The CREAGEST Project (ANR) :

Linguistic and methodological issues involved in creating a corpus of French Sign Language (LSF) and natural gesture

Brigitte GARCIA, Ivani FUSELLIER-SOUZA, Marie-Anne SALLANDRE
Dominique BOUTET, Marie-Thérèse L’HUILLIER, Christian CUXAC
(Paris 8 University/UMR 7023)
Cyril COURTIN (Caen and Paris 5 Universities/UMR 6194)
Antonio BALVET (Lille 3 University/ UMR 8163)

Sign Language Corpora : Linguistic Issues Workshop
London, UCL, July, 24-25th 2009
CREAGEST (2007-2012)

- **ANR** (National Research Agency) program

**OBJECTIVES**

- **Corpus building**
  - Large scale: 250 hours/over 140 signers
  - Sociolinguistic and discourse variations

- **Linguistic modeling**
  - Refine the hypothesis of a semiogenesis of SL (Cuxac, 2000)
  - Better understand the processes at work in the emergence, and semiotization of human gestuality
CREAGEST (2007-2012)

Partners/Members

Paris 8 University
Linguistics

C. Cuxac (H)
B. Garcia (H)
MA Sallandre (H)
I. Fusellier (H)

D. Boutet (H)
MT L’Huillier (D)
S. Fiore (CODA)
P. Guitteny (H)
E. Ciret (H)

Paris 5 and Caen Universities
Cognitive Psychology

Lille 3 University
Linguistics and NLP

C. Courtin (D)

A. Balvet (H)
CREAGEST (2007-2012)

**SUBPROJECTS (SP)**

- **SP1**: LSF acquisition: Deaf children
- **SP2**: Interrelations between human gesture system and SL: hearing and Deaf adults
- **SP3**: Development and understanding of processes at work in LSF Lexicogenesis: Deaf adults
- **SP4**: Administering a collaborative database
- **SP5**: Data processing tools: transcription and mining
SP1 : LSF Acquisition
**SP1 : LSF Acquisition**

- **Goals** => Data collection for:
  - academic research: understanding the **cognitive and linguistic development** of Deaf children
  - research in pedagogy: laying the grounds of a **child-directed grammar** for LSF

- **Linguistic issues**
  - What are the **common roots** between **natural gestures** and Deaf **child LSF**?
  - What **structures** are **available** throughout **Deaf child developmental stages**?
  - How does child-LSF express **verbs of movement**? By manner or path?
SP1: LSF Acquisition

Researchers and investigators from Paris

I. Fusellier (H)
M.T. L’Huillier (D)
(M. A. Sallandre (H)
(Nat. coordinator)
P. Palacios (D)
C. Courtin (D)

Deaf investigators from 4 different regions

P. Palacios (D)
South-west
S. Heouaine (D)
Centre
N. Boursin (D)
West
C. Fiegenwald (D)
East
• **Methodology**:

- **POPULATION**: 56 Deaf children (3 to 11) from different regions with SL as main communication

- **PROTOCOL**
  - 2 cameras (dialogue between child and interviewer or child alone)
  - 30 minutes of interview => free conversation, video-stimuli, image stimuli
  - Outside actor (a **doll** for young children or a **person** for others)
  - Deaf investigators (mostly teachers) from the same geographical area as children
### Metadata and variables

<table>
<thead>
<tr>
<th>CONSTANTs</th>
<th>Transversal data</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 children by age bracket</td>
<td></td>
</tr>
<tr>
<td>Deep deafness</td>
<td></td>
</tr>
<tr>
<td>SL main communication and taught language</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>AGE BRACKETS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3 - 4</td>
</tr>
<tr>
<td>Parents</td>
<td></td>
</tr>
<tr>
<td>Cochlear Implant</td>
<td></td>
</tr>
<tr>
<td>Monomodal bilingualism</td>
<td></td>
</tr>
<tr>
<td>Deaf</td>
<td></td>
</tr>
<tr>
<td>Bimodal bilingualism</td>
<td></td>
</tr>
<tr>
<td>Hearing</td>
<td></td>
</tr>
<tr>
<td>Bimodal bilingualism</td>
<td></td>
</tr>
<tr>
<td>Hearing</td>
<td></td>
</tr>
<tr>
<td>Bimodal bilingualism</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TOTAL</th>
<th>56</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>56</td>
</tr>
</tbody>
</table>
SP1: LSF Acquisition

→ 4 stimuli  Aim: narratives and spontaneous discourse

a) Free conversation

b) “Voluntary movement” stimuli (Hickmann protocol, manner and path verbs, cf. Slobin)

c) Narrative with a cartoon (1 minute)

d) Picture-based narrative (The Horse Story, Hickmann)
SP1: LSF Acquisition

**Preparatory phase (May 08 - Jun 09)**

- 4 collective methodological training sessions with investigators
- 17 pilot corpora conducted by lead team and investigators
  - 8 children, 4 teenagers, 5 adults control group

**Objectives**

- Validation of methodological protocol
- Investigator training (task steps and technical material)
- Preserving children’s natural behaviour in experimental tasks
SP1 : LSF Acquisition

- Video examples
Methodological discussions

- Have a 3rd actor in the film: person or doll?
- Minimize methodological bias (Ex. “how”/”comment” from protocol proposed by M. Hickmann for acquisition of vocal languages)
- Prepare the ground and mobilize parents, professionals, teachers
  - Video clip in LSF/French subtitles
Bienvenus sur CREAGEST- SP1
Sous-projet 1 - Acquisition de la Langue des Signes Française par des enfants

Voici un clip vidéo descriptif du projet en LSF sous-litté en français.
Bon visionnage et n'oubliez pas de nous contacter.

Objectifs de CREAGEST/SP1 - Acquisition de la langue des signes par des enfants.
SP2 : Gesture and SL
SP2 : gesture and SL

• **Goals**
  - Validation of form-meaning pairs, both for gesture and SL
  - Redefinition of Kendon's continuum in the light of the common semiogenetic hypothesis for gesture systems and SLs

• **Linguistic issues**
  - Status of the gestural units, of emblems
  - Relationships between form and meaning (semantics of segmental organizational centers)
  - Towards a standard for physiological annotation (Boutet, 2000)
**Methodology**:

- **POPULATION**: 3 kinds of dialogues between hearing and Deaf interviewees, aged between 18 and 60.
- **PROTOCOL**:
  - Gesture recognition tests (with 100 hearing people).
  - Explicative dialogues for 3 different pairs of subjects (H ↔ H; H ↔ D, D ↔ D) with 2 tasks:
    - Explain differences between eclipse of moon and sun
    - Explain different forms of a sailboat
- **PRINCIPLES**:
  - Adaptability of hearing and Deaf individuals in gesture communication
  - Different strategies in atypical gesture communication
SP3 : LSF Lexicogenesis
**SP3 : LSF Lexicogenesis**

- **Goals**
  - Data collection of:
    - recently stabilized (-ing) lexical signs in a discursive context
    - discursive genres less represented in existing SL corpora: dialogues and metalinguistic register

- **Linguistic issues**
  - role of High Iconicity in lexicogenesis
  - morphemic-iconic compositionality hypothesis
  - emergence and stabilization of lexemes
  - linguistic and extralinguistic factors at play in lexicalization
  - role of eye-gaze in the metalinguistic genre
Sous-Projet 3 – CREAGEST

Team

Brigitte GARCIA
enseignant-chercheur
coordinatrice de CREAGEST

Marie Thérèse L’Huillier
coordinatrice des enquêteurs
Enquêtrice Région Ile-de-France

Enquêteurs des régions

Benoît BLANDIN
Région Centre

Laurence COUTON
Région Est

Olivier DeLanghe
Région Sud-Ouest

Patrick VIVET
Région Sud
SP3: LSF Lexicogenesis

• Methodology:

  ▫ **POPULATION:** 80 Deaf subjects from 3 age brackets, with **LSF as main language**, 4 regions, varied professional settings

  ▫ **PRINCIPLES**

    ▫ **Peer interactions:** Deaf investigator from the same area as interviewee

    ▫ **Diversify areas** of investigation
SP3: LSF Lexicogenesis

**PROTOCOL**

- semi-directed interviews
- interview guidelines
- 3 cameras
- 1 h 30 interview for each interviewee (= 120 h of data)
  - 45 mns: discussion on a pre-established topic (recently emerged themes on the social domain)
  - 45 mns: metalinguistic discussion on signs from the first step
  - 5-10 mns free interaction
Preparatory phase:
- 9 training sessions with investigators in Paris (Jan. 2008/July 2009)
- 3-6 pilot interviews conducted by each investigator => 20 pilot interviews

Evaluation phase:
- Investigators ready
- Interviewees metadata validation
- Corpus recording phase: Sept. 2009-Jan. 2010
SP3 = LSF Lexicogenesis

video examples

Benoît (Centre-west)

Laurence (East)

Marie-Thérèse (Paris-IDF)

Patrick (South-east)
Methodological discussion

Investigator-training issues

- “To form”, “to inform”, “to deform” (formatting)
  - What level of language awareness (training in linguistics)?
  - What kind of advice on the type of questions asked? (eg.: avoid the sign [WORD] for [SIGN]? Avoid dactylology?...)
  - Interview guidelines vs. “inquisition” (spontaneity, naturalness)

- Durable commitment of investigators: heavy responsibility, several resignations
SP4 : Administering a collaborative database
SP4: Administering a collaborative database

Team: Dominique Boutet, Antonio Balvet, Pierre Guitteny and Evariste Ciret

- Training project members in corpora-related issues
  - data collection and metadata
  - ELAN platform and corpus annotation
  - ethical and legal aspects

- Hosting the database on a web server
  - cutting?/encoding of data and metadata
  - ensuring quality, compatibility and interoperability of collected/transcribed data

- Setting up a web portal
  - supervised download of corpora and metadata (cf. ADONIS)
  - online collaboration of project members, investigators and identified teams (French Deaf subjects and interpreters) working on neologisms
  - systematic collection of neologisms, peer-validation

=> creation of an observatory for LSF
SP4: Editing workflow

TGE ADONIS
Datawarehouse for Humanities

IN2P3

CRDO

OAIster
Open Archive Initiative

TEXT

AUDIO

META DATA

VIDEO

DATA (DV)

META DATA

OLAC

CREAGEST

ELAN
Annotation & Analysis

CREAGEST CORPUS
SP5: Tools for data transcription and mining
SP5: Tools for data transcription and mining

- Elan companion tools development
  - "lexicometry" tools: indexing, counting, sorting annotated structures
  - concordancer-like tools: complex queries in context
    - eg: [get all fragments matching a STD sign in a range of +/- Transfer units && where there exists at least 1 TP]
  - (transcription) data mining on an arbitrary range of transcription tiers
    - eg.: measure correlations between a given STD sign and any Transfer unit
  - transcription helpers: enforcing transcription rules for consistency
  - transcription corpus search engine
  - transcription versioning
SP5: Tools for data transcription and mining

- Similarity-based pattern extraction tools
  - based on CoPT project (A. Balvet, STL)
    - aim: identifying \textbf{Constructions} (Cx G) in transcribed data
    - recursively detecting similar pairs of symbols (// minimal pairs)
    - extracting recurring patterns for a given set of pairs => candidates for \texttt{constructions}
Discussion
Discussion


- How authentic are the collected data? (e.g. Schembri 2008)

- Training of Deaf interviewers, recruited as “accomplices” essential in creating the corpora
Discussion

Annotation issues

- Long tradition of transcription/annotation (LS-COLIN, LS-SCRIPT projects)
- Foreseen to **September 2010**
- **Pilot phase of corpora annotation** required
- **Absence of real system of transcription** (limit of gloses) = how to transcribe authentic iconic units respecting SL visual/gesture structure?
- **Need to exchange and share methodological tools and criteria** with hearing and Deaf researchers of SL
MANY THANKS:

CREAGEST TEAM

CREAGEST DEAF SIGNERS

BSL CORPUS PROJECT => A. SCHEMBRI