Document: http://mulce.univ-fcomte.fr/metadata/mce_LETECorpus-fr.pdf, modifié: 2009-03-27

Ce document introduit la notion de *Corpus d'Apprentissage* (LETEC, *Learning and TEACHING Corpus*), en vue d'étendre la typologie *olac-linguistic-type*, de la communauté OLAC (*Open Language Archives Community*, http://www.language-archives.org)

Notion de Corpus d'Apprentissage (LETEC)

Définition

Un *corpus d'apprentissage* (*LEarning & TEaching Corpus, LETEC*) 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, complété 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 (incluant pré-requis, objectifs pédagogiques, contenus et tâches), 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 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 & Teaching Corpus*) nous amène à la distinguer de celle de *corpus d'apprenants* (*Learner Corpus*) (Granger *et al.,* 2001), (Beltz, 2004). Ce dernier ne représente qu'un cas particulier et restreint de corpus d'apprentissage. En effet, il

regroupe uniquement des productions d'apprenants, pas celles d'autres acteurs. Le contexte d'apprentissage n'est pas considéré comme partie intégrante du corpus. Souvent les données n'ont pas été recueillies en situation d'apprentissage mais plutôt de contrôle des connaissances. L'objet d'étude de ces corpus est plus orienté vers celui de l'interlangue des apprenants que vers la situation d'apprentissage. Dans un corpus d'apprentissage, la notion d'acteur englobe toutes les personnes physiques participant à la session de formation (apprenants, natifs, 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. L'enseignant (auteur, concepteur) a lui aussi, à travers le scénario pédagogique, une influence déterminante sur le déroulement, mais s'il ne participe pas (comme tuteur ou coordinateur pédagogique) à la session elle-même, il ne sera pas considéré comme un acteur dans un corpus. Cependant, il sera référencé comme contributeur (concepteur, auteur) du scénario pédagogique.

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* 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.
- 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. 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.

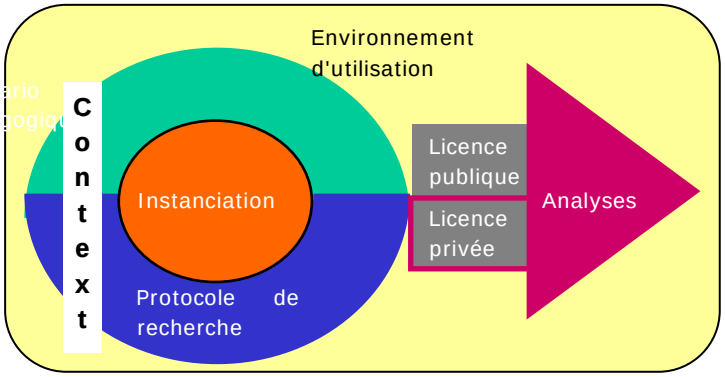

Figure 1 - Les grandes parties d'un corpus d'apprentissage de la banque Mulce

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 fouille 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.

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 ou au contraire être décrit formellement dans un langage et des concepts précis. Ce type de format, très peu lisible directement par les humains, offre 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). Dans ce niveau de descriptions structurées figurent bien sûr celles correspondant aux interactions et productions des participants à la formation, dans la partie "Instanciation".

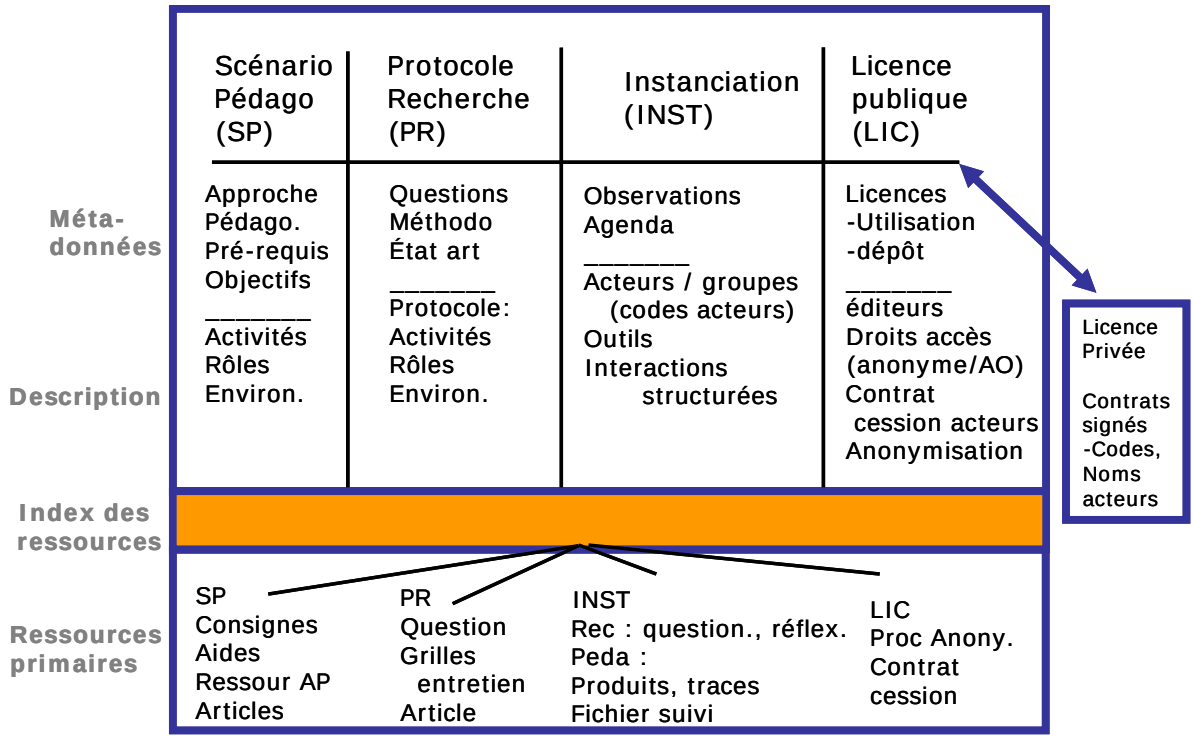

Figure 2 - Structure *Mulce-struct* d'un corpus d'apprentissage

2) *Un ensemble de ressources primaires*. Ces ressources sont "primaires" au sens où elles sont dans l'état (au processus d'anonymisation près), 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". Il en va de même pour les tests (pré et post) qui sont administrés pour mesurer les gains d'apprentissage.

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 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.

Structuration des interactions

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.

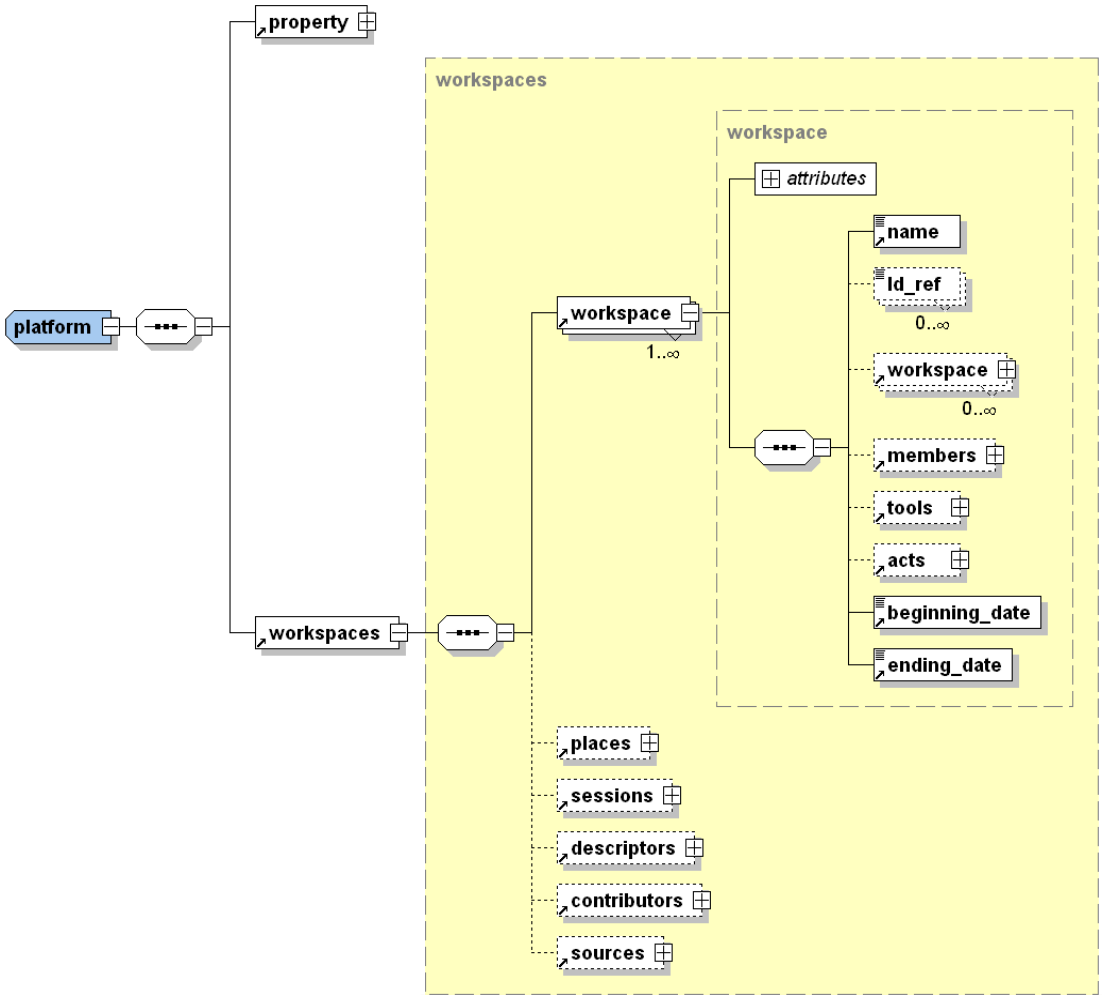

Figure 3. Extrait du schéma XSD de la partie instanciation

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.

De plus, la définition récursive de la notion d'espace de travail (*Workspace*) 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 la formation, à une étape, à une activité qu'à une session de travail (notion correspondant plus à des formations synchrones). Elle permet d'organiser les interactions recueillies au cours des différentes activités et à travers les outils à disposition. La notion de *Workspace* laisse aux concepteurs de corpus la liberté d'organiser les interactions selon leur point de vue : par activité, par tranche temporelle, ou encore par type ou espace d'interaction : forum, clavardage, ⋯

Chaque espace de travail (*Workspace*) peut être lié par des références à des objets identifiés tels que des structures d'activité du scénario pédagogique qui définissent le contexte des traces recueillies dans cet espace de travail. Bien sûr, ces liens seront plus faciles à définir si les espaces de travail correspondent précisément aux structures d'activités définies dans le scénario pédagogique. 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 possèdent un nom permettant de différencier deux outils de même type. À chacun de ces espaces/outils correspondront ensuite des actes stockés dans l'élément.

References

Reffay, C, Chanier, T., Noras, M. & Betbeder, M.-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. In Basque, J. & Reffay, C. (dir.), *numéro spécial EPAL (échanger pour apprendre en ligne)*, *Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (STICEF),* 15, [http://sticef.univ-lemans.fr/num/vol2008/01-reffay/sticef_2008_reffay_01p.pdf , http://edutice.archives-ouvertes.fr/edutice-00159733]

Mulce (2009). Version en anglais de ce document [http://mulce.univ-fcomte.fr/metadata/mce_LETECorpus-en.pdf]

Ce document peut être cité et utilisé ainsi :
Mulce (2009). Notion de corpus d'apprentissage. http://mulce.univ-fcomte.fr/metadata/mce_LETECorpus-fr.pdf

Le projet Mulce (ANR-06-CORP-006) est soutenu par ANR AGENCE NATIONALE DE LA RECHERCHE