Communication au colloque EPAL : Echanger Pour Apprendre en Ligne, 7-9 juin 2007, Grenoble, France

Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche

Christophe REFFAY (LIFC, Besançon), Thierry CHANIER (LASELDI, Besançon), Muriel NORAS (LIFC, Besançon), Marie-Laure BETBEDER (LIFC, Besançon)

RÉSUMÉ

Du point de vue méthodologique, pour permettre une analyse des interactions situées, il convient de relier les différentes données issues de formations en ligne, pour construire un objet d'analyse, exploitable par différentes équipes et disciplines. Le constat actuel est que les données sont souvent décontextualisées, parcellaires ou simplement inaccessibles. Nous définissons le corpus d'apprentissage, en identifiant l'information qu'il doit contenir et une certaine structuration pour rendre possible son échange et la capitalisation des analyses. Le protocole de recherche, le scénario pédagogique, les interactions, productions et traces, les licences et les analyses capitalisables en sont les principaux constituants. Nous illustrons la démarche de construction d'un tel corpus sur l'exemple de la formation Simuligne. Ce travail est ensuite positionné au regard des questions d'éthique et de droit, des efforts de standardisation et des avancées sur l'analyse des traces en EIAH pour rendre les outils d'analyse interopérables.

MOTS CLÉS

apprentissage en ligne, contexte, corpus d'apprentissage, interactions verbales et non verbales, échange de données de recherche.

Introduction

Étudier l'apprentissage en ligne, que cela soit à des fins de compréhension de cette forme d'apprentissage humain situé, d'évaluation des scénarios et dispositifs pédagogiques associés ou encore d'amélioration des environnements technologiques, requiert la disponibilité de données d'interaction provenant des différents acteurs, apprenants et formateurs, participant aux situations d'apprentissage. Les publications et évènements scientifiques en rapport avec ce sujet ne manquent pas en France ou dans le monde. Mais les communautés pluridisciplinaires de chercheurs impliqués dans cette thématique n'ont pas encore réussi à caractériser un véritable objet d'étude scientifique, ni une démarche méthodologique en rapport.

D'une part, les données d'étude sont parcellaires, donc décontextualisées, en regard des éléments constitutifs du dispositif de formation, ou encore inextricablement imbriquées au sein des environnements technologiques sous des formats propriétaires. Créer un objet d'étude scientifique de l'apprentissage en ligne ne peut se limiter à collecter des données d'interaction d'apprenants comme le rappellent ces auteurs oeuvrant dans le champ de l'apprentissage des langues :

Researchers must carefully document the relationships among media choice, language usage, and communicative purpose, but they must also attend to the increasingly blurry line separating linguistic interaction and extralinguistic variables. […] Studies of linguistic interaction will likely need to account for a host of independent variable: the instructor's role as mediator, facilitator, or teacher; cross-cultural differences in communicative purpose and rhetorical structure; institutional

convergence or divergence on defining course goals; and the affective responses of students involved in online language learning projects. (Kern, Ware & Warshauer, 2004).

Notre domaine de recherche s'intéressant, non seulement à l'apprentissage mais également à la pédagogie, il convient pour mener à bien des études en vue de "[…] *gather evidence about the effects of instructional conditions of instruction*" (Chapelle, 2004 : 594). Cela implique de rassembler des éléments du contexte, notamment, mais pas seulement, ceux qui ont caractérisé le dispositif pédagogique. Du point de vue méthodologique, il convient de relier les différents types de données pour avoir un objet digne d'analyse, comme le souligne cet extrait à propos des interactions produites dans les forums de discussion :

La recherche sur le forum de discussion en contexte éducatif tente de rendre compte de phénomènes complexes à l'aide de méthodes d'analyse de contenu qui n'éclairent qu'un aspect de la réalité. […] La méthode d'analyse devrait être capable de traiter le discours comme une interaction verbale située, dans ses dimensions linguistiques […], situationnelles (liées à l'univers de référence et à la situation d'interaction) et des contraintes hiérarchiques (liées à la structure hiérarchique du discours). (Henri et Charlier, 2005)

D'autre part, la non accessibilité aux données de recherche, qui est l'état de fait quasi général au sein de notre communauté internationale, est un frein de premier ordre à la reconnaissance des situations d'apprentissage en ligne comme un objet d'étude scientifique : elle empêche les vérifications ou infirmations, la réplication, le raffinement, les analyses multiples etc. La réplication d'expériences de formations à partir d'un scénario innovant est des plus réduite pour les raisons évoquées précédemment. Les cas de réanalyse de données d'apprentissage sont tellement rares qu'on n'hésitera pas à citer l'étude de Kramsch et Thorne qui, à partir des données de Kern, ont produit une interprétation différente du premier auteur pour expliquer l'échec des échanges en ligne dans une formation associant des apprenants de langues maternelles différentes (Kern et al., 2004 : 251). La réanalyse peut être motivée par différents facteurs comme la vérification (dans l'exemple précédent), l'utilisation de méthodes alternatives d'analyse de contenus (cf. l'éventail présenté dans le numéro thématique de la revue Computers & Education (Valcke & Martens, 2006)), la comparaison de résultats provenant de démarches disciplinaires distinctes (Corbel et al., 2006), etc.

Mais, hormis cette vision contrastive ou alternative des choses, on peut considérer que la démarche d'analyse est par nature un processus cumulatif qui se joue entre équipes de recherche distinctes, s'appuyant sur l'exploitation d'analyses précédentes, chacune apportant leur lot d'annotations. Ainsi le processus de transcription d'interactions audio-orales ou multimodales provenant de plateformes synchrones (Chanier & Vetter, 2006) est l'étape préalable à la conduite des premières analyses. De même, si l'on considère un ensemble de forums ou de clavardages (chat), une première étape, souvent obligée du fait des structures propriétaires des plateformes, consiste à structurer ces données (balisage des tours de paroles, locuteurs / scripteurs, etc.). Cette étape peut être suivie par un premier niveau d'annotations provenant d'une analyse conversationnelle, puis d'un second niveau orienté vers une analyse du discours. De telles pratiques de recherche sont déjà bien ancrées dans des domaines comme le TAL (Traitement Automatique des Langues) où à partir de textes extraits d'un corpus, des chercheurs distincts opèrent des descriptions cumulatives enchaînant niveaux morphologique, syntaxique, sémantique, anaphorique, etc. (Salmon-Alt et al, 2004).

Une communauté de recherche s'affirme en partageant des ressources (contextualisées), des outils et des pratiques. Une partie de ces éléments (ressources et outils) sont les constituants, aux côtés des publications, de la contribution scientifique que nos directions de recherche nous ont donné mandat de déposer en accès libre (Berlin, 2003 ; Chanier, 2004 : 121). L'échange entre équipes de recherche a donc pour condition nécessaire l'accès libre qui doit non seulement s'appuyer sur un ensemble de protocoles et techniques (standardisation, interopérabilité, métadonnées, etc.) mais aussi fixer les questions de droit relatives au domaine et, plus encore, dans ce domaine reposant sur l'étude de l'apprentissage / formation, celle longtemps négligée, de l'éthique :

Any discussion of technology in second language research would not be complete without raising the ethical challenges that researchers face in SLA [Second Language Acquisition] research in general

and particularly in research involving the collection and archiving of personal performance data that reveal personal attributes (Chapelle, 2004 : 599).

Dans cet article, aux fins d'élargir et de consolider la démarche scientifique dans le domaine de l'apprentissage en ligne, et particulièrement, des interactions en ligne en situation d'apprentissage, nous présentons la notion corpus d'apprentissage. Elle est au cœur du projet Mulce (MULtimodal Corpus Exchange) (Mulce, 2007) soutenu par l'Agence Nationale de la Recherche dans le cadre du programme "Corpus et Outils de la Recherche en Sciences Humaines et Sociales" . Après un détour vers les communautés de recherche qui ont largement développé et instrumentalisé les notions de corpus et banque de corpus associés, nous examinerons les éléments constitutifs d'un corpus d'apprentissage qui regroupe et structure les données et traces issues d'une expérimentation de formation, enrichies d'informations, elles aussi structurées, sur l'environnement technologique, pédagogique et scientifique et les descriptions résultats d'analyses. Nous illustrerons ensuite une partie de ces composants et des formalismes associés à partir d'une formation donnée. Nous aborderons enfin les conditions de la réalisation de l'échange de tels corpus et de leur accès libre.

2. Exploration de la notion de corpus d'apprentissage

2.1. Notion de corpus en linguistique et en interactions orales

Bommier-Pincemin (Bommier-Pincemin, 1999) étudie les points de vue des différents champs disciplinaires qui constituent et exploitent les corpus textuels. D'après elle, "le corpus se définit de fait comme l'objet concret auquel s'applique le traitement, qu'il s'agisse d'une étude qualitative ou quantitative". L'auteur parle d'emblée d'un corpus dans une perspective de traitement et ne se contente pas de le définir comme un ensemble de textes présentant une certaine homogénéité. En réponse à ceux qui ont cette vision réductrice, elle répond "mais les données ont un nom trompeur : elles ne s'imposent pas, elles sont construites". Elle ajoute ensuite :

Le corpus est un tout, un vaste ensemble, qui constitue à lui seul le cadre et le référentiel de l'analyse. Il met en présence les éléments, il fait qu'ils sont aussi considérés dans leur interrelation globale.

Ce cadre est donc indispensable. On peut le rapprocher de la notion de contexte, terme fortement polysémique mais que l'on saisira du point de vue des chercheurs en interactions verbales. Pour Goodwin et Duranti (Goowin & Duranti, 1992), le contexte se détermine à partir de la perspective des acteurs participant à l'interaction, agissant dans l'univers où ils sont impliqués. Le contexte s'étudie en se concentrant sur les activités que les participants construisent en vue de se constituer des univers sociaux culturellement et historiquement organisés. Les notions d'acteurs, d'activité, voire de communauté avec les règles auto-construites sont familières dans notre domaine. Suivant ce courant interactionniste, l'analyste ne peut directement invoquer un contexte donné que s'il peut être mis en rapport avec le point de vue des participants. L'analyste doit alors trouver des manifestations du contexte dans les occurrences verbales (lorsque les participants parlent d'un objet de connaissance visé, d'un point d'organisation de l'activité, etc.) ou dans leurs actions. La nature du contexte invoqué lors de l'analyse est donc un sujet de débat entre chercheurs. Comme le dit Schegloff (Schegloff, 1992 : 194) : « "putting something in context" can take the proposed context as the "news" and as the object of analysis (rather than as the "given" relative to the object of analysis) ».

Ne peut être invoqué que ce qui est connu. Il incombe donc au collecteur du corpus de renseigner ces éléments du contexte. Dans les domaines des interactions verbales, les chercheurs constituent des corpus autour d'un ensemble d'objets protéiformes dans lequel le contexte joue un rôle important. Ainsi dans le projet Clapi (Plantin et al., 2005), les corpus sont constitués d'objets multimédias documentant une ou plusieurs interactions qui présentent une certaine homogénéité (de site, de terrain, entre participants, etc.). Le corpus contient des données primaires (enregistrement audio et vidéo) accompagnées de données collectées dans l'environnement (comme les documents lus ou produits par les participants), de données secondaires (transcriptions, éventuellement en plusieurs versions, accompagnées des conventions, des notes d'observation), des métadonnées sur le corpus, les contributeurs à son élaboration, des données documentaires comme les articles de recherche expliquant des analyses associées.

Alors que les interactionnistes se "contentent" d'observer et interpréter des situations qu'ils n'ont pas provoquées, notre milieu est directement intéressé par les rapports entre contenus des formations, fonctions des dispositifs pédagogiques et environnements technologiques, ainsi que sur les interactions en situation d'apprentissage, toute chose sur lesquelles nos chercheurs ont un pouvoir d'intervention en tant que concepteur de formation ou développeur d'environnements technologiques.

2.2. Banques de corpus, modèles, échanges et instrumentalisations

Examinons succinctement des projets nationaux qui se sont donnés pour objectif l'échange de corpus et la constitution d'une communauté de recherche conséquente.

Freebank (Salmon-Alt et al., 2004) (Freebank, 2007) est une banque de corpus du français, corpus annotés à plusieurs niveaux, libre d'accès, codée selon des schémas normalisés, intégrant des ressources existantes et ouvertes à l'enrichissement progressif. Le modèle de corpus associé décrit celui-ci comme étant composé d'un ensemble de ressources et d'un ensemble de niveaux de description. Un corpus se constitue autour d'une couverture linguistique donnée. La collection de ressources rassemble les unités physiques de dépôt de données relatives à cette couverture. Elle contient aussi bien les documents récoltés dans leur état primaire que ceux ayant fait l'objet d'une série d'étiquetages linguistiques. Les auteurs, remarquant qu'il est difficile de départager ce qui relève de la représentation de ce qui relève de l'interprétation, rassemblent les notions de "recueil de données", "transcription" et "annotation" dans celle de niveau de description. Les liens de dépendance entre niveaux de description permettent de gérer les séries d'analyses successives sur un même corpus accomplies par des chercheurs différents.

La banque de corpus Clapi (Plantin et al., 2005) (CLAPI, 2007) sur les interactions verbales a constitué son fonds d'origine à partir des recueils de données orales, des transcriptions, analyses et publications de chercheurs en passe de quitter leur activité professionnelle. Ce fonds historique est aujourd'hui régulièrement complété par le produit des nouveaux projets de recherche, qu'ils soient de taille nationale, ou résultant du travail de jeunes doctorants. C'est en examinant l'interface en ligne de la banque de données Clapi que le travail de collecte et d'organisation des corpus prend tout son sens. En effet, outre les procédures de dépôt de nouveaux corpus, le site offre un ensemble d'outils de consultation, sélection, recherche. Même si on peut regretter à ce stade que seuls des extraits de corpus, et non les corpus entiers, soient téléchargeables directement par le chercheur, l'exploration, elle, s'effectue à l'aide de requêtes qui fouillent les corpus dans leur intégralité. Des outils standard de traitement textuels (lemmatiseur, concordanceur) permettent des recherches pointues à partir d'éléments lexicaux ou de structures du discours (tours de parole, etc.). Cette fouille inter-corpus conduit les chercheurs à la découverte de nouveaux objets études et à la formulation d'hypothèses originales.

De ces projets de banque de corpus (et d'autres que nous n'avons pas place d'examiner ici, dont (Jacobson, 2004)), nous retiendrons les traits communs suivants. Tout d'abord, les efforts de constitution de ces banques ont des visées triples : patrimoniales (préservation de données sociétales anciennes ou contemporaines), appliquées (dont des visées éducatives pour la formation des jeunes chercheurs) et, prioritairement, scientifiques au sens où ces banques contribuent de façon essentielle à un approfondissement du travail de recherche dans le partage et à une confrontation des idées entre équipes dispersées. Ensuite une base de corpus peut être vue de trois façons différentes, chaque vue étant reliée à des rôles d'acteurs distincts :

- **Dépôt**. Dans les rôles associés on distinguera le responsable du corpus, personne ou entité qui a déposé le corpus et est garante du respect des droits le concernant, et les contributeurs (transcripteurs, collecteurs, etc.)
- **Organisation et diffusion**. L'entité, souvent une communauté, responsable du site de la banque de corpus. Elle joue le rôle d'éditeur et a pour fonction de créer et organiser les structures, modèles pour représenter et documenter les données déposées en vue de leur étude ou traitement ultérieur. La démarche de ceux qui ont créé les ressources de départ (ressources de base, niveaux de description, analyses) doit donc être renseignée.

- **Utilisation**. Parmi les utilisateurs de la base on distinguera les internautes anonymes, de ceux identifiés, chercheurs ou éducateurs. La question de l'accès libre intégral ne s'appliquant, à notre sens, que sur les personnes identifiées.

Notons, sans pouvoir le développer plus ici, que des critères de qualité, tels que signifiance, acceptabilité, exploitabilité sont utilisés pour juger de la recevabilité d'un dépôt. L'existence de tels critères n'implique pas pour autant que seuls les gros corpus sont acceptables. Des contributions de taille modeste permettent d'élargir le champ des acteurs impliqués dans le partage.

Dernier point d'importance, l'acte d'échange qui met en rapport ceux qui déposent avec ceux qui utilisent requiert une attention particulière aux problèmes d'interopérabilité, de formats, de standards (pour les métadonnées, comme pour les données) dont nous reparlerons en section 4.

2.3. La notion de corpus d'apprentissage

Un corpus d'apprentissage est constitué autour de l'objet d'étude résultant d'une situation de formation / apprentissage en ligne. Le corpus primaire rassemble l'ensemble des données d'interaction, de production des acteurs engagés dans la formation, complétées par les traces des actions laissées par ces acteurs dans le système. On y trouve donc des éléments comme courriels, forums, clavardages, interactions issues d'environnements audio-vidéo graphique synchrone, vidéo d'écran, données audio, traces (logs) système etc.

Le cadre (ou contexte) qui permet au chercheur à la fois de donner du sens à ces données (offrir un cadre interprétatif) et d'ouvrir la porte aux analyses est constitué principalement par :

- le cadre pédagogique : scénario pédagogique, données sur les acteurs ;
- le cadre de recherche (s'il existe), qui peut lui aussi apporter son lot de données primaires sur les acteurs (questionnaires, entretiens, etc.), ainsi qu'un scénario (ou protocole) de recherche, qui a mis à contribution les acteurs de la formation dans des activités spécifiques, planifiées en pré-, post-formation ou au cours de son déroulement.

Le tout (données et contexte) est organisé en vue de l'analyse de ces situations d'apprentissage en ligne. Une banque de corpus d'apprentissage doit disposer, à l'image du projet Clapi d'un environnement d'utilisation également en ligne, que nous intitulerons sommairement "système de fouille".

La détermination d'un objet d'étude, de données primaires répondant aux critères de qualité des corpus, d'un cadre / contexte et d'un système de fouille sont indissociables pour définir la notion de corpus d'apprentissage. Le qualificatif "apprentissage" se rapporte à l'objet d'étude et aux types de données primaires (produits d'une situation de formation), au cadre ou contexte (qui relie approche pédagogique et recherche sur l'apprentissage). Les outils du système de fouille, quant à eux, n'ont pas nécessairement besoin d'être spécialement conçus pour cet objet d'étude. Par exemple, de "simples" outils de concordances peuvent apporter un service notable au chercheur.

Cette introduction à la notion de corpus d'apprentissage (learning corpus) nous amène à la distinguer de celle de corpus d'apprenants (learner corpus) (Granger et al, 2001), (Beltz, 2004). Dans un corpus d'apprentissage, la notion d'acteur englobe aussi bien les apprenants que les formateurs (tuteurs). Les données relatives à ces derniers sont indissociables du corpus d'apprentissage dans la mesure où, d'une part, ils ont interagi avec les apprenants et ont donc influencé l'ensemble de la situation d'apprentissage et, d'autre part, l'étude de leur comportement est une des clés pour comprendre ce qu'est un bon formateur en ligne.

2.4. Les constituants d'un corpus d'apprentissage

La figure 1 schématise les constituants principaux d'un corpus d'apprentissage, à savoir :

- Le noyau du corpus, encore appelé Instanciation (pour des raisons définies en section 3) comprend l'objet d'étude à savoir l'ensemble de données d'interactions, de production des acteurs de la situation de formation / apprentissage en ligne, complété par les traces système.

- Le Contexte préexistant ou cadre référentiel, lui-même composé des : scénario pédagogique et protocole de recherche (élément facultatif). Une autre partie du contexte portant sur la définition des environnements technologiques et sur les acteurs se trouve dans l'instanciation (comme nous l'expliquerons en section 3).
- Une partie Licence qui indique à la fois les droits de l'éditeur du corpus et des utilisateurs et les éléments de respect de l'éthique vis-à-vis des acteurs de la formation (cf. section 4). Cette partie ouvre la voie à l'utilisation du corpus et à la production d'analyses. Une partie du contenu licence est privée, détenue seulement par le responsable du corpus et contient les informations nécessaires à la preuve de l'existence des personnes et du respect des droits et de l'éthique (cf. figure 2).
- Une partie Analyses qui contient les niveaux de description au sens de la Freebank. Les transcriptions en font donc partie.

Un corpus d'apprentissage est associé à l'environnement d'utilisation qui intègre le système de fouille déjà évoqué.

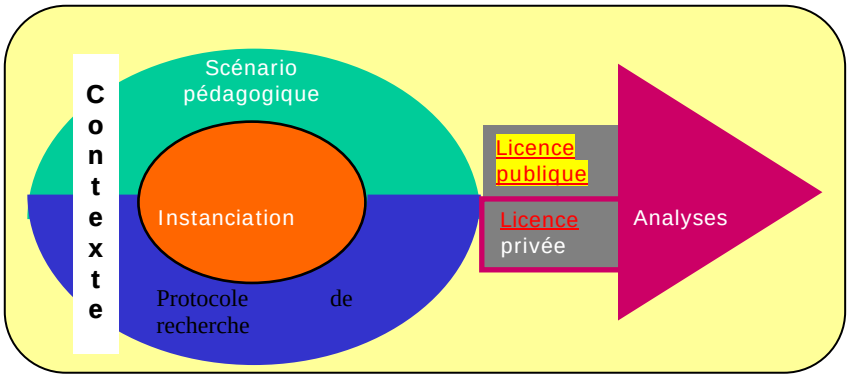

Figure 1 - Les grandes parties d'un corpus d'apprentissage de la banque Mulce

2.5. La structuration d'un corpus d'apprentissage

La structuration d'un corpus d'apprentissage a une double fonctionnalité, d'une part, organiser et structurer les données de façon à pouvoir établir des liens entre interactions, production et contexte, à permettre au système de fouilles d'opérer dans un ensemble cohérent et, d'autre part, à autoriser l'exportation du corpus d'apprentissage entier (ou de chacun de ses sous-corpus distinguables) dans un format d'échange ou format pivot. La structure adoptée, encore dénommée Mulce-struct dans notre terminologie, est schématisée en figure 2.

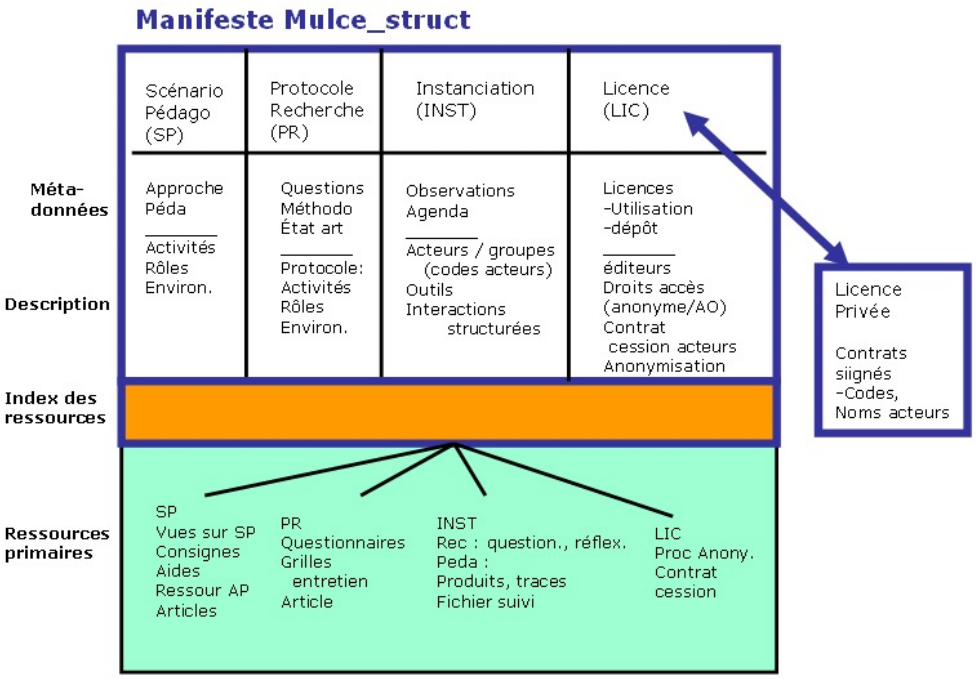

Figure 2 - Structure Mulce-struct d'un corpus d'apprentissage

Dans la partition horizontale, le lecteur y reconnaîtra 4 constituants principaux du corpus cités précédemment (scénario pédagogique, protocole de recherche, instanciation et licence, la partie Analyses n'étant pas représentée). La structure est stratifiée verticalement suivant 3 niveaux :

1. Un ensemble de descriptions structurées (suivant des schémas XML) contenant les descriptions propres à chaque constituant du corpus d'apprentissage, accompagnées de métadonnées. Ces dernières informent sur ces constituants en déclarant de façon synoptique l'approche pédagogique choisie, les principales questions de recherche, etc., et citent les auteurs et contributeurs de chaque description. Des vues alternatives peuvent être présentes. Ainsi le scénario pédagogique peut être illustré graphiquement dans un format lisible par un humain (cf. les graphes MOTPLUS en section 3) ou au contraire être décrit formellement dans un langage et des concepts précis. Ce type de format, très peu lisible directement par les humains, est en revanche déterminant dans le but d'offrir une description détaillée, aisément traitable automatiquement pour mettre, en particulier la description d'un constituant en rapport avec celle d'un autre (par exemple, un scénario pédagogique décrit en IMS-LD en rapport avec la partie instanciation, comme l'explique la section 3).
2. Un ensemble de ressources primaires. Ces ressources sont "primaires" au sens où elles sont dans l'état (au processus d'anonymisation près, cf. section 4), où elles ont été déposées par le responsable du corpus ou d'une transcription. Les données correspondantes sont rarement à l'état brut mais ont souvent subi un pré-traitement comme des montages audio et vidéo pour les vidéogrammes, des conversions pour des forums de formats propriétaires dans des formats plus ouverts. La répartition de ces ressources primaires en répertoires correspondant aux quatre constituants du corpus, conduit ainsi à placer les formulaires vierges des questionnaires de recherche dans la partie "Protocole de recherche" où ils seront reliés au scénario de recherche et les questionnaires remplis par les acteurs dans les répertoires correspondant à la partie "Instanciation".
3. Un index des ressources où sont listés de façon structurée les liens associant les fichiers figurant dans le niveau "ressources primaires" à leur référencement dans les éléments du niveau description.

Le lecteur averti aura sans doute reconnu immédiatement dans le schéma de la figure 2, une structure de type IMS-CP (2004) : la couche des descriptions structurées couplée à celle de l'index des ressources (abusivement nommée "resources" en IMS-CP) constituant le manifeste, écrit en XML ; le niveau "ressources primaires", correspondant à la partie "content" et le tout étant assemblé dans une archive ("Package Interchange File") permettant le transport de l'ensemble du corpus. Tel est bien le cas, même si notre schéma ne représente pas explicitement la couche de métadonnées, sous-partie du manifeste, propre à l'ensemble du corpus. Les raisons du choix d'IMS-CP tout comme les limites inhérentes à ce format seront exposées en section 4.

2.6. Plateforme Mulce et corpus distinguables

Le premier intérêt de ces efforts de structuration abordés jusqu'ici est de pouvoir offrir au chercheur un environnement de travail lui permettant d'effectuer des fouilles intra ou inter corpus, les corpus en question pouvant avoir la granularité d'un corpus d'apprentissage ou d'un sous-corpus distinguable.

La linguistique de corpus différencie les types corpus d'étude et corpus distingué (Bommier-Pincemin, 1999), le premier type de corpus contenant "l'ensemble des textes sur lesquels porte effectivement l'analyse, pour lesquels on attend des enseignements, des résultats" et, le second, "un groupe de textes du corpus d'étude que l'on veut caractériser dans leur cohésion d'ensemble, par rapport au reste du corpus d'étude". Le corpus distingué est donc tout à la fois un sous-corpus du premier et un corpus en soi. Un corpus d'apprentissage est donc un corpus d'étude. En son sein on peut trouver des corpus distingués chacun correspondant au grain habituellement retenu par un chercheur pour y accomplir une analyse sur un phénomène précis. Ainsi l'ensemble des forums d'une formation est un objet d'étude fréquent dans notre domaine, tout comme une session d'une heure de travail collaboratif dans un environnement audio-graphique synchrone. Comme le montrent les exemples de la section 3, un corpus d'apprentissage sera souvent un méga corpus renfermant des dizaines de corpus distinguables.

La figure 3 schématise "environnement utilisateur" et "structures des données" en rapport avec un corpus distinguable de la plateforme Mulce, en cours de développement. La sélection du corpus s'opère en cheminant soit à travers la structure décrivant la hiérarchie des corpus et sous-corpus, soit à partir du scénario pédagogique dans un format structuré (dans le niveau 1 de Mulce-struct, partie "scénario pédagogique"). Ce cheminement a conduit à sélectionner une session de travail collaboratif opérée dans un environnement audio-graphique synchrone (Chanier et al., 2006).

Le corpus distinguable contient notamment un ensemble de données structurées (avec liens vers la hiérarchie de corpus, le descriptif de l'activité correspondant à la session, le descriptif des acteurs de la formation impliquée, l'ensemble des interactions provenant des différentes modalités de communication de l'environnement de travail organisées suivant un schéma dont nous parlerons dans la section suivante) et une ressource primaire sous forme de vidéogramme. Une procédure d'alignement permet au chercheur de sélectionner un ensemble d'échanges entres les participants et de jouer la partie vidéo correspondante de façon à retrouver le contexte immédiat de l'activité.

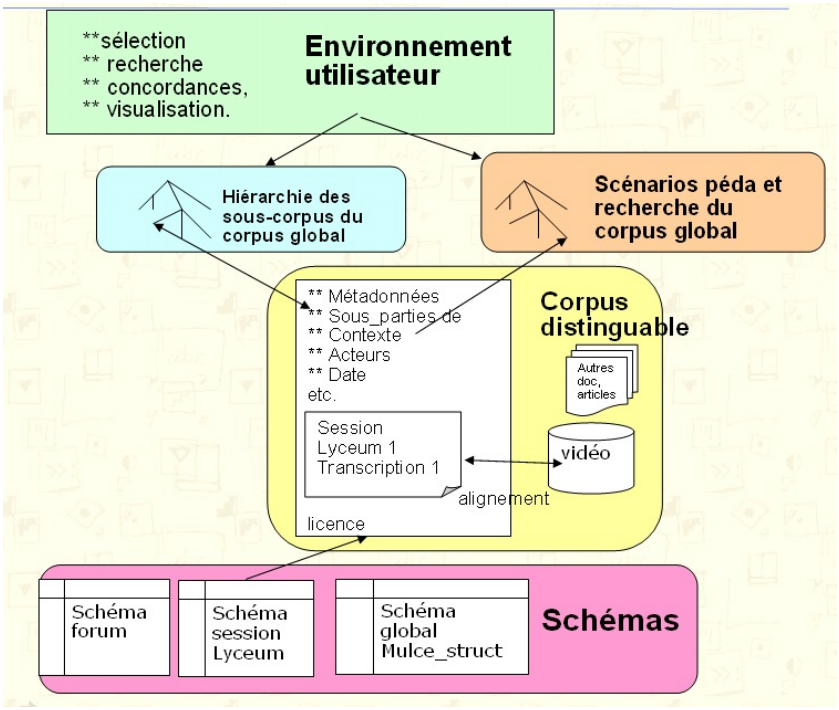

Figure 3 - L'environnement d'utilisation d'un sous-corpus du corpus d'apprentissage

Dans son environnement l'utilisateur peut également sélectionner des outils de recherche de patrons d'interactions au sein de la session (Betbeder et al., 2007a) ou de visualisation (Betbeder et al., 2007b) afin d'afficher, par exemple, la part d'utilisation pour chaque acteur de telle ou telle modalité (audio, clavardage, traitement de texte, carte conceptuelle, etc.).

Il dispose ainsi d'un environnement évolué d'étude d'une situation d'apprentissage remise en contexte et peut également exporter le corpus distinguable afin d'y retravailler avec ses propres outils en utilisant le format d'échange évoqué précédemment.

Nous sommes aux termes du parcours qui nous a permis de fixer le cadre général englobant la notion de corpus d'apprentissage, ses constituants et sa structuration. Nous pouvons maintenant examiner la démarche de construction d'un corpus d'apprentissage en détaillant scénario pédagogique structuré, instanciation des acteurs et environnements.

3. Construction du corpus d'apprentissage d'une formation

Nous présentons dans cette partie, de façon plus détaillée, la structuration d'un corpus d'apprentissage à travers des exemples. Mais avant de décrire cette construction, nous présentons les trois ensembles de données issus des expérimentations "Simuligne", "Copéas" et "Tridem", qui sont à la disposition du projet Mulce.

3.1. Ensembles de données

Nous donnons ici des informations synthétiques sur les 3 ensembles de données déjà en notre possession et qui seront transformés, en corpus d'apprentissage, au sens défini en section 2. Le tableau 1 met en valeur la diversité des contextes pédagogiques (institutions, formations, apprenants, domaines d'apprentissage), des environnements technologiques utilisés (synchrones / asynchrones) seuls ou combinés, le volume, la dispersion et l'hétérogénéité des données recueillies et permettant de décrire le scénario pédagogique, les interactions et productions et les données de recherche.

	Simuligne	Copéas	Tridem
Institutions*	UFC, OU	OU, UFC	CMU, UFC, OU
Participants	- 1 coordinateur - 10 natifs (UFC), - 40 apprenants (OU) - 4 tuteurs (OU). - 4 groupes de 12 - 1 groupe global (60)	- 14 apprenants (UFC) - 2 tuteurs (OU) - 2 groupes de 7+1	- 28 apprenants : 13 USA, 10 FR, 5 GB - 5 tuteurs - 10 tridems,
Environnements technologiques	Asynchrone (WebCT)	Synchrone:(Lyceum) Asynchrone:(WebCT)	Blog Synchrone:(Lyceum)
Interactions Devoirs rendus	- 2686 mess. forum, - 4062 courriels - 5680 tours de clavardage - 93 doc. textuels, - une image - 28 fichiers audio	- 5506 tours de parole audio (8h29 en temps cumulé) - 1529 tours de clavardage - 16 séances Lyceum	-11 blogs archivés (avec 610 messages et 127 photos), - 1030 tours de clavardage - 19 séances Lyceum - 8 éval. indiv.
Productions affichées	342 pages web incluant 115 images et 44 fichiers audio	Documents, cartes conceptuelles et tableaux blancs s	10 documents, 4 cartes concept. 51 tableaux blancs
Ressources pédagogiques	guide apprenant guide tuteur guide natifs	guide apprenant guide tuteur	guide apprenant
Scénario	28 activités réparties en 7 étapes / 12 semaines,	8 activités sur 10 semaines	4 activités sur 10 semaines
Questionnaires, Entretiens	12 questionnaires apprenants,	- 14 quest. app., - 7 entretiens, - 9 Critical event recall (8 app., 1 tuteur)	26 pré-questionn., 13 post-questionn., 13 post-entretiens (12 app., 1 tuteur)
Taille	Total : 650 Mo : - 30 000 fichiers répartis dans 2708 dossiers	Total : 35,3 Go : - 37 vidéos (27h) - 512 autres fichiers dans 117 dossiers.	Total : 7,37 Go : - 16 vidéos (20h) - 939 fichiers
Cession droits	Oui	Oui	Oui

Tableau 1 • Description synthétique des 3 ensembles de données

*UFC (Univ. de Franche-Comté), OU (Open Univ.), CMU (Canerge Melon Univ.)

La dernière ligne du tableau indique que nous possédons, pour chacun de ces ensembles de données, les contrats de cession des droits de la part des acteurs.

Nous nous appuierons ensuite sur des extraits de Simuligne, tant pour le scénario pédagogique, que pour l'instanciation, pour illustrer les différentes parties structurant un corpus d'apprentissage. Nous soulignons la souplesse du modèle en espaces de travail qui permet d'épouser par exemple la structuration du scénario pédagogique.

3.2. Scénario pédagogique comme contexte des données

Le scénario pédagogique est la partie du corpus d'apprentissage qui décrit la formation prescrite : l'organisation en étapes et sous-étapes de l'activité, leurs dépendances, les rôles associés (ex : apprenant, natif, tuteur, etc.), ainsi que les environnements (ressources, espaces et outils) nécessaires à leur déroulement.

Nous préconisons l'utilisation de la spécification IMS-LD (IMS-LD, 2003) comme format de description du scénario. D'une part parce que ce formalisme est utilisé et répandu dans la communauté, d'autre part parce que la structuration des activités du scénario pédagogique (prescrit) peut être reprise dans la partie instanciation pour systématiser la structuration des données recueillies à l'issue du déroulement de la formation (cf section 3.3). Nous avons également choisi IMS-LD parce qu'il est basé sur une spécification XML indépendante des plateformes de téléformation.

La description du scénario inclut la description des environnements dans lesquels les activités ont lieu ; les rôles (qui seront incarnés par les différents acteurs) ; la méthode (IMS-LD, 2003) permet de décrire l'ordonnancement de l'ensemble des activités et tâches à réaliser au cours de la formation.

La figure 4 illustre un extrait de la formation Simuligne décrite en IMS-LD à l'aide de MOTPlus (MOTPlus, 2005). Elle met en évidence les rôles (Apprenant et Tuteur) et la décomposition de la structure d'activité "E0 étape 0 et Interculture Prep" constituée d'une part de la séquence de deux structures d'activité ("Début E0" et "Suite E0") et d'autre part "Interculture Prep" (Préparation).

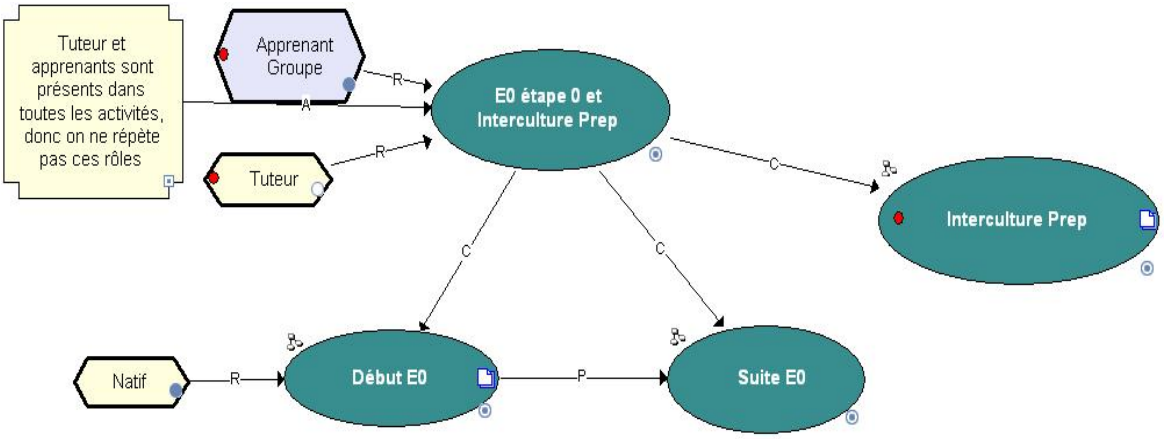

Figure 4 • Niveau étape : Structure de l'activité
Schéma MOTPlus[1] (MOTPlus, 2005) de la formation Simuligne

La figure 5 offre une vue plus complète de la structure d'activité "Début E0". Elle montre en particulier des activités d'apprentissage (E0A1 et E0A2), de support (Correction biographie), les environnements (Forum "Principal" et "E0A3_RDV_bavardage", "A rendre" et "Courriel") inclus dans "l'environnement groupe" et dans lesquels auront lieu ces activités, les intrants et produits de ces activités (ex : Biographie) et les rôles des acteurs (ex : Apprenant, Natif).

Nous reprendrons sur la figure 9 de la section 3.3.2 l'exemple du forum "Principal" pour illustrer la manière dont un message est structuré dans la partie instanciation d'un corpus d'apprentissage.

[1] (1) Le formalisme (MOTPlus) s'interprète comme suit : les ellipses représentent des tâches, les rectangles des objets ou concepts. Les liens C sont des liens de Composition, P de Précédence, et IP Intrant/Produit. Les ellipses foncées dénotent des structures d'activité, le clair des activités d'apprentissage, le jaune les activités de support, le bleu foncé des objets d'apprentissage ou des produits, le bleu moyen des environnements, le bleu clair des services.

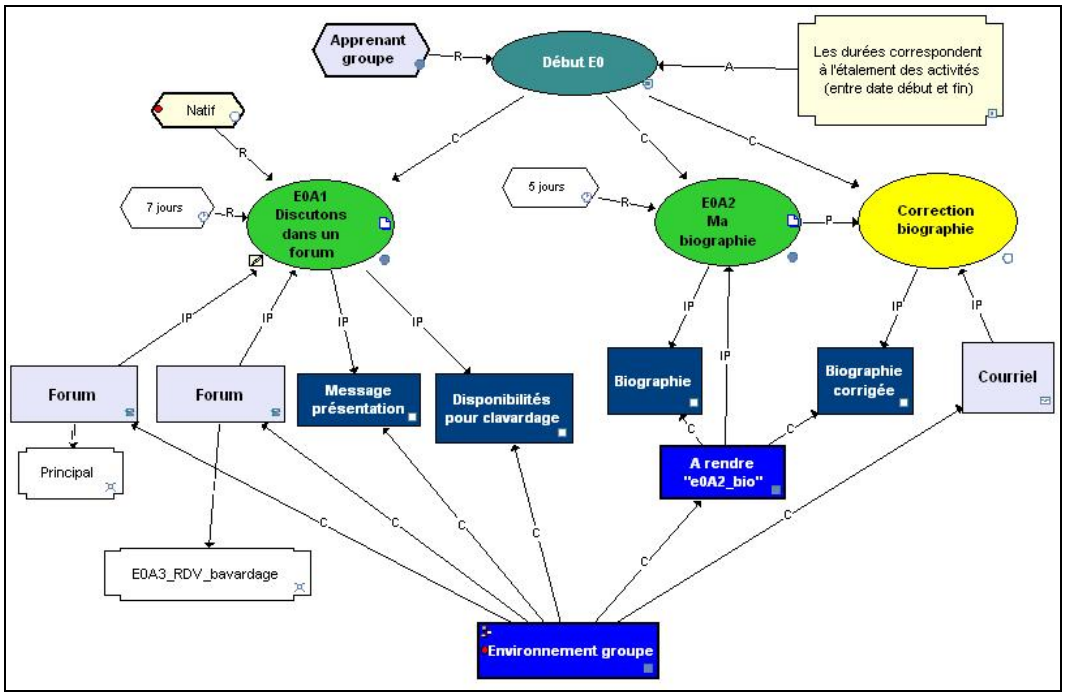

Figure 5 • Niveau structure d'activité : activité, environnements, rôles
Schéma MOTPlus d'une structure d'activité de la formation Simuligne

3.3. Instanciation de la formation

Le terme "Instanciation" fait référence à la modélisation orientée objet. Par analogie, nous pouvons considérer que le scénario pédagogique est une classe tandis que le résultat du déroulement de la formation par un groupe donné d'acteurs, dans une plateforme technologique donnée, à une période donnée, constitue une instance de cette classe. Un même scénario peut être joué par différentes cohortes d'acteurs (formateurs, tuteurs, apprenants, natifs, etc.) au cours d'une même période (en parallèle) ou sur des périodes successives (typiquement des semestres universitaires).

Le cas de Simuligne apporte un nouveau degré de complexité : le scénario comportait 2 niveaux : un niveau global dans le groupe "Monde" dans lequel se déroulait l'activité Interculture mettant en relation tous les acteurs des différents "groupes de base" ayant joué à un niveau local en parallèle une Simulation Globale. Le résultat du déroulement de Simuligne donne donc lieu à une instance du "Monde" et quatre instances du groupe de base nommées respectivement : "Aquitania", "Gallia", "Lugdunensis" et "Narbonensis", chacune peuplée par ses acteurs et recueillant leurs traces, interactions et productions.

Actuellement il n'existe pas de standards définissant la structure des interactions issues de formations en ligne. Cet aspect est détaillé dans la partie 4.2. Pour notre projet, l'instanciation constitue le cœur du corpus : pour pouvoir analyser un corpus il faut pouvoir disposer à la fois d'éléments sur les acteurs et groupes mais également disposer des interactions. Celles-ci doivent pouvoir être liées au contexte pédagogique défini précédemment. Nous proposons donc une structure générique pouvant contenir les informations sur les groupes et acteurs ainsi que les traces, interactions et productions qui résultent du déroulement du scénario.

Le caractère générique de cette structure doit assurer son indépendance vis-à-vis de la plateforme de téléformation utilisée. En effet, chaque plateforme connait une durée de vie limitée dans le temps, évolue d'une version à l'autre, empêchant parfois l'accès à des données anciennes, enregistrées dans un format obsolète. La constitution d'un corpus d'apprentissage ayant des visées à plus long terme, il est essentiel que son accès ne soit pas mis en péril par l'évolution des plateformes.

Cette nécessaire généricité induit un besoin d'expressivité capable de décrire aussi bien des activités (et donc des interactions, traces et productions) issues de scénarios pédagogiques variés impliquant des environnements tant synchrones qu'asynchrones.

Alors que le scénario pédagogique d'une formation, décrit de façon abstraite, avant le déroulement de la formation, l'activité type d'un groupe en définissant les rôles sans les attribuer nominativement et en citant uniquement le type des outils utilisés (ex : forum, clavardage), lors de l'instanciation, il s'agit de préciser les différents acteurs (personnes physiques), leur organisation dans les différents groupes et les outils concrets (ex : forum et clavardage de WebCT). Les interactions et productions peuvent alors être décrites et faire référence aux acteurs.

3.3.1. Définition des acteurs

Pour la définition des acteurs, nous avons choisi d'utiliser la spécification IMS Enterprise (IMS-Enterprise, 2007). Cette spécification est initialement prévue pour gérer les inscriptions des acteurs dans les cours au sein d'une plateforme, ainsi que les informations sur les cours et sur les utilisateurs. Nous restreignons ici son utilisation à la définition des acteurs, (tuteurs, natifs, apprenants) à celle des groupes et à l'affectation des acteurs à un ou plusieurs groupes. Les informations nécessaires à la compréhension des interactions (en excluant les données personnelles non diffusables) sont décrites de même que les différents surnoms ou login sous lesquels on peut retrouver l'acteur dans les interactions/traces. Le produit final est donc un document XML respectant le schéma IMS Enterprise.

Pour la définition de l'environnement de travail, nous proposons une spécification sous forme de schéma XML (cf. figure 6) décrite dans la partie suivante. Ce schéma donne la structure du document XML associé dans lequel seront stockées les interactions.

3.3.2. Instanciation des environnements de travail

Nous avons choisi de décrire un environnement technique (dispositif utilisé pour la formation) comme un espace de travail correspondant à un lieu dans lequel des acteurs disposent d'outils (dotés de certaines fonctionnalités explicitées) et interagissent dans une période donnée. Cet espace de travail peut inclure des sous espaces de travail.

La spécification que nous proposons se veut la plus exhaustive possible quant aux outils de communication utilisables et quant à leurs fonctionnalités. Dans le cas où des outils ou fonctionnalités n'auraient pas été pris en compte par notre spécification, le descripteur peut enrichir ce schéma (par l'insertion de nouveaux outils ou fonctionnalités) pour lui permettre de décrire une variété croissante de corpus.

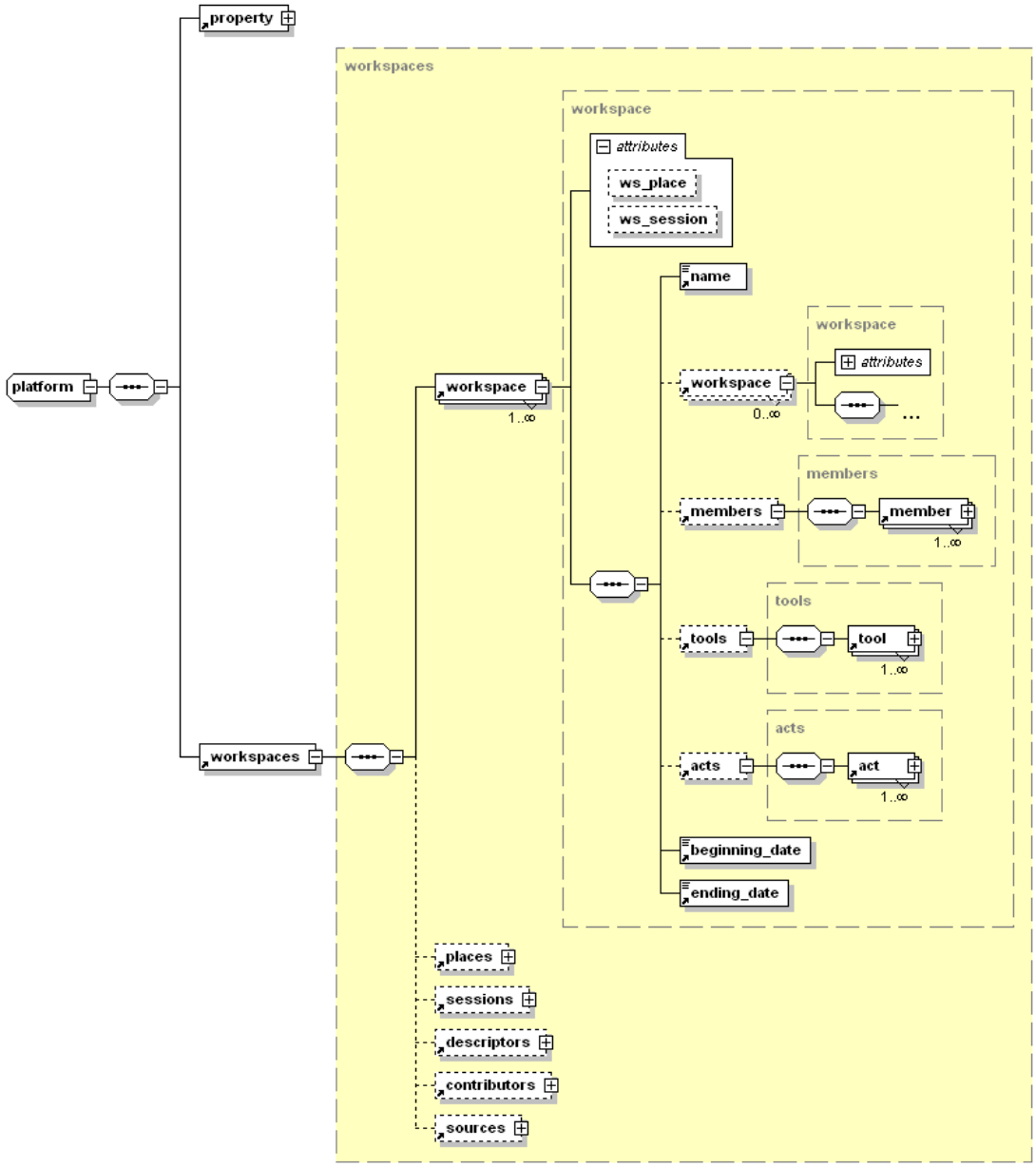

Figure 6. Extrait du schéma XSD d'instanciation d'un environnement

De plus, la définition récursive de la notion d'espace de travail permet au descripteur de corpus de choisir le niveau de granularité qu'il souhaite. Ainsi, l'espace de travail peut donc aussi bien correspondre à une plateforme, à une étape, à une activité qu'à une session de travail (notion correspondant plus à des formations synchrones).

Chaque espace de travail (Workspace) est lié par son nom (élément name) à une structure d'activité du scénario pédagogique qui définit le contexte des traces recueillies dans cet espace de travail. Outre les membres (référence aux acteurs inscrits dans la formation, définis dans la partie précédente) et les dates de début et de fin, un espace de travail contient une liste déclarative des espaces/outils (tools) d'interactions disponibles et la liste des actes (acts), chacun d'entre-eux, faisant référence à l'un des espaces/outils déclarés. Ces espaces/outils sont typés (Forum, Clavardage, Audio, Vote, etc.) et possède un nom permettant de différencier deux outils de même type, tel que dans Simuligne, par exemple : deux forums "Principal" et "E0A3_RDV_Bavardage" utilisés à l'intérieur d'une même activité "E0A2". A chacun de ces espaces/outils correspondront ensuite des actes stockés dans l'élément acts.

Pour permettre une analyse globale des différents actes, il est indispensable de définir une partie générique à tout acte. C'est précisément l'objet de la figure 7 présentant notre définition d'un acte. Tout acte possède donc un identifiant (attribut id), une référence à l'espace/outil dans ou par lequel il a été déposé (ref_tool), une référence à l'auteur (author), une date de début (beginning_date) et un sélecteur de type d'acte.

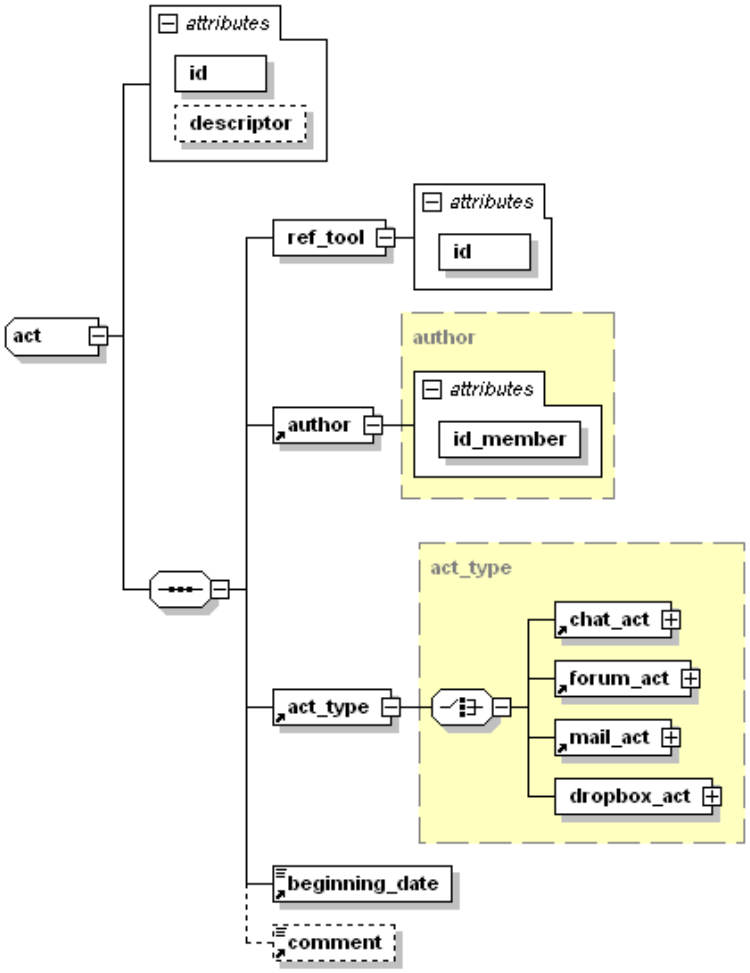

Figure 7. Extrait du schéma XSD – notion d'acte

Ce sélecteur (act_type), précisant le type d'acte (clavardage, forum, mail, etc.) permet d'ajouter à la partie générique d'un acte, une partie propre à ce type d'acte. L'exemple d'un acte de type forum est donné à la figure 8.

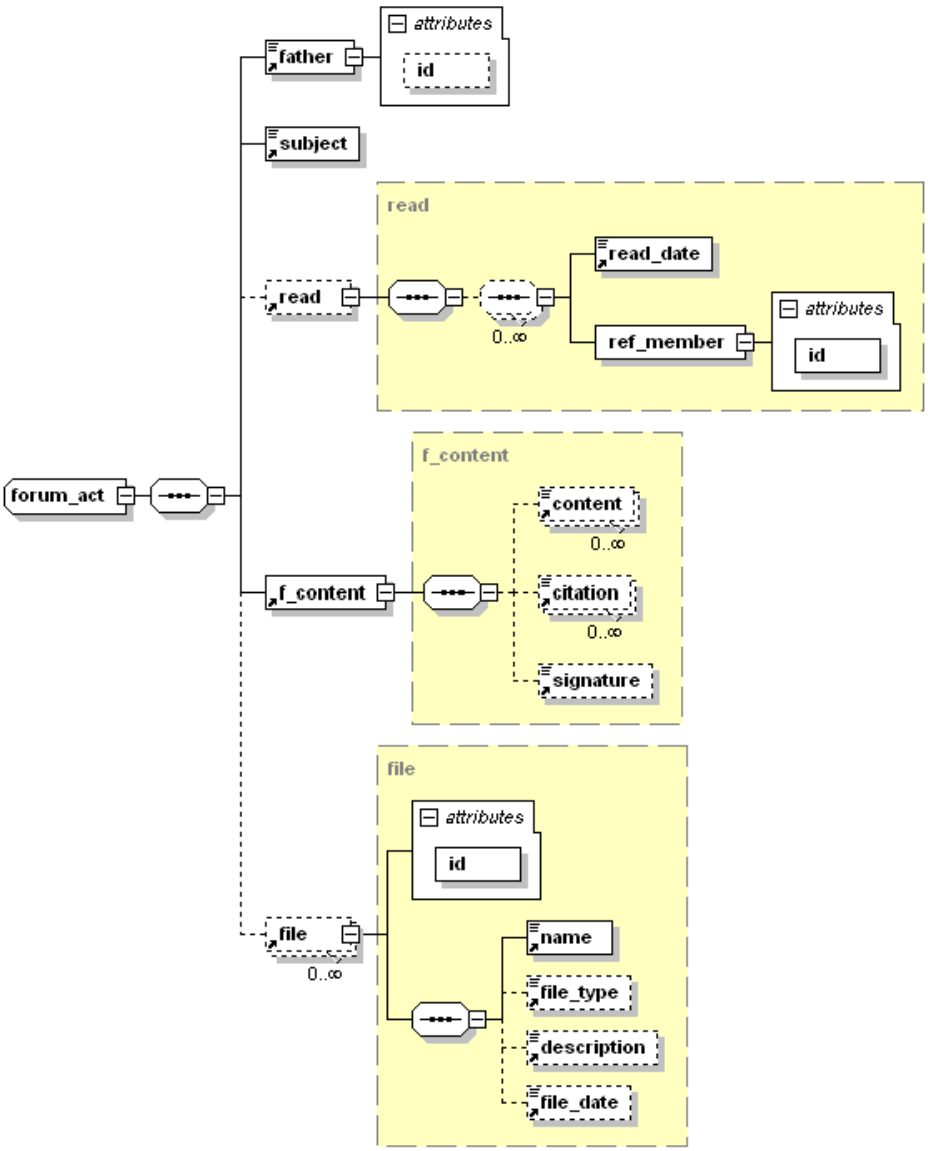

Figure 8. Extrait du schéma XSD – notion d'acte de forum

Un acte de forum (cf. figure 8) correspond au dépôt d'un message dans un fil de discussion. Ce message a un sujet, par exemple le titre du fil de discussion et peut contenir la référence au message père, message auquel il répond. La lecture de ce message peut être décrite par une date ainsi qu'une référence au membre lecteur. Le contenu du message peut se composer de trois éléments : le contenu même du message, une citation et une signature. Il peut également contenir un ensemble de fichiers attachés (nom, type, date du fichier).

La figure 9 droite correspond à la structuration XML de deux actes de forum déposés dans la plateforme WebCT au cours de la formation Simuligne (figure 9 gauche). La mise en évidence des champs "numéro de message", "sujet" et "date" sur chacune de ces figures montre bien que les informations recueillies sont conservées et structurées.

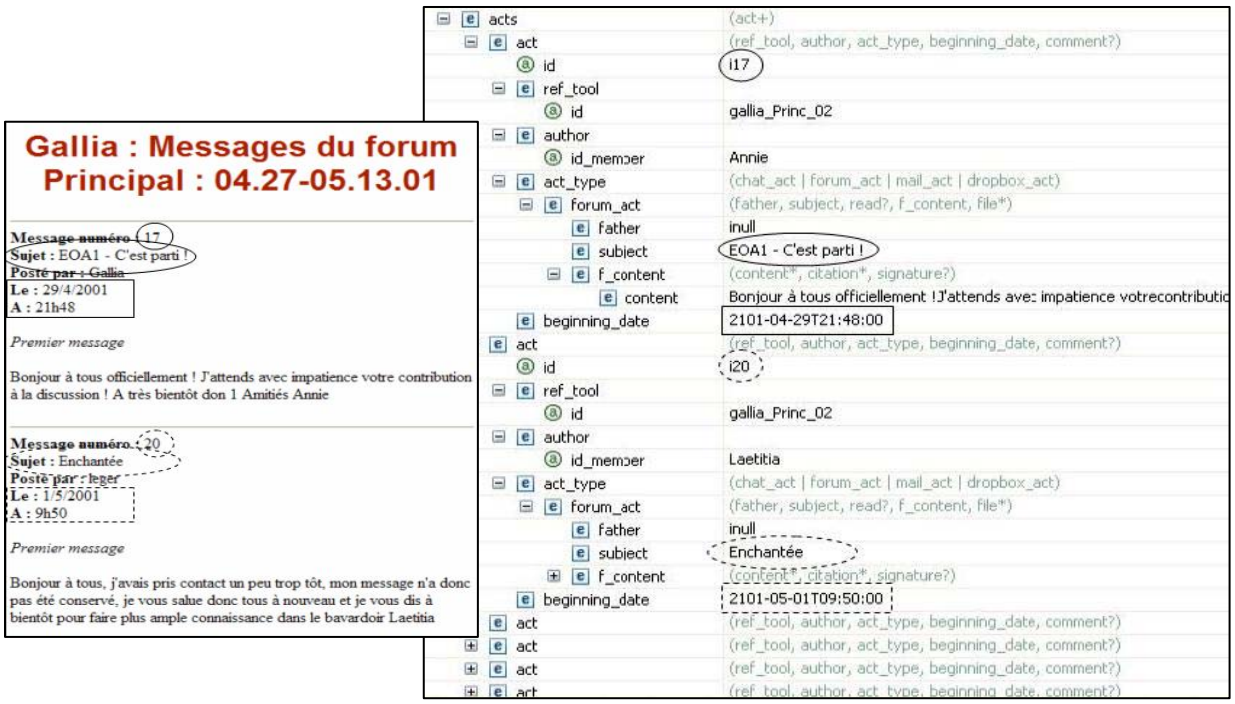

Figure 9 gauche. Exemple de forum à décrire
Droite : Exemple de description d'actes de forum de la figure gauche

4. Les conditions de réalisation de l'échange et de l'accès libre

Dans cette partie, nous abordons les conditions qui nous paraissent prévaloir pour réaliser l'échange et l'accès libre. Nous retenons trois sujets (droits et éthique, standard et normes, analyse de traces) qui méritent l'attention des chercheurs de notre domaine même s'ils sont aujourd'hui peu explorés ou l'ont été seulement très récemment. Pour chacun, nous indiquons l'approche des membres du projet Mulce en la positionnant par rapport à ceux de la communauté et à d'autres communautés connexes. L'obtention de réponses cohérentes, simples, applicables même si elles sont limitées, est indispensable au partage de corpus d'apprentissage et des recherches sur ces objets.

4.1. Droits et éthique

Nous avons mentionné en section 1, l'intérêt de premier plan que certains milieux concernés par l'apprentissage humain et les technologies expriment aujourd'hui sur les questions des droits et de l'éthique. Ces discussions ont pu être éludées jusqu'à récemment pour diverses raisons, la première tenant au fait que la mise sous clef des données produites dans les laboratoires de recherche évitait d'avoir à se poser la question. A un tout autre niveau, dans notre domaine, le fait d'utiliser des systèmes spécifiques plutôt que des systèmes généralistes (voir section 4.3) permettait de travailler sur des traces générées automatiquement, ce qui réduit d'emblée l'éventualité de l'identification des individus et, par la même de l'atteinte éventuelle à leurs droits.

Le passage à l'accès libre et à l'échange de données (données au sein desquelles les interactions, en particulier verbales, des acteurs de situation d'apprentissage) change radicalement les choses, en plaçant les sujets sur l'éthique et les droits avec une priorité élevée dans l'agenda de la recherche. Sous quelle perspective faut-il envisager d'aborder cela ? L'exemple actuel des débats sur les droits dans le milieu de l'édition scientifique et des archives ouvertes (Chanier, 2004) montre l'équilibre qu'il faut préserver, en examinant, d'une part, avec intérêts mais circonspection le point de vue de certains juristes, appuyés par les médias, qui essayent de faire jouer des règlements anciens, non appropriés aux nouveaux contextes pour la défense d'intérêts corporatistes et, d'autre part, la nécessité d'intégrer le nouveau contexte apporté par la société en réseaux (Lessig, 2001).

Alors que le positionnement des communautés scientifiques responsables du développement des archives ouvertes est le résultat de plus de 30 ans de réflexion et développement, notre communauté débute seulement ses réflexions. Les avancées s'accompliront tout à la fois en inscrivant ce thème dans notre recherche, en restant cohérent avec ceux que d'autres communautés ont élaborés sur des thématiques plus large que les nôtres, et en restant pragmatique, au sens où la recherche de solutions simples et immédiatement applicables doit s'imposer afin de ne pas ralentir le travail des chercheurs. Pour ne prendre qu'un exemple, nombres d'universités étrangères ont depuis longtemps adopté des positions sur l'éthique propres à leurs institutions, qui contraignent fortement le montage des expérimentations et plus encore la diffusion des résultats. Ces règlements édictés dans un contexte ancien ont certes besoin d'être considérés, mais surtout reconsidérés à la lumière des nouveaux contextes et enjeux.

Nous nous contenterons ici de lister les différents points qui font l'objet d'un travail spécifique au sein du projet Mulce, en coopération avec d'autres projets, et sont en rapport avec le constituant "Licence" discuté en section 2.

- La licence de dépôt, celle que les responsables de corpus signeront au moment du dépôt de corpus ou d'analyses. Elle stipulera la reconnaissance du travail des auteurs / contributeurs ayant permis de créer l'objet du dépôt, la cession non exclusive dans les respect des droits sur les matériaux engagés (et l'éventuel retrait de certains matériaux litigieux), l'acceptation de la mise en accès libre conformément à la licence d'utilisation, en distinguant éventuellement au sein des matériaux ce qui est directement accessible aux internautes anonymes et ce qui doit être réservé aux chercheurs, éducateurs identifiés. Enfin, le responsable devra garantir être en possession de l'ensemble des contrats de cession des droits signés par les acteurs de la formation et de la table reliant codes d'acteurs dans la partie diffusée et identités dans la partie conservée par l'institution d'origine (cf. licence privée en figure 3).
- La licence d'utilisation pour les utilisateurs en distinguant celle pour les anonymes et celle pour les institutionnels (chercheur / pédagogue). Cette licence sera choisie parmi les différentes variantes des Creative Commons.
- La procédure d'anonymisation (Reffay et Teutsch, 2007) (Mondada, 2005) dont la spécification doit trouver l'équilibre entre le respect d'un certain type d'anonymat des acteurs, la possibilité de semi-automatiser cette procédure pour retoucher aux ressources primaires en évitant le remplacement de tokens par des informations qui pourraient biaiser l'interprétation des chercheurs (en remplaçant un patronyme par un autre inventé) et en indiquant précisément les sections des données ayant été modifiées.
- Les contrats type de cession des droits, avec d'éventuels libellés différents suivant les variations de conditions expérimentales qui indiqueront aux responsables de corpus les points sur lesquels ils doivent avoir obtenu le consentement signé des acteurs en garantissant, d'un coté, un certain niveau d'anonymat et, de l'autre, une autorisation d'utiliser leurs productions dans le cadre de la licence d'utilisation.

4.2. Échange de corpus d'apprentissage

Il est difficile d'aborder en quelques mots la question de l'échange, qui a des abords trompeurs. En effet, dans les discours des individus ou institutions du milieu de la recherche ou du monde professionnel sont exprimées avec force les nécessités de concevoir l'échange au niveau international, d'utiliser des formats d'échange standard, voire normalisés. Mais dans les faits, il est fréquent de voir les actes en désaccord avec les propos. Or, on peut sérieusement s'interroger sur la possibilité de discuter formats et langages associés sans avoir une pratique d'utilisation systématique à une échelle significative au sein de communautés, pratique qui amène à saisir, non seulement les limitations de tels formats mais surtout le fait qu'ils soient le résultat d'un long processus de coopération, négociation entre cultures différentes (voir, par exemple, sur le thème du catalogage des ressources pédagogiques (La Passardière et Jarraud, 2004)). La question de l'échange est d'abord un enjeu sociétal avant d'être une question de puissance d'expression ou de sémantique de langages. L'échelle du temps est un

paramètre de tout premier ordre. Ainsi les standards et normes du champ corpus et langage (naturel) mentionnés dans les premières sections de cet article sont le résultat de plus de 20 ans de travaux, commencés avant SGML et encore en développement. Comparativement, les efforts de standardisation en e-formation, sont récents, ce que témoignent certains indices révélateurs .

D'autre part, les institutions du domaine de la e-formation (regroupées dans des consortia comme l'IMS) ont pour priorité de créer les conditions à la mise en place de formations en ligne et à leur duplication sur toute plateforme en garantissant dans un premier temps l'échange à partir de la standardisation des scénarios pédagogiques et des ressources pédagogiques (pour ne mentionner que les 2 points les plus importants des chantiers en cours). Les étapes suivant la conception et l'instanciation, étapes intitulées étape de production (Production), étape de diffusion (Delivery), voire legacy pour la partie administrative, sont à l'état de jachère. Il existe donc une sous-détermination importante entre la conception et la conduite effective (run) d'une formation. L'écart devient abyssal si l'on considère la perspective du chercheur-pédagogue qui aimerait estimer l'efficience de la formation et réfléchir sur la pertinence des dispositifs au regard des actions et produits des acteurs de la formation.

Ceci dit, comme dans la sous-section précédente, nous pensons que ces points doivent faire l'objet d'une discussion approfondie dans notre communauté, en s'appuyant au mieux sur les standards existants, en garantissant la comptabilité avec ceux des communautés plus larges que la notre ou contiguës et avançant pragmatiquement sur une échelle significative.

Voici les standards que nous utilisons ou envisageons d'utiliser dans le projet Mulce.

- Pour le serveur Internet en vue de l'échange des données en accès libre : le protocole de l'Open Archive Initiative (OAI-MPH, 2002).
- Pour les métadonnées les plus générales sur les corpus et sous-corpus, nous écartons la LOM destinée au catalogage des ressources pédagogiques, au profit de la norme ISO générique du Dublin Core Metadata Initiative (DCMI, 2006).
- Pour l'échange des données de types verbales et la description des rôles, langues, impliquées dans les corpus, les standards de l'Open Language Archives Community (OLAC, 2007), de la Text Encoding Initiative (TEI, 2007) pour l'exportation des textes et interactions verbales.
- Pour l'exportation des corpus et sous-corpus entiers, nous avons mentionné en section 2, l'utilisation du standard Content Packaging. Conscients des limites de ce standard qui, du fait de sa conception pour l'agrégation et la désagrégation de ressources pédagogiques, a une gestion des liens et sous-manifestes antynomiques de la mise en relation des différents constituants d'un corpus (cf. section 4.8.4 Package Scope du (IMS-CP, 2004)), nous ne l'utilisons ici que comme empaquetage général des données..

En particulier, les chercheurs ne cherchent plus à trouver un accord sur, par exemple, les codes de transcription de l'oral ou les catégories syntaxiques d'étiquetage des textes mais unifie la façon de les décrire, de les renseigner et les utiliser. On mesure l'intérêt d'une telle approche dans notre domaine pour, par exemple, permettre que les différences de point de vue sur la pertinence de tel ou tel système de codage des interactions dans les forums ou clavardage (en termes socio-affectif, cognitif, métacognitif, tutoriel, etc.), n'empêche pas la circulation des données et des analyses correspondantes dans des formats utilisables par toutes les écoles.

4.3. Analyse des traces et standardisation de la formation en ligne

De nombreux travaux de nos communautés (AIED, EIAH et CSCL) sur l'analyse des traces concernent des environnements spécifiques (i.e. : Aplusix (Bouhineau et al., 01), Logic-ITA (Yacef, 2005)) dans lesquels les productions (algébriques, logiques) sont contraintes par des règles calculables, ce qui en permet une analyse sémantique. Ces environnements proposent des activités d'apprentissage

plutôt individuelles, puisque le contenu, riche, interprétable et didactisé peut être traité pour construire automatiquement une rétroaction (feedback) appropriée.

D'autres utilisent des environnements plus généraux (i.e. : Synergo (Avouris et al., 2003), Moodle (Mazza & Milani, 2005)) sur lesquels les chercheurs informaticiens peuvent intervenir pour définir les traces à générer et leur format. La trace obtenue peut alors être visualisée a posteriori par des chercheurs ou à la volée par les acteurs (mirroring). Elle peut subir des traitements ou analyses automatiques pour construire des indicateurs de suivi (monitoring) ou être comparée à un modèle attendu : caractérisant un succès ou une erreur documentée pour construire une rétroaction permettant le guidage (guiding) de l'acteur au cours de l'activité (Jermann et al., 1999).

Les plateformes de téléformation généralistes se multiplient, mais tentent de se normaliser pour intégrer des contenus pédagogiques indépendants (LOM, SCORM) et sont constituées de modules assez génériques (clavardage, forum, wiki, blogues, etc.). Cette diversification de plateformes semble répondre aux exigences culturelles et institutionnelles tandis que leur standardisation et convergence fonctionnelle (au niveau des composants) offrent une meilleure acceptabilité. Cette maturation des plateformes a permis de multiplier les formations en ligne dans de nombreux domaines.

De très nombreuses expérimentations utilisent ces plateformes généralistes WebCT, Dokeos, Moodle, Lyceum, sans avoir la main sur le processus de génération des traces. Les interactions et productions recueillies concernent, pour une part importante, une communication verbale (textuelle ou non) qui s'inscrit dans un contexte. Les interactions humaines médiées par les plateformes y jouent un rôle essentiel en particulier dans le cas des scénarios d'apprentissage collaboratif. Les analyses quantitatives (Reffay & Chanier, 2003) (Betbeder et al., 2007a) apportent quelques indicateurs utiles pour le suivi de l'activité (durée, volume des interactions et productions, fréquence d'accès), mais le besoin d'instrumenter les analyses qualitatives, portant sur le contenu lui-même (Betbeder et al., 2007b), se fait de plus en plus pressant.

Les disciplines des sciences humaines s'intéressent à ces analyses, et utilisent souvent des outils de traitement automatiques, que nous avons mentionnés en section 1 et 2. Ceux-ci fonctionnent assez bien sur des textes bien formés pour tenter d'extraire la sémantique de textes syntaxiquement corrects, mais pour s'adapter aux multiples formes présentes dans les situations écologiques d'apprentissage, ils doivent être enrichis et intégrer des fonctions d'apprentissage automatique dépendantes du contexte (culture, langue, institution, domaine, activité). Pour progresser dans cette voie, la mise à disposition de corpus d'interactions nous semble indispensable. Ces corpus doivent pouvoir intégrer toutes les formes d'interaction effectivement recueillies (Adam et al., 2007), y compris multimodales, éventuellement transcrites.

Pour représenter un ensemble unifié de traces d'interactions, dans un environnement multi-acteurs, intégrant différents outils ou modalités d'interaction, nous rejoignons Kahrimanis sur la nécessité de définir un format partagé. L'auteur montre dans (Kahrimanis et al., 2006) qu'un tel format est nécessaire pour l'interopérabilité bidirectionnelle entre, d'une part, les plateformes de formation à distance et, d'autre part, les outils d'analyses des traces issues de ces plateformes. De notre côté, puisque nous souhaitons échanger les données en vue de leur analyse, nous mettons l'accent sur la contextualisation et sur les besoins multiples (en terme de format de données) pour leur transcription, annotation, traitement automatique ou analyse. Ces travaux pourront se rapprocher grâce à des coopérations initiées lors d'événements scientifiques tels que le symposium (Reffay, 2007) sur l'échange de corpus organisé à l'occasion du colloque EPAL (EPAL, 2007).

Conclusion

La notion de corpus d'apprentissage a été définie en regard des corpus linguistiques et de traitement automatique de la langue. Elle vient élargir celle des traces dans les domaines d'apprentissage assez formels comme les mathématiques et la logique. Les analyses visées sont celles qui cherchent à améliorer les dispositifs d'apprentissage médiatisés, tant dans leur dimension socio-pédagogique, que technique. L'enjeu est celui de la réplicabilité de ces situations d'apprentissage mais aussi et surtout de la validité externe des résultats de recherche. L'objectif central de ce travail est de proposer un cadre suffisant pour rendre accessible à d'autres, les données et résultats de recherche sur ces dispositifs

pour en permettre la capitalisation, la comparaison ainsi que la remise en cause par d'autres chercheurs. Les situations d'apprentissage collaboratif sont la cible privilégiée de cette proposition puisqu'elle leur permet de décrire l'ensemble des interactions placées dans le contexte de leur genèse (scénarios de recherche, scénario pédagogique, environnement technologique, institution et acteurs). La mise à disposition de corpus issus de situations écologiques est le besoin scientifique et sociétal qui a motivé cette proposition, pour que la recherche sur les interactions (situées) issues de formations en ligne puisse travailler sur des données réelles.

Nous avons déjà entamé la structuration de plusieurs ensembles de données issus de situations pédagogiques et technologiques assez variées dans le domaine de l'apprentissage des langues. La construction du corpus d'apprentissage "Simuligne" nous a permis d'illustrer ici les schémas XML que nous avons réutilisés (IMS-CP, IMS-LD, IMS-Enterprise) et surtout ceux que nous proposons pour structurer les interactions. La notion de workspace permettant de situer des interactions diversifiées (synchrones / asynchrones ; verbales, iconiques, graphiques, etc.) selon des approches différentes (structures d'activité ou espaces d'échange), rend possible la description de multiples corpus distinguables dans un corpus d'apprentissage.

Les principes fondamentaux de la composition et de la structuration d'un corpus ont été énoncés. La dimension éthique et juridique y joue un rôle essentiel puisqu'elle doit protéger les acteurs et contraindre les usages. Le défi de l'anonymisation systématique est incontournable dans le domaine des interactions humaines situées. La structure opérationnelle Mulce-Struct doit encore être affinée pour rendre possibles les diverses fouilles, recherches, étiquetages sur les interactions verbales (textuelles ou audio), ainsi que l'alignement entre les données audio ou vidéo et les transcriptions disponibles. La plateforme Mulce devra intégrer des outils standards de traitement ou de fouille des données pour ouvrir la voie à des analyses intra- et inter-corpus.

Enfin, la convergence des modules de communication intégrés dans les plateformes de téléformation ainsi que les recherches de standardisation de traces pour leur analyse automatique sont des efforts visibles qui confortent les orientations de notre projet Mulce.

Remerciements

Mulce (Échange de corpus d'apprentissage multimodaux, http://mulce.univ-fcomte.fr) est un projet soutenu par l'Agence nationale de la Recherche (ANR-06-CORP-006) dans le cadre du programme "Corpus et Outils de la Recherche en Scien-ces Humaines et Sociales". Il rassemble des équipes des laboratoires LASELDI et LIFC (Université de Franche-Comté), CREET (The Open University) et LIP6 (Uni-versité Paris 6), coordonnées respectivement par Thierry Chanier, Christophe Reffay, Marie-Noelle Lamy et Jean-Gabriel Ganascia.

Références

Tous les liens Internet de cette section ont été vérifiés en date du 1er juillet 2007.

Bibliographie

ADAM J.-M., MICHELET S., MARTEL C., DAVID J.-P., GUERAUD V. (2007). Une infrastructure logicielle pour instrumenter l'expérimentation en EIAH. Dans Nodenot, T., Wallet, J., Fernandes E. (Dir.) Conférence EIAH 2007 : Environnements Informatiques pour l'Apprentissage Humain, Lausanne, Suisse, juin, pp. 449-454.

AVOURIS N., FIOTAKIS G., KAHRIMANIS G., MARGARITIS M., KOMIS V. (2007). Beyond logging of fingertip actions: analysis of collaborative learning using multiple sources of data. Journal of Interactive Learning Research JILR, vol. 18(2), Special Issue: Usage Analysis in Learning Systems : Existing Approaches and Scientific Issues. pp.231-250 http://hci.ece.upatras.gr/pubs_files/j46_Avouris_etal_JILR_2007.pdf.

BELZ, J. A. (2004).Learner corpus analysis and the development of foreign language proficiency. System, pp. 32 . 577–591

BERLIN (2003). Appel de Berlin d´octobre 2003 sur "Open Access to Knowledge in the Sciences and Humanities". Institut Max Planck : Munich. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

BETBEDER M.-L., TISSOT R., REFFAY C. (2007a). Recherche de patterns dans un corpus d'actions multimodales. Dans Nodenot, T., Wallet, J., Fernandes E. (Dir.) Conférence EIAH 2007 : Environnements Informatiques pour l'Apprentissage Humain, Lausanne, Suisse, juin, pp. 533-544. http://edutice.archives-ouvertes.fr/edutice-00158881

BETBEDER M.-L., CIEKANSKI M., GREFFIER F., REFFAY C., CHANIER T. (2007b). Comment spécifier, codifier et représenter les interactions multimodales synchrones issues de formations en ligne. Communication au Colloque EPAL : Echanger Pour Apprendre en Ligne, Grenoble, juin.

BOMMIER-PINCEMIN Bénédicte (1999) Caractérisation d'un texte dans un corpus : du quantitatif vers le qualitatif. Chapitre VII : § A "Définir un corpus" de la Thèse de Doctorat en Linguistique "Diffusion ciblée automatique d'informations :conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des document" , Université Paris IV pp. 415-427. http://www.revue-texto.net/Corpus/Publications/pincemin_ad_1999.pdf

BOUHINEAU D., NICAUD J.-F., PAVARD X., SANDER E. (2001) Un micromonde pour aider les élèves à apprendre l'algèbre. Actes des sixièmes journée EIAO, Sciences et Techniques Educative, Hermès, Paris, octobre.

CHAMPIN P.-A., PRIE Y., MILLE A., (2003). MUSETTE: Modelling USEs and Tasks for Tracing Experience. ICCBR'03: Workshop ''From structured cases to unstructured problem solving episodes'' ICCBR'03: NTNU, 279–286 : http://liris.cnrs.fr/yannick.prie/download/iccbr2003b.pdf

CHANIER, T. (2004) Archives ouvertes et publication scientifique. Comment mettre en place l'accès libre aux résultats de la recherche ? Paris : L'Harmattan. 186p. http://archivesic.ccsd.cnrs.fr/sic_00001486.html

CHANIER, T. et VETTER, A. 2006). "Multimodalité et expression en langue étrangère dans une plateforme audio-synchrone". Apprentissage des langues et Système d'Information et de Communication (Alsic), vol. 9. http://alsic.u-strasbg.fr/v09/chanier/alsic_v09_08-rec3.htm

CHAPELLE, C.A. (2004). "Technology and second language learning: expanding methods and agendas. System, 2004, pp. 593-601.

CORBEL, A., GIRARDOT, J.-J., LUND, K. (2006). A method for capitalizing upon and synthesizing analyses of human interactions. In (eds.) W. van Diggelen & V. Scarano, Workshop proceedings Exploring the potentials of networked-computing support for face-to-face collaborative learning. EC-TEL 2006 First European Conference on technology Enhanced Learning, October 1, Crete, pp. 38-47

DE LA PASSARDIERE, B. et JARRAUD, P. (2004). ManUeL, un profil d'application du LOM pour C@mpuSciences. Revue STICEF, Vol. 11, http://sticef.org. http://sticef.univ-lemans.fr/num/vol2004/passardiere-11/sticef_2004_passardiere_11.htm

GOODWIN, C. et DURANTI, A (1992). Rethinking context: an introduction. In Duranti, A et Goowin, C. (Dir.) Rethinking context. Language as an interactive phenomenon. Camridge University Press : Cambridge. p 1-42.

GRANGER S., VANDEVETER A., HAMEL M-J (2001). Analyse de corpus d'apprenants pour l'ELAO base sur le TAL. Traitement automatique du langage (Tal), vol. 42, 2. pp. 609-621.

HENRI, F. & CHARLIER, B. (2005). "L'analyse des forums de discussion Pour sortir de l'impasse". In Baron G-L., Bruillard E., Sidir M. (Dir.) Symposium Symfonic "formation et nouveaux instruments de communication. Amiens, janvier : Université de Picardie. http://archive-edutice.ccsd.cnrs.fr/edutice-00000897

JACOBSON , M. (2004). Corpus oraux en linguistique de terrain. Traitement automatique du langage (Tal), vol. 45, 2. p 63-88

KAHRIMANIS G., PAPASALOUROS A., AVOURIS N., RETALIS S. (2006). A Model for Interoperability in Computer-Supported Collaborative Learning. ICALT 2006 - The 6th IEEE International Conference on Advanced Learning Technologies. Kerkrade, Netherlands, p. 51-55. http://hci.ece.upatras.gr/pubs_files/C114_Kahrimanis_etal_ICALT2006.pdf

KERN, R., WARE, P. & WARSHAUER, M. (2004). Crossing frontiers: new directions in online pedagogy and research. Annual Review of Applied Linguistics, Vol. 24. pp. 243-260.

LESSIG, L. (2005). L'avenir des idées. Le sort des biens communs à l'heure des réseaux numériques. Presses Universitaire de Lyon : Lyon. Traduction de The Future of Ideas, 2001.

MAZZA R., MILANI C. (2005). Exploring Usage Analysis in Learning Systems: Gaining Insights From Visualisations. In Workshop on Usage analysis in learning systems. 12th International Conference on Artificial

Intelligence in Education (AIED 2005). Amsterdam, The Netherlands. 18 July 2005. pp. 65-72. http://www.inf.unisi.ch/assistants/ mazza/Web_area/Pubblicazioni/AIED05/aied-ws2005.pdf

MONDADA, L. (2005). Constitution de corpus de parole-en-interaction et respect de la vie privée des enquêtés : une démarche réflexive. Rapport sur le projet "Pour une archive des langues parlées en interaction". Université Lyon 2 et CNRS. 43 p. http://icar.univ-lyon2.fr/projets/corinte/documents/%20%20Mondada_juridique_MARS05.pdf

NORAS M., REFFAY C., BETBEDER M.-L. (2007). Structuration de corpus de formation en ligne en vue de leur échange, Dans Nodenot, T., Wallet, J., Fernandes E. (Dir.) Conférence EIAH 2007 : Environnements Informatiques pour l'Apprentissage Humain, Lausanne, Suisse, juin, pp. 59-64 . http://edutice.archives-ouvertes.fr/edutice-00154372

PLANTIN C., MONDADA, L. et al. (2005). Statuts juridiques, formats et standards, représentativité. Rapport sur le projet "Pour une archive des langues parlées en interaction." Université Lyon 2 et CNRS. http://icar.univ-lyon2.fr/projets/corinte/bandeau_gauche/Projets/rapport_Archives.pdf

REFFAY C. CHANIER, T. (2003). How social network analysis can help to measure cohesion in collaborative distance-learning. In Procs. of Computer Supported Collaborative Learning Conference (CSCL'2003), Bergen, Norway, pages 343-352, June 2003. Kluwer Academic Publishers : Dordrecht(nl). http://edutice.archives-ouvertes.fr/edutice-00000422

REFFAY C. (2007). Symposium : Corpus d'apprentissage en ligne : Conception, réutilisation, échange, Colloque Echanger Pour Apprendre en Ligne (http://w3.u-grenoble3.fr/epal/), Grenoble, juin. http://mulce.univ-fcomte.fr/epal_symposium/

REFFAY C., TEUTSCH P. (2007). Anonymisation de corpus réutilisables. Prépublication, soumise à EIAH2007. 12 pages. http://edutice.archives-ouvertes.fr/edutice-00158877

SALMON-ALT, ROMARY, L. & PIERREL, J.-M. (2004). "Un modèle générique d'organisation des corpus en ligne". Traitement automatique du langage (Tal), vol. 45, 3. pp. 145-169.

SCHEGLOFF , E.A. (1992). In another context. In Duranti, A et Goowin, C. (Dir.) Rethinking context. Language as an interactive phenomenon. Camridge University Press : Cambridge. p 191-229.

VALCKE, M. & MARTENS, R. (2006). Methodological Issues in Researching CSCL , Special issue of Computers & Education, Vol. 46, 1. pp. 1-104.

YACEF K. (2005). The Logic-ITA in the classroom: a medium scale experiment. International Journal on Artificial Intelligence in Education, vol 15, pp 41-60.

Références à des logiciels

MOTPlus (2005). Modélisation par Objets Typés, MOTPlus version 1.6.3. Editeur de modèles de connaissances et de scénarios pédagogiques, développé au LICEF (Laboratoire en Informatique Cognitive et Environnement de Formation de la Télé-Université du Québec). http://www.licef.teluq.uqam.ca/fr/realisations/mot1.htm

Références à des sites Internet

CLAPI (2007) Site de la banque de corpus sur les interactions verbales. Université Lyon 2 / Cnrs. http://clapi.univ-lyon2.fr

DUBLIN CORE (2006). Site du Dublin Core Metadata Initiative. http://dublincore.org/

FREEBANK (2007). Site de la banque de corpus constitué en vue d'échanges entre chercheurs en linguistique et traitement automatique du langage. Atilf / Cnrs. http://www.loria.fr/projets/freebank

IMS-CP (2004) : IMS Content Packaging Best Practice and Implementation Guide, version 1.1.4. IMS Global Learning Consortium, Inc. http://www.imsglobal.org/content/packaging/cpv1p1p4/imscp_bestv1p1p4.html .

IMS-LD (2003) : Instruction Management System, Learning Design Specification Version 1, final specification. IMS Global Learning Consortium, Inc. http://www.imsglobal.org/learningdesign/index.html

MULCE (2007): Site du projet Multimodal Learning Corpus Exchange (2007-2009). http://mulce.univ-fcomte.fr

OAI-MPH (2002). The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. Document Version 2002/07/05. Open Archive Initiative. www.openarchives.org/OAI/openarchivesprotocol.html

OLAC (2007). Site de l'Open Language Archives Community. University of Pennsylvania. http://www.language-archives.org/

SCORM (2000). Sharable Content Object Reference Model. V1.0 (2000), SCORM 2004 3rd edition. The advanced Distributed Learning. http://www.adlnet.gov/Scorm/index.aspx

TEI (2007) Site du standard The Text Encoding Initiative. http://www.tei-c.org/

TELOS : Technology Enhanced Learning Operating System. LORNET – Thème 6. http://www.lornet.org/Default.aspx?tabid=512

WEBCT (1998). Site de la plateforme de téléformation.. http://www.blackboard.com/webct