



Corpus PAROLE

Parallèle Oral en Langue Étrangère

**Architecture du corpus &
conventions de transcription**

Laboratoire LLS – Équipe Langages
Université de Savoie

juin 2008
Heather Hilton

Corpus PAROLE

Parallèle Oral en Langue Étrangère

Architecture du corpus & conventions de transcription

Membres du projet PAROLE

Équipe *Langages*
du laboratoire *Langages, Littératures, Sociétés*
UFR-LLSH, Université de Savoie

Heather HILTON

(corpus anglais, corpus français, vérifications du corpus italien)

John OSBORNE

(corpus anglais, corpus français, vérifications du corpus italien)

Marie-Jo DERIVE

(corpus français)

Nejma SUCCO

(corpus anglais, vérifications du corpus français)

Jean O'DONNELL

(corpus anglais)

Sandra BILLARD

(corpus italien)

Sandrine RUTIGLIANO-DASPET

(corpus italien)

Nous remercions le laboratoire LLS pour son soutien du projet,
ainsi qu'Alice Henderson, pour son aide lors des enregistrements.

Nous remercions également (par ordre chronologique de leur passage à Chambéry)
les experts externes, qui ont accepté de débattre avec nous de la conception du corpus :

Florence MYLES, Daniel VERONIQUE, Colette NOYAU.

Toute faiblesse dans l'architecture de l'ensemble relève, bien sûr,
entièrement de la responsabilité de l'équipe PAROLE.

Table des matières

PRESENTATION DU CORPUS.....	4
OBJECTIFS.....	4
PARTICIPANTS	5
RECOLTE DES DONNEES	5
LES TACHES DE PRODUCTION.....	5
ENREGISTREMENT DES PRODUCTIONS.....	5
TACHES & SUPPORTS	6
TACHE A : « LE FRIGO »	7
TACHE B : « LA DAME AUX BEIGNETS »	7
TACHE C : « L'ÉLEPHANT »	7
TACHE D : « LES PHRASES COMPLEXES »	7
TACHE E : « L'ACCIDENT »	8
LES DONNEES COMPLEMENTAIRES.....	8
QUESTIONNAIRES.....	8
PROFIL DES PARTICIPANTS	8
QUESTIONNAIRE DE MOTIVATION	8
PASSAGE DES TESTS	9
LES TESTS UTILISES.....	9
TEST DE COMPREHENSION DE L'ORAL (DIALANG)	9
TEST DE CONNAISSANCES GRAMMATICALES (DIALANG)	9
TESTS DE CONNAISSANCES LEXICALES (DIALANG ET FORUMÉDUCATION)	9
TEST D'APTITUDE A L'ANALYSE GRAMMATICAL	10
TEST DE MEMOIRE PHONOLOGIQUE (REPETITION DE PSEUDOMOTS)	10
PREPARATION A LA TRANSCRIPTION DANS CHAT.....	11
INSERER LES « BALISES » (SEGMENTER LE FICHIER SON)	11
PREPARATION A LA TRANSCRIPTION (ORGANISATION DES FICHIERS)	11
DEMARRER UNE NOUVELLE TRANSCRIPTION : POSER D'ABORD LES BALISES	11
RAJOUT DES RUBRIQUES (HEADERS)	12
TRANSCRIPTIONS EN « MODE SONIQUE ».....	13
IDENTIFIER & ECOUTER UN SEGMENT BALISE.....	13
RETROUVER UN SEGMENT BALISE DANS LA BARRE SONIQUE.....	13
SELECTIONNER ET ECOUTER UN PETIT SEGMENT	13
CHANGER LE « RELIEF » DE LA BARRE SONIQUE	13
CHRONOMETRER AVEC LA BARRE SONIQUE	13
MAINTENIR INTACTES LES BALISES « COMPLETES »	14
RAFISTOLER LES BALISES	14
ÉLARGIR OU REDUIRE UN SEGMENT BALISE (SUR LA BARRE SONIQUE)	14
ÉLARGIR, REDUIRE OU FUSIONNER BALISES (DANS LA TRANSCRIPTION)	15
RESUME DES COMMANDES DE LA BARRE SONIQUE	15
DELIMITER LES ENONCES DANS PAROLE.....	16
CRITERES DE BASE	16
PROPOSITION INDEPENDANTE SIMPLE.....	16
PROPOSITION INDEPENDANTE + COMPLETIVE(S).....	16
PROPOSITION INDEPENDANTE + AUTRES SUBORDONNEES	16

CAS PARTICULIERS	16
EFFACEMENT DU SUJET	16
AUTRES PROPOSITIONS COORDONNEES.....	17
PROPOSITIONS CONCATENEES (<i>RUN-ON SENTENCES</i>).....	17
INCISES	17
TRANSCRIRE PAUSES ET HESITATIONS.....	18
PAUSES SEULES.....	18
LES PAUSES SILENCIEUSES SEULES (<i>SP – SILENT PAUSE</i>).....	18
LES PAUSES VOCALISEES (<i>FP – FILLED PAUSE</i>).....	18
AUTRES PHENOMENES D'HESITATION	18
BRUITAGES PARALINGUISTIQUES.....	18
ALLONGEMENTS (VOYELLES & CONSONNES).....	19
GROUPES D'HESITATION.....	19
CHRONOMETRER LA DUREE D'UNE PAUSE OU D'UN GROUPE D'HESITATION	20
TRANSCRIRE LES REPRISES.....	20
CODAGE DES REPRISES DANS PAROLE	20
[/] = REPETITION SIMPLE, SANS CHANGEMENT	20
[//] = REFORMULATION SIMPLE	21
[///] = REDEMARRAGE (RESTART) [REFORMULATION SYNTAXIQUE]	21
[/-] = ABANDON (<i>FALSE START</i>)	21
REPRISES & ERREURS	21
RESUME DES CONVENTIONS CHAT UTILISES DANS PAROLE	22
POSE DE BALISES (= SEGMENTATION DU FICHIER-SON) : RAPPEL.....	22
EN-TETES.....	22
LIGNE PRINCIPALE	23
ÉNONCES.....	23
ÉNONCES INCOMPLETS & INTERROMPUS* ; RECOUVREMENTS.....	23
CITATIONS, COMMENTAIRES METALINGUISTIQUES	24
PONCTUATION	24
HESITATIONS, REPRISES, APPROXIMATIONS, MOTS INCOMPLETES, BRUITAGES... ..	24
ERREURS	25
LIGNES SECONDAIRES.....	26
CODER LES ERREURS	26
LES PERLES (<i>GEMS</i>)	29
VERIFICATION EN COURS DE TRANSCRIPTION.....	30
RESULTATS.....	31
CONCLUSIONS & PERSPECTIVES	31

Présentation du corpus

Objectifs

En 2003, au sein d'une équipe pluridisciplinaire, nous avons décidé de mettre sur pied un projet qui pouvait fédérer des chercheurs s'intéressant à trois L2 différentes (l'anglais, le français et l'italien), et à différents domaines de la maîtrise linguistique : phonologie, lexique, morphosyntaxe. L'idée de base était de constituer un corpus oral de productions à différents niveaux de compétence, afin d'investiguer et de comparer les caractéristiques linguistiques de ces différents niveaux (intra et inter-langues).

L'idée d'un corpus *PARallèle Oral en Langue Etrangère* – le Corpus PAROLE – était lancée.

Un certain nombre des tâches proposées aux participants allaient être contraintes, nous permettant de comparer plus aisément les productions ; une dernière tâche plus ouverte allait permettre à nos sujets de s'exprimer plus personnellement et plus naturellement.

Dès les premières transcriptions, nous nous sommes rendu compte de l'importance des aspects temporels des productions aux différents niveaux ; après consultation des travaux sur l'aisance en production orale (en L1 et en L2), nous avons intégré dans PAROLE un codage minutieux des phénomènes d'hésitation et de reprise.

Nous avons voulu des stimuli – ou « déclencheurs » (*triggers*) de la parole – drôles et motivants, d'où le choix de supports vidéo courants (vidéo gag, caméra cachée, publicité), plutôt que de supports papier. Ceci nous a empêché d'utiliser les mêmes supports que d'autres créateurs de corpus en L2 (*The Frog Story* ou *Modern Times*, par exemple) ; ces supports ne semblaient pas adaptés à nos besoins (public de jeunes adultes, pour qui *the Frog Story* risquait d'être infantilisante et paradoxalement trop sophistiquée sur le plan lexical ; l'extrait de *Modern Times* utilisé dans le projet ESF était trop long pour l'intégrer à notre protocole, déjà chronophage). Nous avons pré-testé nos tâches et nos déclencheurs ; ces premiers participants (n= 10) ont tous réagi très favorablement aux supports.

Pour pouvoir croiser les caractéristiques des productions dans PAROLE avec d'autres informations pertinentes, nous avons constitué une batterie de six « tests complémentaires », mesurant les connaissances linguistiques des sujets dans leur L2, leur niveau dans d'autres compétences communicatives dans cette langue et quelques aptitudes langagières.

Depuis nos toutes premières discussions concernant PAROLE, nous savions que le travail s'effectuerait à l'aide des outils de transcription et d'analyse proposés par CHILDES, car l'un de nos objectifs majeurs est de partager nos données avec d'autres chercheurs s'intéressant à l'acquisition des compétences discursives en L2.

NOTE: For the moment, we are not planning to provide the contents of this manual in English; descriptions of the project and of our findings will be available to the research community in the form of articles and other publications, of course. We welcome questions in English from fellow language researchers who are not able to understand the explanations given here in French: hilton@univ-savoie.fr or osborne@univ-savoie.fr.

Participants

Les participants au projet (jeunes adultes inscrits à l'Université de Savoie) furent recrutés sur la base du volontariat, au sein de cours se déroulant sur nos campus (automne 2005, printemps 2006, et printemps 2007). Le tableau présente le nombre et l'origine des contributeurs au corpus PAROLE :

L2	n=	L1 des sujets	âge (moyen)	sexe	corpus « natif »	sexe	âge (moyen)
anglais	33	français (24) allemand (9)	21	F (29) M (4)	n= 9	F (6) M (3)	21
italien	23	français	20	F (17) M (6)	n=10	F (8) M (2)	23
français	12	espagnol (5) chinois (3) suédois (2) polonais (1) anglais (1)	23	F (8) M(4)	n= 8	F (7) M (1)	21
TOTAUX	68				27		

Chaque contributeur au corpus non-natif fut rémunéré (au SMIC ou plus, selon ses diplômes) pour les trois heures consacrées au projet. Nous avons attribué à chaque sujet un numéro d'identité à trois chiffres, pour un corpus complètement anonyme (tests, enregistrements, transcriptions) :

- non-natifs, anglais L2 – ID commençant par 0 (012, 006, 034...)
- non-natifs, italien L2 – ID commençant par 2 (212, 206, 221...)
- non-natifs, français L2 – ID commençant par 4 (412, 406...)
- natifs, toutes langues – ID commençant par N (N0x pour l'anglais, N2x pour l'italien, N4x pour le français)

Récolte des données

La première heure de participation fut consacrée au remplissage du dossier administratif (nécessaire aux rémunérations) et des deux questionnaires (profil, motivation), avec 20 minutes pour l'enregistrement des productions pour le corpus. Pour économiser du temps, les sujets étaient convoqués par groupes de 2 ou 3 participants, une ou deux personnes remplissant les questionnaires et formulaires pendant qu'une autre passait à l'enregistrement.

Une ou deux semaines plus tard, chaque sujet a consacré deux heures au passage de six « tests complémentaires », dans le but d'obtenir des informations sur son niveau en L2, ses connaissances lexicales et grammaticales dans la L2, ainsi que ses capacités de traitement phonologique et d'analyse grammaticale. A part la répétition de pseudomots, les tests étaient administrés en format numérique, dans une salle informatisée.

Les participants natifs ont répondu à des questions concernant leur profil linguistique ; ils n'ont pas passé de tests, et ont consacré environ 15 à 20 minutes au projet. Ils n'ont pas été rémunérés, mais ont reçu des cadeaux (magazines et chocolats) à titre de remerciement.

Les tâches de production

Enregistrement des productions

Les sujets ont été accueillis individuellement dans une petite salle aménagée pour l'enregistrement des productions. Ils ne connaissaient pas l'interviewer, natif de la L2 en question (et enseignant-chercheur ou doctorant à l'Université de Savoie).

Les enregistrements ont été effectués avec un enregistreur numérique (Marantz PMD660, *très satisfaisant*), et un microphone directionnel Sony (IM Stage Line ECM-925P, également *très performant*). Les fichiers mp3 générés furent convertis en fichiers .wav pour le travail de transcription (version 2005 de CLAN au début du projet).

Le sujet fut installé devant un ordinateur portable ; l'interviewer – en face – était incapable d'en voir l'écran. Pour les tâches A, B et C, le sujet résumait ou commentait le contenu de la séquence pour l'interviewer, qui ne pouvait pas la voir.

Les instructions pour chaque tâche de production, ainsi que les « déclencheurs » visuels étaient intégrées à un ensemble de pages html, dans lesquelles le sujet avançait en cliquant. Après un écran de bienvenue en français, les textes des diapositives étaient tous en L2 (consignes, boutons de pilotage, etc.). Le sujet était invité à adresser toute question concernant la procédure à suivre à l'interviewer (qui vérifiait également la compréhension des consignes, de façon rapide et informelle, mais en L2 : « Do you understand what to do ? Fine, go ahead. »). Nous avons opté pour un protocole « tout en L2 » pendant les enregistrements, afin d'optimiser les conditions de production de la langue (limiter la traduction, le va et vient entre L1 et L2, etc.).

A l'intérieur des pages html, le sujet pilotait les séquences vidéo pour les trois premières tâches lui-même ; il pouvait revenir en arrière, regarder la séquence plusieurs fois, poser ses questions à l'interviewer, etc. La plupart des sujets ont regardé chaque séquence une seule fois, et se sont ensuite lancés dans la tâche de production sans hésitation.

Aucun mot de vocabulaire n'était fourni dans les pages de présentation.

Les interviewers avaient comme consigne de laisser parler le sujet, en l'encourageant si nécessaire, mais en évitant de fournir mots et structures. Malheureusement, et malgré notre pré-test du dispositif, nous n'avons pas adopté un protocole suffisamment rigoureux cadrant l'interaction (linguistique) de l'interviewer avec le sujet ; nous avons une certaine variation dans le comportement des interviewers, qui a, bien évidemment, un effet sur les productions des sujets. Nous conseillons, pour de futures recherches de ce genre, un comportement linguistiquement discret de la part de l'interviewer, mais toutefois communicatif, pour optimiser la quantité de langage produit :

- Il ne faut pas rester silencieux (comme l'un de nos interviewers), car le sujet se comporte alors comme dans une situation d'examen, tentant de boucler sa production le plus rapidement possible, et renonçant à coder un certain nombre d'informations, sans doute par crainte de commettre des erreurs.
- Il ne faut pas non plus *trop* parler – posant des questions et fournissant des mots – car ensuite on a du mal à distinguer (sur les plans lexical et propositionnel) ce qui émane du sujet de ce qui est « soufflé » par l'interviewer. En plus, ces échanges plus conversationnels sont diablement difficiles à transcrire.
- Il faut, en fait, un juste milieu : bruits d'assentiment (se situant entre deux énoncés), petits « oh ! » en réaction à une information surprenante, etc. – qui montrent au sujet que l'on écoute, que l'on capte et traite normalement les informations qu'il fournit. On peut réagir de façon un peu plus expansive à la fin du résumé ou du récit, afin de démontrer une implication interactive dans la construction du sens (ces petits échanges semblent remotiver le sujet pour la prochaine tâche).
- L'interviewer doit également veiller à la cohérence de son comportement, dans le cadre de la ruse sur laquelle l'échange est basé (le sujet résume pour l'interviewer ce qu'il n'a pas vu). Le sujet est bien sûr conscient de l'artificialité de la situation, mais il vaut mieux préserver les apparences ; à plusieurs reprises, un interviewer a démontré par ses questions une trop grande connaissance des contenus à l'écran.

Tâches & supports

Chaque participant à PAROLE (sujets natifs et non-natifs) a complété cinq tâches de production, décrites ci-dessous dans l'ordre dans lequel elles ont été effectuées. La version en ligne de PAROLE comportera les transcriptions pour chaque sujet des tâches A (frigo), C (éléphant) et E (accident) : deux résumés de séquences vidéo, donc, et un court récit autobiographique. Le corpus non-natif pour ces trois tâches contient 20.000 mots, et le corpus natif environ 10.000 mots, avec de trois à six minutes de production par sujet.

Les séquences vidéo utilisées dans PAROLE seront disponibles sur le site du laboratoire *Langages* de l'Université de Savoie (nous contacter pour le lien). Nous attirons l'attention de la communauté scientifique sur le fait que nous n'avons pas négocié de droits de diffusion de ces clips ; merci d'exercer la plus grande prudence dans leur utilisation.

Tâche A : « le frigo »

Résumé après visionnage d'une courte séquence vidéo (muette), de type « vidéo gag ».

Nous avons édité la séquence, enlevant la bande de son, ainsi que la reprise de la scène finale. Le clip montre une scène de déménagement, où un réfrigérateur tombe du haut d'un immeuble sur une voiture dans la rue.

Comme toutes les séquences vidéo, nous avons choisi celle-ci pour les défis particuliers qu'elle présente à l'encodage linguistique des événements vus à l'écran :

- description de relations spatiales et mouvements
- l'objet saillant (le frigo) est d'abord patient, ensuite agent des événements mis en scène
- la séquence pose quelques défis lexicaux (mots infréquents, comme *grue*, ainsi que le lexique de l'émotion nécessaire à la description de la réaction d'un homme qui observe la scène).



Tâche B : « la dame aux beignets »

Description de la séquence pendant le visionnage (*running commentary*), avec un résumé final.

Séquence extraite de *Candid Camera Classics* (KTel Video, 1988) : la caméra cachée espionne une ouvrière débutante confectionnant des pâtisseries à la chaîne, qui est confrontée à une machine déréglée. Séquence choisie pour :

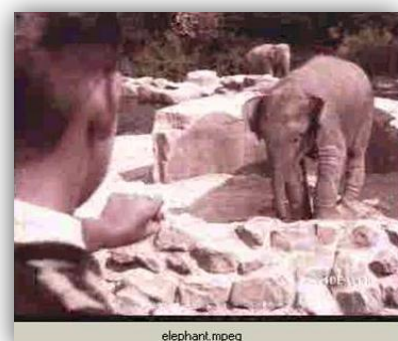
- son personnage central féminin (les autres séquences ne mettant en scène qu'hommes, garçons et animaux au masculin [en italien et en français])
- la nature répétitive des événements à décrire (codage de l'itération).

Tâche C : « l'éléphant »

Résumé après visionnage d'une séquence publicitaire (comportant un peu de langage, non-essentiel à la compréhension de l'ensemble).

Séquence choisie pour :

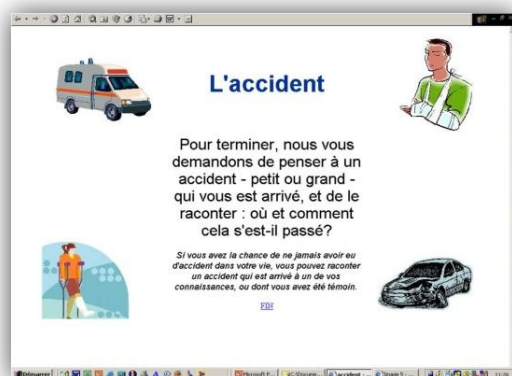
- sa structure en deux parties (l'une au passé, l'autre au présent) ; le même personnage central est d'abord garçon, ensuite adulte ; nous nous intéressons à l'encodage de ce contraste temporel, et à l'organisation du discours
- pour comprendre l'humour de la publicité il faut interpréter les intentions des participants, et donc expliquer les motifs des comportements
- l'encodage des échanges entre les deux protagonistes de la publicité nécessite l'utilisation de structures ditransitives
- ses défis lexicaux.



Tâche D : « les phrases complexes »

Cette tâche est dérivée du « Formulated Sentences Test » (Semels et al. 1994, *CELF 3 – Clinical Evaluation of Language Fundamentals*, London : The Psychological Corporation). Dans notre version

de la tâche, le sujet doit formuler neuf phrases complexes, à partir d'une image fixe, en incorporant dans chaque phrase une conjonction de coordination ou de subordination (*mais, si, parce que, quand...*) affiché en haut de l'écran. La tâche fut introduite afin d'obliger les sujets à formuler des phrases complexes, car les pré-tests de nos déclencheurs avaient révélé que certains sujets – stratégiquement et tout à fait intelligemment – se limitaient à un enchaînement de phrases simples. Pour l'instant, nous n'avons pas transcrit les productions liées à cette tâche ; elles sont notées (selon le barème du test), et le score obtenu est considéré comme une mesure complémentaire de la compétence productive en L2.



Tâche E : « l'accident »

Récit autobiographique, d'un accident, à partir d'un écran donnant quelques idées et quelques images, à titre d'inspiration. Tâche proposée pour :

- provoquer des productions plus libres
- susciter un récit personnel
- parlant d'événements passés.

Les données complémentaires

Il nous semblait indispensable de disposer d'informations concernant les connaissances en L2 des sujets non-natifs ; nous aurions aimé disposer de plus d'informations encore – capacités mnésiques des sujets, connaissances collocationnelles dans la L2, préférences perceptuelles, comportements stratégiques... Le manque d'outils standardisés (en différentes langues) limite la qualité et la quantité de données de ce genre que l'on peut obtenir en un temps raisonnable ; la batterie de tests et questionnaires proposée à nos sujets est loin d'être complète, mais elle fournit quelques informations intéressantes et qui manquent, généralement, dans l'étude des productions en L2.

Questionnaires

Les deux questionnaires étaient proposés en format papier, pour permettre aux participants de les remplir facilement en « salle d'attente » des enregistrements. Dépouillement manuel, donc.

Profil des participants

Un questionnaire pour relever les informations les plus pertinentes concernant le contact du sujet avec la L2 du projet, ainsi que son profil linguistique général (autres langues étudiées, plurilinguisme en dehors du contexte scolaire), a été conçu par les membres de l'équipe PAROLE. Nous avons limité le questionnaire à une feuille recto-verso.

Questionnaire de motivation

Le questionnaire de motivation comporte 47 affirmations (positives et négatives) auxquelles le sujet doit réagir selon une échelle de six réactions possibles (le score de motivation maximale étant donc de 282 points) ; encore une fois, nous nous sommes limités à une feuille recto-verso.

Adapté de Gardner, R. (2004) *Attitude & Motivation Test Battery* (London, Ontario : University of Western Ontario), avec la permission de Robert Gardner.

Passage des tests

Cinq des six tests complémentaires ont été administrés dans une salle informatique interactive, accueillant 18 participants à la fois. Chaque outil fut rapidement présenté au groupe, qui disposait également d'instructions écrites pour les tests, en français. Un membre de l'équipe PAROLE pilotait la séance, et répondait aux questions des participants. Il aidait les sujets à noter les résultats obtenus aux tests sur une fiche individuelle, et vérifiait la précision de ces notations. Chaque sujet travaillait à son rythme ; il n'y avait pas de limitation de temps pour les tests (tous les sujets y ont passé entre 90 et 120 minutes).

Une fois les cinq tests informatisés terminés, les sujets quittaient la salle interactive, pour passer un test de répétition de pseudomots, dans une petite salle insonorisée. Voir « Données complémentaires », ci-après.

Les tests utilisés

Quand on veut faire passer des tests comparables dans le cadre d'un projet comportant différentes L2, le nombre d'instruments disponibles est malheureusement très restreint. Pour les tests de compétence et de connaissances en L2, nous avons adopté les instruments proposés en ligne par le consortium européen DIALANG (www.dialang.org), car ils existent dans les trois langues du projet. La version statique de chaque test fut utilisée (version obtenue par défaut, si l'on saute le « test de niveau » proposé en début de séance), afin de rendre les résultats plus comparables.

Pour les tests informatisés, les sujets devaient relever – avec l'aide du coordonnateur des tests – leurs résultats sur une fiche individuelle. Le résultat obtenu à chaque test de DIALANG est éphémère (il ne rentre pas dans une base de données, devenant donc impossible à récupérer dès que l'on passe au prochain test) ; le coordinateur des tests veillait étroitement à l'avancement des sujets et aux fiches de résultats. Malgré cette vigilance, nous avons perdu quelques scores – parce que le sujet avait quitté trop vite la fenêtre des résultats, mais également parce que le test en ligne a bloqué à quelques reprises, sans fournir de score au test.

Le passage des tests de DIALANG étant chronophage (environ 80 minutes pour trois tests), nous nous sommes limités à quatre tests sur les connaissances et compétences en L2, et deux tests concernant des aptitudes réputées pertinentes dans le niveau atteint en L2 (un test de l'aptitude à l'analyse grammaticale et un test de la mémoire phonologique).

Test de compréhension de l'oral (DIALANG)

Test « Compréhension de l'oral » de DIALANG, comportant 30 questions.

A la fin du test, le logiciel indique le niveau européen du sujet en compréhension. Le test étant conçu dans une optique d'évaluation diagnostique qualitative, aucun « score » n'est donné, mais le détail des réponses correctes et incorrectes est fourni. Nous avons donc compté les réponses correctes, et noté ce montant comme le « score » du sujet au test. Cette donnée quantitative s'est avérée très utile dans les analyses statistiques effectuées par la suite.

Test de connaissances grammaticales (DIALANG)

Test « Structures » de DIALANG, qui comporte 30 questions en anglais et en italien et 32 en français. Les sujets ont noté leur niveau européen « en grammaire », ainsi que leur « score » (nombre de réponses correctes).

Tests de connaissances lexicales (DIALANG et ForumEducation)

Chaque sujet non-natif a passé deux tests de vocabulaire en L2.

- Tous ont passé le test « Vocabulaire » de DIALANG dans leur L2. Le test comporte 30 questions ; le niveau européen et le « score » ont été notés comme avant.

- Les apprenants du français et de l'italien L2 ont également complété le « Test de niveau » de DIALANG (basé sur *The Eurocentres Vocabulary Size Test*, Meara & Jones (1990), conçu pour quantifier rapidement l'étendu du lexique en L2). 75 items, score sur 1000 points.
- Les apprenants de l'anglais L2 ont pu passer un autre test informatisé estimant l'étendu du lexique en anglais L2 : B. Hever (sans date) *General English Vocabulary Test*, « ordinary level », ForumEducation, Suède. (Ce test est actuellement diffusé sous le nom *Lemma*, sur le site *Words&Tools* : http://www.wordsandtools.com/vocdemo/index_uk.htm). Chaque sujet reçoit un score (sur 600 points) pour le test, qui indique le nombre de mots reconnus à l'écrit en anglais (exprimé en milliers de mots) : par exemple, un score de 250 indique que le sujet reconnaît environ 4000 mots d'anglais. Les deux chiffres – score et taille estimée du vocabulaire – furent notés pour chaque sujet (anglais L2).

Test d'aptitude à l'analyse grammaticale

Une bonne capacité d'analyse grammaticale est censée faciliter la réussite en AL2. Nous avons donc demandé à nos sujets de compléter le « Test C » de la batterie de tests d'aptitude à l'apprentissage des langues *Lognostics* : Meara, P. M., Milton, J. & Lorenzo-Dus, N. (2001) *Test C, Lognostics Language Aptitude Tests* (Newbury : Express Publishing). L'interface de ces tests est en anglais ou en français, au choix.

Cet outil donne (et stocke dans une base de données, sous le numéro du sujet) le pourcentage de réponses correctes obtenu par chaque sujet.

Nous nous sommes limités à ce seul test de la batterie *Lognostics LAT*, toujours pour limiter raisonnablement le temps consacré aux tests complémentaires.

Test de mémoire phonologique (répétition de pseudomots)

En plus des cinq tests informatisés, chaque sujet non-natif a complété un test de répétition de pseudomots en L2 ; les apprenants anglophones et francophones ont également passé le test en L1. La répétition de pseudomots (en L1) est un test psychométrique utilisé pour déceler d'éventuelles déficiences de la mémoire phonologique (dans le dépistage de la dyslexie, par exemple). Nos recherches ayant déjà démontré que la mémoire phonologique en L2 peut varier significativement des capacités de la mémoire phonologique en L1, nous avons décidé d'incorporer ce test à notre batterie.

La répétition de pseudomots est un test oral, qui fut administré individuellement par un examinateur natif de la langue testée selon le protocole stricte qui régit ce test psychométrique. Les sujets furent notés pendant le test, mais également enregistrés (pour permettre une vérification du score obtenu). Instruments utilisés :

- Français L1 et français L2 : Casalis, S. (2000). *Répétition de logotomes*. Lille : Université de Lille.
- Anglais L2 : adaptation de Gathercole & Baddeley (1996). *The Children's Test of Nonword Repetition*. London : Psychological Corporation.
- Italien L2 : adaptation de Sartori, G., Job, R., Tressoldi, P. E. (1995). *Batteria per la valutazione della dislessia e della disortografia evolutiva*. Firenze : O. S.

Préparation à la transcription dans CHAT

Les sections qui suivent dans ce livret contiennent les consignes qui ont été transmises aux transpositeurs intervenant dans le projet PAROLE. Une partie de ces précisions résume tout simplement les contenus du manuel CHAT (MacWhinney 2005-2007), pour les membres de l'équipe PAROLE qui ne sont pas suffisamment à l'aise avec l'anglais pour tout y comprendre¹. Et, bien sûr, une assez grande partie de ces consignes précise la façon dont notre équipe a interprété et adapté les conventions de transcription CHAT, selon nos hypothèses concernant la production orale en L2 et les analyses que nous voulons effectuer dans PAROLE. Ces pages concernent donc les détails techniques de nos transcriptions ; elles peuvent servir d'explication aux systèmes de codage adoptés dans PAROLE, ainsi qu'à un document-conseil pour des utilisateurs francophones de CHAT.

Insérer les « balises » (segmenter le fichier son)

Préparation à la transcription (organisation des fichiers)

Avant de commencer à travailler, créer (dans Windows Explorer) un dossier par tâche, où tu mettras tous les fichiers (son et CLAN) pour la tâche en question. Organisation suggérée :

dossier principal : un dossier par langue

transcriptions_eng
transcriptions_fra
transcriptions_ita

sous-dossiers (utiliser le mot *task*, *tâche* ou *compito* selon la L2) : un dossier par tâche

task_A
task_B
task_C
task_D
task_E

Dans chaque dossier, coller *une copie* des fichiers son pour la tâche en question, tous sujets (tous les fichiers XXXa.wav dans le dossier « task_A », tous les fichiers XXXb.wav dans le dossier « task_B », etc.).

C'est dans le dossier de la tâche que tu sauvegarderas aussi tous les fichiers .cha de transcription. Créer aussi / maintenir un dossier Enregistrements, dans lequel tu mets (sans sous-dossiers, en vrac) TOUS les fichiers son (segmentés) pour la langue. Histoire de garder une version centralisée de ces fichiers pour des raisons de sécurité. (Ne pas oublier d'effectuer une sauvegarde de ces mêmes fichiers, version segmentée définitive, sur un jeu de CD-ROMs, pour archivage au labo.) S'il y a rafistolage des fichiers son par la suite, il faudra mettre à jour les CD-ROMs.

Démarrer une nouvelle transcription : poser d'abord les balises

Les « balises » (*bullets*) correspondent aux repères temporels (renvoyant au fichier son) de chaque énoncé. (Voir la section « Délimiter les énoncés », ci-après, pour nos définitions de ce qui constitue un énoncé.)

Créer un nouveau document .cha, en passant par le menu **File > New**.

Sauvegarder le nouveau fichier tout de suite : **Fichier > Save As...** > remplacer « newfile » dans le nom de fichier proposé par « 003c.cha », par exemple (sujet 003, tâche C [éléphant]). N'oublie pas d'effectuer cette sauvegarde dans le bon sous-dossier (ici, **transcriptions_eng > task_C**).

Menu **Mode > Transcribe sound or movie**.

¹ Ce qui explique le tutoiement qui apparaît dans les consignes – que nous avons décidé de maintenir, car nous nous adressons à une communauté restreinte de chercheurs comme nous.

Une fenêtre s'ouvre, te demandant d'identifier le fichier son qui sera lié à la transcription : **Please locate movie, sound, picture or text file**. Dans la case « regarder dans », choisir le bon sous-dossier. (Si tu as suivi les consignes ci-dessus, le fichier son qui correspond au segment à transcrire devrait déjà se trouver dans le même sous-dossier pour la tâche.) Dans cet exemple, je sélectionnerais le fichier 003c.wav. Avant de cliquer sur « Ouvrir », vérifier que ton casque/ tes haut-parleurs sont branchés, car tu passeras immédiatement en écoute active du fichier son.

Le fichier se mettra donc tout de suite en route ; poser les balises en appuyant sur la barre espace à chaque nouvel énoncé.

Si dans la foulée tu te rends compte que tu as mal posé une balise, cliquer tout de suite dans la fenêtre CLAN, pour arrêter le balisage. Revenir sur la ligne à changer avec le curseur & cliquer. La touche **F5** (touches fonction, en haut du clavier) relancera la procédure de balisage à cet endroit.

Toutes les balises au-dessous de ce point seront automatiquement effacées (les balises « en amont » resteront inchangées).

Avec un peu de pratique, ta capacité à identifier les énoncés s'améliora, et le balisage deviendra moins stressant.

Dès la fin du balisage, **SAUVEGARDER !**

Rajout des rubriques (*headers*)

Avant de transcrire, il faut établir les rubriques (*headers*), qui permettront d'identifier l'auteur de chaque production.

Créer un fichier-type pour les rubriques

Pour simplifier la saisie des rubriques (un peu fastidieuse), j'ai suivi une procédure de copier-coller, à partir d'un petit fichier-type. Créer – à l'intérieur du dossier pour chaque tâche – un fichier CLAN avec les headers pour chaque groupe de sujets (selon la filière et la tâche), intitulé « headers_LEAa » ; « headers_CAPESc », etc. A la place du numéro du sujet dans ce fichier générique, je mets **000** (trois zéros) ; à la place de l'âge je mets **âge;00.00** ; à la place du sexe je mets **sexe**. Je rentre déjà la date des enregistrements, et toute autre information stable (nom du projet, de la tâche, etc.).

J'ouvre le fichier *headers* pertinent avant de commencer les transcriptions du jour ; il reste ouvert tout le temps.

Coller les rubriques

Une fois un nouveau balisage terminé (et sauvegardé), je reviens dans le fichier « header » (par le menu **Window**, où on voit les fichiers CLAN ouverts). Je sélectionne et copie les headers (**contrôle A > contrôle C**), je reviens dans le fichier de transcription en cours, et j'y colle les headers, en début de la transcription (j'insère une nouvelle ligne avant la première balise avec la touche « entrée », je reviens au tout début de cette ligne vierge, et je colle – **contrôle V**).

Taper **@End** après la dernière ligne balisée, et sauvegarder.

Renseigner les rubriques pour chaque sujet

Poser le curseur au début du document ; menu **Edit > Replace** ; remplacer automatiquement les 000 des *headers* par le numéro du sujet sur lequel tu travailles. Rentrer l'âge et le sexe du sujet (informations qui figurent dans le document PAROLE qui s'appelle « **Liste_transcriptions.xls** ») aux bons endroits, vérifier la date de l'enregistrement. Sauvegarder.

Préparer la génération automatique des « *tiers* » (identifiants des participants à l'échange) : menu **Tiers > Update**.

Tu es prêt/e à transcrire.

Transcriptions en « mode sonique »

Identifier & écouter un segment balisé

Pour écouter un énoncé balisé, c'est simple : poser le curseur sur la ligne et faire **F4** (touche fonction en haut du clavier).

Si tu cliques dans la fenêtre CLAN pendant l'écoute, l'enregistrement s'arrêtera (plutôt embêtant pour transcrire !) ; pour redémarrer l'écoute, refaire F4 (tu entendras le segment balisé depuis le début).

Pour fonctionner de façon plus efficace, il faut passer au **Sonic Mode** : menu **Mode > Sonic Mode**. Une barre d'ondes sonores (the *waveform bar*) apparaîtra en bas de l'écran. C'est un outil indispensable.

Sur cette **barre sonique**, tu vois la totalité du fichier son sous forme d'ondes sonores ; tu peux rapidement y situer chaque segment balisé, ainsi que des phrases, voire des mots et des pauses individuels. La petite ligne noire qui se trouve juste au-dessus de la barre sonique, à gauche, donne des informations importantes. Si tu cliques sur cette zone d'information, tu verras qu'elle change, donnant différentes informations avec chaque clique (trois lignes différentes en tout).

Retrouver un segment balisé dans la barre sonique

Pour visualiser un segment balisé dans la barre sonique, poser le curseur à côté d'une balise (matérialisée par un petit point noir) dans la ligne de transcription, et cliquer deux fois. Le segment s'affiche en noir sur la barre (on passe automatiquement à la FIN de l'énoncé ; pour voir le début, il faut utiliser le glisseur en bas de la barre).

Tu peux relancer l'écoute d'un segment balisé en pointant le curseur sur la barre (sur le segment en noir), ensuite faire **Contrôle clique** (avec la souris). Tu peux arrêter le son en cliquant simplement.

Sélectionner et écouter un petit segment

Le véritable intérêt de la barre sonique est de pouvoir écouter des segments *plus petits* que l'énoncé balisé. Avec la souris, sélectionner n'importe quel fragment sur la barre (un mot, un groupe de mots, un phonème, un groupe d'hésitation...), pointer dessus + **contrôle clique**.

Changer le « relief » de la barre sonique

Tu peux augmenter le relief des ondes sonores (essentiel pour s'y repérer et utile pour « visualiser » l'articulation de tel son difficile à capter (on voit bien les « t » finaux, par ex., les bilabiaux...). Pour ce faire, il suffit de cliquer sur la petite barre verticale noir, à l'extrême droite de la barre ; cliquer (une, deux, trois fois) sur la petite case **+V**. La case **-V** diminue le relief des ondes ; à gauche, les cases **+H** et **-H** augmentent ou diminuent l'étendue horizontale des ondes. (Pour les cases **S**, **L**, et **R** ; voir le manuel ; je ne les ai pas utilisés.)

Chronométrer avec la barre sonique

La barre sonique permet également de chiffrer la durée des pauses et des groupes d'hésitation – tâche laborieuse, mais importante pour notre investigation des caractéristiques temporelles de l'aisance et la disflue dans la production orale en L2. Voir la section « *Hésitations & pauses* ».

Maintenir intactes les balises « complètes »

Dans la fenêtre de transcription, en mode simple, la balise est matérialisée par un petit point noir.

Fais la touche **Esc** en même temps que la lettre **a** : les valeurs numériques des balises ainsi que la référence du fichier son apparaîtront – c'est le **balisage complet**. (On peut aussi passer par le menu **Mode > expand bullets**). C'est une commande bascule : si tu refais **Esc a** les chiffres disparaîtront/ la balise est réduite au simple point noir (ou bien, **Mode > hide bullets**).

Les perfectionnistes se serviront du mode d'affichage complet, pour rafistoler les débuts/ fins de balises (voir ci-dessous). Mais même les approximatifs devront se rendre compte qu'il **ne faut jamais effectuer des touches clavier** (retour, espace, etc.) **quand le curseur se trouve à droite d'une balise simple** (mode d'affichage non-expanded). Car si tu le fais, tu fractionneras la balise, et le fichier son ne sera plus associé à la transcription à cet endroit.

TRES IMPORTANT DONC : Pour l'insertion des lignes dépendantes (%err, %act...), il faut insérer la nouvelle ligne **à la fin d'une balise complète**. Procédure conseillée : cliquer plutôt au début de la ligne principale qui *suit*, avant de faire *entrée*.

Rafistoler les balises

Hypothèse scientifique importante : que le *rapport temps / mots énoncés* est un indicateur précieux de l'aisance – de la maîtrise communicative (ou du moins linguistique). Il est donc important de faire un balisage très précis des énoncés de chaque sujet, car CLAN va calculer la durée des énoncés à partir de ces balises temporelles.

Les balises posées lors de la phase initiale du balisage (en **Mode > Transcribe sound or movie**) constituent de bons repères pour la transcription. Tu verras assez rapidement que – les temps de réaction étant ce qu'ils sont – tes balises tombent souvent un peu après la fin ou avant le début d'un énoncé.

Tu vas aussi te rendre compte que ce que tu prenais pour un seul énoncé lors du balisage initial en comporte en réalité deux, etc. Il faudra donc pouvoir repositionner certaines balises, en insérer d'autres, en se servant de la barre sonique.

Élargir ou réduire un segment balisé (sur la barre sonique)

Sélectionner le segment, en double-cliquant à côté de sa balise dans la transcription ; le segment s'affichera en noir sur la barre sonique.

- Sur la barre sonique, utiliser le glisseur pour te positionner en début du segment.
- En enfonçant la touche **Maj**, cliquer là où tu veux que le segment débute ; il s'étendra automatiquement à cet endroit.
- Utiliser le glisseur pour te positionner à la fin du segment.
- Utiliser **Maj clique** pour repositionner la fin du segment.
- Revenir (sans cliquer ailleurs sur la barre sonique !) dans la transcription.
- Sélectionner la balise (simple).
- Fais **Ctrl i**, et l'ancienne balise (complète) sera automatiquement remplacée par les nouvelles valeurs pour le segment.

Ce même principe (sélection d'un segment sur la barre sonique ; insertion de la balise correspondant en tapant **Ctrl i** dans la ligne de transcription) te permettra de scinder des énoncés, de peaufiner les plages attribuées au sujet et à l'interviewer, etc. Il te permet également de rétablir un balisage effacé accidentellement, etc.

Élargir, réduire ou fusionner balises (dans la transcription)

A certains moments, il est moins fastidieux de changer les valeurs d'une balise directement :

- **Mode > expand bullets** ; on voit les valeurs temporelles de chaque balise (référence en millisecondes au fichier son).
- On peut copier-coller la valeur finale du segment précédant comme valeur de départ du segment suivant, pour que les valeurs de deux énoncés se suivent précisément, par exemple.
- Ce mode est utile si on veut fusionner deux segments balises en un seul énoncé.

Les balises complètes sont également utiles en cas de problème de « dialogue » entre la transcription et le fichier son (suite au copiage des fichiers, il peut y avoir des choses à changer dans la partie de la balise qui y réfère.). On peut changer l'intitulé du fichier son dans les balises par une opération « Rechercher – remplacer », par exemple.

Résumé des commandes de la barre sonique

Ctrl clique	= j'entends ce qui est sélectionné sur la barre sonique (en pointant sur la sélection)
Ctrl i	= insertion d'une balise (à l'endroit où se trouve le curseur dans la transcription)
Maj clique	= je déplace l'une des extrémités du segment sélectionné (selon la position du curseur sur la barre sonique)

A la fin de la transcription (après vérification sous « Check »), tu peux voir les fruits de ton travail : menu **Mode > Continuous playback**. Chaque énoncé est lié au fichier son ; on le voit en écoutant. Chic !

Délimiter les énoncés dans PAROLE

Dans les consignes qui suivent², nous avons tenté de tenir compte des conventions suivies par les autres membres de la communauté CHILDES, mais nous avons également pris en compte les caractéristiques prosodiques et temporelles des productions de nos sujets.

Le découpage du flux parlé en énoncés conditionne certaines analyses. Tous les cas délicats de délimitation des énoncés furent débattus au sein de l'équipe PAROLE, dans un souci de cohérence entre les transcriptions et entre les langues.

Critères de base

Transcrire chaque énoncé sur une nouvelle ligne. Un énoncé est constitué d'une proposition indépendante + le cas échéant toutes les propositions subordonnées qui en dépendent. Les exemples suivants illustrent les différents cas de figure qu'on transcrit comme un seul énoncé ; ces extraits des transcriptions ont été simplifiés (autour des codes d'hésitation), afin de les rendre plus lisibles (seule la durée des groupes d'hésitation figure ; des ellipses remplacent les reprises).

Proposition indépendante simple

Proposition simple = un verbe conjugué avec ses arguments + éventuellement d'autres expansions directes (adverbes, adjectifs, groupes prépositionnelles).

*004: I can see: [#1_747] a fridge↑ .

032: #0_273 suddenly it fall [] down onto the car↑ .

Proposition indépendante + complétive(s)

Proposition principale + proposition(s) complétive(s) attachée(s) à un élément de la principale.

009: [#1_208] and the man [#0_325] is [#2_125] trying [...] to catch↑ [] the fridge↑ .

023: and [#2_305] a pedestrian or [...] a man in the street [#2_613] is furious to see that the fridge has just bumped [] [...] his car .

Proposition indépendante + autres subordonnées

Proposition principale + une ou plusieurs subordonnées (circonstancielles, relatives).

003: yes [#1_823] [...] I have seen [#1_881] mans [] [#3_263] who were [#1_979] at the window .

*418: #0_522 mais [...] quand les hommes essaient [...] de: prendre le frigidaire [...] [#0_528] pour le faire [#0_214] entrer [...] dans la maison↑ [#0_859] le frigidaire il tombe .

Cas particuliers

A l'oral, des propositions syntaxiquement indépendantes peuvent être produites sans aucune discontinuité perceptible. Pour conserver cette caractéristique de l'oral et éviter des découpages artificiels, nous transcrivons en un seul énoncé les cas illustrés ci-dessous.

Effacement du sujet

Transcrire comme un seul énoncé lorsqu'une proposition est suivie d'une ou plusieurs autres propositions (généralement coordonnées) dont le sujet est effacé.

*N02: #0_261 but [/] [#0_319] and then it fa:lls a:nd lands on some guy's car .

*N13: #0_540 a:nd [#0_621] then the elephant hits (h)im and [#0_384] eats the Rolo himself and says +"/.

² Rédigées par John Osborne.

Autres propositions coordonnées

Proposition coordonnée = deux propositions simples liées par une conjonction de coordination (and, but, or... / et, mais, ou... / e, ma, o...)

Ce sont les cas les plus difficiles à résoudre. La solution classique consiste à transcrire chaque proposition sur une nouvelle ligne, et donc de traiter les conjonctions de coordination comme marquant les frontières des énoncés. Mais à l'oral la fonction des conjonctions de coordination est variable ; il arrive assez souvent que les propositions coordonnées apparaissent comme une seule unité, sans discontinuité perceptible.

Transcrire en un seul énoncé **seulement si les trois conditions suivantes sont remplies** :

1. La conjonction de coordination n'est pas *précédée* d'une pause de plus de 450ms (ne pas tenir compte d'éventuelles pauses après la conjonction) ;
2. la prosodie à la fin de la première proposition indique que l'énoncé n'est pas terminé ;
3. il n'y a pas de rupture thématique majeure entre les propositions.

Exemples : un seul énoncé

Unité thématique, absence d'hésitation avant and et continuité prosodique :

- *019: [...] so: [...] the(r)e are two people↑ [#0_505] looking out of the window↑ a:n(d) they try to catch [*] the fridge and the fridge falls .
- *N12: #0_587 (a)nd there's the guy who's like #0_447 &ge gesturing madly with his ha:nds and the fridge is just ruined and the car is also ruined .
- *N11: #0_459 and he was eating a Rolo #0_529 a:s [...] [#1_655] a: circus parade went by and [...] #0_255 a full+grown elephant [#0_552] slapped him on the face with his trunk !

Exemples : différents énoncés

Changement thématique, présence d'hésitation ou de prosodie de fin d'énoncé :

- *N03: #0_200 and then you see the man when he's older↑ and (1) he's [...] watching a: [#0_226] parade down the street .
- *N03: #0_465 and (2) he gets tapped on the shoulder by this elephant who's now grown up and (3) then he smacks him in the face &=rire with his trunk !

(1) absence de pause + proposition coordonnée à l'intérieur d'une relative = 1 seul énoncé ;

(2) pause de plus de 450 ms avant la conjonction = nouvel énoncé ; (3) absence de pause + continuité thématique (participants identiques à ceux de la proposition précédente) = 1 seul énoncé.

*024: and he's like [#1_927] watching a parade↑ .

*024: a:nd (1) [#1_480] you can see the elephant which [...] happens to be the same one as the one [#0_302] that was in the zoo↑ +/.

(1) absence de pause avant la conjonction, mais discontinuité thématique (aucun élément commun avec la proposition précédente) = nouvel énoncé

Propositions concaténées (run-on sentences)

Deux (ou plusieurs) propositions simples juxtaposées sans conjonction de coordination et sans hésitation. Coder en un seul énoncé seulement s'il n'y a aucune hésitation à la jonction des propositions.

- *019: [...] uh the boy (h)as grown up obviously:↑ [*] now he's a man↑
- *N12: #0_726 and they almost got it in it was very [?] [*] up to the window [...]
- *N47: [#0_696] e:t elles n' y arrivent pas tout à fait [...] le frigo [...] #0_517 tombe sur une voiture↑ .

Incises

Propositions contenant des parenthèses ou des apartés, avec retour au thème de départ :

- *N47: alors donc <on voit [...] un enfant> [/] [#0_575] donc uh &je bon alors là je sais plus [...] uh si ça faisait ancien ou non↑ mais en tout cas on voit un enfant [#0_459] qui mange donc une [#0_307] barre chocolatée on sait pas trop ce que c' est exactement↑ .

Transcrire pauses et hésitations

Chronométrer toutes les hésitations de > 200ms de durée.

Pauses seules

Il existe 2 types de pauses : silencieuses, vocalisées.

Les pauses silencieuses seules (SP – *silent pause*)

Les pauses silencieuses seront marquées par le symbole #, directement suivi de la durée de la pause (en secondes_milisecondes, selon les conventions CHILDES, p. 61) : par exemple #0_956 ; #1_023. [CLAN « bug » un peu avec les zéros qui se trouvent directement après le soulignement ; nous tenterons d'alerter Carnegie Mellon et de résoudre le problème.]

Pour mesurer la durée d'une pause, voir ci-après.

Si ce qui semble être une pause silencieuse dure moins de 200ms, ne pas le coder, ni le transcrire. (Souvent c'est soit la fin d'un allongement vocalique, soit une petite préparation à l'articulation du son suivant.)

Les pauses vocalisées (FP – *filled pause*)

Les pauses vocalisées (ou remplies) seront transcrites selon la liste suivante. Nous avons élargi la liste des « fillers » dans CHILDES ; les remplisseurs marqués d'une étoile doivent être rajoutés au depfile avant de procéder aux analyses dans CLAN.

- uh** tout son vocalique central
- um** son vocalique avec finale bilabiale /m/ ou /n/ (visible dans Sonic Mode)
- er** consonne liquide à la fin d'un son vocalique central
- eu*** son vocalique plus arrondi et plus avant, comme en français
- eh*** son vocalique plus fermé et plus avant que « uh » ; caractéristique des productions des sujets hispano- et suédophones
- em*** idem « eh », avec finale bilabiale

Si la pause vocalisée dure moins de 200ms, la transcrire, sans mesurer sa durée.

Si la pause vocalisée dure plus de 200ms, le transcrire, marquer la longueur vocale avec le symbole « : » immédiatement après la voyelle (**u:m u:h eu:h e:h**) et indiquer ensuite sa durée, entre crochets :
409: quand il **eu:h** [#0_296] Oobj met↑ [*]

NOTE : système à améliorer pour de futurs travaux, car il ne distingue pas les pauses remplies seules des groupes d'hésitation (ci-dessous).

Autres phénomènes d'hésitation

Bruitages paralinguistiques

- &=bouche** claquement de la langue, « pff », autre bruitage paralinguistique meublant une difficulté de traitement langagier (« événement local simple », manuel CHAT, 59)
- ahem** raclement de gorge communicatif ; doit figurer dans un groupe d'hésitation (au même titre qu'un bruitage paralinguistique), si équivalent à une pause remplie
- hum** variant anglophone de « um » ; compris dans les groupes d'hésitation
- well, ben...** par contre, les petits mots remplisseurs seront transcrits mais non pas chronométrés (car pas vraiment paralinguistiques...)

Ne pas coder les **inspirations** (*intake of breath*), qui figureront parmi les pauses silencieuses (#).

Raisonnement : il est facile d'entendre une inspiration en début d'énoncé chez certains sujets, mais

beaucoup plus difficile à d'autres moments/ pour d'autres sujets. Mieux vaut donc « banaliser », en ne pas codant spécialement ce phénomène, et inclure le temps de respiration dans le temps total d'hésitation. Les chercheurs s'intéressant à ce type de phénomène sont invités à nous contacter avec leurs suggestions de codage.

Allongements (voyelles & consonnes)

Les allongements syllabiques (prolongation non-phonémique d'une voyelle (ou consonne) constituent également un phénomène d'hésitation à coder :

- : le symbole suit directement la voyelle (consonne) qui dure plus que 200ms
- :: (suggestion pour de futures transcriptions ; ceci n'a pas été fait dans PAROLE)
utiliser le double symbole quand l'allongement d'un phonème dépasse [?] 600ms

Faut-il chronométrer certains allongements prolongés ? C'est un vrai phénomène d'hésitation, notamment en production francophone ; mais il faut encore en débattre (et il ne faut surtout pas en abuser, car le sujet est quand même en train de produire du langage). Voici quelques pistes :

Allongements : mots référentiels

- coder l'allongement (phonème qui dure >200ms) avec le symbole « : »
- ne pas chronométrer ces allongements.

Allongements : mots grammaticaux

Deux cas de figure (système de codage perfectible...) :

- **MG allongé suivi d'une pause** : si la durée du MG dépasse 500ms, inclure la dernière partie du mot (là où la courbe acoustique s'aplatit, comme un « vrai » remplisseur, tel *eu*h) dans le temps total de la pause qui suit. [Procédure qui ne distingue pas entre ce cas de figure et les pauses réellement silencieuses.]
- **MG allongé suivi directement d'un mot** : ne chronométrer que si la durée totale du MG est de >800ms ; dans ce cas, mettre la durée de la partie finale (aplatie) de l'allongement entre crochets après le MG [Le codage donc ressemble à celui des groupes d'hésitation, malheureusement, et des pauses vocalisées longues.]

Dans les deux cas, faire attention de ne pas surestimer l'hésitation que représente cet allongement, car l'articulation d'un mot grammatical représente surtout un moment de production.

Groupes d'hésitation

Le **groupe d'hésitation** (GH) : un ensemble d'au moins deux phénomènes d'hésitation, qui s'enchaînent, ininterrompus par la production (ou une tentative de production) langagière – pause silencieuse suivie d'une pause remplie, suivie d'un bruitage paralinguistique, par exemple.

Dans PAROLE, de tels enchaînements ont été regroupés (« *scoped* ») dans une seule entité complexe chiffrée, avec crochets <> autour du groupe, suivis de la durée totale du groupe entre crochets [] :

*025: and in the end <# uh> [#0 354] the: fridge f:ell #1_103 on a car &=rire .

*021: <u:m # u:h # u:m # &=bouche # u:m uh #> [#11 049] <l lack the vocabulary> ["] !

Ce système nous permet de chiffrer plus aisément le temps total d'hésitation pour chaque production, ainsi que d'identifier des hésitations disfluentes (dépassant le seuil communicatif fatidique de 2,5 ou 3 secondes).

Ne pas inclure les fragments (&f &eb) dans les groupes d'hésitation, car ils constituent des tentatives (même ratées) de production.

Inclure les rires nerveux (&=rire signifiant « je tente de trouver le langage qu'il me faut ») mais non pas les rires comiques dans les GH. En cas de doute, demander un 2^e avis.

Chronométrer la durée d'une pause ou d'un groupe d'hésitation

C'est assez laborieux, mais il faut minutieusement chronométrer les pauses en transcrivant dans CLAN. Il y aura des vérifications par la suite du chiffre rentré.

En **Mode sonique** (Menu **Mode > Sonic Mode**) – où tu vois la barre sonique en bas de l'écran de CLAN, sélectionner une pause ou un groupe d'hésitation avec le curseur sur la barre sonique.

Sur la ligne d'information noire (en haut, à gauche) de la barre sonique, cliquer une fois.

La valeur temporelle (à la milliseconde près) s'affiche dans la deuxième partie de la ligne, après la lettre « D » : D 00:00:04:585 par exemple (pour un groupe d'hésitation qui dure 4,585").

N'oubliez pas que dans la ligne principale de la transcription, il faut toujours inclure les secondes (même si la pause dure moins de 1000ms), de la façon suivante : #0_858 ou #2_655.

IMPORTANT : ne pas surestimer le temps d'hésitation

Ne pas mesurer pile jusqu'à la courbe où le prochain mot commence ; laisser une minuscule zone-tampon (quelques millisecondes) après le dernier mot et avant le prochain.

Il faut aussi s'assurer que l'on ne prenne pas pour silence ce qui est en effet un début ou une fin d'articulation, à cause d'un manque de relief dans la barre sonique. Il faut s'assurer donc qu'il y ait suffisamment de relief (case +V à droite de la barre sonique), pour voir si certaines articulations (consonnes dentales ou fricatives, surtout, /s/ ou /f/, /t/ final) ne prolongent la production d'un mot. Il faut, bien sûr, éviter d'inclure ces sons dans la durée des pauses.

Par contre, on inclut les inspirations (et autres préparations à l'articulation) dans nos temps d'hésitation (car c'est impossible de distinguer dans tous les cas entre la préparation à l'articulation, et une vraie hésitation).

Transcrire les reprises

Il n'y a pas que les pauses et hésitations qui signalent un manque d'aisance en production orale, il y a aussi les répétitions, reformulations, etc. – que nous appellerons globalement « reprises » (*retracings*).

Codage des reprises dans PAROLE

MacWhinney identifie les reprises suivantes, et quatre symboles de transcription (2006 : 71sv.) :

[/] – retracing without correction

[//] – retracing with correction : changes the syntax but maintains the same idea ; usually moves closer to the standard form (but can move away from it)

[///] retracing with reformulation : full and complete reformulations of the message ; when none of the material being corrected is included in the retracing

[/-] – false start (without retracing) : on commence un nouvel énoncé

Ces définitions donnent l'impression que chaque reformulation génère une correction, ce qui n'est pas le cas. A la page 73 (« *false start without retracing* »), MacWhinney admet que l'utilisation des symboles dépend de ce qui est ciblé dans le corpus. Nous avons donc développé un système légèrement différent, pour mieux refléter les caractéristiques observées dans PAROLE.

[/] = répétition simple, sans changement

Reprise à l'identique d'un ensemble (groupe de phonèmes, mot, groupe de mots). Le matériel repris sera regroupé (entre <>) s'il comporte plus d'un mot.

La répétition peut concerner :

- un mot grammatical
- un mot référentiel
- un groupe de mots (les regrouper avec <>)

Ne pas coder le bégaiement/ la répétition d'un fragment phonologique avec [/] : &f &f frigo

ATTENTION : les **dédouplements augmentatifs** ne sont pas des reprises (= hésitation/ disfluence) :

- *il est très très en colère*

[//] = reformulation simple

Répétition, avec un seul changement ; deux cas de figure :

1. reformulation linguistique :

- changement phonologique
- changement lexical
- changement morphologique
- changement syntaxique

2. reformulation sémantique : enrichissement (ou réduction) du contenu sémantique de la proposition

- *he sees <an elephant> [///] a baby elephant .*

Dans de futurs corpus, il pourrait être intéressant de prévoir un codage spécifique pour ces reformulations sémantiques (afin de pouvoir les retrouver rapidement), par exemple [//s]

[///] = redémarrage (restart) [reformulation syntaxique]

Changement de deux éléments (ou plus) dans l'expression de l'idée qui est reprise. Une partie de la syntaxe change ; une partie de l'énoncé est maintenue (des mots ou la structure syntaxique de base) :

- *il veut essayer de lui donner [///] enfin l'éléphant croit qu'il va Opro donner*

[/-] = abandon (false start)

La syntaxe initiale est abandonnée, pour une nouvelle structure syntaxique de l'énoncé.

Différence avec l'auto-interruption [+//.] : le parleur ne change pas de sujet.

- *he er there's [-] the elephant is going down the side of the street*

NOTE : Pour [///] et [-] ce n'est pas nécessaire de « scoper » ce qui est repris (car souvent long et/ ou compliqué).

Reprises & erreurs

Nous avons décidé de ne pas coder les erreurs qui se trouvent dans la première partie d'un segment repris (avant la répétition ou la reformulation). Ceci simplifie la transcription un peu, et évite de surestimer le nombre d'erreurs produites par le sujet.

- Coder l'erreur après la **répétition**, ou une erreur dans la deuxième partie d'une **reformulation** (le cas échéant).
- On peut coder les erreurs dans un segment qui est ensuite **abandonné**.

Résumé des conventions CHAT utilisés dans PAROLE

Voici le document de travail que nous avons utilisé lors de nos transcriptions : il constitue une synthèse des codes les plus utilisés. Il peut servir à d'autres utilisateurs (francophones) de CHAT et, bien sûr, aux chercheurs investiguant PAROLE.

Ce résumé ne remplace pas le manuel CHAT, qui doit être consulté régulièrement par les transscripteurs, et qui répond aux cas non-traités ici.

Pose de balises (= segmentation du fichier-son) : RAPPEL

Rappel des fonctions de base, une fois les balises posées :

- Écouter un segment balisé : poser le curseur dans la ligne de transcription + **touche F4**.
- Pour voir la barre sonique : menu **Mode > Sonic Mode** (ou **Esc 0**).
- Trouver un segment balisé sur la barre sonore : double-cliquer à côté d'une balise.
- Rafistoler des balises : trouver le segment sur la barre sonique, **Maj clic** pour étendre le segment à gauche, ensuite à droite.
- Poser manuellement des balises : sélectionner un segment sur la barre sonique, positionner le curseur à la fin de la ligne de transcription concernée, et faire « **contrôle i** ».
- Pour voir les valeurs des balises : « **Esc a** » (possibilité de modifier chiffres).

En-têtes

Après chaque rubrique de l'en-tête un : suivi d'une **tabulation**.

Exceptions : les commandes @Begin et @End ne sont suivies de rien (ni : ni espace).

rubrique	code à entrer, selon le cas	précisions
@Begin		aucune ponctuation
@Languages:	en it fr	= anglais = italien = français
@Participants:	021 Subject, INV investigator	numéro du sujet (à 3 chiffres) ; NB présence virgule (ne pas coder l'identité de l'investigateur, car c'est sans importance scientifique)
@ID:	en parole 021 22;00.00 female CAPES Subject	langue corpus n°sujet âge sexe groupe rôle NB : trait vertical obtenu avec les touches "Alt Gr" et "6" (simultanément) ; chaîne sans espaces ; doubles traits autour de "rôle" (car on laisse deux cases vides). L'AGE doit comporter les mois et jours, selon le format indiqué (mettre des 00)
@Coder:	Hilton	nom de famille de la personne qui transcrit
@Language of 000:	fr de	L1 du sujet : mettre le numéro du sujet, et l'abréviation CHILDES pour la langue parlée (p. 25)
@Date:	27-FEB-2006	suivre ce format ; abréviations des mois en anglais : JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
@G:	frigo commentaire_beignets résumé_beignets éléphant image_1 image_2 (etc.) accident	nom de la tâche : (indiquer les deux tâches liées à "La Dame aux beignets")
@End		à la fin de la transcription, aucune ponctuation (mettre en dernière ligne tout au long de la transcription, si tu veux utiliser la fonction "CHECK" en cours de route)

Une fois ces informations rentrées, programmer CLAN pour qu'il génère automatiquement le « code participant » en début de chaque ligne de transcription : menu **Tiers > Update**.

Ensuite, en transcrivant :

- pour obtenir le code sujet : touche **Ctrl** et la touche **&** (même touche que le chiffre 1)
- pour le code investigator ; **Ctrl** et **é** (même touche que le chiffre 2).

IMPORTANT : Penser à vérifier le bon codage de l'identité des participants à la fin de chaque transcription (**mode** « **Continuous playback** »), car il est facile de se tromper de touche en se servant des raccourcis.

Ligne principale

Énoncés

Pour CLAN, l'énoncé est déterminé par l'un des trois marques de ponctuation finale : . ! ?

ATTENTION : ces marqueurs finaux sont toujours précédés d'un espace .

Énoncés incomplets & interrompus* ; recouvrements

+...	" <i>trailing off</i> " ; énoncé abandonné (sans interruption), souvent avec pauses ; le même énonciateur peut reprendre (nouvelle ligne) : *407: et il y a [#0_923] deu:x #0_285 gra:nds éléphants e:t [#0_401] u:n #0_203 petit [#0_726] éléphant [#1_173] qui: <eh # uh> [#1_167] +... *407: il mont(r)e à le [*] petit éléphant qui: [#0_801] venir [*] à: [*] manger [#0_546] les bonbons .
+..?	" <i>trailing off</i> ", de type interrogatif
+//.	" <i>self-interruption</i> " ; énoncé abandonné brusquement par l'énonciateur, sans pause, pour en commencer un autre
↑ .	prosodie montante. Pour obtenir la flèche : ouvrir la fenêtre des caractères spéciaux – menu Window > Special characters . Pour l'insérer : placer le curseur à la fin du mot (où l'intonation monte) dans la transcription ; double-cliquer sur la flèche dans la fenêtre des caractères. Version 2008 de CLAN : placer la flèche à la fin du mot, sans espace.
+^	" <i>quick uptake</i> " ; en début de ligne quand un énoncé suit très rapidement l'énoncé précédant
++	quand un énoncé abandonné est complété par qq'un d'autre (en début de cette ligne, suivi d'un espace) *MOT: si j'avais su +... *FAT: ++ tu ne serais pas venue .
+/.	énoncé incomplet INTERROMPU par qq'un d'autre
+,	reprise (par l'énonciateur original) de l'énoncé interrompu (en début de ligne, suivi d'un espace) : *417: il [*] a: revendiqué [*] non ["] +/. *INV: oui . *417: +, [#0_673] sa vengeance sur lui↑ +/.
+<	" <i>lazy overlap</i> " ; marqueur sommaire de recouvrement, quand deux morceaux de discours sont prononcés en même temps (suffisant pour nos besoins dans PAROLE). *418: le frigidaire il tombe +/. *INV: mh ! *418: +< sur une voiture .

* Pour les **mots** incomplets/ interrompus, voir ci-dessous (« **Hésitations...** »)

Citations, commentaires métalinguistiques

+"/. +"	citation directe (fréquente dans le récit du "frigo", pour l'homme qui crie) : *025: he was um raising his hands and he was um screaming +"/. *025: +" what's happened to my car ?
['']	commentaire métalinguistique (utiliser <> s'il s'agit d'un groupe de mots) Utiliser ce marqueur chaque fois que le sujet commente ses choix linguistiques : *019: [''] [#2_305] so he comes up to: [#1_106] the elephant [#2_107] <how can I call it> [''] #0_581 field +...

Ponctuation

Pas de ponctuation interne aux énoncés!

En FRANCAIS, il faut ABSOLUMENT **mettre un espace après chaque apostrophe** d'élision. Sinon, toutes les analyses de CLAN seront faussées !

*989: et là il **n'**est pas content .

*989: je pense que **c'**est un oiseau .

*989: et là l' oiseau **s'**est envolé .

Si la virgule fait partie d'un mot, comme **aujourd'hui**, ne pas mettre d'espace. Par contre : **d'+accord** ; **d'+ailleurs** ; **d'+après** (selon la « bibliothèque » de CLAN).

En ANGLAIS, on peut transcrire les contractions normalement, sans espace.

I'm she's he's they're you're it's don't can't won't wouldn't aren't ...

Pas de trait d'union dans les inversions, non plus : **n' est ce pas dit il vas y a t il dit**

Dans les mots composés, utiliser + à la place du trait d'union (ou espace) : **parce+que peut+être nah+nah+nah+nah+nah**

Les majuscules sont réservées aux noms propres (**Rolo**) ; pas de majuscule en début d'énoncé.

En anglais, on peut mettre le pronom « I » en majuscule.

Si jamais nous en avons, les sigles seront transcrits selon leur prononciation : **E_T** si chaque lettre est prononcée ; **Fifa** si prononcé comme un mot).

S'il y a des chiffres dans le discours, les écrire en toutes lettres, sans trait d'union : **cela s' est passé en mil neuf cent quatre vingt huit .**

Hésitations, reprises, approximations, mots incomplètes, bruitages...

#	Voir la section sur la transcription des pauses
um uh er	(<i>hésitations vocalisées/ pauses remplies</i>) : se limiter aux transcriptions suivantes : um (présence d'une finale bilabiale), uh (tout son vocalique central), eu (pauses remplies à la française, avec voyelle plus fermée et avancée que la valeur centrale "uh", et arrondissement des lèvres), er (présence d'une liquide), eh , em (pauses remplies hispanophones et suédophones, voyelle proche de /e/)
mmhm uhhuh mh	bruits d'assentiment
[/]	simple répétition (Voir la section sur la transcription des reprises)
[//]	reprise avec un seul changement (Voir la section sur la transcription des reprises)
[///]	reprise avec plus d'un changement (mais éléments syntaxiques maintenus)
[-]	"redémarrage" : abandon de la syntaxe initiale (si l'amorce ne constitue pas un énoncé abandonné – qui serait codé +... ou +//.)

< >	<p>"scoping" (regroupement) : Attention! Si un groupe de mots est repris, il sera placé entre crochets pointus (pour permettre à CLAN d'effectuer ses analyses correctement) ; si un seul mot est repris, pas besoin de crochets :</p> <ul style="list-style-type: none"> reprise d'un seul mot *021: there' s someone # putting # sugar or flour on [/] on a donut reprise d'un groupe de mots *021: it' s in a white building # and <the fridge> [/] the fridge is # um going up . reprise avec changement *021: and <she now> [/] now she adds chocolate .
------------------	---

Quand une reprise comprend des hésitations, on met le symbole de reprise AVANT les symboles d'hésitation :

*019: a:nd &th there is a man on [/] #0_982 on the: [/] #0_702 the pavement .

&	symbole qui signale à CLAN de ne pas prendre en compte ce qui suit (dans l'étiquetage et certaines analyses).
&=rire	transcription de rires (invariable)
&=bouche	transcription de tous les bruits paralinguistiques, sans distinction (sauf "ahem") (claquement de la langue, bruit de lèvres, "pff", reniflement, etc.)
&=cherche: aide	possibilité pour les "événements locaux" – quand sujets demandent un mot de vocab par le regard? (p. 60)
&(lettre)	<p>bégaiements, fragments phonologiques (sublexicaux) :</p> <p>003*: and now the &f &f frigo [*] is going up .</p> <p>(représenter le son répété selon UNICODE, ou meilleure approximation alphabétique)</p>

@	formes intermédiaires
@fs	<p><i>filler syllable</i> : "mots pivots", caractéristiques de la production en L1 (18-36 mois) – et sans doute de l'AL2 – formes approximatives qui deviendront/ remplaceront des mots grammaticaux.</p> <p>Utiliser @fs pour les formes intermédiaires (concernant les mots grammaticaux seulement)</p> <ul style="list-style-type: none"> the@fs – forme intermédiaire entre <i>the/ that/ this</i> (anglais) le@fs – forme intermédiaire entre <i>le/ la/ les</i> (français) un@fs – entre <i>un/une</i>, etc
@n	<p>néologisme, c'est-à-dire, création de mot référentiel :</p> <ul style="list-style-type: none"> I think it's a frigo@n [phonologie anglophone] ; he gives him a # giffel@n.
@l1	<p>après chaque mot en L1/ langue maternelle :</p> <ul style="list-style-type: none"> I think it is a [/] a mover [/] remover um je@l1 sais@l1 pas@l1 .
@s	après chaque mot dans une langue autre que la L1 ou la L2-cible.

xx	mot incompréhensible
xxx	groupe de mots incompréhensibles
[?]	interprétation plausible d'une forme produite peu claire
()	<p>autour d'un segment absent</p> <ul style="list-style-type: none"> N41: il tombe sur la voiture qu' i(l) y a juste en dessous
www	phrase/ énoncé(s) sans importance, que l'on ne transcrit pas. Mettre une ligne de commentaire résumant ce qui se passe (si vraiment nécessaire, pour le sujet seulement).

Erreurs

<p>[*]</p> <p>< ></p>	<p>Signale d'erreur dans la ligne principale (un mot entier, un groupe de mots, phonème ou accent tonique). La nature de l'erreur sera systématiquement indiquée sur la ligne secondaire (voir cette section).</p> <ul style="list-style-type: none"> • um and then # um on [*] the film in the street # um there's a man [au lieu de in] <p>ATTENTION, si l'erreur implique plus d'un seul mot, il faut mettre le groupe de mots concernés entre < > :</p> <ul style="list-style-type: none"> • 021: and she<'s always pushing> [*] a button . [au lieu de keeps pushing]
erreurs particulières : (coder pour le bon fonctionnement de MOR/ POST)	
()	autour d'un segment (de mot) absent ; transcrire un vrai fragment avec & (phonème), ci-dessus.
<p>0</p> <p>(chiffre zéro)</p>	<p>devant un mot qui manque (important, sinon CLAN bloquera dans l'étiquetage morphologique) :</p> <ul style="list-style-type: none"> • I want 0prep go (pour l'énoncé "I want go") <p>Pour éviter trop d'extrapolation, coder le type de mot qui manque seulement (selon abréviations de CLAN), par exemple :</p> <ul style="list-style-type: none"> • 0v verbe manque • 0aux auxiliaire manque • 0det article/ déterminant manque • 0subj sujet manque • 0obj complément d'objet manque • 0pro pronom manque • 0prep préposition manque ; etc.

Lignes secondaires

La ligne secondaire concerne systématiquement un seul énoncé. Elle est placée immédiatement après l'énoncé en question, et commence par le symbole % suivi d'un code à trois lettres. Après le code, deux-points tabulation (comme tout début de ligne).

Nous allons nous limiter (sauf débat et accord du groupe) aux lignes secondaires suivantes :

%com :	commentaire du transcripteur
%act :	décrire une action ou un geste (avec signification convenue, soit en L1 soit en L2)
%err :	<p>reprend l'erreur et donne sa correction ; l'erreur est systématiquement codée (par une abréviation après le symbole \$; voir ci-après) :</p> <p>*021: um and then um on [*] the film in the street um there's a man . %err: on = in \$MOR \$PREP</p> <p>Si deux erreurs dans un même énoncé, les séparer dans la ligne secondaire par ;</p> <p>021: and now she doesn't have time to put [*] cream so she put [*] them on the table . %err: put = put on \$LEX \$PHR ; put = puts \$MOR \$AGR</p>

Coder les erreurs

Il y a différents « couches » dans le codage des erreurs.

RAPPEL : Ne pas coder une erreur dans la première partie d'un segment repris (avant la répétition ou la reformulation).

008: #0_476 a:nd u:h [#0_336] the: [] fridge <u h #> [#3_280] fall [/]/1 #0_546 falls .

Coder d'abord le niveau d'erreur

Se limiter aux catégories d'erreur suivantes. Veillez à bien analyser chaque erreur ; on peut débattre ensemble (même en petit comité) de la nature d'une erreur compliquée à étiqueter (car il y en a dans chaque enregistrement). Dans la première phase de transcription, on peut se limiter au seul « niveau d'erreur », sans détailler le type d'erreur.

Niveaux d'erreur : pour toute erreur, coder d'abord l'une de ces 5 catégories. Un mot peut, bien sûr, comporter plus d'un niveau d'erreur (\$PHO et \$MOR, par exemple).

\$PHO	erreurs de prononciation, que ce soit au niveau des sons ou de la prosodie (Note : dans le manuel CHILDES, les erreurs d'intonation sont indiquées comme un autre niveau d'erreur)
\$LEX	erreurs dans le choix d'un mot référentiel , (nom, verbe, adjectif, adverbe) comprend les erreurs au niveau des préfabriqués (tel <i>put on</i> ci-dessus)
\$MOR	erreurs dans la forme d'un mot référentiel (affixes), dans le choix du temps ou de la forme aspectuelle d'un verbe ; erreur concernant tous les mots grammaticaux (<i>function words</i> - déterminants, prépositions, conjonctions, pronoms, auxiliaires ☹). [Pour de futures transcriptions, nous aimerions proposer une distinction plus rigoureuse entre erreurs de morphologie inflexionnelle et erreurs autour des mots grammaticaux.]
\$SYN	erreurs dans l'ordre des mots : <i>she explained me it</i> (au lieu de <i>she explained it to me</i>) ; <i>une blanche voiture</i> (= <i>une voiture blanche</i>), etc. Comprend mots manquants [0subj , par ex].
\$REF	erreurs au niveau du contenu référentiel d'un résumé

Coder ensuite le type d'erreur

Pour chaque « niveau » d'erreur, il y aura différents « types » d'erreurs. La plupart de ces codes sont pris dans le manuel CHILDES (en mélangeant un peu les niveaux de codage) ; certains sont des nouveautés (marqués avec *). Ce codage peut se faire dans la deuxième phase de transcription ; encore une fois, faites une liste des erreurs dont le codage est à discuter en équipe.

Il peut y avoir des erreurs sans codage de type ; il peut aussi y avoir 2 (voire 3) codes pour le type d'erreur, dans certains cas.

\$PHO	\$VOW	erreur de son vocalique
	\$CON	erreur de consonne
	\$CC	erreur dans l'articulation d'un groupe de consonnes (<i>pemier</i> = <i>premier</i> , 407a)
	\$INT	erreur d'intonation [code « niveau d'erreur » dans le manuel]
	\$STS	erreur d'accentuation (de mot)
	\$SYL	erreur dans le nombre de syllabes
	\$ELI	erreur d'élision ou de liaison
	\$CWFA	peut signaler un effort d'articulation particulier
	\$SEM	erreur phonologique qui peut générer un malentendu sémantique
\$LEX	\$CWFA	« <i>complex word finding attempts</i> » (recherche de mot laborieuse)
	\$DER	erreur de morphologie dérivationnelle (<i>bravitude</i> au lieu de <i>bravoure</i>)
	\$PHR*	erreur phraséologique ou collocationnelle : absence ou choix erroné d'un (ou plusieurs) élément(s) d'un préfabriqué (au sens large, y compris <i>phrasal verbs</i>)
	\$L1*	mot utilisé calqué sur la L1 du sujet

	\$L3*	mot utilisé calqué sur une autre LE du sujet (extrapolations intelligentes, où possible)
	\$EMO*	erreur dans le choix d'un mot pour exprimer une émotion (concerne les adjs, surtout)
	\$HYP*	utilisation d'un hyperonyme (mot « générique ») à la place d'un mot plus spécifique (<i>animaux</i> à la place de <i>chameaux</i> , <i>nez</i> à la place de <i>trompe</i>)
	\$VCP*	erreur de verbe pronominal (français, italien)
\$MOR	\$AGR	erreur d'accord (nom, adj, verbe, dét, pronom...)
	\$ASP	erreur aspectuelle (perfectif/ simple/ imperfectif...)
	\$TNS	erreur de temps (futur/ présent/ passé)
	\$NFL	autres erreurs inflexionnelles
	\$PREP	erreur dans le choix (ou la forme) d'une préposition
	\$PRO	erreur dans le choix (ou la forme) d'un pronom
	\$CONN	erreur dans le choix d'un connecteur (mots de liaison)
	\$CONJ	erreur dans le choix d'une conjonction (de subordination, etc.)
	\$DET	erreur dans le choix (ou la forme) d'un déterminant
	\$AUX	erreur dans le choix (ou la forme) d'un auxiliaire (<i>avoir</i> au lieu d' <i>être</i> , par ex.)
	\$MOD	erreur dans le choix d'un modal
	\$CWFA	recherche laborieuse d'une forme inflexionnelle/ d'un morphème
\$SYN	\$POS	erreur de position
	\$NP	erreur dans le groupe nominal
	\$CONJ	absence ou présence inutile d'une conjonction
	\$REL	erreur dans l'ordre des mots dans une proposition relative
	\$L1*	structure syntaxique calquée sur la L1 du sujet
	\$PHR*	erreur phraséologique
	\$VC*	erreur de complémentation du verbe
	\$VCP*	erreur de verbe pronominal (français ; italien)
	\$FOC	erreur de focalisation
\$REF*	\$PRO	erreur ou ambiguïté dans le référent d'un pronom
	\$DIS*	erreur de type discursif
	\$SEM*	erreur de type sémantique (le sens ne correspond pas à ce qui fut montré)
	\$COMP*	erreur dans la compréhension d'une séquence, par exemple

Tout ceci concerne la forme des mots référentiels, surtout... peut concerner les articles, en FR et ITA

Un * indique un code créé par/ pour *PAROLE*, à rajouter au *deffile*.

On remarquera que certains types d'erreurs reviennent à différents niveaux.

Tenter de faire un codage assez « minimaliste » des erreurs :

- Ne pas proposer (dans la correction, ligne %err) une phraséologie native complètement différente de ce qui a été produit.
- Cibler le mot ou les mots qui ne vont vraiment pas (noyau de l'erreur).

- Ne pas coder comme erreur ce qui découle des choix erronés effectués en amont : dans **une petite éléphant** coder l'erreur d'article (la = le \$MOR \$DET \$AGR), mais non pas de l'adjectif, qui est juste par rapport au choix effectué par le sujet).

AVERTISSEMENT : Il y a une liste (assez exhaustive) de codes d'erreurs dans le manuel CHAT (aux alentours de la page 115). Aux membres de l'équipe de voir quels codes ils utiliseront régulièrement. Tenir l'équipe au courant.

Utilisation du système [: forme correct] (MacWhinney, juillet 2007)

Réserver ce système pour des cas un peu particuliers, où le sujet produit une forme intéressante, que l'on ne veut pas « noyer », en le mettant dans la ligne d'erreur. Pour l'instant, nous l'avons utilisé pour des « *blends* » – formes composites intéressantes, qui semblent dépasser la simple erreur phonologique, lexicale ou morphologique :

010c: [...] this elephan(t) [*] againt↑ [: again] [*] #0_200 but the elephant [...]
 035c: the: piece of chocolate that he was helding [: holding] [*] <in front of> [...]
 407c: il y a: #0_610 u:n délépha:nt [: éléphant] [*] qui [... *forme que le sujet produit 2 fois*]
 409c: je crois qu' il a peut-être: huit ou neuf ant [: ans] [*] .

NOTE : La forme qui se trouve immédiatement devant l'ensemble "[: " n'est PAS reconnu par CLAN, dans une recherche par "kwal", par exemple. Donc, transcription à limiter à des cas assez spéciaux. On peut ensuite trouver les formes "invisibles" en effectuant la recherche sur "+s":["*]", qui marche.

Les Perles (Gems)

Lors de la transcription et des vérifications des transcriptions, pensez à marquer les « perles » : exemples intéressants de certaines procédures d'encodage, que l'équipe pourra ensuite identifier et analyser grâce aux commandes liées aux « *Gems* » dans CLAN.

Nous avons codé les perles suivantes (juin 2008 – liste qui pourra être augmentée) :

paraphrase

Exemples de compensation lexicale intéressantes.

@Bg: **paraphrase**

005: [#0_383] a:nd [#1_033] the [] man [#1_620] (h)ave [*] <u:h #> [#2_066] [///] it's [/] <u:h # &=snap> [#1_254] it's [/] <u:h # &=bouche> [#1_173] it's [*] not [*] a: [/] #0_777 a boy↑ .

%err: ze = the \$PHO \$CON; (h)ave = has \$PHO \$CON \$MOR \$AGR; it = he \$MOR \$PRO; not = no longer \$LEX \$CWFA

*005: #0_917 <l: [/] <&=rire #> [#0_586] voilà@s> ["] !

@Eg: **paraphrase**

lexsearch

Exemples d'activation ou de recherche lexicales intéressantes.

@Bg: **lexsearch**

023: +, I mean the fridge which <u:m # &=bouche #> [#1_590] co:me [///] <comes up> [] thanks to <u:m # &=snap #> [#1_254] <it's not an elevator> ["] it's <u:m #> [#1_718] <I don't know> ["] .

%err: comes up = is lifted \$LEX \$V \$HYP

*INV: uhuh .

023: +, <# u:m> [#1_097] <it's not a tow> ["] <# # &=bouche #> [#3_362] <no I don't know> ["] [] .

%err: On [tow, elevator] = crane \$LEX \$CWFA

@Eg: **lexsearch**

lexlearning

Le sujet tente d'apprendre un nouveau mot, avec l'aide de l'interviewer.

@Bg: lexlearning

***INV:** <on dit la trompe> ["] .

***415:** <ah on dit <la &tâ> [//]> ["] #0_616 avec le [*] trompe .

%err: le = la \$MOR \$DET

@Eg: lexlearning

phosearch

Tentatives d'encodage phonologique intéressants.

@Bg: phosearch

***407:** [...] peut-être c' est le &pruperit #0_325 &pri: #0_488 propri(é)tai:re [*] uh de la voiture et il a crié .

@Eg: phosearch

morsearch

Opérations morphologiques intéressantes.

@Bg: morsearch

***009:** [...] and u:h [#0_302] he: [/] #0_738 &e #0_200 <he #0_342 eats the chocolate↑> [//] +/.

***INV:** mmhm mmhm .

***009:** +, #1_904 <he ates> [//] #0_331 <he eat> [//] he ate the chocolate .

%err: ates = ate \$MOR \$AGR; eat = eats \$MOR \$AGR \$CWFA

@Eg: morsearch

synsearch

Tentatives de formulation syntaxique intéressantes.

@Bg: synsearch

***017:** [#3_129] [...] (h)e: [#0_540] [/] (h)e [*] sees an elephant↑ and u:h [#0_360] (h)e: [/] #0_934 (h)e [*] calls (h)im↑ [*] [//] #0_366 it #1_086 a:nd [//] <u:m #> [#1_282] for [/] [#0_447] &b <for u:h [#0_255] give> [/] #0_673 for [*] give it <# u:h> [#1_045] a chocolate↑ [*] .

%err: e = he \$PHO \$CON; im = him \$PHO \$CON; for = to \$MOR \$PREP \$SYN \$VC; choc/leit = /chocolate \$PHO \$STS

@Eg: synsearch

Vérification en cours de transcription

N'oubliez pas que dans le menu « Mode », on peut activer une vérification de la transcription :

Mode > Check open file

- recherche des erreurs de transcription ; à faire au fur et à mesure que l'on transcrit.

Il faut avoir déjà rentré l'ensemble « @End » à la fin de la transcription, avant de pouvoir effectuer une vérification.

Garder une liste d'éventuels dysfonctionnements de « Check » ; à transmettre à Carnegie Mellon.

Ce document est proposé
pour faciliter le travail de transcription.
Il ne peut en aucune façon être considéré
comme un remplacement de la dernière version
des manuels pour CHAT et pour CLAN.

Tout transcripteur se doit
de consulter le manuel en cas de doute
ou d'insuffisance de ce résumé !

Résultats

Nous mettrons en ligne la bibliographie de publications présentant les résultats des analyses effectuées dans PAROLE, et nous invitons, bien sûr, d'autres chercheurs en acquisition des L2 de s'y plonger, et de nous tenir au courant de leurs analyses et découvertes.

Conclusions & perspectives

Au fur et à mesure que nous avançons dans les transcriptions, le *Cadre européen commun de référence pour les langues* (CECRL) prend de plus en plus d'importance dans la conception des programmes d'enseignement des langues en France et en Europe. Pourtant, les caractéristiques concrètes – linguistiques, prosodiques, temporelles – des différents niveaux de référence restent à définir, ainsi que la nature des tâches pédagogiques qui amèneront les apprenants au prochain niveau de compétence communicationnelle, de façon efficace.

PAROLE ayant été conçu précisément pour explorer les caractéristiques de différents niveaux de production, nous espérons contribuer aux inventaires nécessaires à la mise en place pratique/concrète du CECRL. Nous espérons bientôt inclure aux rubriques du corpus le niveau de référence européen attribué à chaque participant, selon un panel de juges experts³, pour avancer dans ce sens.

Dans le domaine de la recherche appliquée à la classe de L2, les techniques apprises dans la transcription de ces productions (et notamment la prise en compte des facteurs temporels de la production orale) pourront servir dans des démarches expérimentales plus resserrées, testant les effets de telle ou telle démarche en classe de langues, des séjours à l'étranger, etc.

Pour les besoins de l'évaluation, ainsi que de la recherche, il serait intéressant de développer des outils qui permettront des mesures plus automatiques de l'hésitation en production orale, car les caractéristiques temporelles de la production orale semblent être des indicateurs fiables du niveau d'aisance atteint en L2.

³ Les partenaires linguistiques de l'équipe européen WebCEF s'attendent actuellement à la tâche, et nous espérons bientôt rajouter ces informations aux fichiers en ligne. Pour une description du projet WebCEF, voir : <http://www.webcef.eu/>.