Projet ANR-06-CORP-012

CREAGEST

Programme ANR-Corpus 2006

A IDENTIFICATION

Acronyme du projet	CREAGEST
Titre du projet	Réalisation de corpus de données visuelles pour l'analyse des processus de création d'unités gestuelles (LSF et gestualité naturelle).
Coordinateur du projet (société/organisme)	Christian CUXAC et Brigitte GARCIA — UMR 7023 SFL (Université Paris 8 et CNRS)
Période du projet (date de début – date de fin)	22 janvier 2007-21 janvier 2012
Site web du projet, le cas échéant	

Rédacteur de ce rapport	
Civilité, prénom, nom	Brigitte Garcia Liste des autres rédacteurs du rapport : Sallandre, M.-A., Balvet, A., Boutet, D., L'Huillier, M.-T., Courtin, C., Fusellier-Souza, I., Vincent, C., Schoder, C., Makouke, D.
Téléphone	
Adresse électronique	Brigitte.Garcia@univ-paris8.fr
Date de rédaction	Mars 2012

Si différent du rédacteur, indiquer un contact pour le projet	
Civilité, prénom, nom	
Téléphone	
Adresse électronique	

Liste des partenaires présents à la fin du projet (société/organisme et responsable scientifique)	—UMR 7023 SFL (U. Paris 8 et CNRS), C. Cuxac et B. Garcia (Partenaire 1) —UMR 8163 STL (U. Lille 3 et CNRS), Antonio BALVET (Partenaire 2) *Nota* : décès en cours de projet (décembre 2010) du responsable du Partenaire 3, Cyril COURTIN (UMR GIN, U. Caen et Paris 5 et CNRS).

B RESUME CONSOLIDE PUBLIC

B.1 RESUME CONSOLIDE PUBLIC EN FRANÇAIS

Corpus LSF et Gestualité : comprendre la sémiotisation de la gestualité humaine

1-Constitution de corpus de LSF (enfants et adultes) et de gestualité naturelle pour l'analyse des processus de création d'unités gestuelles. Le propos était de *constituer et documenter un corpus de données gestuelles* incluant *discours en LSF* d'enfants et d'adultes sourds *et gestualité naturelle,* avec pour enjeu de mettre à disposition des communautés françaises et internationales concernées des corpus représentatifs et pérennes. Il s'agissait de

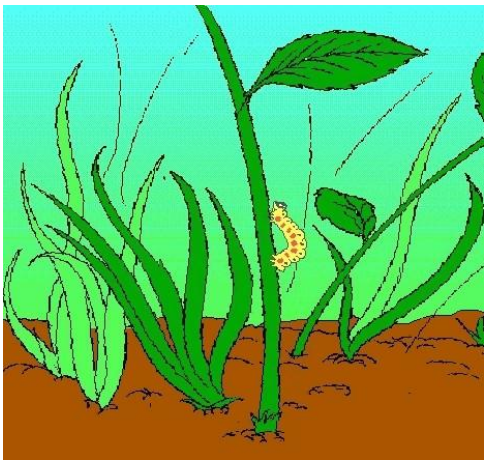
pallier les manques majeurs en la matière et de complémenter l'existant, aussi bien en termes d'âges représentés que de genres discursifs et d'implantation régionale des locuteurs. Les choix méthodologiques opérés visaient à apporter des réponses à deux types d'objectifs :
1) l'un, de modélisation linguistique : affiner sur la base du *modèle sémiologique* (Cuxac 2000) la description des processus opérant dans la sémiotisation de la gestualité humaine, tant au plan des interrelations entre gestualité naturelle et LS, qu'à ceux de l'acquisition de la LSF et des émergences et évolutions dans la LSF adulte ;
2) l'autre, centré sur la documentation des corpus visant à explorer les méthodologies d'annotation, à pérenniser les données et à les rendre accessibles aux communautés de chercheurs et de locuteurs et surtout, parmi ces derniers, les enseignants de et en LSF.

2-Choix techniques et méthodologiques pour la constitution, la documentation, la pérennisation et l'édition des corpus. Les méthodes de recueil des données ont été élaborées selon les objectifs propres à chaque sous-projet (LSF enfantine : stimuli pour des tâches descriptives et interaction avec une poupée, en présence d'un adulte ; gestualité naturelle : protocole de test selon une méthode des juges et dialogue dirigé par une tâche de description ; LSF adulte : entretien semi-directif conduit par un enquêteur sourd). Pour les corpus de LSF, la captation est toujours faite selon plusieurs angles de vue (2/3 caméras). Les corpus sont renseignés selon la norme préconisée par le TGE-ADONIS selon les recommandations du CINES pour l'archivage pérenne (norme de métadonnées OLAC, archivage ouvert selon la norme OAI-PMH) en vue d'un archivage, d'une indexation et d'un référencement des corpus sur des métaportails ouverts (Google Scholar, OAIster database). Les annotations de corpus sont faites sous ELAN, selon des normes ouvertes, partageables, interrogeables y compris à distance (mise en ligne des annotations à moyen terme). L'absence de système de transcription *stricto sensu* contraint à recourir à des logiciels empilant sur une même image les moments clés d'un signe en mouvement.

3- Résultats majeurs du projet
Nous avons constitué un corpus de données gestuelles d'une ampleur quantitative et qualitative inédite (voir **4.**). Outre la valeur patrimoniale de ce corpus et le vivier de ressources pédagogiques ainsi créé, *CREAGEST* a permis d'élaborer et tester des schémas d'annotation originaux (langue sans tradition écrite, geste) et d'ouvrir des pistes novatrices (cf. Étude préliminaire d'un module d'annotation en 3D permettant d'enregistrer les coordonnées des positions des 3 degrés de liberté de la main et de l'amplitude des mouvements). Le projet a par ailleurs permis des avancées descriptives notables (expression du déplacement en LSF enfantine, structuration du geste, processus de la lexicogenèse). Enfin, il a donné lieu au développement partiel et/ou à l'élaboration d'un cahier des charges plus ajusté pour des outils d'optimisation du logiciel d'annotation utilisé, ELAN.

4-Production scientifique et brevets depuis le début du projet
La production scientifique majeure est un corpus numérisé et pérenne, partiellement annoté (± 1h), de 365 h de LSF adulte et enfantine et 5h de gestualité naturelle, assorti de métadonnées détaillées débouchant sur des propositions de complémentation et normalisation des critères existants (OLAC et IMDI). Les données de LSF ont une représentativité sans précédent par la diversité des âges (enfants de 3 à 15 ans, adultes de 18 à 60 ans), des origines géographiques (4 régions de France) et des genres discursifs (descriptif, explicatif, métalinguistique, narratif).

5-Illustration



Légende : Arrêt sur image de l'enfant E37, 10.3 ans, corpus CELEM, caméra en plan rapproché.
Description : (Image de gauche) Arrêt sur image de l'item vidéo « la chenille grimpe à la tige » (d'après Hickmann et al. 2009, avec autorisation des auteurs)
(Image de droite) L'enfant signe le déplacement volontaire correspondant à l'image de gauche ; sa main droite figure la chenille qui monte tandis que sa main gauche figure le support de la tige. Cette structure est un transfert situationnel, dans la terminologie de Cuxac (2000).

6-Informations factuelles
Le projet *Creagest* est un projet alliant recherche fondamentale (meilleure compréhension de la sémiogenèse de la gestualité et des LS) et développement exploratoire (procédures d'annotation d'une langue visuo-gestuelle et non écrite et du geste, et de pérennisation de corpus de données visuo-gestuelles). Coordonné par B. Garcia et C. Cuxac (UMR 7023 SFL, U. Paris 8 et CNRS), le projet a associé 3 partenaires : l'UMR SFL, l'UMR 8163 STL (U. Lille 3 et CNRS, Antonio Balvet) et l'UMR GIN (U. Caen et Paris 5, Cyril Courtin, disparu en décembre 2010). Commencé en janvier 2007, CREAGEST a duré 60 mois. Ayant en effet bénéficié d'une année de dérogation (avenant N°1 du 5 février 2009 à la décision attributive d'aide (n° ANR-06-CORP-012-01 du 22 janvier 2007), il s'est achevé le 21 janvier 2012.
Ce projet a bénéficié d'une aide de l'ANR de 200 000 €, pour un coût global de 486 945 €.

B.2 RESUME CONSOLIDE PUBLIC EN ANGLAIS

A corpus study of LSF and Gesturality: understanding the semiotisation process of human gesturality

1-Building an LSF and a natural gesturality corpus for analysing the creation of gestural units
In this project, the goal was to set up and document a corpus of gestural data, including LSF discourse, both from children and adult deaf speakers, in order to provide representative corpora to SL research communities. We strove to sustain research in the area of human

gesturality, by extending the range of existing realizations as well as creating original corpora where needed. We therefore paid special attention to experimental variables such as speakers's ages and regions of origin, as well as discourse genres.
Our methodology options addressed two central issues:
1) modeling linguistic data; based on the semiological model in (Cuxac 2000), we wished to support a finer-grained description of the semiotization process in human gesturality. We therefore devised three sub-corpora, aimed at the interplay between natural gesturality and LSF, at the LSF acquisition process, as well as at the creative processes and evolution of adult LSF.
2) documenting corpora; we explored annotation methodologies so that our corpora could best be used and accessed by a large audience, bearing in mind sustainability issues as these data will be stored over long periods of time. In this process, we strove to ensure maximum accessibility to our data for the different SL research communities, as well as for deaf speakers and LSF teachers in classroom situations.

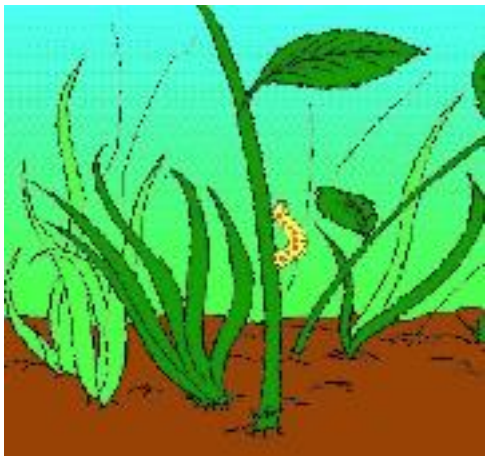
2-Technical and methodological options for setting up, documenting, and editing sustainable corpora
The methodology for data collection is adapted to each subproject's main objective: child Sign Language SP1: description tasks and interaction with a doll in the presence of an adult; natural gesturality SP2: tests following an ??[assessor-based method]?? and elicited dialogue based on a description-oriented task; adult LSF SP3: semi-directive interviews led by a deaf investigator.
For LSF corpora, recordings were always performed using different viewing angles (2/3 cameras). Each corpus is documented following the method proposed by TGE-Adonis, based on the recommendations from the CINES for sustainable archiving (OLAC metadata, open archives following the OAI-PMH standard), so that open meta-portals (Google Scholar, OAIster database) can archive, index and reference the collected data.
Corpora are annotated with ELAN, enforcing open, shared and searchable annotations (both local and remote searches can be performed) based on international standards; annotations will eventually be made available on the web in a near future. Due to the absence of established transcription systems, we had to resort to software allowing us to associate the different key moments of a sign to a given image.

3-Main outcomes of the project
We have set up an outstanding corpus of gestural data, both from the quantitative and the qualitative viewpoints (cf. 4.). This corpus is both a testimony on contemporary LSF and a repository of teaching material. In addition, Creagest allowed us to define and test original annotation schemes (language lacking a written tradition, gesture) and to explore radically new venues in SL research (cf. preliminary study on a 3D annotation module for the hand's 3 degrees of freedom, as well as for movement intensity). This project has also sustained progress from the descriptive viewpoint (expression of movement in child Sign Language, gesture structure, lexicogenesis process). Finally, it has allowed the definition of a specification for future annotation tools, extending the ELAN platform.

4-Scientific outcomes and patents since the project's onset
The central scientific outcome is a partially annotated (circa 1h), digitized and sustainable corpus of 365 hours of child and adult LSF, with the addition of 5 hours of natural

gesturality. This body of data is associated with detailed metadata, which complement and extend existing metadata definition frameworks (OLAC and IMDI). The LSF data in this corpus exhibit unprecedented representativeness thanks to the span of speakers's ages (children from 3 to 15 years of age, adults from 18 to 60 years of age), regions of origin (4 French regions) and discourse genres (descriptive, explicative, metalinguistic and narrative genres).

5-Illustration



Caption: stop-motion picture of child E37, 10.3 years of age, corpus CELEM, close-up camera.
Description: (leftmost image) stop-motion picture of the video item "the caterpillar is climbing on the stalk" (from Hickmann et al. 2009, authors's permission granted). (rightmost image) the child is signing voluntary movement corresponding to the leftmost image. His right hand represents the caterpillar, while his left hand stands for the stalk's support. This structure is a situational transfer, in Cuxac (2000) terminology.

6-Factual information

The Creagest project has a fundamental nature, as it allows for a better understanding of semiogenesis, both in natural gestures and in SLs. Moreover, it has sustained exploratory developments: annotation procedures for a visual-gestural language and for gesture in general, sustainability issues for corpora and visual-gestural data. It is coordinated by B. Garcia and C. Cuxac (UMR 7023 SFL, U. Paris 8 and CNRS) and comprises 3 partners: UMR

SFL, UMR 8163 STL (U. Lille 3 and CNRS, Antonio Balvet) and UMR GIN (U. Caen and Paris 5, Cyril Courtin, deceased in December 2010). The project started in January 2007, and lasted for 60 months, as the original length of the project has been extended (cf. addendum 1 February 2009 to the original attributive decision (ANR-06-CORP-012-01, January 22nd 2007)). The project stopped on January 21st 2012. This project was alloted a budget of 200,000 € by the ANR, for a total cost of 486,945 €.

C MEMOIRE SCIENTIFIQUE

Mémoire scientifique confidentiel : non

C.1 RESUME DU MEMOIRE

L'objectif de *Creagest* était de *constituer et documenter un corpus de données gestuelles* incluant *discours* d'enfants et d'adultes sourds *en LSF et gestualité naturelle*. Il s'agissait de complémenter l'existant, en termes d'âges et de genres discursifs représentés comme d'implantation régionale des locuteurs. Outre cet enjeu, les choix des types de corpus étaient tributaires d'un objectif descriptif transversal : approfondir, sur la base du *modèle sémiologique* (Cuxac 2000), notre compréhension des processus opérant dans la sémiotisation de la gestualité humaine (liens gestualité naturelle/LS, acquisition, lexicogenèse).
Les méthodes de recueil ont été élaborées selon les divers objectifs spécifiques (e-g stimuli pour des tâches descriptives, entretiens semi-dirigés). La captation vidéo est toujours faite selon plusieurs angles de vue. Les corpus sont renseignés selon la norme préconisée par le TGE-ADONIS selon les recommandations du CINES (norme de métadonnées OLAC) en vue d'un archivage, d'une indexation et d'un référencement des corpus sur des métaportails ouverts (Google Scholar, OAIster database). Les annotations sont réalisées sous Elan, selon des normes ouvertes, partageables, interrogeables y compris à distance.
Est ainsi constitué au final un corpus pérenne de données gestuelles d'une ampleur inédite (365 h de LSF adulte et enfantine et 5h de gestualité naturelle), partiellement annoté (± 1h), et assorti de métadonnées détaillées. Les données de LSF ont une représentativité sans précédent par la diversité des âges (de 3 à 60 ans), des origines géographiques (4 régions) et des genres discursifs. Outre la valeur patrimoniale de ce corpus et le vivier de ressources pédagogiques ainsi créé, *Creagest* a permis d'élaborer et tester des schémas d'annotation originaux pour cette langue gestuelle et non écrite et d'ouvrir des pistes novatrices (ex. Étude d'un module d'annotation en 3D permettant d'enregistrer les coordonnées des positions des 3 degrés de liberté de la main et de l'amplitude des mouvements). Il a par ailleurs permis de notables avancées descriptives (expression du déplacement en LSF enfantine, structuration du geste, processus de la lexicogenèse). Enfin, il a donné lieu au développement partiel ou à l'élaboration d'un cahier des charges plus ajusté pour des outils d'optimisation du logiciel d'annotation utilisé, Elan.

C.2 ENJEUX ET PROBLEMATIQUE, ETAT DE L'ART

CREAGEST est un projet de constitution et documentation de corpus de Langue des Signes Française (LSF) et de gestualité naturelle. Le projet devait avant tout répondre *à un important enjeu patrimonial et de mise à disposition de ressources linguistiques réellement représentatives.* Non écrite et non standardisée, la LSF est en effet une langue menacée. Bien qu'elle ait juridiquement recouvré droit de cité dans les écoles (« Loi Fabius » 1991, Loi 2005-102), la réalité du terrain est celle d'une forte pression faite aux familles vers l'implantation

cochléaire précoce et l'intégration individuelle de l'enfant sourd en milieu 'ordinaire', souvent aux dépens de la LSF. Or, malgré l'explosion récente des recherches linguistiques sur la LSF et des actions comme le Projet *LS-Colin* (Cuxac *et al* 2002), les corpus de LSF disponibles demeuraient très fragmentaires, en termes qualitatifs et quantitatifs. Il était donc urgent, *a fortiori* dans un contexte international de développement de corpus de LS, d'œuvrer *à l'illustration de la LSF dans sa diversité* et *à une mise à disposition pérenne des ressources constituées.* Dans ce cadre, *CREAGEST* s'est fixé *deux niveaux corrélés d'objectifs* :

1/ Un objectif de description linguistique : l'approfondissement de notre connaissance des processus à l'œuvre dans la sémiotisation de la gestualité humaine signifiante. Intégrant pour ce faire une prise en compte de la gestualité coverbale, le projet se proposait : de rendre compte de *l'hétérogénéité linguistique de la LSF* (LSF enfantine et adulte, émergences lexicales) ; d'évaluer *la pertinence de l'hypothèse d'une sémiogénèse de la gestualité communicative fondée sur un processus d'iconicisation de l'expérience perceptivo-pratique* (« Modèle sémiologique », Cuxac 2000) en en précisant les modalités aux plans ontogénétique, phylogénétique et de synchronie dynamique ; *d'affiner la compréhension des liens entre gestualité naturelle et LS* en explorant l'hypothèse d'une continuité entre leurs structures respectives (Kendon 1988, McNeill 1992,Volterra & Erting 1994).

2/ Un objectif de constitution/documentation des corpus**.** *Trois sous-projets* (1, 2, 3) *ont été dédiés à la constitution de ces corpus* : l'un à la LSF enfantine, un deuxième à la LSF adulte, un troisième à la gestualité naturelle. Les méthodologies de recueil devaient intégrer les 2 types d'enjeux indiqués : a) la représentativité, *i-e compléter l'existant pour rendre compte de la diversité effective de la LSF* : privilégier les dialogues, l'existant étant surtout monologique, et les genres discursifs peu représentés (métalinguistique, descriptif, explicatif) ; diversifier les âges (pallier notamment la rareté de corpus de LSF enfantine) et les régions (surreprésentation préexistante de signeurs parisiens), et b) les enjeux descriptifs, *i-e collecter des données relatives aux objectifs spécifiques*. *Deux SP étaient consacrés aux aspects techniques* : création d'une plate-forme (4) et optimisation des outils d'annotation (5) .

Deux ordres de facteurs nous ont toutefois conduits à faire évoluer nos objectifs initiaux : 1/ des facteurs externes graves touchant les personnes impliquées dans le projet : sur ces points (congés longue durée de deux membres clés, décès du responsable et seul représentant du Partenaire 2), voir Annexe F1) ; 2/ des raisons liées au caractère pionnier de l'entreprise et aux spécificités de l'objet (LS) et de la population concernée (communauté sourde). *CREAGEST* est une première à divers titres : ampleur inédite des corpus de LSF constitués, panels d'enquêtés issus de tout le territoire, choix de faire réaliser ces corpus par des enquêteurs sourds non chercheurs formés dans le cadre du projet (pour des raisons historiques et socio-éducatives, la majorité des adultes sourds n'ont bénéficié d'aucune formation académique ; la communication à distance est plus complexe avec des personnes sourdes). Le temps dédié à la formation des enquêteurs et le nombre de corpus pilotes requis ont de ce fait été plus importants que prévus, décalant d'autant le planning d'ensemble.

Nous avons donc opéré les modifications suivantes, en cherchant à préserver l'essentiel de nos objectifs : réduction du nombre d'enquêtés et de régions pour les corpus de LSF ; focalisation, conformément à la vocation exploratoire du projet, sur l'approfondissement qualitatif des modalités d'annotation, au détriment de la mise en place d'une plate-forme collaborative de collecte des néologismes et de la constitution des corpus de gestualité naturelle. Par ailleurs, les objectifs du sous-projet du Partenaire 2 étaient de développer et tester des outils d'aide à l'annotation sous Elan et d'exploitation de corpus annotés sous ce logiciel. Or, la dernière version de Elan, élaborée en cours de projet, a rendu caducs plusieurs

des développements logiciels prévus et la mise à disposition par l'institut Max Planck (MPI Nimègue) de la plate-forme *Language Archiving Tools* a également modifié l'offre logicielle en matière d'exploitation des données annotées sous Elan : plusieurs des besoins initialement ciblés se trouvent ainsi couverts par Elan et/ou la plate-forme LAT ; enfin, les entreprises contactées pour assurer les développements logiciels n'ont pu s'engager sur une prestation couvrant l'ensemble du cahier des charges final (cf. Annexe F6).

C.3 APPROCHE SCIENTIFIQUE ET TECHNIQUE

Dans le *modèle sémiologique*, les unités couramment appelées « signes » et que nous désignons comme « unités lexématiques » (UL) ne sont que l'un des 2 grands types d'unités des LS, l'autre étant celui des unités actualisées par les structures dites de « transfert », les unités de transfert (UT).

La question scientifique majeure associée spécifiquement à la constitution de corpus de LSF enfantine est celle de l'émergence et du développement de la référence actancielle et spatiale en LSF, question peu étudiée et toujours sur un faible nombre de sujets. Le modèle sémiologique ouvre de nombreuses hypothèses quant à l'âge d'apparition et de maîtrise des UL et surtout des UT par l'enfant, que ces corpus doivent explorer, en intégrant les apports de la psychologie cognitive (Partenaire 3, C. Courtin). Une part conséquente du travail a consisté d'une part à adapter les méthodologies de recueil des données à cette population particulière (nombreux tests pilotes, cf. Annexe F2), d'autre part à repérer les enfants susceptibles d'être filmés et à collecter les autorisations de tournage, spécifiquement difficiles à obtenir pour cette population sensible (enfantine et sourde). Les tournages ont eu lieu dans des lieux divers (école, domicile, laboratoire), et toujours strictement selon la même méthodologie de passation des stimuli (récits à partir de vidéos animées, récit à partir d'une histoire dessinée et dialogue libre).

Le sous-projet consacré à la LSF adulte aborde la question transversale de l'émergence sur un autre plan, celui des émergences lexicales. Contrairement aux modélisations courantes, de type phonologique (e.g Kegl et al 1999, Sandler 2003), nous proposons de décrire les UL comme relevant d'une compositionnalité de type morphémique et non phonémique. Cette hypothèse du modèle sémiologique reste toutefois à valider et modéliser pour la synchronie. Une étude antérieure sur 4000 signes a amorcé un inventaire des composants morphémiques (Projet RIAM-ANR *LS Script*, Garcia 2005-2007). Conduite sur des signes décontextualisés et stabilisés, cette étude devait être complétée par une observation *in vivo* des principes à l'œuvre dans la genèse et la stabilisation d'UL récentes et/ou en voie de stabilisation, en contexte discursif. Les corpus constitués à cette fin ont, parallèlement, cherché à privilégier des corpus de type dialogique. Le protocole, original, consiste en entretiens semi-directifs, captés par 3 caméras et conduits par un enquêteur sourd vivant dans la région d'origine des enquêtés (cf. Métadonnées Annexe F3). Le guide d'entretien, qui devait être intériorisé par l'enquêteur pour préserver une spontanéité maximale, contient deux phases : a) discussion sur un thème préétabli (secteur de compétences propre à l'enquêté et susceptible d'avoir engendré des émergences lexicales récentes), b) échange métalinguistique à partir des UL émergentes produites, ré-introduites en *stimuli*.

La méthodologie élaborée pour traiter de la gestualité naturelle et de ses interrelations avec la LS était, elle, en deux volets. Il s'agissait d'une part de constituer un corpus de genre explicatif de dyades d'entendants/de sourds/mixtes, qui visait à explorer les stratégies d'inscription gestuelle dans l'espace pour la gestualité symbolique humaine selon l'hypothèse de continuités gestuelles entre gestualité entendante et LS. Un objectif corollaire

est d'élaborer un cadre de notation commun à la gestualité et à la LS. Le second volet consiste en des tests de reconnaissance de gestes par appariement visant à valider une catégorisation gestuelle sémantique basée sur la physiologie articulaire (Boutet 2001). Le protocole mis en place consiste en la présentation vidéo de 148 réalisations gestuelles isolées et hors contexte constituées à partir des 20 Unités Gestuelles du modèle. 427 sujets passent 9 tests d'appariement sémantique (choix d'une étiquette dans un exemplier d'étiquettes pour chaque réalisation gestuelle, chacune des 20 UG étant présentée sous forme de réalisations gestuelles couvrant l'ensemble des combinaisons possibles entre les segments doigts, main, avant-bras, bras et parfois épaules). Le modèle prévoit une stabilisation d'UG fondée sur une auto-organisation articulaire du substrat physiologique. Ces UG répondent à des schémas d'action de mise en mouvement ordonnée et contrainte des degrés de liberté de chaque segment. Elles sont organisées par segment et diffusent sur tout ou partie du membre supérieur ; cette diffusion dépend du segment organisateur du schéma d'action. Si celui-ci est le bras, l'inertie est si forte que l'UG peut être transférée sur la main et les doigts en maintenant son identité (sémantique) *vs* si la main est le centre d'organisation, le transfert du mouvement vers des segments de plus grande inertie (avant-bras, bras) maintiendra plus difficilement la signification de cette UG. On s'attend donc à ce que les réalisations gestuelles proposées aux tests soient identifiées de manière significative par un label et à ce que cette reconnaissance diminue à mesure que les réalisations gestuelles mettent en mouvement des segments éloignés du centre d'organisation de l'UG correspondant.
Quant au sous-projet consacré aux outils d'annotation, le cahier des charges final des développements logiciels prévus pour optimiser l'annotation sous Elan, élaboré en partenariat avec l'institut Max Planck (MPI Nimègue), développeur du logiciel Elan, comprenait 4 tâches : 1/ la francisation de l'interface graphique du logiciel Elan, l'élaboration 2/ de modules logiciels (servlets) pour l'affichage synthétique des principales métadonnées et données de chaque corpus sous une forme graphique, 3/d'un langage de scripts pour l'aide à l'annotation, 4/ d'un moteur de recherche adapté aux données annotées sous Elan, permettant la recherche par l'exemple.

C.4 RESULTATS OBTENUS

Au sein du sous-projet relatif à la LSF enfantine, 82 locuteurs sourds ont été filmés, dont 65 enfants de 3 à 15 ans (population cible) et 17 adultes (population référence) [cf. Annexe F2]. Chaque locuteur est filmé de 20 à 30 minutes en plans large et rapproché, le corpus constitué représentant un total 70 heures de données vidéo. La majorité des enfants filmés vit en région parisienne, pour les raisons humaines évoquées plus haut (congés maladie de certains membres et difficultés du terrain) et pour des raisons sociologiques (nombre d'établissements accueillant des jeunes sourds sont en Ile de France). Numérisation, organisation des métadonnées et conception/expérimentation d'un schéma d'annotation sous Elan (cf. Annexe F5) ont été effectuées en parallèle. Des annotations ont été réalisées ou sont en cours pour les productions de 6 enfants et d'un adulte.
Une première étude de cas (Sallandre et al 2010), menée sur l'expression de la manière et de la trajectoire chez deux enfants sourds de 8 ans et un adulte sourd en comparatif avec des données du français montre une densité sémantique importante dans les énoncés en LSF, marquée par l'utilisation de deux structures de transferts (UT) principales : les transferts personnels expriment plutôt la manière tandis que les transferts situationnels expriment simultanément la trajectoire et la manière, là où en français le verbe exprime le plus souvent

la trajectoire seule. De plus, les sujets sourds enchaînent très souvent deux structures dans le but d'exprimer deux perspectives du locuteur, fait peu fréquent en français. L'étude de Sallandre et Schoder (2011) corrobore ces tendances (chez les mêmes sujets) pour le suivi de la référence aux personnes dans un récit.

Dans le sous-projet dédié à la constitution de corpus de LSF adulte, 58 entretiens ont été réalisés, de 100 mns en moyenne, dans 4 régions de France, auprès d'enquêtés hommes et femmes de 18 à 60 ans (cf. Annexes F3). Chaque entretien étant capté sous trois angles, le corpus représente 290 heures de données vidéo. Le sous-corpus d'unités néologiques compte plusieurs centaines d'UL émergentes (recensement en cours). Outre la constitution et la documentation de ces corpus (numérisation, formalisation des métadonnées, identification des UL émergentes, cf. Annexe F5), le travail mené a été de trois ordres : conception/expérimentation d'un schéma d'annotation sous Elan et annotation effective de 35 mns ; analyse d'un sous-corpus d'unités émergentes ; analyse des procédés de l'autonymie (registre métalinguistique), en transversal de 2 LS, la LSF et la LSFB (collaboration L. Meurant, FUNDP-FNR Namur). Les analyses, qualitatives, menées sur un sous-corpus d'UL récentes corroborent le caractère de forme-sens (morphèmes) des composants mobilisés, attestent de la récurrence de matrices lexicogéniques et du fait que les principes de stabilisation mettent en jeu une logique relevant d'une sémiologie du visuel (cf. Garcia et Perini 2010, Garcia 2010). L'analyse de la mention autonymique d'une UL a permis d'identifier des marqueurs réguliers (direction du regard, mimique faciale neutre, ± procédés multiples de contrôle phatique, comme la répétition multiple, la labialisation, une structuration de type parenthétique), distinguant notamment entre *l'ensemble du signe* en mention (regard sur l'interlocuteur) et le focus sur le *signifiant du signe* en mention (regard sur le signe) (Garcia et Meurant 2010).

Concernant le sous-projet centré sur la gestualité naturelle et pour les raisons indiquées ci-dessus (C.2), le corpus de genre explicatif de dyades d'entendants, de sourds et mixtes, n'est pas achevé : 16 dialogues sont en cours de réalisation, impliquant 3 sourds et 4 entendants (une même personne participant à plusieurs dialogues). Le schéma d'annotation élaboré spécifiquement prend en compte divers aspects (syntaxe, spatialisation, compositionnalités paramétriques, mesures de changements morphologiques et d'interactions). Devant permettre d'affiner le codage et donc l'annotation de ces corpus dyadiques, les tests de reconnaissance de gestes prévus ont été réalisés. 19 des 20 UG sont reconnues de manière significative (moins de 1% que le hasard intervienne). On peut donc dire que les UG ont une identité significativement reconnue basée sur une organisation physiologique. D'autres résultats montrent une structuration des UG par centre d'organisation. Ainsi de « accepter » dont le substrat est manuel et de « s'en ficher » organisé sur le bras (Annexe F4) : leur identification décroît à mesure qu'on s'éloigne de leur centre organisateur respectif. Toutes les UG ne montrent cependant pas de telles caractéristiques, sans doute parce que les différences entre les réalisations sont si ténues qu'il est difficile de les effectuer avec précision. Une perspective consiste dès lors à enregistrer ces réalisations en Mocap : les données physiologiques seraient alors parfaitement contrôlées (Boutet 2008, 2010).

Les tâches prévues au titre du sous-projet consacré à l'élaboration d'outils d'optimisation de l'annotation n'ont pu donner lieu à des développements professionnels mais leur développement partiel a été assuré et des maquettes sont en cours de constitution pour chacun des points figurant dans le cahier des charge final (cf. C2, F6). La francisation du logiciel Elan (1) est en voie de finalisation, pour mise à disposition des développeurs de Elan. Un prototype est en cours d'élaboration (2), exploitant des transformations XSLT des

annotations Elan (XML), associées à des librairies open-source telles que JFreeChart, Google chart API ou autres API équivalentes. Des tests préliminaires (3) ont permis d'identifier le langage de script jython comme ressource exploitable à courte échéance: Jython étant un langage de script compatible avec Java, disposant d'une spécification mature et d'une communauté de développeurs, nous écartons l'idée d'un langage de script *ad hoc*, intégré au logiciel Elan, pour explorer les possibilités de ce langage dans l'optique d'une aide à l'annotation sous Elan par l'automatisation de certaines tâches : pré-annotation de pistes, génération de nouvelles pistes à partir des annotations existantes, vérification et mise en cohérence des données annotées. Dans cette optique, les informations fournies par les utilisateurs avancés de Elan, les retours des annotateurs recrutés au cours du projet, ainsi que les besoins spécifiques au SP2 (actuellement couverts grâce à des macros Excel) permettent d'envisager l'élaboration d'un prototype opérationnel à court terme, qui pourra servir de base à des développements plus poussés. Enfin (4), des tests en cours détermineront l'intérêt des fonctionnalités de clustering (carrot2) intégrées au moteur de recherche open-source Lucene SolR et permettront d'identifier les développements logiciels spécifiques nécessaires pour permettre aux utilisateurs de regrouper automatiquement des productions (ou sous-parties de productions) en fonction d'un exemple-cible. L'exploitation des outils de la plate-forme expérimentale CoPT développée par A. Balvet est tributaire du développement à venir des annotations.

C.5 EXPLOITATION DES RESULTATS

Au-delà des analyses en cours décrites plus haut, le potentiel de recherches ouvert par les corpus constitués est considérable. Ainsi de la description de la grammaire de la LSF enfantine et des stades d'acquisition de la langue, de l'analyse linguistique de l'interaction en LSF adulte (et entre autres des modalités de partage de l'espace), de la conception d'une dictionnairique fondée sur la stratification morphémique propre à la LSF (vs entrées actuelles par des mots français), de l'élaboration d'une notation commune LS/geste, et de toutes les exploitations didactiques afférentes. L'expérience construite par le projet sur la constitution/documentation/édition pérenne de corpus de données visuo-gestuelles doit par ailleurs fortement contribuer à la rédaction d'un "guide des bonnes pratiques" en la matière.

C.6 DISCUSSION

Parmi les nombreuses perspectives ouvertes dans et par *CREAGEST* (qu'il est impossible de lister ici) celles relatives à l'annotation des données visuo-gestuelles sont centrales : dans la communauté nationale et internationale des spécialistes de LS et/ou de gestualité, ceci est un verrou majeur. Outre la question essentielle pour le traitement automatique des critères de segmentation (énoncés, unités des divers niveaux), un enjeu fondamental est d'éviter le recours à un étiquetage (filtrage) par les mots de la langue vocale, palliatif le plus répandu mais d'autant plus problématique que c'est sur ces "gloses" en LV (notamment impuissantes à rendre compte des UT) que portent les requêtes. Des notations fondées sur la physiologie du geste ou sur une modélisation de l'organisation morphémique de la LS, d'une part, des annotations plurilinéaires multiparamétriques combinant plusieurs niveaux/outils de codage, d'autre part, expérimentées dans ou à l'horizon des recherches menées dans *CREAGEST*, ouvrent des alternatives qu'il faudra approfondir.

C.7 CONCLUSIONS

Compte tenu d'une part des lourds facteurs externes ayant entravé le déroulement du projet, d'autre part de son caractère pionnier lié aux spécificités de l'objet (modalité visuo-gestuelle, langue non standardidée et non écrite) et de la population linguistique concernée (communauté sourde), le projet nous semble avoir atteint de manière très satisfaisante l'essentiel des objectifs initiaux. Un corpus de LSF d'une ampleur et d'une représentativité sans précédent a été constitué, déjà considéré comme référence (visibilité internationale et nationale, cf. l'intégration de plusieurs responsables au CS de l'IRCOM) et offrant un riche vivier de ressources pour la recherche linguisique et pour l'exploitation didactique. Outre des avancées descriptives notables (LSF enfantine, modélisation du geste, émergences lexicales), le projet a permis la conception de schémas d'annotation novateurs et débouche sur des propositions relatives à la formalisation des métadonnées.

C.8 REFERENCES

Boutet, D. 2001. *Approche morpho-dynamique du sens dans la gestuelle conversationnelle.* Thèse de Doctorat en Sciences du Langage, Université Paris 8.

Boutet, D. 2008. Une morphologie de la gestualité : une structuration articulaire. in D. Boutet, et C. Cuxac (dir.), *Le signifiant gestuel*, *Les Cahiers de Linguistique Analogique*, 5, Dijon, 81-115.

Boutet, D. 2010. « Structuration physiologique de la gestuelle : modèle et tests » in J-M. Colletta, A.Millet et C. Pellenq (eds.), *Multimodalité de la communication chez l'enfant*, LIDIL, n°42, Grenoble.

Cuxac, C. 2000. *La Langue des Signes Française ; les Voies de l'Iconicité. Faits de Langues,* 15-16, Ophrys, Paris.

Cuxac, C. *et al*, 2002. *Projet LS-Colin, Rapport de fin de recherche*, Programme Cognitique 2000, « Langage et Cognition », IRIT-TCI, LIMSI-CNRS, Université Paris 8, Université Paris Sorbonne.

Garcia, B. 2010. *Sourds, surdité, langue(s) des signes et épistémologie des sciences du langage. Problématiques de la scripturisation et modélisation des bas niveaux en Langue des Signes Française (LSF)*. Thèse d'Habilitation à Diriger les Recherches, Université Paris 8.

Garcia, B. et Meurant, L. 2010. "Signing about signing. Sign metalanguage in LSF and LSFB", Theoretical Issues in Sign Language Research Conference (TISLR 10), *Research Methodologies in Sign Language Linguistics,* Purdue University, Indianapolis, 30 sept-2 oct. 2010.

Garcia, B. et Perini, M. 2010. « Normes en jeu et jeu des normes dans les deux langues en présence chez les sourds locuteurs de la Langue des Signes Française (LSF) ». In B. Garcia et M. Derycke (eds.), *Sourds et langue des signes. Norme et variations*, Langage et Société, n° 131, mars 2010, 75-94.

Garcia, B., Brugeille J.L., Kellerhals, M.P., Braffort, A., Boutet, D., Dalle, P., Mercier, H. 2007. Rapport du Projet RIAM-ANR *LS Script*, 2005-2007, Agence Nationale de la Recherche.

Hickmann, M., Taranne, P., et Bonnet, P. 2009. "Motion in first language acquisition: manner and path in French and in English." *Journal of Child Language, 36*, n°4, 705-741.

Kegl, J., Senghas, A. et Coppola, M., 1999, « Creation through contact : sign language emergence and language change in Nicaragua », *in* M. DeGraff (éd.), *Language Creation and Language Change,* Cambridge, Mass., MIT Press : 179-237.

Kendon, A. 1988. How gestures can become like words. In F. Poyatos (ed.), *Cross-Cultural Perspectives in Nonverbal Communication.* Toronto: Hogrefe, 131-141.
McNeill, D. (1992). *Hand and Mind, What Gestures Reveal about Thought*. Chicago, London: The University of Chicago Press.
Sallandre, M-A, Courtin, C., Fusellier Souza, I., et L'Huillier, M-T. 2010. « L'expression des déplacements chez l'enfant sourd en langue des signes française », LIA1:1 (*Langage Interaction Acquisition*), Amsterdam : John Benjamins, 41-66.
Sallandre, M.-A. et Schoder, C. 2011. « Acquisition des références actantielle et spatiale en langue des signes française et en français ». In P. Trévisiol-Okamura et G. Komur-Thilloy (eds.), Discours, acquisition et didactique des langues. Les termes d'un dialogue. Editions Orizons, Paris. 277-293.
Sandler, W. 2003. Sign Language Phonology. In Frawley W. (ed.), *The Oxford International Encyclopaedia of Linguistics*.
Volterra, V., Erting, C. (Eds) (1994). *From gesture to language in hearing and deaf children*. Washington, D.C.: Gallaudet University Press.

D LISTE DES LIVRABLES

*Quand le projet en comporte, reproduire ici le tableau des livrables **fourni au début** du projet. Mentionner l'ensemble des livrables, y compris les éventuels livrables abandonnés, et ceux non prévus dans la liste initiale.*

Date de livraison	**N°**	**Titre**	**Nature** (rapport, logiciel, prototype, données, …)	**Partenaires** (souligner le responsable)	**Commentaires**
	1				

E IMPACT DU PROJET

E.1 INDICATEURS D'IMPACT

Nombre de publications et de communications

		Publications multipartenaires	Publications monopartenaires
	Ouvrages ou chapitres d'ouvrage	0	4
	Communications (conférence)	5	10
France	**Revues à comité de lecture**	0	3
	Ouvrages ou chapitres d'ouvrage	0	1
	Communications (conférence)	2	17
Actions de diffusion	**Articles vulgarisation**	0	2
	Conférences vulgarisation	2	9

Autres valorisations scientifiques (à détailler en E.3)

	Nombre, années et commentaires (valorisations avérées ou probables)
Brevets internationaux	

obtenus	
Brevet internationaux en cours d'obtention	
Brevets nationaux obtenus	
Brevet nationaux en cours d'obtention	
Licences d'exploitation (obtention / cession)	
Créations d'entreprises ou essaimage	
Nouveaux projets collaboratifs	**Mise en place de collaborations :** - Avec L. Meurant, chercheur au FNRS-Belgique - Série de workshops *Sign Language Corpus Network* organisée par O. Crasborn - Projet MARQSPAT (projet franco-québécois) - Projet COLAJE (ANR)
Colloques scientifiques	**Organisation de journées d'études en lien direct avec le travail mené sur les corpus Creagest :** - Journée d'études, Paris, mars 2010. - Défi DEGELS (DEfi GEstualité et Langues des Signes Montpellier, 1 juillet 2011. - Atelier TALS 2010 Traitement automatique des langues des signes, Montréal, 23 juillet 2010.
Autres travaux	**1/ Recherches doctorales** Thèse de HDR Thèses de doctorat en cours (2) **2/ Expertise scientifique** (IRCOM)

E.2 LISTE DES PUBLICATIONS ET COMMUNICATIONS

International

- **Revues à comité de lecture**

1. Boutet, D. 2010. « Structuration physiologique de la gestuelle : modèle et tests » in J-M. Colletta, A.Millet et C. Pellenq (eds.), *Multimodalité de la communication chez l'enfant*, LIDIL, n°42, Grenoble.

2. Garcia, B. et Perini, M. 2010. « Normes en jeu et jeu des normes dans les deux langues en présence chez les sourds locuteurs de la Langue des Signes Française (LSF) ». In B. Garcia et M. Derycke (eds.), *Sourds et langue des signes. Norme et variations*, Langage et Société, n° 131, mars 2010, 75-94.

3. Cuxac, C. et Antinoro-Pizzuto, E. 2010. « Emergence, norme et variation dans les langues des signes : vers une redéfinition notionnelle ». In B. Garcia et M. Derycke (eds.), *Sourds et langue des signes. Norme et variations*, Langage et Société, n° 131, mars 2010, 37-53 et annexes en ligne.

4. Sallandre, M-A, Courtin, C., Fusellier Souza, I., et L'Huillier, M-T. 2010. « L'expression des déplacements chez l'enfant sourd en langue des signes française », LIA1:1 (*Langage Interaction Acquisition*), Amsterdam : John Benjamins, 41-66.

- **Ouvrages ou chapitres d'ouvrage**

1. Garcia, B. & Sallandre, M-A. (in press). "Transcription systems for sign language: a sketch of the different notation systems for sign language and their characteristics." In C. Müller, A. Cienki, E. Fricke and D. McNeill (eds.), Handbook "Body-Language-Communication". Berlin, New York: Mouton De Gruyter. 12 p.

2. Antinoro Pizzuto, E. et Garcia, B. (en préparation, parution prévue en 2013), « Annotation tools for Sign Languages (SL) : the crucial problem of the graphical representation of forms ». In Terry Janzen & Sherman Wilcox (eds.). Berlin/New York: Mouton De Gruyter, [contact : Pr Sherman Wilcox, Université de New Mexico : wilcox@unm.edu], 31 p.

3. Garcia, B. et Sallandre, M-A. (en préparation, parution prévue en 2012), « Reference and Definiteness in French Sign Language (LSF) ». In Patricia Cabredo Hofherr & Anne Zribi-Hertz (eds.), Syntax and Semantics Series, Emerald Publishing limited, [contact : Patricia Cabredo Hofherr, SFL-CNRS : patricia.cabredo-hofherr@sfl.cnrs.fr], 23 p.

4. Garcia, B., Sallandre, M-A. et Cuxac, C. (en préparation, parution prévue en 2013), «Epistemological issues in the semiological model for the annotation of sign language ». In Laurence Meurant, Aurélie Sinte, Mieke Van Herreweghe & Myriam Vermeerbergen (eds.) Sign Language research, uses and practices, Crossing views on theoretical and applied sign language linguistics » (Sign Language and Deaf Communities), Berlin/New York: Mouton De Gruyter, [Contact : Laurence Meurant, FUNDP-FNR U. Namur : laurence.meurant@fundp.ac.be], 25 p.

- **Communications (conférences – dont invitées)**

1. Boutet, D. 2011, « Gestuality as support and substratum for meaning » Poster, The Third Conference of the Scandinavian Association for Language and Cognition, SALCIII, E. Engberg-Pedersen (org.), Copenhague, 14-16 juine 2011.

2. Sallandre, M-A. 2011. "Iconicity in French Sign Language (LSF) and acquisition of motion events by deaf children." Invited speaker at the *Sign Language Colloquium* (seminar), Radboud University, Nijmegen, Netherlands, 25 Janvier 2011.

3. Balvet, A. 2010. Issues underlying a common Sign Language Corpora annotation scheme, 4th workshop on the representation and processing of Sign Languages: Corpora and Sign Language technologies. The Seventh international conference on Language Resources and Evaluation (LREC 2010), Malte, 17-23 mai 2010.

4. L'Huillier, M.Th., Sallandre, M-A. et Courtin, C. 2010. "Deaf Participation in Sign Language Corpus Projects." Invited speaker to the *Sign Language Corpus Network 4, Workshop.* O. Crasborn (org.). Berlin, 4 décembre 2010. http://www.ru.nl/slcn/workshops/4_exploitation/

5. Sallandre, M-A., L'Huillier, M.Th., Garcia, B. et Courtin, C. 2010. "Links between data collection and deaf education: some perspectives with a corpus in French Sign Language". Poster, *Sign Language Corpus Network 4, Workshop.* O. Crasborn (org.). Berlin, 4 décembre 2010. http://www.ru.nl/slcn/workshops/4_exploitation/

6. Garcia, B. et Meurant, L. 2010. "Signing about signing. Sign metalanguage in LSF and LSFB", Theoretical Issues in Sign Language Research Conference (TISLR 10), *Research Methodologies in Sign Language Linguistics,* Purdue University, Indianapolis, 30 sept-2 oct. 2010.

7. Garcia, B., Cuxac, C., Fusellier, I., Sallandre, M-A., Boutet, D., Courtin, C., L'Huillier, M-T et Balvet, A. 2010. « Sign Language and human gestuality corpora : what is at stake ? The French Creagest Project ». Poster, Theoretical Issues in Sign Language Research Conference (TISLR 10), *Research Methodologies in Sign Language Linguistics*, Purdue University, Indianapolis, 30 sept-2 oct. 2010.

8. Boutet, D. 2010. « Tests d'assignation d'étiquette sur des gestes dits coverbaux ». Journée de conférences scientifiques, UQAM, Montréal, 19 juillet 2010.

9. Sallandre, M-A. 2010. Annotation of Highly Iconic Structures (HIS). Invited speaker in the *Sign Language Corpus Network 3, Workshop,* O. Crasborn (org.), University of Stockholm, 15 juin 2010, http://www.ru.nl/slcn/workshops/3_annotation/

10. Sallandre, M-A. 2010. Annotation of Highly Iconic Structures (HIS). Tutoriel in the *Sign Language Corpus Network 3, Workshop,* O. Crasborn (org.), University of Stockholm, 15 juin 2010, http://www.ru.nl/slcn/workshops/3_annotation/

11. Balvet, A., Garcia, B., Boutet, D., Courtin, C., Cuxac, C., Fusellier-Souza, I., L'Huillier, M-T. et Sallandre, M-A. 2010. « The *CREAGEST* Project: a Digitized and Annotated Corpus for French Sign Language (LSF) and Natural Gestural Languages », The Seventh international conference on Language Resources and Evaluation (LREC 2010), Malte, 17-23 mai 2010.

12. Sallandre, M.-A., Braffort, A. 2009. « LSF resources ». Sign Linguistics Corpora Network 1, Workshop. O. Crasborn (org.), University College London, 26 juillet 2009, http://www.ru.nl/slcn/workshops/1_data_collection/

13. Garcia, B., Fusellier, I., Sallandre, M.-A., Balvet, A. 2009. "The ANR *Creagest* Project: Linguistic and methodological issues involved in creating a corpus of French Sign Language (LSF) and natural gesture". *Sign Language Corpora: Linguistic Issues Workshop*, London, UCL, 24 juillet 2009.

14. Garcia, B., Sallandre, M-A et Fusellier-Souza, I. 2009. « Rôle du pointage dans l'expression de la définitude en langue des signes », Colloque international « Du geste au signe, le pointage dans les langues orales et signées », organisé par l'UMR 8163 "Savoirs,

Textes, Langage" (CNRS et universités Lille 3 & Lille 1), 4 et 5 juin 2009, Université de Lille 3, Villeneuve d'Ascq.

15. Garcia, B. 2008. « Investigations about the internal structure of lexical signs in LSF (French sign language) », Colloque franco-norvégien *Basic and applied issues in sign language research* (Vonen, A. et Sallandre, M-A, org.), Fondation Maison des Sciences de l'Homme, Paris, 11-12 février 2008.

France

- **Revues à comité de lecture**

1. Boutet, D., Blondel, M., Caët, S., Beaupoil, P. et Morgenstern, A. 2011. « Tu pointes ou tu tires?! Annotation sous ELAN des pointages d'un 'entendant vocalo-gestualisant' ». In Boutora, L. et Braffort, A. (eds.), 15–27.

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

3. Boutet, D., et Cuxac, C. 2008. « Le signifiant gestuel : langue des signes et gestualité ; Avant-Propos », *Cahiers de linguistique analogique* n°5, Abell, 2-15.

- **Ouvrages ou chapitres d'ouvrage**

1. Sallandre, M.-A. et Schoder, C. 2011. « Acquisition des références actantielle et spatiale en langue des signes française et en français ». In P. Trévisiol-Okamura et G. Komur-Thilloy (eds.), Discours, acquisition et didactique des langues. Les termes d'un dialogue. Editions Orizons, Paris. 277-293.

- **Communications (conférences– dont invitées)**

2. Garcia, B. et Sallandre, M.-A. 2011. « Modalités de l'instanciation d'une entité référentielle et marquage du défini/indéfini en LSF ». Journée inter-équipe de l'UMR SFL, Paris, 9 mai 2011.

3. Boutet, D. 2011. Organisation de 3 journées de formation sur le logiciel ELAN en lien notamment avec le plan de Formation de l'Unité 7023 (11 janvier et 20 janvier 2011) et avec le LIMSI, sous l'égide de l'ATALA : Organisation et formation en présentiel et en différé par la réalisation de tutoriaux, 21 janvier 2011. http://tals.limsi.fr/tuto/tuto.html

4. Boutet, D. et Braffort, A. 2011. « ELAN et ANVIL, logiciels d'annotation multimodale : principes et différences ». Journée de formation organisée sous l'égide de l'ATALA (Association pour le Traitement Automatique des Langues des Signes), 22 janvier 2011. http://tals.limsi.fr/tuto/1.Annotation_principes.pdf

5. Garcia, B. 2011. « Problématiques du lexique et de l'unité linguistique dans les langues des signes. Facteurs convergents d'opacification et vision alternative », Séminaires des *Cahiers du Français Moderne*, Collège de France, 16 février 2011.

6. Blondel, M., Boutet D. 2010. « Echantillons de recherches et questions liées aux corpus », séminaire de M.-A. Sallandre sur les corpus en LS, Université Parie 8, 3 décembre 2010.

7. Boutet, D., Sallandre, M-A. et I. Fusellier-Souza, I. 2010. « Gestualité humaine et langues des signes : entre continuum et variations ». Journée d'études (org. B. Garcia, avec J. Boutet et M. Derycke) autour du n° 131 de Langage et Société (mars 2010), *Sourds et Langues des Signes, Norme et variations* (Garcia, B. et Derycke, M., eds.), Paris, Maison des Sciences de l'Homme, 17 juin 2010.

8. Cuxac, C. et Antinoro Pizzuto, E. 2010. « Emergence, normes et variation en langues des signes : pour une redéfinition conceptuelle ». Journée d'études (org. B. Garcia, avec J. Boutet et M. Derycke) autour du n° 131 de Langage et Société (mars 2010), *Sourds et Langues des Signes, Norme et variations* (Garcia, B. et Derycke, M., eds.), Paris, Maison des Sciences de l'Homme, 17 juin 2010.

9. Garcia, B. et Perini, M. 2010. « Normes en jeu et jeu des normes dans les deux langues en présence chez les sourds locuteurs de la Langue des Signes Française ». Journée d'études (org. B. Garcia, avec J. Boutet et M. Derycke) autour du n° 131 de Langage et Société (mars 2010), *Sourds et Langues des Signes, Norme et variations* (Garcia, B. et Derycke, M., eds.), Paris, Maison des Sciences de l'Homme, 17 juin 2010.

10. Boutet, D., L'Huillier, M-T, Courtin, C. 2010. « Le projet Creagest : objectifs, enjeux, questions méthodologiques », Séminaire de l'EHESS (A. Benvenuto, A. Karacostas et M. Coutant, org.), *LSF : Questions de terrain*, 7 juin 2010.

11. Garcia, B, Cuxac, C., Courtin, C., Boutet, D. 2010. « Le projet *Creagest* », Workshop *Psychologie et apprentissages*, organisé par l'Agence Nationale de la Recherche, siège de l'ANR, 212, rue de Bercy, 28 mai 2010.

12. Boutet, D. 2010. « Compositionnalité gestuelle : modèle et quelques résultats », *Parisian Workshop on Gesture, Sign and Language Acquisition,* organisé par Aliyah Morgenstern, Institut du Monde Anglophone, 15 et 16 avril 2010.

13. Boutet, D. 2009. « Compositionnalité gestuelle », Workshop Autour de la langue des signes organisé par Agnès Millet et Saskia Mugnier, Université Stendhal Grenoble 3, 26 et 27 novembre 2009.

14. Garcia, B., Sallandre, M-A et Fusellier-Souza, I. 2009. « Rôle du pointage dans l'expression de la définitude en langue des signes », Journée d'études *Langues avec / sans articles, UMR 7023 SFL*, Université Paris 8, 6 mars 2009.

15. Boutet, D. 2008. « Creagest, un projet ANR comprenant la constitution d'un corpus sur la LSF. De la formulation de la demande à la mise en place du projet : outils, méthodes et

pistes de réflexion », Ecole Thématique *Préservation et diffusion numériques des sources de la recherche en sciences humaines et sociales*, Département SHS du CNRS et Adonis (org.), Frejus, 19-24 novembre 2008.

16. Garcia, B. 2008. « Scripturisation des langues des signes et contraintes liées à la modalité », conférence invitée, Journées *Réseau Thématique Européen « Langage et Cognition », 4 ans après ?*, Colette Grinevald et Barbara Köpke, org., Maison de la Chimie, Paris 7e, 1er-2 octobre. 2008.

17. Garcia, B., 2008. « Situation institutionnelle, socio-éducative et sociolinguistique des sourds français et de leur langue, la Langue des Signes Française (LSF) », Colloque *Langues et identités finlandaises : Langues non finno-ougriennes de Finlande*, Institut Finlandais, Paris, 14 et 15 novembre 2008.

18. Boutet, D. et Garcia, B. 2007. « Structure morphophonétique de la LSF, étude à partir d'une base de données relationnelle », *Traitement Automatique des Langues des Signes* TALS 2007, journée d'études dans le cadre du Colloque International TALN'07, ERSS-Université Toulouse Le Mirail / IRIT-Université Paul Sabatier, Toulouse, 5-8 juin 2007.

19. Garcia, B., 2007. « Ecrire les langues des signes, langues visuo-gestuelles : spécificités, conditions et enjeux », Colloque Gramagicom *La langue et ses écritures dans les palliations langagières*, ENST-Bretagne, Technopole de Brest-Iroise, 28 fév-1er mars 2007.

Actions de diffusion

Articles de vulgarisation

1. L'Huillier, M.-T. 2008. « Le projet Creagest : recueil de vidéos en langue des signes. » Compte-rendu de la Journée d'ouverture de l'Année du Handicap à Paris 8, *8ème Sens, le magazine de Paris 8*, n°2, octobre-décembre 2008, page 6.

2. Sallandre, M.-A. 2008. « Mise en accessibilité de l'université aux personnes en situation de handicap », Présentation des projets de recherche en rapport avec le handicap, Journal de l'université Paris 8, page 5.

Conférences de vulgarisation

1. Wauquier, S., Hickmann, M., Soroli, E., Schoder, C., Garcia, B. et Sallandre, M.-A. 2012. « Acquisition des langues : Diversité linguistique et enjeux cognitifs. » Salon de la valorisation de la recherche en SHS, Paris, 1er et 2 octobre 2012.

2. Sallandre, M.-A., L'Huillier, M.-T. et Heouaine, S. 2011. « Acquisition du langage et développement cognitif de l'enfant sourd : Construction d'une méthodologie dans le cadre du projet Creagest ». Journée d'hommage à Cyril Courtin, Cité des Sciences et de l'industrie, Paris, 9 décembre 2011.

3. L'Huillier, M.-T., Sallandre, M.-A. et Heouaine, S. 2011. « Le projet Creagest : Enjeux et perspectives pour l'éducation des jeunes sourds ». Après-midi de sensibilisation auprès

des équipes éducatives en vue du tournage. Ecole maternelle « Les Coquelicots » et école primaire « Roux Tenon », Massy, avril 2011.

4. L'Huillier, M.-T., Sallandre, M.-A. et Heouaine, S. 2011. « Le projet Creagest : Enjeux et perspectives pour l'éducation des jeunes sourds ». Après-midi de sensibilisation auprès des équipes éducatives en vue du tournage. Etablissement CELEM, Paris, mars 2011.

5. L'Huillier, M.-T., Sallandre, M.-A. et Heouaine, S. 2009 et 2011. « Le projet Creagest : Enjeux et perspectives pour l'éducation des jeunes sourds ». Après-midi de sensibilisation auprès des équipes éducatives en vue du tournage. Ecole maternelle et primaire « Les Deux Parcs », association Laurent Clerc, Champs sur Marne, septembre 2009 et février 2011.

6. Garcia, B. 2010. « Le projet Creagest : enjeux linguistiques et sociaux pour les sourds et la LSF», colloque *Les 30 ans de l'Académie de la Langue des Signes Française (ALSF)*, 6 février 2010, Paris.

7. L'Huillier, M.-T., Sallandre, M.-A., Courtin, C. et Heouaine, S. 2009. « Le projet Creagest : Enjeux et perspectives pour l'éducation des jeunes sourds ». Après-midi de sensibilisation auprès des équipes éducatives en vue du tournage. Classe bilingue de l'école primaire du 3ème arrondissement de Paris, septembre 2009.

8. Fusellier, I., L'Huillier, M.-T., Sallandre, M.-A., Courtin, C. et Heouaine, S. 2008. « Le projet Creagest : Enjeux et perspectives pour l'éducation des jeunes sourds ». Après-midi de sensibilisation auprès des équipes éducatives en vue du tournage. Institut Département des Jeunes Sourds Gustave Baguer, Asnières, mars et septembre 2008.

9. Garcia, B., et L'Huillier, M-T. 2008., « Projet *Creagest* : importance des corpus de LSF et construction de la place des sourds dans la recherche linguistique », colloque *Les 10 ans de l'association Visuel-LSF*, Paris, 21 novembre 2008.

10. Sallandre, M.-A. 2008. « Evolution des recherches sur l'iconicité des LS colloque *Les 10 ans de l'association Visuel-LSF*, Paris, 22 novembre 2008.

11. L'Huillier, M.-T. 2008. « Le projet Creagest : recueil de vidéos en langue des signes ». Communication pour la Journée d'ouverture de *l'Année du Handicap à Paris 8*, Université Paris 8, Saint-Denis, 15 avril 2008.

E.3 LISTE DES ELEMENTS DE VALORISATION

1/ Mise en place de collaborations :

- Garcia, B. Collaboration avec L. Meurant, chercheur au FNRS-Belgique : étude comparative des marqueurs de l'autonymie en LSF et LS francophone de Belgique (LSFB) ; reprise pour la LSFB du protocole établi pour les corpus du SP3.
- Sallandre, MA. Collaboration dans le cadre de la série de workshops *Sign Language Corpus Network* organisée par O. Crasborn pour la mise en place de bonnes pratiques au niveau européen pour l'archivage, l'annotation et l'exploitation des corpus de LS.
- Boutet, D. Collaboration dans le cadre du projet MARQSPAT (projet franco-québécois) notamment pour la mise en place d'un schéma d'annotations pour le gestualité.
- Boutet, D. Collaboration dans le cadre du projet COLAJE (ANR) notamment pour la mise en place d'un schéma d'annotations de la gestualité pour l'étude de la négation.

2/ Organisation de journées d'études en lien direct avec le travail mené sur les corpus Creagest :

- Journée d'études (org. B. Garcia, avec J. Boutet et M. Derycke) autour du numéro spécial de Langage et Société (n° 131, mars 2010), Sourds et Langues des Signes, Norme et variations (coord. Garcia, B. et Derycke, M.), Maison des Sciences de l'Homme, mars 2010.
- Organisation du Défi DEGELS (DEfi GEstualité et Langues des Signes) (org. L. Boutora, A. Braffort, D. Boutet, P. Dalle, M. Blondel) dans le cadre de la conférence TALN 2011 Montpellier, Montpellier, 1 juillet 2011.
- Organisation de l'Atelier TALS 2010 Traitement automatique des langues des signes, (org. A.-M. Parisot, J. Rinfret, A. Voghel, D. Machabée, D. Boutet, M. Blondel, L. Boutora, J. Dalle et P. Dalle) dans le cadre de la conférence internationale TALN, Montréal, 23 juillet 2010.

3/ Expertise scientifique

Intégration en 2011 de B. Garcia et D.Boutet au titre de :

- membres du conseil scientifique du Consortium national *Infrastructure de Recherche Corpus Oraux et Multimodaux* (IRCOM, Fédération *Typologie et Universaux Linguistiques*, fédération de recherche FR 2559) dans le cadre du *TGIR-Corpus*, 2011-2014
- et co-pilotes (avec M. Hickmann et H. Jisa) du Groupe de Travail *Multimodalité et modalité visuo-gestuelle*, constitué par le CP de l'IRCOM.

4/ Autres travaux :

—Thèse de HDR

Garcia, B. 2010. *Sourds, surdité, langue(s) des signes et épistémologie des sciences du langage. Problématiques de la scripturisation et modélisation des bas niveaux en Langue des Signes Française (LSF)*. Thèse d'Habilitation à Diriger les Recherches, Université Paris 8.

—Thèses de doctorat suscitées dans le cadre du projet *Creagest*:

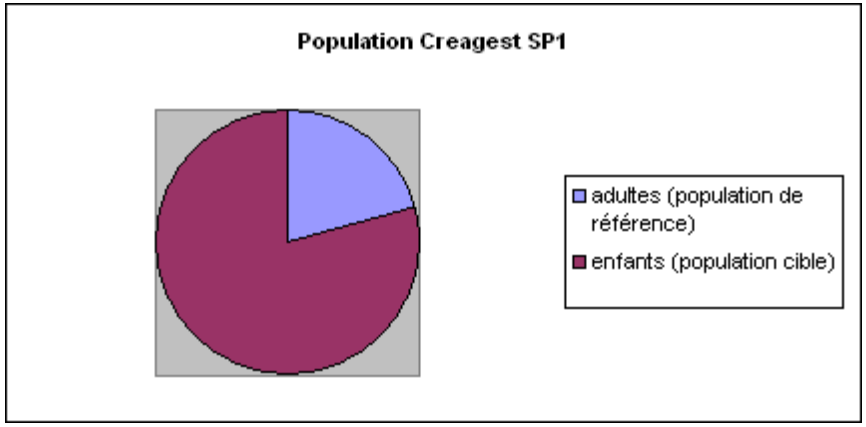
Makouke, D. : thèse commencée en 2010 portant sur la problématique d'une morphologie de la LSF/des LS.
Schoder, C. : thèse commencée en 2011 portant sur l'acquisition de l'espace en LSF par des enfants sourds. Cette thèse bénéficie d'un contrat doctoral de l'université Paris 8 depuis avril 2011.

F **ANNEXES** (PARTIES NON CONFIDENTIELLES)

F.1 COMPLEMENTS RELATIFS AU SOUS-PROJET DE CONSTITUTION DE CORPUS DISCURSIFS DE LSF ENFANTINE : DETAIL DES METADONNEES

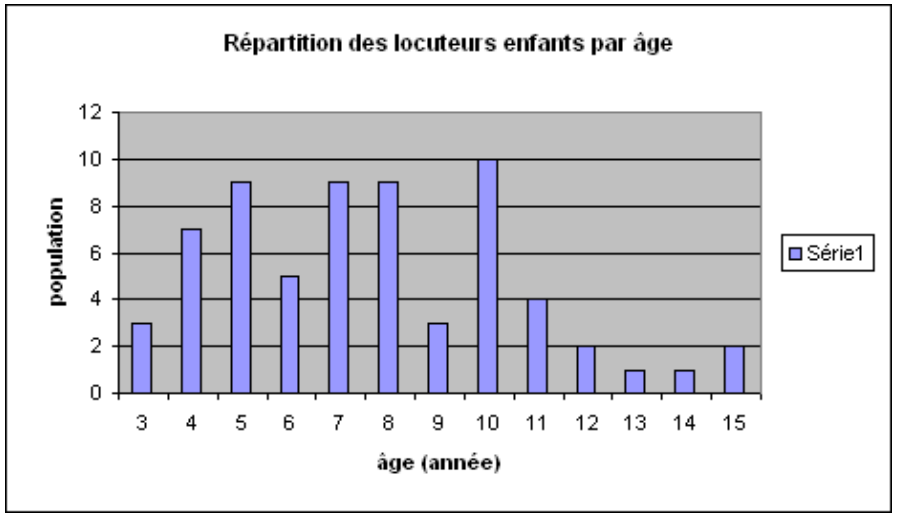
Dans ce sous-projet, des enquêteurs sourds de régions différentes ont été recrutés et formés durant plusieurs sessions aux techniques de passation des stimuli et d'enregistrement vidéo. Des corpus « pilotes » ont été réalisés afin de tester la faisabilité de notre protocole et d'affiner la méthodologie, inédite auprès de cette population. Certains corpus pilotes particulièrement réussis ont ensuite été intégrés aux corpus définitifs et ont été annotés.

Population SP1	
adultes (population de référence)	17
enfants (population cible)	65
Total	**82**



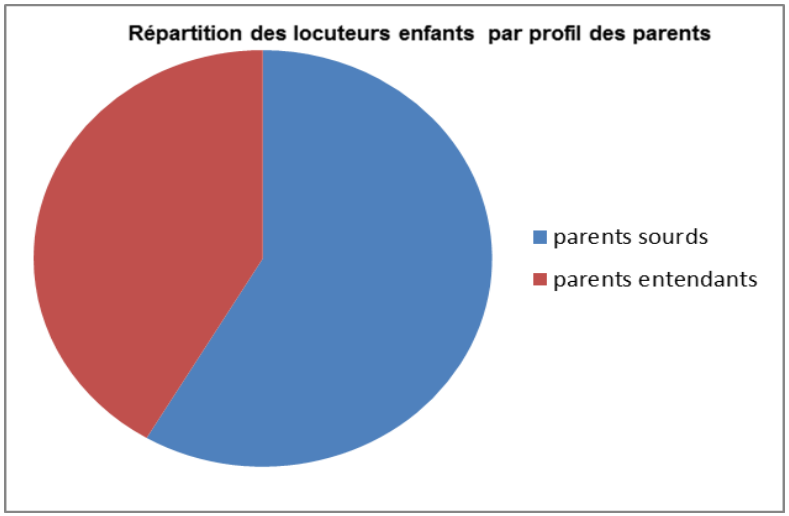
Répartition des locuteurs enfants par âge

âge (en année)	**population**
3	3
4	7
5	9
6	5
7	9
8	9
9	3
10	10
11	4
12	2

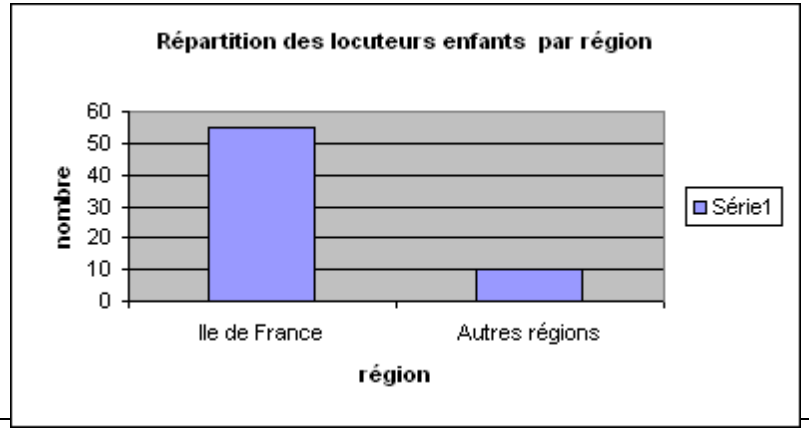


13	1
14	1
15	2
Total	**65**

Répartition des locuteurs enfants par profil des parents	
parents sourds	38
parents entendants	27
Total	**65**



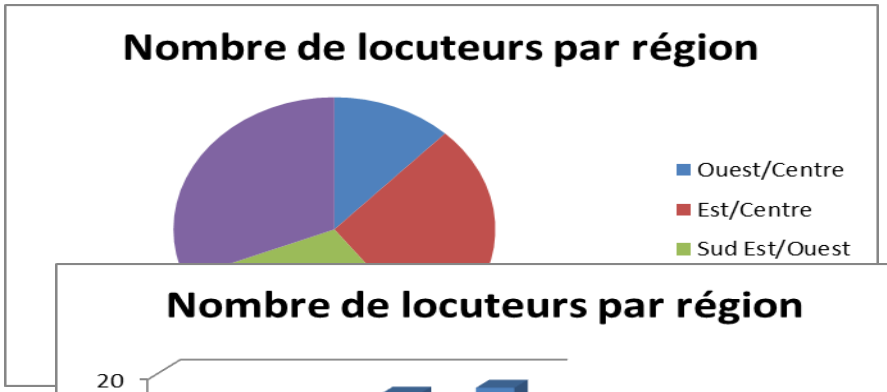
Répartition des locuteurs enfants par région	
Ile de France	55
Autres régions	10
Total	**65**

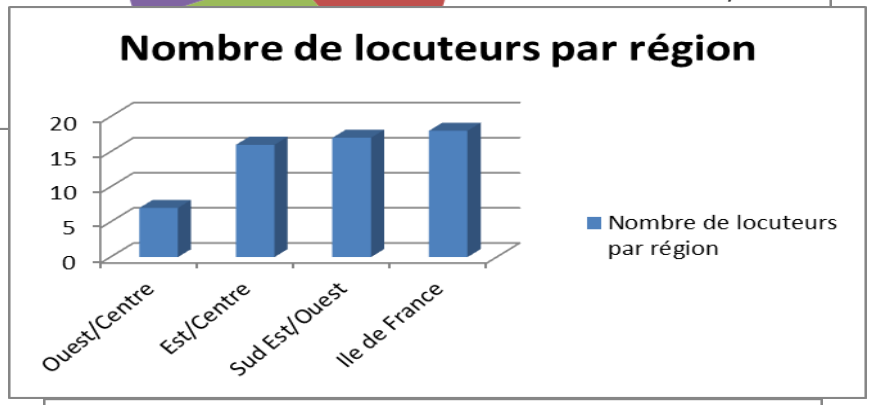


F.2 COMPLEMENTS RELATIFS AU SOUS-PROJET DE CONSTITUTION DE DIALOGUES ENTRE ADULTES SOURDS : DETAIL DES METADONNEES

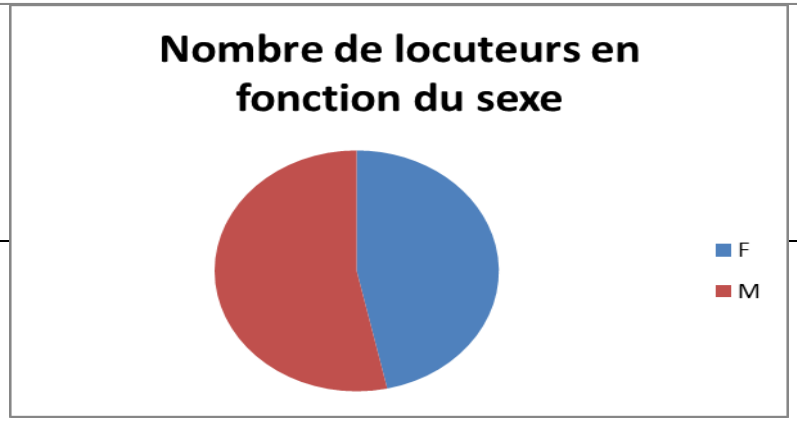
Dans ce sous-projet, des enquêteurs sourds de régions différentes ont été recrutés et formés durant plusieurs sessions aux techniques de l'entretien semi-directif et d'enregistrement vidéo. Des corpus « pilotes » ont été réalisés afin de tester la faisabilité de notre protocole et d'affiner la méthodologie, inédite auprès de cette population. Certains corpus pilotes particulièrement réussis ont ensuite été intégrés aux corpus définitifs et ont été annotés.

régions	Nombre de locuteurs par région
Ouest/Centre	7
Est/Centre	16
Sud Est/Ouest	17
Ile de France	18
total:	**58**

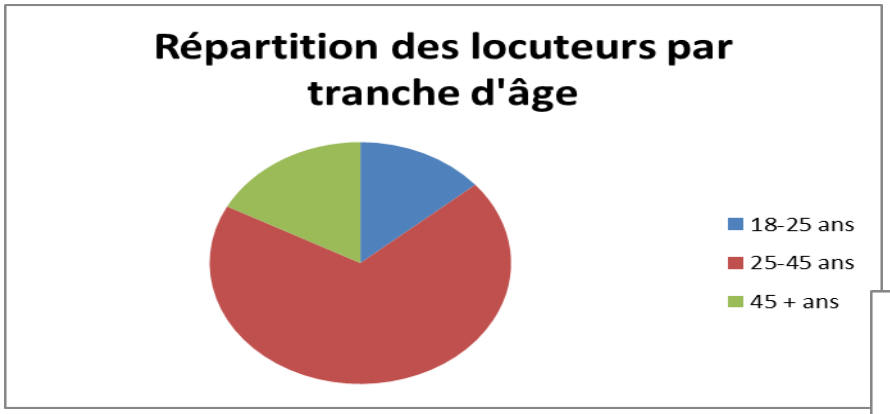




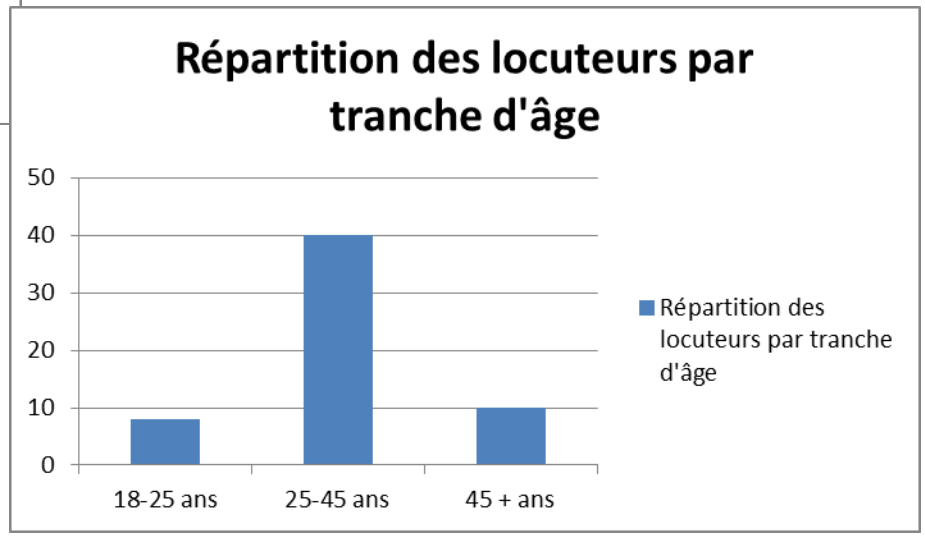
sexe	Nombre de locuteurs en fonction du sexe
F	27
M	31
total:	**58**



...



	Répartition des locuteurs par tranche d'âge
18-25 ans	8
25-45 ans	40
45 + ans	10
total:	**58**



F.3 COMPLEMENTS RELATIFS AU SOUS-PROJET CENTRE SUR L'ETUDE DE LA GESTUALITE COVERBALE

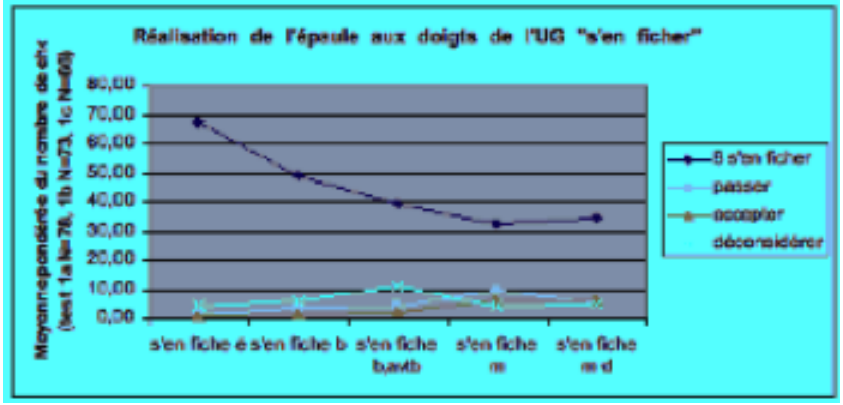
Les 20 UG ont été labellisées et reconnues comme le montre le tableau ci-dessous (P>0,05). Si on abaisse le seuil de significativité à 1%, alors 90% des UG sont reconnues. Une seule réalisation gestuelle par UG a été présentée dans ce test, celle mettant en mouvement la main seulement.

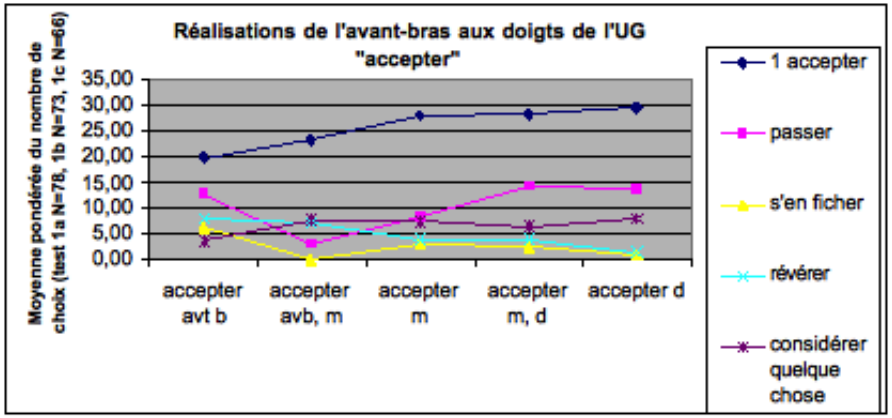
test 1	Prob. Hasard	test2	Prob. Hasard
apparaître	$6{,}997*10^{-6}$	impuissance	$1{,}116*10^{-9}$
disparaître	0,02	mettre de côté	$3{,}31*10^{-26}$
s'en ficher	$3{,}885*10^{-31}$	rejeter	$2{,}955*10^{-23}$
constater	$5{,}761*10^{-3}$	passer	$1{,}773*10^{-12}$
refuser	$1{,}499*10^{-6}$	se préserver	$7{,}202*10^{-35}$
offrir	$1{,}577*10^{-20}$	omettre	$5{,}761*10^{-3}$
arrêter	0,02	commencer	$1{,}043*10^{-3}$
considérer qu.chose	$1{,}088*10^{-4}$	mépriser	$3{,}593*10^{-4}$
déconsidérer	$1{,}701*10^{-13}$	révérer	$1{,}436*10^{-10}$
accepter	$1{,}116*10^{-9}$	considérer quelqu'un	$1{,}116*10^{-9}$

Tableau 1 : Probabilités d'intervention du hasard pour les tests 1 et 2 (P<0,05)

Six autres séries de tests (voir graphiques ci-dessus) ont permis de mesurer la validité du lieu d'émergence des UG et leur diffusion sur le membre supérieur. L'étiquetage est significatif (P<1%) pour la totalité des réalisations gestuelles présentées alors même que dans ces tests les UG ont été regroupées par proximité formelle *i-e* avec des labels très proches (p.e. « rejeter », « refuser », « mépriser », « déconsidérer »). Chaque UG a été déclinée gestuellement en 4 ou 5 réalisations mettant en mouvement un ou plusieurs segments du membre supérieur. Selon le modèle (Boutet 2001), trois segments constituent ici des centres d'organisation des UG : le bras, l'avant-bras et la main. Les réalisations gestuelles des UG organisées par le bras sont toutes bien reconnues (4 sur 4, une seule étiquette est choisie de manière significative pour les réalisations). 10 des 12 UG organisées sur la main ont également été bien reconnues. Pour les 2 UG manuelles non reconnues, une étiquette très proche sémantiquement a été préférée par les sujets (« apparaître » plutôt que « commencer » qui était attendue) dans un cas, et l'étiquette attendue « omettre » a été choisie pour une des 4 réalisations gestuelles, tandis que l'étiquette « arrêter » a été retenue de manière significative pour une des 4 réalisations gestuelles de l'UG « omettre » et l'étiquette « disparaître » une fois. Enfin, une seule UG organisée sur l'avant-bras sur les 4 a été reconnues et bien étiquetée.

Une autre hypothèse n'a été que partiellement validée dans cette série de 6 tests : la reconnaissance de l'UG diminue à mesure qu'on s'éloigne du centre d'organisation. Si ce dernier est au niveau de la main, alors une réalisation gestuelle ne mettant en mouvement que l'avant-bras devrait être moins bien identifiée. Dans le même ordre d'idée, une UG dont le centre d'organisation se situe sur le bras devrait voir diminuer la reconnaissance d'une réalisation gestuelle ne mettant en mouvement que la main. Deux graphiques concernant chacun une Unité Gestuelle montre ce qui se passe pour « accepter » dont le centre d'organisation est la main et pour « s'en ficher » organisée sur le bras.





F.4 CHOIX TECHNOLOGIQUES OPERES POUR LA CAPTATION, LA NUMERISATION ET LE STOCKAGE DES DONNEES ET POUR LA FORMALISATION DES METADONNEES (ELEMENTS COMPLEMENTAIRES)

Pour des raisons pratiques, les corpus sont captés grâce à des caméras « grand public » sur support mini-DV, dans des conditions (éclairage naturel, pas de maquillage etc.) aussi proches que possible d'une interaction « spontanée ». Les corpus collectés nécessitent donc une phase de numérisation avant exploitation. Ces corpus ne présentent pas, pour les raisons indiquées, une qualité HD.

De nombreuses séances de réflexion collective ont permis d'identifier 1) le nombre, le type et la structure des éléments de métadonnées pertinents pour chaque sous-corpus, 2) le format de structuration de métadonnées le plus adéquat. Les métadonnées et les annotations associées à chaque production sont pensées pour être explorables et moissonnables par des dispositifs tels que OAI-ster. De ce fait, l'ensemble de ces informations est saisi sous la forme de fichiers XML : fichiers de transcription EAF sous Elan et fichiers XML *ad hoc* pour les métadonnées « pures ». Au sens strict, les métadonnées renseignent les différentes variables sociolinguistiques (âge, sexe, région d'origine, latéralisation, type de sous-corpus, type de stimulus etc.) nécessaires à l'identification de chaque production, ainsi qu'à une exploration détaillée des corpus constitués. Le but est de pouvoir constituer des sous-corpus en exploitant les différentes métadonnées ainsi renseignées, par ex., ou de permettre de croiser paramètres linguistiques observés (structures annotées) et variables sociolinguistiques. Le format de métadonnées retenu est OLAC, augmenté des informations manquantes (ex. : âge des locuteurs).

La pérennisation est assurée de façon générale par l'adoption d'outils et de formats de données ouverts et de standards internationaux. Ainsi, les corpus sont numérisés au format vidéo MPEG4, les annotations et métadonnées sont saisies sous la forme de fichiers XML, en utilisant des outils utilisés par une communauté d'utilisateurs étendue : ELAN + LAT du MPI. Enfin, les corpus, annotations et métadonnées sont stockés en interne sur différents supports numériques (cassettes mini-DV, serveur + NAS, disques durs autonomes). Toutefois, une chaîne éditoriale a été élaborée en partenariat avec le CRDO et l'IN2P3 dans le but d'assurer une visibilité et un accès le plus large possible aux données collectés dans le cadre du projet, ainsi qu'à un archivage le plus pérenne possible.

L'édition des fichiers vidéo est assurée par des logiciels propriétaires, le format de sortie étant toutefois ouvert et standard : MPEG4. Les enregistrements donnent lieu à une édition minimale : insertion d'un titre et du logo Creagest, calage sur clap de début, recalage des bandes enregistrées par différentes

caméras, reformatage si nécessaire (4/3 ou 16/9). Une fois les fichiers numérisés, l'essentiel du processus d'édition concerne les annotations associées à chaque enregistrement. Ces annotations sont réalisées sous ELAN, afin de garantir une diffusion maximale de ces données auprès de la communauté des linguistes des LS structurée par l'action SLCN (Univ. Radboud, Pays-Bas). Les fichiers EAF produits par ELAN sont des fichiers XML, qu'il est possible de transformer afin d'exporter les données dans des formats propres à d'autres outils : Anvil, Shoebox, Childes, html, texte tabulé etc.

F.5 COMPLEMENTS RELATIFS AU SOUS-PROJET D'OPTIMISATION ET DE TESTS DES OUTILS D'ANNOTATION

A — CAHIER DES CHARGES SOUMIS POUR PRESTATION DE SERVICE

Définition des tâches de développement

Le MPI développe une plate-forme web destinée à faciliter l'upload, l'archivage, la structuration des archives et la diffusion des données annotées grâce à ELAN, via un serveur web. Cette plate-forme, LAT (Language Archiving Tool) sera installée sur le serveur Creagest dédié au projet, et administré par les ingénieurs-systèmes de l'UPS Pouchet : Evariste Ciret en particulier (evariste.ciret@pouchet.cnrs.fr).
Dans les développements envisagés, nous distinguerons les développements destinés à une exploitation en local (poste personnel d'un utilisateur) de ceux destinés à une exploitation sur le serveur Creagest. Pour chaque tâche, une estimation du temps de développement nécessaire est fournie (nb. de jours), à titre purement indicatif.

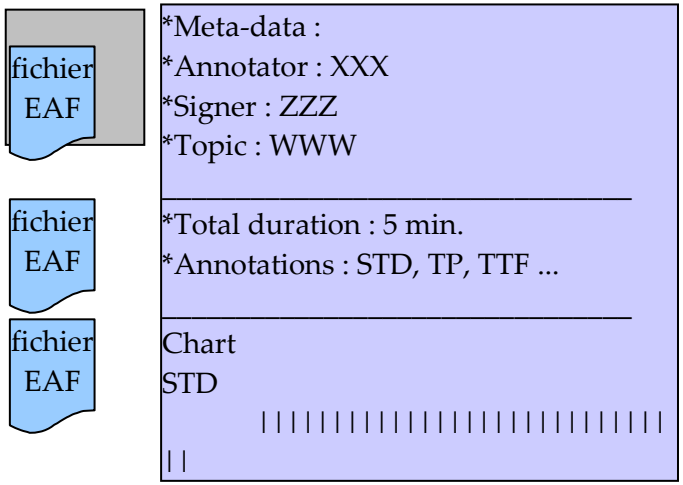
Fonctionnalités utilisables en local

- **Localisation du logiciel ELAN** : dans sa version actuelle, le logiciel ELAN dispose de fichiers de localisation en français, toutefois cette localisation est incomplète et le choix des termes n'est pas toujours satisfaisant pour les utilisateurs linguistes. Il conviendrait donc, en partenariat au minimum avec le groupe de recherche Creagest, de développer une version française exploitable du logiciel ELAN. Le logiciel étant open-source, cette localisation est envisageable, moyennant la fourniture par l'UMR STL des fichiers concernés. **Durée estimée du développement : 1 jour.**
- **Développement d'un langage de script compatible avec ELAN** : afin d'optimiser le temps de travail manuel pour chaque transcription, il est nécessaire de développer un langage de script compatible avec la structure des annotations saisies. Ce langage doit permettre à des **utilisateurs non programmeurs** de développer simplement des "macros", leur permettant par exemple de vérifier la cohérence des annotations manuelles, ou de générer de nouvelles annotations à partir d'annotations existantes. De fait, ce langage de script doit également être capable de s'interfacer avec l'API ELAN, afin de bénéficier des fonctionnalités déjà développées, ex. : construction d'un objet "TRANSCRIPTION", disposant de toutes les informations structurées nécessaires. À défaut, le langage de script devra permettre de parser les fichiers XML afin d'en extraire les informations

nécessaires (i.e. réimplémenter des fonctionnalités déjà disponibles dans l'API ELAN). Ce langage peut être envisagé comme une "sur-couche" au minimum à l'API ELAN, voire à XPath : l'utilisateur doit pouvoir spécifier de façon simple des règles procédurales de type `SI Condition(s) ... ALORS Action`. Deux options sont envisageables pour cette tâche, qui a une vocation essentiellement **exploratoire** : 1) développement de zéro d'un langage de script ou 2) adaptation d'un langage de script tel que Groovy. Des exemples de scripts réalisés en perl par l'université Radboud peuvent être fournis. **Durée estimée des développements : 10 jours.**

Fonctionnalités utilisables sur le serveur Creagest (ou en local)

1/ Développement/adaptation de fonctionnalités de visualisation synthétique des données annotées pour chaque fichier eaf stocké sur un répertoire. Cette fonctionnalité vise à représenter de façon synthétique et aussi graphique que possible les informations et annotations présentes dans chaque fichier d'annotation stocké sur un répertoire du serveur (ou d'un pc utilisateur). À partir d'une transformation XSLT, il est envisageable de récupérer un tableau structuré des principales méta-données (ex. : identifiant de l'annotateur, du locuteur, du thème de l'entretien), ainsi qu'une vue globale des annotations (ex. : types d'annotations, volume d'annotations par type, durée totale, 10 annotations les plus fréquentes). Les API candidates pour cette tâche sont : Google chart API (http://code.google.com/intl/fr/apis/chart/), JFreeChart (http://www.jfree.org/jfreechart/ samples.html) ou API équivalente (java, open-source). L'illustration ci-dessous donne une vision de la fonctionnalité recherchée.



Ici, pour chaque fichier du répertoire courant, une vue synthétique est proposée (calculée à la volée ou stockée dans un fichier de log) sous la forme d'un objet mêlant texte et éléments graphiques. Le but est de fournir pour chaque fichier eaf une fiche récapitulative. Ici, les différents types d'annotations sont présentés, accompagnés d'un graphique en bâtons (bar chart). **Durée estimée du développement : 3 jours.**
2/ Développement/adaptation de fonctionnalités de recherche par l'exemple. La plate-forme LAT du MPI propose des fonctionnalités de recherche permettant d'extraire tous les passages jugés pertinents par un linguiste, dans un répertoire contenant des fichiers eaf. Ces fonctionnalités permettent, par exemple, de décrire de façon systématique le comportement linguistique d'un ensemble d'unités linguistiques étudiées (ex. : la construction des verbes). Toutefois, il manque des fonctionnalités de clustering, d'une part, et de recherche par l'exemple d'autre part. Les fonctionnalités de clustering permettraient de fournir une vue globale des données collectées, en regroupant les fichiers eaf les plus proches (i.e. les annotations les plus similaires) autour d'un noyau central. La recherche par l'exemple permettrait à un utilisateur de sélectionner un sous-ensemble de fichiers eaf, les plus proches d'un fichier sélectionné. Ceci permettrait d'autres modes d'exploration et d'exploitation des données que ce que les fonctionnalités actuelles de recherche proposent. Le moteur de recherche Apache SolR dispose de fonctionnalités de clustering et de recherche par l'exemple, toutefois il reste à évaluer l'intérêt de ces fonctionnalités, pensées pour des documents textuels "classiques", confrontées à des fichiers d'annotation, comportant aussi bien des phrases utilisant un vocabulaire ouvert (ex. : traductions en français ou en anglais) que des annotations basées sur un vocabulaire fermé (redondance élevée). Cette tâche a une dimension essentiellement exploratoire. **Durée estimée du développement : 6 jours**.
Les différentes tâches détaillées ci-dessus peuvent être classées, de la plus critique à la plus facile : 1) développement d'un langage de script, 2) développement de fonctionnalités de clustering + recherche par l'exemple, 3) développement de fonctionnalités de "fiche récapitulative" pour chaque fichier d'annotations, 4) localisation en français.

B—PROSPECTIVE

Comme indiqué en C4, les objectifs de ce sous-projet ont dû être amendés en fonction de l'évolution de l'offre logicielle proposée par le Max Planck Institute (cf. *infra*). Restent les points « durs » du programme prévu : l'application d'algorithmes de découverte de structures récurrentes au sein de corpus annotés, la comparaison entre productions ou passages annotés et leur ordonnancement le long d'un continuum de similarité. Ces deux points sont tributaires d'une masse critique d'annotations, en cours de constitution. Un autre domaine restant à aborder de façon approfondie est celui de la méta-annotation, c'est à dire la génération de nouvelles annotations à partir des annotations existantes. Pour ce faire, nous testons actuellement le langage de script jython, à partir d'annotations produites tant dans le cadre de *CREAGEST* que dans celui d'autres projets (travaux de l'université Radboud notamment). Sur ce point précis, nous pouvons affirmer que l'approche reposant sur un langage de script, impliquant une démarche procédurale, restera limitée aux cas facilement traitables par l'application de règles d'annotation ordonnées, ex : SI annotation1 ET SI annotation2 ALORS annotation3. Toutefois, il apparaît de nos discussions sur ce point, tant avec les membres du présent projet que les collègues étrangers, qu'au-delà de cette approche procédurale une approche plus dynamique sera nécessaire. Le problème est le suivant : si on considère une transcription comme une structure de données comportant un ensemble de « pistes » dédiées à un type d'unités, constituant elles-mêmes des ensembles de couples attribut-valeur, le problème de la méta-annotation peut se

reformuler dans celui de la production de structures de symboles, à partir d'un ensemble de symboles lui-même structuré (au minimum par la succession temporelle). De ce fait, la génération de nouvelles annotations (symboles structurés) à partir d'annotations existantes semble s'apparenter au moins en partie au problème plus général de l'analyse syntaxique, autrement dit la dérivation d'une structure de données typée à partir de séquences de symboles eux-mêmes typés (ex : compilateurs des langages de programmation, analyse syntaxique des langues naturelles). Ce problème demandera donc de déployer des approches s'inspirant de la théorie des compilateurs, *i-e* le développement d'un parser spécifique aux structures de données contenues dans les transcriptions Elan. Sur ce point, des tests sont en cours afin de prouver la faisabilité et l'intérêt de l'approche dynamique pour la méta-annotation, en complément de l'approche procédurale.