The Creagest Project

A Digitized and Annotated Corpus for French Sign Language (LSF) and Natural Gestural Languages

A. Balvet (Lille 3), B. Garcia (Paris 8)

C. Courtin (Paris 5)

D. Boutet, C. Cuxac, I. Fusellier-Souza,
M-T. L'Huillier, M-A. Sallandre
(Paris 8)

Outline

1. On Sign Languages
2. Objectives of the Creagest corpus
3. Methodological issues
4. Technical aspects
5. Theoretical/technical perspectives
6. Summary

On Sign Languages

- Visuo-gestural languages
  - ➔No standardized written form
  - ➔Variation

- Vocal language / SL
  - Some influence from the vocal language (French)
  - But 2 distinct linguistic types

On Sign Languages

- Main typical linguistic features
  - 2 signifying strategies
  - lexical signs = say without showing
  - "Highly Iconic Structures": Transfers = say by showing
  - Multi-parametric and multi-linear structures
  - Parameters: facial expressions + eyegaze + body movement + manual parameters
  - Each parameter is linguistically specialized

Objectives of the Creagest corpus project

- 3 main objectives
  - representativity
    - + complement existing LSF corpora
  - interoperability, sustainability
    - comparing SL corpora
    - accessing the digitized archives + transcriptions over long stretches of time (> 50 years)
  - Linguistic description
    - «Semiological model» (Cuxac)
    - Semiogenesis

3 sub-corpora

- Child LSF (ontogenesis)
  - 3-11 years old children (72 participants)
- Dialogues (lexicogenesis)
  - deaf/deaf interactions
- Natural gesturality (phylogenesis)
  - Natural gestures as a matrix for SL structures
  - explanation task: deaf/deaf, hearing/hearing, mixed dyads

Still pictures

- Child LSF

- Dialogues

Methodological issues

- ~300 h of digitized corpora, 250 signers
  - breakthrough for LSF
  - comparable with other large-scale projects
    - Auslan, BSL, NGT etc.
  - but crucial methodological options
    - not restricted to non-native speakers
      - < 5% of deaf children have LSF as their first language
    - accounting for HIS (Transfers)
      - ~ 40% in average
      - never transcribed, generally not glossed or annotated
      - glosses are not felicitous for lexical signs, even less for HIS
      - challenge for LS corpora annotation

Methodological issues

- Deaf interviewers

LSF child-acquisition team

- Deaf interviewers

Deaf investigators from 4 different regions

P. Palacios
SW

S. Heouaine
Center

N. Boursin
W

C. Fitzenwald
E

Lexicogenesis team

- Deaf interviewers

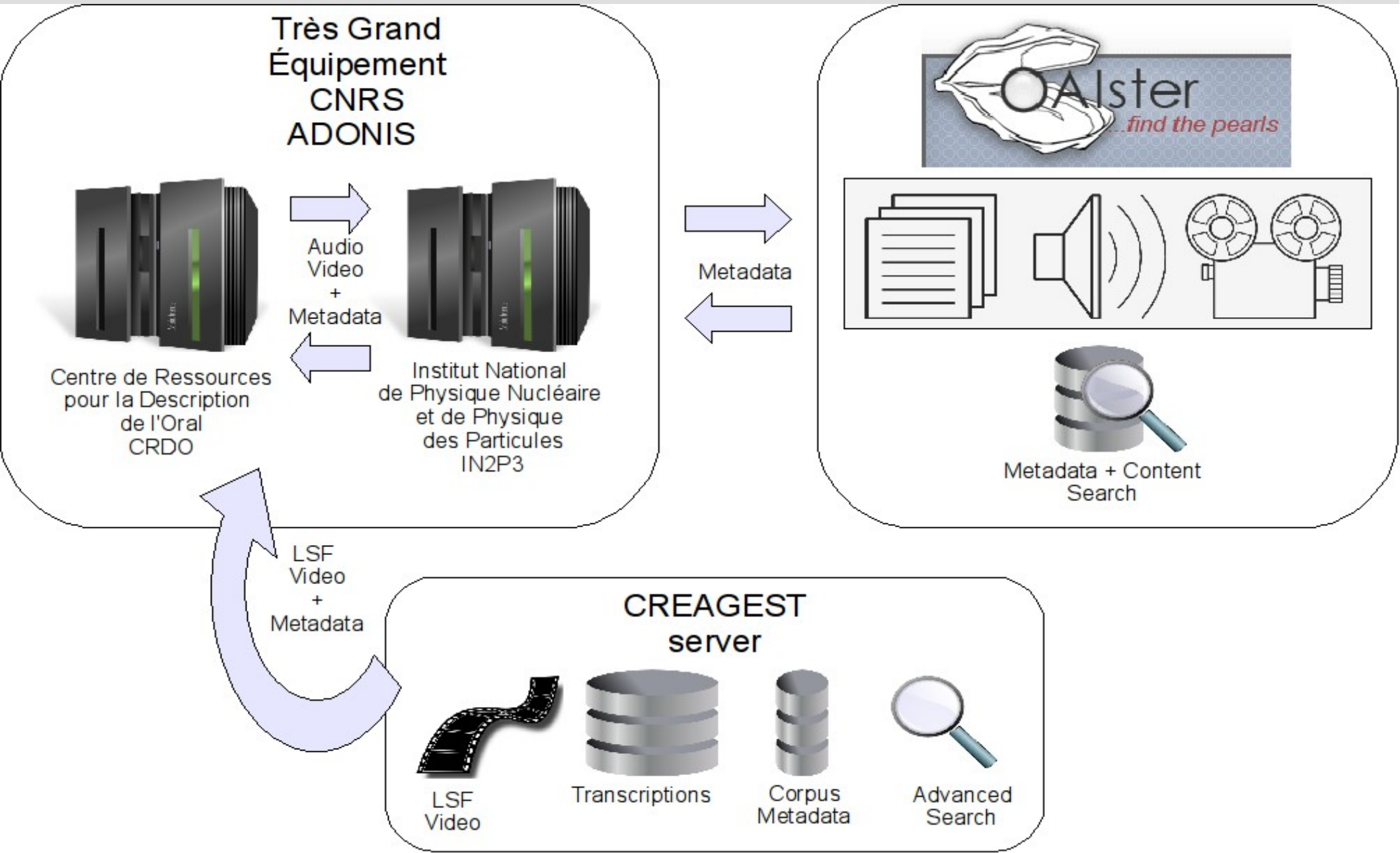
B. Blandin
Center-W

L. Couton
E

M-T. L'Huillier
Paris IDF

P. Vivet
S-SW

Technical aspects



Technical aspects

- A web-based collaborative and federative platform for corpus distribution
  - Archiving and search platform
  - Extended querying and search features
    - Elan companion tools
    - Adaptation of existing large corpora querying tools (eg. CQP)
  - Observatory for LSF
    - Sign creation

Theoretical/technical Perspectives

- Interaction between theoretical frame-work and practical aspects
  - New annotation tools + annotation scheme(s?)

➔Towards a computer-aided corpus-based LSF grammar

- Using annotations as a corpus
  - Spotting recurrent structures
  - Similarity assessment between emerging/established signs

➔[DESSIN/DESSINER] / [INFOGRAPHIE]

Summary

- ~300 h, 250 speakers, 3 sub-corpora
- Crucial methodological choices
  - eg.: Deaf interviewers, non-native speakers, HIS
- A technical infrastructure for the observation, description and dissemination of LSF data and analysis

Acknowledgments

- Main funding
  - ANR (Agence Nationale de la Recherche) Corpus
- Complementary financial support
  - DGLFLF (Délégation Générale à la Langue Française et aux Langues de France): visa #17852, november 2009

CREAGEST

Thank you for your attention