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

Ciara R. Wigham, Aurélie Bayle

Wigham, C.R. & Bayle, A. (à paraître). Enjeux, outils et méthodologie de constitution de corpus d'apprentissage, in Damiani M., Dolar K., Florez-Pulido C., Magnier J. & Loth R. (dir.) *Actes de Coldoc 2012*. Paris : Modyco

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

Ciara R. Wigham[1] *Aurélie Bayle*[2]
(1) Clermont Université, Laboratoire de Recherche sur le Langage
(2) Clermont Université, Laboratoire de Recherche sur le Langage
ciara.wigham@univ-bpclermont.fr, aurelie.bayle@univ-bpclermont.fr

RESUME

Notre recherche porte sur des interactions multimodales collaboratives issues de situations d'apprentissage de langue étrangère (L2) dans des mondes synthétiques (virtuels). Cet article décrit l'intérêt de construire des corpus d'apprentissage dans le cadre de nos travaux de thèses ainsi que la méthodologie employée. Un corpus d'apprentissage assemble un ensemble de données structurées issues d'une expérimentation de formation en ligne dont le contexte est décrit par un scénario pédagogique et un protocole de recherche. Nous présentons les étapes de constitution d'un tel corpus ainsi que les outils utilisés en s'appuyant sur des exemples concrets issus des recueils de nos données, de leur structuration et des analyses faites. Nous montrons que les analyses sont facilitées grâce à la vue d'ensemble donnée par un corpus structuré et qu'un tel corpus aide à la visibilité des travaux scientifiques de thèse.

Abstract

Building a LEarning and TEaching Corpus: Issues, Tools and Methodology

Our research focuses on multimodal, collaborative interactions during foreign language learning situations in synthetic (virtual) worlds. This article describes the benefits of building LEarning and TEaching Corpora (LETEC) in the framework of our doctoral studies and the methodology employed to do so. A LEarning and TEaching Corpus is defined as a structured entity containing all the elements resulting from an online learning situation, whose context is described by a pedagogical scenario and a research protocol. Through the use of concrete examples from our data collection, structuration and analyses, we present the stages and tools involved when building this type of corpus. We show that our analyses are facilitated due to the global view the structured corpus offers and how a LEarning and TEaching Corpus helps to increase the visibility of the scientific work undertaken during our doctoral studies.

MOTS-CLES : corpus d'apprentissage, didactique des langues-cultures, interactions multimodales en ligne, mondes synthétiques, outils
KEYWORDS: LEarning and TEaching corpus (LETEC), language didactics, multimodal online interactions, synthetic worlds, tools

1 Introduction

Dans le domaine des sciences humaines et plus particulièrement des sciences du langage, un problème fréquemment soulevé est celui de l'importance de rendre les données de recherche visibles et les publications accessibles, dans le but d'accroitre la validité scientifique et la facilitation d'exploitation des publications (Reffay, Betbeder et Chanier, 2012). Dans le cadre de cette problématique, pour notre recherche portant sur des interactions multimodales collaboratives issues de situations d'apprentissage de langue étrangère (L2) dans des mondes synthétiques (virtuels), nous adoptons l'approche méthodologique des corpus d'apprentissage (LEarning and TEaching Corpora, LETEC). Dans cet article, nous présentons la notion de corpus d'apprentissage, le contexte de nos thèses et les raisons pour lesquelles nous optons pour l'approche LETEC avant d'entrer dans les détails des composants d'un corpus d'apprentissage et la structuration de ce dernier. Nous expliquons, à travers des exemples concrets issus de nos travaux de recherche, les processus de recueil, de structuration et d'analyse des données portant sur le rôle du clavardage dans les interactions multimodales et l'influence du comportement de l'animateur sur les interactions des étudiants. Nous montrons qu'une méthodologie adoptant l'approche LETEC débute dès la mise en place du dispositif d'apprentissage et du protocole de recherche et se poursuit au-delà du processus d'analyse. Nous verrons également de quelle manière un corpus d'apprentissage facilite les analyses

Wigham, C.R. & Bayle, A. (à paraître). Enjeux, outils et méthodologie de constitution de corpus d'apprentissage, in Damiani M., Dolar K., Florez-Pulido C., Magnier J. & Loth R. (dir.) *Actes de Coldoc 2012*. Paris : Modyco

et les rend visibles grâce à la mise en parallèle des données ayant servi aux analyses avec les résultats de ces analyses présentés dans les publications scientifiques.

2 Corpus d'apprentissage (LEarning and TEaching Corpora, LETEC) : vue d'ensemble

Un corpus d'apprentissage est un ensemble de données structurées composé de tous les éléments issus d'une situation de formation en ligne recueillis en fonction d'un protocole de recherche :

> *Un corpus d'apprentissage en ligne assemble de façon systématique et structurée un ensemble de données, particulièrement d'interactions, et de traces issues d'une expérimentation de formation partiellement ou totalement en ligne, enrichies par des informations techniques, humaines, pédagogiques et scientifiques permettant leur analyse en contexte.* (Mulce-documentation, 2011)

Selon l'approche LETEC, un corpus n'est pas simplement une collection de documents numériques mais doit couvrir un paradigme composé de quatre facettes : le recueil systématique des documents, la description du contexte, l'organisation et l'instrumentalisation en vue de traitements et les dispositions en vue de l'échange et du partage (table 1).

Recueil systématique des documents liés à l'objet d'étude	Description du contexte	Organisation et instrumentalisation en vue de traitements	Dispositions en vue de l'échange et du partage
-Productions des participants -Documents concernant le contexte d'élaboration -Documents décrivant le déroulement de la formation	-Aspects technologiques -Aspects pédagogiques -Documents en rapport avec le protocole de recherche -Métadonnées décrivant les caractéristiques de l'œuvre, ses acteurs suivant les standards	-Documents recueillis numérisés dans des formats ouverts -Documents adaptés aux différents outils -Documents organisés dans des langages de balisage ouverts aux traitements (XML) et structurés suivant des schémas standards (TEI) ou accessibles à tous	-Dépôt du corpus en accès libre dans une banque de corpus qui a un serveur indexable -Règles du jeu de l'utilisation du corpus sous forme de licence

TABLE 1- Quatre facettes d'un corpus (d'après Chanier et Ciekanski, 2010)

Par rapport à un corpus d'apprenants (*learner corpora,* Granger, Hung et Petch-Tyson, 2002), un corpus d'apprentissage ne s'intéresse pas seulement aux apprenants mais considère tous les participants, y compris les tuteurs. Plutôt que de focaliser sur des productions, souvent issues de contrôles de connaissances (Reffay *et al.*, 2008), et de comparer ces productions avec celles d'interlocuteurs natifs (Belz et Vyatkina, 2009), un corpus d'apprentissage comprend en plus les interactions entre participants durant la formation et le contexte dans lequel ces interactions ont lieu.

La notion de corpus d'apprentissage a été développée dans le cadre de l'*ANR Mulce* (Mulce-documentation, 2011) dans le but de structurer et de contextualiser des données issues de formations en ligne. Elle répond aux critères énoncés dans la table 1 en rendant explicites les liens entre les données d'instanciation (en incluant les traces d'interactions), le contexte technique (décrit dans le protocole de recherche), le contexte d'apprentissage (décrit dans le scénario pédagogique) et éventuellement les analyses (figure 1).

Etant donné que la réplication du contexte écologique d'une situation d'apprentissage collaboratif en ligne est pratiquement impossible en raison du nombre de variables incontrôlables (Reffay, Betbeder et Chanier, 2012), des analyses cumulatives et contrastives sont difficiles à réaliser. Même si le scénario pédagogique est réappliqué avec un groupe d'apprenants différents, les phénomènes observables ne seront pas nécessairement identiques. Un deuxième objectif est donc de rendre disponibles les données structurées à la communauté scientifique[1] pour permettre aux chercheurs extérieurs à l'expérimentation de conduire de nouvelles analyses facilitées par le fait que les données d'un corpus d'apprentissage sont structurées dans un formalisme indépendant (IMS-CP, 2007) et pérenne en éliminant les formats propriétaires des logiciels ou plateformes d'apprentissage.

[1] Les corpus sont disponibles dans la banque de données Mulce (http://mulce.univ-bpclermont.fr:8080/PlateFormeMulce/)

Wigham, C.R. & Bayle, A. (à paraître). Enjeux, outils et méthodologie de constitution de corpus d'apprentissage, in Damiani M., Dolar K., Florez-Pulido C., Magnier J. & Loth R. (dir.) *Actes de Coldoc 2012*. Paris : Modyco

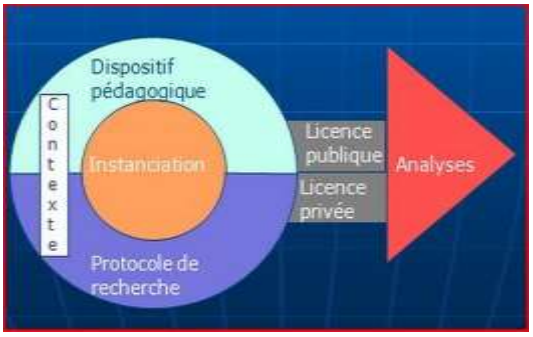

FIGURE 1 - Schéma des composants d'un corpus d'apprentissage (Mulce- documentation, 2011)

3 Contexte de notre recherche

Dans cette partie, nous introduisons les mondes synthétiques, l'environnement qui est l'objet de nos études avant de présenter nos problématiques de recherche et l'intérêt d'utiliser un corpus d'apprentissage dans nos méthodologies. Nous présentons ensuite nos terrains de recherche.

3.1 Environnement d'apprentissage étudié

Un monde synthétique, tel que *Active Worlds* (Active Worlds, Inc., 1997), *Moove Online* (Moove, 1999) ou *Second Life* (Rosedale, 2011) est un environnement synchrone auquel on accède à partir d'une interface graphique en trois dimensions, qui est persistant et interactif. Les utilisateurs se connectent à travers un réseau et interagissent par le biais d'un avatar (Book, 2004 ; Bell, 2008). Ces environnements sont nommés de différentes manières dans la littérature. Le terme "monde virtuel" est peut-être le plus courant. Nous préférons opter pour le terme "monde synthétique" car dans ces environnements, l'interaction et l'apprentissage se produisissent réellement. Nous faisons donc le choix d'éviter le terme "virtuel" qui est souvent opposé au "réel".

3.2 Intérêt des mondes synthétiques pour la formation en langues étrangères

Antonacci *et al.* (2008) suggèrent plusieurs potentialités des mondes synthétiques pour l'apprentissage. Ici nous les mettons en lien avec l'apprentissage d'une langue étrangère. Premièrement, ils suggèrent que ces mondes offrent des possibilités pour l'accomplissement de tâches considérées comme difficiles dans le premier monde[2]. Cela peut aider des apprenants à contextualiser la L2 et, à partir de la contextualisation, à transférer les connaissances acquises aux situations de communication authentiques dans le premier monde. Deuxièmement, pour Antonacci *et al.*, l'accès persistant favorise les interactions sociales. Concernant l'apprentissage des langues, il permet des interactions avec des interlocuteurs natifs et donc les possibilités pour des tâches authentiques. La persistance de l'environnement permet également à un apprenant de retourner à l'environnement d'apprentissage pour interagir à nouveau avec des objets d'apprentissage. Finalement, Antonacci *et al.* soulignent l'intérêt de ces environnements pour la collaboration. En effet, dans ces mondes synthétiques, les apprenants peuvent être actifs mais également acteurs, participant à la co-construction du monde. Peterson (2011), qui s'intéresse à l'apprentissage de l'anglais L2, souligne que la possibilité de créer des objets qui ont une signifiance personnelle pourrait stimuler l'engagement et l'investissement dans l'environnement d'apprentissage et ainsi la motivation de l'apprenant. Dalgarno et Lee (2009) expliquent également que les multiples modes de communication favorisent la collaboration car le mode spatial et le non verbal augmentent le sentiment d'appartenance des apprenants et donc, les relations et la communication efficace entre eux.

Le fait que la communication en L2 passe à travers un avatar a également été cité comme affordance de ce type d'environnement pour l'apprentissage de L2. Sanchez (1996) et Schweinhorst (2002)

[2] Nous utilisons la notion de "premier monde" en opposition à "monde synthétique", souvent appelé "deuxième monde"

suggèrent que l'avatar peut réduire l'appréhension d'un apprenant à s'exprimer en L2. De ce fait, il pourrait se sentir plus à l'aise pour prendre des risques dans la L2 (Teoh, 2007) : un comportement qui facilite l'apprentissage des langues (Peterson, 2011).

3.3 Problématiques communes à notre recherche et intérêt d'un corpus d'apprentissage dans nos travaux

Nos thèses portent sur des interactions multimodales collaboratives issues de situations d'apprentissage de L2 dans des mondes synthétiques (virtuels), 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 de 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, le type de tâche mis en place 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 et dans lequel les interactions entre tous les participants, et pas seulement les productions des apprenants, sont mises en avant en prenant en compte leur contexte de réalisation. Le deuxième intérêt d'utiliser un corpus dans nos travaux de recherche vient du fait que la structuration des données dans un format XML facilite les analyses. La structuration permet à un chercheur de gagner du temps s'il veut employer des outils de traitement et d'analyse différents à partir du même ensemble de données. Les analyses sont également facilitées grâce à la vue d'ensemble donnée par un corpus structuré qui permet de faire des comparaisons entre séances, entre groupes et entre les différents outils utilisés par les apprenants. Finalement, nous voyons l'intérêt d'une approche par corpus du fait que nous travaillons dans des projets avec des partenaires extérieurs qui pourront être amenés eux aussi à travailler sur les données pour leur propres objectifs de recherche ou pour compléter nos analyses, par exemple dans le cadre du projet *Archi21* avec des analyses dans le domaine de l'architecture.

3.3.1 Le projet *Archi21*

Le projet européen *Archi21* a pour but d'étudier l'approche Emile pour pallier le manque de cours spécialisés dans le domaine de l'architecture en L2 généralement observé dans l'enseignement supérieur. Le projet cherche à aider les étudiants à maitriser une L2 pour valider leurs diplômes au niveau Master (Joint Quality Initiative, 2004) afin de faciliter leur mobilité. Dans le cadre de ce projet, une formation hybride appelée "Building Fragile Spaces" a été menée en février 2011. Comportant des séances en présentiel et à distance, la formation s'est déroulée sur 5 jours, de façon intensive, avec 17 étudiants qui avaient l'anglais ou le français comme L2. Ces étudiants étaient accompagnés par deux enseignants d'architecture en présentiel, un anglophone et un francophone, et deux tuteurs de langue à distance dont un enseignait l'anglais et l'autre le français.

Lors de la formation, les étudiants travaillaient en petits groupes organisés par L2 et devaient créer un modèle dans le monde synthétique *Second Life* qui répondait à une problématique désignée par les enseignants d'architecture. Pour les accompagner dans cette tâche, des séances réflexives, entre autres (voir Rodrigues *et al.,* à paraître), étaient menées par les tuteurs de L2. Les objectifs de ces séances étaient à la fois architecturaux et linguistiques. En L2, les étudiants devaient s'exprimer sur leur contribution personnelle au travail de groupe, sur les retours faits par les enseignants d'architecture sur leur modèle et sur la façon dont ils allaient les prendre en compte dans l'avancement de leur travail. Le point culminant de la formation était la présentation des modèles en L2 par chaque groupe le dernier jour de la formation, sur laquelle les étudiants étaient notés.

3.3.2 Le projet *Slic*

Le projet *Slic* (Second Life InterCulturel) est le fruit d'une collaboration entre l'Université Blaise Pascal (UBP) à Clermont-Ferrand et Carnegie Mellon University (CMU) à Pittsburg, Etats-Unis (Bayle,

Foucher et Youngs, 2012). Ce projet a mis en relation 21 apprenants de français de CMU avec 14 étudiants de Master en Didactique des Langues et Cultures, spécialité Français Langue Etrangère et Seconde (DLC-FLES) à l'UBP. L'expérimentation s'est déroulée entre septembre et décembre 2011. Sept groupes de 4 à 6 étudiants (dont 2 de l'UBP) ont mené des tâches collaboratives dans le monde synthétique *Second Life*. Le projet s'est structuré en 5 étapes précédées d'une introduction au monde synthétique. De plus, les participants avaient accès à la plateforme *Moodle*, utilisée comme espace de ressources, de consignes et d'échanges asynchrones entre les séances synchrones. L'objectif de la formation était, pour tous les participants, le développement de compétences interculturelles. Pour les étudiants de CMU, il s'agissait également de développer leurs compétences orales en français et d'approfondir le cours qu'ils suivaient à CMU. *Slic* a permis aux étudiants de l'UBP de découvrir l'utilisation pédagogique d'un environnement informatique, la FOAD, et d'avoir une première expérience de prise de responsabilités dans une situation pédagogique à distance et d'utilisation des moyens de communication. En effet, ils avaient la tâche supplémentaire, à tour de rôle, d'animer les séances synchrones.

Les tâches collaboratives ont été conçues à partir du programme de l'enseignante de français à CMU qui enseignait le module "Introduction to French culture". Les grandes thématiques de ce programme (langue, identité, symboles, actualités) ont constitué le canevas du projet *Slic* puisque chaque étape correspondait à une thématique. A la fin de chaque étape, les étudiants devaient collaborer pour produire divers documents (compte-rendu de séance, diaporamas, cartes conceptuelles, tableaux).

4 Démarche méthodologique

L'adoption d'une approche par corpus d'apprentissage est constituée de quatre phases chronologiques : avant l'expérimentation, pendant l'expérimentation, post-expérimentation et post-recherche. Nous décrivons ici chaque phase de cette approche.

4.1 Avant l'expérimentation

L'élaboration d'un dispositif de formation implique de réfléchir, en amont, aux objectifs, étapes, méthodes, modalités de travail, environnements, rôles des participants, etc. L'établissement d'une progression pédagogique prenant en compte tous ces éléments constitue le scénario pédagogique qui décrit et guide le déroulement de la formation.

Si ce dispositif doit faire l'objet d'un travail de recherche, il est également nécessaire d'élaborer un protocole de recherche qui se compose des questions de recherche sur lesquelles vont se baser les analyses ainsi que du protocole de recueil de données. C'est en effet au moment où l'on conçoit le dispositif de formation que l'on définit quelles données seront recueillies et de quelle manière, quel sera le rôle des chercheurs et comment vont se dérouler les activités propres à la recherche.

4.2 Recueil de données pendant l'expérimentation

Les protocoles de recherche élaborés pour suivre les formations dans le cadre des projets *Archi21* et *Slic* impliquaient de recueillir des données à l'intérieur du monde synthétique ainsi qu'à l'extérieur de l'environnement d'étude.

Les données d'interaction provenant des mondes synthétiques sont multimodales et donc très diverses (Wigham et Chanier, à paraître-a). Le mode verbal produit des données dans la modalité audio ainsi que dans la modalité clavardage. Dans le mode non verbal, il faut prendre en compte la modalité proxémique (l'orientation de l'avatar, mouvement de l'avatar vers un autre) et la modalité kinésique (le regard, les gestes et les expressions faciales des avatars) ainsi que les productions effectuées qui peuvent être la production et l'utilisation des artefacts dans l'environnement (apparition ou construction) ou la production d'un texte, par exemple dans un éditeur de texte collaboratif, sur une note ou sur un tableau blanc interactif (figure 2).

Vu la diversité des données provenant du monde synthétique, leur gestion n'est pas facile. Leur mode de recueil peut être semblable à celui d'autres environnements ou peut nécessiter des compétences

spécifiques pas forcement maitrisées par les chercheurs, telles que l'enregistrement vidéo ou la programmation d'objets dans le monde synthétique (Yee et Bailenson, 2008).

Le recueil nécessite également la présence d'un avatar dans le monde synthétique pour capter les interactions. Bayle et Foucher (2011) soulignent les avantages et les contraintes générées par le choix d'avoir soit un avatar-chercheur présent dans le monde synthétique, soit de capter les données à partir des écrans des participants dans la formation. En ce qui nous concerne, dans les formations Building Fragile Spaces et *Slic*, nous avons opté pour le choix d'un avatar-chercheur. Pour diminuer l'impact potentiel de "l'observation participante" (Blanchet, 2011) sur les dynamiques entre participants (cf. "le paradoxe de l'observateur", Labov, 1972 :209), nous avons choisi un avatar dont la forme était celle d'un petit animal (figure 3) pensant que les étudiants adresseraient moins la parole à un avatar non-morphologique (cf. Wigham et Chanier, à paraître-b) et de ce fait que le chercheur serait aussi discret que possible.

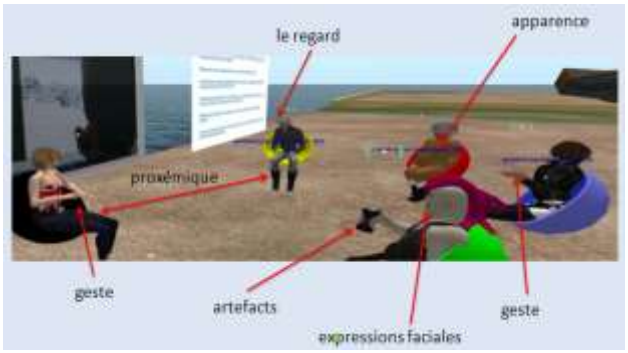

FIGURE 2 – Modalités de communication

FIGURE 3 - L'avatar-chercheur

Avant les formations, dans le protocole de recherche, nous avons décrit la façon dont le chercheur allait positionner son avatar, en le plaçant à un endroit où il est possible de capter un maximum d'éléments. C'est primordial car le clavardage public n'apparait que dans la fenêtre de clavardage des avatars qui sont à 20 mètres autour de la personne qui communique et l'audio public ne peut être entendu que par les avatars à 60 mètres de l'interlocuteur. L'avatar-chercheur a procédé à des enregistrements des séances en utilisant des logiciels d'enregistrement vidéo d'écran : *Fraps* (Beepa, 2012) et *Camtasia* (TechSmith Corporation, 2010). Le compte de l'avatar-chercheur dans le monde synthétique a également été paramétré pour enregistrer le clavardage en texte brut.

Nos dispositifs de formation ont également donné lieu à des données ne provenant pas des mondes synthétiques. Pour les besoins de la recherche, nous avons obtenu l'accord des participants pour utiliser les données qu'ils ont produites. Nous leur avons également soumis des questionnaires pré- et post- formation en ligne (*KwikSurveys*, n.d.) pour faciliter le recueil et le traitement des réponses. Certains étudiants ont participé à des entretiens après la formation, conduits dans un environnement audio-graphique synchrone et enregistrés en utilisant le logiciel *SkypeRecorder* (Nikiforov, 2011). S'ajoutent à ces données les descriptions des acteurs, des environnements utilisés ainsi que le scénario pédagogique. Pour le projet *Slic*, *Moodle* a été utilisé en complément de *Second Life*. Les données provenant de cette plateforme d'apprentissage (messages dans les forums, consignes, productions et ressources) ont été extraites au format XML.

4.3 Post-expérimentation : Constitution du corpus global

Dans cette section, nous décrivons les composants d'un corpus d'apprentissage (section 4.3.1) avant d'expliquer comment ils sont structurés dans le corpus (section 4.3.2). Pour mieux comprendre cette section, le lecteur pourra télécharger un corpus global, par exemple celui d'*Archi21* (Chanier et Wigham, 2011).[3] Il verra que les composants du corpus correspondent à des répertoires et pourra explorer le "manifeste" en parallèle de sa lecture de la section 4.3.2 en ouvrant le fichier

[3] La création d'un compte sur Mulce-repository (2011) sera nécessaire.

Wigham, C.R. & Bayle, A. (à paraître). Enjeux, outils et méthodologie de constitution de corpus d'apprentissage, in Damiani M., Dolar K., Florez-Pulido C., Magnier J. & Loth R. (dir.) *Actes de Coldoc 2012*. Paris : Modyco

"imsmanifest.xml" dans un éditeur XML tel qu'*Oxygen* (SyncRO Soft SRL, 2012*).*

4.3.1 Composants du corpus

Les données primaires d'un corpus d'apprentissage sont organisées au sein de quatre répertoires correspondant aux quatre constituants du corpus : instanciation, scénario pédagogique, licences, protocole de recherche. Ces données sont dites "primaires" car elles ne sont pas brutes, elles ont été extraites, anonymisées et converties dans un format ouvert (Reffay et al, 2008).

4.3.1.1 Instanciation

Le composant "instanciation" est le noyau d'un corpus d'apprentissage. Premièrement, il regroupe les enregistrements des interactions des participants lors la formation, sous forme vidéo, audio ou textuelle, et les productions des participants de la situation d'apprentissage en ligne, par exemple, dans le cadre du projet *Slic*, les comptes rendus écrits à chaque séance par le groupe, ou dans le cadre du projet *Archi21* les images des modèles finaux créés à partir de la problématique. Ce composant peut également rassembler les traces système, par exemple, dans le cadre du projet *Slic,* le temps de connexion des participants à la plateforme *Moodle* et les statistiques sur leur participation. Deuxièmement, le composant instanciation rassemble les questionnaires remplis, les enregistrements des entretiens post-formation et les documents utilisés lors de l'expérimentation. (par exemple, les grilles utilisées pour conduire les entretiens ou des images et vidéos utilisées lors des entretiens pour provoquer une réponse du participant). Dans le projet *Archi21*, nous avons montré des images d'avatars aux étudiants lors des entretiens en leur demandant d'expliquer leur choix d'apparence d'avatar et nous avons également montré des clips des activités dans une activité d'auto-confrontation pour solliciter leurs explications sur la non-réussite de la tâche.

Avant de regrouper toutes les données et documents primaires (les ressources) dans la partie "instanciation", une phase de prétraitement est nécessaire. Ils sont tout d'abord anonymisés. L'anonymisation consiste à la fois à remplacer l'utilisation des patronymes par les codes acteurs et, si nécessaire, de modifier toutes les informations qui pourrait conduire à l'identification d'un participant ou qui pourraient biaiser l'interprétation des chercheurs. Ensuite, les données et documents sont convertis en formats ouverts et pérennes. L'élimination des formats propriétaires rend les ressources adaptables à différents outils que les chercheurs, à la fois à l'intérieur du projet et extérieurs à l'expérimentation, pourraient être amenés à employer.

4.3.1.2 Scénario pédagogique et protocole de recherche

Ces éléments sont considérés comme des éléments de contexte. Ils sont décrits en se référant à la norme IMS Learning Design (2003) et permettent de comprendre et de pouvoir traiter les données issues d'une expérimentation dans une situation de formation en ligne.

Nous avons choisi de modéliser nos scénarios pédagogiques et nos protocoles de recherche à l'aide du logiciel *MotPlus* (Paquette, 2009). *MotPlus* permet de créer des modèles correspondant à la norme IMS-LD. Il permet de décrire une situation d'apprentissage en faisant apparaitre les relations entre les micro-tâches, les environnements de communication où elle a lieu et le rôle des participants (tuteurs et apprenants) ainsi que d'exposer les étapes du protocole de recherche. De plus, comme *MotPlus* se réfère à la norme IMS-LD, la description dans ce format permet de rendre les scénarios pédagogiques et les protocoles de recherche interopérables et compréhensibles pour les chercheurs n'ayant pas participé à nos expérimentations. Dans le corpus, les scénarios pédagogiques et les protocoles de recherche peuvent être consultés en format html ou MotPlus.

4.3.1.3 Licences

Le composant "licence" est constitué de deux parties, l'une privée, l'autre publique. La partie privée contient tous les contrats de consentement signés avant l'expérimentation par les participants. Elle contient également un fichier tableur dans lequel figurent les coordonnées des participants et les patronymes liés aux codes acteurs utilisés pour l'anonymisation des données. Comme cette partie privée de la licence concerne le respect des droits et de l'éthique des participants, elle n'est pas

intégrée directement au corpus mais est conservée par le responsable de ce dernier.

La partie publique comprend un exemplaire du contrat de consentement remis aux participants et la licence d'utilisation du corpus qui indique les droits de l'éditeur du corpus et des futurs utilisateurs (chercheurs, praticiens). Pour les corpus déposés dans la banque de données Mulce-repository (2011), la licence *Creative Commons* est employée. Elle détaille les conditions sous lesquelles les analyses par des chercheurs extérieurs à l'expérimentation sont permises à partir du corpus.

4.3.2 Structuration du corpus

Une fois les composants du corpus rassemblés et traités, ils sont structurés. Un corpus d'apprentissage s'organise en trois parties (figure 4).

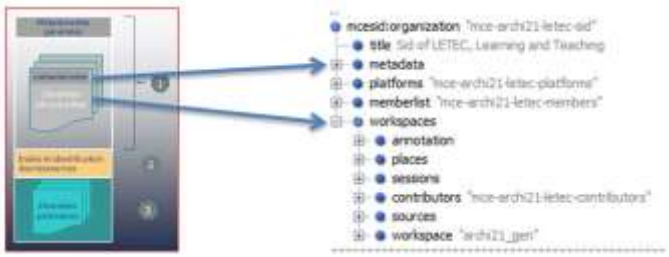

FIGURE 4 – Structuration du corpus d'apprentissage

Dans la partie 3 se trouvent les données primaires qui ont été prétraitées. A chaque ressource est attribué un identifiant. Un index regroupe chaque ressource avec un identifiant et un résumé de celle-ci (partie 2). Ce dernier est structuré dans le sens où les ressources sont regroupées dans l'index. Par exemple, l'ensemble de données concernant une activité ou concernant les entretiens se retrouveront ensemble. Finalement, les données sont structurées dans la partie 1 du corpus (le "manifeste").

Le manifeste est structuré en langage de balisage XML suivant plusieurs schémas. Nous avons utilisé le logiciel *Oxygen* (SyncRO Soft SRL, 2012). Le manifeste contient des informations sur chaque composant du corpus d'apprentissage vu dans la section 4.3.1. Il inclut un "memberlist" (voir figure 4) qui fournit des informations sur la biographie langagière de chaque participant et une liste de division des participants (apprenants-tuteurs-chercheurs / groupes de travail). La liste "platforms" décrit les environnements employés dans la formation.

L'exemple 1 montre un exemple d'entrée dans le "memberlist" du corpus *Archi21* (Chanier et Wigham, 2011). Le participant dont le code acteur est *Tfrez1,* est une tutrice de l'Université Blaise Pascal. Elle a 24 ans, le français est sa langue maternelle et l'anglais sa L2. Tous les autres participants sont décrits de la même manière dans le manifeste.

(1) <actor id="*Tfrez1*" designation="xxx" status="teacher" institution="*Université Blaise Pascal*" country="fra" gender="female" age="24" L1="fra " L2="eng" L3="esp"/>

La partie interactions est structurée de manière hiérarchique selon le "Structured Interaction Data model" (Mce_sid, 2011). Elle est organisée à partir de "workspace elements" qui correspondent à des "lieux" dans lesquels "des acteurs disposent d'outils (dotés de certaines fonctionnalités explicitées) et interagissent dans une période donnée" (Reffay *et al.,* 2008) (voir Figure 4). Dans nos corpus, un "workspace" correspond à une structure d'activités définie dans le scénario pédagogique.

Dans chaque "workspace", la description de l'environnement technologique où l'activité a eu lieu est donnée à côté des dates de début et de fin de l'activité, des acteurs qui participaient et des outils à leur disposition. Les ressources qui correspondent à la structure d'activité (partie 3) sont liées.

Un ensemble de métadonnées générales sur le corpus est aussi inclus dans le manifeste. Il contient les métadonnées concernant les contributeurs au corpus selon les standards déterminés par OLAC (Open Language Archives Community, Olac-metadata, 2008).

Les trois parties du corpus d'apprentissage sont enveloppées dans un conteneur ("content packaging") qui correspond à un format prescrit par IMS. Cela permet le dépôt dans une banque de

données, par exemple Mulce-repository (2011). L'intérêt de faire cela ne vient pas simplement des avantages du partage en libre accès des données structurées. Le dépôt permet également de rendre le travail visible par le référencement dans des réseaux tels que OLAC (2011) ou CLARIN (2012) et d'obtenir un identifiant OAI (Open Archives Initiative) permettant de le citer de la même façon qu'un article scientifique.

4.3.3 Post-rechercheDans cette section, nous décrivons les opérations d'analyse ayant lieu suite à la constitution et au dépôt du corpus global et, à partir de ces analyses, la notion de corpus distinguable.Transcription des interactions

La transcription des données recueillies dans le monde synthétique a été faite sur le logiciel *ELAN* (Max Planck, 2001) et selon une méthodologie prédéterminée (Saddour, Wigham et Chanier, 2011). Ce logiciel est bien adapté pour la transcription de la langue des signes et nous intéresse particulièrement pour transcrire les aspects non verbaux des séances de travail. Nos transcriptions rendent compte de tous les actes verbaux et non-verbaux effectués par les participants dans l'environnement. Chaque transcription indique une modalité liée à un acteur. Pour les actes verbaux, le contenu de la production est transcrit. Pour les actes non-verbaux, un code d'annotation est attribué à chaque type d'acte. Par exemple, le code "move(arm_L)" est utilisé si un avatar fait l'acte kinésique de bouger son bras gauche. Chaque annotation est alignée temporellement à la source vidéo. Le contenu écrit des annotations est en Unicode et la transcription conservée dans un format XML.

Tout fichier de transcription est lié à un fichier de métadonnées réalisé sur le logiciel *IMDI MEtadata* Editor (Max Planck, 2000-3). Il est conçu pour décrire des ressources et des données multimodales en format XML et permet de saisir plusieurs informations sur les acteurs (participants dans la séance vidéo, chercheurs, collecteurs, déposants, diffuseurs, éditeurs) et de fournir une description des différentes ressources (fichier vidéo, fichier clavardage, fichiers images).

4.3.4 Analyses

Une fois que les données sont structurées au sein d'un corpus, elles sont reliées et contextualisées et peuvent être analysées. Le format XML permet d'utiliser différents logiciels d'analyse tels que *Tatiana* (2008) pour l'aide à l'analyse d'interactions ou *Calico* (2009) pour l'analyse de forums. Il permet également d'annoter facilement les données et d'effectuer un certain nombre de calculs et de requêtes complexes (voir exemples d'analyse dans la section 5).

4.3.5 Constitution de corpus distinguables

A partir d'une analyse sur une question de recherche précise et à partir des données du corpus global, un chercheur peut produire un "corpus distinguable". L'intérêt de cette procédure est soit d'associer une publication avec les données analysées, soit de partager l'analyse dans laquelle les données sont mises en forme pour un outil employé lors de cette analyse. Chanier et Ciekanski (2010) expliquent qu'un corpus distinguable constitue, en même temps, un sous-corpus du corpus global d'apprentissage et un corpus en soi. Il utilise le même format qu'un corpus global LETEC mais ne contient que les données modifiées lors de l'analyse. Une description structurée du corpus est donnée par rapport au corpus global. Elle prend la forme de "commentaires libres et d'index précis renvoyant sur chacune des sous-parties du corpus global" (Chanier et Ciekanski, 2010 : para. 33).

Concernant nos projets, un corpus distinguable a été produit pour chaque séance transcrite. Le corpus contient dans son "manifeste" les données modifiées lors de la transcription. Dans la partie "workspace", pour la séance transcrite, chaque acte d'interaction est décrit par un identifiant, une référence à l'outil avec lequel il a été effectué (forum, clavardage), le type d'acte (modalité), une référence à l'auteur de l'acte et une date de début et de fin. Le contenu de l'acte est inclus. Pour les actes non verbaux, le code pour décrire l'acte figure à la place du contenu. Chaque acte, que l'interaction ait eu lieu dans l'audio, le clavardage ou dans une modalité non verbale, est donc encodé de façon homogène, ce qui permet la recherche ou la visualisation des données d'interaction de façon différente selon l'analyse à effectuer.

Wigham, C.R. & Bayle, A. (à paraître). Enjeux, outils et méthodologie de constitution de corpus d'apprentissage, in Damiani M., Dolar K., Florez-Pulido C., Magnier J. & Loth R. (dir.) *Actes de Coldoc 2012*. Paris : Modyco

Un corpus distinguable a également été produit à partir d'une analyse qui mettait en lien la communication non-verbale des avatars des apprenants avec la participation verbale en L2 dans le cadre du projet *Archi21* (Wigham et Chanier, 2012) dans le but d'associer la publication de l'analyse (Wigham et Chanier, à paraitre-b) avec les données employées.

Nous sommes actuellement en train de préparer des corpus distinguables concernant les analyses que nous allons maintenant présenter dans les sections 5.1 et **Erreur ! Source du renvoi introuvable.**.

5 Corpus d'apprentissage et facilitation d'analyse à travers des exemples

Nous présentons dans cette section deux exemples d'analyses tirés des formations Building Fragile Spaces et *Slic* afin de montrer comment l'approche par corpus a facilité nos analyses.

5.1 L'utilisation des modes de communication par des groupes d'apprenants différents

Dans le cadre de la formation Building Fragile Spaces, nous avons étudié l'utilisation des modalités verbales (audio et clavardage) par les participants (Wigham et Chanier, 2012b). Nous nous sommes intéressés à la place et au rôle du clavardage dans un monde synthétique où cette modalité est non seulement en compétition avec l'audio mais également avec des modalités non verbales. Nos questions de recherche étaient les suivantes :

- S'il est utilisé, pour quelles fonctions discursives le clavardage est-il employé ?
- Le clavardage offre-t-il aux tuteurs la possibilité de proposer de la rétroaction ?
- Vu la nature multimodale du monde synthétique, les étudiants arrivent-ils à répondre à la rétroaction éventuelle dans le clavardage ou cela présente-t-il une surcharge cognitive ?

Nous avons analysé les données de 6 séances réflexives (voir 3.3.1) à partir de 5 corpus distinguables dans lesquels les interactions sont transcrites (Chanier, Saddour et Wigham 2012a-e) et une ressource "resource-archi21-lact-slrefl-av-avi" du corpus global (Chanier et Wigham, 2011) qui a été transcrite mais pour laquelle nous mettons actuellement en place le corpus distinguable. 3 séances concernaient des groupes de travail dont la L2 était le français et 3 séances concernaient des groupes dont la L2 était l'anglais.

Pour chaque séance nous avons, à partir des transcriptions des séances, annoté les données en XML. Trois couches d'annotation ont été effectuées. Pour répondre à notre première question de recherche, nous avons annoté chaque acte de clavardage selon sa fonction discursive. 5 catégories et codes ont été employés : socialisation (soc), technique (tech), gestion du discours (cm), forme (form) et activité (task). Ensuite, pour analyser l'utilisation du clavardage pour la rétroaction, pour chaque acte de clavardage concernant la forme, nous avons annoté la rétroaction selon leur type (à partir de la classification de Bower et Kawaguchi, 2011) et l'auteur de la rétroaction (tuteur, pair, étudiant). Nous avons également annoté le type de production non standard auquel la rétroaction répondait (erreur typologique, lexicale, grammaticale, pragmatique, idiomatique ou de prononciation). Finalement, pour établir si les étudiants répondaient aux rétroactions et comment, nous avons annoté chaque instance de réponse selon quatre catégories: répétition de la rétroaction, intégration réussie de la rétroaction dans le discours, intégration non-réussie de la rétroaction dans le discours et l'accusé de réception ('acknowledgement') de la rétroaction.

L'exemple 2 illustre notre méthodologie d'annotation. Un participant, *Arnaudrez,* effectue un acte audio (tpa) d'une durée de 26 secondes. 12 secondes après le début de cet acte audio, le tuteur *Tfrez2* intervient dans le clavardage (tpc). La fonction discursive de son acte concerne la forme. L'annotation 37 montre que le tuteur corrige une erreur de type grammaticale (ntl="gram") en utilisation une reformulation (cf="rec"). Cette reformulation concerne l'annotation 36 dans la production d'*Arnaudrez.* Nous notons, dans l'annotation 38, que l'étudiant répète la rétroaction offerte (type="cf-rpt").

(2) tpa, *Arnaudrez* [12:31-12:57]: and this is a very personal work so +++ Brad gave some ways to to begin and + then our reflection <anno id="an36">lead lead us</anno> hm + different different ideas <anno id="an38"type="cf-rpt" ref="an37">led us</anno>

Wigham, C.R. & Bayle, A. (à paraître). Enjeux, outils et méthodologie de constitution de corpus d'apprentissage, in Damiani M., Dolar K., Florez-Pulido C., Magnier J. & Loth R. (dir.) *Actes de Coldoc 2012*. Paris : Modyco

tpc, <form>, *Tfrez2* [12:53-12:53]:<anno id="an37" function="form" ntl="gram" type="cf-rec" author="tut" ref="an36">led us</anno>

Le fait d'avoir structuré nos données d'interaction dans un corpus LETEC nous a permis de les annoter en XML. De ce fait, nous avons pu utiliser un outil d'analyse quantitative de corpus, *Comptage* (Lotin, 2012), sur les données structurées de chaque séance ce qui a facilité les comparaisons des données entre séances, entre groupes (français/anglais) et les différentes approches employées par les tuteurs.

Notre analyse a montré une différence entre l'utilisation des modalités audio et clavardage selon la L2 du groupe. Les groupes dont la L2 était l'anglais (sc et es) ont utilisé en moyenne 141 actes de clavardage par séance comparé à 150 actes audio en moyenne. Les groupes français (ls et av) ont utilisé en moyennne 21 actes de clavardage par séance comparé à 128 actes audio.

Concernant la fonction discursive des actes de clavardage, le clavardage n'était pas simplement utilisé quand des problèmes techniques survenaient dans l'audio. Dans 5 des 6 séances analysées, la majorité des actes de clavardage concernaient l'activité. Dans les séances en anglais, en moyenne 22% des actes de clavardage étaient des interventions concernant la forme. En revanche, dans les séances en français, seulement 1 ou 2 actes de clavardage concernaient la forme lors d'une séance donnée. Nous nous sommes donc concentrés sur les groupes de L2 anglais pour poursuivre l'analyse.

Pour les séances en anglais, la plupart de ces interventions faisaient suite à des erreurs de lexique (51%) ou de grammaire (36%). Nos données ont montré 3 exemples d'autocorrection dans le clavardage et 3 exemples de correction par un pair. 43 actes de rétroaction étaient offerts par le tuteur, dont 32 qui étaient une reformulation de la production audio d'un étudiant. 58% des rétroactions offertes par le tuteur d'anglais ont reçu des réponses de la part des étudiants. Le plus souvent, ces réponses prenaient la forme d'une répétition de la rétroaction (9) ou d'un accusé (9). Nos données ont montré 7 exemples de reprise de la forme corrigée dans le clavardage dont quatre étaient correctes. 20 des 25 réponses à la rétroaction étaient dans des actes audio. Cela montre la capacité des étudiants à jongler entre les deux modalités.

Dans Wigham et Chanier (2012b) nous détaillons beaucoup plus cette étude. Ici, nous souhaitons surtout souligner que l'analyse a été possible, voire facilitée, grâce à la vue d'ensemble donnée par le corpus structuré. Cela nous a permis, premièrement, de concevoir une méthodologie pour coder des séances en XML et donc, par la suite, d'utiliser des outils d'analyse de corpus et, deuxièmement, de faciliter la comparaison des analyses entre séances, groupes et tuteurs. L'utilisation d'un format d'annotation de balisage XML permet ensuite à d'autres chercheurs d'approfondir notre analyse. Par exemple, l'étude de Rodrigues et Wigham (2012) a repris les données annotées dans le cadre de cette étude sur le clavardage pour ajouter une quatrième couche d'annotations XML dans le but d'étudier l'aide à la résolution des points de vocabulaires problématiques. Nous voyons donc le gain d'un corpus structuré pour que dans une équipe de chercheurs, chacun avec ses propres questions de recherche puisse travailler ensemble : l'analyse de chacun enrichit le corpus et par la suite, des analyses croisées aident à mieux comprendre les interactions dans le monde synthétique.

5.2 Analyse des interactions animateur-étudiants

A partir des données issues de la dernière étape du projet *Slic*, nous avons analysé et comparé les interactions entre animateurs et étudiants dans deux groupes différents (Bayle et Youngs, 2012). Notre hypothèse était que les animateurs avaient différentes "techniques" d'animation, plus ou moins similaires à la structure d'interaction "traditionnelle" enseignant-apprenant et que cela avait une influence sur les interactions des étudiants entre eux et avec l'animateur.

Notre analyse a donc focalisé sur les différents styles d'animations de deux animateurs ainsi que sur l'influence de ceux-ci sur l'interaction dans le groupe entre l'animateur et les étudiants ainsi qu'entre les étudiants entre eux. Nos questions de recherche étaient les suivantes :

- Quel comportement l'animateur adopte-t-il durant la séance étudiée ?
- Comment se réalise l'interaction dans le groupe en fonction du comportement de l'animateur ?

Wigham, C.R. & Bayle, A. (à paraître). Enjeux, outils et méthodologie de constitution de corpus d'apprentissage, in Damiani M., Dolar K., Florez-Pulido C., Magnier J. & Loth R. (dir.) *Actes de Coldoc 2012*. Paris : Modyco

Nous avons analysé les données des transcriptions de la dernière séance du projet pour deux des sept groupes. Chaque acte de parole a été annoté en fonction de plusieurs critères. Dans un premier temps, nous avons annoté le destinataire de chaque acte de parole (animateur, groupe, étudiant). A un niveau plus précis, nous avons annoté la fonction de l'acte de parole (salutations, aspects techniques, contribution, nouvelle proposition, question de compréhension, réponse fermée, etc.). Nous avons également effectué des comptages généraux (nombre de mots, nombre de tours de parole, fréquence de prise de parole au cours de la séance, etc.) pour dresser un profil des séances et les comparer.

Notre analyse a révélé une différence de style d'animation entre les deux animateurs ainsi qu'une différence dans la manière dont l'interaction se déroulait dans le groupe. En effet, dans le premier groupe étudié, les deux étudiants de master ont pris le rôle d'animateur, d'où un déséquilibre préalable. L'animateur "officiel" contrôlait totalement la discussion et l'interaction était du type de celle que l'on retrouve traditionnellement en classe : question de l'enseignant, réponse de l'étudiant, rétroaction de l'enseignant (McCarthy, 1991). De plus, les questions des animateurs étaient souvent directives, fermées et nominatives, ce qui empêchait les étudiants américains de développer leurs idées. A l'inverse, dans le deuxième groupe, l'animatrice donnait aux étudiants la liberté d'interagir entre eux, de prendre des responsabilités dans le déroulement des tâches, ses questions étaient souvent ouvertes et appelaient à un dialogue, une négociation. Son comportement se rapprochait de ce que Shrum et Glisan (2010) appellent un comportement positif de l'enseignant.

Du côté des étudiants américains, on retrouve également des différences dans les manières d'interagir liées aux comportements des animateurs. Dans le premier groupe, des comportements directifs et dominants incluant des questions dirigées nominativement à un étudiant ont empêché les étudiants d'émettre de nouvelles idées, d'interagir entre eux. Les étudiants de ce groupe se sont également positionnés comme simples apprenants et non comme membre d'un groupe et se sont contentés de répondre aux questions des animateurs sans prendre d'initiatives. Dans le deuxième groupe, les étudiants américains ont pu prendre des initiatives, s'exprimer librement et se sont adressés à la fois à l'animatrice et à leurs pairs.

Il semblerait que les comportements directifs et dominants des animateurs ne créent pas un environnement dans lequel le travail collaboratif peut prendre forme. Des analyses sont en cours pour déterminer le lien entre les styles d'animation des animateurs de chaque groupe et la dimension collective, sinon collaborative de la réalisation des tâches dans les groupes.

Le partage des données et des annotations effectuées par les deux chercheurs a été facilité par l'organisation des données. Les données de transcription au format XML ont pu être traitées par le logiciel Tatiana (2008) qui permet de visualiser les interactions et donc d'aider à l'analyse. Les analyses et comparaisons entre les deux groupes ont été facilitées grâce à la structuration préalable des données et la possibilité de lier données d'interactions, productions et scénario pédagogique.

6 Conclusion

Nous avons montré dans cet article les différentes étapes de constitution d'un corpus d'apprentissage, de l'élaboration du dispositif d'expérimentation jusqu'aux phases d'analyse et de diffusion des résultats. La structuration des données en corpus permet de travailler à différents niveaux d'analyse en ayant une vue d'ensemble de la formation étudiée.

L'approche par corpus d'apprentissage n'est pas encore totalement développée dans le domaine de la recherche en didactique des langues. Elle peut sembler coûteuse en temps, demander des compétences que tous les chercheurs n'ont pas nécessairement. Pourtant ses avantages sont nombreux, que ce soit au niveau des analyses mais également en ce qui concerne la diffusion, le partage des données, la visibilité et la continuité des travaux de recherche. En effet, le temps passé à s'approprier la méthodologie et à structurer les données est rapidement compensé par la possibilité d'effectuer des analyses à partir de différents outils sans avoir besoin d'un formatage spécifique à chaque fois et par la reconnaissance du travail du chercheur grâce à la mise à disposition du corpus à la communauté scientifique (Chanier et Ciekanski, 2010).

Wigham, C.R. & Bayle, A. (à paraître). Enjeux, outils et méthodologie de constitution de corpus d'apprentissage, in Damiani M., Dolar K., Florez-Pulido C., Magnier J. & Loth R. (dir.) *Actes de Coldoc 2012*. Paris : Modyco

Dans le cadre d'une thèse, la méthodologie à adopter n'est pas toujours évidente. L'approche LETEC accompagne le doctorant tout le long de son parcours et facilite donc son travail. Elle permet également de valoriser son travail de recherche par le dépôt en ligne du corpus. De plus, le fait de pouvoir mettre en parallèle les données utilisées avec les résultats dans les publications permet d'assurer la validité des analyses qui peuvent, grâce à la mise à disposition du corpus, être vérifiées par la communauté de chercheurs. Comme un travail de thèse s'inscrit aujourd'hui de plus en plus dans le cadre de projets incluant différents partenaires, au sein d'un même laboratoire ou avec différentes institutions, la constitution d'un corpus d'apprentissage permet également de faciliter l'accès aux données, le partage des analyses entre chercheurs et l'approfondissement d'analyses déjà effectuées.

Références

Toutes les URL étaient valides le 2 septembre 2012

Antonacci, D., Dibartolo, S., Edwards, N., Fritch, K, Mcmullen, B. et Murch-Shafer, R. (2008). *The Power of Virtual Worlds in Education: A Second Life Primer and Resource for Exploring the Potential of Virtual Worlds to Impact Teaching and Learning*. Angel Learning. [http://www.angellearning.com/products/secondlife/downloads/The%0Power%20of%20Virtual%20Worlds%20in%20Education_0708.pdf]

Active Worlds Inc. (1997). *Active Worlds* [logiciel]. Las Vegas : Active Worlds Inc. [http://www.activeworlds.com]

Bayle, A. et Foucher, A.-L. (2011). Comment étudier les interactions d'apprenants de langue dans les mondes virtuels ? Dans Dejan, C., Mangenot, F. et Soubrié, T. (dir.), *Actes du colloque Echanger pour apprendre en ligne* (EPAL). Grenoble, 24-26 juin 2011. [http://w3.u-grenoble3.fr/epal/dossier/06_act/actes2011.htm]

Bayle, A., Foucher, A-L. et Youngs, B. (2012). *SLIC*: Second Life as a Collaborative Tool for Graduate Teacher Training and Developing Intercultural Communicative Competences. Communication à *CALICO 2012*, 12-16 juin, 2012, Notre-Dame : Etats-Unis. [http://edutice.archives-ouvertes.fr/edutice-00688378]

Bayle, A. et Youngs, B. (2012). Patterns of Interaction Between Moderators and Learners during Synchronous Oral Discussions Online. Document de travail. [http://edutice.archives-ouvertes.fr/edutice-00726762]

Beepa (2012). *Fraps* version 3.5.9 [logiciel]. Beepa Pty Ltd. [http://www.fraps.com]

Bell, M. (2008). Toward a Definition of "Virtual Worlds". *Journal of Virtual Worlds Research, 1*(1). [http://journals.tdl.org/jvwr/article/view/283/237]

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

Blanchet, P. (2011). Les principales méthodes et leurs techniques de construction des observables. Dans Blanchet, P. et Chardenet, P. (dir.) *Guide pour la recherche en didactique des langues et des cultures* (p. 73-192). Paris : Editions des archives contemporaines.

Book, B. (2004). Moving Beyond the Game: Social Virtual Worlds. Communication à *State of Play 2 Conference 2004*, 13-15 Novembre, 2004, New York : Etats-Unis. [http://www.virtualworldsreview.com/papers/BBook_SoP2.pdf]

Bower, J. et Kawaguchi, S. (2011). Negotiation of meaning and corrective feedback in Japanese/English eTandem. *Language Learning* et *Technology (LLT), 15*(1), 41-71.

Calico (2009). Communautés d'apprentissage en ligne, instrumentation, collaboration. ERTé Calico. [http://woops.crashdump.net/calico]

Castronova, E. (2005). S*ynthetic Worlds: the Business and Culture of Online Games,* Chicago : University of Chicago Press.

Chanier, T. et 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. *ALSIC,* 13, [doi : 10.4000/alsic.1666]

Wigham, C.R. & Bayle, A. (à paraître). Enjeux, outils et méthodologie de constitution de corpus d'apprentissage, in Damiani M., Dolar K., Florez-Pulido C., Magnier J. & Loth R. (dir.) *Actes de Coldoc 2012*. Paris : Modyco

Chanier, T., Saddour, I. et Wigham, C.R. (dir.) (2012a). *Distinguished Corpus: Transcription of Verbal and Nonverbal Interactions of the Second Life Reflection archi21-slrefl-es-j3*. Mulce.org : Clermont Université. [oai : mulce.org:mce-archi21-slrefl-es-j3 ; http://repository.mulce.org]

Chanier, T., Saddour, I. et Wigham, C.R. (dir.) (2012b). *Distinguished Corpus: Transcription of Verbal and Nonverbal Interactions of the Second Life Reflection archi21-slrefl-av-j2*. Mulce.org : Clermont Université. [oai : mulce.org:mce-archi21-slrefl-av-j2 ; http://repository.mulce.org]

Chanier, T., Saddour, I. et Wigham, C.R. (dir.) (2012c). *Distinguished Corpus: Transcription of Verbal and Nonverbal Interactions of the Second Life* Reflection archi21-slrefl-ls-j3. Mulce.org : Clermont Université. [oai : mulce.org:mce-archi21-slrefl-ls-j3 ; http://repository.mulce.org]

Chanier, T., Saddour, I. et Wigham, C.R. (dir.) (2012d). *Distinguished Corpus: Transcription of Verbal and Nonverbal Interactions of the Second Life Reflection archi21-slrefl-sc-j2*. Mulce.org : Clermont Université. [oai : mulce.org:mce-archi21-slrefl-sc-j2 ; http://repository.mulce.org]

Chanier, T., Saddour, I. et Wigham, C.R. (dir.) (2012e). *Distinguished Corpus: Transcription of Verbal and Nonverbal Interactions of the Second Life* Reflection archi21-slrefl-sc-j3. Mulce.org : Clermont Université. [oai : mulce.org:mce-archi21-slrefl-sc-j3 ; http://repository.mulce.org]

Chanier, T. et Wigham, C.R. (dir.) (2011). Learning and Teaching Corpus (LETEC) of ARCHI21. Mulce.org : Clermont Université. [oai : mulce.org:mce-*Archi21*-letec-all, http://repository.mulce.org]

Clarin (2012). *Virtual language observatory* [banque de données] [http://catalog.clarin.eu/]

Dalgarno, B., et Lee, M. J. W. (2010). What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology*, *41*(1), 10–32. [doi:10.1111/j.1467-8535.2009.01038.x]

Granger, S., Hung, J. et Petch-Tyson, S. (dir.) (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. (p. 3-33). Amsterdam : Benjamins. [http://hdl.handle.net/2078.1/75823]

IMS-CP (2007). Schema for IMS Content Package. *IMS Global Learning Consortium 1999-2007*. [http://www.imsglobal.org/xsd/imscp_v1p1.xsd]

Joint Quality Initiative Informal Group (2004). *Shared 'Dublin' descriptors for Short Cycle, First Cycle, Second Cycle, and Third Cycle Awards.* [rapport] [http://www.jointquality.org/content/descriptors/CompletesetDublinDescriptors.doc]

Kwiksurveys (n.d.). Kwiksurveys. [http://kwiksurveys.com/]

Labov, W. (1972). *Sociolinguistic patterns.* Oxford : Blackwell.

Lotin, P. (2012). *Comptage.* Clermont-Ferrand : Clermont Université.

Max Planck (2000-2003). *IMDI Editor*. Nijmegen: Max Planck Institute for Psycholinguistics.[http://www.lat-mpi.eu/tools/imdi/editor/]

Max Planck (2001). ELAN. Nijmegen: Max Planck Institute for Psycholinguistics. [http://www.lat-mpi.eu/tools/elan/]

Mce_sid_letec (2011). Schéma décrivant les différentes structures d'interactions. [http://lrl-diffusion.univ-bpclermont.fr/mulce/metadata/mce-schemas/mce_sid.xsd]

McCarthy, M. (1991). *Discourse Analysis for Language Teachers*. Cambridge : Cambridge University Press.

Moove (1999). *Moove Online.* [logiciel]. Köln : Moove. [http://www.moove.com]

Mulce-Documentation (2011). *Site web expliquant la méthodologie Mulce et les informations autour du projet Mulce.* Mulce.org : Clermont Université. [http://mulce.org]

Mulce-Repository (2011). *Banque de données Mulce*. Mulce.org : Clermont Université. [http://repository.mulce.org]

Nikiforov, A. (2011). *MP3 Skype Recorder* version 1.9.0 [logiciel]. London : Nikiforov. [http://voipcallrecording.com/]

Olac (2011). *OLAC: Open Language Archives Community.* [http://www.language-archives.org/]

Paquette, G. (2009). Mot Plus version 1.6.7 [logiciel]. Québec : Licef, Télé-Université.

Peterson, M. (2011). Towards a Research Agenda for the Use of Three- Dimensional Virtual Worlds in Language Learning. *Calico Journal*, *29*(1), 67–80. [https://calico.org/html/article_893.pdf]

Reffay, C., Chanier, T., Noras, M. et 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]

Reffay, C., Betbeder, M.-L. et Chanier, T. (2012). Multimodal learning and teaching corpora exchange: lessons learned in five years by the Mulce project. *International Journal of Technology Enhanced Learning (IJTEL), 4*(12), 11-30.

Rodrigues, C. et Wigham, C.R. (2012). Second Life et apprentissage d'une langue étrangère dans une approche Emile : quels apports d'un environnement synthétique pour l'apprentissage du vocabulaire ? *Colloque ACEDLE*, 7-9 juin 2012, Nantes. [http://edutice.archives-ouvertes.fr/edutice-00703113]

Rodrigues, C., Wigham, C.R., Foucher, A-L. et Chanier, T. (à paraître). Architectural design and language learning in Second Life. Dans Gregory, S., Lee, M.J.W., Dalgarno, B. et Tynan, B. (dir.) *Virtual Worlds in Online and Distance Education,* Edmonton : Athabasca University Press

Rosedale, P. (2011). *Second Life,* version 2.7.2(233432) [logiciel]. San Fransisco : Linden Lab. [http://www.secondlife.com]

Saddour, I., Wigham, C.R. et Chanier, T. (2011). *Manuel de transcription de données multimodales dans Second Life*. Document interne, Laboratoire de Recherche sur le Langage. [http://halshs.archives-ouvertes.fr/edutice-00676230/]

Sanchez, B. (1996). MOOving to a new frontier in language teaching. Dans Warschauer, M. (dir.). *Telecollaboration in foreign language learning*. Honolulu, HI: Second Language Teaching and Curriculum Center, University of Hawai'i.

Schwienhorst, K. (2002). Why virtual, why environments? Implementing virtual reality concepts in computer-assisted language learning. *Simulation* et *Gaming*, *33*(2), 196-209. [http://dx.doi.org/10.1177/1046878102332008]

Shrum, J. L. et Glisan, E. W. (2010). *Teacher's Handbook: Contextualized Language Instruction*. Boston: Heinle, Cengage Learning.

SyncRO Soft SRL (2012) *Oxygen XML Editor* version 14.0. Craiova: Syncro Softsrl. [http://www.oxygenxml.com/]

Tatiana (2008). Trace Analysis Tool for Interaction ANAlysts [logiciel].[http://lead.emse.fr]

TechSmith Corporation (2010). Camtasia Studio version 8.0 [logiciel] [http://www.techsmith.com/camtasia.html]

Teoh, J. (2007). Second Life, a simulation: barriers, benefits, and implications for teaching. Technology, *Colleges & Community (Tcc) Worldwide Online Conference 2007 Proceedings*.

Wigham, C.R. et Chanier, T. (à paraître-a). A study of verbal and nonverbal communication in Second Life - the *ARCHI21* experience. *ReCALL, 25*(1).

Wigham, C.R. et Chanier, T. (à paraître-b). Architecture students' appropriation of avatars – relationships between avatar identity and L2 verbal participation and interaction. Dans Lamy, M-N. et Zourou, K. (dir.). *Social Networking for Language Education*. Basingstoke : Palgrave Macmillan.

Wigham, C.R. et Chanier, T. (2012) (dir.) *Distinguished Corpus: Influence of nonverbal communication on verbal production in the Second Life Reflective Sessions*. Mulce.org: Clermont Université. [oai : mulce.org:mce-Archi21-modality-interplay, http://repository.mulce.org]

Wigham, C.R. et Chanier, T. (2012b). Interactions between text chat and audio modalities for L2 communication in the synthetic world Second Life. *15th International CALL Research Conference*, 25-27 mai 2012, Taichung : Taiwan. [http://hal.archives-ouvertes.fr/hal-00660865]

Yee, N. et Bailenson, J.N. (2008). A method for Longitudinal Behavioural Data Collection in Second Life. *Presence: Teleoperators and Virtual Environments*, 17(6), 594-596.