ANR-06-CORP-006

Échange de corpus d'apprentissage multimodaux (MULCE)

Rapport d'activité

Tâche Tmult1 : Multimodal, modèles, transcription

Annexe technique : Multimodal, conventions de transcription et outils de transcription

Modification du 3 novembre 2008 (MC)

Coordinateur de la tâche

Maud CIEKANSKI, LASELDI

Participants

Thierry CHANIER, LASELDI

MarieLaureBETBEDER, Françoise GREFFIER, Christophe REFFAY, LIFC

Anna VETTER, Sandra TOMC, Projet CoPéAs, 2005/Projet Tridem, 2006

1. Table des matières

2. Sous-tâche Tmult1.1 : Modèle transcription

2.1. Historique des phases de transcription

2.1.1. Le Corpus CoPéAs

Une première phase de transcription a été effectuée par Anna Vetter, dans le cadre du projet CoPéAs. Elle a concerné le découpage des sessions de formation qui ont eu lieu dans la plateforme audiographique synchrone Lyceum©[1] en séquences et la transcription des actions dans les modalités audio, clavardage, vote, et des entrées/sorties dans les salles. Ce travail a donné lieu à un premier ensemble de données transcrites en juillet 2005.

Une deuxième phase de transcription a été effectuée par l'équipe du LIFC. La première passe a concerné la définition des espaces-temps et la transcription (1) des entrées/sorties visibles dans les modules où les acteurs réalisent leurs productions (traitement de texte, document, carte conceptuelle, correspondant aux types de modules disponibles dans la dite plateforme) et les actes iconiques de « lever la main » ; (2) les actions visibles ayant lieu dans les espaces de production (pour l'un des deux groupes, celui de TutT), tels que l'édition de texte, la sélection de concept ou la création d'objet. Ce travail a donné lieu à un deuxième ensemble de données transcrites en juin 2006.

L'ensemble du travail de transcription a donné lieu à la rédaction d'une première notice sur les conventions de transcription par Muriel Noras (LIFC) le 5/07/06.

Les conventions de transcription arrêtées en 2005 pour les modalités audio, clavardage, vote et entrées/sorties dans les salles ont été traduites en anglais par Chris Jones (CMU) en avril 2007.

La transcription des actions visibles ayant lieu dans les espaces de production tels que l'édition de texte, la sélection de concept ou la création d'objet pour le deuxième groupe (celui de TutR) a été finalisé en juillet 2008 par Sandra Tomc.

La base de données (format Access et MySQL), finalisée en juillet 2006 et mise en jour en 2008, reprend l'ensemble des données transcrites pour le corpus CoPéAs.

2.1.2. Le Corpus Tridem

Une première phase de transcription a été effectuée par Anna Vetter, dans le cadre du projet Tridem 2006. Elle a concerné le découpage d'une partie des sessions de formation qui ont eu lieu dans la plateforme audiographique synchrone Lyceum© en séquences et la transcription de l'ensemble des actions de communication et des productions. Ce travail a donné lieu à un premier ensemble de données transcrites en octobre 2007 et a enrichi les conventions de transcription pour les données correspondant au mode Parole et aux productions.

Une deuxième phase de transcription a été effectuée par Sandra Tomc sur une partie complémentaire des sessions de formation du corpus Tridem en juillet 2008.

[1] http://kmi.open.ac.uk/projects/lyceum/ et http://lyceum.open.ac.uk/

2.2. Conventions de transcription

(A partir de la « Notice sur les conventions de transcription », document de travail rédigé par Muriel Noras, version du 05/07/06, et du travail de transcription d'Anna Vetter en 2005 sur le corpus CoPéAs). Les exemples donnés dans ce rapport sont issus de la transcription du corpus CoPéAs, complétés par les quelques ajouts de la transcription du corpus Tridem.

2.2.1. Introduction

Ce document présente les conventions générales de transcription et de codage à utiliser lors de toute transcription.

La transcription se fera à partir d'une trame définie pour tout corpus, en utilisant un tableur Excel qui reprend les différentes tables de la base de données (Cf. Fig. 1).

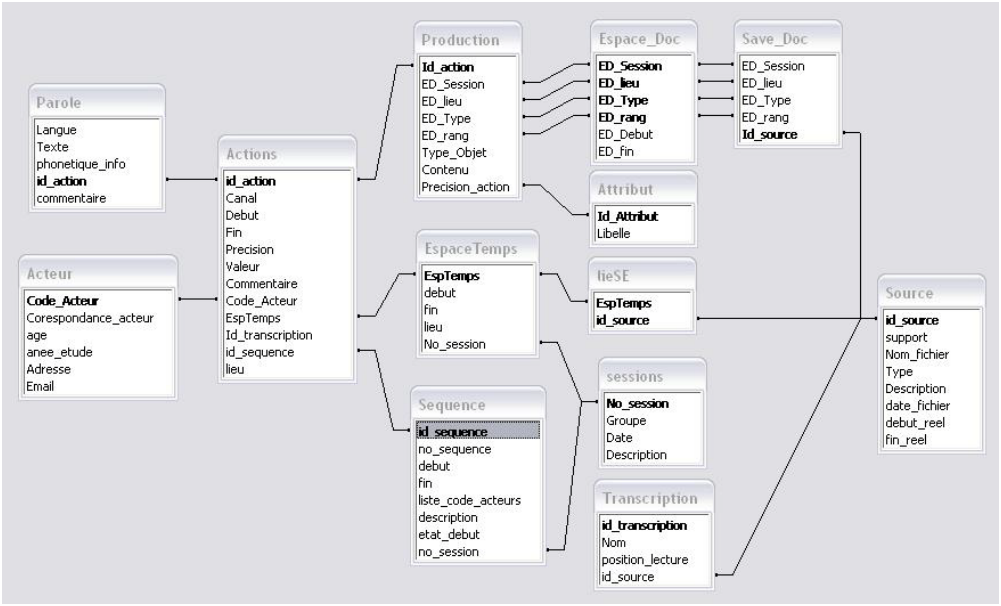

Figure 1 : Schéma de la base de données

Le tableur est organisé comme suit :

- une feuille "actions" qui reprend une partie des différentes informations des tables "action", "session", "parole", "production", "espace-temps" et "source" de la base de données Access ;
- une feuille "espace-temps" pour le découpage en espace-temps (identifiant espace-temps, début/fin et numéro de la session) ;
- une feuille "source" pour l'identifiant source du fichier, le nom du fichier, le type de fichier, le début réel et la fin réel ;
- une feuille "sessions" pour le début de la session, le groupe, la date et la description ;
- une feuille "séquence" pour l'identifiant de la séquence, le numéro de la séquence, le début et la fin de la séquence, la liste des codes acteurs, la description, l'état au début de la séquence et le numéro de session.

La plupart des informations renseignées pour la feuille "actions" seront ensuite reprises dans les autres feuilles mentionnées. Il est inutile de transcrire deux fois les mêmes données, la duplication de ces données se fera en fin de transcription.

2.2.2. Table « Session », « Séquence » et « Espace_Temps »

- **Avertissement sur la notion de temps**

La figure 2 présente la configuration de connexion des différentes machines utilisées lors des sessions pédagogiques. Une horloge est représentée sur chaque machine susceptible de fournir des traces ou enregistrements des actions de la session. Nous pouvons constater que les différentes unités utilisées par le serveur ont chacune une horloge autonome. La machine de chaque chercheur ayant enregistré les clavardages et la vidéo d'écran des espaces-temps a aussi sa propre horloge. Ainsi, lors du recueil des traces, les estampilles temporelles des actions ne se réfèrent pas à un temps universel : elles ne sont pas synchronisées. En utilisant la redondance de certaines traces, nous avons pu déterminer le décalage existant entre les différentes machines.

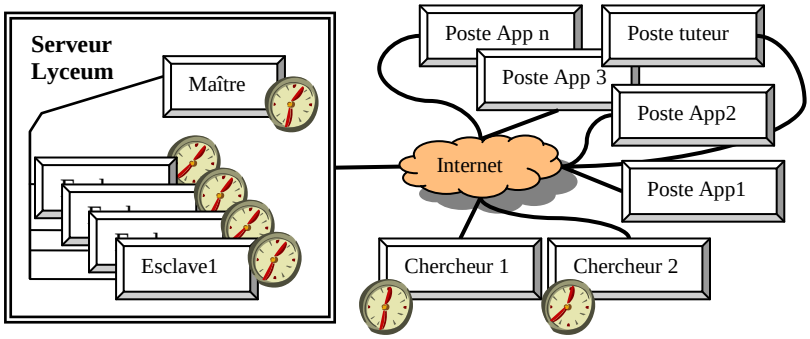

Figure 2 : Connexion des différentes machines Serveurs et Clients

Cette contrainte dans la précision du temps est extrêmement forte dans le cas d'interactions synchrones puisque deux actions consécutives peuvent être séparées de moins d'une seconde, tandis que les horloges étaient désynchronisées de plusieurs minutes.

Lors de la transcription des vidéos d'écran, la date (hh:mm:ss) de chaque action a été donnée dans le référentiel local de la durée de la vidéo donné par le lecteur vidéo ; la date zéro étant le début de l'enregistrement de la vidéo. **Il a fallu définir ensuite avec précision,** la date et heure (jj/mm/aa hh:mm:ss) du début de l'enregistrement vidéo dans le référentiel temporel de la machine du chercheur. Ainsi, la date de chaque action a pu être replacée dans un temps de référence choisi en appliquant un décalage rigoureusement calculé. Tous ces calculs de synchronisation ont été faits en supposant que chaque horloge n'a subi aucun décalage au cours de l'expérimentation.

Par ailleurs, nous avons noté que l'exportation de dates et heures de Excel vers Access puis MySQL peut engendrer des petites imprécisions. Pour pallier aux risques d'introduction d'imprécisions dans les valeurs temporelles nous opérons une vérification sur un échantillon, à chaque étape de collecte, transcription ou transformation, afin de contrôler qu'en particulier l'ordre des évènements n'est pas affecté.

➢ **Découpage en session et séquence**

L'unité première de découpage des enregistrements vidéo est celle de **session**. Chaque session est composée d'**espaces-temps** qui caractérisent le lieu et la date/heure d'une action. La notion d'espace-temps ET = (S, t_0, t_1) est définie comme un lieu S (salle ou espace virtuel) où un groupe se retrouve effectivement dans un intervalle de temps [t_0, t_1] avec t_0 : la date d'entrée de la *première* personne dans l'espace et t_1 : la date de sortie de la *dernière* personne de cet espace (ces personnes pouvant être différentes). Pour qu'il y ait un espace-temps, il faut qu'une personne présente dans un lieu soit au moins suivie d'une autre personne (signifiant une co-présence).

Chaque session est également découpée en **séquences pédagogiques.** Une séquence pédagogique est un temps pédagogique qui s'identifie par un **objectif** (la tâche ou la consigne), une **durée** plus ou moins prévue, un **lieu** et l'ensemble des **modalités** et **outils** mobilisés par les utilisateurs (formellement négocié ou non) pour réaliser la tâche. Ces données sont retranscrites comme suit dans la table « session » et « séquence » :

➢ **Table session**

Id_transcripteur	***No_session***	***Groupe***	***Date***	***Description***	***Code_transcript***
	R1	TutR	18/01/05	salutations et essai clavardage	Mul_Tmult1-3_codetranscription_v3_080415.doc
	R2	TutR	25/01/05	salutations et raconter sa semaine	Mul_Tmult1-3_codetranscription_v3_080415.doc
	R3	TutR	01/02/05	salutations + météo	Mul_Tmult1-3_codetranscription_v3_080415.doc

➢ **Table séquence**

id_sequence	***no_sequence***	***debut***	***Fin***	***liste_code_acteurs***	***description***	***etat_debut***	***no_session***
385	s1	17/01/05 10:23:00	17/01/05 10:41:00	mnl, chris, tim, al, herve, ed, ghis, res1, remy	salutations, test son (pb Rémy)	Clavardage affiché. Chris, al, herve, ghis ont leur vote Yes affiché. On prend la conversation en cours.	T1
386	s2	17/01/05 10:41:00	17/01/05 11:13:40	chris, res1, mnl, tim, ghis, amand, herve, ed, al, remy	présentations	Deux TB vierges sont affichés. Pas de clavardage.	T1

Chaque séquence est décrite par un cartouche d'en-tête ; elle est suivie de sa retranscription.

➢ **En-tête d'une séquence**

La transcription d'une séquence est d'abord identifiée par un en-tête qui indique :

Description	***Champ français***	***Champ anglais***	***Exemple***
Le numéro de la session pédagogique	no_session	no_session	R4

La date	date	date	15/02/05
La référence du fichier vidéo ou audio	ref_fichier_video	filename_video	Robin4_S102_CR_050208_monte.avi
La référence du fichier de clavardage	ref_fichier_clavardage	filename_chat	Robin5_Lobby+S101_MLB_chat_correct_050215.txt
Le lieu dans lequel se déroule la séquence	lieu	place	Lobby
Le numéro de la séquence à l'intérieur de la session	no_sequence	no_sequence	S1
L'heure de début	heure_debutS	time_startS	00:04:58
L'heure de fin	heure_finS	time_endS	00:37:52
La liste des codes des acteurs qui sont présents au début de la séquence	liste_code_acteurs	list_code_actors	sand, amel, res3, tim, aur, agnes, isa
Une description de la séquence	description	description	groupB : organiser CM modes de paiement
Un état de la séquence au début de l'enregistrement	état_début	state_start	Pas de chat affiché

2.2.3. Table « Action »

Sont détaillés ici les différents champs à renseigner correspondant à la partie « table action » de la feuille "actions" de la trame.

- **Les acteurs**

Les acteurs (apprenants, tuteurs, chercheurs) présents lors d'une session sont codifiés de façon spécifique pour chaque corpus et anonymisés. Les codes acteurs spécifiques ne seront pas développés ici et font l'objet d'une discussion entre les chercheurs de chaque corpus et le transcripteur. Il est important de bien respecter l'anonymisation des acteurs dans le contenu des échanges au moment de la transcription.

Certains aspects liés à l'audio ou au déroulement de la session comme les silences, la fin de la séquence et l'indétermination au sujet d'un acteur ont été codifiés comme **des acteurs** et correspondent aux codes suivants :

Code	***Code (version en)***	***Description***
sil	sil	Silence
fin	end	Fin de séquence
ind	ind	indéterminé

L'acteur "sil" renvoie à un silence dont la durée est supérieure à trois secondes entre chaque tour de parole audio.

L'acteur "fin" correspond au temps "T" qui marque la seconde exacte de la fin de la durée audio de chaque séquence. Cet acteur est utilisé pour le traitement des données audio.

L'acteur "ind" correspond à un acteur non identifié par le transcripteur.

➢ **Les lieux**

Les lieux où les actions se passent sont codifiés. Par exemple, pour l'environnement audiographique synchrone *Lyceum*, les codes adoptés sont les suivants (colonne Lieu) :

Lieu	***description***
Lobby	Lobby
S101	Salle 101
S102	Salle 102
S103	Salle 103

➢ **Les actions : généralités**

Les actions peuvent avoir lieu dans différentes modalités (audio, clavardage, textuel, iconique, graphique). Chaque action correspond ainsi à un type donné :

id	***Description***
tpc	clavardage
lm	lever de main
tpa	audio (dont silences)
v	vote
as	entrée sortie
abs	sortie momentanée
prod	action dans un module (Traitement de Texte (TT), Tableau Blanc (TB), Carte Conceptuelle (CC))
prod	gestion des espaces-documents (modules) : ouvrir, fermer, entrer, sortir

Chaque action correspond à un **identifiant (Id)**, à une **modalité (canal)** et à une **valeur**. Le numéro correspondant à la modalité/canal est donné de façon automatique par la base de données.

canal (id)	***modalité/canal***	***valeur***	***Description***
abs	1	abs_d abs_f	Sortie momentanée (sm début) Sortie momentanée (sm fin)
as	2	a/s	Entrée/sortie
tpa	3		Audio et silences
tpc	4		Clavardage *ou* Ouverture de la zone de clavardage
v	5	yes/no	Vote
ERREUR	6		Erreur (code utilisé pour les macros)
lm	7	le ba	Visualisation d'un lever la main Fin visualisation d'un lever la main
prod	8	prendre dans table	Transcription des actions dans les modules TT, TB, CC.

Pour chaque action correspond un temps de **début** et de **fin**, ainsi qu'une **séquence** et un espace-temps durant lesquels l'action a lieu. Par exemple :

Id_action	*canal*	*valeur*	*heure_debut*	*heure_fin*	*Code_acteur*	*EspTemps*	*id_sequence*	*lieu*
prod3	prod	CC(entrer)	15/02/2005 11:45:13	15/02/2005 11:45:13	TutR	E5R5	448	S102

Toutes les actions ont une durée, sauf le vote, le clavardage, les entrées/sorties et quelques actions de productions (dans ces cas, heures_debut = heure_fin).

Le transcripteur a la possibilité d'apporter dans la colonne « **commentaire** » un commentaire général sur l'action décrite (généralement sur la nature ou la forme de ce qui a été fait).

- **Avertissement sur Ouverture/fermeture du chat**

L'ouverture/fermeture du clavardage est transcrit comme une action de la modalité/canal clavardage donc **tpc**.

Ainsi, il existe 2 types de tpc :

- ceux dont la valeur est l'énoncé tapé dans la zone de clavardage ;
- ceux décrivant une ouverture ou une fermeture de la zone de clavardage.

Dans le second cas, pour éviter toute ambiguïté avec les tours de parole contenant des énoncés, la valeur de cette action (champ « valeur » dans la table Actions) sera **<<ouverture/fermeture du clavardage>>.**

2.2.4. Table « Parole »

Elle rassemble les informations concernant les champs suivants :

- la langue de la communication (la valeur de ce champ doit être conforme au code ISO 639-3, avec séparateur si plusieurs langues) ;
- le texte (ce qui a été écrit dans le clavardage);
- des informations phonétiques (si besoin pour souligner un phénomène particulier);
- des commentaires concernant les actes de parole (audio + clavardage).

- **Exemple de transcription du flux audio d'une séquence**

Faisant suite à l'en-tête, la transcription audio de la séquence est notée selon les rubriques suivantes :

Description	*champ*	*Champ*	*Exemple*
L'identifiant d'un tour de parole ou d'une action. Chaque identifiant est numéroté de façon incrémentale.	id	Id	1, 2, 3 etc.
L'heure de début du tour de parole ou de l'action	heure_debut	time_start	01:55:18
Le code de l'acteur	code_acteur	code_actor	Angel

La transcription audio	texte	Audio	sorry I did not understand +
Le codage phonétique API	phonetique[2]	phonetic	[↔≅γρι ⌉əɛIβλ]
Ce qui est écrit dans le clavardage	texte	Chat	that some old web site
L'action de vote (oui/non)	vote	Vote	yes/no
L'action d'arriver ou de sortir de la salle	arriv_sort	ent_exit	a/s
Les commentaires du transcripteur	commentaires	comments	Begaie
Les commentaires du transcripteur sur des actions autres que linguistiques	autre_com	other_com	tous les votes sont effacés
Le numéro de séquence dans cette session	id_sequence	id_sequence	385
Le codage de l'espace-temps	EspTemps	SpaceTime	R6s4 (correspond au codage de la séquence identifiant la session)

➢ **Les conventions spécifiques de transcription audio**

Les conventions adoptées pour l'audio reprennent pour partie de celles de la convention ICOR[3]. Le lecteur se reportera au document ANR-06-CORP-006- Annexe technique : mode d'emploi pour transcription multimodale.

2.2.5. Table « Production »

Ces actions regroupent les déplacements des acteurs d'un espace de production à un autre (i.e. les entrées/sorties de modules), les actions visibles ayant lieu dans chaque espace de production.

➢ **Rappel des types d'objet par espace de production**

Ce tableau indique pour chaque type (TB, CC, TT) correspondant au champ **« ED_Type »** de la table Production de la trame, quel type d'objets peut être décrit dans chaque espace de production. Ces dernières informations correspondent au champ **« Type_Objet »** de la table Production de la trame.

Tableau blanc (TB)	***Carte conceptuelle (CC)***	***Document texte (TT)***
Rectangle	Concept	Paragraphe
Ellipse	Relation	
Forme libre (crayon, pinceau,		

[2] La police utilisée est SilDoulos IPA93. Elle est téléchargeable depuis : http://www.sil.org/computing/fonts/encore-ipa.html

[3] Groupe ICOR 2006, *La convention ICOR*, site CORINTE, http://icar.univ-lyon2.fr/projets/corinte/

marqueur)		
Trait (ligne, flèche)		
Punaise		
Image		
Zone texte		

Les codes de transcriptions pour les productions sont présentés dans le document ANR-06-CORP-006- Annexe technique : mode d'emploi pour transcription multimodale.

- **Actions sur les objets par espace de production**

Ce tableau liste l'ensemble des actions réalisables et, par conséquent, visualisables dans chaque espace de production.

Actions	***TT***	***CC***	***TB***
Créer	Paragraphe (saisie ou peut être le résultat d'un coller issu d'un copier)	Concept, relation	Rectangle Ellipse Forme libre Trait Punaise Image Z_texte
Sélectionner	Paragraphe	Concept, relation	Rectangle Ellipse Forme libre Trait Punaise Image Z_texte
Editer (Contenu, Forme, Position, Taille)	Paragraphe (modifier en supprimant ou ajoutant du texte qui peut provenir (ou non) d'un copier ou d'un couper)	Concept, relation	Rectangle (modif couleur) Ellipse (modif couleur) Forme libre (modif couleur) Trait (modif couleur) Punaise (modif couleur) Z_texte (modif couleur et texte)
Supprimer	Paragraphe	Concept, relation	Rectangle Ellipse Forme libre Trait Punaise Image Z_texte
Charger	Un fichier texte	Un fichier CC	Une image

Entrer	Dans le module traitement de texte	Dans le module carte conceptuelle	Dans le module tableau blanc
Sortir	du module traitement de texte	du module carte conceptuelle	Du module tableau blanc
Ouvrir	le module traitement de texte	Le module carte conceptuelle	Le module tableau blanc
Fermer	le module traitement de texte	Le module carte conceptuelle	Le module tableau blanc
Renommer	le module traitement de texte	Le module carte conceptuelle	Le module tableau blanc
Sauvegarder	le module traitement de texte	Le module carte conceptuelle	Le module tableau blanc

- **La notion de rang**

La numérotation du rang d'un module se fait de gauche à droite. Un module garde son numéro de rang tout au long de sa « vie » et même après : un numéro de rang ne peut pas être réaffecté à un autre module qui serait ouvert ensuite. Ainsi, il n'est pas nécessaire de renuméroter les rangs des modules si un module intermédiaire vient d'être fermé. Autrement dit, tout module créé après le module 3 de la figure1 portera le numéro 4, que le module 3 soit fermé ou non. De même, si le module de rang 2 est fermé alors que les modules 3 et 4 existent, il n'y a aucune renumérotation. C'est la notion de moment de création du module qui importe pour affecter un numéro au module. Cette numérotation correspond au champ **« ED_rang »** de la trame.

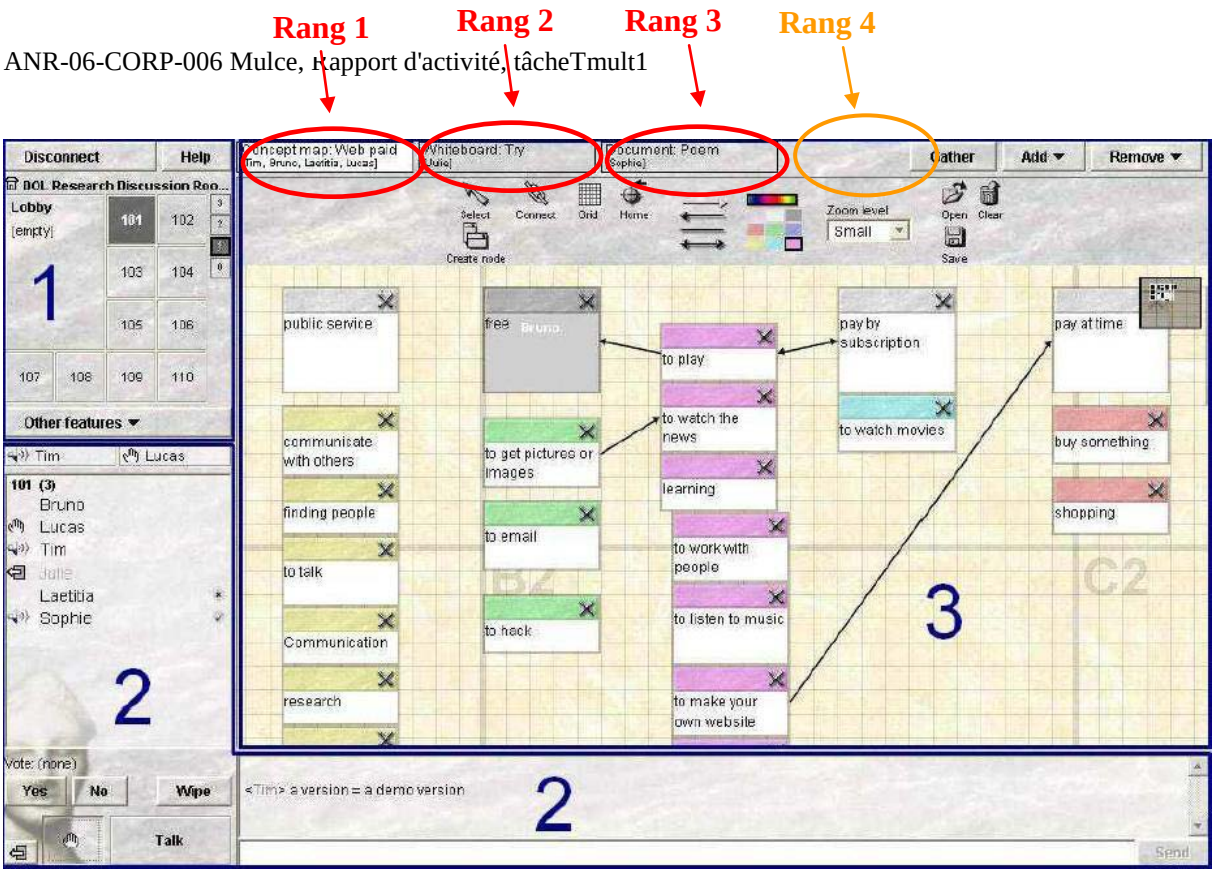

Figure 3 : Ecran Lyceum[4]

➢ **Le champ « precision_action »**

Des précisions sur l'action effectuée telles que : contenu/mise en forme/position (TT et TB), titre/contenu/position/mise en forme/dimensions (CC) sont des informations contenues dans le champ « **Precision_action »** de la table Production de la trame.

Les actions déplacer et redimensionner sont des instances particulières de l'action générique éditer. L'action **Editer** peut affecter différents attributs de l'objet :

- Dans le document textuel, le paragraphe possède les attributs suivants :
  - **Contenu** : le texte du paragraphe
  - **Forme** : couleur, gras, italique, surlignage, etc.

- Dans la carte conceptuelle, le concept possède les attributs suivants :
  - **Contenu** : texte intérieur au concept
  - **Forme** : couleur du bandeau de titre
  - **Position** : => éditer la position signifie déplacer l'objet
  - **Dimensions** : => éditer la dimension signifie redimensionner l'objet
  - **Titre** : => éditer titre du concept

- Dans le tableau blanc : la plupart des objets ont les attributs suivants :
  - **Contenu** : seulement pour zone de texte
  - **Forme** : couleur du trait, du texte ou du remplissage
  - **Position** : => éditer la position signifie déplacer l'objet
  - **Dimensions** : => éditer la dimension signifie redimensionner l'objet

[4] http://lyceum.open.ac.uk/

Le contenu des actions portant sur les concepts d'une carte conceptuelle est renseigné de la façon suivante :

- le code « ind » est utilisé lorsque le transcripteur ne voit pas le titre du concept.
- le code « » est utilisé lorsque le concept ne porte aucun titre.
- le séparateur « \ » est utilisé pour séparer les items du contenu d'un concept.

Exemples de transcriptions :

valeur	***Code_acteur***	***Type_objet***	***Precision_action***	***Contenu***
CC (éditer)	Ind	concept	dimensions	Ind
CC (créer)	Ind	concept		(;)
CC (éditer)	Ghis	concept	position	(elearning-website ;)
CC (éditer)	Ind	concept	position	(; NetLanguages)
CC (sélectionner)	Ghis	concept		(Contenu;materials\activities,exercices\difficulties levels\grammar,vocabulary,phonetic,heading(classement)\resources documents (externs links)\date of update\instructions for exercices\explanations of objectives)

- **Les valeurs dans la table Actions**

Cette table présente l'ensemble des valeurs, telles que transcrites dans les fichiers Excel, que peut prendre le champ **« valeur »** de la table Actions.

Valeur	***Description***
TT(éditer)	Edition d'un paragraphe
TT(supprimer)	Suppression d'un paragraphe
TT(sélectionner)	Sélection d'un paragraphe
TT(créer)	Création d'un paragraphe
TT(ouvrir)	Ouverture d'un module traitement de texte
TT(fermer)	Fermeture d'un module traitement de texte
TT(renommer)	Renomination d'un module traitement de texte
TT(sauvegarder)	Sauvegarde d'un module traitement de texte
TT(charger)	Téléversement un contenu dans le module traitement de texte
TT(entrer)	Entrée dans un module traitement de texte
TT(sortir)	Sortie d'un module traitement de texte
TB(éditer)	Edition d'un objet dans le tableau blanc
TB(supprimer)	Suppression d'un objet dans le tableau blanc
TB(sélectionner)	Sélection d'un objet dans le tableau blanc
TB(créer)	Création d'un objet dans le tableau blanc

TB(ouvrir)	Ouverture d'un module tableau blanc
TB(fermer)	Fermeture d'un module tableau blanc
TB(renommer)	Renomination d'un module tableau blanc
TB(sauvegarder)	Sauvegarde d'un module tableau blanc
TB(charger)	Téléversement un contenu dans le module tableau blanc
TB(entrer)	Entrée dans un module tableau blanc
TB(sortir)	Sortie d'un module tableau blanc
CC(éditer)	Edition d'un concept ou relation
CC(supprimer)	Suppression d'un concept ou relation
CC(sélectionner)	Sélection d'un concept ou relation
CC(créer)	Création d'un concept ou relation
CC(ouvrir)	Ouverture d'un module carte conceptuelle
CC(fermer)	Fermeture d'un module carte conceptuelle
CC(renommer)	Renomination d'un module carte conceptuelle
CC(sauvegarder)	Sauvegarde d'un module carte conceptuelle
CC(charger)	Téléversement un contenu dans le module carte conceptuelle
CC(entrer)	Entrée dans un module carte conceptuelle
CC(sortir)	Sortie d'un module carte conceptuelle

- **Récapitulatif des différents champs de la trame à renseigner, quelques exemples**

Table Production					
ED_Session	***ED_lieu***	***ED_Type***	***ED_rang***	***Type_Objet***	***Contenu***
T1	S101	TB	4		Whiteboard: Untitled 4
T1	S101	TB	1		Whiteboard: Untitled 1
T3	S101	CC	4	concept	(;Not professional)
T3	S101	CC	4	concept	(;Not professional)
T4	S101	TT	4		Document: Untitled 1
T5	S101	TT	1	paragraphe	the advantages are: real life, students area more implicated,

2.2.6. Table « Espace_Document »

Cette table apparaît dans la trame de transcription issue des différents champs de la base de données.

- **Définition des espaces documents**

Les espaces documents correspondent aux modules ouverts. Un espace document a une « vie » qui débute par une action ouvrir. Si aucune action ouvrir n'a été transcrite, l'heure de début est donc celle de l'espace temps auquel est rattachée l'action. L'heure de fin

correspond à l'heure de l'action de fermeture du module en question. Si aucune action de fermeture n'a eu lieu, l'heure de fin indiquée dans la base de données est celle de la fin de l'espace temps auquel est rattachée l'action **+ 4 heures**. Les 4 heures ajoutées correspondent au temps après lequel la plateforme de travail ferme automatiquement les modules ouverts.

Espace Document	
ED_Debut	***ED_fin***
17/01/05 10:23:38	17/01/05 14:41:30
17/01/05 10:23:38	17/01/05 14:41:30

2.2.7. Table «Source»

La table « Source » reprend les différentes informations concernant les fichiers enregistrés. L'id_source est donné automatiquement par la base de données. Le nom de fichier est celui donné lors de l'enregistrement. Ce n'est pas le transcripteur qui nomme le fichier.

Pour un **enregistrement audio**, le fichier avi est enregistré sous le nom précis : NomGroupe_NumeroSalle_AAMMJJ. Ex : GpTim_S101_050117.avi

Pour un enregistrement de texte (clavardage ou de la production dans le traitement de texte), le **fichier texte** est enregistré sous le nom précis : NomGroupe_NumeroSalle_canal_AAMMJJ. Ex : GpTim_S101_chat_050117.txt

Pour un **enregistrement de document html** (un document en html dans le tableau blanc par exemple), le fichier est enregistré sous le nom suivant : NomProduction_NomGroupe_NumeroSalle_AAMMJJ. Ex : TableauBlanc2-GPTim_S101_050117.html

Début réel pour une vidéo : date (jour jheure min sec) coorespondant au zéro de la vidéo

id_source	***Nom_fichier***	***Type***	***date_fichier***	***debut_reel***	***fin_reel***
85	GpTim_S101_monte_050117.avi	video/msvideo	17/01/05 00:00:00	17/01/05 10:17:32	17/01/05 11:56:10
174	GpTim_S101_chat_correct_050117.txt	text/plain	17/01/05 00:00:00		

2.2.8. Marquage de l'anonymisation

Suivant les licences Creative Commons, les indications concernant les individus doivent être masquées sauf indications contraires (adresses, patronymes, etc.). Les prénoms peuvent être conservés.

Voici un extrait du rapport Mulce Tâche Tstruct1.2 « Propositions pour l'anonymisation et application aux interactions textuelles de Simuligne ». Dans ce rapport la procédure de marquage complet est abordée. Pour une procédure simplifiée, voir en fin de cette section. Le choix de la procédure simplifiée appartient aux collecteurs du corpus.

- **Marquage complet**

A l'intérieur du texte, les « chaînes » identifiant un acteur seront systématiquement encadrées par des balises <Actordesignation> chaîne </Actordesignation>. La « chaîne » pourra (ou non) avoir subit une transformation selon qu'elle contenait ou non des informations non diffusables sur l'identité d'une personne physique. Aucun attribut ni élément ne seront obligatoires dans cette structure.

```
<actordesignation actorref = "{actor_code}" person = "real/fictitious" process = "{process_mark}">
    <firstname type = "initial/abbreviated/shortname/complete" correct = "exact/modified/wrong" modified = "true/false">Ici se trouve le texte de remplacement du prénom</firstname>
    {ici il peut y avoir (ou non) des caractères séparateurs : espaces, fin de ligne, fin de paragraphe, ou encore « „.;:!?=+-/\~*_["]#(')&@$% », autres ?}
    <surname type = "initial/abbreviated/complete" correct = "exact/modified/wrong" modified = "true/false">Ici se trouve le texte de remplacement du surnom</surname>
    {ici il peut y avoir (ou non) des caractères séparateurs}
    <lastname type = "initial/abbreviated/complete" correct = "exact/modified/wrong" modified = "true/false">Ici se trouve le texte de remplacement du patronyme</lastname>
</actordesignation>
```

Notons que le texte de remplacement du prénom ou du surnom peuvent être identiques à la chaîne originelle, mais celui du patronyme réel doit obligatoirement être différent dans une version diffusable.

➢ **Exemple**

Extrait (raccourci) d'un forum du corpus Simuligne. Pour des exemples portant sur des exemples de clavardage et de courriel, coir le document sur l'anonymisation.

<discussion nom="**les sons du Suffolk**">

<message idmess="**i78**" mess_pere="**i75**" date="**4 5 2001**" jour="**Vendredi**" heure="**9h50**" nom_auteur="**Marja GIEJGO**">**Merci Anna! J'ai les deux navigateurs sur mon ordi - Netscape et Internet Explorer, et donc la prochaine fois je vais télécharger avec Netscape. J'ai toujours préféré Internet Explorer, je ne sais pas pourquoi - mais je peux toujours m'arranger!! Marja Giejgo** </message>

<message idmess="**i74**" mess_pere="**i60**" date="**4 5 2001**" jour="**Vendredi**" heure="**1h40**" nom_auteur="**Anna Vetter**">**Moi, j'ai pu le lire. Peut-être que les autres n'ont pas encore installé PureVoice ?**</message>

- <message idmess="**i73**" mess_pere="**i59**" date="**4 5 2001**" jour="**Vendredi**" heure="**1h39**" nom_auteur="**Anna Vetter**">

Ah, des charolais ! je me disais bien que j'en avais déjà vu des comme ça. En franche-comté, les vaches sont tachetées, comme dans le document attaché. Le coucou sufflokais est vraiment génial ! Maintenant, je me représente beaucoup mieux ta région : avec le son et l'image ! Anna

<attachement reference="**988933128**" nom="**zanvett.jpg**" type="**image/jpeg**" />

</message>

Version transformée :

<discussion nom="**les sons du Suffolk**">

<message idmess="**i78**" mess_pere="**inull**" date="**4 5 2001**" jour="**Vendredi**" heure="**9h50**" nom_auteur="**Al5**">**Merci** <Actordesignation ActorRef = "At" person = "Real" process = "demoCR070513"><Firstname Type = "Complete" Correct = "exact" Same = "True">**Anna**</firstname></Actordesignation>**! J'ai les deux navigateurs sur mon ordi - Netscape et Internet Explorer, et donc la prochaine fois je vais télécharger avec Netscape. J'ai toujours préféré Internet Explorer, je ne sais pas pourquoi - mais je peux toujours m'arranger!!** <Actordesignation ActorRef = "Al5" person = "Real" process = "demoCR070513"><Firstname Type = "Complete" Correct = "exact" Same = "True">**Marja**</firstname><Lastname Type = "Complete" Correct = "exact" Same = "False">Golly</Lastname></Actordesignation></message>

<message idmess="**i74**" mess_pere="**i60**" date="**4 5 2001**" jour="**Vendredi**" heure="**1h40**" nom_auteur="**At**">**Moi, j'ai pu le lire. Peut-être que les autres n'ont pas encore installé PureVoice ?**</message>

<message idmess="**i73**" mess_pere="**i59**" date="**4 5 2001**" jour="**Vendredi**" heure="**1h39**" nom_auteur="**At**">

Ah, des charolais ! je me disais bien que j'en avais déjà vu des comme ça. En franche-comté, les vaches sont tachetées, comme dans le document attaché. Le coucou sufflokais est vraiment génial ! Maintenant, je me représente beaucoup mieux ta région : avec le son et l'image ! <Actordesignation ActorRef = "At" person = "Real" process = "demoCR070513"><Firstname Type = "Complete" Correct = "exact" Same = "True">**Anna**</firstname></Actordesignation>

```
        <attachement          reference="988933128"          nom="zanXXX.jpg"
        type="image/jpeg" />
      </message>
</discussion>
```

- **Marquage simplifié**

Cette procédure de marquage simplifiée a été utilise lors de la transcription du corpus de Tridem 06 :

```
merci <acd cd="afb01_2">Aida</acd> + merci <acd
cd="amc01_3">Greg</acd> + merci <acd cd="afb02_4">Sophie</acd> + je
dois aller euh chercher mon fils donc à bientôt ++
```

2.3. Récapitulatif des codes pour transcription

2.3.1. Type action dans espace de production

Valeur
TT(éditer)
TT(supprimer)
TT(sélectionner)
TT(créer)
TT(ouvrir)
TT(fermer)
TT(renommer)
TT(sauvegarder)
TT(charger)
TT(entrer)
TT(sortir)
TB(éditer)
TB(supprimer)
TB(sélectionner)
TB(créer)
TB(ouvrir)
TB(fermer)
TB(renommer)
TB(sauvegarder)
TB(charger)
TB(entrer)
TB(sortir)
CC(éditer)
CC(supprimer)
CC(sélectionner)
CC(créer)
CC(ouvrir)
CC(fermer)
CC(renommer)
CC(sauvegarder)
CC(charger)
CC(entrer)
CC(sortir)

2.3.2. Type_objet

Tableau blanc (TB)	***Carte conceptuelle (CC)***	***Document texte (TT)***
Rectangle	Concept	Paragraphe
Ellipse	Relation	
Forme libre (crayon, pinceau, marqueur)		
Trait (ligne, flèche)		
Punaise		
Image		
Zone texte		

2.3.3. Action sur les objets par espace de production

Actions	***CC***	***TB***
Créer	Concept, relation	Rectangle
		Ellipse
		Forme libre
		Trait
		Punaise
		Image
		Z_texte
Sélectionner	Concept, relation	Rectangle
		Ellipse
		Forme libre
		Trait
		Punaise
		Image
		Z_texte
Editer	Concept,	Rectangle

	relation	Ellipse
		Forme libre
		Trait
		Punaise
		Z_texte
Supprimer		Rectangle
		Ellipse
		Forme libre
		Trait
		Punaise
		Image
	Concept, relation	Z_texte

2.3.4. Actions dans l'espace de communication

canal	id_action
1	abs
2	as
3	tpa
4	tpc
5	v
6	ERREUR
7	lm
8	prod