Enjeux, outils et méthodologie de constitution de corpus d'apprentissage

Ciara R. Wigham, Aurélie Bayle

Enjeux, outils et méthodologie de constitution de corpus d'apprentissage

Ciara R. Wigham & Aurélie Bayle, Clermont Université

Mots-clés : corpus d'apprentissage, LETEC, didactique des langues-cultures, interactions multimodales en ligne, mondes synthétiques, outils

Dans le domaine de l'enseignement-apprentissage des langues, des corpus d'apprenants (*learner corpora*) sont exploités pour la recherche qui porte sur l'acquisition d'une langue étrangère (L2), souvent en recueillant des données de contrôle des connaissances (Reffay *et al.*, 2008) et en faisant des études comparatives avec des productions d'interlocuteurs natifs (Belz & Vyatkina, 2009). Ce type de corpus focalise uniquement sur les productions des apprenants et ne prend pas en compte celles des autres acteurs de la formation ni le contexte d'apprentissage.

Nos thèses portent sur des interactions multimodales collaboratives issues de situations d'apprentissage de L2 dans des mondes synthétiques (virtuels), ceci dans le cadre des projets *Archi21* et *Slic*. Dans le projet *Archi21*, une formation a été conçue dans une approche Emile (Enseignement d'une Matière Intégré à une Langue Etrangère) pour mêler l'apprentissage en architecture et en langues étrangères (français et anglais). Le projet *SLIC* a mis en relation des apprenants du L2 avec des futurs enseignants pour réaliser des tâches collaboratives sur des thématiques culturelles. Nos problématiques de recherche s'intéressent directement aux rapports entre les affordances des dispositifs pédagogiques dans ce nouveau type d'environnement d'apprentissage et les interactions entre participants (apprenants, tuteurs, natifs). Il nous parait donc nécessaire de constituer un objet de recherche complet qui rassemble tous les éléments issus du dispositif de formation dans lequel les interactions entre tous les participants, et pas seulement les productions des apprenants, sont mises en avant. Les données provenant du monde synthétique *Second Life* sont multimodales et donc très diverses (audio, vidéo, clavardage, production d'objets, écriture collaborative, communication non verbale des avatars). Ceci rend les études difficilement comparables et la ré-analyse d'une situation d'apprentissage par un chercheur extérieur difficile s'il ne connait ni le contexte d'apprentissage ni le protocole de recueil de données. D'où l'intérêt de constituer un corpus d'apprentissage (*LEarning and TEaching Corpora* - LETEC).

Un corpus d'apprentissage relie, en suivant des standards internationaux, tous les éléments provenant d'une situation de formation en ligne (Chanier & Ciekanski, 2010). Il est principalement constitué d'un fichier XML appelé "manifeste" qui décrit les composants du corpus : le protocole de recherche (questionnaires, entretiens), le scénario pédagogique, toutes les interactions, productions et traces extraites de la situation de formation ainsi que les licences (des participants et des utilisateurs du corpus). A cela s'ajoutent un index de ressources et l'ensemble des ressources de la formation et de l'expérimentation (fichiers vidéo, audio, texte…).

Notre communication portera sur la méthodologie, les étapes de constitution d'un corpus d'apprentissage ainsi que les outils utilisés (*Fraps, MotPlus, ELAN, Oxygen*) en s'appuyant sur des exemples concrets issus des recueils de données, leur structuration (Chanier & Wigham, 2011) et des exemples d'analyses faites dans nos travaux de thèse. Nous montrerons que les analyses sont possibles, voire facilitées, grâce à la vue d'ensemble donnée par un corpus structuré. Ce rassemblement de données nous permet, par exemple, dans le projet *Slic* de lier les productions issues de la plateforme *Moodle* avec les interactions dans *Second Life* pour étudier la réalisation complète des tâches collaboratives et faire des comparaisons intergroupes. Dans le projet *Archi21* la vue d'ensemble nous permet d'employer des outils d'analyse quantitatives sur toutes les sessions de la formation pour analyser l'utilisation des modes (verbal, non verbal) et modalités diverses (audio, clavardage, kinésique, production…) et ensuite de comparer, dans les groupes (français/anglais), les différentes approches employées par les tuteurs des deux groupes pour rétroagir dans le clavardage.

Les analyses effectuées sur les données/ressources peuvent être intégrées dans un corpus global ou, dans notre cas, font l'objet de sous-corpus que l'on appelle corpus distinguables. De ce fait, un corpus d'apprentissage se constitue tout au long du travail de thèse et dans sa continuité. C'est un travail qui débute dès la mise en place du dispositif d'apprentissage et du protocole de recherche et qui se poursuit au-delà du processus d'analyse. La constitution et la diffusion d'un corpus d'apprentissage permettent de mettre en parallèle les données ayant servi aux analyses avec les résultats de ces analyses présentés dans les publications scientifiques. Cela valorise également le travail de thèse en le rendant visible par le référencement à l'extérieur *(OLAC, CLARIN)* et facilite l'approfondissement des recherches après la thèse.

Références :

Belz, J.A. & Vyatkina, N. (2009). The pedagogical mediation of a developmental learner corpus for classroom-based language instruction. *Language Learning & Technology (LLT)*, 12(3), 33-52. [http://llt.msu.edu/vol12num3/belzvyatkina.pdf]

Chanier, T. & Ciekanski, M. (2010). Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage : un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage. *Apprentissage des Langues et Systèmes d'Information et de Communication (ALSIC)*, 13. [oai : edutice.archives-ouvertes.fr:edutice-00486676]

Chanier, T. & Wigham, C.R. (2011). (Dir.) *Learning and Teaching Corpus ARCHI21.* Mulce.org : Clermont Université. [oai : mulce.org:mce-archi21-letec-all ; http://repository.mulce.org]

Reffay, C., Chanier, T., Noras, M. & Betbeder, M.-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. *Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (Sticef)*, 15. [oai : edutice.archives-ouvertes.fr:edutice-00159733]