The Creagest Project: a Digitized and Annotated Corpus for French Sign Language (LSF) and Natural Gestural Languages

Antonio Balvet, Cyril Courtin, Dominique Boutet, Christian Cuxac, Ivani Fusellier-Souza, Brigitte Garcia, Marie Thérèse L'Huillier, Marie Anne Sallandre

The *Creagest* Project: a Digitized and Annotated Corpus for French Sign Language (LSF) and Natural Gestural Languages

Antonio Balvet, Cyril Courtin, Dominique Boutet, Christian Cuxac, Ivani Fusellier-Souza, Brigitte Garcia, Marie-Thérèse L'Huillier, Marie-Anne Sallandre

UMR 8163 STL (Université Lille 3 & CNRS),
UMR 6232 (Universités Caen/Paris Descartes & CNRS)
UMR 7023 SFL (Université Paris 8 &CNRS)
Université Lille Nord de France F-59653 Villeneuve d'Ascq,
Université Paris 5, Université Paris 8
antonio.balvet@univ-lille3.fr,cyril.courtin@paris5.sorbonne.fr,
dominique_jean.boutet@orange.fr, ccuxac@club-internet.fr,
ivani.fusellier@univ-paris8.fr,bridge.garcia@wanadoo.fr,
mt.lhuillier@wanadoo.fr, marie-anne.sallandre@univ-paris8.fr

Abstract

In this paper, we discuss the theoretical, sociolinguistic, methodological and technical objectives and issues of the French Creagest Project (2007-2012) in setting up, documenting and annotating a large corpus of adult and child French Sign Language (LSF) and of natural gestural language. In section 2., we address theoretical and practical issues, emphasizing the outstanding features of the Creagest Project. In section 3., we discuss methodological issues for data collection. Finally, in section 4., we cover technical aspects of LSF video data editing and corpus annotation, in the perspective of setting up a corpus-based formalized description of LSF.

1. Introduction

In this paper, we discuss the theoretical, sociolinguistic, methodological and technical objectives and issues of the French Creagest Project (2007-2012) in setting up, documenting and annotating a large corpus of adult and child French Sign Language (LSF) and of natural gestural language. The main objective of this ANR[1]-funded research project is to set up a collaborative web-based platform for the study of semiogenesis in LSF (French Sign Language), *i.e.* the study of emerging structures and signs, be they used by Deaf adult signers, Deaf children, or even by Deaf and hearing subjects in interaction.

In section 2., we will address theoretical and practical issues, emphasizing the outstanding features of the Creagest Project. In section 3., we will deal with methodological issues for data collection. Finally, in section 4., we will examine technical aspects of LSF video data editing and corpus annotation, in the perspective of setting up a corpus-based formalized description of LSF.

2. Objectives of the Creagest corpus

2.1. Devising a large-scale, digitized and annotated LSF corpus

The Creagest corpus project is a large-scale undertaking: over 500 hours of digital video are being recorded, involving over 250 signers: hearing and Deaf adult signers, Deaf children signers. To our knowledge, the Creagest Project is the first large-scale project for LSF and natural gesturality

2.2. Outstanding features of the Creagest Project

The first outstanding feature of this project is that it aims at devising a representative controlled corpus of LSF, as used all over metropolitan France, encompassing as much linguistic and sociolinguistic diversity as possible. The Creagest corpus is thus designed to complement existing LSF corpora, issuing from previous research projects. (Sallandre and Braffort, 2009) proposed a thorough census of LSF corpora existing to this day. One of the conclusions of their study is that, even though the majority of existing corpora were aimed from the onset at the analysis of long utterances (see 2.3.), most of the time, they were *in situ* corpora, lacking control on their technical quality and systematic metadata documentation (speaker description). In their vast majority, LSF corpora are monologic productions (stories, conferences), by adult Deaf signers. More controlled corpora do exist, though, e.g: *DGLFLF-UMR SFL 2004* (Garcia, 2005), *TALS IRIS-LIMSI 2005*[2], but they lack either speaker or genre diversity, or their size incompatible with the objective of a truly corpus-based formalized description of LSF.

The first large-scale systematic LSF corpus collection is the LS-Colin corpus (Cuxac and *alii.*, 2002): it is composed of 90 productions, by 13 adult Deaf signers, with an emphasis on genre diversity (stories, cooking recipes, personal opinions). Though it is the largest LSF corpus existing as of today, it is still made up of monologic productions, by adult signers. One of the crucial objectives of the Creagest Project is to complement existing corpora. Therefore, we strive to:

- ensure effective representativity of our data, by collecting corpora from a variety of speakers: origins, ages, family backgrounds;
- provide dialogic productions as well as monologic ones, with an emphasis on less-represented genres

[1] Agence Nationale de la Recherche.

[2] See http://tals.limsi.fr/.

(metalinguistic, explicative, descriptive genres);

- devise the first systematic corpus of natural gesturality, by confronting Deaf and hearing signers.

The other outstanding feature of this project is our intention to provide both the research and Deaf/signers' community with complete and free access to the digitized and annotated corpora. In order to achieve this goal, all the technical options we selected for data recording, digitizing, editing, archiving and metadata edition are meant to ensure maximum accessibility and interoperability. The same holds true for the choice of annotation tools and the annotation methodology itself (cf. section 4.).

2.3. Theoretical issues

Along with preservation and archiving objectives, and together with fine-grained description of LSF in its variety, the future Creagest corpus is meant to support the theoretical developments, initiated by (Cuxac, 1985), Cuxac (1996) and Cuxac (2000), in what he terms the "semiological model". This novel approach to Sign Language was designed in a truly corpus-based approach[3]. In this approach, iconicity is considered as an organizing principle of every Sign Language. His model poses a common iconicization principle of human perception and practical experience, which is supposedly shared by both natural human gesturality and semiogenesis, *i.e.* the emergence and structuring of signifying gestures in Sign Languages. His hypothesis mentions four main factors influencing the final form of emerging signs:

- the genetic origin of deafness, and the subsequent atypical modality of its transmission: 95% of Deaf children are born to hearing non-signing families, and therefore do not have SL as their mother tongue;
- the sociological situation of Deaf people and their family, and the attitude towards her/his spontaneous gestural creations by her/his community;
- the practical conditions of linguistic communication for a non-hearing individual: maximum exploitation of the only accessible channel, the visuo-gestural one;
- the different steps in cognitive development.

According to Cuxac, this hypothesis helps account for the emergence and differentiation, from an initial iconicization process, of two of the main linguistic features which define Sign Languages:

- structures aiming to saying while showing, labelled **Highly Iconic Structures** or **Transfers** in Cuxac's terminology;
- structures aiming solely to saying, labelled **Standard Signs** (or, more recently, "lexemes"[4].

This hypothesis is also an explicative model for the discursive (as opposed to syntax) orientation of contemporary institutionalized Sign Languages like LSF (Cuxac, 2000), (Sallandre, 2003). It is also a predictive model of their diachronic evolution dynamics (lexical emergence).

From the perspective of the Creagest team, **Highly Iconic Structures** are a central linguistic device of Sign Languages. This position is not shared by the vast majority of Sign Language linguists around the world, who generally assume these structures to be peripheral at best, or even outside the range of language altogether (Garcia, 2010) and (Boutet et al., 2010) [5].

As an illustration of the first factor mentioned above, the most frequent situation for a signing Deaf person is to be born to a non-signing family, which means signing Deaf people are seldom "native speakers" of their own language. In our theoretical perspective, this situation is of great import for the emergence, structuring and development of a Sign Language. Therefore, being a native signer of LSF (which represents less than 5% of the signing Deaf community) is not a selection criterion in our project; it is nothing more than a piece of metadata among others. This position sets Creagest apart from the majority of other SL-centered corpora projects, focusing on native (or "near-native") speakers. From our viewpoint, the only relevant criterion is that speakers must have LSF as their main (or reference) language[6].

The theoretical developments the Creagest corpus is meant to support, in the framework of the semiological model, are divided into three sub-corpora, each of which focusing on less often studied aspects of the semiogenesis process.

- SP 1: emergence of LSF in the ontogenetical perspective of its acquisition by Deaf children;
- SP 2: characteristics and respective potentials of SL and human natural gesturality (with a focus on illustrative coverbal gestures) and their interrelation in the emergence of SLs;
- SP 3: processes and composition rules underlying the emergence and stabilization of new lexical units in LSF (dynamic synchronicity and linguistic change).

The nature of each sub-corpus is therefore highly dependent on the research objectives underlying each sub-project:

- SP1: discourse utterances of child LSF, collected from Deaf children, ranging from 3 to 11 years old;
- SP2: discourse and descriptive type corpora, collected in a parallel fashion from adult Deaf adult signers and non-Deaf people;
- SP3: LSF dialog corpora between Deaf adult signers, on the topic of lexical creation in LSF.

[3] As opposed to *corpus-driven*, as defined by (Tognini-Bonelli, 2001)): Cuxac posed that since most Sign Language researchers are non native speakers of SLs, no corpus-driven (Competence-based) description of their grammar can be provided.

[4] see (Cuxac and Pizzuto, 2010).

[5] See section 4. for more details on this topic.

[6] See (Cuxac, 1996), Cuxac (2000), (Sallandre, 2003), (Fusellier-Souza, 2004), (Jacob, 2007), (Pizzuto et al., 2007), (Cuxac and Pizzuto, 2010) and (Garcia, 2010) for more detail on this topic.

Each sub-corpus is devised according to different methodological options, namely with regard to elicitation procedures and the material used as stimulus.

3. Methodological issues

All three sub-corpora have the common goal of complementing existing corpora (see section 2.) in the general framework of the semiological model. In this section, we summarize each sub-project's main objective. In the last subsection, we discuss some of the common methodological issues to each sub-project.

3.1. SP1

This sub-project aims at collecting free as well as induced LSF production in 72 deaf children, using four different tasks –this project is not designed to assess children's comprehension. First, a free LSF dialog occurs during individual sessions between the children and a Deaf interviewer. Then children are shown some items designed to trigger the expression of path and manner in verbs of motion, in order to identify the structures they use for this purpose (e.g., personal transfers, situational transfers, or frozen signs). The last two tasks are devoted to narratives: in the first task children are shown a familiar cartoon, while in the second task picture drawings are used as prompts. In both cases the children are asked to tell a naïve person what they have been presented. In SP1, the experimental constants are as follows:

- 2 children per age bracket;
- all children are profoundly deaf;
- Sign Language is used non-exclusively as the main communication language, other languages may be used by the family and child;
- SL input: SL is learned in a classroom context: monomodal (SL) or bimodal (SL + French or other oral language) education (classes in SL, classes of SL).

Figure 1 below shows a still picture taken from one of the pilot recordings, where Deaf interviewer M-T. L'Huillier asks a Deaf child to tell her a story based on a visual stimulus.

Figure 1: SP1 pilot corpus recording with a deaf child

3.2. SP2

This sub-project aims at collecting explicative genre dialogs, where pairs of 5 hearing-hearing, 5 Deaf-Deaf and Deaf-hearing individuals are involved. In the latter case, 15 dialogs are being recorded, of which 10 involve a directionality of the explanations: 10 Deaf –¿ hearing dialogs, of which 6 involve Deaf "accomplices" who are asked to pretend either to help the hearing individual, or to simulate not being able to understand her/him. Two main explanation tasks were defined:

- explaining the difference between a moon and a sun eclipse;
- explaining the angle a sailboat must take under different wind directions, in order to achieve forward motion (i.e. point of sail).

3.3. SP3

Sub-project 3 aims at collecting 53 semi-directive interviews, of 90 minutes each. A 45 minute phase of guided interview on a set of pre-established topics is meant to collect newly created lexical units in a discursive context. It is followed by another 45 minute phase of metalinguistic discussions on the lexical units collected. Each interview is concluded by an informal discussion of at least 5 minutes. The panel of interviewees was devised in order to gather a representative sample of new LSF units, from all over metropolitan France, across all age brackets (from 18 to 45 years old), socio-professional settings and linguistic backgrounds. Figures 2 and 3 below show still pictures of the interviews described above.

Figure 2: SP3: South of France area

3.4. Overall methodological issues

For all three sub-projects (SP1, SP2 and SP3), the data collected are, in their vast majority, long utterances, but methods of data collection are adapted to each sub-project. They range from guided interaction in a carefully controlled experimental setting (SP2 tests, targeted stimuli for SP1) to informal interviews (the informal phases of SP1 and SP3), which are meant to provide a baseline for our experiments. One methodological choice we wish to emphasize is our calling on Deaf interviewers who have a strong connection to the region in which interviews are conducted (because

Figure 3: SP3: Center-West of France area

of personal and family history) for guided or semi-directed phases of data collection in SP1 and SP3. In order to ensure the quality of the data collected, we deliberately included the training of Deaf interviewers as a necessary step, prior to recording each subcorpus. These Deaf interviewers, recruited as "accomplices", are essential in creating the corpora and in ascertaining that the LSF speaking community as a whole adheres to the research objectives underpinning our project. Therefore, each sub-corpus includes a preparatory Deaf interviewer training phase in field linguistics. We consider this field linguistics training phase as providing the beginning of an answer to a question inherent to the exponential development of vast controlled corpora of SL discourse: how authentic are the data collected[7]? This question is crucial for all corpora-based or corpora-driven linguistic research, but even more so in our theoretical perspective, since the semiogenetic model stems from an original tradition of LSF descriptions based mainly on extensive spontaneous discourse corpora (Garcia, 2000), rather than elicited material. In SP3, for instance, Deaf interviewers are the keystone of the whole project, as they are in charge of all the interviews performed. Therefore, a large amount of preparation was necessary prior to undertaking this sub-project: extensive training in both technical and methodological aspects of field linguistics was provided to each interviewer, so as to guarantee unimpeded and spontaneous interaction between Deaf interviewers and interviewees. Interviewers therefore had to follow a strict interviewing guide, while appearing to interact freely with interviewees[8]: they were specifically asked not to use any written or other visual support during interviews.

We believe it is of utmost importance that the interviewer be a signing Deaf individual, given that in this sub-project we are striving to collect metalinguistic data from "naïve" Deaf signers, who have never had the opportunity to phrase such metalinguistic considerations in and on LSF, since it is a minority language in competition with a well-established national oral language (French), that has had strong institutional backup for more than two centuries[9].

[7] See (Schembri, 2008) on that topic.

[8] See (Schembri, 2008) on this topic.

[9] The use of LSF in France was in fact forbidden in specialized (mostly vocational) schools for the Deaf run under the French ministry of Health, up till 1977, and in the overall education system until 1991.

4. Technical aspects

In this section, we discuss practical and computational issues related to the archiving, distribution and exploitation of annotated LSF video corpora.

4.1. A web-based collaborative platform for corpus distribution

The Creagest website is available at: http://www.creagest.cnrs.fr/[10]. It is intended to provide a web-based platform for data archiving, indexing and distribution for all the sub-projects presented above. In addition, it is meant as a long-term open neologism repository for the Deaf community.

4.2. A web-based archiving and search platform

In order to be of service to the community, a project of the magnitude of the Creagest Project must deal with the practical issue of giving access to the finalized corpora, together with their transcription(s), while enforcing access restriction policies (public, protected, private). Therefore, a subproject is specifically concerned with deploying and administering a collaborative database framework underlying and supporting each subcorpus. This sub-project is mainly concerned with the technical infrastructure for the digitization, archiving, video editing, compression and distribution of our corpora over the internet. A major concern for us is ensuring the quality, compatibility and interoperability of the data collected/transcribed. So far, this issue has been dealt with by adopting the OAI standards[11],in collaboration with the TGE Adonis facility[12] for storing, indexing and distributing large digital file collections[13].

Upon completion of the Creagest Project, we will therefore be able to provide supervised access and downloading of corpora and metadata through our website, which is essential for enforcing the legal and ethical aspects of the project. For example, corpora involving children will not be publicly available before the subjects turn 18 and give written permission to allow public access to their productions. In the meantime, the metadata associated to their productions will nevertheless be publicly accessible, while the footage and annotations will only be accessible by researchers of the Creagest team. The website will therefore enforce different levels of privileges and access rights, following a classical public/restricted/private hierarchy. The website will also provide online collaboration features for project members, investigators and identified research or end-user groups. For example, Deaf speakers working on neologisms (LSF teachers, interpreters or parents) will be

[10] A password is required.

[11] Open Archive Initiative, see: http://www.openarchives.org/

[12] Très Grand Équipement (Very Large Facility), a CNRS-funded collection of technical facilities, encompassing telescopes, computer grids, computing centers etc. TGE Adonis is dedicated to the storage, archiving, format re-encoding and distribution of digital data for different scientific domains, including the humanities.

[13] See figure 4 for an overall view of the interrelationships between Creagest, TGE Adonis and OAIster

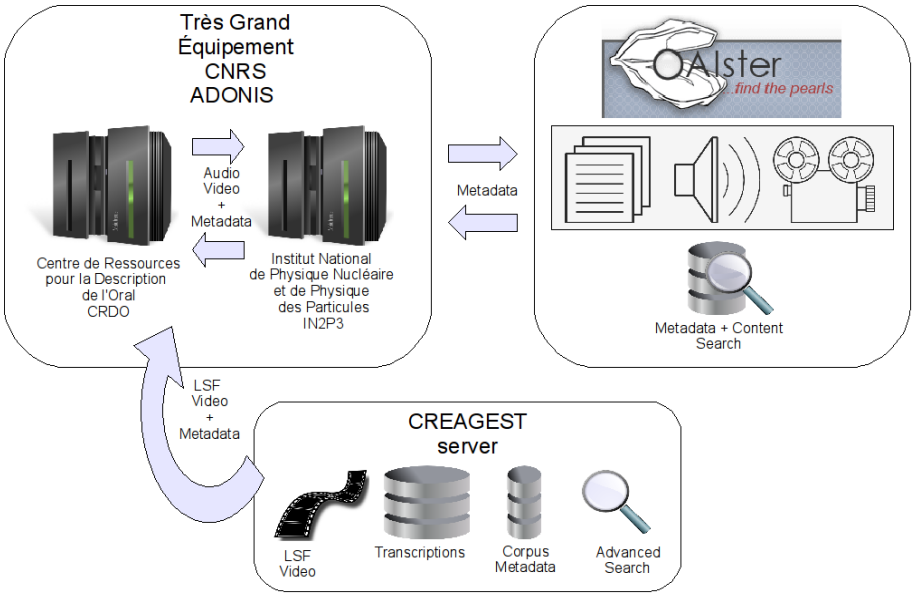

Figure 4: LSF data and metadata archiving and distribution architecture for the Creagest platform

able to upload "new signs" (neologisms) to the website, which will be discussed among peers, thus offering a platform for the systematic collection of new lexical units in LSF.

4.3. Extended querying and search features

Alongside the technical infrastructure outlined in the preceding section, software development is in progress, which is aimed at easing (and ensuring the coherence of) the annotation process, and providing data mining and other corpus-linguistics tools such as concordancers, adapted and customized for SL research. Moreover, we hypothesize that (manually and automatically) spotting recurring patterns among annotations together with assessing degrees of similarity among transcriptions are essential to a formal description (*i.e.* a grammar) of LSF, one of the long-term goals of the project. Therefore, two development sub-tasks have been devised.

Elan companion tools

The ELAN platform[14] is our main corpus annotation tool, the participants in this sub-project are therefore in close connection with the ELAN development group at Max Planck Institute. The main reason for choosing Elan as a central annotation tool is that it is becoming a standard among Sign Language linguistics researchers, in place of other multimedia-annotation tools, such as Anvil[15] or more general annotation tools such as the Nite-XML Toolkit[16]. The constant development efforts by the MPI throughout the years and the close interaction between the MPI development team and Sign Language research groups[17] have yielded an annotation tool which is free, open-source[18], based on standard formats (e.g. XML, MPEG, but also comma-separated text if the need should arise), and capable of importing from and exporting to other widespread text or speech annotation tools (e.g. Shoebox, Praat, Childes). Elan appears therefore to be specifically oriented toward Sign Language data annotation, and is thoroughly integrated in the MPI's general infrastructure for corpus archiving and distribution, in a manner none of the aforementioned tools could provide. Elan is but one of several modules composing the Language Archiving Technology project at MPI[19], which also includes a web-based lexicon tool, a manual syntactic annotation tool, and other corpus-related software. For all these reasons, from the viewpoint of the Creagest Project, Elan represents the best choice in the (small) world of multimedia annotation tools, even though it lacks certain useful features. One of the most important features for the long-term goal underpinning the present corpus project, i.e. devising a corpus-based grammar of LSF, is a robust indexing and search engine, together with concordancer-like features[20]. Granted, these features do exist in the current version of the tool, but they are clearly meant for a restricted usage scenario: an isolated researcher, working on a finite (and not too large) set of annotated files, wishes to extract all occurrences of a given pattern, either on one isolated file or on a collection of Elan files stored on a local disk, with results of the search presented in the form of a concordance[21], linked to the original video data and its annotations.

Upon completion of the Creagest Project, we will be confronted with several hundred annotated files, stored on a remote server (with different access privileges), to which

[14] Available at http://www.lat-mpi.eu/tools/elan/.

[15] Available at: http://www.anvil-software.de/.

[16] See http://groups.inf.ed.ac.uk/nxt/.

[17] To name but a few: Nijmegen University and the NGT corpus project, London University and the BSL corpus project, Macquarie University/London University and the Auslan corpus project.

[18] Though not intended to be really modular or plug-in oriented.

[19] See http://www.lat-mpi.eu/tools/ for more information.

[20] A tool capable of extracting units matching a given query, and of presenting them in their original context.

[21] A list of matching occurrences, generally presented in a limited context of a given span.

concurrent queries might be applied. This represents a totally different usage scenario from the one briefly described above. Moreover, in our perspective, once the annotation process is complete for a collection of files, researchers will typically try to use actual LSF usage to devise parts of a grammar of LSF, and to confront existing hypotheses to the data recorded, in order to comply with the scientific objectives outlined in section 2.3.. Concordances built on any given sub-corpus[22] will therefore be a central tool for this task, which means, again, a wholly different usage scenario than the one underlying Elan's existing search and concordancing features.

We have therefore planned to hire professional developers from a subcontracting company in order to develop a set of "companion tools" to Elan, in order to provide researchers with the minimum set of corpus-linguistics tools necessary for exploiting the data collected and their annotations. In addition to these tools, we have also planned to develop annotation helpers, capable of identifying inconsistencies in the annotated files, or to automatically propose annotation completions, based on past annotation behavior and possibly on a set of annotation rules. For example, in our descriptive framework, Highly Iconic Structures are seldom used without being introduced by Standard units: in stories, actants are generally introduced by Standard signs (FROG, HORSE, etc.) and then referred to using HIS (e.g. Personal Transfers) or pointings. We hope to use these discourse-level regularities together with Elan's Controlled Vocabulary feature to ensure maximum consistency of annotations among annotators from a given sub-project. These tools will, of course, be developed under an open-source license so as to be further extended and modified by other research groups.

Towards a computer-aided corpus-based LSF grammar

As mentioned above, one of the long-term research goals of the corpus collection projects mentioned in this paper is to propose a corpus-based formalized description (a grammar) for LSF. Even though members of the Creagest research group, and more widely Sign Language linguists are aware of Chomsky's rebuttal of a bottom-up approach to grammar, LSF cannot be considered the same as any other (vocal) language: it has no standard written form, it therefore possesses almost no written grammatical tradition (as opposed to French), from which a body of carefully constructed rules could be derived, as is the case for oral languages such as French or English. Due to its multi-segmental properties (SL structures are generally composed of hand gestures, facial expressions, eye-gaze and other non manual parameters at least), its visual modality and inherent diversity, the question of unit identification and delimitation arises even from the transcription/annotation stage. Moreover, most SL linguists are not native speakers of the language they are attempting to describe, which bars resorting to Sign Language linguistic Competence (in Chomsky's terms, i.e. an internal model of the language). It follows that no consistent (formal or non formal) grammar of SL can be devised, based on internal SL Competence. It could be said that, in this respect, SLs should best be thought of as an unknown language. Therefore, representative and controlled corpora are the only means to achieve any consistent description of a given SL. In the framework of the Creagest Project, we posit that in order to address the issues mentioned above, namely: 1) delimiting LSF units based on explicit criteria (rules), 2) proposing a corpus-based formal description of LSF structures, the identification of naturally recurring patterns among annotation levels is a necessary step, in order to supplement linguistic intuition. In the framework of the Creagest Project, along with Elan companion tools, we plan to use and further develop existing corpus processing tools for the task of detecting recurring patterns, extending well-known approaches (collocation extraction) as well as novel algorithms. For instance, by using similarity measures between two strings (two annotation tiers from two different files), we can automatically identify optimal alignments, in other words minimal pairs, among these strings. We posit that by applying this basic process on all similar pairs of annotation symbols taken from our annotated files, we will be able to spot recurring structures, in a semi-automatic fashion. As of today, this approach is clearly not optimized, as it requires examining a $\frac{n(n-1)}{2}$ function on the number of strings (i.e. annotation tiers). This means that, for any reasonable corpus, say 500 annotated corpora of which we extract a single (possibly long) tier, the amount of data to be processed becomes quite large: 124,750 candidate pairs, of which a small percentage yields meaningful patterns. If we perform the same operation on tier segments (a tier is composed of 1 segment or more, e.g.: beginning, unfolding and end of story), we are likely to be confronted with possibly untractable computing times, even though the similarity computation using standard algorithms (e.g. Levenshtein edit-distance) is a linear function of the size of the longest string: if we consider an average of 30 segments (propositions) per utterance (story, interview), we would have to process over 110 million pairs. Nevertheless, even in its present implementation, this approach yields interesting insights into the structure of LSF utterances. Moreover, it allows for a crude yet efficient "query by example" search feature: the overall similarity of a given (set of) tier(s) can be computed on two annotation file pairs, yielding the most similar pairs of the corpus for the tier under consideration. This allows for mining a corpus of Elan annotation files, and ranking it according to their similarity to a given file. We plan to extend this simple pattern-detection approach in order to perform optimal alignments on an arbitrary number of tiers at the same time. In other words, we wish to perform the same task not only in a pair-wise fashion, but also on an arbitrary range of tiers, and detect whatever recurring patterns are found, regardless of the tiers. By following this approach, we wish to support intuitions and corpus-experience of SL linguists by actual linguistic patterns found in annotated corpora, following an explicit methodology.

5. Conclusion and future research

We have outlined the main objectives of the Creagest corpus building and annotation project, an ongoing project for

[22] Using metadata, e.g. a sub-corpus made up of "all productions from the South and West of France where a cochlear-implanted Deaf child uses Highly Iconic Structures in the initial phases of his stories".

the description, formalization and dissemination of French Sign Language. This project is centered on the recording of three sub-corpora: the first one centered on LSF acquisition processes, the second on the relationships between natural gestural languages (of both hearing and Deaf people) and SL, and the third on lexicogenesis (the genesis of signs). This project also addresses technical, sociocultural as well as ethical issues. One of the outstanding methodological options in this project is the involvement of the LSF speaking community, by recruiting and training Deaf interviewers as well as by providing a technical infrastructure for the observation, description and dissemination of LSF data and analysis on a long-term basis. By so doing, the Creagest Project intends not only to be a Sign Language research effort, but also to pave the way to an observatory of LSF and natural gestural languages.

6. Acknowledgments

The Creagest Project is an ANR-funded research project, involving three academic research laboratories: UMR 7023 SFL (Université Paris8 &CNRS), UMR 6232 (Universités Caen/Paris Descartes & CNRS), UMR 8163 STL (Université Lille 3 & CNRS). It also receives financial support from DGLFLF (Délégation Générale à la Langue Française et aux Langues de France): visa #17852, november 2009.

7. References

Dominique Boutet, Marie-Anne Sallandre, and Ivani Fusellier-Souza. 2010. Gestualité humaine et langues de signes : entre continuuum et variations. In B. Garcia and M. Derycke, editors, *Langage et Société*, number 131, pages 55–74. Maison des sciences de l'homme.

Christian Cuxac and Elena Antinoro Pizzuto. 2010. Émergence, norme et variation dans les langues des signes : vers une redéfinition notionnelle. In B. Garcia and M. Derycke, editors, *Langage et Société*, number 131, pages 37–53. Maison des sciences de l'homme.

Christian Cuxac and *alii.* 2002. Projet LS-Colin, Rapport de fin de recherche. Research report, Université Paris 8–Saint-Denis and Université Paris 5.

Christian Cuxac. 1985. Esquisse d'une typologie des langues des signes. In *Autour de la langue des signes, Journées d'Études 10*, pages 35–60. UFR de linguistique générale et appliquée, Université René Descartes.

Christian Cuxac. 1996. *Fonctions et structures de l'iconicité. Analyse descriptive d'un idioloecte parisien de la Langue des Signes Française*. Ph.D. thesis, Université Paris 5.

Christian Cuxac. 2000. *La Langue des Signes Française (LSF), Les voies de l'iconicité*. Ophrys.

Ivani Fusellier-Souza. 2004. *Sémiogenèse des langues des signes. Primitives conceptuelles et linguistiques des langues des signes primaires*. Ph.D. thesis, Université Paris 8–Saint-Denis.

Brigitte Garcia. 2000. *Contribution à l'histoire des recherches linguistiques sur la Langue des Signes Française (LSF). Les travaux de Paul Jouison*. Ph.D. thesis, Université Paris 5.

Brigitte Garcia. 2005. Projet LSF : quelles implications pour quelles formes graphiques ? Research report, DGLFLF, Ministère de la Culture et de la Communication.

Brigitte Garcia. 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)*. Habilitation thesis, Université Paris 8–Saint-Denis.

Stéphanie Jacob. 2007. *Description des procédés référentiels dans des narrations enfantines en Langue des Signes Française : maintien et réintroduction des actants*. Ph.D. thesis, Université Paris 8–Saint-Denis.

Elena Pizzuto, Paola Pietrandrea, and Raffaele Simone, editors. 2007. *Verbal and Signed Languages. Comparing Structures, Constructs and Methodologies*. Mouton de Gruyter.

Marie-Anne Sallandre and Annelies Braffort. 2009. LSF resources (French Sign Language). In Onno Crasborn, editor, *Sign Linguistics Corpora Network 1*, july.

Marie-Anne Sallandre. 2003. *Les unités du discours en Langue des Signes Française (LSF). Tentative de catégorisation dans le cadre d'une grammaire de l'iconicité*. Ph.D. thesis, Université Paris 8–Saint-Denis.

Adam Schembri. 2008. British sign language corpus project: Open access archives and the observer's paradox. In *LREC (Language Resources and Evaluation Conference) 2008, Proceedings of the workshop on the representation and processing of Sign Languages*, pages 165–169.

E. Tognini-Bonelli. 2001. *Corpus Linguistics at Work*. John Benjamin's Publishing.