Structuration de corpus de formation en ligne en vue de leur échange

Muriel Noras, Christophe Reffay, Marie-Laure Betbeder

Structuration de corpus de formation en ligne en vue de leur échange

Muriel Noras*, Christophe Reffay*, Marie-Laure Betbeder*

** LIFC – Université de Franche-Comté*
16 route de Gray
F-25030 Besançon cedex
{prenom.nom}@lifc.univ-fcomte.fr

RÉSUMÉ. *Dans cet article, nous proposons une définition des corpus de formation en ligne permettant leur structuration en vue de leur échange. La structuration prévoit d'inclure les éléments caractérisant le dispositif de formation ainsi que le dispositif de recherche afin de permettre à un chercheur n'ayant pas participé à l'expérimentation d'accéder au contexte des interactions pour en saisir le contenu. Ce papier décrit les composants d'un corpus échangeable : les données de recherche, le scénario pédagogique, les acteurs et outils, les traces, interactions et productions elles-mêmes et les contrats de cessation et d'utilisation de ces données. La démarche de structuration s'appuie sur deux corpus existants très différents : l'un essentiellement asynchrone et textuel, l'autre synchrone et multimodal.*

MOTS-CLÉS : *spécifications, structuration, corpus, échange.*

1. Introduction

La pratique de la formation à distance et en ligne s'intensifie, se généralise. Les scénarios pédagogiques qui l'utilisent se diversifient, touchant divers publics à travers des institutions chaque jour plus nombreuses. Cette variabilité de contextes pédagogiques et institutionnels se multiplie par l'offre exponentielle de plateformes technologiques pour créer d'innombrables situations de formation en ligne. Chaque publication vient, avec ses questions de recherche, son protocole expérimental, sa méthodologie et ses outils d'analyse, rapporter dans nos communautés un résultat d'analyse portant sur les données issues de l'une de ces situations dites écologiques, qui ne peut être décrite qu'à trop gros traits dans nos formats de publication. Les données elles-mêmes étant inaccessibles à une expertise indépendante, ces résultats ne peuvent recevoir une validation plus objective sur ces données. Le contexte expérimental étant souvent insaisissable, la réplication de l'expérience est impossible. Comme le soulignent les auteures dans [HENRI & CHARLIER 05], ce sont l'objectivité, la fidélité, la réplicabilité et la systématicité qui sont en question dans nos recherches.

1.1. Contexte et enjeux

Pour faire un pas vers la réplicabilité, le projet Mulce[1] propose de rendre accessibles nos ensembles de données. Cet accès permettrait à d'autres chercheurs d'appliquer leurs propres modèles ou analyses et de discuter de leurs résultats. Les outils de traitement ou d'analyse de traces [CHOQUET et al. 05] [MILLE & PRIÉ 06] issues de situations similaires pourraient venir tester ou étendre leur validité sur ces corpus rendus accessibles. Les outils similaires viendraient se comparer sur des données qui deviendraient une référence (Benchmark). Les traitements (automatiques ou non) appliqués à ces données engendreraient des résultats qui, rendus accessibles à leur tour, viendraient enrichir le corpus initial dans une double perspective : comparer les analyses concurrentes et capitaliser les traitements complémentaires selon une logique similaire à celle proposée dans [SALMON-ALT et al. 04] [DAOUST et al. 06] en traitement automatique du langage sur des données plus simples que celles que nous visons.

La spécification IMS-LD est connue, outillée et permet de décrire le scénario pédagogique prescrit : constituant majeur du contexte des données recueillies. Les composants de communication (forum, chat, wiki, etc.) semblent converger, rendant la structure des données d'interaction plus consensuelle. Le moment semble donc opportun pour proposer un cadre intégrant l'ensemble des données et informations nécessaires à la compréhension et au traitement (automatique ou non) des interactions issues d'une expérimentation de formation ou apprentissage à distance et

[1] Mulce : http://mulce.univ-fcomte.fr

en ligne. Cette contribution vient donc proposer une définition de la notion de corpus.

1.2. Démarche

Nous cherchons un cadre technique pérenne, extensible et si possible répandu et outillé pour contenir et organiser les différentes parties d'un corpus. Ce cadre doit permettre au chercheur (détenteur des données recueillies) de décrire tous les éléments de son corpus. Leur organisation et leur description doivent permettre aux chercheurs (n'ayant pas participé à l'expérimentation) de trouver et traiter les données qui les intéressent et leur contexte pour en interpréter correctement les analyses.

Nous disposons de données issues de deux formations en ligne dont les modalités d'interactions sont très différentes : Simuligne [CHANIER 01] est asynchrone textuelle, tandis que Copéas [BETBEDER et al. 06] est synchrone et multimodale. Les caractéristiques (scénarios, interactions) de ces corpus étant très éloignées, nous pensons qu'ils constituent de bons candidats pour valider le caractère générique de la définition et de la structuration d'un corpus proposées ici.

Ayant choisi un cadre technique répondant à nos critères, nous avons caractérisé les différents constituants d'un corpus, et pour chacun d'eux, nous proposons d'en définir une structure qui soit compatible avec nos deux candidats. Après cette introduction, nous proposons dans la deuxième partie de cet article, une définition et la structure générale d'un corpus.

2. Définition et composition d'un corpus

Nous définissons un corpus comme un ensemble de données et de traces issues d'une expérimentation, enrichies par des informations techniques, humaines, pédagogiques et scientifiques permettant leur analyse en contexte.

Si le cœur du corpus concerne les données d'interaction recueillies, son partage (et donc sa compréhension) implique de décrire le contexte de l'expérimentation en incorporant les éléments définitoires du dispositif de formation [KERN et al. 04] ainsi que du dispositif de recherche. Certaines données sont extraites du dispositif tel que conçu avant la formation (scénario pédagogique, partie de matériaux pédagogiques, consignes, etc.). D'autres résultent de la mise en œuvre du dispositif et rendent compte des écarts avec la situation initialement planifiée (absences, événements imprévus, apports de données extérieures, etc.). Les données provenant du dispositif de recherche comprennent les informations sur les acteurs, questionnaires, entretiens, prises de notes, enregistrements vidéo, etc. La description du contexte doit permettre à des chercheurs n'ayant pas participé à la formation, de comprendre, à partir d'une interaction, comment (quel outil, quel acteur, quelle activité) et pour quoi (dans quel but) elle a été produite pour en permettre l'analyse. Pour cela, il est également nécessaire de décrire l'institution et le cadre pédagogique

dans lequel s'est déroulée la formation ainsi que les acteurs et de leur rôle dans la formation.

Pour mettre en évidence les différents registres de données nécessaires à la compréhension d'un corpus, nous proposons une organisation en six parties (cf. figure 1).

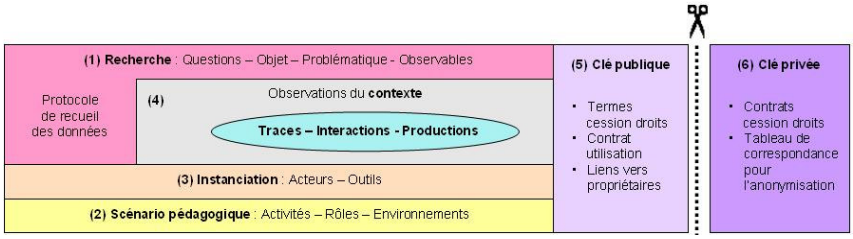

Figure 1. *Les six parties d'un corpus*

La partie (1) de la figure 1 contient l'**orientation de recherche**. Elle représente : la problématique de la recherche, son objet (ce qui la motive), ses hypothèses ; les observables et le protocole de recueil des données et également les enregistrements des entretiens ou des questionnaires. Les informations de cette partie conditionnent l'ensemble de l'expérimentation.

La partie (2) correspond à la phase de description du **scénario pédagogique** prescrit. Au cours de cette phase, les rôles et les activités pédagogiques sont décrits en lien avec les objectifs pédagogiques, les outils génériques (ex. forum, tests) ou spécifiques (ex. simulation) et les contenus pédagogiques nécessaires. Il s'agit d'une description du scénario à un niveau abstrait excluant les références à des acteurs humains et environnements précis. Pour décrire cette partie, nous utilisons IMS-LD.

La partie (3) intitulée **instanciation** contient les données concrétisées de la phase précédente. Elle permet, d'une part, d'inscrire les acteurs (nom, prénom, login, adresse, etc.) qui joueront les rôles décrits dans la partie (2) ; et d'autre part, d'instancier les environnements abstraits (ex. forum) de la partie (2) par des outils concrets (ex. phpBB). Ce n'est qu'à partir de ce moment que nous pouvons connaître précisément le type d'interactions à recueillir et leurs caractéristiques. Pour cette partie, nous utilisons IMS Enterprise pour l'instanciation des acteurs et nous avons défini une spécification détaillée pour décrire les traces issues de ces outils.

La partie (4), centrale dans le corpus, concerne les données recueillies lors de la **formation**. Au cours de la formation, les acteurs se retrouvent dans le dispositif technologique pour réaliser les activités. Cette partie du corpus permet de stocker deux types d'informations, recueillies selon le protocole décrit dans la première partie. D'une part, sont stockés l'ensemble des **logs système** [JERMANN 01], des **interactions humaines médiées** (acte de parole, iconique, graphique, vidéo, textuel – message de forum, courriel) et des **productions** des acteurs (productions de

l'activité d'apprentissage au sens de la théorie de l'Activité [KUUTTI 96]). D'autre part, cette partie contient les éléments du **contexte,** capturés en vue d'une meilleure interprétation des données et des analyses. On y trouvera aussi un relevé des disfonctionnements ou autres événements ayant affecté le déroulement de la formation. Les données d'interaction seront décrites en utilisant la structure spécifiée pour les outils dans la partie 3.

Les parties (5) et (6) sont des clés et concernent les devoirs des utilisateurs et les droits des acteurs du corpus. Dans l'optique d'une diffusion du corpus, la partie (5) est incluse, elle contient les termes de cession des droits, les contrats d'utilisation du corpus ainsi qu'un lien vers les propriétaires du corpus. La partie (6) est détenue uniquement par le groupe de recherche propriétaire du corpus et porte sur les données privées : contrats de cessions des droits signés par les acteurs, tableau de correspondance pour l'anonymisation des interactions.

Enfin, les données des six parties sont inter-reliées. Le but est de tisser les liens entre les ressources, les traces, les productions et les interactions issues des phases précédentes et contenues dans les parties correspondantes du corpus pour constituer un tout cohérent. Chaque élément est alors mis en relation avec les données de son contexte.

3. Conclusion

L'objectif de cet article est de présenter une proposition permettant concrètement la mise à disposition de corpus dans un format commun en vue d'améliorer la réplicabilité des modèles et analyses des interactions dans nos communautés de recherche.

Cette contribution définit la notion de corpus d'interactions en situation d'apprentissage en ligne. Elle spécifie les cinq parties qu'il faut ajouter au cœur du corpus (traces, interactions, productions, contexte), pour rendre le corpus échangeable : les orientations de recherche, le scénario pédagogique, l'instanciation, les observations du contexte, les clés publique et privée.

Sur le plan technique, pour bénéficier des avantages de la spécification IMS-CP (extensibilité, dissémination, outillage) nous avons choisi de fonder notre structuration dans un conteneur IMS-CP : soit en réutilisant des spécifications existantes (IMS-LD, IMS Enterprise), soit en en proposant de nouvelles.

Remerciements

Mulce (Échange de corpus d'apprentissage multimodaux) 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 Sciences Humaines et Sociales". Il rassemble des équipes des laboratoires LASELDI et LIFC (Université de Franche-

Comté), CREET (The Open University) et LIP6 (Université Paris 6), coordonnées respectivement par T. Chanier, C. Reffay, M.-N. Lamy et J.-G. Ganascia.

4. Bibliographie

[BETBEDER et al. 06] Betbeder, M.-L., Reffay, C., Chanier, T., « Environnement audio graphique synchrone : recueil et transcription pour l'analyse des interactions multimodales », *Actes des premières journées communication et apprentissage instrumenté en réseau : JOCAIR 2006*, Amiens, 6-7 juillet 2006, p. 406-420.

[CHANIER 01] Chanier, T., « Créer des communautés d'apprentissage à distance. » *Les dossiers de l´Ingénierie Educative*, vol. 36, Centre National de Documentation Pédagogique, Montrouge, 2001, p. 56-59.

[CHOQUET et al. 05] Choquet, C., Luengo, V., Yacef, K., Usage Analysis in Learning Systems, *Actes de workshop de la 12th Conference on Artificial Intelligence in Education*, Amsterdam, Pays-Bas, 18-22 juillet 2005.

[DAOUST et al. 06] Daoust, F., Dobrowolski, G., Dufresne, M., Gélinas-Chebat, C., « Analyse exploratoire d'entrevues de groupe : quand ALCESTE, DTM, LEXICO et SATO se donnent la main », *Actes des 8èmes journées internationales d'Analyse statistique des Données Textuelles : JADT 2006*, Besançon, 19-21 avril 2006.

[HENRI & CHARLIER 05] Henri, F., Charlier, B., « L'analyse des forums de discussion pour sortir de l'impasse », *Symposium, formation et nouveaux instruments de communication*, Amiens, janvier 2005.

[JERMANN 01] Jermann, P., Soller, A., Muehlenbrock, M., « From Mirroring to Guiding : A Review State of the Art Technology for supporting Collaborative Learning. », proceedings of *the First European Conference on Computer-Supported Collaborative Learning*, 2001.

[KUUTTI 96] Kuutti, K., « Activity Theory as a Potential Framework for Human Computer Interaction Research », dans B. A. Nardi (Ed.), Context and Consciousness, pp.17-44.

[KERN et al. 04] Kern, R., Ware, P., Warshauer, M., « Crossing frontiers: new directions in online pedagogy and research », *Annual Review of Applied Linguistics*, vol. 24, p. 243-260.

[MILLE & PRIÉ 06] Mille, A., Prié, Y., « Une théorie de la trace informatique pour faciliter l'adaptation dans la confrontation logique d'utilisation/logique de conception », *Actes des 13èmes journées de Rochebrune*, Rochebrune, janvier 2006.

[REFFAY & TEUTSCH 07] Reffay, C., Teutsch, P., « Anonymisation de corpus réutilisables », *Annexes aux actes de la conférence EIAH 2007*, 2 p., Lausanne, Suisse, 27-29 juin 2007.

[SALMON-ALT et al. 04] Salmon-Alt, S., Romary, L., Pierrel, J.-M., « Un modèle générique d'organisation des corpus en ligne ». *Traitement Automatique des Langues (TAL)*, vol. 45, n°3, p. 145-169.