Conventions de transcription au format CHAT pour le programme CLAN[1]

Ces conventions sont compatibles avec les conventions standards du système CHILDES. Toutes les extensions créées pour le projet Transferts sont compatibles avec les *guidelines* de CHAT (MacWhinney, 2000).

Il est recommandé à tous les utilisateurs de mettre régulièrement à jour leur version du logiciel CLAN en téléchargeant le programme disponible sur la page Internet http://childes.psy.cmu.edu/clan/. Il faut au moins faire cela une fois avant de commencer à utiliser CLAN pour le projet. (Ne pas utiliser une version antérieure à janvier 2012).

INTRODUCTION

CHILDES (*Child Language Data Exchange System,* voir http://childes.psy.cmu.edu) est un système d'échange et de description du langage de l'enfant mis sur pied en 1984 par Brian MacWhinney et Catherine Snow (Université Carnegie Mellon de Pittsburgh, USA).

CHILDES est constitué de trois éléments:

1 - CHAT, un format de transcription et de codage qui permet d'informatiser le corpus.
2 - Une banque de données.
3 - CLAN, une série de programmes informatiques pour traiter et analyser les données : mots, grammaire, erreurs, contextes, prosodie, accentuation, pauses,...

CHILDES permet donc notamment la mise à disposition de transcriptions sur l'Internet et le partage de fichiers son et vidéo et compte actuellement 130 corpus (acquisition L1 et L2, récits, bilingues, données cliniques) dans 26 langues différentes.
Plus d'une centaine de groupes de chercheurs répandus dans le monde entier travaillent avec ces outils. Ils transcrivent des entretiens dans le format CHAT ce qui enrichit la banque de données, et les analysent grâce aux différents programmes de CLAN.

Principe de transcription :

Le programme **CLAN** (*Computerized Language Analysis*), disponible sur http://childes.psy.cmu.edu/clan/ permet de réaliser des transcriptions de langage spontané dans le format CHAT (*Codes for the Human Analysis of Transcripts*) alignées sur des vidéos ou du son. Il s'agit de transcrire des corpus de langage spontané de la façon la plus complète possible.

[1] Document réalisé par Christophe Parisse et Colette Noyau. Des éléments de ce document sont repris du guide de transcription du projet ANR COLAJE, rédigé par Françoise Bourdoux et Stéphanie Caët.

Format de base de CHAT :

Trois types d'information sont contenus dans une transcription :
- des informations à caractère général qui se rapportent à tout l'enregistrement, ce sont principalement les en-têtes (lignes commençant par @) ;
- des transcriptions réalisées énoncé par énoncé (ou par tours de parole) sur les lignes principales (lignes commençant par *) ;
- des indications complémentaires se rapportant à un énoncé ou à un tour de parole précis, ce sont les lignes dépendantes (lignes commençant par %).

Marche à suivre

1) Compléter les en-têtes
2) Découpage en énoncés
3) Transcrire les lignes principales
4) Ajouter des lignes dépendantes
5) Vérifier sa transcription (y compris en utilisant le programme CHECK).

Les en-têtes

Les lignes d'en-tête commencent par @.
Les en-têtes suivantes sont obligatoires et toujours dans l'ordre suivant :

@Begin
@Languages: [tabulation] fr pour français, ful pour fulfulde, zar pour zarma, jul pour jula, hau pour hausa, bam pour bambara et son pour songhay
@Participants: [tab.] MTR Teacher, ELV Student, ELV2 Student, ELVS Student
(ELVS permet de coder plusieurs élèves parlant en même temps)
@ID: [tab.*] (reprend des informations sur chaque participant, y compris les langues utilisées (une ligne par participant)[2])*
@ID: fra, ful|transferts|MTR|||||Teacher|| (par exemple)
@ID: fra, ful|transferts|ELV1|||||Student|| (par exemple)
@ID: fra, ful|transferts|ELV2|||||Student|| (par exemple)
@Media: nom-du-fichier-media-sans-extension, video *(ou audio)*
@Date: [tab.] (= date d'enregistrement)
@Location: [tab] (= lieu d'enregistrement)
@Time duration: [tab.] (= durée de l'enregistrement en minutes :secondes

[2] Il est possible de ne considérer qu'un seul participant générique ELV recouvrant n'importe quel élève prenant la parole dans le cas notamment où des élèves mal identifiés s'expriment. CLAN peut accepter ce format, mais les analyses ne pourront faire la différence entre les enfants. Soulignons que l'utilisation des champs @ID et du rôle « Student » permet de faire des traitements sur l'ensemble des participants ayant le même rôle. Il est donc techniquement possible de séparer les élèves lors de la transcription et de les regrouper plus tard lors du travail scientifique.

@Transcriber: [tab.] + date transcription
Reviser + date révision de la transcription
@Situation: [tab.]
… *(transcription)*
@End *(dernière ligne)*

Les lignes @G : permettent de scinder les séquences enregistrées en sous-séquences. Toutes les sous-séquences peuvent être décrites par des mots clés ou du texte libre, ce qui permet de récupérer les parties intéressantes automatiquement à l'aide d'une commande de recherche des sous-séquences.
Une sous-séquence commence par une ligne @G : et se termine à la ligne @G : suivante ou à la fin du fichier.

NB : Pour stabiliser le vocabulaire de notre domaine : une **séquence de classe** est un moment d'activité continu et poursuivant un objectif, à l'intérieur de l'enseignement d'une matière ou d'un domaine (dure généralement une vingtaine de minutes) Ainsi, une **séquence** de calcul consacrée à la multiplication à 2 chiffres peut contenir la révision d'un type déjà maîtrisé de multiplication, par ex. par 10 + l'explication de la réalisation de la multiplication par 11 + des exercices d'entraînement sur ardoise par groupes + la restitution et l'élucidation des difficultés au tableau + l'annonce que c'est fini, qu'on passe à une autre activité. *'séance'* est un terme profane non défini à ne pas utiliser, ex. la séance du matin = ??? une séance chez le dentiste ?? Une **sous-séquence** est une activité définie à l'intérieur d'une séquence, par ex. : Révision / Explication de la nouvelle modalité de multiplication / Exercices en groupe / Restitution et élucidation au tableau sont 4 sous-séquences d'une séquence de calcul consacrée à la multiplication par 11. Un fichier de transcription correspond normalement à une séquence de classe (activités groupées et en continu concourant à un objectif d'un domaine ou une matière, durée moyenne 20 mn.). D'où l'utilité de recourir à la notation par @G des sous-séquences. @G peut aussi concerner des événements plus ponctuels, intervenant dans une séquence ou sous-séquence, qui serviront de jalon lorsqu'on effectue d'abord une **description** rapide de l'enregistrement, avant de passer à la **transcription** proprement dite.

ATTENTION : FORMAT A RESPECTER
- Pas d'espace entre « @ » et « en-tête » ni entre « en-tête » et « : » et mettre une tabulation après le « : ».

Exemple de transcription complète :

@Begin
@Languages: fra, jul
@Participants: MTR Teacher, ELV Child, ELV2 Child
@ID: fra, ful|change_me_later|MTR|||||Teacher||
@ID: fra, ful|change_me_later|ELV|||||Child||
@ID: fra, ful|change_me_later|ELV2|||||Child||
@Dependent: com, gls, act, pho, sit, fra
@Media: bjul_a3_obse_l2, video
@Date: 10-NOV-2011
@Transcriber: Guiré Inoussa 17-02-2012
@Situation: nous sommes à Bobo Dioulasso dans une classe de 3ème année

en train de suivre un cours d'observation L2.
@G: début du cours
*ELV: pour que le bon couscous soit prêt +...
*ELV: femme debu@u[3] [: debout] et du courage .
*ELV: pilons panpan@i .
*ELV: pilons panpan@ .
*ELV: pilons gaiement .
*ELV: pilons gaiement .
*MTR: asseyez-vous .
*ELV: 0
@G: rappel du thème du cours
*MTR: quelle est la dernière leçon en observation ?
*MTR: Alima .
*ELV: le soleil .
*MTR:observation le soleil .
@End

Les lignes principales

La ligne principale est symbolisée par « * ». On y code les énoncés en format orthographique (sauf rares exceptions).

FORMAT A RESPECTER : REGLES DE BASE DE TRANSCRIPTION

- **Un seul énoncé par ligne**.
- Si la transcription ne peut tenir sur une ligne, commencer la ligne suivante avec une [TAB] (le logiciel CLAN crée automatiquement des tabulations en début de chaque nouvelle ligne, mais il faut faire attention à ne pas les supprimer ou les déplacer).
- Chaque ligne principale commence par *NOM: [TAB]
  Le nom (ou rôle) du participant est un code de 3 à 7 caractères majuscules (ELV, MTR)
- Pas d'espace entre * et nom du participant
- Pas d'espace entre nom et « : »
- Tabulation après les « : »

    *ELV: le soleil se lève à l'Est .

- Un énoncé se termine toujours par un marqueur de fin d'énoncé (voir tableau ci-dessous)
- L'espace avant le marqueur de fin d'énoncé est conseillé.
- Espace obligatoire après le marqueur de début d'énoncé.
- Espace obligatoire entre marqueur de fin d'énoncé et balise.
- La balise figure toujours sur la ligne principale.
- Les mots sont codés orthographiquement sur la ligne principale.
- **Aucune ponctuation au sein de l'énoncé (sauf marqueur de fin d'énoncé)** (ne jamais utiliser de « , » virgule ou « ; » point virgule ou « . » point à l'intérieur d'un énoncé). Un ensemble de marqueurs de fin d'énoncé permettent de caractériser l'intonation globale de l'énoncé.

[3] Le codage « debu@u » est un code phonétique API correspondant à /debu/ (voir ci-dessous).

- Il ne faut pas utiliser de gras ou d'italique dans la transcription.
- Les marques de silence et d'intonation peuvent être utilisées pour rendre compte du discours oral.
- Les crochets [] et chevrons < > sont utilisés pour du métacodage (cf ci – après).
- **Majuscules seulement aux noms propres.**
- Il faut absolument transcrire les chiffres et nombres en toutes lettres et ne pas utiliser d'abréviations.

*MTR: qu'est+ce+qu' on voit à l' image numéro un ?

Mots composés

On les écrit avec un « + » (par exemple : pomme+de+terre).

*ELV: c'est une pomme+de+terre .

Dans le cas des apostrophes, si deux mots sont collés, ils sont considérés comme appartenant au même mot (comme « aujourd'hui »). Il faut utiliser un espace pour indiquer qu'il s'agit de deux mots.

*MTR: voici l' image présentée aujourd'hui .
« aujourd'hui » forme un seul mot, « l' » et « image » forment deux mots.

Les tirets sont réservés pour marquer la morphologie des mots (par exemple travaill-er) mais leur usage n'est pas nécessaire dans le projet.

Les sigles s'écrivent en séparant les lettres par des _ (*underscores*, = touche du 8 sous Windows clavier français). Il faut toutefois utiliser des minuscules pour différencier les sigles des noms propres.

*MTR: l' a_u_f soutient le projet .

Les noms propres composés s'écrivent comme les sigles mais en utilisant les majuscules.

*CHI: c'est le Père_Noël .

Codage des tons et codes API

En cas de besoin, l'écriture des tons et des codes phonétiques de l'API doit être réalisée à l'aide du clavier AFU (disponible sur http://llacan.vjf.cnrs.fr/sec_comput.htm) pour les tons et du clavier AFU ou FreeKey SIL IPA keyboard (disponible sur http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=UniIPAKeyboard) pour les codes phonétiques. Ceci permet, grâce à l'usage des polices UNICODE, d'assurer une complète compatibilité avec tous les logiciels de traitement de corpus et d'édition de textes.

Exemples : codes orthographiques/API ɛ ɔ -- tons ɛ̌ ɔ̀ -- codes API ɛ̃ ɔ̃

LES CODAGES A UTILISER DANS LA LIGNE PRINCIPALE

Absence de codage de la ligne.

Une ligne de corpus peut ne pas contenir de transcription, soit parce qu'elle décrit des faits ou des gestes et pas du langage, soit parce que l'on a décidé d'ignorer cette ligne.
Ce codage peut se représenter par « 0 » qui signifie : pas de transcription ou par « www » qui signifie transcription ignorée.

*MTR: 0 .
*ELV: www .

Marqueurs de fin d'énoncé

MARQUEUR DE FIN D'ENONCE	SIGNIFICATION
.	énoncé affirmatif
?	énoncé interrogatif
!	énoncé exclamatif
+…	Énoncé en suspend
+..?	Question en suspend
+//.	Auto-interruption par le locuteur qui s'interrompt spontanément et commence immédiatement à parler d'autre chose. Il n'y a pas de pause dans la conversation.
+//?	Le locuteur auto-interrompt son interrogation.
+/.	Interruption par quelqu'un d'autre
+/?	Le locuteur est interrompu par quelqu'un d'autre dans son interrogation.
+"/.	Une citation ou du discours direct suit
+"	Une citation ou du discours direct précède

Marqueurs de début d'énoncé

MARQUEUR DE DEBUT D'ENONCE	SIGNIFICATION
+,	énoncé complétant un énoncé précédent produit par le même locuteur.
++	énoncé complétant un énoncé précédent produit par un autre interlocuteur.
+^	le locuteur embraye tout de suite sans que l'autre ait le temps de prendre la parole (le schéma conversationnel n'est pas respecté).
+"	discours direct, citation

Exemples : Interruption et reprise

*ELV: lorsqu'il se couche +...
*MTR: alors ?
*ELV: +, il fait nuit .

*MTR: tu vois des +..?

*ELV: ++ zɛʀb [: herbes] .

Exemples : Discours direct

*ELV: il dit +"/.
*ELV: +" je viendrai demain .

*ELV: tout s'est bien fini +"
*ELV: +" il m'a dit

Citation, discours direct

Les citations et discours directs peuvent être indiqués aussi directement dans l'énoncé en étant suivis de ["].

*MTR: rabougries ["] ça veut dire quoi ?

Mot(s) ou énoncés inintelligibles

- yy : Production d'un mot que l'on ne peut pas écrire orthographiquement sur la ligne principale mais qui sera ou pourrait être transcrit sur la ligne %pho (cf ci- après). Quand il y a plusieurs "mots" (identifiés du point de vue sémantique) indécodables, on note autant de yy que de mots.
- yyy : énoncé entier (ou un long bout d'énoncé) impossible à transcrire en orthographe.
- xx : mot inintelligible, non transcriptible ni orthographiquement ni phonologiquement (par exemple masqué par un bruit).
- xxx : production inintelligible, sur tout l'énoncé (pas de ligne %pho dans ce cas là).
- www : tout un énoncé est volontairement non-transcrit (par exemple une intervention externe ou une interruption externe n'ayant rien à voir avec la situation enregistrée).
- ww : un mot est volontairement non transcrit.

FORMAT A RESPECTER

- xx / xxx / yy / yyy / www sont toujours écrits en lettres minuscules
- Comme pour tout énoncé, ne pas oublier le marqueur de fin d'énoncé après yyy ou xxx.

Elisions

Les élisions réalisées à l'oral et les parties de mots non prononcées en cas d'ébauches phonologiques sont marquées entre parenthèses.

*ELV: i(l) voit le p(e)tit chat ?

*ELV: les sais(ons) .
*MTR: saisons !

Répétition et reprise d'un ou plusieurs mot(s) au sein d'un énoncé

[/] pour répétition, reprise sans correction de ce qui est dit avant,

*ELV: il [/] il fait nuit +...

[//] pour reprise avec correction (correction syntaxique)

*MTR: <des herbes> [//] les herbes .

[///] pour reprise avec reformulation (auto-correction sémantique)

*MTR: vous allez compléter par <le mot qui manque> [///] la réponse qui manque .

Si la répétition porte sur plusieurs mots, mettre ceux – ci entre < > (les chevrons ne sont pas nécessaires si la répétition ne porte que sur un mot, par défaut le mot juste avant ce symbole.)

FORMAT A RESPECTER

- [/] écrit sans espace mais espace avant et après
- pas d'espace entre les < > et le groupe de mots qui y figure (par exemple : <le chocolat>)

Evénements para-linguistiques

L'indication figure entre crochets (voir format à respecter)

exemples : [=! rit], [=! sourit], [=! crie], [=! pleure], [=! chuchote], [=! chante], [=! tousse], [=! petits bruits], [=! bruits], [= ! lit], [= ! discours rapporté], …

Cette indication porte sur le mot juste avant. Si elle porte sur plusieurs mot ou l'énoncé en entier, mettre le groupe de mots entre < > .

L'indication peut également figurer seule, précédée de 0 (absence d'énoncé).

*ELV: l'agriculture [=! plusieurs élèves répètent] .

*ELV: 0 [=! silence].

Complément de description

Lorsqu'un évènement nécessite un éclaircissement ou une explication, on le note entre crochets en utilisant le symbole [= texte]. Ce codage peut aussi être réalisé en utilisant le champ %com. Il doit être réservé aux explications très courtes, sinon l'utilisation de %com est plus claire (voir plus loin).

*MTR: quelle est la dernière leçon [= regarde le tableau] en observation ?

Remplacement

Lorsqu'une forme utilisée n'est pas standard et peut être difficile à comprendre pour des personnes extérieures à l'école ou la famille, il est possible de proposer un remplacement par une forme traditionnelle. Utiliser [: texte] après l'élément à remplacer.

*ELV: j'a [: j'ai] une grande soeur .

Transcription alternative

Lorsque la transcription n'est pas certaine (par exemple le son était difficile à entendre et un doute subsiste) il est possible de proposer une transcription alternative, c'est-à-dire de faire une seconde proposition de transcription. Utiliser [=? texte] après l'indication à clarifier.

*MTR: qu'est+ce [=? est+ce] qu'on voit encore ?

Transcription incertaine

Lorsqu'une transcription est difficile et que le texte transcrit est incertain, on peut le spécifier en l'indiquant par un [?] en fin de transcription.

*ELV: il fait nuit [?] .

Pauses

Les pauses et hésitations sans reprise ni retour en arrière peuvent être notées directement dans l'énoncé principal. Plusieurs marqueurs existent à utiliser en fonction de la longueur de la pause.

Pause courte : (.)
Pause moyenne : (..)
Pause longue : (…)
Pause avec indication de durée : (18.5) – 18,5 secondes

*ELV: le jour (.) il éclaire et réchauffe la terre .
*ELV: moi je [/] je vois des (..) dessins .

Chevauchements

Les chevauchements peuvent être indiqués de manière complexe (voir le guide officiel de CLAN). Toutefois pour le projet Transferts, nous n'avons besoin que du marqueur dit sommaire « +< » qui doit être utilisé au début de l'énoncé qui recouvre le précédent.

*ELV : moi !
*ELV2 : +< moi !

Codage du bilinguisme

Deux principes de codage peuvent être utilisés : le codage implicite et le codage explicite. Dans le codage implicite, on définit les langues utilisées dans l'en-tête du corpus avec en première position la langue « dominante » (L1), et ensuite la seconde langue (L2).

@Languages : fra, ful
(où le français est L1 et le fulfulde est L2).
@Languages : zar, fra
(où le zarma est L1 et le français est L2).

La notion de L1 et de L2 ici correspond à la nature de l'enregistrement, pas au parler des locuteurs. Un cours donné en français sera français L1, un cours donné en bambara sera bambara L1.

Une fois que L1 et L2 sont définis, dans le codage implicite, on note les énoncés en L1 de façon habituelle sans rien spécifier. Par contre pour dire qu'un mot n'est pas en L1, mais en L2 on le complète par @s.

*MTR: an@s ti@s yaa@s fo@s que les hommes pratiquent (.) l' agriculture .

Il est possible de spécifier en début d'énoncé qu'un énoncé est dans une autre langue que la langue par défaut. Dans ce cas, tout l'énoncé sera de manière implicite dans une autre langue.

*MTR: [- jul] bien tile man logotira do ?

Et dans le cas d'un énoncé principalement en L2, codé par [- jul] (ou une autre langue) en début d'énoncé, un mot isolé en L1 sera codé suivi de @s.

*MTR: [- jul] est+ce@s que@s bi nunu kelele xx fla be la ?

Enfin il est possible de spécifier directement la langue d'un mot spécifique, même si cette langue n'est pas décrite dans l'en-tête du fichier. Pour cela il faut simplement préciser le code de langue précédé de « : » après @s.

*ELV: les herbes sont dry@s:eng .
dans cet exemple fictif, l'élève utilise un mot anglais « dry » spécifié par @s:eng

Formes particulières ne figurant pas dans le dictionnaire

Pour coder des formes particulières, le mot est codé orthographiquement sur la ligne principale directement suivi du symbole (sans espace):

- @c : mot inventé par le locuteur
- @f : mot spécifique d'une communauté réduite (famille par exemple)
- @i: interjection (ex : ouh@i)
- @o: onomatopées (ex : coincoin@o)
- @l : lettre (ex : elle dit le r@l maintenant)
- @u : code phonétique (voir ci-dessous).

Les interjections et onomatopées figurant dans le dictionnaire n'ont pas besoin d'être suivies de @i et @o.

Les codages à utiliser dans les lignes secondaires

Les lignes secondaires suivent, complètent et spécifient les lignes principales. Elles ne doivent donc concerner que le temps de l'énoncé de la ligne principale. Des indications à caractère général doivent être indiquées par l'intermédiaire des codes @G et @Situation.

Les lignes secondaires commencent par %.

FORMAT A RESPECTER

- On y code des indications complémentaires correspondant à la ligne principale située juste au-dessus.
- On peut avoir plusieurs lignes dépendantes pour une même ligne principale, mais une seule ligne de chaque type (par exemple une seule ligne %act dépendante d'une ligne principale donnée).
- Chaque ligne dépendante commence par % suivi de lettres minuscules, ex : %pho: [TAB]
  Pas d'espace entre % et nom de la ligne
pas d'espace entre nom et :
TAB après les :
- Si l'information ne peut contenir sur une ligne, commencer la ligne suivante avec une [TAB] (normalement ceci est réalisé de façon automatique par le logiciel).

La détermination des lignes dépendantes dépend des buts de l'étude. Pour le projet Transferts, on a déterminé un minimum de lignes dépendantes qui correspondent :

- à la glose en français des segments en langue autre → %gls:
- à des commentaires optionnels → %com:
- à une description optionnelle des actes du locuteur → %act:
- à une description optionnelle de la situation → %sit:

- à une description phonologique pour les variations phonologiques et en fonction des besoins de travaux spécifiques → %pho:

Ligne %gls:

La ligne secondaire contient une glose en français du sens global des énoncés. Il n'y a pas lieu de faire une traduction « collant » à la forme de la langue traduite. Les études ultérieures visant ce but doivent pour cela utiliser le champ %mor: qui est conçu pour cela.

Ligne %com:

La ligne %com: contient toutes les informations inattendues et pouvant aider à l'explication qui ne sont pas déjà codées dans les lignes %act et %sit. Il peut s'agir tout aussi bien de commentaires techniques, linguistiques, métalinguistiques.

Ligne %act:

Coder les actions des interlocuteurs
La ligne dépendante %act se rapportera à la ligne principale juste au-dessus. Ainsi, pour coder une action sans énoncé, il faudra mettre en ligne principale « 0 ».

*ELV: 0.
%act: ELV pointe sur le dessin au tableau.

Cette technique peut aussi être utilisée pour décrire l'action qu'une personne fait en même temps que l'autre parle.

*ELV: 0 .
%act: ELV écrit la réponse au tableau.
*MTR: +< s@l (.) faut pas créer de en plus .
%sit: MTR corrige ELV.

La ligne %act est destinée à coder les gestes et les actions des interlocuteurs, surtout lorsque cela aide à comprendre la situation. Il n'est pas nécessaire de toujours le coder. Il ne faut pas non plus confondre avec le code %sit (voir ci-dessous).

Ligne %sit:

Coder les situations du discours
Il s'agit de décrire la situation ou des actions non impliquées dans la conversation (plus général que %act). Les lignes %sit peuvent concerner quiconque et pas nécessairement l'interlocuteur de la ligne principale juste avant, donc il ne sera pas nécessaire d'ajouter une ligne principale « 0 » pour ce codage.

*MTR: 0 .
%act: MTR donne un coup de bâton sur la table.
%sit: les élèves lèvent leur main tenant la craie.

Cet exemple permet de montrer la différence entre %act et %sit.

Ligne %pho:

Il y a plusieurs moyens de noter les valeurs phonétiques ou phonologiques du langage produit par les interlocuteurs. La méthode exhaustive consiste à coder la totalité des énoncés en phonétique dans une ligne dépendante appelée %pho.
Dans ce cas, on découpe la transcription sur la ligne phono en « mots » identiques à ceux de la ligne principale. Il faut exactement le même nombre d'éléments sur les deux lignes. On code les xx par X sur la ligne %pho. (Rappelons que les yy, eux, seront transcrits phonétiquement sur la ligne %pho).

*CHI: ah les xx .
%pho: a lɛ X

*CHI: ah les yy .
%pho: a lɛ putu

Ce système complet est nécessaire pour les études portant sur la phonétique ou sur la phonologie. Toutefois dans le projet Transferts où la phonologie n'est pas le but premier de l'étude, il peut être suffisant de coder seulement les mots « mal prononcés ». Les variantes systématiques liées à l'oral ou à des variations régionales ne nécessitent pas de codage phonologique dans le cadre du projet.

Dans cette optique, la pratique demandée est la suivante pour l'exemple du français (la même politique devrait être adoptée dans les autres langues): respect des conventions standards de l'orthographe lorsqu'il s'agit de la prononciation ou non de 'e' muet et des liaisons. Les formes simplifiées comme 'il y a' prononcées 'y a' sont transcrites 'y a' (et toutes leurs variantes). Il faut faire toutefois attention à ne pas rajouter d'éléments « omis » naturellement dans la langue orale.

Pour les éléments ayant une prononciation s'écartant de la norme ou des variantes standard, le codage phonétique figure directement dans la transcription, suivi de la mention du mot cible sous la forme suivante [: mot-cible] et enfin d'un code d'erreur optionnel pouvant prendre les valeurs suivantes :
[* p:w] – l'erreur produite mène à un mot existant
[* p:n] – l'erreur produite mène à un mot inexistant
[* s:r] – l'erreur produite correspond à un mot lié sémantiquement
[* n:uk] – l'erreur produite ne correspond à aucun mot existant

CHI: c'est mon mɔ̃mi@u [n:uk]
CHI: c'est un kato@u [: gateau] [p:w]

Vérification du format : check

Une fonction du programme CLAN permet de vérifier que les transcriptions suivent exactement le format officiellement défini dans CHAT (ci-dessus). Ce suivi du bon format permet d'analyser les corpus avec toutes les commandes de CLAN et d'être compatible avec toutes les extensions futures des corpus, notamment leur diffusion et leur utilisation avec d'autres outils. Il est donc absolument nécessaire d'utiliser CHECK.

L'utilisation de l'outil de vérification de format impose la mise à jour, tout d'abord du logiciel CLAN qui doit être une version récente (télécharger sur http://childes.psy.cmu.edu/clan/) et la mise à jour du fichier de définition des noms de langues. Pour cela, ouvrir avec le programme CLAN le fichier nommé ISO-639.cut et figurant dans le répertoire c:\childes\clan\lib\fixes\. Il faut insérer en fin de fichier les lignes suivantes :

jul	ju	Jula
zar	za	Zarma
bam	ba	Bambara
ful	fu	Fulfulde
hau	ha	Hausa
son	so	Songhay

Mettre une tabulation entre chaque mot. Un simple copier-coller depuis les lignes ci-dessus suffit.

Le logiciel de vérification s'utilise en lançant le menu « Mode » puis « Check opened file ». On peut y accéder directement en tapant la touche ESC puis la touche « L » (en minuscule).

Le logiciel vérifie le fichier ouvert dans CLAN et s'arrête à l'emplacement de la première erreur trouvée. Le curseur se place alors la plupart du temps directement sur l'erreur à corriger ou à défaut sur la même ligne. Une fois l'erreur corrigée, il suffit de relancer la même procédure jusqu'à obtenir le message « Success! No errors found. »

Attention le logiciel gère mal les cas d'erreurs où il manque un crochet ouvrant « [» avant une crochet fermant «] ». Il suffit de chercher les crochets au préalable pour vérifier d'abord ce type d'erreur (une prochaine version du programme devrait résoudre ce problème).

Quelques astuces

Ces quelques astuces devraient faciliter vos transcriptions/ vérifications/ corrections.

Affichage des numéros de lignes
Vous pouvez afficher les numéros des lignes en sélectionnant dans le menu « Mode » puis « Show line numbers ».

« Go to »
Vous pouvez aller directement à une ligne dont vous connaissez le numéro (utile lors du check) en sélectionnant dans le menu « Edit » puis « Go to » ou via le raccourci ctrl+L

Correction « systématique »
Pensez aux fonctions « find » (raccourci ctrl+f), « replace » (raccourci ctrl+r) et « find/replace same » (raccourci crtl+g) lorsque vous avez identifié une erreur que vous voulez corriger à travers tout le fichier.

Commandes complémentaires de CLAN

Un des avantages de CLAN est qu'il permet, une fois la transcription réalisée, d'utiliser une large gamme de commandes de manipulation de corpus permettant en particulier d'extraire

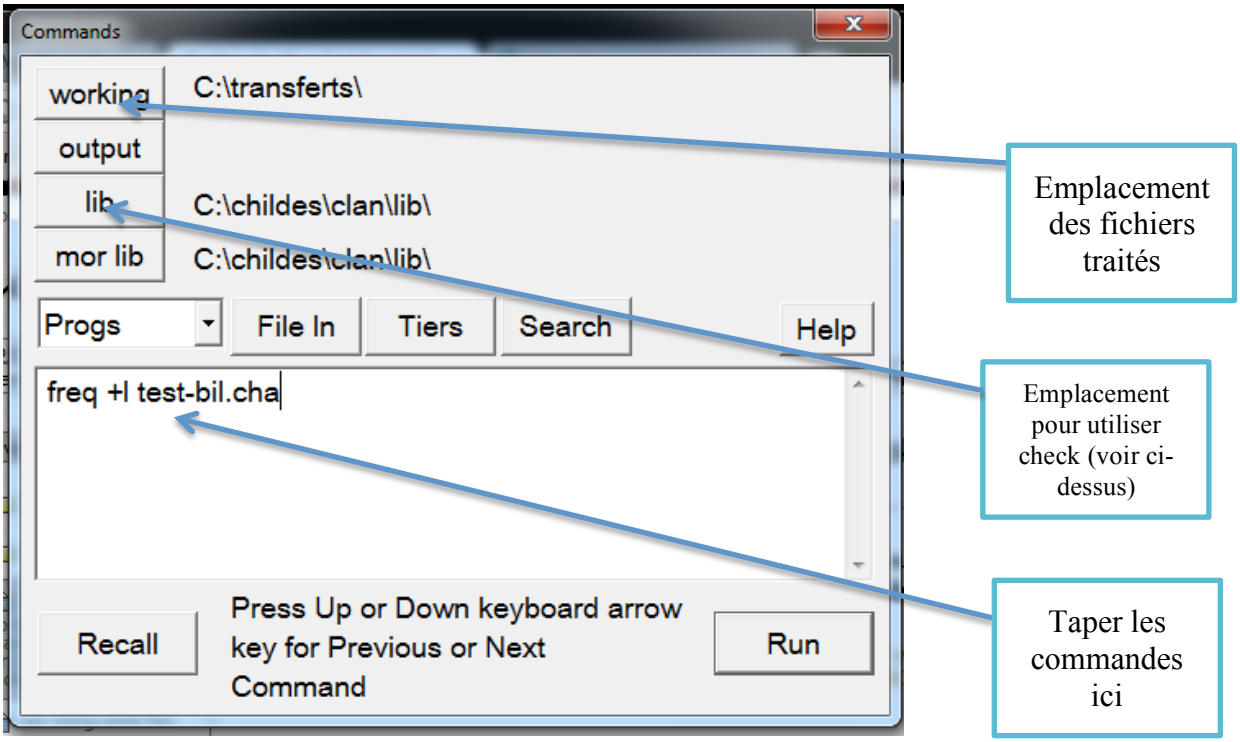
automatiquement tout le lexique, de chercher automatiquement toutes les occurrences d'un mot, d'une suite de mot, d'extraire une partie du corpus (certaines séquences par exemple).

Commande FREQ

La commande FREQ permet de récupérer tout le lexique d'un corpus et de faire directement le tri entre les différentes langues utilisées. Cette commande permet de comprendre l'intérêt qu'il y a à coder le corpus avec soin dans les cas de multilinguisme.

Principes d'utilisation des commandes : Les commandes se lancent dans une fenêtre séparée de la fenêtre principale de CLAN accessible par la commande CTRL+d ou par le menu « Windows » puis « Commands ». Les résultats des commandes sont affichés dans la fenêtre principale de CLAN ou sauvegardés directement dans un fichier qui peut par la suite être lu par CLAN ou par EXCEL ou par WORD.

La fenêtre de commande permet de déterminer l'emplacement (le répertoire) des fichiers sur lesquels on va travailler et de taper les commandes à effectuer. L'emplacement des fichiers à traiter se détermine en cliquant sur le bouton « working » et en choisissant l'emplacement avant de valider. Le résultat s'affiche à côté. Les fichiers résultats générés seront au même endroit à moins que le champ « output » ne soit rempli (s'il est vide, il est identique à « working »).



La commande à effectuer pour générer le lexique d'un texte est simplement :

freq nom-de-fichier.cha

où « nom-de-fichier.cha » est le nom complet de la transcription. En ajoutant « -l » à la commande on obtient un code langue pour tous les mots. Par exemple, si on a un fichier test-bil.cha qui contient :

*MTR: ounhoun@i .
*ELV: ou de la soleil .

*MTR: pendant la saison sèche +...
*MTR: [- jul] tle mana wagatira +...
*MTR: [- jul] an lam nyina mulo ba yira xx xx tle mana wagatila ?
*ELV: 0 [=! tousse] .
*MTR: ounhoun@i .
*ELV: les herbes sont sèches .
*MTR: les herbes sont (.) sèches.

Le résultat de la commande « freq -l test-bil.cha » donnera :

```
freq -l test-bil.cha
Tue Apr 10 10:03:50 2012
freq (28-Mar-2012) is conducting analyses on:
  ALL speaker tiers
****************************************
From file <test-bil.cha>
  1 an@s:jul
  1 ba@s:jul
  1 de@s:fra
  2 herbes@s:fra
  2 la@s:fra
  1 lam@s:jul
  2 les@s:fra
  2 mana@s:jul
  1 mulo@s:jul
  1 nyina@s:jul
  1 ou@s:fra
  2 ounhoun@i@s:fra
  1 pendant@s:fra
  1 saison@s:fra
  1 soleil@s:fra
  2 sont@s:fra
  1 sèche@s:fra
  2 sèches@s:fra
  2 tle@s:jul
  1 wagatila@s:jul
  1 wagatira@s:jul
  2 xx@s:jul
  1 yira@s:jul
------------------------------
   23  Total number of different item types used
   32  Total number of items (tokens)
0.719  Type/Token ratio
```

On voit qu'il y a un sigle, @s:fra ou @s:jul, par mot pour indiquer la langue originale. CLAN a généré automatiquement la totalité des extensions de langue (le @s étant une forme réduite de la forme complète comprenant l'abréviation de la langue). Les chiffres correspondent au nombre d'occurrences des mots. Ce résultat peut être filtré ou intégré dans EXCEL.

Ce résultat peut être sauvegardé en utilisant le menu « file » puis « save as ». On peut aussi directement générer un fichier résultat en faisant :

```
freq -l +f test-bil.cha
```

Dans ce cas, le logiciel affichera :

```
From file <test-bil.cha>
Done with file <test-bil.frq.cex>
```

ce qui veut dire qu'il a traité le fichier test-bil.cha et que le résultat a été mis dans le fichier test-bil.frq.cex (ce fichier peut être visualisé par CLAN, ou intégré dans WORD ou dans EXCEL : pour EXCEL, utiliser la fonction « importation de données texte »).