



# CHILDES + Chat + Clan + Commandes

*Christophe Parisse*  
*MoDyCo-Paris Ouest Nanterre*



# Organisation du cours

- 1) Présentation de CHILDES
  - Base de corpus : CHILDES
  - Format de transcription : CHAT
  - Logiciel de transcription et utilisation des corpus : CLAN + commandes
- 2) Le format et la transcription : CHAT
- 3) Le logiciel et la transcription : CLAN (éditeur)
- 4) Le logiciel et l'exploitation des corpus : commandes CLAN

# (1) Qu'est-ce que CHILDES

- CHILDES (Child Language Data Exchange System)
  - ensemble d'éléments permettant à la communauté scientifique d'échanger des corpus de langage d'enfants
  - créé en 1984 par Brian MacWhinney et Catherine Snow
  - soutenu et hébergé par l'Université Carnegie Mellon de Pittsburgh (USA)

# Éléments

- des données pouvant être partagées par la communauté scientifique
  - 4500 membres
  - 1500 articles publiés
  - un format commun : **CHAT** (Codes for the Human Analysis of Transcripts)
  - une base de données accessible sur Internet (**<http://childes.psy.cmu.edu>**)
    - 130 corpus (développement, récits, bilingues, données cliniques)
    - 26 langues différentes
- des outils pour manipuler les données
  - un logiciel **CLAN** (Computerized Language Analysis)
    - éditeur de corpus, insertion d'audio ou de vidéo
  - des sous-**commandes** utilitaires

# Types de données pouvant être traitées

## ■ actuellement

- ☐ des transcriptions simples (textes seulement, avec ou sans phonétique)
- ☐ utilisation de la phonétique (y compris au format API), lien avec le son, analyse des fichiers sons
- ☐ liens avec la vidéo

## ■ futur (améliorations en cours)

- ☐ nouvelle base de données : TALKBANK,
- ☐ nouveaux outils : analyse phonologique (PHON)
- ☐ autres outils plus ou moins compatibles (ELAN, PRAAT, EXMARaLDA)



# Comment analyser une transcription ?

- commandes pour traiter les corpus transcrits
- commandes évoluées
  - MOR-POST (syntaxe)
  - PRAAT (son)



CHILDES is the child language component of the [TalkBank](#) system. TalkBank is a system for sharing and studying conversational interactions.

# Le site CHILDES

## System

[Ground rules](#)

[Guidelines for Contributors](#)

[Overviews and Introductions](#)

[Membership list](#)

How to subscribe to [Mailing Lists](#)

## Links

The [TalkBank](#) Database

[IASCL](#) information

[Other Child Language sites](#)

[Research based on CHILDES](#)

## Phonology and Fonts

[Phon & PhonBank](#)

Unicode and IPA for [Mac](#)

Unicode and IPA for [Windows](#)

## Special Procedures

Procedures and tools for [CA analysis](#)

Working with [digitized video](#)

Working with [digitized audio](#)

The [Computerized Comprehension Task](#)

## Programs and Database

The [Database](#)

The [CLAN](#) Program

[WebCLAN](#)

[Training Videos](#)

[Related Software](#)

## Manuals

[CHAT Transcription](#)

[CLAN Programs](#)

[Database Manuals](#)

[BTS sign transcription system](#)

## Teaching with CHILDES

[Topics](#) in language acquisition.

Teaching [Tips](#) and [Resources](#).

[Child Language Bibliographies](#)

## Morphology and Lexicon

[Part of Speech Analysis by MOR](#)

Parental frequency count: [view](#),  
[download](#), [documentation](#)

[MRC lexical dictionary](#)

Syntactic analysis by [GRASP](#)

# A récupérer sur le site CHILDES

- La dernière version de CLAN
- Les documentations en anglais de CHAT, CLAN et celle en français de Christian Champaud
- Les corpus
- Des présentations PPT de CHILDES
- Les grammaires pour l'analyse syntaxique (MOR + POST)
  
- Sur le site du GDR acquisition
  - [www.modyco.fr/gdr/](http://www.modyco.fr/gdr/)
    - **Outils de création de corpus de langage oral**
  - [www.modyco.fr/gdr/spip.php?rubrique1](http://www.modyco.fr/gdr/spip.php?rubrique1)
    - CLAN
    - Analyse morphosyntaxique avec CLAN (MOR + POST français)





The MOR program provides a method for automatic tagging of corpora in the CHAT format. To make this work, it is necessary to construct a separate MOR grammar for each language. After analysis with MOR, users can then use the POST program to disambiguate the %mor line. We provide a POST disambiguation database for English, but for other languages, users will need to do the work of training a POST database for themselves. This whole system is described in a [recent article on morphosyntactic analysis](#) in CLAN.

We have working MOR grammars for these languages:

- [Cantonese \(yue\)](#): This grammar was built by Brian MacWhinney relying on a Cantonese-English lexicon provided by K. K. Luke.
- [Chinese \(zh\)](#): This grammar was built by Brian MacWhinney and Twila Tardif. Thanks to K. J. Chen and the CKIP Group of the Academia Sinica for providing an [Excel listing](#) of the 20,000 highest frequency forms of Puotonghua along with their English translations and romanizations.
- [Dutch \(nl\)](#): This grammar was contributed by Steven Gillis.
- [English \(en\)](#): This grammar was built initially by Brian MacWhinney and Mitzi Morris. It covers all the forms in the CHILDES English database.
- [French \(fr\)](#): This grammar was contributed by Christophe Parisse.
- [German \(de\)](#): This grammar was contributed by Heike Behrens.
- [Hebrew \(he\)](#): This grammar was contributed by Sigal Uziel-Karl and Bracha Nir-Sagiv of Tel-Aviv University in Israel.
- [Japanese \(ja\)](#): This grammar was constructed by Norio Naka and Susanne Miyata. The [Wakachi](#) system is helpful for reference.
- [Italian \(it\)](#): This grammar was built by Livia Tonelli and Brian MacWhinney.
- [Spanish \(es\)](#): This grammar was built by Brian MacWhinney and Monica Sanz Torrent.

Four of these grammars (English, Chinese, Japanese, Spanish) also include POST databases created by Christophe Parisse's POSTTRAIN program. After running MOR, you run POST to automatically disambiguate the output of MOR. The Chinese version is functional, but needs a bit more training and clarification of part of speech categories to improve accuracy.

To help those interested in building their own MOR grammars, we provide these two examples of minMOR grammars. One is the [basic example](#) and the other indicates how to build a grammar that targets only a [few word forms](#), such as the German article.

[http://  
childes.psy.c  
mu.edu/  
morgrams/](http://childes.psy.cmu.edu/morgrams/)



# CHAT : format de présentation

- au départ: texte brut qui peut être édité et manipulé avec n'importe quel logiciel
- plus tout à fait vrai depuis qu'il est possible d'insérer du son ou de la vidéo, ainsi que d'utiliser des symboles API pour représenter la phonétique



# Principes de CHAT

- il y a trois types d'information stockés dans une description
  - 1) des informations à caractère général qui se rapportent à tout l'enregistrement
  - 2) des transcriptions réalisées énoncé par énoncé (ou par tours de parole)
  - 3) des indications complémentaires se rapportant à un énoncé ou à un tour de parole précis

# Premier niveau de format

- le type d'information est indiqué par le premier caractère d'une ligne
  - @ informations de type général
  - \* énoncé (transcription « principale » le plus souvent en caractères graphémiques) – **lignes principales**
  - % indications complémentaires correspondant à la ligne \* située juste au-dessus dans la transcription – **lignes dépendantes**
  - tabulation → continuation de la ligne du dessus lorsqu'elle ne peut tenir sur une seule ligne

# Exemple

@Participants: CHI Ross Enfant,

—FAT Brian Père

\*CHI: tu veux que je le ferme

—ton sac ?

%pho: ty v9 k9 Z9 l9 ferm to~

—sak

# Second niveau de format

- chacun des caractères @, \* et % est suivi d'un mot indiquant la signification de la ligne, suivi d'un « : » et d'une tabulation
- pour \* et %, le « mot » qui suit doit être un code d'exactly 3 caractères

# Exemple

**@Participants:** CHI Ross Enfant,  
FAT Brian Père

**\*CHI:** tu veux que je le ferme  
ton sac ?

**%pho:** ty v9 k9 Z9 l9 ferm to~  
sak

# Respect des conventions

- peu de marge d'erreur dans le codage (voir instrument CHECK)
- surtout faire attention aux espaces et tabulations :
  - pas d'espace avant @, \*, %
  - pas d'espace après @, \*, % et avant le mot qui suit (\*CHI:)
  - pas d'espace après le mot qui suit et avant le : (\*CHI:)
  - ne pas oublier le caractère : et la tabulation qui suit
  - ne pas oublier de mettre une tabulation pour continuer une ligne
- CLAN possède plusieurs caractéristiques qui simplifient beaucoup ces aspects !!!



# Organisation d' une transcription

@Begin

@Languages: fr

@Participants: CHI Charles Child, MOT Mère Mother

@ID: fr|exemple|CHI|3;06.00|male|group|ses|Child|mat|

@ID: fr|exemple|MOT||female|group|ses|Mother|high|

\*CHI: c'+est maman.

\*MOT: c'+est moi, ça?

\*CHI: oui.

\*CHI: xx jouer.

\*MOT: tu joues?

%com: répétition de la mère

\*CHI: moi je jouer.

@End

# Codes obligatoires

- @Begin
- @Languages: suivi d'un code langage
- @Participants: liste des participants
- @Dependent: liste des champs optionnels
- @ID: identification d'un des participants
- ...
- @End

# Codes obligatoires

## ■ @Participants:

- élément de la liste en trois parties
- code de trois lettres majuscules
- optionnel: nom de l' enfant, indication utile
- role: indication du rôle du locuteur dans la situation → limité à une liste (très large) comprenant Target\_Child, Child, Mother, ..., Investigator, ..., Camera\_Operator, ..., Unidentified, ...

## ■ @Dependent:

- pas nécessaire mais bien utile

# Codes obligatoires

@ID: language| corpus| code| age| sex|  
group| SES| role| education|

- pas d' espace avant ou après |
- les champs language, corpus, code, et role sont obligatoires
- les autres champs sont optionnels – pour les ignorer, ne rien insérer ||



# Quelques codes optionnels

en début de corpus

@Birth of XXX: 01-01-1999

@Coder: nom de la personne qui transcrit

partout dans le corpus

@Comment:

@Date:

@Location:

@Situation:

# Les codes de lignes dépendantes

- il existe de nombreux codes
- %act: action
- %com: commentaire
- %pho: ligne phonétique
- %mod: modèle pour la ligne %pho:
- %mor: ligne morphosyntaxique
- %sit: situation
- %spa: actes de langage (communication)
- %tim: temps déroulé

# Codages des transcriptions

- les transcriptions principales doivent se terminer par une ponctuation . ? ! ou un symbole spécial indiquant un énoncé inachevé +... ou interrompu +/.
- on peut aussi utiliser des marqueurs de fin caractérisant le contour intonatif
  - ? montant
  - ! exclamatif
  - . descendant
  - '. montant-descendant
  - ,. descendant-montant

# Codage des mots

- un mot se code normalement en signe alphabétique → maman
- les apostrophes doivent être suivies d' un espace → d' abord
- les '-' des mots composés sont remplacés par des '+' → pommes+de+terre
- les '-' des clitiques post-verbaux sont supprimés → donne le



# Codage des mots

- les mots incomplets peuvent être complétés → (pe)tit (en)core
- les pauses peuvent être marquées → (.)
- les reprises peuvent codées
  - <le pe> [/] le petit bébé
  - <le pe> [//] les petits bébés
  - <le pe> [///] un garçon
- le mot précédent est incertain → [?]
- le mot précédent est erroné → [\*]

# Codage des mots

- les mots peuvent complétés par des indicateurs → ding@o
- un codage phonétique optionnel est codé comme → SeSe@u
- les énoncés incompréhensibles sont codés comme xxx → \*CHI: xxx.
- les mots incompréhensibles sont codés comme xx → \*CHI: xx tombe.

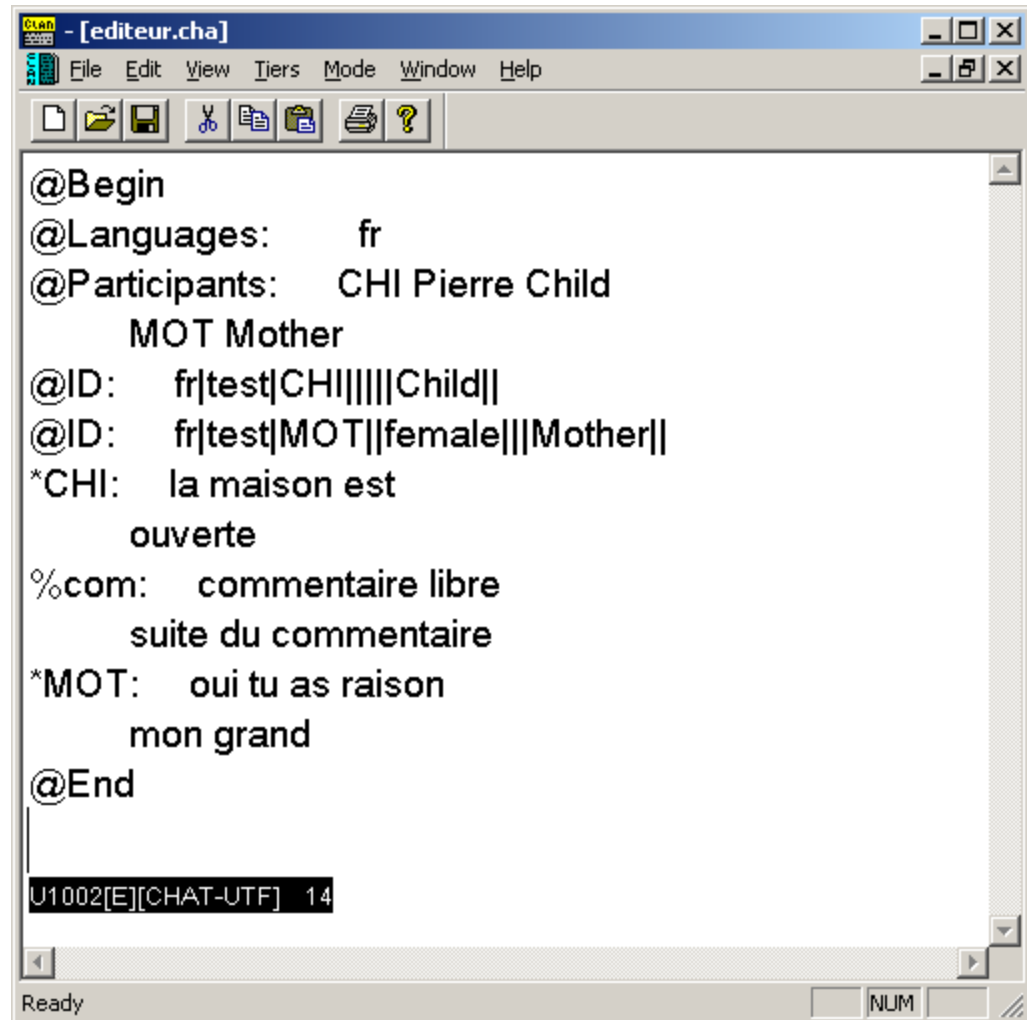
# Autres codages

- superposition d' énoncés
- continuation de l' énoncé précédent
- variété de pauses → (..) (2.3)
- acronymes → w\_c
- lettres isolées → a@l
- omission de mots
- commentaires dans une ligne → [= xy]

# Utiliser CLAN

- CLAN est un logiciel gratuit disponible sous Windows ou Macintosh
  - (<http://childes.psy.cmu.edu>)
- il permet d'éditer et de transcrire les corpus → **l'éditeur**
  - **Y compris**
    - phonétique
    - syntaxe
- il possède un ensemble de fonctions qui facilitent le travail de transcription et d'analyse des transcriptions  
→ **la fenêtre de commandes**
- il permet de travailler avec du son ou de la vidéo

# L'éditeur



The screenshot shows a window titled "[editeur.cha]" with a menu bar (File, Edit, View, Tiers, Mode, Window, Help) and a toolbar with icons for file operations. The main text area contains a chat script with the following content:

```
@Begin
@Languages:      fr
@Participants:   CHI Pierre Child
                  MOT Mother
@ID:   fr|test|CHI||||Child|
@ID:   fr|test|MOT||female||Mother|
*CHI:   la maison est
        ouverte
%com:   commentaire libre
        suite du commentaire
*MOT:   oui tu as raison
        mon grand
@End
```

At the bottom of the text area, a status bar displays "U1002[E][CHAT-UTF] 14". The window's status bar at the very bottom shows "Ready" and a "NUM" button.

# CLAN aide l'utilisateur !

@Begin

@Languages: fr

@Participants: CHI Charles Child, MOT Mère Mother

@Dependent: %pho, %sit, %act

Lancer check pour créer les @ID

Le retour chariot insère automatiquement une tabulation (continuer les énoncés)

Taper \* , % , @ ramène automatiquement au début de la ligne

CTRL+1 ramène au début de ligne et génère \*CHI:

CTRL+2 ramène au début de ligne et génère \*MOT:

remplace le participant déjà écrit

...

CTRL+7 ramène au début de ligne et insère %pho:

CTRL+8 ramène au début de ligne et insère %sit:

...

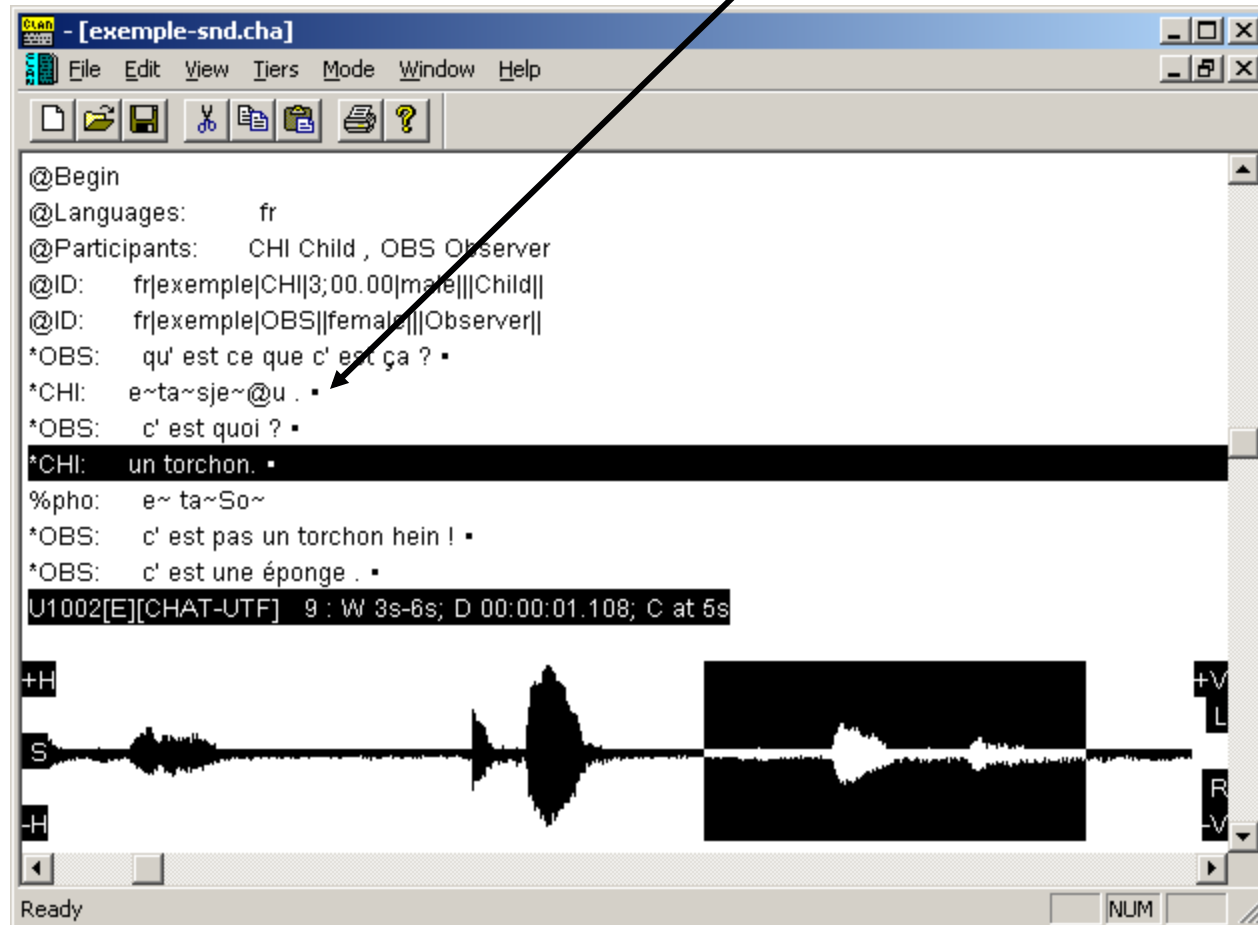


# Utilisation du son ou de la vidéo

- si la transcription reste un processus manuel fastidieux (surtout pour les longs corpus), elle peut être facilitée par CLAN avec l'intégration du son ou de la vidéo
- le but est de permettre de manière instantanée de réécouter ou de revoir l'enregistrement en cours de transcription sans avoir à changer de programme ou d'écran, sans utilisation de magnétophone et sans avoir à réaliser des retours en arrière fastidieux

# CLAN & Son

les bullets



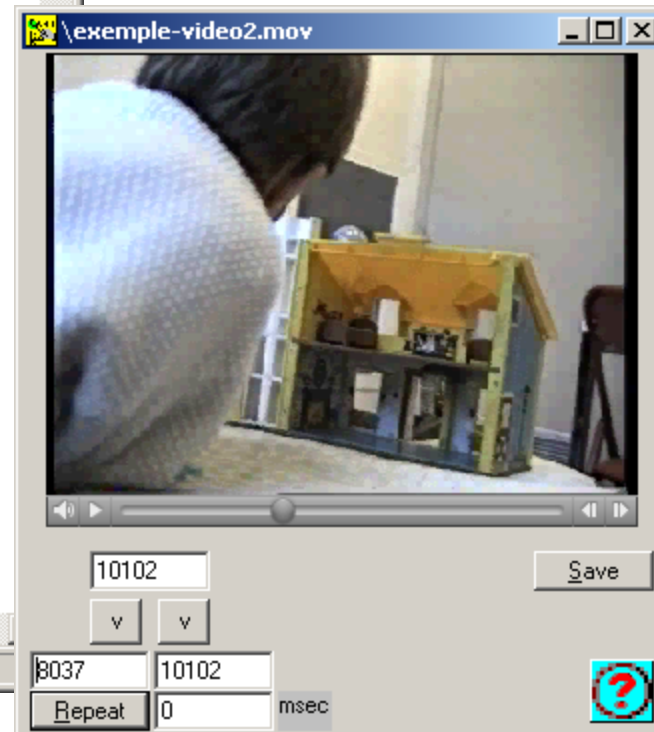
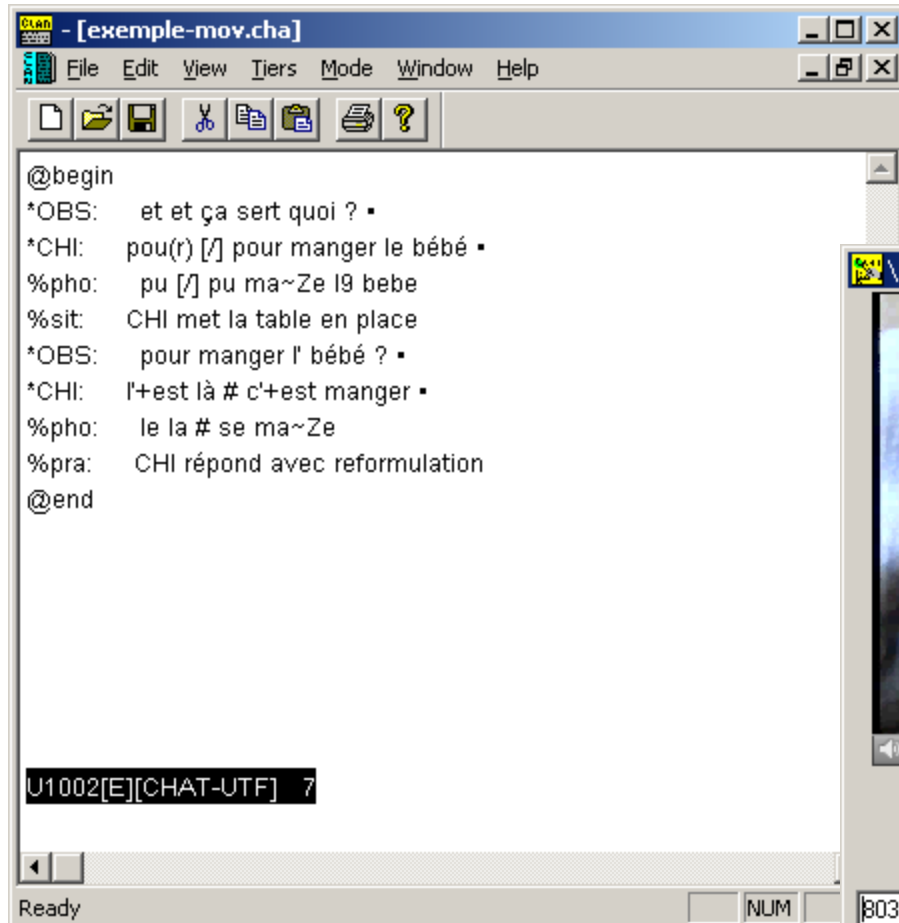
The screenshot displays the CLAN software interface for a file named [exemple-snd.cha]. The window includes a menu bar (File, Edit, View, Tiers, Mode, Window, Help) and a toolbar with icons for file operations. The main text area contains a transcript of a conversation between a child (CHI) and an observer (OBS). The transcript is as follows:

```
@Begin
@Languages:      fr
@Participants:   CHI Child , OBS Observer
@ID:   fr[exemple]CHI|3;00.00|male|||Child||
@ID:   fr[exemple]OBS||female|||Observer||
*OBS:   qu' est ce que c' est ça ? .
*CHI:   e~ta~sje~@u .
*OBS:   c' est quoi ? .
*CHI:   un torchon .
%pho:   e~ ta~So~
*OBS:   c' est pas un torchon hein ! .
*OBS:   c' est une éponge .
U1002[E][CHAT-UTF] 9 : W 3s-6s; D 00:00:01.108; C at 5s
```

Below the transcript is a waveform visualization of the audio. The waveform shows the amplitude of the speech over time. A black rectangular box is overlaid on the waveform, corresponding to the time segment 3s-6s mentioned in the transcript. The status bar at the bottom indicates 'Ready' and includes a 'NUM' button.



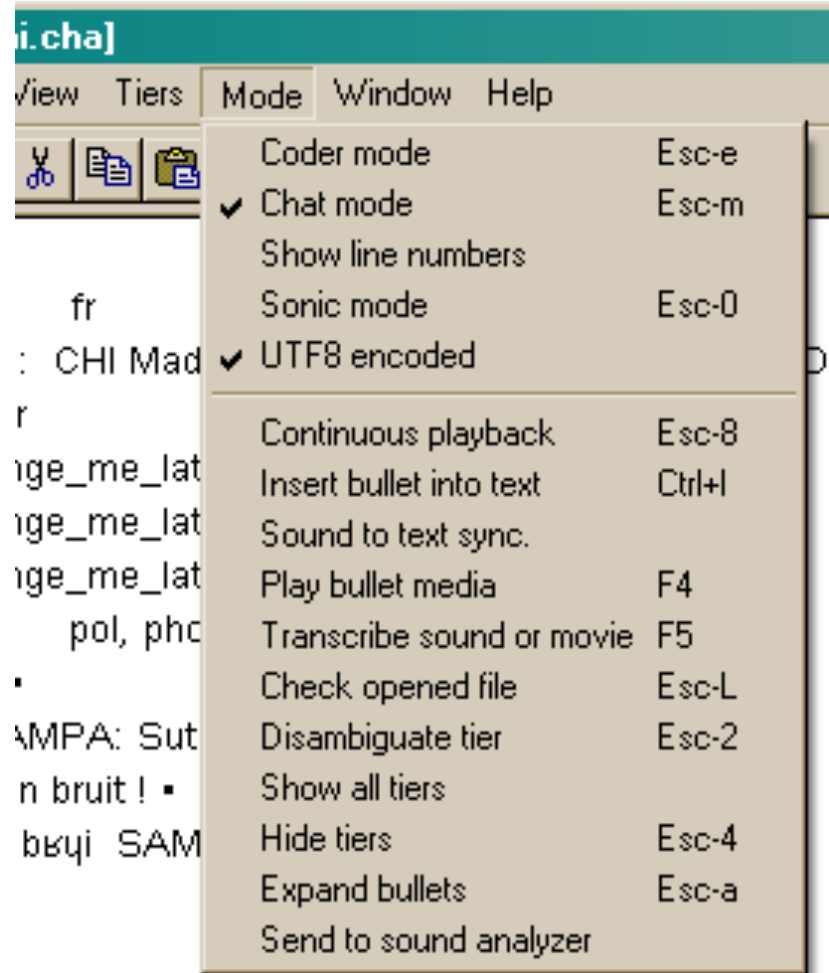
# CLAN & Vidéo



# Alignement de son ou de vidéo

- Les deux fonctions sont assez proches l'une de l'autre

- Toutes les commandes liées à l'alignement des transcriptions sont dans le menu « Mode »



# Commandes d'alignement

- Continuous playback      ESC – 8
- Insert bullet into text      Ctrl + i
- ...
- Play bullet media      F4
- Transcribe sound or movie      F5
- ...
- Expand bullets      ESC – a

# Aligner : commande F5

- Cette commande sert
  - à découper et créer une matrice pour du matériel nouveau
  - à réaligner du corpus déjà transcrit
    - que l'on est en train de transcrire
    - qui est issu d'un travail précédent



# Appuyer sur F5 en étant sur une ligne \*LOC:

- Si pas de « bullet » sur la ligne
  - CLAN demande de choisir un fichier son ou vidéo
- Lancement de la transcription
  - Chaque fois que l' on appuie sur la barre d' ESPACE, on indique à CLAN le moment où est une **fin d' énoncé**.
    - Dans l' usage par défaut il considère qu' il s' agit aussi du début de l' énoncé suivant.

\*CHI: lə@fs [/] lə@fs loup. •

%mov:"12-1\_11\_13"\_135606\_137880• →

%sit: MOT a sélectionné un/des livres, posé CHI à terre,  
CHI tient le/les livres en main.

\*CHI: lə@fs loup. •%mov:"12-1\_11\_13"\_137880\_139340• →

\*CHI: lə@fs loup. •%mov:"12-1\_11\_13"\_139340\_141508• →

%nom: sing/inanimé/cas\_sujet/tout/mnipulé/manipulable

\*OBS: t(u) aimes les loups Mad(e)leine ? •

%mov:"12-1\_11\_13"\_141508\_144338• →

\*CHI: <ouh+la>[>] ! •

%mov:"12-1\_11\_13"\_144338\_146458• →

%sit: un livre tombe.

\*CHI: <apoum@i>[<].

%sit: CHI; suivie de MOT et OBS, entre dans la salle de  
jeux et s'installe sur le canapé vert; ramasse un livre aux  
pieds du canapé.

\*CHI: xx. •%mov:"12-1\_11\_13"\_146458\_147636• →

\*CHI: pas fait moi ai pas fait il est tombé mon [/] mon livre  
il est tombé. •%mov:"12-1\_11\_13"\_147636\_154871• →

\*CHI: lə@fs [/] lə@fs loup. •

%sit: MOT a sélectionné un/des livres, posé CHI à terre,  
CHI tient le/les livres en main.

\*CHI: lə@fs loup. •

\*CHI: lə@fs loup. •

%nom: sing/inanimé/cas\_sujet/tout/mnipulé/manipulable

\*OBS: t(u) aimes les loups Mad(e)leine ? •

\*CHI: <ouh+la>[>] ! •

%sit: un livre tombe.

\*CHI: <apoum@i>[<].

%sit: CHI; suivie de MOT et OBS, entre dans la salle de  
jeux et s'installe sur le canapé vert; ramasse un livre aux  
pieds du canapé.

\*CHI: xx. •

\*CHI: pas fait moi ai pas fait il est tombé mon [/] mon livre  
il est tombé. •

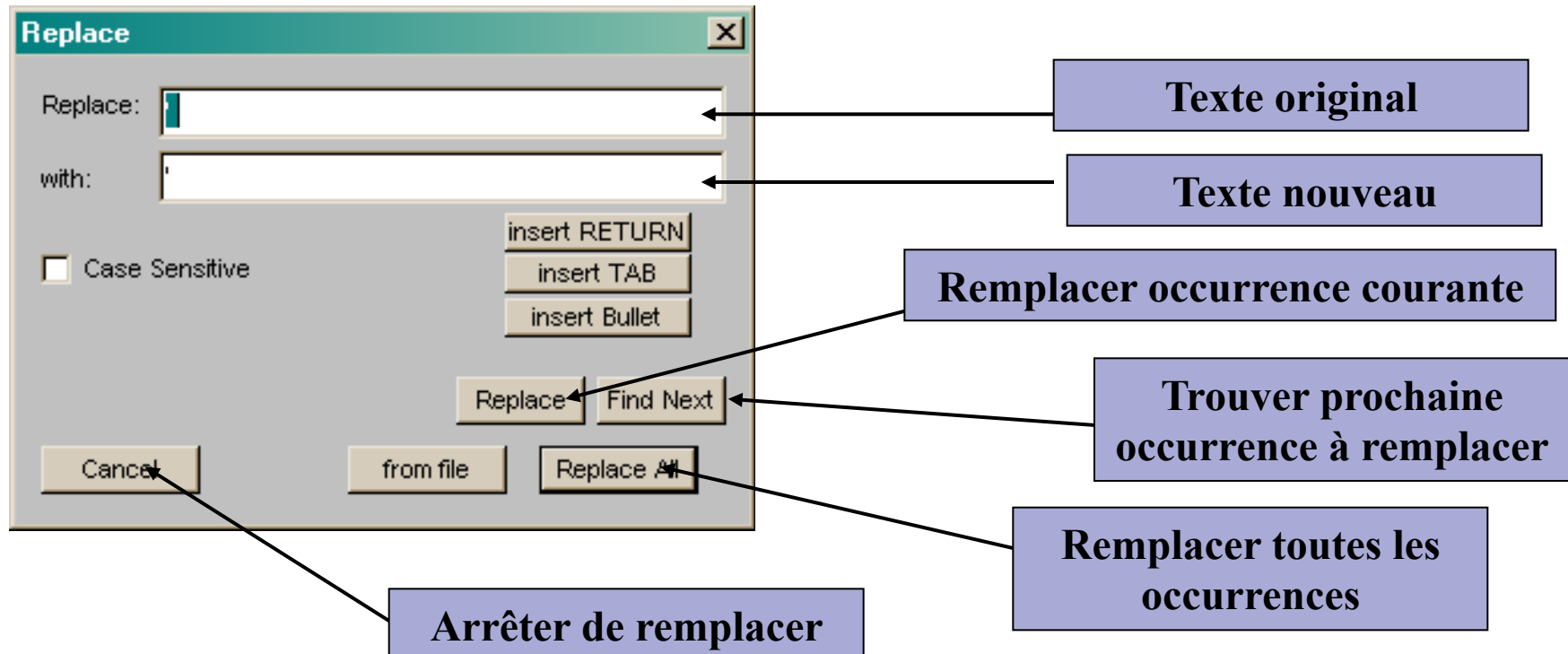
# Revoir et modifier un alignement

- Jouer une bullet:
  - F4 (à partir de la ligne du curseur)
  - CTRL + Clic Gauche
- Jouer plusieurs bullets:
  - ESC – 8 (à partir de la bullet juste au dessus du curseur).
- Refaire l'alignement
  - F5 (à partir de la ligne du curseur)
- Visualiser le codage (caractères cachés) des bullets:
  - ESC – a
- Modifier une bullet à partir de la sous-fenêtre son ou de la fenêtre vidéo
  - CTRL + i



# Remplacements dans CLAN

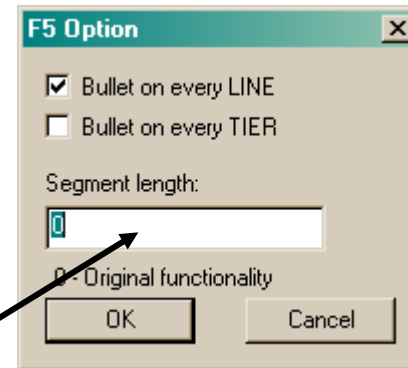
- Faire Menu : Edit → Replace... ou CTRL+R



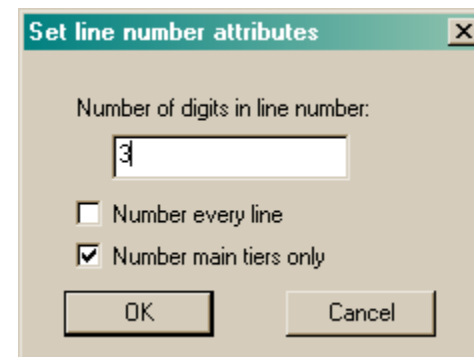
# Paramètres de CLAN (menu Edit)

## ■ Select F5 option

Pour ceux qui ne veulent transcrire que des enregistrements où les productions ne sont pas continues (durée de chaque bullet)



## ■ Define line numbers



# Fonctionnalités « cachées »

## ■ depfile.cut

- ☐ Rajouter des codes

- %etc:, @quelque chose:
- Rajouter des codes de fin de mots ciao@l2

## ■ 0hide.cut

- ☐ Voir aussi Menu Mode: « Hide tiers » et « Show all tiers »
- ☐ Cacher certains champs non-utilisés ou utilisés par d'autres ou gênants à un moment donné

## ■ Coder mode

- ☐ Pour des tiers %, entrer facilement un système de codes fixes et organisés

# La fenêtre commande

dossier de travail

dossier résultats

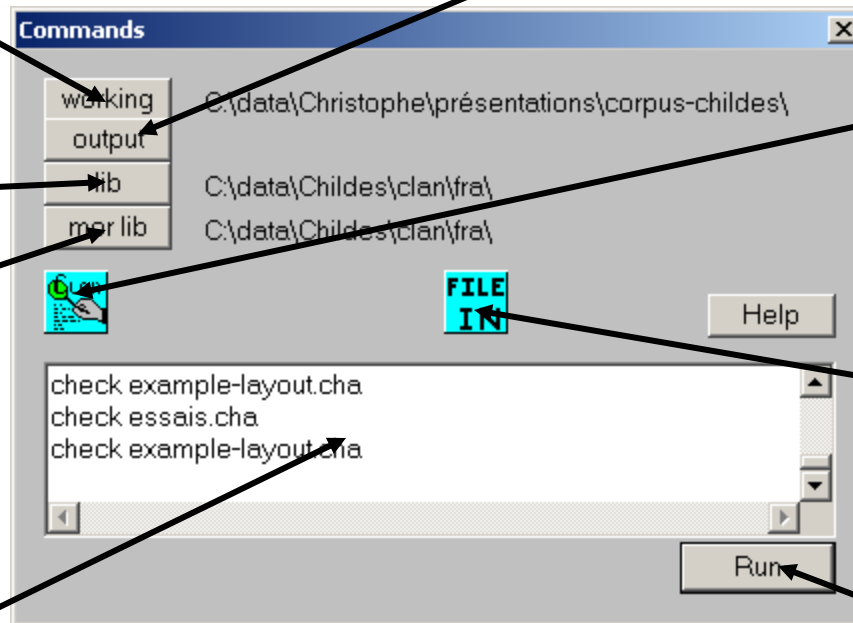
dossiers des  
fichiers  
paramètres

choix d'une  
commande

choix des  
fichiers  
à traiter

taper et éditer les commandes

lancer les  
commandes



# Liste des commandes

The screenshot shows the CLAN software interface. The main window is titled 'Clan - [CLAN Output]' and contains a list of CLAN programs and utility programs. Several programs are circled in blue: 'check', 'ship', 'combo', 'freq', 'gemlist', 'keymap', 'kwal', 'mlu', 'modrep', 'mor', 'post', 'postlist', 'postmodrules', 'posttrain', 'chstring', 'compound', and 'dir'. A 'Commands' dialog box is open on the right, showing a list of commands and a 'Run' button. The dialog box has a 'FILE IN' button and a 'Help' button. The 'Run' button is highlighted.

Clan - [CLAN Output]

File Edit View Tiers Mode Window Help

> ?

You can run any of the following CLAN programs:

chains	check	ship	combo	cooccur
dist	dss	freq	freqmerg	freqpos
gem	gemfreq	gemlist	keymap	kwal
maxwd	mlt	mlu	modrep	mor
phonfreq	post	postlist	postmodrules	posttrain
rely	statfreq	timedur	vocd	wklen

Or any of the following CLAN utility programs:

chat2ca	chat2elan	chat2praat	chat2xmar	chstring	combtier
compound	cp2utf	dataclean	dates	delim	elan2chat
fixbullets	fixit	fixmp3s	flo	indent	insert
lipp2chat	longtier	lowcase	makemod	megrasp	ort
postmortem	praat2chat	quotes	repeat	retrace	rtfin
saltin	textin	tierorder	trnfix	uniq	utf2cp

Supplementary commands available:

batch	cd	copy	del	dir
info	list	ren(ame)	rm	ty(pe)

>

020109[E][TEXT] 24

Ready

Commands

working C:\Colaje-Leonard\codage semantique\

output

lib C:\CHILDES\CLAN\fra\

mor lib C:\CHILDES\CLAN\fra\

FILE IN

Help

combo +y +d +s"%pol\*" \*.recup.cex

combo +y +d +s"%pol\*" \*.recup.cex > aa

?

Run

# Le programme CHECK

- vérifie que le format utilisé est exactement celui de la norme officielle
- peut poser des problèmes pour les personnes voulant modifier la norme CHILDES pour l'adapter à leur travail
- est nécessaire pour utiliser certaines commandes
- utilisation interne (depuis l'éditeur): ESC – |
- utilisation externe grâce à la fenêtre de commande

# Exemple

**check exemple-layout.cha**

**depfile.cut being used from: "C:\data\Childes\clan\fra\"**

**check exemple-layout.cha**

**Fri Nov 11 19:59:42 2005**

**check (10-Feb-2005) is conducting analyses on: ALL speaker tiers  
and those speakers' ALL dependent tiers and ALL header tiers**

**\*\*\*\*\***

**From file <exemple-layout.cha>**

**First pass DONE.**

**Second pass DONE.**

**Success!**

**No errors found.**



# Le programme MLU

- calcule le nombre d' énoncés, le nombre de mots, et la longueur moyenne d' énoncés, ainsi que l' écart-type de cette moyenne



# Exemple

**mlu exemple-layout.cha**

**Fri Nov 11 19:52:12 2005**

**mlu (10-Feb-2005) is conducting analyses on:**

**ALL speaker tiers**

**\*\*\*\*\***

**From file <exemple-layout.cha>**

**MLU for Speaker: \*MOT:**

**MLU (xxx and yyy are EXCLUDED from the utterance and morpheme counts):**

**Number of: utterances = 2, morphemes = 6**

**Ratio of morphemes over utterances = 3.000**

**Standard deviation = 1.000**

**MLU for Speaker: \*CHI:**

**MLU (xxx and yyy are EXCLUDED from the utterance and morpheme counts):**

**Number of: utterances = 4, morphemes = 9**

**Ratio of morphemes over utterances = 2.250**

**Standard deviation = 0.829**



# Le programme FREQ

- génère la liste des mots utilisés dans un corpus
- calcule le nombre d'occurrences pour chaque mot
- calcule le ratio types/token (TTR)

# Exemple

freq exemple-layout.cha

Fri Nov 11 19:52:19 2005

freq (10-Feb-2005) is conducting  
analyses on:

ALL speaker tiers

\*\*\*\*\*

From file <exemple-layout.cha>

2 c'

2 est

1 je

2 jouer

1 joues

1 maman

2 moi

1 oui

1 tu

1 xx

1 ça

-----  
11 Total number of different  
word types used

15 Total number of words  
(tokens)

0.733 Type/Token ratio



# Le programme COMBO

- recherche un ou plusieurs mots utilisés dans un ou des corpus
- possibilité de chercher un mot seul
- de chercher un couple de mots (ou même plus)
- de chercher dans une des lignes dépendantes

# Exemple

combo +smoi exemple-layout.cha

Fri Nov 11 20:23:04 2005

combo (10-Feb-2005) is conducting analyses on:

ALL speaker tiers

\*\*\*\*\*

From file <exemple-layout.cha>

-----

\*\*\* File "exemple-layout.cha": line 7.

\*MOT: c' est (1)moi , ça ?

-----

\*\*\* File "exemple-layout.cha": line 11.

\*CHI: (1)moi je jouer .

Strings matched 2 times

# Exemple 2

((moi^je))

combo +s"moi^je" exemple-layout.cha

Fri Nov 11 20:26:07 2005

combo (10-Feb-2005) is conducting analyses on:

ALL speaker tiers

\*\*\*\*\*

From file <exemple-layout.cha>

-----

\*\*\* File "exemple-layout.cha": line 11.

\*CHI: (1)moi (1)je jouer .

Strings matched 1 times

# Transcription phonétique

- utilisation d'un alphabet phonétique SAMPA (UNIBET) qui utilise les caractères alphabétiques standards → **SERSe** (chercher)
  - pas de problèmes de clavier mais apprentissage nécessaire
- utilisation des caractères IPA (International Phonetic Alphabet) grâce la police UNICODE Arial MS → **ʃɛRʃe** (chercher)
  - problèmes d'utilisation du clavier
    - Utilitaire sur Windows (voir <http://latlcui.unige.ch/phonetique/>)
    - Utilitaire sur Mac (voir )



# Le programme PHONFREQ

- calcule le répertoire phonétique à partir de la ligne %pho:
- fournit des statistiques pour le début des mots, le milieu des mots, et la fin des mots



# Exemple

phonfreq +b%pho exemple-mod.cha

Fri Nov 11 20:14:50 2005

phonfreq (10-Feb-2005) is conducting analyses on:

ALL speaker tiers

and those speakers' ONLY dependent tiers matching: %PHO;;

\*\*\*\*\*

From file <exemple-mod.cha>

1 # initial = 1, final = 0, other = 0  
2 . initial = 2, final = 0, other = 0  
1 9 initial = 0, final = 1, other = 0  
2 Z initial = 0, final = 0, other = 2  
1 a initial = 0, final = 1, other = 0  
2 a~ initial = 0, final = 0, other = 2  
2 b initial = 1, final = 0, other = 1  
6 e initial = 0, final = 5, other = 1  
3 l initial = 3, final = 0, other = 0  
2 m initial = 2, final = 0, other = 0  
2 p initial = 2, final = 0, other = 0  
1 s initial = 1, final = 0, other = 0  
2 u initial = 0, final = 2, other = 0

# Modèle phonétique

\*CHI: pou(r) [/] pour manger le bébé .

%pho: pu [/] pu ma~Ze l9 bebe .

%mod: puR [/] puR ma~Ze l9 bebe .

\*CHI: l+est là # c+est manger .

%pho: le la # se ma~Ze .

%mod: lE la # sE ma~Ze .



# Le programme MODREP

- calcule les différences entre la phonétique produite (ligne %pho:) et le modèle phonétique de l'adulte (ligne %mod:)

# Exemple

modrep +b%MOD +c%PHO +k exemple-mod.cha

Fri Nov 11 20:10:10 2005

modrep (10-Feb-2005) is conducting analyses on:

ALL speaker tiers

and those speakers' ONLY dependent tiers matching: %MOD; %PHO;

\*\*\*\*\*

From file <exemple-mod.cha>

1 #

1 #

2 .

2 .

1 bebe

1 bebe

1 I9

1 I9

1 IE

1 le

1 la

1 la

2 ma~Ze

2 ma~Ze

1 puR

1 pu

1 sE

1 se

# La syntaxe avec MOR et POST

- MOR est un dictionnaire listant toutes les formes morphosyntaxiques possibles pour un mot
- commande:
  - `mor syntaxe.cha`

# Exemple

\*LOC: tu veux faire à manger ?

%mor: v:pp|taire&\_MASC&\_SING^pro:subj|tu&2S v:mdllex|  
vouloir&IMP&2SV^v:mdl|vouloir&IMP&2SV^v:mdllex|  
vouloir&PRES&2SV^v:mdl|vouloir&PRES&2SV^v:mdllex|  
vouloir&PRES&1SV^v:mdl|vouloir&PRES&1SV v:mdllex|  
faire&INF^v:mdl|faire&INF adv|à^prep:art|à^prep|à v:inf|manger^n|  
manger&\_MASC ?

\*LOC: bah oui xx les autres ils dorment encore .

%mor: co|bah v|ouïr&IMP&2SV^v|ouïr&PRES&3SV^n|oui&\_MASC^adv:yn|  
oui undef|xx det|les&PL^pro:obj|les&PL det:gen|autres&\_PL^pro|  
autres&\_PL^adj|autre&\_PL pro:subj|ils&MASC&\_3P v|dormir&SUBJV  
+PRES&3PV^v|dormir&PRES&3PV adv|encore .



# Génération du dictionnaire

- Utilisation de MOR +xl pour rechercher tous les mots inconnus
- Correction éventuelles à appliquer
- Édition d' un dictionnaire complémentaire
  - Format des entrées du dictionnaire
  - Emplacement des fichiers dictionnaire
- Utilisation de MOR +xi

# La syntaxe avec MOR et POST

- POST est un analyseur choisissant automatiquement la forme la plus plausible parmi les formes morphosyntaxiques proposées par MOR
- commande:
  - `post +dfra.db +ttags.cut syntaxe.mor.cex`





# Analyse morphosyntaxique

## ■ Commande POST

- ☐ Paramètres de base
- ☐ Nettoyage des catégories non-désirées
- ☐ Visualisation des ambiguïtés
- ☐ Corrections éventuelles :
  - CHSTRING
  - Sur-apprentissage

# Exemple

\*LOC: tu veux faire à manger ?

%mor: pro:subj|tu v:mdl|vouloir v:mdllex|faire&INF prep:art|à n|manger

\*LOC: bah oui xx les autres ils dorment encore .

%mor: co|bah adv:yn|oui undef|xx det|les&PL pro|autres pro:subj|  
ils&MASC v|dormir adv|encore .

# Utilisation de Praat (étude fine des sons)

- logiciel annexe (Institut sciences phonétiques à Amsterdam) qui fonctionne en parallèle de CLAN
- lorsque des données sonores sont insérées dans CLAN, ces données (le signal sonore) peuvent être transférées immédiatement à Praat
- Praat permet de faire des études fines des sons
  - F0, énergie, spectrogramme, fréquences fondamentales