeurs convolutionnels testés

La profondeur du réseau (nombre de couches) et la taille de chaque couche (nombre de neurones) peuvent influencer directement la qualité des représentations temporelles apprises à partir des données de mouvement, ainsi que la capacité de généralisation du modèle. Dans le contexte d'un corpus annoté de taille limitée, ces choix architecturaux impliquent un compromis entre capacité de modélisation et complexité du modèle, afin d'éviter à la fois le sous-apprentissage et le surapprentissage.

| Id | Encodeurs testés | Tailles de sortie par RB | Durée | N param. | Taille |
|---|---|---|---|---|---|
| Ba | *Baseline* | 128, 256, 128, 96 | 3min | 1M | 4 Mo |
| R1 | Réduction uniforme | 64, 128, 64, 96 | + 0 s | -326K | -1,3 Mo |
| A1 | Augmentation modérée | 192, 384, 192, 96 | + 35 s | +600K | +2,4 Mo |
| BN1 | Bottleneck réduit | 128, 192, 128, 96 | +1 min 23s | -88K | -0,3 Mo |
| BN2 | Bottleneck accentué | 128, 384, 128, 96 | + 1 min | +300K | +1,4 Mo |
| D-3 | Profondeur réduite | 256, 96 | + 7 s | -26K | -0,5 Mo |
| D+6 | Profondeur augmentée | 28, 192, 256, 192, 128, 96 | + 1 s | +400K | +1,7 Mo |

*Tableau 47. Conditions testées pour la profondeur du réseau (N param. = nombre de paramètres)*

Plusieurs variantes architecturales ont été explorées à partir du modèle *baseline* (Tableau 47). Une réduction uniforme du nombre de canaux (R1) a été conçue pour évaluer la capacité minimale nécessaire pour la tâche de segmentation, tandis qu'une augmentation modérée (A1) a été testée afin d'explorer si une capacité accrue permet d'améliorer la détection au prix d'une complexité plus élevée. Deux variantes centrées sur la couche intermédiaire (BN1 et BN2) permettent plus spécifiquement d'examiner le rôle du goulot d'étranglement (*bottleneck*) dans la structuration des représentations internes, en évaluant respectivement l'effet d'une expansion plus ou moins grande de cette couche centrale. Enfin, des architectures de profondeur réduite (D−3) et augmentée (D+6) ont été étudiées afin de déterminer l'influence du nombre de blocs résiduels sur la capacité du modèle à capturer la dynamique temporelle des gestes, ainsi que sur les risques de surapprentissage associés à des réseaux plus profonds.

Le nombre total de paramètres et la taille mémoire des variations architecturales de l'encodeur permettent de mettre en évidence l'impact des choix de profondeur et de nombre de neurones sur la capacité de modélisation et le coût computationnel. Par rapport au modèle de référence (1M paramètres, 4 Mo), certaines variantes présentent une réduction substantielle de la complexité, notamment R1 (–326K ; –1,3 Mo), tandis que d'autres augmentent sensiblement

le nombre de paramètres et la taille (A1 : +600K, +2,4 Mo ; BN2 : +300K, +1,4 Mo ; D+6 : +400K, +1,7 Mo). Certaines variantes restent proches de la configuration initiale (BN1 et D-3).

### 4.1.2 Impact de la complexité et de la profondeur

L'analyse des valeurs de F1-Score (*frame-wise* et *segment-wise*) et de la moyenne des IoU met en évidence des différences de performance selon les configurations testées (Figure 60). Dans l'ensemble, les performances restent systématiquement inférieures à celles de la *baseline*, avec un impact particulièrement prononcé sur le F1-segment, tandis que la diminution est moins marquée pour le F1-*frame* et le IoU moyen. Seul le IoU moyen du modèle D+6, correspondant à l'architecture la plus profonde, dépasse légèrement celui de la *baseline*, mais de manière marginale. Il est à noter que la variante BN2, avec une couche centrale élargie par rapport à celle du modèle *baseline*, n'a pas produit de métriques exploitables. La cause exacte n'est pas déterminée, mais il est possible que la taille excessive de la couche intermédiaire ait perturbé l'entraînement ou l'inférence.

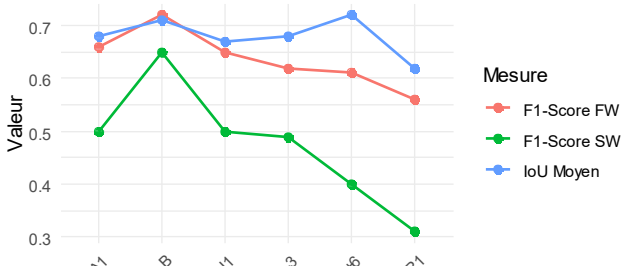



*Figure 60. Valeurs des métriques de F1 et IoU en fonction de la complexité des modèles (FW = frame-wise ; SW = segment-wise ; IoU = intersection over union)*

Ces résultats indiquent que la réduction de la capacité (R1, BN1) ou de la profondeur (D-3) du modèle compromet sa faculté à apprendre des représentations riches. À l'inverse, l'augmentation de la capacité (A1) ou de la profondeur (D+6) n'entraîne pas d'amélioration significative des performances, tout en générant un surcoût computationnel lié à l'accroissement du nombre de paramètres et de l'empreinte mémoire. Collectivement, ces observations suggèrent que l'encodeur de la *baseline* est bien dimensionné par rapport à la taille du corpus, offrant un compromis optimal entre complexité du modèle et capacité de modélisations.

## 4.2 Taux d'apprentissage

### 4.2.1 Valeurs testées

L'impact du taux d'apprentissage a été évalué afin de mesurer son influence sur les performances du modèle *baseline,* dont l'architecture restait inchangée. Le taux d'apprentissage (lr) contrôle l'amplitude des mises à jour des paramètres pendant l'entraînement : un lr trop élevé peut provoquer des oscillations et empêcher la convergence, tandis qu'un lr trop faible peut ralentir l'apprentissage ou bloquer le modèle dans un optimum local. Tester différentes valeurs permet donc de trouver un compromis entre vitesse d'apprentissage et stabilité de convergence. Trois valeurs de lr ont été explorées en comparaison avec la valeur du lr (=1e-3) du modèle *baseline* : 1e-5, 1e-4 et 1e-2.

### 4.2.2 Impact du taux d'apprentissage

L'évaluation de l'impact du taux d'apprentissage sur les performances du modèle (Tableau 48) montre que la valeur initiale de lr = 1e-3 offre les meilleurs performances, avec un F1-*frame* de 0,72, un F1-segment de 0,65 et un IoU moyen de 0,71. Une réduction du lr à 1e-4 entraîne une chute importante des performances (F1-*frame* = 0,33 ; F1-segment = 0,26 ; IoU moyen = 0,64), indiquant que le modèle n'a pas suffisamment convergé au cours du nombre d'époques prévu pour optimiser ses paramètres. Les autres taux de lr testés (1e-5 et 1e-2) n'ont pas permis d'obtenir de résultats exploitables sur le jeu de données de test, pour des raisons non élucidées, mais probablement liées à un apprentissage trop lent (1e-5) ou à une instabilité de convergence (1e-2). La valeur du taux d'apprentissage choisie pour le modèle *baseline* apparaît adaptée, permettant au modèle de converger correctement sur notre corpus.

| | ***Baseline*** *Lr = 1e-3* | **Lr=1e-4** |
|---|---|---|
| F1-*frame* | 0,72 | 0,33 |
| F1-segment | 0,65 | 0,26 |
| IoU moyen | 0,71 | 0,64 |

*Tableau 48. Performances en fonction du learning rate (lr)*

## 4.3 Hyperparamètres contrôlant le surapprentissage

### 4.3.1 Méthodes pour contrôler le surapprentissage

#### Régularisation des poids de la fonction de perte

Au cours de l'apprentissage du modèle *baseline*, une augmentation progressive de la *loss* en validation a été observée, suggérant un risque de surapprentissage sur le jeu d'entraînement. Dans ce contexte, le modèle peut présenter des performances réduites sur les données de test, en raison d'une capacité limitée à généraliser au-delà des exemples rencontrés pendant l'entraînement. Pour limiter ce phénomène, une régularisation peut être appliquée lors de l'entraînement (L1 ou L2). L2 a été choisie afin de pénaliser fortement les poids de grande amplitude tout en les réduisant de manière continue et lisse. Contrairement à L1, qui tend à annuler certains poids (ce qui peut entraîner une perte d'informations importante pour un corpus de petite taille), L2 préserve l'ensemble des poids tout en limitant leur amplitude, ce qui permet de conserver la contribution de nombreuses *features* tout en réduisant la complexité globale du modèle. La régularisation L2 est calculée comme suit :

$$loss_{finale} = Loss_{originale} + \lambda \sum_i w_i^2$$

où $w_i$ sont les poids du modèle et $\lambda$ un hyperparamètre contrôlant la force de la régularisation. Plusieurs valeurs de $\lambda$ ont été testées : 1e-3, 1e-4, et 1e-5. L'objectif de cette comparaison, avec ou sans régularisation, est d'évaluer l'impact effectif de la régularisation sur les performances du modèle.

#### Valeurs de dropout

Pendant l'entraînement, le *dropout* consiste à ignorer temporairement, de manière aléatoire, un certain pourcentage de neurones afin d'obliger le réseau à apprendre des représentations plus robustes et distribuées. Varier le taux de *dropout* permet d'évaluer la capacité du modèle à limiter le surapprentissage et à améliorer la généralisation sur des données non vues. Trois valeurs de dropout ont été testées : 10 %, 20 % et 30 %, correspondant à la probabilité qu'un neurone soit désactivé aléatoirement à chaque itération. C'est trois valeurs ont ainsi été testées sans changer l'architecture du modèle *baseline*.

### 4.3.2 Impact de la régularisation L2

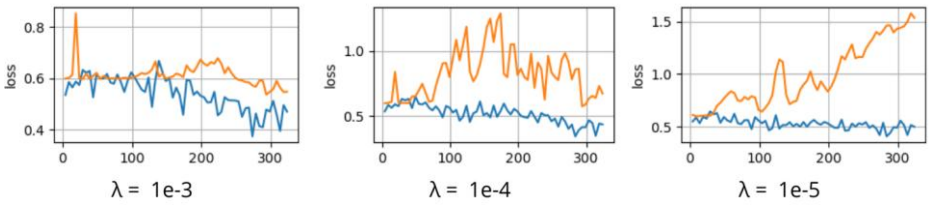



*Figure 61. Valeurs de loss en fonction des régularisations L2 appliquées*

On peut observer un effet clair de la régularisation sur l'évolution de la *loss* pendant la validation (Figure 61). Pour un taux de régularisation de 1e-3, la *loss* apparaît peu bruitée et converge de manière relativement stable après 200 *steps*, ce qui suggère un risque de surapprentissage moins prononcé. En revanche, pour les deux autres valeurs testées, la *loss* de validation présente un bruit important et une difficulté à converger, indiquant que le modèle a du mal à stabiliser ses prédictions. Enfin, pour $\lambda$ = 1e-5, aucune convergence n'est observée : la *loss* augmente continuellement, reflétant un échec d'apprentissage. Ces observations mettent en évidence l'impact de la régularisation sur la stabilité de l'entraînement.

| | ***Baseline*** | **Régularisation L2** | | |
|---|---|---|---|---|
| | *Pas de L2* | 1e-3 | 1e-4 | 1e-5 |
| F1-*frame* | 0,72 | 0,55 | 0,38 | 0,26 |
| F1-segment | 0,65 | 0,37 | 0,08 | 0,15 |
| IoU moyen | 0,71 | 0,63 | 0,67 | 0,65 |

*Tableau 49. Performances du modèle baseline selon la régularisation L2 appliquée (F1-frame = F1-Score appliqué aux frames ; F1-segment = F1-Score appliqué aux segments ; IoU moyen = interaction over union moyen)*

L'analyse des performances du modèle avec régularisation L2 met effectivement en évidence un sous-apprentissage significatif par rapport au modèle sans régularisation (Tableau 49). Cette baisse est particulièrement marquée pour les F1-scores par *frame* (0,26 à 0,55) et par segment (0,08 à 0,37), et s'accentue à mesure que la régularisation L2 devient plus faible, indiquant que le modèle peine à prédire avec précision tant au niveau des *frames* qu'au niveau des gestes. En revanche, l'IoU moyen montre une diminution plus modérée, suggérant que, pour les gestes prédits et réels dont le recouvrement temporel dépasse 50 %, ce recouvrement reste globalement correct en moyenne.

Ces résultats suggèrent qu'une régularisation L2 peut, dans ce contexte, se révéler contre-productive. Nous supposons que l'effet de la L2 n'est pas monotone : pour une régularisation trop faible, les poids ne sont pas suffisamment pénalisés pour guider efficacement l'optimisation, ce qui se traduit par une *loss* instable ou une convergence lente. Une régularisation plus forte favorise en revanche une convergence plus stable. Ce comportement indique l'existence d'une zone critique de régularisation, dans laquelle un petit intervalle de valeurs de $\lambda$ peut temporairement gêner l'apprentissage avant que l'effet bénéfique de la régularisation ne prédomine. Bien que le modèle *baseline* montre un certain surapprentissage au cours de l'entraînement, il conserve un équilibre satisfaisant entre capacités à s'ajuster aux données d'entraînement et performances sur des données non vues, indiquant que la généralisation demeure effective, même en l'absence de régularisation. Néanmoins, il pourrait être pertinent d'explorer un entraînement sur un plus grand nombre d'époques avec $\lambda$ = 1e-3, afin de vérifier si cette valeur de régularisation pourrait stabiliser encore davantage la convergence sur le long terme sans nuire à la performance.

### 4.3.3 Impact de l'activation du dropout

| | ***Baseline*** | **Dropout** | |
|---|---|---|---|
| | *Dropout = 0* | *10%* | *30%* |
| F1-*frame* | 0,72 | 0,62 | 0,66 |
| F1-segment | 0,65 | 0,47 | 0,54 |
| IoU moyen | 0,71 | 0,67 | 0,68 |

*Tableau 50. Performances du modèle baseline selon le dropout (F1-frame = F1-Score appliqué aux frames ; F1-segment = F1-Score appliqué aux segments ; IoU moyen = interaction over union moyen)*

Lorsque le *dropout* est activé, on observe une diminution des performances plus ou moins marquée selon le taux appliqué (Tableau 51). La F1-*frame* diminue par rapport à la *baseline* (0,72), atteignant 0,62 pour un dropout de 10 % et 0,66 pour 30 %, indiquant une légère sous-performance dans la détection des *frames* de classe 1 et une augmentation relative des erreurs. La F1-segment suit la même tendance, passant de 0,65 pour la *baseline* à 0,47 et 0,54, suggérant que le *dropout* rend la détection des gestes plus difficile. En revanche, l'IoU moyen reste stable, avec des valeurs proches de la *baseline* (0,71 contre 0,67 et 0,68), ce qui indique que, lorsque les gestes sont correctement détectés, le recouvrement temporel avec la vérité terrain n'est pas affecté par la désactivation neuronale. Il convient de noter que le *dropout* à 20 % n'a pas fourni de métriques exploitables.

# 5 Analyses de l'impact des données sur les performances

## 5.1 Évaluation des performances sur des données d'un locuteur non entraîné

Afin de réaliser une première évaluation de la capacité du modèle *baseline* à produire des représentations des gestes qui soient généralisables, et non uniquement corrélés aux gestes des

fichiers du seul locuteur utilisés pour l'entraînement, le modèle a été testé sur des données de mouvement provenant d'un locuteur inédit. Pour cette analyse, un fichier du locuteur Scott issu du corpus BEAT a été annoté manuellement, en suivant exactement le même protocole de segmentation des gestes que celui appliqué aux données de Wayne. Les données de mouvement et les annotations ont ensuite été normalisées afin de produire des fichiers synchronisés à l'échelle de la *frame*, comme précédemment. Au total, la séquence réannotée dure 2 min 30 s, comprenant 5 279 *frames* correspondant à des gestes et 12 721 *frames* sans geste, soit une durée et une répartition comparable à celles de l'ensemble de tests de Wayne, également de 2 min 30 s. Ces données ont enfin été analysées par le modèle *baseline*, permettant de générer les métriques de performance pour la prédiction de gestes sur ce nouveau locuteur.

| **Métrique** | **Accuracy** | **Précision** | **Rappel** | **F1 Score** | **F1 Aléatoire** | **IoU Moyen** |
|---|---|---|---|---|---|---|
| *Frame* | 0,67 | 0,44 | 0,38 | 0,41 | 0,30 | - |
| *Segment* | - | 0,27 | 0,19 | 0,22 | - | 0,67 |

*Tableau 51. Performances du modèle sur un locuteur inédit (IoU moyen = interaction over union moyen)*

Sur le locuteur inédit, les valeurs de précision (0,44) et le rappel (0,38) au niveau de la *frame* indiquent que le modèle parvient à identifier une partie des *frames* correspondant à un geste, mais avec une proportion non négligeable d'erreurs de classification. Le F1-frame (0,41), supérieur au F1 aléatoire (0,30), suggère néanmoins que le modèle capture une information pertinente au-delà du hasard, bien que la généralisation demeure partielle. Cette sous-performance est encore plus marquée à l'échelle du segment (F1-segment = 0,22), suggérant que la plupart des gestes prédits ne dépassent pas le seuil de recouvrement temporel requis pour être considérés comme des détections correctes. Toutefois, l'IoU moyen de 0,67 indique que, lorsque des gestes prédits peuvent être appariés à un geste annoté manuellement, le recouvrement temporel entre ces deux gestes reste relativement élevé en moyenne. Ces résultats suggèrent que le modèle capture certaines régularités dynamiques du geste, mais que l'absence de variabilité inter-locuteurs dans les données d'entraînement contraint la généralisation des représentations internes. Cette chute de performance illustre un phénomène classique de *Domain Shift* (décalage de domaine). Le modèle, entraîné exclusivement sur Wayne, a appris ses idiosyncrasies motrices, sa posture de repos et ses tics de langage corporel, et non une représentation universelle du « geste humain ». Ces résultats confirment que sans apprentissage multi-locuteurs, le modèle doit être envisagé stricto sensu comme un modèle spécifique à un locuteur (*speaker-dependent*). Une solution future consisterait à appliquer un *Few-Shot Fine-Tuning* (ré-entraînement sur un très petit échantillon) sur les premières minutes d'annotation de Scott.

## 5.2 Evaluation des performances avec d'autres données que les positions

### 5.2.1 Types de données testées

Dans la mesure où l'annotation experte des gestes ne se base pas uniquement sur le mouvement des articulations, mais s'inscrit dans un contexte multimodal, d'autres modalités de données ont été explorées afin d'évaluer leur impact sur les performances d'un modèle de segmentation de gestes, en comparaison avec le modèle *baseline*. Parmi ces modalités, la fréquence fondamentale (F0) et les *Mel-Frequency Cepstral Coefficients* (MFCC) ont été extraites des fichiers audios afin de capturer les caractéristiques acoustiques pertinentes de la parole. Les quaternions (quat) ont également été explorés comme une représentation alternative du mouvement. Ils permettent de coder les rotations dans l'espace 3D, offrant ainsi une description de l'orientation des segments corporels indépendamment de leur position absolue.

| Modalité | Taille de données | RB de l'encodeur |
|---|---|---|
| F0 | 1 | 27, 64, 128, 64 |
| MFCC | 27 | 64, 128, 64, 27 |
| Quat | 252 | 128, 256, 128, 96 |

*Tableau 52. Tailles des données et des couches de leur encodeur correspondant pour les nouvelles modalités de données testées (RB = blocs résiduels ; F0 = fréquence fondamentale ; MFCC = mel-frequency cepstral coefficient ; Quat = quaternions)*

Deux types de modèles ont été évalués : des modèles unimodaux, utilisant une seule modalité de données, et des modèles multimodaux, dans lesquels les représentations latentes issues de chaque encodeur étaient concaténées avant d'être transmises au décodeur, ce dernier restant identique à celui de la *baseline* pour tous les modèles testés. Les modalités audio (F0 et MFCC) ainsi que les quaternions, tout comme les positions 3D utilisées dans le modèle *baseline*, ont été encodées avec des encodeurs CNN 1D. Plusieurs configurations d'encodeurs CNN 1D ont été explorées pour chaque modalité, et seuls les modèles présentant les meilleures performances ont été retenus pour cette analyse exploratoire (Tableau 52). L'évaluation des performances a été limitée aux métriques *frame-wise*, sans considérer les mesures à l'échelle des segments.

### 5.2.2 Performances des modèles unimodaux et multimodaux testés

| | Modèle | Accuracy | Précision | Rappel | F1 *frames* |
|---|---|---|---|---|---|
| Unimodal | *Baseline* | 0,86 | 0,69 | 0,76 | 0,72 |
| | F0 | 0,68 | 0,33 | 0,33 | 0,33 |
| | MFCC | 0,64 | 0,31 | 0,44 | 0,37 |
| | Quat | 0,81 | 0,61 | 0,58 | 0,60 |
| Multimodal | Ba + F0 | 0,81 | 0,60 | 0,63 | 0,61 |
| | Ba + Quat | 0,82 | 0,64 | 0,55 | 0,59 |

| | | | | |
|---|---|---|---|---|
| Quat + F0 | 0,80 | 0,58 | 0,55 | 0,57 |
| Ba + Quat + F0 | 0,82 | 0,65 | 0,55 | 0,60 |

*Tableau 53. Métriques frame-wise des modèles testés avec d'autres types de données d'entraînement (F1 frames = F1-score appliqué aux frames ; Ba = baseline ; F0 = fréquence fondamentale ; MFCC = mel-frequency cepstral coefficient ; Quat = quaternions)*

L'analyse des performances des modèles unimodaux et multimodaux met en évidence un sous-apprentissage en comparaison au modèle *baseline* entraîné sur les positions 3D (Tableau 53). Ce sous-apprentissage est particulièrement marqué pour les modèles unimodaux basés sur les données audios, qu'il s'agisse de la F0 (F1-Score = 0,33) ou des MFCC (F1-Score = 0,37), suggérant qu'ils n'ont pas été capables de générer des représentations suffisamment pertinentes pour détecter les *frames* de gestes à partir de ces seules modalités. Plusieurs facteurs peuvent expliquer cela. Tout d'abord, la capacité du modèle à extraire des *features* discriminantes depuis l'audio semble limitée : les CNN 1D utilisés pour encoder ces signaux peuvent avoir du mal à identifier des *patterns* pertinents pour la détection des gestes, car les informations les plus saillantes du geste ne sont pas directement codées dans cette modalité. Bien que la linguistique mette en évidence un lien étroit entre le discours et la gestualité, le signal audio (ici F0 ou MFCC) ne transmet essentiellement que des informations acoustiques, sans fournir d'indication directe et univoque des frontières cinématiques du geste. De plus, la relation entre l'acoustique et la motricité est fortement non-linéaire : un même profil intonatif peut être produit avec ou sans geste. Cela rend la tâche de segmentation temporelle au niveau de la *frame* particulièrement complexe. En conséquence, la distribution des *features* extraites ne présente pas de corrélation suffisamment forte avec les labels de geste pour permettre au réseau de discriminer efficacement les *frames* gestes et non-gestes. Ce signal contient une importante variabilité non liée directement au geste (intonations où il n'y a pas de gestes associés, pauses, bruits de fond, etc.) qui introduit du bruit dans le signal et complique la convergence vers des représentations généralisables.

Les quaternions, en revanche, subissent une dégradation moins marquée que les données audios seules. En codant directement la rotation des articulateurs, ils produisent des représentations latentes saillantes pour la segmentation des gestes, potentiellement parce que ces données reflètent directement la dynamique du mouvement, contrairement aux signaux audios, dont la relation avec le mouvement reste indirecte.

On observe également un sous-apprentissage pour tous les modèles multimodaux testés en comparaison. Cette limitation peut s'expliquer par la combinaison de plusieurs facteurs liés à la taille du corpus. Dans le cadre des modèles multimodaux, les représentations latentes issues de chaque encodeur sont concaténées pour être transmises au décodeur, créant ainsi un vecteur d'entrée de grande dimension. Cette approche, bien que potentiellement riche en informations, pose plusieurs difficultés lorsqu'elle est appliquée sur un corpus de taille limitée. L'augmentation de la dimension effective de l'entrée du décodeur entraîne une augmentation implicite du nombre de paramètres que le réseau doit ajuster, notamment dans les couches du décodeur qui doivent combiner toutes les dimensions latentes. Avec peu de données, ces paramètres ne

peuvent pas être correctement estimés, ce qui se traduit par un apprentissage incomplet : certaines dimensions latentes ne sont pas exploitées, ou le signal qu'elles apportent est noyé dans le bruit. L'échec relatif des modèles intégrant l'audio (F0, MFCC) ne saurait être interprété comme une absence d'informations pertinentes dans la modalité acoustique au regard de la gestualité. Il est probable que la méthode de fusion tardive employée, fondée sur une simple concaténation des espaces latents issus du mouvement et de l'audio avant le décodeur, soit à l'origine de ces performances limitées. L'alignement de flux de nature hétérogène, tels que la cinématique 3D et les descripteurs acoustiques, requiert généralement des mécanismes de fusion plus sophistiqués, notamment des dispositifs de type *cross-attention*. Or, la taille restreinte de notre corpus ne permettait pas d'entraîner de tels mécanismes dans des conditions suffisamment stables.

Toutefois, bien que ces modèles présentent une performance inférieure au modèle *baseline*, l'utilisation de représentations du mouvement en quaternions et l'intégration multimodale restent prometteuses notamment pour leur capacité à fournir un signal de supervision pertinent pour un modèle de segmentation de gestes (F1-Score entre 0,55-0,63).

# 6 Conclusion de chapitre

La détection automatique des gestes constitue un enjeu central pour l'enrichissement de corpus multimodaux de grande taille. L'annotation experte des gestes étant particulièrement coûteuse en temps et en ressources, le développement d'outils capables d'automatiser ce processus, au moins partiellement, représente une étape clé pour faciliter l'exploitation de données à large échelle. Ainsi, l'un des défis de cette thèse consiste à concilier la production d'un modèle de segmentation semi-automatique fiable avec la limitation du coût de l'annotation manuelle nécessaire à la supervision du modèle. Pour répondre à cet objectif, nous avons développé COSMOS, un modèle semi-automatique dédié à la segmentation des gestes (en tant que modèle « compagnon »), reposant sur une architecture CNN-GRU légère et peu profonde. L'analyse des métriques de classification et des segmentations produites indique que le modèle détecte les gestes de manière fiable. Bien que ces prédictions ne reproduisent pas exactement la segmentation experte, l'essentiel des *frames* correspondant aux gestes est correctement identifié. Cette approche s'inscrit pleinement dans un cadre semi-automatique reposant sur le principe du *human-in-the-loop* : l'objectif n'est pas de remplacer l'expertise humaine, mais de proposer un outil d'aide à l'annotation, capable de produire des segmentations préliminaires cohérentes, susceptibles d'être validées ou ajustées par un annotateur expert. Cette approche vise ainsi à réduire significativement le coût temporel de l'annotation manuelle tout en produisant des segmentations automatiques dont la précision et la cohérence sont comparables à celles d'un annotateur expert.

Il convient toutefois de souligner que ce modèle a été entraîné sur un seul locuteur et que ses performances sont donc limitées à ce contexte spécifique. Il ne s'agit pas d'un outil directement généralisable à d'autres locuteurs. L'intérêt principal de cette expérience réside plutôt dans la démonstration du principe : il est possible de former un modèle de segmentation fiable à partir d'un volume restreint d'annotations expertes, ce qui ouvre la voie à l'extension

du paradigme à des corpus plus larges. Dans cette continuité, un autre axe à explorer concerne la généralisation à plusieurs locuteurs. Le modèle *baseline*, entraîné sur un seul locuteur, présente un sous-apprentissage lorsqu'il est testé sur un nouveau locuteur, révélant un effet de *domain shift* : les représentations apprises ne sont pas entièrement généralisables au-delà du style gestuel spécifique de l'individu d'entraînement. Il reste à déterminer si un entraînement multi-locuteur nécessite un volume de données équivalent pour chaque individu, ou si un *fine-tuning* avec un petit nombre de données d'un nouveau locuteur suffirait pour adapter le modèle. Cette étape est essentielle pour que le modèle apprenne des représentations robustes et généralisables, capables de capturer les motifs gestuels indépendamment du style individuel, et pour que les analyses associées à ces représentations puissent porter sur la nature du geste lui-même plutôt que sur les particularités d'un locuteur unique.

D'un point de vue technique, l'apprentissage de COSMOS s'est uniquement appuyé sur les positions 3D des articulations. Comme première investigation, nous avons également testé d'autres modalités, telles que l'audio, ainsi que, bien que non présentés dans ce manuscrit, des *embeddings* BERT de la transcription du discours ou des affiliés lexicaux des gestes. Dans le cas des modèles incluant le texte, ces expérimentations ont montré que la taille des vecteurs de représentation rendait le modèle trop volumineux et complexe à entraîner compte tenu de la taille limitée du corpus et de l'architecture choisie. De plus, même si les performances des modèles multimodaux testés révèlent un sous-apprentissage par rapport au modèle *baseline*, leur F1-Score reste néanmoins relativement satisfaisants. Ces observations mettent en évidence le potentiel des différentes modalités et justifient des investigations plus approfondies. Étant donné que le geste se produit dans un contexte multimodal, il est pertinent d'étudier des architectures capables d'intégrer plusieurs modalités simultanément. Une approche multimodale permettrait également d'adopter une démarche exploratoire des représentations internes : en analysant comment le réseau intègre et pondère les informations issues de chaque modalité pour extraire les *features* discriminantes du geste, on pourrait ainsi mieux comprendre la contribution relative de chaque modalité et, plus largement, la nature multimodale du geste. De la même manière, il serait intéressant de poursuivre les investigations sur différentes manières de représenter ces modalités. Dans cette étude, nous avons utilisé les positions 3D et testé les quaternions, mais d'autres représentations, notamment l'utilisation de vecteurs 6D pour coder les rotations des articulations [68], pourraient être explorées. Ces approches permettraient de capturer de manière plus riche la dynamique gestuelle.

Un autre axe d'amélioration pour COSMOS concerne les architectures utilisées. Il serait intéressant d'intégrer une prise en compte explicite de la structure du corps, afin de respecter les contraintes anatomiques et les relations entre articulations. Les *Graph Neural Networks* (GNN) [111] permettent de représenter le squelette humain comme un graphe, où chaque articulation est un nœud et les liens représentent des connexions anatomiques ou fonctionnelles. Dans ce contexte, les ST-GCN (*Spatial-Temporal Graph Convolutional Networks*) constituent également une extension prometteuse, car elles permettent de modéliser à la fois les relations spatiales entre articulations et la dynamique temporelle des gestes. Ces approches présentent plusieurs avantages : en codant explicitement la structure corporelle dans le modèle, elles

permettent d'intégrer simultanément la topologie du squelette et sa dynamique temporelle, tout en conservant cette information tout au long de l'apprentissage. Les représentations latentes restent ainsi cohérentes avec la physique du corps, permettant de suivre les poids des représentations associées à chaque articulation et d'identifier les segments anatomiques saillants du geste, ce qui facilite l'interprétation de ce que le modèle encode.

Le chapitre met donc en évidence la capacité d'un réseau de neurones à segmenter un flux de mouvement, à apprendre des représentations gestuelles fiables et, par conséquent, à proposer des segmentations cohérentes sur la base d'annotations expertes comme données supervisées, même en quantité restreinte. Ces résultats illustrent pleinement le fruit d'une démarche interdisciplinaire, combinant expertise en sciences du langage et en sciences computationnelles. Ainsi, ces résultats ouvrent la voie à des applications concrètes, notamment pour l'enrichissement semi-automatique de corpus multimodaux, ainsi qu'à des investigations futures variées, portant sur la généralisation multi-locuteurs et l'exploitation de représentations multimodales.

# Bilan de partie

L'émergence de nouveaux paradigmes méthodologiques, combinée au développement d'outils d'analyse, de traitement et de représentation du mouvement, renouvelle profondément l'étude de la gestualité. La capture de mouvement illustre particulièrement cette évolution : en fournissant des représentations tridimensionnelles précises du squelette et des articulations, elle permet une représentation directe du mouvement lui-même, sous forme de données géométriques et cinématiques exploitables. Ces représentations constituent une base sur laquelle les traitements automatiques peuvent s'appuyer : loin de se limiter à l'extraction de mesures, ces paradigmes offrent des cadres permettant de formaliser et de structurer les données, et d'en exploiter la complexité multidimensionnelle. En s'appuyant sur des méthodes computationnelles capables de traiter de grands volumes de données, il devient possible d'identifier des régularités, des corrélations ou des structures dynamiques difficilement perceptibles ou généralisables par l'observation directe. Elles rendent possible la modélisation de relations complexes entre différentes dimensions du geste en intégrant ces paramètres au sein de représentations unifiées et manipulables. Cette capacité à formaliser et à quantifier des interdépendances ne se limite pas à une meilleure compréhension des mécanismes sous-jacents à la production gestuelle ; elle ouvre également la voie à des applications en modélisation prédictive, et en génération automatique de gestes. Ainsi, les traitements automatiques ne se réduisent pas à une amélioration instrumentale des analyses : ils instaurent un cadre permettant de formaliser des relations complexes, d'en explorer les dynamiques à grande échelle et d'ouvrir des perspectives applicatives variées. Ces nouveaux moyens ne se substituent pas à l'expertise analytique, mais élargissent considérablement les possibilités d'exploration, de validation et de généralisation des phénomènes gestuels.

L'enrichissement des corpus multimodaux constitue un préalable indispensable pour caractériser avec précision les gestes et soutenir des analyses fines et rigoureuses de la gestualité. Toutefois, dans le domaine des sciences du langage, l'annotation des gestes demeure encore majoritairement manuelle. Ce processus, exigeant en temps et en expertise, limite fortement la taille des corpus exploitables. Les paradigmes d'analyse et de traitement automatique offrent un levier pour dépasser ces limites : en exploitant à la fois les données de mouvement et l'expertise humaine (même sur un volume réduit de données pour entraîner un modèle supervisé), il devient possible de produire des annotations fiables, systématiques et reproductibles à plus grande échelle. La présente partie illustre précisément cette démarche, en

proposant une méthodologie permettant d'exploiter conjointement les données de mouvement et l'expertise humaine pour enrichir les corpus. Elle s'inscrit pleinement dans un dialogue interdisciplinaire, à l'interface des sciences du langage et de l'informatique, en adoptant une approche centrée sur les données, où l'annotation multimodale constitue le noyau méthodologique. Cette démarche permet à la fois une analyse fine des gestes et la production de ressources exploitables pour la modélisation, la prédiction et la génération de comportements gestuels.

Bien que les différentes méthodologies proposées constituent une avancée significative pour l'enrichissement de corpus de gestes, elles restent encore en cours de développement et ne couvrent pas l'ensemble des annotations qui pourraient être pertinentes à produire de manière semi-automatique ou automatique. Plusieurs axes de développement restent envisageables : affiner et diversifier les descripteurs formels pour capturer une plus grande richesse du geste, tester leur robustesse sur différents locuteurs, ou encore explorer plus systématiquement des aspects saillants du geste, tant au niveau formel que fonctionnel. Dans le cadre de l'interface de visualisation et de calcul des différents descripteurs formels, certaines fonctionnalités sont déjà opérationnelles, mais elles restent encore préliminaires, laissant de nombreuses possibilités d'extension et d'amélioration. De même, pour le modèle de segmentation semi-automatique, il serait pertinent de tester si la méthodologie adoptée peut s'appliquer à d'autres corpus, permettant ainsi d'étendre l'expertise d'annotateurs qualifiés à de grands ensembles de données, même en s'appuyant sur un volume limité de données initialement annotées pour l'entraînement supervisé de ces modèles. Ces différents axes ouvrent des perspectives variées et prometteuses pour enrichir les corpus multimodaux et approfondir la modélisation de la gestualité. Ainsi, ce travail constitue un point de départ prometteur, offrant un socle méthodologique stable tout en laissant la place à de futurs développements et améliorations.

Les paradigmes computationnels constituent aujourd'hui un levier puissant pour l'analyse du geste, offrant de nouvelles perspectives analytiques, permettant d'explorer des aspects du geste difficilement accessibles par des approches traditionnelles. Parmi ces axes, l'exploration des représentations latentes apprises par les modèles neuronaux constitue un champ particulièrement prometteur. Ces représentations internes codent des régularités présentes dans les données qui ne sont pas toujours accessibles à l'observation directe ou à l'analyse traditionnelle. L'analyse de ces représentations peuvent prendre plusieurs formes : visualisation des *embeddings* par des techniques de réduction de dimensionnalité comme l'ACP, t-SNE ou UMAP pour identifier des clusters de gestes similaires ; corrélations entre dimensions latentes et descripteurs formels extraits du geste pour interpréter les informations apprises par le modèle ; manipulation des vecteurs latents pour observer l'effet sur la prédiction ou la génération des gestes. Ces approches permettent non seulement de valider et d'expliquer les comportements des modèles, mais aussi de relier directement les informations extraites à des caractéristiques formelles et fonctionnelles des gestes. En conséquence, l'analyse des représentations latentes pourrait constituer un outil puissant pour mieux comprendre les mécanismes sous-jacents à la production gestuelle. Pour être pleinement exploitée, cette démarche doit toutefois s'inscrire dans un cadre interdisciplinaire, combinant informatique et sciences du langage. En effet, si ces

analyses offrent un accès inédit à des structures latentes et des régularités du geste, elles restent aujourd'hui largement inaccessibles aux chercheurs en sciences du langage. Une collaboration interdisciplinaire permettrait de rendre ces outils exploitables pour enrichir les corpus, tester des hypothèses sur la production des gestes et éclairer des questions fondamentales sur l'organisation et la coordination des gestes avec la parole.

Cette partie illustre clairement qu'il est possible de produire des annotations multimodales à la fois fiables et pertinentes, en s'appuyant sur des paradigmes computationnels. Ces ressources constituent, en retour, un levier pour le renforcement des performances des modèles dédiés à l'étude et au traitement automatique du geste (ex. : classification des types fonctionnels, génération de gestes). Les différents paradigmes issus de l'approche computationnelle constituent ainsi un levier méthodologique puissant, permettant d'étudier le geste à grande échelle tout en garantissant la qualité et la systématicité de l'enrichissement des corpus. En combinant exploitation des données de mouvement et expertise humaine, même à partir de volumes de données limités, cette démarche ouvre la voie à des analyses plus fines, à une meilleure reproductibilité et à des perspectives variées pour l'étude de la gestualité.

# Conclusion : Les gains d'une approche interdisciplinaire

L'étude de la gestualité suscite un intérêt croissant dans de nombreux domaines, qu'il s'agisse des sciences du langage, des sciences cognitives, des sciences du mouvement, de la psychologie ou encore de l'informatique. De nombreux travaux cherchent en effet encore à mieux comprendre les productions gestuelles et leur logique d'articulation au discours, qu'il s'agisse d'approches issues de l'analyse multimodale, telle que développées depuis plusieurs décennies dans le champ des sciences du langage, ou de méthodes computationnelles plus récentes. Malgré cet engouement, le geste demeure un phénomène dont de nombreux aspects restent encore insuffisamment compris. Les approches computationnelles ont indéniablement permis des avancées majeures : traitements automatiques de grands volumes de données, formalisations mathématiques de relations complexes, modélisation prédictive ou encore génération de gestes synthétiques. Toutefois, ces modèles semblent aujourd'hui atteindre un certain palier, notamment lorsqu'il s'agit de rendre compte de ce qu'est le geste dans toute sa complexité. De même, les analyses issues des sciences du langage offrent des descriptions riches et théoriquement rigoureuses, mais se heurtent à des limites d'échelle, de systématicité et de reproductibilité. C'est à partir de ces constats que cette thèse défend une approche résolument interdisciplinaire de l'étude du geste. L'enjeu est de tirer parti de l'articulation de ces expertises, afin de répondre aux obstacles méthodologiques identifiés au sein des différents champs envisagé isolément. La problématique centrale est donc de concevoir une méthodologie d'analyse du geste manuel co-verbal qui soit à la fois qualitative et outillée. Dans une approche fondée sur les données, la richesse des cadres théoriques et pratiques développés dans les *gesture studies* assure la fiabilité et la précision de l'analyse multimodale, tandis que les approches computationnelles ouvrent la voie à la formalisation, la systématicité et la capacité de généralisation. Cette complémentarité constitue ainsi le socle de la démarche proposée dans ce travail de recherche, dont l'objectif était de démontrer en quoi une approche interdisciplinaire peut faire progresser l'étude et la modélisation du geste.

Les différentes parties de ce manuscrit mettent ainsi en évidence la mise en œuvre concrète de l'approche interdisciplinaire défendue, notamment via les contributions scientifiques qui en découlent. La partie II de cette thèse, concentrée sur les données et les ressources qui constituent la base des analyses multimodales et des traitements automatiques du geste, a porté l'attention sur la qualité, la représentativité et la pertinence des corpus multimodaux, tant en ce qui concerne les données primaires que secondaires. L'analyse critique du corpus BEAT, présenté ici comme un exemple des limites d'un enrichissement insuffisamment qualitatif, a permis de démontrer comment vérifier la pertinence d'un enrichissement proposé via des statistiques descriptives, que l'expertise issue des *gesture studies* a ensuite permis d'interpréter pour conclure sur son manque de fiabilité et de cohérence. Dans une perspective

computationnelle, où l'objectif est notamment d'exploiter ces données pour des traitements automatiques, cette évaluation est alors précieuse : elle permet de déterminer si les données reflètent correctement les phénomènes gestuels à étudier et, par conséquent, si leur utilisation est exploitable dans le cadre d'un apprentissage automatique. Ce travail s'inscrit ainsi dans un dialogue interdisciplinaire et un partage d'expertise, approche que nous avons poursuivie dans la conception du protocole d'enrichissement de reBeat-14 présenté dans la suite de la partie II. Si ce protocole s'inscrit principalement dans les pratiques d'annotation formalisées dans le champ des *gesture studies*, il a aussi été pensé pour permettre une exploitation computationnelle des annotations multimodales qui en résultent. Ainsi, l'objectif était double, à savoir identifier et caractériser de manière précise et rigoureuse les gestes au sein du flux de mouvement, et s'assurer que ces données secondaires puissent être utilisées dans des modèles automatiques, avec le potentiel d'améliorer leurs performances dans un objectif de reconnaissance et de génération. Dès la conception du protocole, le croisement des expertises a notamment permis d'identifier et d'exploiter des outils et méthodes issus des traitements automatiques, peu, voire rarement, mobilisés en sciences du langage. Ces ressources ont ensuite été intégrées au protocole, contribuant à enrichir ou à faciliter les annotations multimodales. Parmi ces outils figurent, par exemple, la reconnaissance automatique de la parole (via l'ASR), sa segmentation et son alignement (via MFA) pour la transcription du signal audio, ainsi que le traitement automatique du langage (via spaCy) pour la catégorisation lexicale. L'approche interdisciplinaire permet ainsi le partage des connaissances et des méthodes entre disciplines, ouvrant de nouvelles possibilités pour enrichir et améliorer la conception et l'exploitation des corpus multimodaux.

La troisième partie de cette thèse examine plus spécifiquement les apports et perspectives issus des sciences du langage dans le cadre d'une démarche pluridisciplinaire. Elle met en évidence le rôle déterminant de l'annotation multimodale, envisagée non comme une simple étape descriptive, mais comme un levier méthodologique pouvant structurer les approches computationnelles dédiées à l'analyse du geste. L'analyse multimodale des annotations fonctionnelles et formelles du corpus reBeat-14 met ainsi en évidence leur capacité à rendre compte de premières propriétés constitutives du geste. L'annotation multimodale experte propose ainsi une modélisation fine et structurée du phénomène gestuel, constituant une formalisation qui dépasse la simple abstraction quantitative en s'appuyant sur une lecture contextualisée et interprétative, que les modèles computationnels actuels ne sont pas en mesure de reproduire sans supervision experte. Cette modélisation permet également de caractériser les gestes générés automatiquement, comme le montre l'analyse comparative des productions du modèle STARGATE, et de mettre en évidence les dimensions du geste correctement reproduites par le modèle, ainsi que celles qui restent insuffisamment capturées. Cette modélisation a conduit à la définition de critères d'évaluation à la fois précis et objectifs des gestes de synthèse, en complément des évaluations perceptives. La modélisation du geste, telle que proposée dans le cadre des *gesture studies*, devrait donc constituer un référentiel pour les approches computationnelles. En effet, en identifiant précisément les aspects les plus saillants du geste, cette modélisation experte fournit un cadre permettant d'orienter la conception et

l'optimisation des systèmes de traitement automatique, afin de reproduire plus fidèlement les caractéristiques du geste humain.

De manière réciproque, la quatrième et dernière partie de ce manuscrit se concentre sur les apports et perspectives issus des sciences computationnelles, en mettant particulièrement l'accent sur les nouvelles dimensions et informations qu'il devient possible d'extraire en traitant automatiquement les données multimodales. Les avancées dans la capture de mouvement ont considérablement enrichi les formes de représentation numériques des gestes, en offrant un accès direct à leurs dimensions cinématiques et dynamiques à chaque instant. La richesse de cette représentation tridimensionnelle permet ainsi de produire des mesures quantitatives précises et exploitables. Le traitement automatique de ces données, tel que mis en œuvre dans le cadre de la conceptualisation du *framework* d'annotation automatique des aspects formels du geste (et son interface dédiées GUI), a conduit au développement d'une méthode objective, cohérente et fiable pour l'enrichissement formel. Il devient ainsi possible de s'affranchir des limites majeures de l'annotation formelle manuelle basée sur l'observation vidéo, souvent restreinte par la perception et la subjectivité humaine. La méthodologie de segmentation automatique du geste développée dans le cadre de COSMOS met également en évidence la capacité des modèles à produire des prédictions robustes, directement fondées sur l'annotation experte. Ces contributions constituent un apport majeur pour l'enrichissement systématique des corpus, en permettant de produire des annotations reproductibles en plus grandes quantités, sans perte de qualité du point de vue de l'expertise des *gesture studies*. L'accroissement des données enrichies décuple les possibilités d'étude et d'analyse au plus long terme. Il devient possible d'affiner nos connaissances sur les caractéristiques formelles et fonctionnelles du geste, tout en alimentant des modèles prédictifs ou génératifs. Ces avancées ouvrent la voie au développement d'outils automatiques plus sophistiqués, tels que la détection des types fonctionnels ou d'autres dimensions du geste, ainsi qu'à des modèles de génération capables de reproduire la complexité de la gestuelle co-verbale en intégrant l'ensemble de ses aspects saillants, jusqu'ici inaccessibles par les systèmes existants. L'intégration de ces aspects permet également d'étudier les représentations générées par ces modèles, d'évaluer leur fonctionnement et de tester des hypothèses concernant la coordination geste-parole ou la formalisation des gestes. En ce sens, l'enrichissement systématique et automatique des données, rendu possible par l'exploitation des paradigmes de traitement et d'analyse issus de l'approche computationnelle, dépasse largement le simple gain quantitatif ; il ouvre un véritable champ des possibles, jusque-là limité par la disponibilité des corpus multimodaux enrichis de manière fiable et précise, tant en sciences du langage qu'en sciences computationnelles.

L'ensemble des éléments présentés précédemment met en évidence les gains substantiels qu'offre une approche interdisciplinaire pour l'analyse et la modélisation du geste. En articulant, des cadres théoriques, des méthodologies et des outils issus de champs disciplinaires distincts, cette démarche permet de saisir la complexité du geste dans ses dimensions à la fois formelles, fonctionnelles et dynamiques. Cette approche interdisciplinaire bénéficie à l'ensemble des acteurs impliqués dans l'étude de la gestualité, en enrichissant les analyses et en favorisant les perspectives méthodologiques pour chaque discipline. Néanmoins, la mise en œuvre d'une telle

approche sur l'ensemble du doctorat a également représenté un défi substantiel. Elle a impliqué un effort constant de clarification conceptuelle, de traduction terminologique et d'ajustement méthodologiques, dans la mesure où les disciplines mobilisées ne partagent ni exactement les mêmes finalités, ni parfois les mêmes définitions opératoires, ni les mêmes contraintes pratiques. La construction d'un cadre commun intelligible pour l'ensemble des interlocuteurs a ainsi constitué un enjeu central, à la fois méthodologique et scientifique. Si ces efforts ont pu constituer une difficulté réelle tout au long de ces travaux de recherche, ils ont ainsi contribué à structurer une dynamique de recherche particulièrement riche. La confrontation des approches ne s'est pas limitée à une mise en compatibilité des cadres existants, mais a généré de nouvelles questions de recherches, de nouvelles articulations méthodologiques et de nombreuses pistes d'analyse qui restent encore à explorer.

L'approche interdisciplinaire adoptée ne se limite pas au cadre strict de cette thèse ; elle apparaît aujourd'hui comme non seulement nécessaire, mais également de plus en plus reconnue et explicitement encouragée au sein de la communauté scientifique consacrée à l'étude de la gestualité. Des instances, telles que l'*International Society for Gesture studies* (ISGS), témoignent ainsi d'un intérêt croissant pour les approches computationnelles, à travers l'organisation de sessions dédiées à la modélisation automatique, à la génération de gestes ou à l'exploitation de nouvelles technologies. Lors de la dixième édition, qui s'est tenue en 2025 sous le thème *Motion to Meaning : Innovations in Multimodal Language Research*, une session consacrée au *Machine Learning/AI/VR*, dirigée par Essam Ghaleb (Max Planck Institute, AI researcher), a ainsi été organisée, illustrant concrètement l'ouverture de la communauté des *gesture studies* aux méthodes computationnelles. Une partie des travaux développés dans le cadre de cette thèse y a notamment été présentée. À l'échelle nationale, les rencontres de l'ISGS France participent également à cette dynamique au travers des journées d'étude en favorisant le partage de protocoles d'annotation, de méthodologies émergentes et de cadres théoriques entre linguistes, spécialistes du geste et chercheurs de différents domaines. Il est par ailleurs prévu que certaines des nouvelles approches développées dans la présente recherche y soient présentées, en lien notamment avec les contributions méthodologiques apportées au projet *Syncogest* dans le cadre duquel s'inscrit cette thèse. Ce projet constitue d'ailleurs un exemple concret de démarche interdisciplinaire pour l'étude du geste, et son financement récent par l'ANR témoigne de l'intérêt croissant porté à de telles démarches mobilisant plusieurs expertises complémentaires. Parallèlement, des événements centrés sur les approches computationnelles, tels que le workshop GENEA, illustrent une ouverture réciproque vers les sciences du langage et les traditions analytiques qualitatives. L'invitation de chercheuses issues des *gesture studies* comme conférencières invitées – par exemple lors de l'édition 2025 avec la présence d'Aslı Özyürek, directrice du *Multimodal Language Department* de l'Institut Max Planck de psycholinguistique de Nimègues (Pays-Bas) – souligne cette volonté croissante de dialogue entre communautés historiquement distinctes. Ces évolutions institutionnelles et communautaires mettent en évidence l'émergence d'un espace de recherche véritablement transversal, dans lequel la confrontation des cadres théoriques, des pratiques d'annotation et des méthodes de

modélisation apparaît comme un levier central pour progresser dans la compréhension et la simulation du geste co-verbal.

L'ensemble des travaux réalisés au cours de cette thèse constitue une contribution notable à l'étude de la gestualité, tant sur le plan théorique que méthodologique, offrant des bases solides pour approfondir l'analyse, la modélisation et la compréhension des gestes humains. La démarche interdisciplinaire, combinant sciences du langage et sciences computationnelles et telle qu'appliquée dans ce travail, ouvre ainsi un large éventail de perspectives pour la recherche future dans ce domaine.

# Annexes

## 1 Annexe C8-S5-gui : Architecture logicielle et manuel d'utilisation de GUI

### 1.1 Architecture fonctionnelle

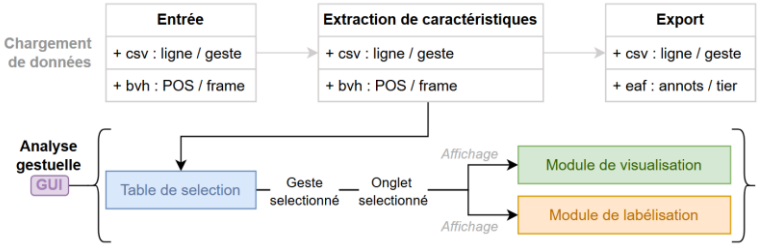



*Figure 62. Architecture fonctionnelle conceptuelle pour le développement du prototype*

La Figure 62 illustre l'architecture fonctionnelle réfléchie du *pipeline* d'analyse gestuelle, composée de plusieurs modules, chacun remplissant un rôle technique spécifique.

- **Chargement de données** : l'utilisateur doit sélectionner un fichier (ou un dossier de fichiers) nécessaire au fonctionnement de l'interface. GUI fonctionne à partir de fichiers annotés et prend en compte spécifiquement les segmentations de gestes, les données de mouvement correspondantes, ainsi que les annotations (qui ne sont pas formelles) réalisées concernant le geste. Pour l'instant, cette interface ne fonctionne donc que pour les fichiers tels qu'annotés dans reBeat-14. Les annotations manuelles des aspects formels dans reBeat-14 ne sont toutefois pas réutilisées dans l'interface, qui a pour but d'objectiver la qualification de ces aspects selon la conceptualisation présentée précédemment dans ce chapitre.

- **Extraction de caractéristiques** : les informations d'annotation et de mouvement sont extraites des fichiers chargés. Les métadonnées associées à chaque geste (identifiant, fichier d'enregistrement d'origine, temps de début et de fin, annotations fonctionnelles, et lien au discours) sont d'abord récupérées depuis le fichier d'annotation. Ces informations permettent ensuite d'extraire, dans les fichiers de mouvement correspondants, les segments de données associés à chaque geste, notamment les positions 3D sur l'intervalle temporel concerné.

- **Analyse formelle automatique** : l'ensemble des gestes chargés est analysé en arrière-plan selon les procédures d'analyse conceptualisés dans le *framework* présenté dans ce chapitre pour la catégorisation des aspects formels du geste (la localisation, la dynamique et l'orientation du geste, la configuration et l'orientation de la main, l'orientation de la paume, et la configuration des doigts). Chaque analyse automatique produite peut ensuite être consultée via différents onglets (présentés plus loin), chacun étant constitué de modules de visualisation et de labélisation automatique des aspects formels du geste. Ces modules

constituent le cœur du *pipeline* de GUI, conçu pour une annotation automatique, une validation des méthodologies et une visualisation interactive des gestes analysés.

- **Export de fichiers** : À l'issue des analyses, l'export des annotations automatiques générées pour les gestes traités dans l'interface (notamment aux formats `.eaf` ou `.csv`) est envisagé. Cette fonctionnalité n'est cependant pas implémentée à ce stade.

## 1.2 Prototype

### 1.2.1 Onglet « Home »

L'interaction initiale avec l'interface s'effectue via l'onglet « Home » (Figure 63), à partir duquel l'utilisateur peut charger les fichiers dans lesquels se trouve les gestes à analyser. Le chargement de ces fichiers déclenche l'extraction des métadonnées et des données de mouvement, puis l'exécution des analyses formelles automatiques, sur la base des annotations proposées dans le *framework*. Une fois cette opération effectuée, le tableau interactif de cet onglet est mis à jour pour afficher les métadonnées extraites pour chaque geste analysé par le système. Ces métadonnées comprennent : l'identifiant du geste, les bornes temporelles (en secondes), sa durée (en seconde), sa phase gestuelle, son type fonctionnel (en première et en seconde dimension), et son affiliation lexicale.

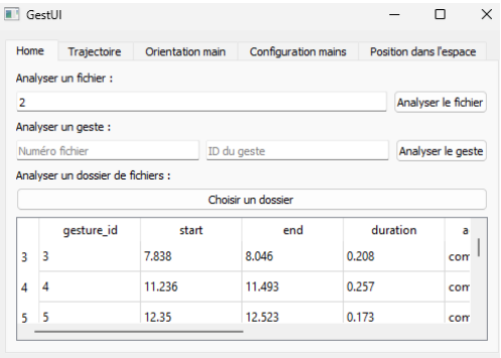



*Figure 63. Onglet Home et table de sélection de geste*

La sélection d'un geste au sein de ce tableau geste déclenche alors la mise à jour dynamique du contenu des différents onglets situés en haut de la fenêtre, par lesquels sont accessibles les visualisations et annotations automatiques produites pour ce geste. Ces onglets sont intitulés « Trajectoire », « Orientation de la main », « Configuration de la main » et « Position dans l'espace » et sont détaillés dans la suite de cette section.

### 1.2.2 Onglet « Trajectoire »

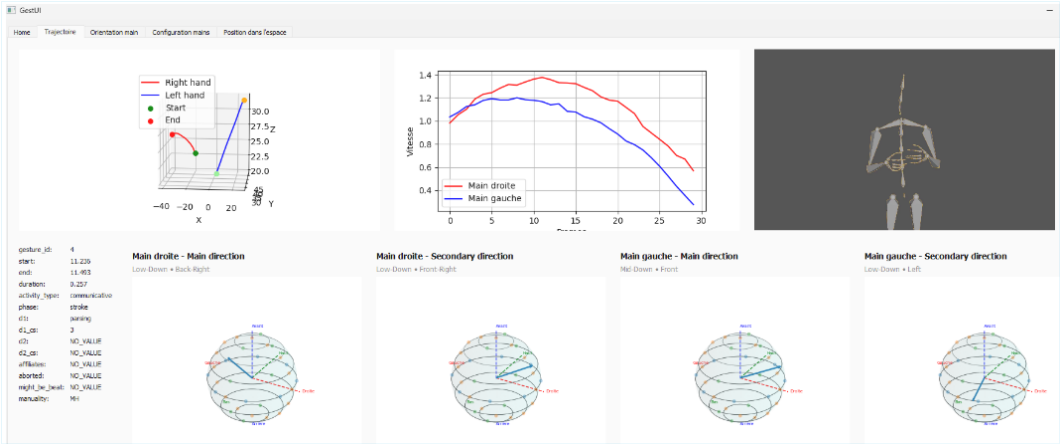



*Figure 64. Onglet Trajectoire*

L'onglet « Trajectoire » permet de visualiser et d'analyser les trajectoires des mains, plus précisément de l'extrémité de chaque index, au cours du geste sélectionné. Il se compose de deux sections (Figure 64) : la partie supérieure regroupe la vidéo et les trajectoires, tandis que la partie inférieure présente les métadonnées du geste ainsi que les sphères de catégorisation de sa direction.

La partie supérieure est composée de trois éléments : (1) un graphique interactif 3D (`TrajectoryWidget`) des trajectoires de l'index droit (en rouge) et de l'index gauche (en bleu), (2) un graphique de la vitesse instantanée des index au cours de l'exécution du geste (`SpeedWidget`), et (3) la vidéo du geste sur l'ECA (`VideoWidget`). Ces éléments permettent une première analyse de la dynamique et une visualisation du geste analysé.

La partie inférieure est composée d'un tableau récapitulatif des métadonnées du geste et de quatre sphères interactives (`SphereWidget`), représentant les projections des vecteurs d'orientation de la trajectoire de chaque index. Au-dessus de chaque sphère sont affichées les étiquettes correspondant aux orientations principale et secondaire du geste, déterminées automatiquement sur la base de la discrétisation de la sphère 3D proposée dans le *framework*.

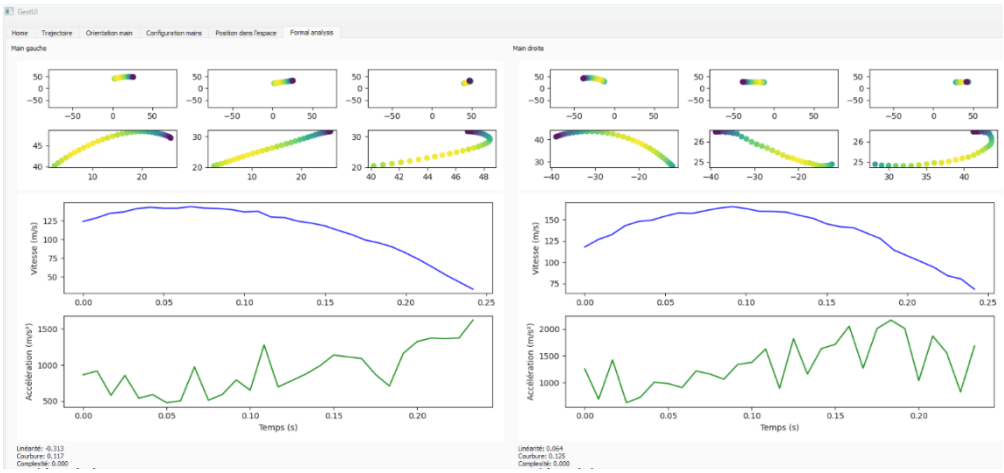



*Figure 65. Onglet Formal analysis*

Depuis l'onglet « Trajectoire », il est possible de d'accéder à l'onglet « Formal analysis » (Figure 65). Cet onglet fourni des visualisations et analyses complémentaires à celles proposées dans l'onglet « Trajectoire ». Dans la partie supérieure de l'onglet, les trajectoires des index sont représentées selon trois vues planes :

- **Vue frontale (plan XY)** : montre la trajectoire de face, en mettant en évidence les déplacements latéraux et verticaux.
- **Vue horizontale (plan XZ)** : montre la trajectoire vue du dessus, mettant en évidence les déplacements avant-arrière et latéraux.
- **Vue sagittale (plan YZ)** : montre la trajectoire de côté, mettant en évidence les déplacements verticaux et avant-arrière.

Les `TrajectoryWidget` permet d'explorer librement les trajectoires dans un espace tridimensionnel interactif, en faisant pivoter et déplacer la vue pour les examiner sous différents angles. Cependant, il peut être difficile de se positionner exactement sur une vue spécifique. Les vues planes fixes de l'onglet « Formal analysis » viennent alors compléter cette exploration en offrant des points de vue standardisés, permettant de visualiser précisément la forme de la trajectoire selon chaque plan. L'onglet propose également des graphiques représentant l'accélération des index au cours du temps, complétant ainsi le graphique de vitesse présent dans l'onglet « Trajectoire ».

### 1.2.3 Onglet « Orientation main »

L'onglet « Orientation main » permet de déterminer l'orientation de chaque main et chaque paume lors de la dernière *frame* du geste sélectionné, quel que soit les mains impliquées dans le geste. L'onglet comprend quatre `SphereWidget`. Pour chaque main, une sphère permet de visualiser la projection du vecteur d'orientation de l'index, et une autre celle du vecteur d'orientation de la paume. Au-dessus de chaque sphère, les orientations sont indiquées automatiquement, déterminées à partir de la discrétisation de la sphère 3D.

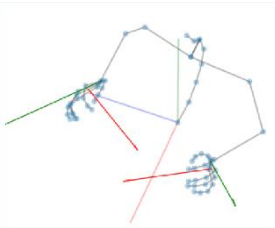


*Figure 66. `SkeletonWidget` pour l'affichage des vecteurs d'orientation de la main (vert) et de la paume (rouge)*

L'onglet comprend également un modèle interactif du squelette du haut du corps (`SkeletonWidget`, Figure 66), permettant de visualiser la posture et l'orientation des mains du locuteur. Les vecteurs calculés automatiquement pour la détermination des orientations des mains (en vert) et des paumes (en rouge) sont également affichés dans cette représentation

### 1.2.4 Onglet « Configuration mains »

L'onglet « Configuration mains » est dédié à la visualisation et à l'analyse de la configuration des doigts, là encore lors de la dernière *frame* du geste uniquement (Figure 67).




*Figure 67. Onglet Configuration mains*

Cet onglet comprend deux `SkeletonWidget`, un pour chaque main, quel que soit le ou les mains impliquées dans le geste. Cette représentation permet une visualisation directe de la configuration des mains. Plusieurs boutons sont également intégrés à l'onglet, permettant d'afficher ou de masquer, dans le `SkeletonWidget`, les vecteurs (représenté en rouges et en bleus dans la Figure 67)permettant de déterminer les différentes qualifications d'aspects relatifs à la configuration des doigts tels que conceptualisés dans le *framework*. Pour chaque doigt, trois vecteurs sont calculés entre les marqueurs d'intérêt et utilisés pour le calcul des distances, déterminant alors l'écartement relatif (*spacing*), le degré de fermeture (*closure)* et la courbe de chaque doigt (*curvature*).

### 1.2.5 Onglet « Position dans l'espace »

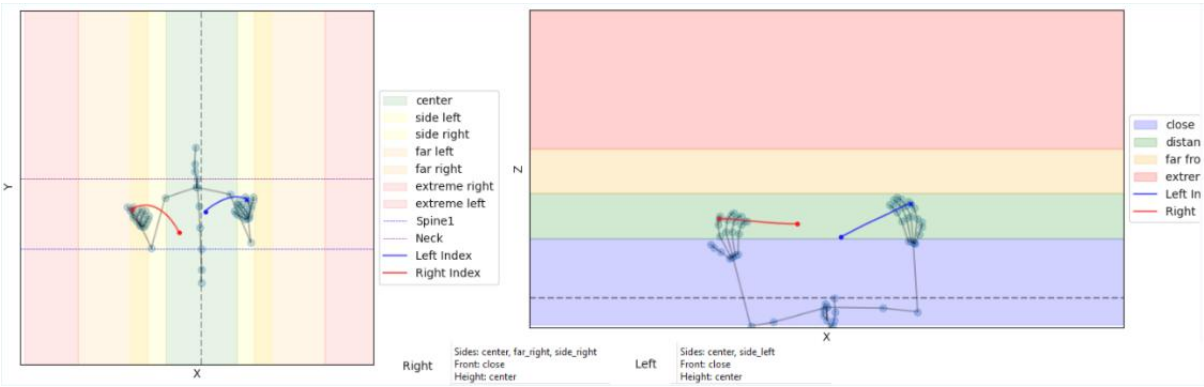



*Figure 68. Onglet Position dans l'espace*

L'onglet « Position dans l'espace » (Figure 68) permet d'analyser et de déterminer la position de l'extrémité des index dans l'espace gestuel au cours de l'exécution du geste sélectionné, quel que soit la ou les mains impliquées. Cet onglet comprend deux `PosWidget` (vues fixes affichant la trajectoire des index) : le premier présente la vue frontale (de face), et le second la vue horizontale (du dessus) de l'espace gestuel discrétisé, tel que conceptualisé dans le cadre du *framework*. Ces deux vues offrent une visualisation claire de l'utilisation de l'espace par le locuteur dans les trois dimensions : la latéralité, la verticalité et la profondeur. Pour chaque main, les étiquettes des zones traversées sont déterminées automatiquement pour ces trois dimensions, à partir de la détection des positions 3D dans cet espace discrétisé, à chaque instant du geste.

## 1.3 Axes de développement pour la diffusion initiale

### 1.3.1 Documentation

Une documentation complète est indispensable pour accompagner la diffusion initiale du logiciel. Elle doit inclure un guide utilisateur détaillant les principales fonctionnalités, des tutoriels pas à pas (éventuellement sous forme de vidéos), des exemples de fichiers d'entrée et de sortie, ainsi qu'un `README` décrivant l'installation, les dépendances et les usages courants. La mise en place d'un dépôt GitHub permettra également de centraliser le code, de signaler les bugs et de proposer des améliorations, favorisant ainsi une appropriation progressive de l'outil par la communauté.

### 1.3.2 Gestion de la standardisation des annotations chargées

À l'heure actuelle, le chargement initial des données ne peut s'effectuer que pour des fichiers d'annotation respectant le *template* d'enrichissement spécifique proposé dans le cadre de cette thèse. Afin d'élargir les possibilités d'utilisation du logiciel, celui-ci devra être en mesure de charger des annotations issues de différentes pratiques ou conventions, notamment celles produites dans ELAN. Une attention particulière devra être portée à la standardisation des champs (noms de tiers, types fonctionnels, phases gestuelles, dimensions annotées, etc.), afin de garantir la compatibilité avec les *widgets* d'analyse présentés précédemment et d'éviter les incohérences lors du chargement. Pour faciliter cette intégration, des mécanismes de validation ou de normalisation automatique des annotations pourraient être mis en place. En complément, le logiciel pourrait proposer aux utilisateurs un *template* de chargement configurable, à adapter en amont aux fichiers d'annotation à analyser.

### 1.3.3 Chargement de données 3D

Actuellement, l'analyse repose sur des fichiers `.bvh`, dont le squelette correspond à la configuration utilisée lors de la captation des données par les auteurs de BEAT. À terme, le logiciel pourrait prendre en charge d'autres sources de données 3D, telles que des *keypoints* générés par MediaPipe ou par d'autres systèmes de capture. Une telle extension nécessiterait la définition explicite d'un *mapping* du squelette afin d'assurer la correspondance avec les articulations attendues par les *widgets* existants. Le bon fonctionnement des fonctionnalités actuelles – notamment le chargement des données, la sélection et la visualisation des gestes, ainsi que la mise à jour dynamique des onglets – devra être systématiquement vérifié lors de l'ajout de nouveaux formats de données 3D.

### 1.3.4 Ajout de fonctionnalité d'export d'annotation et de fichiers

À l'heure actuelle, le logiciel est principalement orienté vers la visualisation et l'exploration des annotations existantes, bien qu'il intègre déjà des mécanismes de calcul et de catégorisation. L'ajout de fonctionnalités d'export des données analysées constitue ainsi un axe essentiel pour permettre la production et la diffusion d'annotations formelles automatiques des

gestes, fondées sur les calculs réalisés par le système. Les annotations et résultats devront pouvoir être exportés dans des formats standards tels que `CSV`, `EAF` ou `JSON`, afin de permettre leur intégration ou leur validation dans des outils d'annotation comme ELAN, ou leur réutilisation dans des *pipelines* d'analyse ultérieurs. Cette évolution renforcerait l'interopérabilité du logiciel et assurerait la pérennité des données produites.

### 1.3.5 Ajout de fonctionnalités d'analyses

Au-delà de l'analyse formelle, il est également envisageable d'étendre l'interface avec des méthodologies adaptées aux besoins spécifiques des utilisateurs. Par exemple, l'intégration d'outils de segmentation semi-automatique, paramétrables directement sur les données propres à chaque corpus (en tout cas aux données chargées), permettrait d'adapter finement les algorithmes aux styles gestuels, aux types de capteurs ou aux objectifs d'annotation. Enfin, le logiciel est conçu pour pouvoir intégrer de nouvelles couches d'analyse dans la poursuite de nos travaux, dès lors que celles-ci auront été suffisamment conceptualisées et formalisées. D'autres analyses automatiques des différents aspects du geste, telles que sa fonction, son affiliation lexicale ou d'autres niveaux descriptifs, pourront ainsi être progressivement intégrées, au fur et à mesure de l'avancement des cadres théoriques et des affinements méthodologiques. Cette approche garantit l'évolutivité du système et son adéquation avec les besoins émergents de l'analyse multimodale du geste.

# Bibliographie


[1] W. Wundt, The Language of Gestures, Berlin, New: De Gruyter Mouton, 1973.

[2] F. E. X. Dance, «The “Concept” of Communication,» *Journal of Communication,* vol. 20, pp. 201-210, 1970.

[3] M. Bullowa, Before speech, Cambridge, England: Cambridge University Press, 1979.

[4] H. Frings et M. Frings, «Animal communication. New York. Toronto, London, Blaisdell, 1964,» *Revue d'Écologie (La Terre et La Vie),* vol. 19, p. 400–401, 1965.

[5] M. Argyle, «Non-verbal communication in human social interaction,» chez *Non-Verbal Communication*, R. Hinde, Éd., Cambridge, Cambridge University Press, 1972, p. 243–270.

[6] Y. Winkin and Y. Winkin, La nouvelle communication, 1981.

[7] A. Kendon, Gesture and Speech, Cambridge: Cambridge University Press, 1980.

[8] A. Kendon, Gesture: Visible Action as Utterance, Cambridge ; New York: Cambridge University Press, 2004.

[9] D. McNeill, Hand and Mind: What Gestures Reveal about Thought, Chicago and London: The University of Chicago Press, 1992.

[10] D. McNeill, Gesture and Thought, Chicago and London: The University of Chicago Press, 2005.

[11] G. Calbris, Elements of Meaning in Gesture, M. M. Copple, Éd., John Benjamins Publishing Company, 2011.

[12] I. Mittelberg, «Methodology for multimodality: One way of working with speech and gesture data,» chez *Methods in Cognitive Linguistics*, John Benjamins Publishing Company, 2007, p. 225–248.

[13] C. Debras, «Petits et grands corpus en analyse linguistique des gestes,» *Corpus,* July 2018.

[14] M. Tellier, «Quelques orientations méthodologiques pour étudier la gestuelle dans des corpus spontanés et semi-contrôlés,» *Discours - Revue de linguistique, psycholinguistique et informatique,* vol. 15, 2014.

[15] *ELAN (Version 6.8) [Computer software],* Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive, 2024.

[16] P. Boersma et D. Weenink, *Praat: doing phonetics by computer,* 2024.

[17] I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio et B. De Carolis, «Greta. a believable embodied conversational agent,» chez *Multimodal intelligent information presentation*, Springer, 2005, p. 3–25.

[18] J. Cassell, «Embodied conversational interface agents,» *Communications of the ACM,* vol. 43, p. 70–78, April 2000.

[19] J. Cassell, «Embodied Conversational Agents: Representation and Intelligence in User Interfaces,» *AI Magazine,* vol. 22, p. 67, December 2001.

[20] J. Svartvik, Ed., Directions in corpus linguistics, Berlin, Germany: De Gruyter Mouton, 1992.

[21] R. Nagy, H. Voss, Y. Yoon, T. Kucherenko, T. Nikolov, T. Hoang-Minh, R. McDonnell, S. Kopp, M. Neff et G. E. Henter, *Towards a GENEA Leaderboard – an Extended, Living Benchmark for Evaluating and Advancing Conversational Motion Synthesis,* arXiv, 2024.

[22] R. Nagy, H. Voss, T. Hoang-Minh, M. Tsakov, T. Nikolov, Z. Zhang, T. Ao, S. Yang, S. Huang, Y. Cheng, M. H. Mughal, R. Dabral, K. Chhatre, C. Theobalt, L. Liu, S. Kopp, R. McDonnell, M. Neff, T. Kucherenko, Y. Yoon et G. E. Henter, *Towards Reliable Human Evaluations in Gesture Generation: Insights from a Community-Driven State-of-the-Art Benchmark,* 2025.

[23] S. D. Kelly, A. Özyürek et E. Maris, «Two Sides of the Same Coin: Speech and Gesture Mutually Interact to Enhance Comprehension,» *Psychological Science,* vol. 21, p. 260–267, December 2009.

[24] A. Özyürek, *Role of Gesture in Language Processing: Toward a unified account for production and comprehension,* S. Rueschemeyer et M. G. Gaskell, Éds., Oxford University Press, 2018, p. 591–607.

[25] E. Morsella et R. M. Krauss, «The Role of Gestures in Spatial Working Memory and Speech,» *The American Journal of Psychology,* vol. 117, p. 411, 2004.

[26] A. B. Hostetter, M. W. Alibali et S. Kita, «I see it in my hands' eye: Representational gestures reflect conceptual demands,» *Language and Cognitive Processes,* vol. 22, p. 313–336, April 2007.

[27] W.-M. Roth, «From action to discourse: The bridging function of gestures,» *Cognitive Systems Research,* vol. 3, p. 535–554, September 2002.

[28] S. Kita, «How representational gestures help speaking,» chez *Language and Gesture*, Cambridge University Press, 2000, p. 162–185.

[29] S. Rowbotham, J. Holler, D. Lloyd et A. Wearden, «Handling pain: The semantic interplay of speech and co-speech hand gestures in the description of pain sensations,» *Speech Communication,* vol. 57, p. 244–256, February 2014.

[30] T. Jenkins, M. Coppola et C. Coelho, «Effects of gesture restriction on quality of narrative production,» *Gesture,* vol. 16, p. 416–431, December 2017.

[31] H. Holle et T. C. Gunter, «The Role of Iconic Gestures in Speech Disambiguation: ERP Evidence,» *Journal of Cognitive Neuroscience,* vol. 19, p. 1175–1192, July 2007.

[32] J. Cassell, D. McNeill et K.-E. McCullough, «Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information,» *Pragmatics and Cognition,* vol. 7, p. 1–34, 1999.

[33] H. Holle, C. Obermeier, M. Schmidt-Kassow, A. D. Friederici, J. Ward et T. C. Gunter, «Gesture Facilitates the Syntactic Analysis of Speech,» *Frontiers in Psychology,* vol. 3, 2012.

[34] H. McGurk et J. O. H. N. MacDonald, «Hearing lips and seeing voices,» *Nature,* vol. 264, p. 746–748, December 1976.

[35] P. Barkhuysen, E. Krahmer et M. Swerts, «The interplay between the auditory and visual modality for end-of-utterance detection,» *The Journal of the Acoustical Society of America,* vol. 123, p. 354–365, January 2008.

[36] P. Barkhuysen, E. Krahmer et M. Swerts, «Predicting end of utterance in multimodal and unimodal conditions,» chez *Interspeech 2005*, 2005.

[37] C. Goodwin, Conversational Organization: Interaction Between Speakers and Hearers, 1981.

[38] C. Goodwin, «Notes on story structure and the organization of participation,» *M. Atkinson & J. Heritage (Eds.), Structures of social action,* January 1984.

[39] C. Garitte, «Comparaison entre l'acquisition des gestes conventionnels et les expressions semi-figées chez l'enfant,» *Linx,* p. 71–85, December 2005.

[40] D. Efron, Gesture, race, and culture, The Hague: Mouton, 1941.

[41] P. Ekman et W. V. Friesen, «The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding,» *Semiotica,* vol. 1, p. 49–98, 1969.

[42] L. Payrató, «A pragmatic view on autonomous gestures: A first repertoire of Catalan emblems,» *Journal of Pragmatics,* vol. 19, p. 193–216, 1993.

[43] Olmos, «Étude sur les gestes en Espagne,» 1971.

[44] Badouel-Mathon, «Étude sur la gestuelle en Afrique,» 1971.

[45] M. Virolle-Souibes, «Gestes d'Algérie,» 1985.

[46] R. L. Saitz et E. J. Cervenka, «Handbook of Colombian and U.S. Gestures,» 1972.

[47] G. Calbris, «La gestuelle,» chez *Le langage du corps*, M. J. Bruneau, Éd., Paris, PUF, 1980, p. 29–46.

[48] P. Wagner, Z. Malisz et S. Kopp, «Gesture and speech in interaction: An overview,» *Speech Communication,* vol. 57, p. 209–232, February 2014.

[49] S. Kita, I. van Gijn et H. van der Hulst, «Movement phases in signs and co-speech gestures, and their transcription by human coders,» chez *Gesture and Sign Language in Human-Computer Interaction*, Berlin, 1998.

[50] S. Müller, A. Abeillé, R. D. Borsley et J.-P. Koenig, *Head-Driven Phrase Structure Grammar,* Language Science Press, 2024.

[51] D. Boutet, «Une morphologie de la gestualité : structuration articulaire,» *Cahiers de linguistique analogique,* pp. 81-115, December 2008.

[52] C. Müller, "Forms and uses of the Palm Up Open Hand: A case of a gesture family?," in *The Semantics and Pragmatics of Everyday Gestures: The Berlin Conference April 1998*, C. Müller and R. Posner, Eds., Berlin, Weidler, 2004, p. 233–256.

[53] G. Ferre, «Annotation multimodale du français parlé. Le cas des pointages,» chez *Proceedings of TALN - Atelier Degels*, Montpellier, 2011.

[54] T. Hanke, «HamNoSys—Representing sign language data in language resources and language processing contexts,» 2004.

[55] L. A. Friedman, Éd., On the other hand, New York [u.a.]: Academic Press, 1977.

[56] D. Morris, Manwatching: A Field Guide to Human Behaviour, London, England: Triad Books, 1978.

[57] G. Calbris, L'expression gestuelle de la pensée d'un homme politique, 2003.

[58] J. Florit-Pons, P. L. Rohrer, I. Vilà-Giménez, U. Tütüncübasi, N. E. Gibert, P. Ren, S. Shattuck-Hufnagel et P. Prieto, «The MultiModal MultiDimensional (M3D) labeling system,» 2020.

[59] M. Kipp, M. Neff et I. Albrecht, «An annotation scheme for conversational gestures: How to economically capture timing and form,» *Language Resources and Evaluation,* vol. 41, pp. 325-339, December 2007.

[60] E. Velloso, A. Bulling et H. Gellersen, «AutoBAP: Automatic Coding of Body Action and Posture Units from Wearable Sensors,» chez *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013.

[61] C. Fidelman, G. Calbris et O. Doyle, «The Semiotics of French Gestures,» *The Modern Language Journal,* vol. 75, p. 370, November 1992.

[62] Y. Ferstl, M. Neff et R. McDonnell, «Understanding the predictability of gesture parameters from speech and their perceptual importance,» chez *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, 2020.

[63] P. Ekman et W. W. Friesen, «Hand movements,» *The Journal of Communication,* vol. 22, p. 353–374, 1972.

[64] J. Cosnier et J. Vaysse, «Sémiotique des gestes communicatifs,» *Nouveaux actes s'emiotiques,* p. 7–28, 1997.

[65] E. A. Schegloff, «On some gestures' relation to talk,» chez *Structures of Social Action*, J. M. Atkinson, Éd., Cambridge University Press, 1985, p. 266–296.

[66] G. Ferre, Analyse de discours multimodale: Gestualité et prosodie en discours, UGA Editions, 2019, p. 360.

[67] F. S. Grassia, «Practical Parameterization of Rotations Using the Exponential Map,» *Journal of Graphics Tools,* vol. 3, p. 29–48, January 1998.

[68] Y. Zhou, C. Barnes, J. Lu, J. Yang et H. Li, *On the Continuity of Rotation Representations in Neural Networks,* arXiv, 2018.

[69] T. Kucherenko, R. Nagy, M. Neff, H. Kjellström et G. E. Henter, *Multimodal analysis of the predictability of hand-gesture properties,* arXiv, 2021.

[70] Y. Ferstl, M. Neff et R. McDonnell, «Multi-objective adversarial gesture generation,» chez *Motion, Interaction and Games*, 2019, p. 1–10.

[71] L.-P. Morency, A. Quattoni et T. Darrell, «Latent-Dynamic Discriminative Models for Continuous Gesture Recognition,» chez *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[72] F. Yunus, C. Clavel et C. Pelachaud, «Gesture Class Prediction by Recurrent Neural Network and Attention Mechanism,» chez *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, New York, NY, USA, 2019.

[73] E. Ghaleb, I. Burenko, M. Rasenberg, W. Pouw, P. Uhrig, J. Holler, I. Toni, A. Özyürek et R. Fernández, «Co-Speech Gesture Detection through Multi-Phase Sequence Labeling,» 2023.

[74] P. Molchanov, S. Gupta, K. Kim et J. Kautz, «Hand gesture recognition with 3D convolutional neural networks,» chez *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015.

[75] F. Yunus, C. Clavel et C. Pelachaud, *Sequence-to-Sequence Predictive Model: From Prosody To Communicative Gestures,* arXiv, 2020.

[76] B. G. Gebre, P. Wittenburg et P. Lenkiewicz, «Towards Automatic Gesture Stroke Detection,» chez *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, 2012.

[77] S. Alexanderson, D. House et J. Beskow, «Automatic annotation of gestural units in spontaneous face-to-face interaction,» chez *Proceedings of the Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*, 2016.

[78] R. Bryll, F. Quek et A. Esposito, «Automatic Hand Hold Detection in Natural Conversation,» August 2002.

[79] O. Köpüklü, A. Gunduz, N. Kose et G. Rigoll, «Online Dynamic Hand Gesture Recognition Including Efficiency Analysis,» *IEEE Transactions on Biometrics, Behavior, and Identity Science,* vol. 2, pp. 85-97, 2020.

[80] W. W. Kong et S. Ranganath, «Towards subject independent continuous sign language recognition: A segment and merge approach,» *Pattern Recognition,* vol. 47, p. 1294–1308, March 2014.

[81] G. Zhu, L. Zhang, P. Shen, J. Song, S. A. A. Shah et M. Bennamoun, «Continuous Gesture Segmentation and Recognition Using 3DCNN and Convolutional LSTM,» *IEEE Transactions on Multimedia,* vol. 21, pp. 1011-1021, 2019.

[82] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree et J. Kautz, «Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks,» *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4207-4215, 2016.

[83] Y. Song, D. Demirdjian et R. Davis, «Continuous body and hand gesture recognition for natural human-computer interaction,» *ACM Trans. Interact. Intell. Syst.,* vol. 2, March 2012.

[84] D. Hasegawa, N. Kaneko, S. Shirakawa, H. Sakuta et K. Sumi, «Evaluation of speech-to-gesture generation using bi-directional LSTM network,» chez *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018.

[85] T. Kucherenko, P. Jonell, S. van Waveren, G. E. Henter, S. Alexandersson, I. Leite et H. Kjellström, «Gesticulator: A framework for semantically-aware speech-driven gesture generation,» chez *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020.

[86] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens et J. Malik, *Learning Individual Styles of Conversational Gesture,* arXiv, 2019.

[87] F. Nihei, Y. Nakano, R. Higashinaka et R. Ishii, «Determining Iconic Gesture Forms based on Entity Image Representation,» chez *2019 International Conference on Multimodal Interaction*, 2019.

[88] A. Lücking, A. Mehler, D. Walther, M. Mauri et D. Kurfürst, «Finding Recurrent Features of Image Schema Gestures: the FIGURE corpus,» 2016.

[89] P. J. Yazdian, M. Chen et A. Lim, «Gesture2Vec: Clustering Gestures using Representation Learning Methods for Co-speech Gesture Generation,» 2021.

[90] K. Bergmann et S. Kopp, «GNetIc – Using Bayesian Decision Networks for Iconic Gesture Generation,» 2009.

[91] R. Ishii, T. Katayama, R. Higashinaka et J. Tomita, «Generating Body Motions using Spoken Language in Dialogue,» chez *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018.

[92] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko et H. Kjellström, «Analyzing Input and Output Representations for Speech-Driven Gesture Generation,» 2019.

[93] C. Ahuja et L.-P. Morency, *Language2Pose: Natural Language Grounded Pose Forecasting,* arXiv, 2019.

[94] L. Abel, V. Colotte et S. Ouni, «Towards realtime co-speech gestures synthesis using STARGATE,» chez *Interspeech 2024*, 2024.

[95] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson et H. Vilhjálmsson, «Towards a Common Framework for Multimodal Generation: The Behavior Markup Language,» chez *Intelligent Virtual Agents*, Springer Berlin Heidelberg, 2006, p. 205–217.

[96] J. Cassell, H. H. Vilhjálmsson et T. Bickmore, «BEAT: the Behavior Expression Animation Toolkit,» chez *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 2001.

[97] J. Lee et S. Marsella, «Nonverbal Behavior Generator for Embodied Conversational Agents,» chez *Intelligent Virtual Agents*, Springer Berlin Heidelberg, 2006, p. 243–255.

[98] M. Lhommet et S. Marsella, «Metaphoric Gestures: Towards Grounded Mental Spaces,» chez *Intelligent Virtual Agents*, Springer International Publishing, 2014, p. 264–274.

[99] B. Ravenet, C. Pelachaud, C. Clavel et S. Marsella, «Automating the Production of Communicative Gestures in Embodied Characters,» *Frontiers in Psychology,* vol. 9, July 2018.

[100] J. J. Hopfield, «Neural networks and physical systems with emergent collective computational abilities.,» *Proceedings of the national academy of sciences,* vol. 79, p. 2554–2558, 1982.

[101] M. Schuster et K. K. Paliwal, «Bidirectional recurrent neural networks,» *IEEE Transactions on Signal Processing,* vol. 45, pp. 2673-2681, 1997.

[102] S. Hochreiter et J. Schmidhuber, «Long short-term memory,» *Neural computation,* vol. 9, p. 1735–1780, 1997.

[103] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk et Y. Bengio, *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,* arXiv, 2014.

[104] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim et G. Lee, «Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots,» chez *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[105] J. Woo, C. I. Pelachaud et C. Achard, «Creating an interactive human/agent loop using multimodal recurrent neural networks,» chez *WACAI 2021*, 2021.

[106] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser et I. Polosukhin, «Attention is all you need,» *Advances in neural information processing systems,* vol. 30, 2017.

[107] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera et D. Manocha, «Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents,» chez *2021 IEEE virtual reality and 3D user interfaces (VR)*, 2021.

[108] J. Devlin, M.-W. Chang, K. Lee et K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,» *arXiv preprint arXiv:1810.04805,* 2018.

[109] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard et L. Jackel, «Handwritten digit recognition with a back-propagation network,» *Advances in neural information processing systems,* vol. 2, 1989.

[110] N. T. V. Tuyen, A. Elibol et N. Y. Chong, «Conditional generative adversarial network for generating communicative robot gestures,» chez *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2020.

[111] T. N. Kipf et M. Welling, «Semi-supervised classification with graph convolutional networks,» *arXiv preprint arXiv:1609.02907,* 2016.

[112] S. Yan, Y. Xiong et D. Lin, «Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,» *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 32, April 2018.

[113] K. Zhou, Z. Cheng, H. P. H. Shum, F. W. B. Li et X. Liang, «Stgae: Spatial-temporal graph auto-encoder for hand motion denoising,» chez *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2021.

[114] J. Ho, A. Jain et P. Abbeel, *Denoising Diffusion Probabilistic Models,* arXiv, 2020.

[115] T. Kucherenko, R. Nagy, Y. Yoon, J. Woo, T. Nikolov, M. Tsakov et G. E. Henter, «The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in

monadic and dyadic settings,» chez *Proceedings of the 25th International Conference on Multimodal Interaction*, New York, NY, USA, 2023.

[116] A. Deichler, S. Mehta, S. Alexanderson et J. Beskow, «Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation,» chez *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, 2023.

[117] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu et L. Yu, *Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation,* arXiv, 2023.

[118] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl et B. Yu, «Interpretable machine learning: definitions, methods, and applications,» 2019.

[119] G. E. Hinton et S. Roweis, «Stochastic Neighbor Embedding,» chez *Advances in Neural Information Processing Systems*, 2002.

[120] L. McInnes, J. Healy et J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,* arXiv, 2018.

[121] S. Kornblith, M. Norouzi, H. Lee et G. Hinton, «Similarity of Neural Network Representations Revisited,» chez *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[122] M. Raghu, J. Gilmer, J. Yosinski et J. Sohl-Dickstein, *SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability,* arXiv, 2017.

[123] S. Lundberg et S.-I. Lee, *A Unified Approach to Interpreting Model Predictions,* arXiv, 2017.

[124] M. Sundararajan, A. Taly et Q. Yan, *Axiomatic Attribution for Deep Networks,* arXiv, 2017.

[125] M. T. Ribeiro, S. Singh et C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier,* arXiv, 2016.

[126] S. Abnar et W. Zuidema, *Quantifying Attention Flow in Transformers,* arXiv, 2020.

[127] J. Qi, L. Ma, Z. Cui et Y. Yu, «Computer vision-based hand gesture recognition for human-robot interaction: a review,» *Complex & Intelligent Systems,* vol. 10, p. 1581–1606, July 2023.

[128] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim et G. Lee, «Speech gesture generation from the trimodal context of text, audio, and speaker identity,» *ACM Transactions on Graphics (TOG),* vol. 39, p. 1–16, 2020.

[129] Y. Yoon, P. Wolfert, T. Kucherenko, C. Viegas, T. Nikolov, M. Tsakov et G. E. Henter, «The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation,» chez *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022.

[130] Y. Yoon, K. Park, M. Jang, J. Kim et G. Lee, «SGToolkit: An Interactive Gesture Authoring Toolkit for Embodied Conversational Agents,» chez *The 34th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA, 2021.

[131] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert et G. E. Henter, «A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020,» chez *26th international conference on intelligent user interfaces*, 2021.

[132] M. Kipp, «ANVIL - a generic annotation tool for multimodal dialogue,» chez *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2001.

[133] M. Tellier, B. Azaoui et J. Saubesty, «Segmentation et annotation du geste : Méthodologie pour travailler en équipe,» chez *JEP-TALN-RECITAL 2012 : Atelier DEGELS 2012 (Défi GEste Langue des Signes)*, Grenoble, 2012.

[134] J. Bressem et S. H. Ladewig, «Rethinking gesture phases: Articulatory features of gestural movement?,» *Semiotica,* vol. 2011, January 2011.

[135] B. MacWhinney, «The CHILDES project: tools for analyzing talk,» *Child Language Teaching and Therapy,* vol. 8, p. 217–218, June 1992.

[136] B. Bigi, «SPPAS - MULTI-LINGUAL APPROACHES TO THE AUTOMATIC ANNOTATION OF SPEECH,» *The Phonetician. Journal of the International Society of Phonetic Sciences,* Vols. %1 sur %2111-112, pp. 54-69, 2015.

[137] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner et M. Sonderegger, «Montreal forced aligner: Trainable text-speech alignment using kaldi.,» chez *Interspeech*, 2017.

[138] M. Honnibal et I. Montani, «spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing,» *To appear,* vol. 7, p. 411–420, 2017.

[139] ORTOLANG Consortium, *ORTOLANG – Plate-forme d'outils et de ressources linguistiques pour un traitement optimisé de la langue française,* 2016.

[140] M. Amoyal, B. Priego-Valverde and S. Rauzy, "PACO: a Corpus to Analyze the Impact of Common Ground in Spontaneous Face-to-Face Interaction," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, 2020.

[141] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde et S. Rauzy, «Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle,» *Revue TAL,* vol. 49, p. pp–105, 2008.

[142] Y. Ferstl et R. McDonnell, «Investigating the use of recurrent motion modelling for speech gesture generation,» chez *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018.

[143] S. Alexanderson, G. E. Henter, T. Kucherenko et J. Beskow, «Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows,» *Computer Graphics Forum,* vol. 39, p. 487–496, May 2020.

[144] A. Lücking, K. Bergmann, F. Hahn, S. Kopp et H. Rieser, «The Bielefeld Speech and Gesture Alignment Corpus (SaGA),» 2010.

[145] E. Volkova, S. De La Rosa, H. H. Bülthoff et B. Mohler, «The MPI emotional body expressions database for narrative scenarios,» *PloS one,* vol. 9, p. e113647, 2014.

[146] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee et S. S. Narayanan, «IEMOCAP: Interactive emotional dyadic motion capture database,» *Language resources and evaluation,* vol. 42, p. 335–359, 2008.

[147] K. Takeuchi, S. Kubota, K. Suzuki, D. Hasegawa et H. Sakuta, «Creating a gesture-speech dataset for speech-based automatic gesture generation,» chez *International Conference on Human-Computer Interaction*, 2017.

[148] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André et M. Valstar, «The NoXi database: multimodal recordings of mediated novice-expert interactions,» chez *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017.

[149] L. B. Hensel, *GESRes the gesture dataset,* OSF, 2024.

[150] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt et B. Zheng, «BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis,» chez *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, 2022.

[151] P. Ekman et W. V. Friesen, *Facial Action Coding System,* American Psychological Association (APA), 1978.

[152] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et others, «The Kaldi speech recognition toolkit,» chez *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[153] M. Grondin-Verdon, D. Caillat et S. Ouni, «Qualitative study of gesture annotation corpus : Challenges and perspectives,» chez *ICMI Companion '24: Companion Proceedings of the 26th International Conference on Multimodal Interaction*, San Jose, Costa Rica, 2024.

[154] M. Seyfeddinipur, «Disfluency: Interrupting speech and gesture,» 2006.

[155] J. Nirme, A. Gulz, M. Haake et M. Gullberg, «Early or synchronized gestures facilitate speech recall—a study based on motion capture data,» *Frontiers in Psychology,* vol. 15, March 2024.

[156] E. Jarmołowicz-Nowikow et M. Karpiński, «Boundaries of gestures: Naive segmentation of the stream of human hand movements,» *Language & Communication,* vol. 105, p. 22–36, November 2025.

[157] J. Holler, J. Bavelas, J. Woods, M. Geiger et L. Simons, «Given-New Effects on the Duration of Gestures and of Words in Face-to-Face Dialogue,» *Discourse Processes,* vol. 59, p. 619–645, September 2022.

[158] P. Prieto, A. Cravotta, O. Kushch, P. Rohrer et I. Vilà-Giménez, «Deconstructing beat gestures: a labelling proposal,» chez *Speech Prosody 2018*, 2018.

[159] T. Leonard et F. Cummins, «The temporal relation between beat gestures and speech,» *Language and Cognitive Processes,* vol. 26, p. 1457–1471, December 2011.

[160] J. Cohen, «A Coefficient of Agreement for Nominal Scales,» *Educational and Psychological Measurement,* vol. 20, p. 37–46, April 1960.

[161] B. Butterworth et G. Beattie, «Gesture and Silence as Indicators of Planning in Speech,» chez *Recent Advances in the Psychology of Language*, Springer US, 1978, p. 347–360.

[162] L. Abel, «Expressive audio-visual speech synthesis in an interaction context,» defended 6th February 2025.

[163] L. Abel, V. Colotte et S. Ouni, «Towards interpretable co-speech gestures synthesis using STARGATE,» chez *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, 2024.

[164] M. Grondin-Verdon, D. Caillat, L. Abel et S. Ouni, «Evaluating Automatic Hand-Gesture Generation Using Multimodal Corpus Annotations: The Benefits of a Multidisciplinary Approach,» chez *Proceedings of the International Workshop on Generation and Evaluation of Non-verbal Behaviour for Embodied Agents*, 2025.

[165] R. Benesh et J. Benesh, An Introduction to Benesh Dance Notation, A. and C. Black, 1955.

[166] A. H. Guest, Labanotation: The System of Analyzing and Recording Movement, Routledge, 2013.

[167] N. Dael, M. Mortillaro et K. R. Scherer, «The Body Action and Posture Coding System (BAP): Development and Reliability,» *Journal of Nonverbal Behavior,* vol. 36, p. 97–121, January 2012.

[168] R. Laban et F. C. Lawrence, Effort (Laban Movement Analysis), London: MacDonald and Evans, 1947.

[169] N. Eshkol et A. Wachman, Movement Notation, London: Weidenfeld and Nicolson, 1958.

[170] S. Escalera, X. Baró, J. Gonzàlez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton et I. Guyon, «ChaLearn Looking at People Challenge 2014: Dataset and Results,» chez *Computer Vision - ECCV 2014 Workshops*, Springer International Publishing, 2015, p. 459–473.

[171] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, I. Guyon, V. Athitsos, H. Escalante, L. Sigal, A. Argyros, C. Sminchisescu, R. Bowden et S. Sclaroff, «ChaLearn multi-modal gesture recognition 2013: grand challenge and workshop summary,» chez *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013.

[172] M. Vázquez Enríquez, J. L. A. Castro, L. D. Fernandez, J. C. S. Jacques Junior et S. Escalera, «ECCV 2022 Sign Spotting Challenge: Dataset, Design and Results,» chez *Computer Vision – ECCV 2022 Workshops*, Springer Nature Switzerland, 2023, p. 225–242.

[173] I. Guyon, V. Athitsos, P. Jangyodsuk et H. J. Escalante, «The ChaLearn gesture dataset (CGD 2011),» *Machine Vision and Applications,* vol. 25, p. 1929–1951, February 2014.

[174] P. Wagner, S. Peres, R. Madeo, C. Lima et F. Freitas, «Gesture Unit Segmentation Using Spatial-Temporal Information and Machine Learning,» 2014.

[175] A. S. Ramakrishnan et M. Neff, «Segmentation of hand gestures using motion capture data,» chez *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*, Richland, 2013.

[176] R. Madeo, C. Lima et S. Peres, «Gesture unit segmentation using support vector machines: Segmenting gestures from rest positions,» 2013.

[177] R. C. B. Madeo, S. M. Peres et C. A. d. M. Lima, «Gesture phase segmentation using support vector machines,» *Expert Syst. Appl.,* vol. 56, p. 100–115, September 2016.

[178] S. Okada, M. Bono, K. Takanashi, Y. Sumi et K. Nitta, «Context-based conversational hand gesture classification in narrative interaction,» chez *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, New York, NY, USA, 2013.

[179] Y. Yin et R. Davis, «Gesture spotting and recognition using salience detection and concatenated hidden markov models,» chez *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013.

[180] V. Viitaniemi, T. Jantunen, L. Savolainen, M. Karppa and J. Laaksonen, "S-pot - a benchmark in spotting signs within continuous signing," in *International conference on language resources and evaluation, LREC, Reykjavik, Iceland, May 26-31 2014*, Luxembourg, 2014.

[181] R. Cui, H. Liu et C. Zhang, «Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization,» 2017.

[182] O. Ikne, R. Slama, H. Saoudi et H. Wannous, «Spatio-Temporal Sparse Graph Convolution Network for Hand Gesture Recognition,» chez *The 18th IEEE International Conference on Automatic Face and Gesture Recognition*, Istanbul, 2024.

[183] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez et K. Yanai, *IPN Hand: A Video Dataset and Benchmark for Real-Time Continuous Hand Gesture Recognition,* 2020.

[184] G. Fronteddu, S. Porcu, A. Floris et L. Atzori, «A dynamic hand gesture recognition dataset for human-computer interfaces,» *Computer Networks,* vol. 205, p. 108781, March 2022.

[185] D. P. Kingma et J. Ba, *Adam: A Method for Stochastic Optimization,* arXiv, 2014.

# Résumé

L'étude de la gestualité s'inscrit à l'intersection de plusieurs disciplines. Le champ des *gesture studies* se consacre à la caractérisation des gestes par l'analyse de leurs propriétés formelles et fonctionnelles, de leur coordination avec le discours, et de leur contribution à la construction du sens. Cette approche repose sur une annotation multimodale manuelle, laquelle permet une observation, une description et une interprétation systématiques des gestes. Néanmoins, cette approche demeure extrêmement chronophage, fortement dépendante de l'expertise des annotateurs et sujette à une variabilité inter-projets, limitant la comparabilité des résultats et l'élaboration de corpus annotés à grande échelle. Les sciences computationnelles, quant à elles, se focalisent sur la modélisation, la prédiction et la génération automatique de gestes, visant à animer des agents conversationnels dotés de comportements réalistes. Cependant, les modèles d'apprentissage automatique actuels présentent encore des lacunes : les gestes synthétisés manquent de richesse sémantique et se réduisent à des mouvements rythmiques. Le postulat de cette thèse est qu'une approche résolument pluridisciplinaire de l'étude du geste manuel co-verbal est nécessaire pour favoriser des avancées mutuelles entre disciplines et dépasser leurs limites respectives.

Les contributions scientifiques de ce travail doctoral résultent directement de l'approche pluridisciplinaire adoptée. La première contribution concerne l'analyse critique de l'enrichissement du corpus BEAT, ressource majeure utilisée pour la génération automatique de gestes. Cette étude a révélé des inconsistances significatives dans les annotations multimodales, tant au niveau de la segmentation que de l'interprétation des gestes, et a montré que les statistiques descriptives sur la durée et le nombre de gestes constituent un indicateur initial pertinent de la fiabilité d'un enrichissement. La deuxième contribution est le corpus réannoté *reBeat-14* (13 min), produit dans le cadre de cette thèse. Son protocole d'enrichissement, solidement ancré dans les *gesture studies*, offre des annotations multimodales précises et fiables, et leur analyse met en évidence la finesse de la modélisation du geste, confirmant sa valeur en tant que *gold standard*. La troisième contribution réside dans l'analyse multimodale experte des gestes de synthèse produits par le modèle STARGATE, qui a permis d'identifier des critères plus objectifs pour l'évaluation de la qualité des gestes générés automatiquement, en s'appuyant sur l'analyse des mouvements, incluant les variations de vitesse et d'amplitude ainsi que le degré d'asymétrie des mains. La quatrième contribution consiste en la proposition d'un *framework* d'annotation automatique des aspects formels du geste, incluant la position, l'orientation, la vitesse et l'amplitude du geste, l'orientation et la configuration des mains, ainsi que l'orientation de la paume. Ce système, accompagné d'une interface graphique dédiée (*Gesture User Interface*) répond notamment aux difficultés liées à l'annotation formelle manuelle basée sur l'observation, en permettant une annotation plus objective et reproductible. Enfin, la cinquième et dernière contribution de cette thèse est COSMOS, un modèle de segmentation des gestes, basé sur une architecture CNN-GRU. Ce travail met en évidence la faisabilité d'un modèle « compagnon » performant, malgré un volume limité de données supervisées issues des annotations expertes de *reBeat-14*, et ouvre la voie à

une application semi-automatique (*human in the loop*) permettant d'annoter un plus grand nombre de données sans compromettre la qualité de l'enrichissement.

Ces travaux de thèse soulignent les gains d'une approche interdisciplinaire combinant sciences du langage et sciences computationnelles. Ils offrent de nouvelles perspectives pour l'étude du geste, tant sur le plan descriptif et fonctionnel que pour le développement d'outils et de modèles génératifs fiables.

**Mots clés** : Geste co-verbal manuel, Méthodologies interdisciplinaires, Modélisation computationnelle, Enrichissement de corpus, Annotations multimodales

# Abstract

The study of gesture lies at the intersection of several disciplines. The field of gesture studies is devoted to the characterization of gestures through the analysis of their formal and functional properties, their coordination with speech, and their contribution to meaning construction. This approach relies on manual multimodal annotation, enabling systematic observation, description, and interpretation of gestures. However, it remains extremely time-consuming, highly dependent on annotators' expertise, and subject to variability across projects, thereby limiting the comparability of results and the development of large-scale annotated corpora. Computational sciences, in contrast, focus on the modeling, prediction, and automatic generation of gestures, with the aim of animating conversational agents endowed with realistic behaviors. Nevertheless, current machine learning models still exhibit significant limitations: synthesized gestures often lack semantic richness and are reduced to rhythmic movements. The central premise of this dissertation is that a resolutely interdisciplinary approach to the study of co-verbal manual gestures is necessary to foster mutual advances across disciplines and to overcome their respective limitations.

The scientific contributions of this doctoral work stem directly from the interdisciplinary approach adopted. The first contribution consists of a critical analysis of the enrichment of the BEAT corpus; a major resource used for automatic gesture generation. This study revealed substantial inconsistencies in multimodal annotations, both in gesture segmentation and interpretation, and demonstrated that descriptive statistics regarding gesture duration and frequency constitute a relevant initial indicator of enrichment reliability. The second contribution is the reannotated corpus reBeat-14 (13 minutes), produced as part of this thesis. Its enrichment protocol, firmly grounded in gesture studies, provides precise and reliable multimodal annotations. The analysis of this corpus highlights the granularity and rigor of its gesture modeling, confirming its value as a gold standard. The third contribution lies in the expert multimodal analysis of synthetic gestures generated by the STARGATE model. This analysis made it possible to identify more objective criteria for evaluating the quality of automatically generated gestures, based on movement analysis, including variations in velocity and amplitude, as well as the degree of hand asymmetry. The fourth contribution is the proposal of a framework for the automatic annotation of formal aspects of gesture, including gesture position, orientation, velocity, and amplitude, as well as hand orientation, handshape, and palm orientation. This system, accompanied by a dedicated graphical interface (Gesture User Interface), addresses the challenges inherent in manual observation-based formal annotation by enabling more objective and reproducible annotations. Finally, the fifth and final contribution of this dissertation is COSMOS, a gesture segmentation model based on a CNN-GRU architecture. This work demonstrates the feasibility of a high-performing "companion" model, despite the limited volume of supervised data derived from expert annotations of reBeat-14, and paves the way for a semi-automatic human-in-the-loop application capable of annotating larger datasets without compromising enrichment quality.

Overall, this dissertation highlights the benefits of an interdisciplinary approach combining language sciences and computational sciences. It opens new perspectives for gesture research, both in terms of descriptive and functional analysis and in the development of reliable tools and generative models.

**Keywords**: Co-verbal manual gesture, Interdisciplinary methodologies, Computational modeling, Corpus enrichment, Multimodal annotations

# Table des matières

**INTRODUCTION** - 11 -

*1 La gestualité, un champ d'étude en soi* - 11 -

1.1 Geste et *Gesture studies* - 11 -

1.1.1 Un objet à la croisée de plusieurs disciplines - 11 -

1.1.2 Le geste comme acte communicatif - 12 -

1.2 Geste et traitement informatique - 13 -

1.2.1 L'essor des approches outillées - 13 -

1.2.2 Approche computationnelle fondée sur les modèles d'apprentissage - 14 -

1.2.3 Limites des approches outillées - 15 -

*2 Problématique et questions de recherche* - 16 -

*3 Contexte de recherche* - 18 -

*4 Organisation de la thèse* - 19 -

*5 Valorisations de recherche* - 21 -

5.1 Contributions scientifiques - 21 -

5.2 Publications et médiations scientifiques - 22 -

5.2.1 Articles de conférence associés à une communication en congrès international - 22 -

5.2.2 Communication sur résumé - 23 -

5.2.3 Poster - 23 -

5.2.4 Médiation scientifique grand public - 23 -

**PARTIE I FONDEMENTS THEORIQUES ET ENJEUX DES ETUDES DE LA GESTUALITE** - 25 -

CHAPITRE 1 L'ETUDE DU GESTE DANS LE CHAMP DES *GESTURES STUDIES* - 27 -

*1 Définition du geste* - 27 -

*2 La production gestuelle en contexte* - 28 -

2.1 Coordination entre geste et parole - 28 -

2.2 Ancrage culturel et situationnel - 29 -

*3 Eléments de description de la gestualité* - 30 -

3.1 Segmentation du geste - 30 -

3.1.1 Les unités gestuelles selon Kendon - 30 -

3.1.2 Définition des phases selon Kendon - 30 -

3.2 Approches spatiales et articulaires - 32 -

3.2.1 La description de l'espace gestuel - 32 -

3.2.2 La description articulaire - 32 -

3.3 Descriptions formelles - 33 -

3.4 Approches interprétatives - 34 -

3.5 Complémentarité des approches - 35 -

*4 Conclusion de chapitre* - 36 -

CHAPITRE 2 L'ETUDE DE LA GESTUALITE EN SCIENCES COMPUTATIONNELLES - 39 -

*1 Le geste comme donnée numérique* - 39 -

1.1 Distinction entre gestes et mouvement - 39 -

1.2 Représentation computationnelle du geste - 40 -

*2 Types de modélisation du geste* - 41 -

2.1 Reconnaissance et segmentation - 41 -

2.2 Classification - 42 -

2.3 Génération - 42 -

*3 Outils pour la modélisation* - 43 -

3.1 Modélisation par règles - 43 -
3.2 Modélisation dynamique et temporelle - 44 -
3.2.1 Réseaux de neurones récurrents (RNN) - 44 -
3.2.2 Transformers - 45 -
3.3 Modélisation structurelle et explicite du corps - 46 -
3.3.1 Réseaux de neurones convolutifs (CNN) - 46 -
3.3.2 Réseaux de neurones graphiques (GNN) - 47 -
3.3.3 Modèles de diffusion - 48 -
*4 Interprétabilité et explicabilité des modèles - 49 -*
4.1 Définition - 49 -
4.2 Hyperparamètres et fonction de perte - 50 -
4.3 Outils d'analyse des représentations internes - 50 -
*5 Evaluation des modèles - 52 -*
5.1 Métriques quantitatives - 52 -
5.2 Évaluations perceptives - 53 -
*6 Conclusion de chapitre - 54 -*
BILAN DE PARTIE - 57 -

**PARTIE II RESSOURCES ET DONNEES MULTIMODALES - 61 -**

CHAPITRE 3 CORPUS MULTIMODAUX POUR L'ETUDE DE LA GESTUALITE : ENJEUX ET NECESSITE - 63 -
*1 Introduction - 63 -*
*2 Modes de constitution d'un corpus de gestes - 64 -*
2.1 Situations de communication suscitées par le protocole - 64 -
2.1.1 Monologue et dialogue - 65 -
2.1.2 Spontanée, dirigé, acté - 65 -
2.1.3 Activité libre, élaboration discursive, coopération - 65 -
2.2 Types de captation - 66 -
2.2.1 Extraction automatique - 66 -
2.2.2 Environnement contrôlé - 67 -
2.2.3 Situation authentique et naturelle - 67 -
*3 Mise en œuvre du protocole d'enrichissement - 67 -*
3.1 Types d'annotateurs - 68 -
3.2 Protocole d'enrichissement - 69 -
3.2.1 La rédaction d'un manuel de codage - 69 -
3.2.2 La phase de synchronisation (calibration) des annotateurs - 69 -
3.2.3 Validation des annotations - 69 -
*4 Variétés de formats d'enrichissement d'un corpus multimodal - 70 -*
4.1 Transcription et annotation de la parole - 70 -
4.2 Segmentation des gestes - 71 -
4.3 Catégorisations formelles et fonctionnelles des gestes - 71 -
*5 Outils d'enrichissement de corpus - 72 -*
*6 Partage et science ouverte dans le domaine de la gestualité - 74 -*
6.1 Disponibilité des données et risques - 74 -
6.2 Corpus disponibles - 74 -
6.2.1 PACO - 75 -
6.2.2 CID - 75 -
6.2.3 TSG - 75 -
6.2.4 SaGA - 76 -
6.2.5 MPIEMBM - 76 -
6.2.6 IEMOCAP - 77 -
6.2.7 Aoyama Gakuin - 77 -
6.2.8 NoXi - 77 -
6.2.9 GESRes - 78 -

6.2.10 BEAT - 78 -
7 *Conclusion de chapitre* - 79 -
CHAPITRE 4 ANALYSE CRITIQUE DE L'ENRICHISSEMENT D'UN CORPUS MULTIMODAL - 81 -
1 ***Introduction*** - 81 -
2 *Le corpus BEAT* - 82 -
2.1 Captation et format de données - 82 -
2.2 Enrichissements disponibles - 83 -
3 *Statistiques descriptives de l'enrichissement du corpus* - 84 -
3.1 Analyses statistiques des annotations fonctionnelles des gestes - 84 -
3.1.1 Annotations prises en compte dans le corpus initial - 84 -
3.1.2 Nombre d'annotations dans le corpus - 85 -
3.1.3 Recouvrement temporel des annotations - 85 -
3.2 Analyses statistiques de la durée des gestes - 86 -
3.2.1 Durée des gestes par catégorie fonctionnelle - 86 -
3.2.2 Durées aberrantes - 87 -
3.3 Analyses statistiques de l'affiliation lexicale des gestes - 89 -
4 *Analyse fonctionnelle : exemple de cas* - 90 -
4.1 Méthodologie de réannotation ponctuelle - 90 -
4.2 Analyse comparative de l'enrichissement d'une courte séquence - 91 -
5 *Conclusion de chapitre* - 92 -
CHAPITRE 5 PROTOCOLE D'ENRICHISSEMENT DE CORPUS DE GESTES - 95 -
1 ***Introduction*** - 95 -
2 *Transcription et annotation de la parole* - 96 -
2.1 Transcription orthographique - 96 -
2.2 Révision lexicale - 97 -
2.3 Révision syntaxique - 97 -
2.4 Balisage - 98 -
2.5 Segmentation et alignement des mots et des syllabes - 98 -
2.6 Extractions de traits linguistiques - 99 -
3 *Annotations gestuelles* - 99 -
3.1 Procédure générale d'annotation - 99 -
3.2 Segmentation du flux de mouvement - 100 -
3.3 Annotation fonctionnelle de la gestualité - 101 -
3.3.1 Caractérisation fonctionnelle de l'activité gestuelle - 101 -
3.3.2 Caractérisation du rôle fonctionnel du geste - 101 -
3.3.3 Affiliation lexicale, prosodique et gestuelle - 103 -
3.3.4 Caractérisation de la certitude du choix des types fonctionnels - 103 -
3.4 Annotation formelle de la gestualité - 104 -
3.4.1 Catalogue de trajectoires - 104 -
3.4.2 Labellisation de la manualité - 105 -
3.4.3 Labellisation de la position de la main - 105 -
3.4.4 Labellisation de l'orientation de la paume - 105 -
3.4.5 Labellisation de la configuration de la main - 106 -
4 *Conclusion de chapitre* - 106 -
BILAN DE PARTIE - 109 -

**PARTIE III APPORTS ET PERSPECTIVES DE L'EXPERTISE DES SCIENCES DU LANGAGE - 113 -**

CHAPITRE 6 L'ANNOTATION MULTIMODALE COMME OUTIL DE MODELISATION DU GESTE CO-VERBAL - 115 -
1 ***Introduction*** - 115 -
2 *Statistiques générales* - 116 -
3 *Analyse fonctionnelle des gestes* - 116 -

3.1 Distribution des labels fonctionnels - 116 -
3.1.1 Types fonctionnels par dimension - 117 -
3.1.2 Types fonctionnels appariés en D1 et d2 - 118 -
3.2 Durée des gestes - 119 -
3.3 Scores de certitudes - 120 -
*4 Analyse formelle des gestes annotés - 121 -*
4.1 Manualité - 121 -
4.1.1 Distribution générale des labels de manualité - 121 -
4.1.2 Manualité et type fonctionnel - 122 -
4.2 Forme de la trajectoire - 123 -
4.3 Position de la main - 125 -
4.4 Orientation de la paume - 127 -
4.5 Configuration de la main - 129 -
*5 Analyse cinématique des gestes - 132 -*
5.1 Méthodologies d'extractions de descripteurs du mouvement et de visualisation de trajectoires - 132 -
5.1.1 Calcul de la vitesse - 132 -
5.1.2 Calcul de la distance parcourue et amplitude maximum - 132 -
5.1.3 Visualisation des trajectoires - 133 -
5.2 Vitesse - 133 -
5.3 Utilisation de l'espace - 134 -
5.3.1 Distance parcourue - 134 -
5.3.2 Amplitude maximum - 136 -
5.3.3 Latéralité et hauteur - 137 -
*6 Analyse de l'ancrage des gestes au discours - 138 -*
6.1 Brève description de la production verbale - 138 -
6.2 Affiliation lexicale selon le type fonctionnel - 138 -
*7 Conclusion - 141 -*
CHAPITRE 7 L'ANNOTATION MULTIMODALE COMME OUTIL D'EVALUATION DES GESTES SYNTHETIQUES - 147 -
*1 Introduction - 147 -*
*2 Méthodologie - 148 -*
2.1 Le modèle STARGATE - 148 -
2.2 Sélection des données - 149 -
2.3 Protocole d'annotation et analyses - 150 -
*3 Identification des gestes potentiels - 151 -*
3.1 Désaccord de segmentation entre annotatrices - 151 -
3.1.1 Modes d'analyse de l'accord inter-annotateurs sur les horodatages - 151 -
3.1.2 Mode d'affectation des identifiants des gestes annotés - 151 -
3.1.3 Analyse de l'accord inter-annotateur sur les repères temporels d'annotation - 152 -
3.2 Gestes potentiels (activités) annotés - 153 -
3.2.1 Analyse de la durée des activités - 153 -
3.2.2 Analyse du type de phase d'activité annotées - 153 -
*4 Interprétation fonctionnelle - 154 -*
4.1 Analyse des activités communicatives et non communicatives annotées - 154 -
4.2 Analyse des types fonctionnels des gestes annotés - 155 -
4.3 Analyse de l'affiliation lexicale des gestes - 156 -
4.4 Analyse des scores de certitudes - 157 -
*5 Analyses formelles - 157 -*
5.1 Analyse de la manualité des gestes - 157 -
5.2 Analyse de l'utilisation de l'espace gestuel - 158 -
5.3 Analyse temporelle de la spatialité et de la vélocité - 159 -
5.4 Exemples de gestes ambigus - 161 -
5.4.1 Le geste de synthèse n° 5 - 161 -

5.4.2 Le geste naturel n° 25 - 162 -
6 *Conclusion de chapitre* - 162 -
BILAN DE PARTIE - 165 -

**PARTIE IV APPORTS ET PERSPECTIVES DE L'EXPERTISE DES SCIENCES COMPUTATIONNELLES - 169 -**

CHAPITRE 8 ANNOTATION AUTOMATIQUE DES ASPECTS FORMELS DU GESTE MANUEL CO-VERBAL - 171 -
1 ***Introduction*** - 171 -
2 *Contextualisation* - 172 -
3 *Initialisation du framework* - 174 -
3.1 Identification des variantes de formes dans le corpus reBeat-14 - 174 -
3.2 La manualité : un cas spécifique pour la reconnaissance automatique - 175 -
4 *Qualification formelle automatique* - 176 -
4.1 Localisations au travers des densités de position et des caractéristiques anatomiques - 176 -
4.1.1 Détermination des positions et de l'étendue gestuelle du locuteur - 176 -
4.1.2 Mode de calcul des densités de positions des gestes - 177 -
4.1.3 Mapping théorique de l'espace gestuel - 178 -
**Axe droite-gauche** - 178 -
**Axe haut-bas** - 179 -
**Axe intérieur-extérieur** - 179 -
4.2 Orientations au travers de vecteurs - 181 -
4.2.1 Espace tridimensionnel local discrétisé : une sphère de 64 zones - 181 -
**Quantification des angles** - 181 -
**Qualification des angles** - 182 -
4.2.2 Orientation de la trajectoire du geste - 182 -
4.2.3 Orientation globale de la main et de la paume - 183 -
4.3 Configurations des doigts au travers de calcul de distances - 183 -
4.3.1 La courbure du doigt - 184 -
4.3.2 La fermeture du doigt - 184 -
4.3.3 L'espacement des doigts - 184 -
4.4 Dynamique gestuelle : la vitesse et la taille du geste - 185 -
5 ***GUI : Gesture User Interface*** - 185 -
5.1 Rôle stratégique de l'interface dans la démarche interdisciplinaire - 186 -
5.2 Résolution du biais perceptif - 187 -
5.3 L'interface comme outil d'audit et de validation - 189 -
5.4 Perspectives : vers un écosystème interopérable pour l'annotation multimodale - 189 -
5.4.1 L'interopérabilité et la standardisation des annotations - 189 -
5.4.2 Agnosticisme matériel - 190 -
5.4.3 Intégration de modèles prédictifs continus - 190 -
6 *Conclusion de chapitre* - 191 -
CHAPITRE 9 SEGMENTATION SEMI-AUTOMATIQUE DU GESTE SOUS CONTRAINTE DE DONNEES D'ENTRAINEMENT LIMITEES - 195 -
1 ***Introduction*** - 195 -
2 *Overview de COSMOS* - 196 -
2.1 Données utilisées - 196 -
2.2 Architecture du modèle - 198 -
2.2.1 Encodeur de mouvement - 198 -
**Rôle du décodeur et pertinence des CNN** - 198 -
**Pile de blocs résiduels pour l'encodage du mouvement** - 199 -
**Première couche d'un bloc résiduel** - 199 -
**Seconde couche d'un bloc résiduel** - 200 -

**Champ réceptif total de l'encodeur** - 200 -
2.2.2 Décodeur de mouvement - 201 -
**Rôle du décodeur et pertinence des GRU** - 201 -
**Structure du décodeur** - 201 -
2.2.3 Classification binaire et export des annotations automatiques - 202 -
2.3 Configurations techniques - 203 -
*3 Analyses des performances du modèle baseline - 204 -*
3.1 Métriques utilisées pour l'analyse de performance - 204 -
3.1.1 Métriques *frame*-wise - 204 -
3.1.2 Métriques segment-wise - 205 -
3.2 Performances sur les sets d'entraînement et de validation - 206 -
3.2.1 Valeurs de la fonction de perte à chaque *step* - 206 -
3.2.2 Métriques à chaque *step* - 206 -
3.3 Performances sur le set de test - 207 -
3.3.1 Analyse des métriques *frame-wise* - 207 -
3.3.2 Analyse des métriques *segment-wise* - 208 -
3.3.3 Analyse de la robustesse du modèle - 208 -
3.3.4 Analyse d'une séquence de prédiction - 210 -
**Prédictions partiellement correctes** - 210 -
**Prédictions incorrectes** - 212 -
*4 Analyses de l'impact des hyperparamètres sur les performances du modèle baseline - 213 -*
4.1 Nombre de canaux et nombre de neurones - 213 -
4.1.1 Architectures d'encodeurs convolutionnels testés - 213 -
4.1.2 Impact de la complexité et de la profondeur - 214 -
4.2 Taux d'apprentissage - 214 -
4.2.1 Valeurs testées - 214 -
4.2.2 Impact du taux d'apprentissage - 215 -
4.3 Hyperparamètres contrôlant le surapprentissage - 215 -
4.3.1 Méthodes pour contrôler le surapprentissage - 215 -
**Régularisation des poids de la fonction de perte** - 215 -
**Valeurs de dropout** - 215 -
4.3.2 Impact de la régularisation L2 - 216 -
4.3.3 Impact de l'activation du dropout - 217 -
*5 Analyses de l'impact des données sur les performances - 217 -*
5.1 Évaluation des performances sur des données d'un locuteur non entraîné - 217 -
5.2 Evaluation des performances avec d'autres données que les positions - 219 -
5.2.1 Types de données testées - 219 -
5.2.2 Performances des modèles unimodaux et multimodaux testés - 219 -
*6 Conclusion de chapitre - 221 -*
BILAN DE PARTIE - 225 -

**CONCLUSION : LES GAINS D'UNE APPROCHE INTERDISCIPLINAIRE - 229 -**

**ANNEXES - 235 -**

***1 Annexe C8-S5-gui : Architecture logicielle et manuel d'utilisation de GUI_ - 235 -***
1.1 Architecture fonctionnelle - 235 -
1.2 Prototype - 236 -
1.2.1 Onglet « Home » - 236 -
1.2.2 Onglet « Trajectoire » - 237 -
1.2.3 Onglet « Orientation main » - 238 -
1.2.4 Onglet « Configuration mains » - 239 -
1.2.5 Onglet « Position dans l'espace » - 239 -
1.3 Axes de développement pour la diffusion initiale - 240 -

1.3.1 Documentation - 240 -
1.3.2 Gestion de la standardisation des annotations chargées - 240 -
1.3.3 Chargement de données 3D - 240 -
1.3.4 Ajout de fonctionnalité d'export d'annotation et de fichiers - 240 -
1.3.5 Ajout de fonctionnalités d'analyses - 241 -

**BIBLIOGRAPHIE - 243 -**

**RESUME - 259 -**

**ABSTRACT - 261 -**

**TABLE DES MATIERES - 263 -**

**TABLE DES FIGURES - 271 -**

**TABLE DES TABLEAUX - 275 -**

# Table des figures

FIGURE 1. ILLUSTRATION DES GESTES POUVANT ETRE PROPRES A CHAQUE COMMUNAUTE ..... - 29 -
FIGURE 2. REPRESENTATION DE LA SEGMENTATION DU GESTE DANS LE TEMPS, BASE SUR LES TRAVAUX DE KENDON [7, 8] ..... - 30 -
FIGURE 3. EXEMPLE DE PHRASE GESTUELLE INCLUANT UNE PREPARATION ET UNE REALISATION .... - 31 -
FIGURE 4. ILLUSTRATION DE L'ESPACE GESTUEL DECRIT PAR MCNEILL [9, 10] ..... - 32 -
FIGURE 5. CATEGORIES D'ENCODAGE DE L'ARM SWIVEL [59] ..... - 33 -
FIGURE 6. (A) VISUELS INITIAUX DE L'AGENT GRETA [17] ET L'UNE DE SES MISES A JOUR. (B) ILLUSTRATION DE L'UTILISATION DU BEHAVIOR MARKUP LANGUAGE (BML [95]) POUR LA GENERATION DE SEQUENCE GESTUELLE ..... - 43 -
FIGURE 7. ARCHITECTURE DE RNN STANDARD ET DEPLIE, ET DE RNN BIDIRECTIONNEL (FLECHES : MATRICES DE POIDS MP) ..... - 44 -
FIGURE 8. ARCHITECTURE D'UN TRANSFORMER [102] ..... - 45 -
FIGURE 9. ILLUSTRATION DE L'EXTRACTION DE CARACTERISTIQUES PAR UN CNN, SUR UNE IMAGE 2D, POUR LA CLASSIFICATION ..... - 46 -
FIGURE 10. CAPTURE D'ECRAN DE L'INTERFACE UTILISEE LORS DE L'EVALUATION PERCEPTIVE [115, 130] ..... - 53 -
FIGURE 11. INTERFACE UTILISEE PAR LES AUTEURS DE BEAT POUR L'ANNOTATION DES GESTES [150] - 83 -
FIGURE 12. REPRESENTATION VISUELLE DE LA DISTRIBUTION DES DUREES DE GESTES DANS LE CORPUS [153] ..... - 86 -
FIGURE 13. COMPARAISON ENTRE L'ENRICHISSEMENT FONCTIONNEL ORIGINAL ET CELLE PROPOSEE DANS CETTE ANALYSE [153] ..... - 91 -
FIGURE 14. GESTE ICONIQUE REALISE DANS LA SEQUENCE REANNOTEE [153] ..... - 92 -
FIGURE 15. GESTE METAPHORIQUE REALISE DANS LA SEQUENCE REANNOTEE ..... - 92 -
FIGURE 16. FORMES DE TRAJECTOIRES IDENTIFIEES DANS LE CORPUS ..... - 104 -
FIGURE 17. TYPES DE MANUALITES ANNOTEES, EXEMPLES THEORIQUES DE TRAJECTOIRES ..... - 105 -
FIGURE 18. EXEMPLES ILLUSTRES D'ORIENTATIONS DE LA PAUME ..... - 105 -
FIGURE 19. NOMBRE DE GESTES PAR FICHIER ET DUREE DE FICHIER (SECONDES) ..... - 116 -
FIGURE 20. BOXPLOTS DES DUREES DE GESTES DANS LE CORPUS PAR TYPE FONCTIONNEL ..... - 119 -
FIGURE 21. BOXPLOTS DES SCORES DE CERTITUDE PAR TYPE FONCTIONNEL ..... - 120 -
FIGURE 22. ILLUSTRATION DE FORMES DE GESTES ..... - 124 -
FIGURE 23. ILLUSTRATION DE GESTE PHATIQUE AVEC UNE POSITION DE MAIN DROITE (RH) ANNOTEE COMME VERTICALE (V) ..... - 126 -
FIGURE 24. GESTE DE PARSING AVEC UNE ORIENTATION DE LA PAUME DROITE (RH) PARTIELLEMENT VERS LE HAUT (MU) ..... - 129 -
FIGURE 25. GESTE PHATIQUE (A) AVEC L'INDEX ENGAGE (IE) POUR LA MAIN DROITE (RH) ..... - 131 -
FIGURE 26. ANALYSE TEMPORELLE DE LA VITESSE MOYENNE DES GESTES EN FONCTION DU TYPE FONCTIONNEL ..... - 133 -
FIGURE 27. BOXPLOT DES DISTANCES PARCOURUES (CM) EN FONCTION DU TYPE FONCTIONNEL ANNOTE ..... - 134 -
FIGURE 28. VISUALISATION DES TRAJECTOIRES PAR TYPE FONCTIONNEL ANNOTE ..... - 137 -
FIGURE 29. ILLUSTRATION DE GESTES SANS AFFILIATION LEXICALE EXPLICITE ..... - 140 -
FIGURE 30. EXTRAITS DES TIERS D'ANNOTATION DU FICHIER 2, AVEC L'ANNOTATION COMMUNE DES GESTES NATURELS (NAT), L'ANNOTATION COMMUNE DES GESTES DE SYNTHESE (SYNT), ET LES ANNOTATIONS INDIVIDUELLES DES ANNOTATRICES (_0 ET _1) POUR LES GESTES DE SYNTHESE. .. - 151 -

FIGURE 31. DISTRIBUTION DES SCORES DE CERTITUDE ..... - 157 -
FIGURE 32. TRAJECTOIRES PENDANT LES ACTIVITES ANNOTEES [164] ..... - 158 -
FIGURE 33. DISTANCE (CM) DE L'INDEX DROIT PAR RAPPORT A LA HANCHE ET PAR FRAME [164] .... - 159 -
FIGURE 34. VITESSE (CM/S) DE L'INDEX DROIT PAR FRAME [164] ..... - 160 -
FIGURE 35. ILLUSTRATION DU GESTE N° 5 DANS LE FICHIER DE SYNTHESE ..... - 161 -
FIGURE 36. ILLUSTRATION DU GESTE N° 25 DANS LE FICHIER NATUREL ..... - 162 -
FIGURE 37. SIX ASPECTS IDENTIFIES POUR DISTINGUER LES VARIANTS DANS CHAQUE CATEGORIE DE TRAJECTOIRE ..... - 174 -
FIGURE 38. ILLUSTRATION DE TRAJECTOIRES SELON LA MANUALITE ..... - 175 -
FIGURE 39 .ILLUSTRATION DE L'ETENDUE DE L'ESPACE GESTUEL ET MENSURATIONS DU LOCUTEUR WAYNE (CM) ..... - 176 -
FIGURE 40. PROBABILITES DE POSITIONS DES JOINTS SUR L'AXE X ..... - 178 -
FIGURE 41. PROBABILITES DE POSITIONS SUR L'AXE Y ..... - 179 -
FIGURE 42. PROBABILITES DE POSITIONS SUR L'AXE Y ..... - 179 -
FIGURE 43. SCHEMA DE LA CARTOGRAPHIE GLOBALE, BASEE SUR LES OBSERVATIONS DE DENSITE - 180 -
FIGURE 44. SPHERE 3D POUR LA DETERMINATION DE L'ORIENTATION DANS L'ESPACE ..... - 181 -
FIGURE 45. DETERMINATION DES ETIQUETTES SYMBOLIQUES DEPUIS LES VALEURS D'ANGLE ..... - 182 -
FIGURE 46. ILLUSTRATION DES VECTEURS UTILISES POUR LA DETERMINATION DE L'ORIENTATION DU GESTE (VECTEUR ROUGE = DIRECTION PRINCIPALE ; VECTEUR BLEU = DIRECTION SECONDAIRE) . - 182 -
FIGURE 47. ILLUSTRATION DES VECTEURS POUR LA TRAJECTOIRE GLOBALE DE LA MAIN (VECTEUR BLEU) ET DE LA PAUME (VECTEUR ROUGE) ..... - 183 -
FIGURE 48. ILLUSTRATION DES MARQUEURS UTILISES POUR LE CALCUL DE DISTANCE ..... - 183 -
FIGURE 49. VUE D'ENSEMBLE DE L'INTERFACE ET DE CERTAINS MODULES QUI LA COMPOSE ..... - 187 -
FIGURE 50. TRAJECTOIRES DES MAINS D'UN GESTE SELON LES PLANS FRONTAL ET TRANSVERSAL - 188 -
FIGURE 51. SCHEMA DE L'ENCODEUR DE MOUVEMENT ..... - 199 -
FIGURE 52. SCHEMA DE L'ARCHITECTURE DU DECODEUR ..... - 201 -
FIGURE 53. SCHEMA DE BINARISATION ET DETERMINATION DES GESTES PREDITS ..... - 202 -
FIGURE 54. VALEURS DE LA LOSS PENDANT L'ENTRAINEMENT (BLEU) ET LA VALIDATION (ORANGE) - 206 -
FIGURE 55. VALEURS DES METRIQUES PENDANT L'ENTRAINEMENT (BLEU) ET LA VALIDATION (ORANGE) - 206 -
FIGURE 56. REPRESENTATION GRAPHIQUE DES DISTRIBUTIONS DES VALEURS MOYENNES DE METRIQUES (A = ACCURACY ; F1 = F1-SCORE, P = PRECISION ; R = RAPPEL ; IOU = INTERSECTION OVER UNION) ..... - 209 -
FIGURE 57. EXEMPLE DE PREDICTION ET D'ANNOTATIONS EXPERTES SUR UNE SEQUENCE ..... - 210 -
FIGURE 58 EXEMPLES DE TRAJECTOIRES ENTRE PREDICTIONS ET ANNOTATIONS EXPERTES (G1 ET G3) .. - 211 -
FIGURE 59. EXEMPLES DE TRAJECTOIRES ENTRE PREDICTIONS ET ANNOTATIONS EXPERTES (G5 ET G8) . - 212 -
FIGURE 60. VALEURS DES METRIQUES DE F1 ET IOU EN FONCTION DE LA COMPLEXITE DES MODELES (FW = FRAME-WISE ; SW = SEGMENT-WISE ; IOU = INTERSECTION OVER UNION) ..... - 214 -
FIGURE 61. VALEURS DE LOSS EN FONCTION DES REGULARISATIONS L2 APPLIQUEES ..... - 216 -
FIGURE 62. ARCHITECTURE FONCTIONNELLE CONCEPTUELLE POUR LE DEVELOPPEMENT DU PROTOTYPE ..... - 235 -
FIGURE 63. ONGLET HOME ET TABLE DE SELECTION DE GESTE ..... - 236 -
FIGURE 64. ONGLET TRAJECTOIRE ..... - 237 -
FIGURE 65. ONGLET FORMAL ANALYSIS ..... - 237 -

FIGURE 66. SKELETONWIDGET POUR L'AFFICHAGE DES VECTEURS D'ORIENTATION DE LA MAIN (VERT) ET DE LA PAUME (ROUGE) ..... - 238 -
FIGURE 67. ONGLET CONFIGURATION MAINS..... - 239 -
FIGURE 68. ONGLET POSITION DANS L'ESPACE ..... - 239 -

# Table des tableaux

TABLEAU 1. INSTRUCTIONS D'ANNOTATION SEMANTIQUES POUR LES ANNOTATEURS DE BEAT ..... - 83 -
TABLEAU 2. STATISTIQUES DESCRIPTIVES GLOBALES DES ANNOTATIONS FONCTIONNELLES DE BEAT [153] ..... - 85 -
TABLEAU 3. RECOUVREMENT TEMPOREL DES ANNOTATIONS DANS LE CORPUS [153] ..... - 85 -
TABLEAU 4. ANALYSE DES DUREES DE GESTES (SECONDES) DANS LE CORPUS BEAT [153] ..... - 87 -
TABLEAU 5. ANALYSE DES VALEURS ABERRANTES DE DUREES DE GESTES PAR CATEGORIE FONCTIONNELLE [153] ..... - 88 -
TABLEAU 6. ANALYSE DESCRIPTIVE DE L'AFFILIATION LEXICALE DANS LE CORPUS ET DES DETECTIONS DE SPACY [153] ..... - 89 -
TABLEAU 7. LISTE DES FICHIERS REANNOTES ET DE LEUR DUREE ..... - 95 -
TABLEAU 8. TIERS DEFINIS DANS LE CADRE DE LA TRANSCRIPTION ET DE L'ANNOTATION DE LA PAROLE - 96 -
TABLEAU 9. JEU DE BALISES POUR L'ANNOTATION DE PHENOMENES ORAUX ..... - 98 -
TABLEAU 10. TEMPS D'ANNOTATION POUR CHAQUE FICHIER PENDANT LA PHASE DE SYNCHRONISATION ..... - 99 -
TABLEAU 11. ANALYSE DU TEMPS D'ANNOTATION PENDANT LA PHASE D'ANNOTATION INDEPENDANTE - 100 -
TABLEAU 12. TIERS D'ANNOTATION GESTUELLE ET LEUR USAGE ..... - 101 -
TABLEAU 13. DISTRIBUTION DU NOMBRE DE GESTES PAR FICHIER (MIN = MINIMUM ; MAX = MAXIMUM ; MOY = MOYENNE ; MED = MEDIANE ; SD = ECART-TYPE ; N = NOMBRE D'OCCURRENCES ; TOT = TOTAL). ..... - 116 -
TABLEAU 14. DISTRIBUTION DES ETIQUETTES DE TYPES FONCTIONNELS DANS LE CORPUS ..... - 117 -
TABLEAU 15. DISTRIBUTION CROISEE DES TYPES FONCTIONNELS EN DIMENSION 1 (D1) ET EN DIMENSION 2 (D2) ..... - 118 -
TABLEAU 16. STATISTIQUES DESCRIPTIVES DES DUREES DE GESTES ..... - 120 -
TABLEAU 17. ANALYSE DES SCORES DE CERTITUDE PAR TYPE FONCTIONNEL ..... - 121 -
TABLEAU 18. REPARTITION DES ETIQUETTES DE MANUALITE DANS LE CORPUS ..... - 121 -
TABLEAU 19. DISTRIBUTION DE LA MANUALITE EN FONCTION DU TYPE FONCTIONNEL DU GESTE .. - 122 -
TABLEAU 20. DISTRIBUTION DES LABELS DE LA FORME DE LA TRAJECTOIRE DU GESTE DANS LE CORPUS ..... - 123 -
TABLEAU 21. DISTRIBUTION DES LABELS DE FORME EN FONCTION DES TYPES FONCTIONNELS .... - 123 -
TABLEAU 22. DISTRIBUTION DES LABELS DE LA POSITION DE LA MAIN DANS LE CORPUS ..... - 125 -
TABLEAU 23. DISPOSITION DES LABELS DE POSITION DE LA MAIN EN FONCTION DU TYPE FONCTIONNEL ..... - 126 -
TABLEAU 24. DISTRIBUTION DES LABELS D'ORIENTATION DE LA PAUME DANS LE CORPUS (D = VERS LE BAS ; E = SUR LA TRANCHE ; F = VERS L'AVANT ; MD = PARTIELLEMENT VERS LE BAS ; MU = PARTIELLEMENT VERS LE HAUT ; SO = VERS LE LOCUTEUR ; U = VERS LE HAUT) ..... - 127 -
TABLEAU 25. DISTRIBUTION DES LABELS D'ORIENTATION DE LA PAUME EN FONCTION DU TYPE FONCTIONNEL (D = VERS LE BAS ; E = SUR LA TRANCHE ; F = VERS L'AVANT ; MD = PARTIELLEMENT VERS LE BAS ; MU = PARTIELLEMENT VERS LE HAUT ; SO = VERS LE LOCUTEUR ; U = VERS LE HAUT)- 127 -
TABLEAU 26. DISTRIBUTION DES LABELS DE CONFIGURATION DE LA MAIN DANS LE CORPUS (CH = MAIN FERMEE ; IE = INDEX ENGAGE ; MCH = MAIN PARTIELLEMENT FERMEE ; OH = MAIN OUVERTE ; SEI = INDEX PARTIELLEMENT ENGAGE ; SOH = MAIN PARTIELLEMENT OUVERTE ; TE = POUCE ENGAGE ; TIE = POUCE ET INDEX ENGAGES ; TMIE = POUCE, MAJEUR ET INDEX ENGAGES) ..... - 129 -

TABLEAU 27. DISTRIBUTION DES LABELS DE CONFIGURATION DE LA MAIN EN FONCTION DU TYPE FONCTIONNEL (CH = MAIN FERMEE ; IE = INDEX ENGAGE ; MCH = MAIN PARTIELLEMENT FERMEE ; OH = MAIN OUVERTE ; SEI = INDEX PARTIELLEMENT ENGAGE ; SOH = MAIN PARTIELLEMENT OUVERTE ; TE = POUCE ENGAGE ; TIE = POUCE ET INDEX ENGAGES ; TMIE = POUCE, MAJEUR ET INDEX ENGAGES) ..... - 130 -
TABLEAU 28. STATISTIQUES DESCRIPTIVES DES DISTANCES PARCOURUES (CM) PAR TYPE FONCTIONNEL ..... - 135 -
TABLEAU 29. ANALYSE DESCRIPTIVE DE L'AMPLITUDE MAXIMALE EN FONCTION DU TYPE FONCTIONNEL ..... - 136 -
TABLEAU 30. ANALYSE DU NOMBRE DE MOTS PAR AFFILIE ..... - 139 -
TABLEAU 31. DECOMPTE DE TYPES FONCTIONNELS AVEC UNE AFFILIATION LEXICALE ..... - 140 -
TABLEAU 32. CARACTERISTIQUES MAJORITAIRES DES TYPES FONCTIONNELS DETERMINEES LORS DES DIFFERENTES ANALYSES (RH = MAIN DROITE ; MH = MAINS EN MIROIR ; LH = MAIN GAUCHE ; H = HORIZONTAL ; I = INTERMEDIAIRE ; V = VERTICAL ; E = PAUME SUR LA TRANCHE ; SO = PAUME VERS LE LOCUTEUR ; U = PAUME VERS LE HAUT ; MU = PAUME PARTIELLEMENT VERS LE HAUT ; FR = PAUME VERS L'INTERLOCUTEUR ; SOH = MAIN PARTIELLEMENT OUVERTE ; OH = MAIN OUVERTE ; IE = INDEX ENGAGE) ..... - 142 -
TABLEAU 33. DESCRIPTION DES FICHIERS NATUREL ET SYNTHETIQUE [164] ..... - 149 -
TABLEAU 34. VARIABILITE DES REPERES TEMPORELS D'ANNOTATION DES GESTES [164] ..... - 152 -
TABLEAU 35. PHASE DES ACTIVITES ANNOTEES [164] ..... - 153 -
TABLEAU 36. CARACTERISATION DES ACTIVITES [164] ..... - 154 -
TABLEAU 37. DETAIL DES TYPES FONCTIONNELS ANNOTES [164] ..... - 155 -
TABLEAU 38. ACTIVITES ANNOTEES PAR PHRASES ET NOMBRE D'AFFILIATIONS LEXICALES [164] ... - 156 -
TABLEAU 39. STATISTIQUES DESCRIPTIVES DES SCORES DE CERTITUDE (N = NOMBRE D'OCCURRENCE ; MOY = MOYENNE ; MED = MEDIANE ; D1 = DIMENSION 1 ; D2 = DIMENSION 2) [164] ..... - 157 -
TABLEAU 40. MANUALITES DES GESTES ANNOTES (LH = MAIN GAUCHE ; RH = MAIN DROITE) ..... - 157 -
TABLEAU 41. ANALYSE DU TEMPS D'ANNOTATION ET DE REVISION DES FICHIERS [164] ..... - 163 -
TABLEAU 42. PROPRIETES DYNAMIQUES DU GESTE DANS LE CAS DE REBEAT-14 ..... - 185 -
TABLEAU 43. ENSEMBLES DE DONNEES ..... - 197 -
TABLEAU 44. DIMENSIONS D'ENTREE ET DE SORTIE DES VECTEURS PAR BLOC RESIDUEL (RB) ..... - 199 -
TABLEAU 45. VALEURS DES METRIQUES SUR LE SET DE TEST (IOU = INTERSECTION OVER UNION) . - 207 -
TABLEAU 46. VALEURS MOYENNES DES METRIQUES LORS D'EXECUTIONS MULTIPLES SUR LE SET DE TEST ..... - 209 -
TABLEAU 47. CONDITIONS TESTEES POUR LA PROFONDEUR DU RESEAU (N PARAM. = NOMBRE DE PARAMETRES) ..... - 213 -
TABLEAU 48. PERFORMANCES EN FONCTION DU LEARNING RATE (LR) ..... - 215 -
TABLEAU 49. PERFORMANCES DU MODELE BASELINE SELON LA REGULARISATION L2 APPLIQUEE (F1-FRAME = F1-SCORE APPLIQUE AUX FRAMES ; F1-SEGMENT = F1-SCORE APPLIQUE AUX SEGMENTS ; IOU MOYEN = INTERACTION OVER UNION MOYEN) ..... - 216 -
TABLEAU 50. PERFORMANCES DU MODELE BASELINE SELON LE DROPOUT (F1-FRAME = F1-SCORE APPLIQUE AUX FRAMES ; F1-SEGMENT = F1-SCORE APPLIQUE AUX SEGMENTS ; IOU MOYEN = INTERACTION OVER UNION MOYEN) ..... - 217 -
TABLEAU 51. PERFORMANCES DU MODELE SUR UN LOCUTEUR INEDIT (IOU MOYEN = INTERACTION OVER UNION MOYEN) ..... - 218 -
TABLEAU 52. TAILLES DES DONNEES ET DES COUCHES DE LEUR ENCODEUR CORRESPONDANT POUR LES NOUVELLES MODALITES DE DONNEES TESTEES (RB = BLOCS RESIDUELS ; F0 = FREQUENCE FONDAMENTALE ; MFCC = MEL-FREQUENCY CEPSTRAL COEFFICIENT ; QUAT = QUATERNIONS) ... - 219 -

TABLEAU 53. METRIQUES FRAME-WISE DES MODELES TESTES AVEC D'AUTRES TYPES DE DONNEES D'ENTRAINEMENT (F1 FRAMES = F1-SCORE APPLIQUE AUX FRAMES ; BA = BASELINE ; F0 = FREQUENCE FONDAMENTALE ; MFCC = MEL-FREQUENCY CEPSTRAL COEFFICIENT ; QUAT = QUATERNIONS) ..... - 220 -