Outils et Recherches pour les Corpus d'Acquisition du Langage

Tools and Research based on Language Acquisition Corpora

Version française (see English version below)

Outils et Recherches pour les Corpus d'Acquisition du Langage

18-19 Novembre 2010, PARIS

Centre CNRS Pouchet, 59 rue Pouchet, 75017 PARIS
Contact : cparisse@u-paris10.fr

Les corpus de langues orales ou signées sont aujourd'hui largement utilisés dans les recherches sur l'acquisition du langage. Cette évolution a nécessité de nouveaux outils, de nouvelles pratiques de recherche et de nouvelles théories qui ont amené des changements majeurs dans les recherches en linguistique et en acquisition du langage. Il est nécessaire de partager et diffuser auprès des chercheurs et des étudiants les résultats, les techniques et les transformations récentes qui résultent de cette évolution. En effet, ces produits viennent d'un effort collectif qu'il faut encourager et amplifier afin d'arriver à encore mieux créer, utiliser, partager, diffuser les données de corpus de langage spontané. C'est en utilisant mieux les outils et corpus récents que la recherche, nationale et internationale, pourra s'enrichir et prospérer.

Pour cela, un workshop incluant des tutoriels et des présentations sur les outils et la recherche utilisant les corpus d'acquisition de langage sera organisé par le GDR « Langues, langage oral et cognition : acquisition et dysfonctionnements – nouvelles approches » et financé par le CNRS. Ce workshop est conçu pour :

- présenter et introduire des outils utilisés fréquemment pour créer et éditer des corpus de langage oral (langues orales ou signées). Un tutoriel concernera les usages avancés des corpus ;
- présenter et discuter les évolutions en matière d'outils pour corpus de langage oral ;
- présenter et discuter l'intérêt et les apports des recherches utilisant des corpus de langage oral.

Programme

L'organisation de ces tutoriels pourra évoluer en fonction du nombre de personnes intéressées. Pour cela, toute personne intéressée par ces tutoriels doit écrire à Christophe Parisse (cparisse@u-paris10.fr) pour se préinscrire.

Jeudi 18 novembre 2010 (Matin)

Présentation et introduction aux logiciels CLAN et PHON ; transcription et utilisation pour la recherche. Les présentations se focaliseront sur les principes de base et fourniront des exemples de

recherches effectivement menées avec ces outils. Elles seront organisées sous forme de tutoriels et seront ouvertes à des utilisateurs novices.

9h00 – 10h30: **Stéphanie Caet** (Prismes – Sorbonne Nouvelle Paris 3)
CLAN pour les Néophytes : découvrir, comprendre et utiliser un nouvel outil de transcription et d'analyse des interactions

10h45 – 12h15: **Naomi Yamaguchi (**LPP – Sorbonne Nouvelle Paris 3, SFL – CNRS-Paris 8)
Transcription phonétique et analyse phonologique avec PHON.

Jeudi 18 novembre 2010 (Après-Midi)

Présentation et introduction au logiciel ELAN ; transcription et utilisation pour la recherche. Utilisation de corpus déjà existants : comment traiter ces corpus, comment réaliser des statistiques, comment réaliser une analyse syntaxique ou lexicale.

14h00 – 16h00: **Simon Harrison** & **Vito Evola** (Natural Media & Engineering, HumTec Center, RWTH Aachen University)
Une introduction à ELAN pour l'analyse et la transcription de données multilingues (présentation en anglais)

16h15 – 17h45: **Christophe Parisse** (MoDyCo-INSERM, CNRS-Université Paris Ouest Nanterre)
Utiliser, traiter et interroger des corpus existants

Vendredi 19 novembre 2010 (Matin)

Nouvelles tendances et évolutions dans les outils et les corpus de langage oral, annotation et traitement des corpus. Comment est-il possible d'utiliser efficacement des outils pour travailler sur l'acquisition du langage oral ? Quels avantages ces outils offrent-ils pour la recherche ? Quelles améliorations attendre en particulier dans le traitement de corpus multimédia et de langage oral ? Est-ce que les outils actuels vont évoluer ou de nouveaux outils vont-ils se développer ?

9h30-10h15: **Han Sloetjes** (Max Planck Institute for Psycholinguistics, Nijmegen)
ELAN – développements récents et futurs (présentation en anglais)

10h15-11h00: **Michael Kipp** (Saarland University, Germany)
ANVIL: un outil de codage et d'analyse vidéo multi-niveaux (présentation en anglais)

11h20: 12h05: **Thomas Schmidt** (Universität Hamburg, Germany)
EXMARaLDA : système de formats et outils pour la transcription et l'annotation de la langue parlée (présentation en anglais)

Vendredi 19 novembre 2010 (Après-Midi)

Nouvelles tendances dans les recherches en acquisition du langage. Qu'est-ce que l'utilisation de corpus a apporté aux recherches en acquisition du langage ? Comment ont évolué les liens entre recherche théorique et applications pratiques ? Des présentations seront réalisées par des invités sur

des thèmes allant de la création de données de corpus à l'utilisation de ces données pour la recherche en acquisition du langage ou en linguistique computationnelle :

14h-14h50: **Susanne Miyata** (Aichi Shukutoku University, Nagoya, Japan)
CHILDES: Corpus en langue japonaise et recherches en acquisition du langage (présentation en anglais)

14h50-15h40: **Heidi Waterfall** (Cornell University, NY, USA)
La variété est le sel de l'analyse longitudinale (présentation en anglais)

16h-17h: **Elena Lieven** (MPI-EVA, Leipzig –Germany- and University of Manchester - UK)
Recherches basées sur des corpus: évolutions récentes, difficultés et ouvertures (présentation en anglais)

Le workshop et les tutoriels se dérouleront à Paris, au Laboratoire SFL (UMR 7023, CNRS), Centre CNRS Pouchet, 59 rue Pouchet, PARIS.

Comité d'organisation :
Christophe Parisse (Modyco, INSERM, CNRS/Université Paris Ouest Nanterre la Défense, France)
Aliyah Morgenstern (Prismes, Université Sorbonne Nouvelle, Paris, France)
Maya Hickmann (SFL, CNRS-Paris 8, Paris, France)
Financé par le GDR CNRS « Langue, Langage Oral et Cognition: acquisition et dysfonctionnements - nouvelles approches », responsable Maya Hickmann.

Tools and Research based on Language Acquisition Corpora

November 18-19 2010, PARIS (FRANCE)

CNRS Pouchet center, 59 rue Pouchet, 75017 PARIS (FRANCE)
Contact: Christophe Parisse cparisse@u-paris10.fr

Multimodal language corpora of spoken and signed languages are now widely and extensively used in language acquisition studies. This has required new tools, new research practices, new theories which has brought about ground breaking changes in research in linguistics and in psycholinguistics. These recent transformations, techniques and results need to be shared with a large community of researchers and students in the field of language acquisition. Creating, using, sharing, analyzing spontaneous oral data with the relevant tools for each researcher's specific needs can only help improve and enrich national and international research on language acquisition, and collective efforts are needed to attain this goal.

A Workshop including tutorials and various presentations, will be organized by the GDR "Language, oral language and cognition: Language acquisition – new approaches" and funded by the CNRS. This workshop is designed to:

- present and provide introductions to tools used for creating and analyzing corpora of spontaneous oral (spoken and signed) language (separate tutorials will target beginners and advanced corpora users);
- present and discuss trends about tools for oral language data;
- present and discuss trends about research using oral language corpora.

Program

The organization of the tutorials (Thursday program) may change according to the number of people interested. Please send an e-mail to Christophe Parisse (cparisse@u-paris10.fr) to make a pre-registration.

Thursday November 18th 2010 – Morning session

Introductory presentations and examples of researches conducted using CLAN and PHON. Presentations will present the basic requirements when using these tools and provide examples of actual research conducted with these tools. They will also provide information to more advanced users.

9h00 – 10h30: **Stéphanie Caet** (Prismes – Sorbonne Nouvelle Paris 3)
CLAN for Beginners: discovering and learning how to use new tools for transcribing and analysing interaction data (presentation in French)

10h45 – 12h15: **Naomi Yamaguchi (**LPP – Sorbonne Nouvelle Paris 3, SFL – CNRS-Paris 8)
Phonetic transcription and phonological analyses using PHON (presentation in French)

Thursday November 18th 2010 – Afternoon session

Introductory presentations and examples of researches conducted using ELAN, and other advanced

software for corpus analysis.

14h00 – 16h00: **Simon Harrison** & **Vito Evola** (Natural Media & Engineering, HumTec Center, RWTH Aachen University)
An introduction to ELAN for the analysis and transcription of multimodal linguistic data

16h15 – 17h45: **Christophe Parisse** (MoDyCo-INSERM, CNRS-Université Paris Ouest Nanterre)
Using, processing and querying existing corpora (presentation in French)

Friday November 19th 2010 – Morning session

New trends about tools for oral language corpus annotation and processing. How is it possible to use tools for research on Language Acquisition Corpora efficiently? What do these tools offer to the researcher? Which improvements, which changes are of interest to work with oral and multimodal language corpus? Will existing tools evolve or new tools develop?

9h30-10h15: **Han Sloetjes** (Max Planck Institute for Psycholinguistics, Nijmegen)
ELAN - recent and (possible) future developments

10h15-11h00: **Michael Kipp** (Saarland University, Germany)
ANVIL: A multi-level video coding and analysis tool

11h20: 12h05: **Thomas Schmidt** (Universität Hamburg, Germany)
EXMARaLDA: system of concepts, data formats and tools for the computer assisted transcription and annotation of spoken language

---

Friday November 19th 2010 – Afternoon session

New trends in language acquisition research. What does corpus analysis bring to language acquisition research? How has it changed the relationship between theoretical and applied work?

14h-14h50: **Susanne Miyata** (Aichi Shukutoku University, Nagoya, Japan)
CHILDES: Japanese corpora and language acquisition research

14h50-15h40: **Heidi Waterfall** (Cornell University, NY, USA).
Variety is the Spice of Longitudinal Analysis

16h-17h: **Elena Lieven** (MPI-EVA, Leipzig –Germany- and University of Manchester - UK)
Corpus-based research: Recent developments, problems and possibilities

---

The workshop will be held in Paris, at the SFL Lab (UMR 7023, CNRS), CNRS Pouchet center, 59 rue Pouchet, Paris, FRANCE.

Organizing committee:
Christophe Parisse (Modyco, INSERM, CNRS/Paris Ouest Nanterre University, France)
Aliyah Morgenstern (Prismes, Sorbonne Nouvelle University, Paris, France)
Maya Hickmann (SFL, CNRS-Paris 8, Paris, France)
Funded by the CNRS, GDR « Langue, Langage Oral et Cognition: acquisition et dysfonctionnements - nouvelles approches », scientific director Maya Hickmann.

Abstracts (alphabetical)

Stéphanie Caet (Prismes – Sorbonne Nouvelle Paris 3)

CLAN for Beginners: discovering and learning how to use new tools for transcribing and analysing interaction data

This 1.30-hour workshop aims at familiarizing beginners with the interface and basic principles of the CLAN software. Three main functions of CLAN will be approached:

- alignment between video or sound and text. This not only facilitates the transcription work, it also enables researchers to keep track at all times of the contexts in which utterances were produced;
- quantitative analyses of transcribed utterances or annotations added by the researcher himself, such as frequency counts (number of utterances produced, total number of words, number of different words, number of particular forms of annotations) MLU computation, and automatic localisation of words or constructions in the text and their context;
- automatic morphosyntactic analysis of utterances, to which frequency counts and textual localisation may also be applied.

Illustrative analyses will be guided by a realistic research question and will be performed from data sets available on the CHILDES database (http://childes.psy.cmu.edu/).

French version (this talk will be presented in French)

CLAN pour les Néophytes : découvrir, comprendre et utiliser un nouvel outil de transcription et d'analyse des interactions

Cet atelier d'1h30 a pour but de familiariser des utilisateurs novices à l'interface et aux principes de base d'utilisation du logiciel CLAN. Les trois fonctions principales de CLAN seront abordées:

- l'alignement entre de la vidéo ou du son avec du texte, qui d'une part facilite le travail de transcription et d'autre part permet un va-et-vient permanent entre la transcription et le contexte de production des énoncés ;
- l'analyse quantitative des énoncés transcrits ou des annotations créées par le chercheur (calcul de fréquences : nombre d'énoncés, nombre total de mots, nombre de mots différents, fréquence de certaines formes d'annotations ; calcul de la LME ; repérage automatique de mots ou de constructions dans le texte et dans leur contexte etc.) ;
- l'analyse morphosyntaxique automatique des énoncés, à laquelle sont également applicables calculs de fréquence et repérage textuel.

Les analyses illustratives seront réalisées à partir de données disponibles sur la base de données CHILDES (http://childes.psy.cmu.edu/) auxquelles nous appliquerons une problématique de recherche potentielle qui servira de fil conducteur à ce tutoriel.

Simon Harrison & **Vito Evola** (Natural Media & Engineering, HumTec Center, RWTH Aachen University)

An introduction to ELAN for the analysis and transcription of multimodal linguistic data

In this workshop we acquaint participants with the basic skills required to collect, analyse, and transcribe multimodal linguistic data in ELAN.

Michael Kipp (Saarland University, Germany)

ANVIL: A multi-level video coding and analysis tool

This talk presents ANVIL, a widely used video coding tool that allows the encoding of linguistic data on multiple tiers according to a user-defined coding scheme. The talk's focus will introduce both basic coding functionality and more recent extensions for analyzing coding reliability and event association.

Elena Lieven (MPI-EVA, Leipzig and University of Manchester)

Corpus-based research: Recent developments, problems and possibilities

Corpus-based research has always been central to the study of child language development since it provides evidence of children's speech and their language environment in relatively naturalistic settings. As such, it forms the basis for describing the overall course of language development and for the development of hypotheses that can be tested either using data from the corpora or through experimentation. In this talk I will first briefly discuss two important methodological issues: the nature of the sampling frame and the level of analysis. I then turn to some recent studies that illustrate the various possible approaches to the analysis of corpora, including the production of novel utterances and grammar extraction by young English-speaking children; productivity measures for morphological marking in Spanish and Polish; and calculations of cue validity from the input to predict the development of the transitive in Cantonese, English and German. I return to the issue of sampling and of cross-linguistic and cross-cultural comparisons at the end of the talk.

Susanne Miyata (Aichi Shukutoku University, Nagoya, Japan)

CHILDES: Japanese corpora and language acquisition

The focus on spontaneous speech has some tradition in Japan. In the 1950's the National Institute for Japanese Language started to collect a wide variety of speech data, and Junya Noji began his monumental diary collection of the utterances of his son Sumihare. However, because of a lack of powerful linguistic computer tools, it was only in the late 1990's that these data became object of in-depth research.

My talk describes the process of CHILDES-related tool development for an agglutinative null-argument language with a mixed ideogram-phonogram script without word boundaries, and its impact on current language acquisition research.

Christophe Parisse (MoDyCo-INSERM, CNRS-Université Paris Ouest Nanterre).

Using, processing and querying existing corpora

Presentation on using existing corpora: how to process them, how to obtain statistical results, how to generate syntactic and lexical analyses (using tools such as Excel, R statistical and text processing software, textometric software). This presentation will also be focused on more advanced users, for CLAN as well as other software.

French version (this talk will be presented in French)

Utiliser, traiter et interroger des corpus existants

Cette présentation a pour but de donner des outils et des moyens de traiter les corpus de langage une fois qu'ils sont transcrits : comment les traiter, comment obtenir des résultats statistiques, comment réaliser des analyses linguistiques sur le lexique ou la syntaxe (en utilisant des outils comme par exemple Excel, le logiciel de statistiques et de traitements de corpus R, des outils textométriques). Cette présentation est particulièrement destinée aux utilisateurs avancés de CLAN ou d'autres logiciels de transcription.

Thomas Schmidt (Universität Hamburg, Germany).

EXMARaLDA: system of concepts, data formats and tools for the computer assisted transcription and annotation of spoken language

My talk will give an overview of EXMARaLDA, a system for the computer-assisted construction, management and analysis of spoken language corpora. EXMARaLDA is developed at the Research Centre on Multilingualism at the University of Hamburg with the principal aim of providing researchers with an environment in which spoken language data can be efficiently created, flexibly analysed, and reliably exchanged between people and computer tools. Presently nearing its 10th anniversary, EXMARaLDA is now widely used for doing conversation and discourse analysis, research into first and second language acquisition, dialectology and phonology.
In my talk, I will use examples from learner corpora to demonstrate the functionality of EXMARaLDA. I will devote special attention to questions of interoperability, considering both the present situation and possible future developments.

Han Sloetjes (Max Planck Institute for Psycholinguistics, Nijmegen)

ELAN - recent and (possible) future developments

This talk will spotlight a number of recent and often lesser known improvements on the annotation tool ELAN. Considerable attention will be given to current functionality in terms of querying and modifying multiple files. Extension of this kind of operations on a complete or on a section of a local corpus is planned for future releases. Furthermore developments in the realm of semi-automatic annotation and interaction with lexicon tools will be discussed

---

Heidi Waterfall (Cornell University, NY, USA).

Variety is the Spice of Longitudinal Analysis

In this presentation, I will explore the 'variety' present both in language acquisition corpora and in the methods used to analyze them.
First, I will argue that while it is necessary to calculate measures of quantity (e.g., number of tokens, number of utterances), it is also critical to examine the variety or diversity of structures present in the data (e.g., number of different kinds of complex sentences).
Second, I will demonstrate the importance of analyzing the transcript on various different linguistic levels (e.g., lexical and cross-sentential) in order to understand the input. Lastly, I will address how

using a variety of tools, specifically, the combination of manual and computational analysis can yield interesting new areas of research.

Naomi Yamaguchi (LPP – Sorbonne Nouvelle Paris 3, SFL – CNRS-Paris 8)

Phonetical transcription and phonological analysis using PHON

French version (this talk will be presented in French)

Transcription phonétique et analyse phonologique avec PHON

Cet atelier d'1h30 a pour but de familiariser des utilisateurs novices à l'interface et aux principes de base d'utilisation du logiciel PHON. Nous aborderons :

- la segmentation d'un corpus, la transcription phonétique en utilisateur simple et en double aveugle, la visualisation du signal audio, la syllabification des énoncés ;
- quelques recherches possibles concernant les productions transcrites, comme établir l'inventaire segmental pour une position syllabique donnée, rechercher des traits distinctifs, rechercher des harmonies ou des métathèses, et effectuer des recherches personnalisées.

Les données illustratives font partie de différents projets de recherche en cours, qui seront par la suite données au projet PhonBank.