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

Thierry Chanier, Maud Ciekanski

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

Thierry Chanier & Maud Ciekanski

Version du 8 mai 2010

Article à paraître dans la revue ALSIC : Apprentissage des Langues et Systèmes d'Information et de Communication : http://alsic.revues.org

Résumé : La recherche sur les interactions en ligne en situation d'apprentissage offre encore trop peu souvent la possibilité d'accéder aux données à partir desquelles les chercheurs ont élaboré les analyses présentées dans les publications. Cela restreint, d'une part, la compréhension des phénomènes étudiés et, d'autre part, empêche toute réplication dans le but de comparaisons, d'analyses cumulatives ou contrastives. Dans le projet Mulce, nous défendons le point de vue méthodologique suivant : pour permettre une analyse des interactions situées, il convient de relier les différentes données issues de formations en ligne pour construire un objet d'analyse, exploitable par différentes équipes et disciplines. Le constat actuel est que les données sont souvent décontextualisées, parcellaires ou simplement inaccessibles à la communauté des chercheurs. Nous proposons donc de structurer les données en corpus d'apprentissage (LETEC) de façon à rendre possible leur échange et la capitalisation des analyses. Le protocole de recherche, le scénario pédagogique, les interactions, productions et traces, les licences et les analyses capitalisables en sont les constituants. Cet article présente, dans un premier temps, les questionnements, à la fois théoriques, techniques et méthodologiques soulevés par la conception d'un tel projet. Dans un deuxième temps, nous illustrerons notre démarche à partir d'exemples issus des formations " Simuligne " et "Copéas", en indiquant les processus simples de transformation du format Mulce aux formats requis par deux logiciels d'aide à l'analyse (l'un sur les forums, l'autre sur l'alignement vidéo et transcription). Nous insistons plus particulièrement sur l'intérêt de ces outils pour l'analyse des phénomènes de polyfocalisation et d'écriture multimodale dans l'analyse des interactions multimodales, caractéristiques des environnements d'apprentissage en ligne. Nous conclurons notre propos sur les modes de valorisation scientifique du travail du chercheur confronté à la collecte et à la structuration de corpus d'apprentissage.

Mots-clés : Corpus d'apprentissage, LETEC, interactions en ligne, interactions multimodales, réplication, partage d'outils et d'analyses.

English version: The study of online learning, whether aimed at understanding this form of situated human learning, at evaluating relevant pedagogical scenarios and settings or at improving technological environments, requires the availability of interaction data from all participants in the learning situations. However, usually data are either inaccessible, or of limited access to those who were not involved in the original project. Moreover data are fragmented, therefore decontextualized with respect to the original teaching / learning settings. Sometimes there are buried in a proprietary format within the technological environment. The consequence is that research lacks a scientific basis. In the literature comparisons are often attempted between objects that are ill-defined and may in fact be different. The processes of scientific enquiry, such as re-analyzing, replicating, verifying, refuting or extending the original findings, are therefore disabled. To address this anomaly, we propose to create and disseminate a new type of corpus, contextualized learner corpus,

entitled "LEarning and TEaching Corpus" (LETEC). Such corpora include not only the data that correspond to output of learner activity in online courses, but also their context. Sharing LETEC corpora within the research community implies that: (1) corpora are formatted and structured according to a new model which is compatible with existing standards for corpora and for learning design specifications; (2) corpora are placed on a server offering cross-platform compatibility and free access; (3) an ethics policy is formulated as well as copyright-licences. This paper presents the answers brought by our Mulce project to these questions from a theoretical and methodological standpoint. We give examples extracted from two learning and teaching corpora (Simuligne and Copéas). We show how data structured accordingly to the Mulce format can be transformed and processed by analysis tools available in different research communities. The article sheds light on the potential of such tools for the study of polyfocalisation and multimodal writing. It concludes with the scientific benefits researchers way expect at an institutional level when collecting and structuring data.

Keywords: Learning and teaching corpus, LETEC, online interaction, multimodal interaction, replication, sharing tools and analysis.

1 Introduction

Le récent développement des environnements multimodaux d'échanges synchrones en ligne suscite l'engouement de plus en plus de formateurs et d'apprenants depuis une décennie, en particulier dans le domaine de l'apprentissage des langues. Ces plateformes de formation générant des interactions complexes entre les participants renouvellent la recherche sur les interactions en ligne en situation d'apprentissage, notamment le questionnement sur les traces et leur traitement. En outre, la multimodalité, pourvue de nouveaux atouts, a suscité de nombreuses analyses qualitatives et quantitatives, motivées par les types originaux de dialogues qu'elle favorise, par le soutien qu'elle offre dans la communication en langue-cible (par exemple, l'écrit en soutien de l'oral et inversement), voire dans l'apprentissage de la langue, avec la mise en place de stratégies convoquant plusieurs modalités. Ces travaux ont montré que la multimodalité prend des formes différentes en fonction du type d'environnement technologique choisi[1] : elle peut être uniquement *verbale* (les environnements *audio-synchrones* intégrant seulement audio et clavardage, (Jepson, 2005)), ou combiner une communication *verbale* et *non-verbale* comme dans les environnements de vidéoconférence intégrant la vidéo, l'audio, le clavardage et l'iconique (Wang, 2004), ou plus encore, comme dans les environnements *audio-graphiques synchrones* où s'ajoutent aux précédentes modalités des collecticiels incluant les modes texte, image et graphique (Chanier et Vetter, 2006 ; Lamy, 2007).

Cependant, malgré leur nombre, les publications rapportant ces analyses ne permettent pas d'établir de comparaison entre ces environnements. De plus, la méthodologie de recherche habituellement utilisée dans l'analyse des interactions en ligne, qu'elles soient synchrones ou asynchrones, ne prend souvent en compte qu'une fraction des données, voire qu'un seul type d'interaction privilégiant un outil spécifique (quelques forums, quelques extraits de séances synchrones), et ne les organise pas de façon systématique. L'absence de pratiques partagées et de guides méthodologiques concernant l'élaboration de corpus recueillant ces interactions, tout comme l'absence de corpus organisés à des fins d'évaluation des différentes approches articulant technologie et pédagogie, freinent la recherche. Au premier abord, des raisons évidentes semblent expliquer ces manques. Les corpus peuvent être perçus comme coûteux à collecter et difficiles à organiser. En effet, de multiples niveaux d'annotations s'offrent aux chercheurs du fait que modes et modalités interagissent constamment pour convoyer des actions à caractères verbales ou non verbales ; ces actions peuvent elles-mêmes être organisées dans des schèmes à caractères communicatifs, ou collaboratifs.

Certaines communautés de chercheurs ont cependant décidé de relever le défi, comme celle centrée sur l'apprentissage collaboratif en ligne (CSCL, *Computer-supported Collaborative Learning*). Il s'y développe un intérêt grandissant pour la méthodologie de structuration de corpus et la définition de formalismes pivots, motivées par les perspectives

[1] Sur l'organisation des modes et modalités dans ces différents contextes technologiques, voir (Ciekanski et Chanier, 2008 ; Chanier et Vetter, 2006).

d'échanges, de partage des données, des analyses et des outils. En ce qui concerne la communauté étudiant l'apprentissage des langues en ligne (que nous étiquetterons AL&SIC, pour Apprentissage des Langues et Systèmes d'Information et de Communication), l'intérêt qu'elle accorde à la communication multimodale ne l'a pas encore attirée vers ces questions méthodologiques. La comparaison avec d'autres domaines des sciences du langage montre cependant que leur prise en considération va de pair avec une recherche mieux structurée. Ainsi, en inscrivant ces questions méthodologiques au cœur de nos préoccupations, nous pouvons bénéficier des outils mis au point dans les autres domaines, contribuer à la recherche en général en sciences du langage et, plus simplement, mieux appréhender la nature de la multimodalité et ses potentialités en termes d'apprentissage.

Afin d'ouvrir la discussion sur ce sujet, nous nous appuierons sur l'expérience d'un projet de recherche où ont été élaborés des corpus provenant de différentes formations en ligne, sur ces dix dernières années, projet lors duquel nous avons commencé à partager données, analyses et outils avec d'autres équipes de recherche. L'un des objectifs du projet Mulce (*MULtimodal Corpus Exchange* ; Mulce, 2009a) était de définir la notion de corpus d'apprentissage. Pour ce faire, il s'agissait de constituer un tel corpus à partir de données collectées lors d'une formation expérimentale en ligne, de structurer ces données de façon à rendre possible leurs échanges, enfin de pouvoir capitaliser des analyses faites successivement sur tout ou partie des données. L'étude a été conduite avec l'idée de pouvoir y intégrer une description du protocole de recherche associée à la formation, son scénario pédagogique, les interactions, productions et traces générées par les participants, sans négliger les questions touchant à l'éthique et aux droits. Quant aux analyses, nous avions dès le début songé à pouvoir utiliser les outils mis au point par les communautés scientifiques travaillant sur les interactions et le langage, telles que les EIAH (Environnements Informatiques pour l'Apprentissage Humain), CSCL, ou le TAL (Traitement Automatique du Langage). Cela imposait d'emblée un format de structuration du corpus qui puisse être automatiquement traduit dans les divers formats requis par ces outils.

Ces contraintes étant sommairement posées, l'article expose les réponses que l'équipe Mulce a apportées, ainsi que les réalisations associées. Dans un premier temps, nous montrons l'intérêt de travailler à partir de corpus, en nous appuyant sur des exemples, des références aux travaux provenant de plusieurs domaines des sciences du lange, en particulier celui de l'enseignement/apprentissage des langues. Le terme corpus étant décliné de multiples façons, nous avons évité d'en donner une définition, par nature trop restrictive, et préférons parler de paradigme qui représente mieux toutes les facettes de cet objet complexe. De ce paradigme découle l'approche méthodologique mise en œuvre dans Mulce. La section suivante donne une vue d'ensemble de ce qu'est un corpus d'apprentissage, en abordant les aspects constituants et structuration. Dans un troisième temps, à partir d'exemples issus d'interactions verbales et non-verbales dans deux formations (l'une asynchrone, l'autre synchrone), nous abordons la notion de granularité des corpus en vue de leur partage et de leur analyse. Nous indiquons également le processus de transformation du format Mulce aux formats requis par deux logiciels d'aide à l'analyse (l'un sur les forums, l'autre sur l'alignement vidéo et la transcription), issus des recherches sur l'analyse des traces en EIAH et en CSCL. Nous terminons sur les conditions permettant l'échange de tels corpus et leur libre accès.

2 Intérêt du corpus dans nos domaines de recherche, définition du paradigme et applications

Avant d'aborder la notion de corpus, nous commencerons par considérer un ensemble plus limité de données recueillies lors d'une expérimentation centrée sur une thématique de recherche en pleine expansion en AL&SIC. Cet ensemble de données a fait l'objet de plusieurs études donnant lieu à toute une série de publications et de citations. Leurs auteurs en ont livré des interprétations différentes. Cet exemple permet de s'interroger sur la façon d'effectuer des recherches à partir de données incomplètes, en partie décontextualisées. Nous introduirons alors, en contre-proposition, la notion de corpus. Ce n'est pas une simple définition, mais plutôt un ensemble de traits qui nous permettrons de délimiter le cadre opératoire de cette notion. Les références seront principalement tirées d'un champ particulier de la linguistique de corpus, celui qui s'intéresse à l'apprentissage des langues. Cette section se terminera sur le besoin de définir un nouveau genre dans le paradigme corpus, genre spécifique aux situations d'apprentissage en ligne.

2.1 Interprétations multiples sur données incomplètes

En 1997, une formation en ligne met en rapport une classe de lycée français et une classe d'étudiants aux États-Unis. Lors des communications exolingues par courriel autour de tâches interculturelles survient un incident critique entre les deux groupes. Ce type de phénomènes, qui fait l'objet d'une grande attention dans les recherches sur l'interculturel (Audras et Chanier, 2008) est d'abord expliqué par l'auteur de l'expérimentation comme étant le résultat d'incompréhensions de nature linguistique (Kern, 2000). Par la suite, une description sommaire de la situation de formation et un ensemble fragmentaire des données d'interactions sont transmises par l'auteur à deux autres chercheurs, Kramsch et Thorne, qui n'ont pas participé au projet initial. Ceux-ci en donnent une interprétation différente (Kramsch et Thorne, 2001) en invoquant la prise en compte de compétences de communication spécifiques à l'Internet qui doivent être étudiées en tant que telles à l'échelle de la globalisation des communications. Thorne (2003) poursuit cette réinterprétation en l'orientant vers les cultures et les littératies des participants (voir aussi (Kern, Ware et Warshauer, 2004) pour cette mise en perspective des différentes étapes). Enfin, Basharina (2007 : 84), reprend une partie de cet historique des publications et appuie la dernière interprétation :

> *Thorne (2003) argues that online and other activities […] represent the "culture-of-use" of an artifact (p. 40). He found that the activity of online interaction was different for the French than it was for the Americans, in part because the Internet communication was used differently in each case; e.g., French students were communicating through a surrogate (the teacher who was sending their messages). Thorne concludes that radically different cultures-of-use of Internet communication was one of the major reasons for the tension between the French and American students.*

Cette dernière interprétation pose question. Qu'entend-on par "culture d'usage de l'Internet" ? Veut-on désigner des formes de communication en ligne qui seraient différentes entre adolescents étasuniens et français ? Les jeunes français dans le cadre de leurs loisirs utilisaient-ils en 1997 Internet, par exemple les messageries instantanées et le courriel, d'une manière fondamentalement différente de celles des étasuniens ? Ne pourrait-on pas plutôt parler de cultures institutionnelles différentes ? Car la classe de langues dans le secondaire en France, avec ses programmes, sa culture de formation des enseignants, ses

protocoles d'échanges entre enseignant et élèves était et reste bien différente de celle de la culture d'apprentissage en langues à l'université aux États-Unis. Quant à Internet, son usage à l'époque dans les institutions françaises du secondaire était fortement contrôlé, ce qui pourrait expliquer le choix contre-productif de l'enseignant de faire passer tous les messages des lycéens par son courriel individuel. Que penser en outre de la compétence interculturelle de l'enseignant ? Était-il sensibilisé à l'occurrence et la gestion des incidents critiques, compétence dont l'importance a été soulignée indépendamment de celle concernant la communication sur Internet ?

Beaucoup d'interprétations divergentes sont ainsi rendues possibles par le fait que l'on ne dispose ni de données sur le scénario pédagogique convenu entre les deux enseignants, ni de questionnaires qu'apprenants et enseignants auraient pu remplir, en précisant notamment leurs usages de l'Internet, leurs pré-conceptions sur l'interculturel, et en recueillant leur avis sur le déroulement de la formation.

Notre propos n'est nullement de critiquer le protocole expérimental mis en place par Kern dans une formation exploratoire tout à fait originale en 1997, ni le fait que celui-ci ait transmis une partie des données à d'autres collègues. Bien au contraire, puisque nous développerons plus loin l'importance du partage entre chercheurs. Le problème est plutôt le fait que la recherche se soit développée sur des données trop parcellaires. Le discours scientifique qui s'élabore à partir d'exemples en partie décontextualisés prend alors un caractère impressionniste. Le jeu des citations croisées amplifie le phénomène et oriente le lecteur vers une interprétation stéréotypique, que l'on pourrait qualifier de culturellement marquée (vision nord-américaine sur les usages différenciés d'Internet entre leur pays et les autres). On retrouve ici en quelque sorte, sur un plan épistémique, l'opposition fondamentale entre sciences de *l'exemplum* et sciences du *datum*, telle que la rappelle Laks (2008) en linguistique, et particulièrement en phonologie, où il défend le travail systématique à partir de corpus.

2.2 Le paradigme corpus et son application en apprentissage des langues

Dans l'article (Reffay *et al*., 2008), nous avions introduit la notion de corpus telle qu'elle est perçue dans différents domaines des sciences du langage précurseurs en la matière, à savoir le traitement automatique des langues, la linguistique textuelle et les interactions orales. Le terme "corpus" étant utilisé dans des sens très différents, nous avions pris pour point de départ une définition de Bommier-Pincemin datant de 1999. Nous aurions pu tout aussi bien partir de celle, convergente, de Rastier (2005 : 32) donnée à l'occasion d'une conférence sur "la linguistique de corpus"[2].

[2] Si les références ci-dessus sont francophones, il convient de rappeler que le domaine de la linguistique de corpus a été ouvert et largement exploré par les collègues aux États-Unis, avec par exemple les concepteurs du *Brown Corpus* (Francis & Kucera, 1964) , et en Grande-Bretagne par des personnes comme Biber (1993) ou Sinclair (1987), qui avait pour objectif l'élaboration du premier dictionnaire pour apprenant (*learner dictionary*) en langue étrangère.

Nous préférons aujourd'hui définir cette notion sous forme de paradigme, au sens anglo-saxon, à savoir un modèle explicatif de l'objet considéré[3]. Le paradigme corpus comporte ainsi les quatre points indissociables suivants :

- **Recueil systématique des documents** liés à l'objet d'étude. La couverture et la taille sont alors des indicateurs de la systématicité du recueil. Les documents ne se limitent bien sûr pas à des textes, sauf si on comprend ce mot dans un sens élargi comme le font Halliday (1989) ou Baldry et Thibault (2006), à savoir des documents multimodaux (sur le caractère multimodal des interactions en ligne, voir (Betbeder *et al*., 2008)). Dans le cas d'une situation d'apprentissage en ligne, le point de départ pour la constitution du corpus concerne le recueil des productions des participants, tout comme le contexte d'élaboration et de déroulement de la formation (cf. point suivant). Ainsi se constitue un "objet d'étude", à condition que la mise en forme du recueil respecte les contraintes abordées au point 3 ci-dessous. Les études éventuellement accomplies sur cette situation s'ajouteront ensuite en couches périphériques.
- **Description du contexte**. La nature des objets décrits dans le contexte dépend bien entendu de l'objet d'étude. Dans un corpus d'interactions verbales, la description fournit des éléments d'information sur la situation et les participants. Lorsque l'objet est une situation d'apprentissage provoquée par les chercheurs, comme c'est très souvent le cas en AL&SIC, alors le contexte englobe en plus, ce qui a motivé l'expérience, les éléments liés à son élaboration dans ses aspects technologiques et pédagogiques, à son observation, voire ceux en rapport avec le protocole de recherche le cas échéant (questionnaires, entretiens, journaux de bords, etc.). Le contexte est fortement étoffé de façon à ce que la recherche puisse ensuite étudier le rapport entre le provoqué et le joué. Bien sûr cette description, contenue dans le corpus, n'est pas limitative de ce qui peut être invoqué comme contexte lors des analyses ultérieures. A cette description détaillée du contexte, viennent s'ajouter sous une forme synthétique les métadonnées décrivant en termes précis les caractéristiques de l'œuvre, ses acteurs (collecteurs, contributeurs, etc.) suivant les recommandations de standards, comme celui de OLAC (2008) (*Open Language Archives Community).*
- **Organisation et instrumentalisation en vue de traitements**. Un corpus s'élabore en vue d'analyses multiples. Même si certaines comportent des phases manuelles, elles sont toujours assistées par des outils, à défaut d'être entièrement automatiques. Les travaux de la communauté CATCOD (2008) sur les corpus oraux montrent bien qu'une équipe de recherche aura fréquemment recours à plusieurs outils dans un même projet. Se pose alors la question de la transmission des données produites d'un outil à l'autre. Pour y répondre, le projet européen SACODEYL (2008), producteur de corpus oraux à des fins pédagogiques sur le parler des adolescents de différents pays, a même constitué une chaîne de traitement complète dont les outils peuvent être déployés du laboratoire à la salle de classe. Par ailleurs, tout le monde s'accorde sur le fait qu'un corpus s'organise en vue de permettre aussi bien des analyses qualitatives que quantitatives, comme le soulignent O'Keefe *et al*. (2007 : 2) dans leur ouvrage sur les corpus de langues à destination pédagogique. On accordera alors une grande

[3] Cette acception que l'on relève dans les dictionnaires de langue anglaise se rapproche de celle de Kuhn (1962/1983 ; chap2 et 4) en ce sens que l'on ne peut omettre ou remettre en cause l'un des constituants sans changer complètement l'étude du champ considéré. McEnery & Wilson (1996 : 21) définit lui aussi la notion de corpus à travers ce qu'il intitule des "main headings" : "sampling and representativeness, finite size, machine-readable form, a standard reference".

importance au fait que tous les documents recueillis soient numérisés dans des formats ouverts, adaptables à différents outils, que les données provenant de ces documents d'origine ou des transcriptions ultérieures soient organisées dans des langages de balisage, ouverts aux traitement, comme XML, et soient structurées suivant des schémas standard comme la TEI (*Text Encoding Initiative*), ou, à défaut, des schémas accessibles à tous (voir (Reffay *et al*., *ibid*) pour les références techniques).

- **Dispositions en vue de l'échange et du partage**. Pour qu'une démarche scientifique puisse se dérouler avec ses phases d'analyses multiples, réanalyses, et discussions contradictoires, il est nécessaire que les auteurs du corpus l'organisent en vue de l'échange et du partage. Cela implique le fait d'ouvrir le corpus, ou plutôt la banque de corpus, en accès libre sur un serveur indexable par les autres serveurs de la Toile. Des protocoles standards existent pour ce faire, comme celui des archives ouvertes (Chanier, 2004), repris par OLAC (*ibid*). A cet accès ouvert s'ajoutent les règles du jeu de l'utilisation du corpus qui s'expriment sous forme de licence. Puisqu'il est question de données recueillies sur des humains, voire d'expérimentations les concernant, le corpus intégrera, en outre, les éléments permettant d'apprécier le respect de l'éthique (Oates, 2006) lors de l'expérimentation, des phases de collecte et de pré-traitement (anonymisation). Le site qui héberge le corpus vérifiera que le dépôt s'est effectué dans le respect des droits (Baude *et al*., 2005 : chap 2).

Dans sa définition de la notion de corpus, Rastier (*ibid*) écartait de son objet d'étude les corpus de mots, d'attestations ou d'exemples, tout comme les corpus de fragments. De même, le paradigme corpus décliné ici permet de différencier un corpus d'une base de données. Ainsi une base constituée de documents authentiques numérisés ne constitue pas à nos yeux un corpus, au contraire d'une partie des chercheurs communiquant au colloque "Des documents authentiques aux corpus oraux : questions d'apprentissage en didactique des langues" (CRAPEL, 2007).

On pourrait peut-être craindre que considérer simultanément un tel ensemble de points comme définitoire d'un corpus engendre trop de complexité, que cela décourage les équipes de chercheurs à l'idée de se lancer dans une telle entreprise. Or, paradoxalement en apprentissage des langues, ce sont les chercheurs qui ne s'inscrivent pas dans ce paradigme, mais adoptent au contraire une vision simplifiée, qui doutent de la possibilité d'en constituer :

> *Les programmes de formation linguistique destinés aux étudiants allophones intégrant l'université française s'intéressent […] à la compréhension orale des discours enseignants. Ils tentent pour cela de s'appuyer sur l'enregistrement de cours magistraux ou de travaux dirigés. Mais la transformation de ces enregistrements en supports pédagogiques est loin d'être aisée. Ce type de discours résiste aux habitudes établies par la didactique du FLE en matière d'enseignement de l'oral, […], leur utilisation dans la classe pose encore beaucoup de questions. Ce qui explique la quasi inexistence des discours académiques dans la panoplie des documents authentiques en FLE.* (Parpette, 2007)

Si maintenant on se tourne du côté des chercheurs qui comprennent l'objet corpus sous cette forme paradigmatique, on trouvera, outre le projet SACODEYL déjà cité, la banque de corpus MICASE (2009) ... qui a précisément réussi sur l'anglais académique, ce qui était jugé inatteignable pour une partie du milieu du français langue étrangère. Les équipes européennes et étasunienne qui ont développé ces deux banques de corpus, selon les quatre axes méthodologiques évoqués précédemment, ont créé des objets à double finalité.

Les contenus peuvent servir à des analyses sur la langue (celle des jeunes abordant des thèmes familiers dans sept langues différentes, celle de l'anglais académique), ou bien être utilisés en enseignement des langues. Les travaux exposés dans les différentes conférences et publications de TALC (*Teaching and Language Corpora*) montrent même quels types d'activités sont propices à l'apprentissage de la langue à travers l'étude de ses structures et composants (O'Keefe *et al*., *ibid*).

Les créateurs de ces banques de corpus ont également développé tout un environnement d'apprentissage autour de leurs sites Internet. On y trouve des outils de recherche et de concordance couplés aux contenus augmentés (documents primaires audio et vidéo, transcriptions, annotations linguistiques), ainsi que des guides d'utilisation pour l'apprenant et l'enseignant, voire ,dans SACODEYL, des activités interactives d'apprentissage construites à partir de sélections de matériaux. Enfin, les outils et contenus étant librement accessibles (et conforment aux formats standard), tout utilisateur, apprenant ou enseignant, peut les récupérer et les introduire dans son environnement de travail personnel. On pourrait alors parler de "granularités" différentes dans les corpus : l'objet complexe présenté dans le site Internet et l'ensemble de données extraites par l'utilisateur à des fins de manipulations individuelles. C'est cette même granularité fine qui se trouve dans les objets dénommés "*teacher corpus*" par O'Keefe *et al*. (ibid : chap11) . Ils sont constitués par l'enseignant et basés sur une sélection de matériaux opérée dans les sites précédents ou sur la cueillette des productions de ses propres apprenants (Rézeau, 2007).

Enfin, outre les finalités d'études et d'enseignement de la langue, ces banques de corpus permettent de relancer la recherche en apprentissage des langues dans le contexte des nouveaux environnements offerts par ces sites Internet. Par exemple, Pérez-Llantada (2009) a monté une expérimentation dans laquelle, s'appuyant sur MICASE, elle mesure l'impact de l'utilisation de tels corpus sur l'apprentissage de diverses compétences en L2, en compréhension et en production, lorsque ses apprenants les utilisent suivant des scénarios pédagogiques spécifiques.

2.3 Inscrire un nouveau genre de corpus dans ce paradigme

L'exposé précédent montrait l'intérêt pour un chercheur d'inscrire sa démarche dans le cadre du paradigme corpus en prenant exemple sur les travaux réalisés en apprentissage des langues. Le domaine d'étude que nous abordons dans cet article est celui des apprentissages impliquant des interactions en ligne. Notons que même si les situations évoquées ci-dessous se rapportent à l'apprentissage des langues, nous nous intéressons à toute forme d'apprentissage en ligne, quelle que soit la discipline. L'enjeu est donc de faire entrer dans le paradigme corpus le "genre" des interactions en situations d'apprentissage en ligne avec des études portant sur les différents types de "discours" correspondants (pour reprendre la terminologie de Rastier (*ibid* : 34)).

Pour ce faire, nous allons introduire le genre de corpus, dénommé *corpus d'apprentissage* (*LEarning & TEaching Corpus*), ou LETEC de façon plus concise. Nos objectifs sont doubles, à l'image de ceux évoqués dans la partie précédente : créer un objet pouvant servir à des fins pédagogiques en e-formation ou à des fins de recherche sur ces situations particulières d'apprentissage. Notre démarche se place à l'intersection des études en sciences du langage, en EIAH, et en apprentissage collaboratif (CSCL).

3 Vue d'ensemble de la notion de corpus d'apprentissage dans Mulce

Dans cette section, nous brossons une vue d'ensemble du corpus d'apprentissage tel que développé au sein du projet Mulce, abordant les aspects constituants et structuration. Le lecteur recherchant une information plus détaillée, en particulier sur les modèles de structuration, se reportera à (Reffay *et al*, 2008), (Reffay et Betbeder, 2009) ou au site de documentation Mulce (2010a), les corpus d'apprentissage étant accessibles sur la plateforme Mulce (2010b).

Donnons une première définition :

> *Un corpus d'apprentissage (LETEC) 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.*

3.1 Constituants du corpus

La partie gauche de la figure 1 schématise les différentes parties du corpus.

- Le dispositif pédagogique peut-être librement décrit, mais il est préférable de le faire de façon détaillée en précisant le **scénario pédagogique**, les différents rôles des participants, en particulier des apprenants et enseignants, ainsi que les environnements technologiques retenus avec leurs fonctionnalités, et leurs caractéristiques dédiées aux interactions.
- De la même façon, si l'expérimentation inclut un **protocole de recherche**, le rôle des chercheurs, le séquencement des activités afférentes (administration de questionnaires, entretiens, etc.) seront utilement décrits.
- Les deux parties précédentes correspondent à ce qui était prévu avant le déroulement de la formation, c'est-à-dire à un modèle. Suivant la terminologie des langages objets, le modèle "s'instancie" lors de l'acte pédagogique (avec tous les changements inopinés afférents). La partie **instanciation** assemble donc, d'une part, les enregistrements des interactions des participants (sous forme textuelle, audio ou vidéo), leur productions individuelles (tels que les travaux écrits ou oraux, les journaux de bord) et, le cas échéant, les traces système (temps de connexion, statistiques de participation, etc.). Elle regroupe d'autre part, le cas échéant, les questionnaires remplis, les enregistrements d'entretiens, les matériaux afférents (grille d'entretien, matériaux pour auto-confrontation, etc.).
- La partie publique de la **licence** donne accès aux licences d'utilisation du corpus par la communauté de chercheurs et de praticiens (Mulce a choisi une licence *Creative Commons*) et les formulaires de contrat d'éthique remis aux participants. La partie privée de la licence n'est pas directement intégrée au corpus, mais conservée par le responsable du corpus. Elle incorpore notamment les patronymes et coordonnées des participants, ainsi que les contrats signés.
- Les **analyses** ne font pas en général partie du corpus d'apprentissage, mais seront adjointes ultérieurement sous la forme de corpus distinguables (voir la section suivante). Quant au cas intermédiaire des transcriptions des enregistrements vidéo ou audio, nous les avons intégrées dans les corpus déposés dans la plateforme Mulce, en sachant qu'elles peuvent être recommencées ou modifiées. Le schéma de la figure 1 (partie de gauche) fait apparaître la partie analyse comme l'objectif orientant l'ensemble de l'effort de collecte et d'organisation.

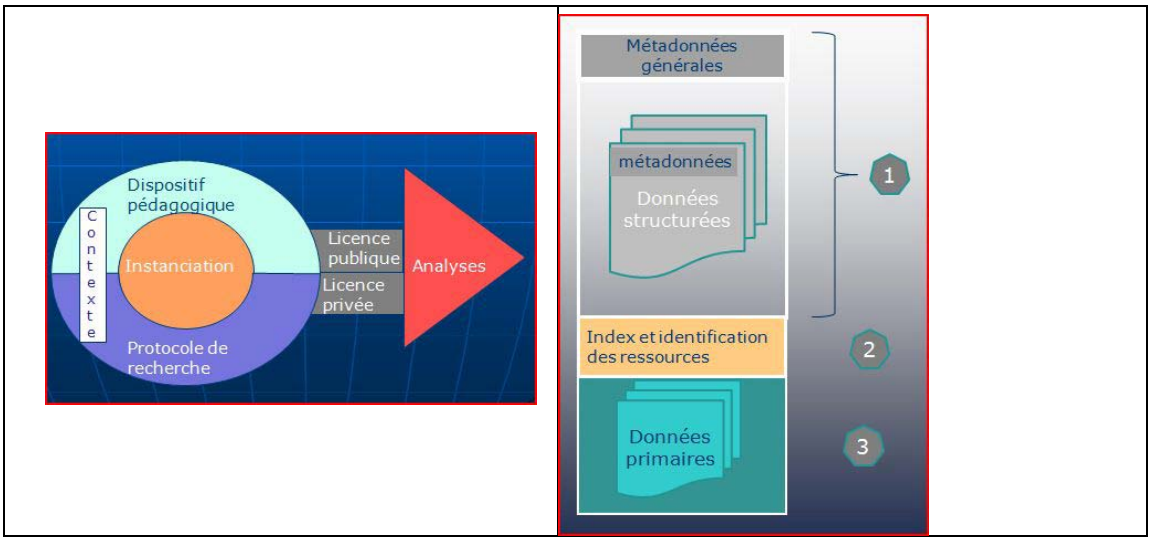

Figure 1 - Les grandes parties d'un corpus d'apprentissage LETEC (partie gauche de la figure) et sa structuration (partie droite).

3.2 Structuration du corpus

La partie droite de la figure 1 indique en trois points comment les données et les informations du LETEC s'organisent. En 3, sont rassemblées les données primaires des différentes parties (consignes pédagogiques, traces des interactions, questionnaires et entretiens, etc.), dans les formats nécessaires à leur conservation et à leur utilisation. Le qualificatif "primaire" adjoint à ces données, suivant la terminologie courante des corpus en linguistique, est quelque peu abusif ici, puisqu'une partie des formats des documents, (textes, vidéogrammes, audiogrammes) est transformée et les documents eux-mêmes anonymés.

En 2, sont regroupées les index, les identifiants et les informations résumées sur chacune des ressources de la partie 3. Cette partie est structurée, laissant ainsi apparaître les groupes de ressources correspondant, par exemple, à un ensemble de consignes pédagogiques, un ensemble de données pour des entretiens semi-directifs, ou un ensemble de fichiers permettant d'aligner transcriptions et vidéo (voir des exemples dans la section suivante).

La partie 1 est entièrement structurée, avec le langage de balisage XML, suivant un ensemble de schémas. Elle contient un premier ensemble de **métadonnées** générales du corpus, au format OLAC, identifiant la plateforme Mulce et ses contenus à l'intention des internautes et des serveurs moissonneurs de la Toile. Apparaissent ensuite, les informations correspondant à chaque constituant du LETEC. Pour ce qui concerne les **interactions**, nous avons mis au point une structure, correspondant au schéma *Mulce-struct* (Reffay *et al*., 2008 : s.3.3.2), dans laquelle sont encodés de façon homogène les messages de courriels, de forums, de blogues, les modalités des environnements audio-graphiques (audio, clavardage, iconique, actes de production dans des tableaux blancs, traitement de texte partagé, carte conceptuelle). Le contenu des actes, résultats des transcriptions (Chanier et Vetter, 2006) ou des traces système, est en correspondance avec les environnements technologiques et les participants. Il est ainsi possible pour un acte donné d'accéder à la description de **l'environnement technologique** dans lequel il a été produit, de même qu'aux caractéristiques de son auteur. Ces **informations ethnographiques** couplées à des éléments

de bibliographie langagière, essentiels pour les analyses linguistiques ultérieures (Belz et Vyatkina, 2008 : 45), sont décrites dans une autre partie de *Mulce-struct*.

De même, si les collecteurs et éditeurs du corpus ont choisi de structurer les informations concernant le **dispositif pédagogique**, alors chaque acte attaché à une interaction peut-être automatiquement mis en rapport avec le contexte pédagogique (activité, rôle des participants, consignes, etc.). Pour les premiers corpus d'apprentissage déposés dans le serveur Mulce (cf. tableau 1), nous avons choisi d'utiliser les modèles développés par la communauté *IMS Global Learning Consortium* (IMS, 2009). Ces modèles sont souvent utilisés dans les cursus de formation d'ingénieur en FOAD (formation ouverte ou à distance).

Signalons que les trois parties étiquetées dans la partie droite de la figure 1 sont enveloppées dans un container (*content packaging*) correspondant à l'un des formats prescrits par IMS. La partie 1 est dénommée le **manifeste**. C'est ce même format qui est souvent utilisé pour échanger des ressources pédagogiques entre les plateformes de téléformation.

Le schéma de LETEC exposé ici répond bien aux 4 aspects du paradigme corpus évoqués dans la section précédente. L'effort de structuration des données est d'abord motivé par l'objectif d'offrir au chercheur accédant au serveur un environnement de travail lui permettant d'effectuer des fouilles **intra** ou **inter corpus**, comme c'est le cas dans le serveur Clapi (2009), banque de corpus dans le champ des interactions verbales. Le second intérêt est de permettre la réutilisation aisée de ces données dans des logiciels d'aide à l'analyse, comme nous allons le développer plus avant. Reste la question du coût d'organisation et de structuration de tels corpus. Elle sera évoquée en fin d'article.

	Simuligne	Copéas
Objectif Pédagogiques	FLE en formation continue pour apprenants anglais	Anglais pour Master2 FOAD
Institutions	UFC, OU	OU, UFC
Participants	- 1 coordinateur - 10 natifs (UFC), - 40 apprenants (OU) - 4 tuteurs (OU). - 4 groupes de 12 - 1 groupe global (60)	- 14 apprenants (UFC) - 2 tuteurs (OU) - 2 groupes de 7+1
Environnements technologiques	Asynchrone (WebCT)	Synchrone (Lyceum) Asynchrone (WebCT)
Interactions Devoirs rendus	- 2686 mess. forum, - 4062 courriels - 5680 tours de clavardage - 93 doc. textuels, - une image - 28 fichiers audio	- 5506 tours de parole audio (8h29 en temps cumulé) - 1529 tours de clavardage - 16 séances Lyceum
Productions affichées	342 pages web incluant 115 images et 44 fichiers audio	Documents, cartes conceptuelles et tableaux blancs
Ressources pédagogiques	guide apprenant guide tuteur guide natifs	guide apprenant guide tuteur
Scénario	28 activités réparties en 7 étapes sur 12 semaines	8 activités sur 10 semaines
Questionnaires, Entretiens	12 questionnaires apprenants	- 14 questionnaires apprenants, - 7 entretiens, - 9 *Critical Event Recall* (8 app., 1 tuteur)
Taille	Total : 650 Mo : - 30 000 fichiers répartis dans 2708 dossiers	Total : 35,3 Go : - 37 vidéos (27h) - 512 autres fichiers dans 117 dossiers. - 180 000 lignes de traces et transcription dans Mulce-struct
Cession droits, contrat éthique	Cession droit (oui), contrat éthique (non)	Oui

Tableau 1 • Description synthétique des ensembles de données de deux corpus d'apprentissage déposés dans le serveur Mulce (2010b). UFC (Université de Franche-Comté), OU (Open University).

4 Analyses et corpus distinguables

Les corpus d'apprentissage ayant été définis précédemment, cette section illustre les utilisations que l'on peut en faire suivant des objectifs de recherche. La place nous manque pour aborder les utilisations pédagogiques de ces corpus. Signalons, toutefois, sans que cela recouvre l'éventail des possibles, que des scénarios pédagogiques et des protocoles de recherche ont été déposés dans la banque de corpus Mulce avec les données associées. Il sont utilisés dans la formation d'étudiants de niveau master.

Dans cette section, nous présentons, à partir d'extraits de nos corpus, des exemples d'analyses assistées à l'aide d'outils développés par des chercheurs investiguant les interactions en ligne. Pour ce faire, nous définissons, dans un premier temps, un type de

corpus de granularité inférieure à celle du corpus d'apprentissage, offrant une manipulation plus aisée au chercheur.

4.1 Des corpus de granularités différentes

Comme le montre les exemples du tableau 1, un corpus d'apprentissage correspondant à une expérience de formation est un méga corpus. Le rassemblement en une seule structure de toutes les interactions offre des possibilités d'analyses quantitatives globales telles que celles effectuées sur les 180 000 lignes de code correspondant à la formation Copéas. Cela nous avait notamment permis de déterminer la durée moyenne des actes de parole, suivant les modalités audio et clavardage, par acteur dans l'environnement audio-graphique *Lyceum* (Vetter et Chanier, 2006). Mais les chercheurs s'intéressent souvent à des phénomènes plus limités. Évoluer au milieu d'un ensemble trop important de données peut alors devenir un obstacle.

Afin de travailler sur des unités intermédiaires, nous avons constitué des **corpus distinguables** (Reffay *et al.*, 2008 : s 2.6). A partir d'un corpus d'apprentissage, que l'on qualifiera de **corpus global**, il est possible de produire des corpus distinguables, chacun correspondant au grain habituellement retenu par un chercheur pour y accomplir une analyse sur un phénomène précis, comme nous l'illustrerons dans les exemples des parties suivantes. Il peut se construire en segmentant le corpus global de deux façons très différentes : soit en ne retenant qu'une sous-partie du corpus global, soit en sélectionnant de façon longitudinale un ensemble de données (par exemple, tous les forums d'un groupe dans une formation, ou de tous les groupes).

Le corpus distinguable est tout à la fois un sous-corpus du corpus d'apprentissage et un corpus en soi. Son container est de même format que celui d'un corpus global (cf. figure 1, partie de droite). Mais à la différence de ce dernier, il est facilement téléchargeable sur un ordinateur personnel. Le chercheur dispose alors d'un ensemble comportant une description structurée du corpus contextualisé par rapport au corpus global (sous forme de commentaires libres et d'index précis renvoyant sur chacune des sous-parties du corpus global), et d'un ensemble de données soit prêtes à l'analyse, soit contenant déjà des résultats d'analyse. Enfin, le corpus global s'insère dans un réseau inter-corpus. Il contient, par définition un ensemble de liens vers le corpus global (comme nous l'avons dit), mais aussi souvent des liens vers d'autres corpus distinguables issus du même corpus global. Ces liens sont des invitations à mener des analyses inter-corpus.

Les corpus distinguables constitués dans Mulce répondent à trois objectifs distincts :

- associer publication scientifique et données (type 1) ;
- rassembler des données prêtes à l'analyse avec la mise en forme pour des outils/logiciels libres (type 2) ;
- partager des analyses avec des outils associés (type 3).

Dans le reste de cette section, nous avons fait le choix d'évoquer succinctement chacun des trois types de corpus distinguables, de façon à monter l'éventail des possibilités. Le lecteur désireux d'en savoir plus, se reportera à la plateforme Mulce (2010b). Il y trouvera les corpus présentés dans l'article, en les choisissant dans la partie "sélection des objets" de

l'interface de consultation, ou en utilisant l'identifiant du corpus cité ci-dessous dans la partie "ressources identifiées" de cette même interface[4].

4.2 Associer publication scientifique et données (type 1)

Lorsque, dans les différentes revues ou colloques organisés par les disciplines travaillant dans le domaine des interactions en ligne, un auteur soumet un article mettant en mots les résultats de sa recherche construite à partir de données, le comité scientifique ne peut accéder aux sources, ni donc de vérifier la qualité du traitement de l'auteur. De même, les lecteurs de l'article publié n'ont aucun moyen systématique d'obtenir toutes les informations ayant suscité ladite publication. Ils ne peuvent ni refaire des analyses sur ces données, ni répliquer l'expérience.

Conscients de ses limitations, nous avions dès 2005 commencé à associer aux dépôts de certaines de nos publications dans les archives ouvertes des fichiers de données. Aujourd'hui grâce à la notion de corpus distinguable, nous avons pu reprendre les publications associées aux deux corpus globaux Simuligne et Copéas (cf. tableau 1) et construire les corpus en rapport. Prenons l'exemple de l'article (Reffay et Chanier, 2003). Nous y avions exposé des modèles de structure des groupes d'apprentissage en ligne, appliquant pour ce faire des résultats mathématiques provenant des réseaux sociaux. Les données provenaient des interactions, courriels et forums, de la formation Simuligne. Nous avions été souvent sollicités par des lecteurs désireux de connaître plus précisément notre démarche. Le corpus distinguable dont l'identifiant est *mce-simu-sna-all* répond précisément à cette demande (voir (Reffay, 2009) pour lire directement sa fiche descriptive). Nous y avons rassemblé et structuré l'ensemble des informations et données spécifiques de Simuligne, ainsi que les résultats, sans oublier les références aux outils de la communauté des réseaux sociaux (SNA) nous ayant servi à faire les calculs.

Un corpus distinguable peut également offrir des perspectives différentes ou complémentaires de celles développées dans un article. Ainsi, notre collègue, T. Lewis, qui était intervenu en qualité de tuteur, enseignant d'anglais de spécialité, lors de la formation Copéas, a publié un article de nature réflexive sur le rôle qu'il y avait joué (Lewis, 2006). Nous avons, avec lui, repris l'ensemble des données à partir desquelles il avait construit son analyse et les avons complétées par des données, toutes extraites du corpus global de la formation, indiquant le point de vue des apprenants. La confrontation de l'article et du corpus distinguable laisse apparaître une appréciation fort différente des apprenants (en l'occurrence bien plus positive) concernant la qualité des processus collaboratifs de celle décrite par le tuteur-chercheur dans l'article (identifiant *mce-copeas-reflexive-tutor-all* ou (Lewis, 2009) pour la fiche descriptive).

Ce type de support (le corpus distinguable), associé à un processus de confrontation croisée avec des tiers est peut-être un moyen de pratiquer plus en profondeur la réflexivité, comme cela se fait déjà en psychologie du travail, que cela soit pour les apprenants ou les enseignants (Chanier et Cartier, 2006). On notera que le corpus distinguable vient conforter

[4] Avant de télécharger le corpus, nous recommandons à l'utilisateur de lire la fiche de description détaillée dudit corpus. Elle résume dans un langage non technique le contenu du manifeste. Ces fiches ont également été extraites des corpus et mis en ligne séparément pour un accès plus direct. Dans cet article, les références indiquées à côté des identifiants y renvoient.

ici une *recherche de type purement qualitative*, ce qui n'exclut donc pas une confrontation avec les données.

Si l'association de publications d'articles et de données n'en est encore qu'à ses prémisses dans notre domaine, elle devient toutefois systématique dans un nombre croissant de champs disciplinaires. Elle est obligatoire depuis des années en médecine expérimentale, où les articles sont reliés par un hypertexte aux données, en biologie pour le décryptage du génome humain, où le préalable à la soumission en vue de publication est le dépôt de la nouvelle séquence dans des banques de données mutualisées. Par ailleurs, elle connaît un essor récent dans les sciences sociales qui développe un nouveau paradigme de publication des travaux scientifiques, autour de la notion **d'ensemble de données pour la réplication** :

> ***Replication data sets*** *include the original data and any other information needed to reproduce the numerical results in a published work. […] making publicly available a replication data set for each of their empirical articles or books. Citation credit should be apportioned both for the original article and separately for the data.* (Gary, 2007 : 145)

Afin de mettre ces propos en application, cette communauté de chercheurs en sciences sociales a développé le réseau *Dataverse* (2009) qui relie les archives de dépôts de données de recherche. Ce réseau s'est doté d'outils libres pour la gestion de ces données, ainsi que d'outils pour les revues désirant changer leur processus de soumission en associant article et données. Ce milieu scientifique offre ainsi une réponse effective à un point fondamental pour tous les chercheurs travaillant sur des corpus, à savoir la reconnaissance scientifique et la valorisation des carrières des chercheurs opérant en sciences du *datum*.

4.3 Rassembler des données prêtes à l'analyse avec mise en forme pour outils/logiciels libres (type 2)

Un corpus d'apprentissage est un objet complexe pouvant donner lieu à de nombreuses analyses sur tout ou partie de ses constituants. Les collecteurs et éditeurs du corpus n'ont en général exploré qu'un nombre limité de possibilités. De plus, des pistes d'investigations plus riches encore peuvent s'ouvrir à des chercheurs, non impliqués dans la création du corpus, mais disposant de données provenant d'autres formations. Les créateurs du corpus global, parce qu'ils en connaissent bien sa structure et ses contenus, sont en mesure d'accomplir le travail préalable à ces recherches ultérieures. Il convient pour cela d'identifier des objets dignes d'intérêts pour les chercheurs du domaine, d'en extraire les données correspondantes, de les documenter et les contextualiser par rapport au corpus global, et enfin de transformer les formats de ces données pour les mettre en correspondance avec ceux des logiciels d'analyse.

Un tel processus d'extraction, de documentation et de mise en forme donne lieu, à partir d'un corpus global, à la création de corpus distinguables de type 2. Nous illustrons ce point par l'exemple des forums de discussion. Cette modalité de communication est au cœur de nombreuses études en sciences du langage, EIAH ou CSCL. Des communautés se sont naturellement constituées comme celle de Calico (2009), où enseignants des écoles et chercheurs en EIAH se coordonnent pour, tout à la fois, spécifier les fonctionnalités des outils d'analyse, les développer et les rendre disponibles à la communauté sur des serveurs qui jouent également la fonction de banque de forums (Bruillard, 2008). Une partie de ces outils permettent de visualiser et quantifier les contributeurs des forums, les fils de discussion, l'organisation temporelle des échanges. D'autres outils permettent de créer des

mini-lexiques décrivant des thématiques et d'observer la reprise de ces thématiques dans les messages et les fils de discussion.

Partant du corpus global de Simuligne, nous avons donc extrait l'ensemble des forums, issus de la simulation globale en langue et de l'activité *Interculture*. Les cent forums correspondants ont été transformés par des traitements automatiques du format *Mulce-struct* au format *XML-forum* de Calico, regroupés dans un corpus distinguable (identifiant *mce-simu-forum-all* ou (Chanier, 2009)), documentés avec les scénarios pédagogiques et les consignes d'utilisation des outils Calico. Ces forums sont tout à la fois disponibles sur le site de Calico (*ibid*), donc librement accessibles et prêts pour les outils de traitement du site, et dans le corpus distinguable, prêts pour une analyse inter-forum avec, par exemple, des concordanceurs ou d'autres outils de traitement du langage.

En dehors des forums, l'ensemble des séances synchrones du corpus d'apprentissage Copéas ont été transformées en corpus distinguables à des fins d'analyses sur la multimodalité. La section suivante détaillera ce point.

4.4 Partager des analyses avec des outils associés (type 3)

Afin de pouvoir analyser les interactions et de tenter de comprendre les phénomènes qui se sont déroulés durant une formation en ligne, il est recommandé de travailler sur "l'organisation, la modélisation et la conceptualisation des traces d'activité, de leur représentation et de leur traitement" (Settouti *et al*, 2006). Un premier obstacle surgit lorsque le chercheur désire trouver un mode de représentation optimale des données de façon à rendre saillante l'interaction humaine en vue de son analyse. Étant donné le volume important de données (cf. tableau 1) et la complexité de ces interactions, les requêtes présentées "à plat", comme dans une base de données, mettent peu en évidence les phénomènes potentiellement intéressants à analyser.

Malgré un intérêt croissant ces dernières années pour la recherche sur les interactions multimodales, notamment dans le domaine des EIAH (Dyke *et al*., 2007 ; Harrer *et al*., 2007), il existe encore peu de logiciels pour l'alignement et la représentation des données. La plupart des logiciels proposent des formats propriétaires de structuration des données, ce qui rend difficile le partage ultérieur des données ou des résultats. En outre, pour réduire la complexité des phénomènes étudiés, les chercheurs ont souvent besoin de pouvoir coupler une première série d'analyses à celle d'un autre niveau, toute chose peu aisée avec ces logiciels.

Le logiciel Tatiana (2008), utilisé dans le projet Mulce, a pour objectif d'aider les chercheurs dans l'analyse des traces d'interactions en ligne. Il permet de synchroniser les interactions multimodales codées sous forme textuelle avec les vidéogrammes où apparaissent leurs manifestations visuelles et auditives, de les annoter et de les catégoriser. Le logiciel favorise une démarche **itérative** de l'analyse et la création multiple d'artefacts, étape préalable aux analyses (Lund et Milles, 2009). Les résultats sont ensuite aisément partageables et exportables (Dyke *et al*., 2008). Nous illustrons cette démarche avec deux exemples provenant de Copéas.

Nous avons repéré deux phénomènes récurrents dans les interactions multimodales issues du corpus Copéas : d'une part, les phénomènes de polyfocalisation (Jones, 2004 ; Ciekanski et Chanier, 2008), dans lesquels apparaissent des négociations sur le contexte, et, d'autre part, les phénomènes de production langagière mettant en jeu plusieurs modalités.

Trois corpus distinguables ont été constitués autour de ces phénomènes. Ils contiennent les vidéogrammes extraits des sessions étudiées, leurs transcriptions adaptées du format *Mulce-struct* vers celui de Tatiana[5], ainsi que les analyses développées dans ce logiciel, le tout étant documenté, comme le montre la lecture des manifestes ou fiches descriptives correspondant aux corpus.

Négociation du contexte en environnement multimodal

Le premier corpus distinguable analyse ces phénomènes de négociation de contexte au cours de la situation d'apprentissage (identifiant *mce-copeas-T5_contexte-all* ou (Ciekanski et Chanier, 2009a)). Il porte sur un extrait du travail d'un sous-groupe de trois apprenants appartenant au groupe des faux-débutants en anglais. Ces apprenants travaillent seuls dans une salle de l'environnement audio-graphique *Lyceum*. Chacun est à distance et intervient depuis son ordinateur personnel à son domicile. Lors de cet épisode, les apprenants évaluent les éléments de réponse à la première question d'un quizz, réponses qu'ils viennent d'élaborer collectivement dans la partie traitement de texte partagé du logiciel de conférence. Au bout de quelques minutes, le tuteur revient dans la salle où travaille le sous-groupe et intervient dans le clavardage, afin d'apporter une aide linguistique aux apprenants qui s'interrogent oralement sur la signification d'une expression. Ses interventions sont toutefois ignorées des apprenants qui poursuivent leur évaluation de leurs réponses aux questions. Après plusieurs tentatives d'intervention dans le clavardage, qui n'attirent pas plus que la précédente l'attention des apprenants, le tuteur quitte la salle.

La visualisation de cet épisode obtenue à partir de Tatiana fait ressortir la répartition des modalités entre participants (apprenants et tuteur). Une série de filtres permet de visualiser chaque modalité sur une ligne différente et d'attribuer des couleurs différentes à chaque participant, à l'image de ce qui apparait dans la fenêtre supérieure de la figure 2. Il en ressort que les apprenants utilisent exclusivement les modalités audio et traitement de texte, alors que le tuteur intervient exclusivement dans le clavardage. La fenêtre *Lyceum*, soit l'espace de travail des participants, apparaît dans le coin inférieur droit de la figure 2. On peut y voir que le traitement de texte partagé y occupe l'espace principal. Le clavardage se devine dans le cadre inférieur de cette fenêtre et il est donc visible de tous les participants. Pourtant, les interactions qui ont lieu dans cette modalité ne sont pas lues par les apprenants.

[5] De même chaque corpus distinguables de type 2, produit à partir d'une séance Copéas, intègre vidéogramme et transcription complète mis au format Tatiana.

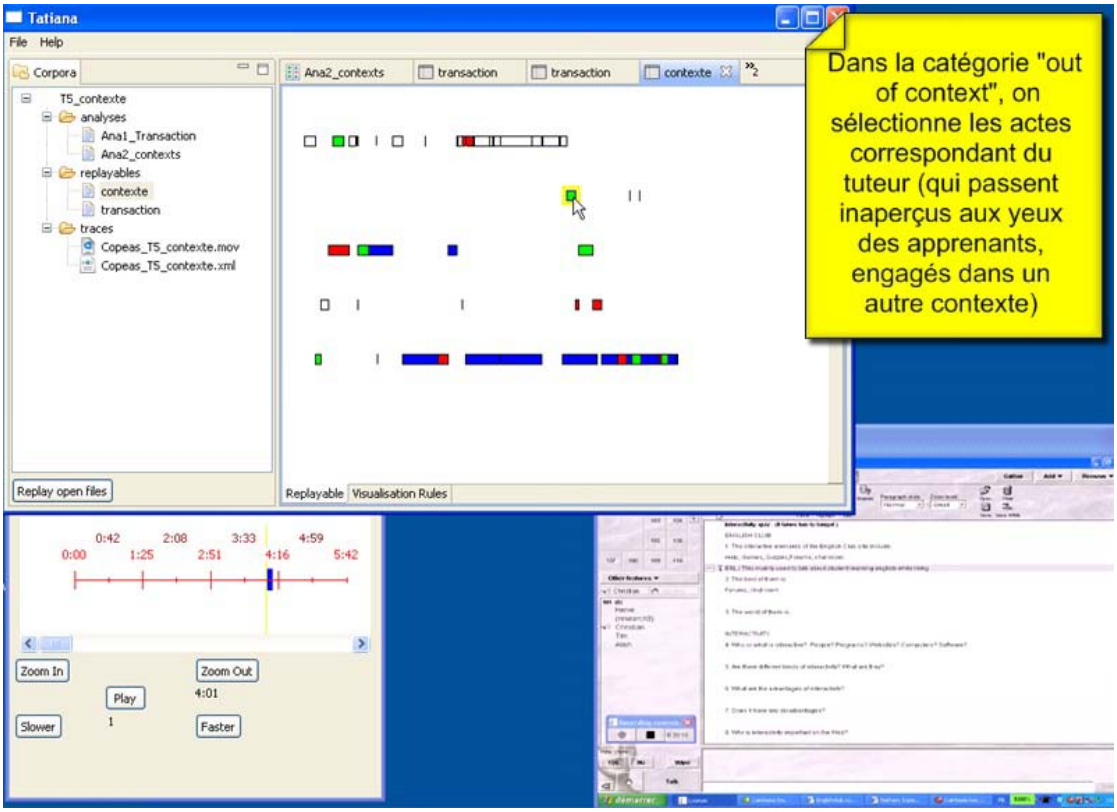

Figure 2- Interface de Tatiana montrant la synchronisation d'une visualisation chronologique des interactions filmées dans Lyceum (fenêtre en bas à droite de l'écran) avec le rejouable obtenu à partir des règles de visualisation du processus de catégorisation étudié dans l'extrait (en haut à gauche de l'écran). La fenêtre en bas à gauche affiche l'axe temporel de la vidéo et les boutons de commande pour lancer la synchronisation entre les deux autres fenêtres.

Ainsi, loin de concevoir le contexte d'interaction comme étant une donnée figée par l'organisation spatiale de la plateforme Lyceum, les apprenants semblent négocier leur espace de communication et de travail au fil de leurs échanges. Pour comprendre la façon dont se construit le contexte lors de cet épisode, nous avons catégorisé à l'aide de Tatiana la transcription multimodale en cinq catégories, rendant compte du processus de contextualisation propre à l'épisode analysé. Les interventions du tuteur apparaissent alors catégorisées comme "hors contexte" (cf. figure 2, deuxième ligne de la fenêtre supérieure). Dans la mesure où le tuteur va et vient entre différentes salles de la plateforme audio-graphique synchrone, il joue un rôle marginal dans la participation au contexte de travail en tant que tel. Il n'a pas négocié l'utilisation du clavardage avec les apprenants et n'a donc pas été inclus dans le contexte partagé.

Stratégies d'écriture multimodales

Les deux autres corpus distinguables constitués se rapportent à une même tâche rédactionnelle réalisée en parallèle par deux sous-groupes (A et B) dans deux salles distinctes de Lyceum (identifiants *mce-copeas-T8_s101_ecriture_multimodale-all* et *mce-copeas-T8_s102_ecriture_multimodale-all* ou (Ciekanski et Chanier, 2009b ; 2009c).

La visualisation de chacun des deux épisodes révèle de grandes différences dans la manière dont les deux sous-groupes recourent à la multimodalité. La focalisation sur les

acteurs met en exergue le fait que l'utilisation de la multimodalité dépend de stratégies individuelles qui doivent être négociées au sein du groupe si l'on vise une collaboration efficace. L'utilisation de la multimodalité est également contextuelle dans la mesure où certains acteurs modifient leur "densité modale" (Norris, 2004) par rapport aux précédentes sessions. La compréhension de l'usage de la multimodalité pour produire un texte en langue étrangère nécessite ensuite une analyse plus fine du processus rédactionnel. Pour ce faire, nous avons eu recours aux étapes du processus rédactionnel tel que le décrit Zimmerman (2000) : formulation, transcription, correction, reformulation et révision. Nous avons catégorisé dans un premier temps la transcription multimodale et recoupé ces informations avec l'utilisation que font les apprenants de la multimodalité lors de leur rédaction.

Pour le sous-groupe A, nous notons que la coexistence de deux modalités pour l'écrit (le clavardage et le traitement de texte) permet de répondre à des fonctions distinctes, selon les apprenants. Ainsi, l'apprenant AT6 semble privilégier le clavardage au traitement de texte pour participer à la formulation et la correction des réponses, sans doute car ce dernier revêt un caractère moins permanent que le texte rédigé directement dans l'outil traitement de texte. En effet, la production dans le clavardage ponctue les interactions qui se déroulent dans les autres modalités (audio et traitement de texte) et n'est pas au centre du discours. Cela permet une prise de risque moins importante par rapport au fait d'écrire directement dans le traitement de texte.

Pour le sous-groupe B on notera que, à la différence des apprenants du sous-groupe A, les apprenants n'utilisent pas le clavardage. Le processus rédactionnel se réalise uniquement dans les modalités audio et texte. En outre, dans la mesure où le sous-groupe B fait un usage inhabituel de la langue maternelle lors de cet épisode, nous avons catégorisé, dans un deuxième temps, l'usage de la langue (L1 et L2) à partir de la première catégorisation sur le processus rédactionnel décrit ci-avant. Cet étayage en L1 concerne principalement l'apprenant AT5, dans les phases de formulation et de révision, cette formulation en L1 étant ensuite prise en charge par un autre apprenant qui reformule et transcrit en L2 dans le traitement de texte.

Tous les phénomènes que nous venons d'évoquer apparaissent clairement dans les visualisations et catégorisations que nous avons réalisées dans Tatiana. Le lecteur intéressé pourra télécharger les corpus correspondants, suivre les procédures d'installation expliquées par des vidéos et observer directement ces résultats.

Perspectives de recherche sur la multimodalité

Le travail sur les interactions multimodales en ligne montre qu'une compréhension de ses phénomènes ne peut se faire que dans un va-et-vient entre plusieurs niveaux de description (*intra-corpus*) et plusieurs corpus (*inter-corpus*) (section 3.2). Les différents niveaux de description requièrent une palette d'outils variés, en fonction des phénomènes étudiés. En effet, l'analyse de processus de groupe demande par exemple de "mesurer" les chevauchements des actes, de calibrer si les échanges représentent des interactions plus ou moins fortes, etc. Ces questions de recherche nécessitent de disposer d'outils traitant des données temporelles ainsi que des outils pour la reconnaissance de forme : reconnaissance de motifs d'actions (de patrons), reconnaissance de contenus d'échanges parlant d'un même "thème", comme le permettraient des outils de traitement automatique du langage (TAL). Par ailleurs, la compréhension des phénomènes étudiés ne peut être fine que si elle confronte plusieurs corpus de même type (en faisant varier les niveaux, les tâches, les

modalités de collaboration, etc.) et de types différents (en faisant varier les environnements par exemple). C'est au prix d'une réflexion sur les formalismes et sur l'intérêt de tous au partage des données de recherche que l'étude des interactions en ligne en situation d'apprentissage permettra de réelles avancées sur le plan théorique et, partant, sur le plan des dispositifs d'apprentissage.

5 Conclusion

La notion de corpus d'apprentissage (LETEC) a été définie au regard des habitudes de recherche du domaine AL&SIC et des développements récents dans les différentes communautés s'intéressant à la recherche sur corpus. Le paradigme corpus comporte les quatre points indissociables suivants : 1) le recueil systématique des documents liés à l'objet d'étude ; 2) la description du contexte ; 3) l'organisation et l'instrumentalisation en vue de traitements ; 4) les dispositions à prendre en vue de l'échange et du partage. Un corpus d'apprentissage en ligne assemble donc de façon systématique et structurée un ensemble de données 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, le tout organisé pour permettre des analyses contextualisées. Ces principes ont été mis en œuvre dans le projet Mulce, dont la plateforme offre un accès libre à la banque de corpus et le site Mulce.org à la documentation méthodologique correspondante.

En phase de constitution de corpus, puis lors de l'accès aux données, les dimensions éthiques et juridiques jouent un rôle essentiel. Elles visent à protéger les acteurs et spécifient les contraintes sur les usages des corpus, conditions sans lesquelles les données ne seraient pas partageables. Ainsi les corpus issus des formations présentées dans cet article ont été systématiquement anonymés. Quant aux licences d'utilisation, elles sont stipulées lors des accès à la banque de corpus.

Depuis 2008, la structure *Mulce-Struct* a été affinée pour rendre possible diverses fouilles, recherches, étiquetages sur les interactions verbales (textuelles ou audio), ainsi que l'alignement entre les données audio ou vidéo et leurs transcriptions. Chaque corpus d'apprentissage inclut un manifeste où sont rassemblées toutes les interactions, les descriptions du scénario pédagogique, des acteurs et environnements. Toutes ces informations sont organisées suivant le schéma *Mulce-struct*. De cette structure nous avons montré comment produire des corpus distinguables, de tailles plus aisément manipulables par le chercheur. Chaque corpus distinguable intègre également une description structurée contextualisée par rapport au corpus global, couplée à un ensemble de données prêtes à l'analyse, voire à des résultats d'analyse. Ces données ont été formatées afin d'être directement traitables dans des outils d'analyse développés par la communauté de recherche sur les interactions en ligne. Enfin, des liens relient un corpus distinguable à son corpus global et, le cas échant, à d'autres corpus distinguables pour des analyses inter-corpus. Actuellement, Mulce propose une trentaine de corpus distinguables prêts à l'analyse, relevant de trois types à finalité différentes illustrés dans cet article.

Les efforts de standardisation convainquent de plus en plus de chercheurs dans nos communautés (AL&SIC, EIAH, CSCL), particulièrement ceux désireux de partager leurs données et de confronter leurs analyses. Notre proposition de structuration offre la possibilité de travailler à des niveaux différents d'analyse, à partir d'un ensemble élargi

d'interactions en ligne, associant interactions synchrones et asynchrones, ce qui est encore peu fréquent dans les recherches actuelles. Reste la question des coûts d'organisation et de structuration de tels corpus. Ils ne deviendront acceptables que si la perspective de partage devient réalité. Les opportunités peuvent provenir de nouveaux projets de recherche incluant des équipes travaillant dans des perspectives interdisciplinaires sur les interactions en ligne. Alors pourrait se développer une nouvelle étape dans laquelle des ensembles de données recueillies par d'autres chercheurs seront adaptés au format LETEC. C'est alors seulement qu'il sera possible de mesurer le coût du travail nécessaire à cette transformation et restructuration.

Notre exposé visait cependant à montrer que le coût d'appropriation d'une telle méthodologie est compensé par les perspectives de gain de temps lorsqu'un chercheur désire, à partir d'un même ensemble de données, mener des analyses dans différents outils, chacun requérant un formatage spécifique. Un autre gain, perceptible dès aujourd'hui, est celui de la reconnaissance du travail du chercheur par ses autorités de tutelle. En effet, la mise à disposition d'un corpus d'apprentissage est un acte clairement identifiable et appréciable au niveau d'une communauté de chercheurs. Le travail correspondant peut alors être présenté comme une production que les directions de recherche savent maintenant prendre en compte dans les procédures d'évaluation des travaux scientifiques, suivant les critères européens ou nationaux. Cette évaluation concerne aussi bien l'individu chercheur que son équipe.

Le projet Mulce s'achèvera dans sa première phase fin 2010. De nombreuses pistes de développements restent ouvertes. Elles concernent les formalismes et les outils permettant d'étendre l'éventail de l'analyse des interactions en ligne. Concernant les formalismes, on peut par exemple discuter la pertinence de la conversion au format TEI des données multimodales. Concernant les outils de communication, une extension de *Mulce-struct* est en cours afin d'y intégrer les blogues et, par là-même de nouveaux corpus d'apprentissage. Mais ces pistes seront réellement confortées lorsque la communauté des chercheurs investira la plateforme Mulce en déposant ses propres données et/ou en confrontant ses analyses avec celles d'autres chercheurs. Tel est l'un des objectifs du projet Mulce pour les mois à venir.

6 Références

Tous les liens Internet de cette section ont été vérifiés en date du 21 avril 2010.

Audras I. & Chanier T. (2008). "Observation de la construction d'une compétence interculturelle dans des groupes exolingues en ligne", *Apprentissage des Langues et Système d'Information et de Communication (Alsic),* vol. 11, n°1. pp. 175-204. http://alsic.revues.org/index865.html

Baldry A. & Thibault P. (2006). *Multimodal Transcription and Text Analysis, a multimedia toolkit and coursebook with associated on-line course*, Equinox, Londres.

Basharina O. K. (2007). "An Activity Theory Perspective On Student-Reported Contradictions In International Telecollaboration", *Language Learning & Technology*, vol. 11, n° 2. pp. 104-127. http://llt.msu.edu/vol11num2/basharina/

Baude O., Blanche-Benveniste B., Calas M.F., Cordereix P., De Lamberterie I., Gourie L., Jacobson M., Marchello-Nizia C. & Mondada L. (dir.) (2005). *Guide des bonnes pratiques pour la constitution, exploitation, conservation et diffusion des corpus oraux*, Éditions du CNRS et DGLF-LF, Paris, http://www.culture.gouv.fr/culture/dglf/Guide_Corpus_Oraux_2005.pdf

Belz J.A. & Vyatkina N.(2008). "The Pedagogical Mediation Of A Developmental Learner Corpus For Classroom-Based Language Instruction», *Language Learning & Technology*, vol. 12, n° 3. pp. 33-52. http://llt.msu.edu/vol12num3/belzvyatkina/

Betbeder M.-L., Ciekanski M., Greffier F., Reffay C. & Chanier T. (2008) "Interactions multimodales synchrones issues de formations en ligne : problématiques, méthodologie et analyses», dans *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)*, vol. 15, J. Basque et C. Reffay (dir.), http://sticef.univ-lemans.fr/num/vol2008/06-betbeder/sticef_2008_betbeder_06.htm

Biber, D. (1993). "Representativeness in corpus design", *Litterary and Linguistic Computing*, vol. 8, n°4. pp. 243-257.

Bruillard E. (2008). "Teacher development, discussion lists and forums : issues and results". In Mcferrin, K., Weber, R., Carlsen, R., & Willis, D.A. (dir.). *Proceedings of Society for Information Technology and Teacher Education International Conference, SITE 2008*. Chesapeake, USA : AACE. p. 2950-2955.

Calico (2009). *Site où sont déposés des forums de discussion, avec les outils associés pour leurs analyses* [site Internet]. ERTé Calico. http://wims.crashdump.net/www/calico/

CATCOD (2008). *Site de la communauté "Catalogage et codage de corpus oraux"* [site Internet]. Réseau Risc/CNRS. http://www.catcod.org/

Chanier T. (2009). *Manifeste du corpus distinguable fournissant des données prêtes à l'analyse des forums de Simuligne* [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/mce-simu-forum.xml

Chanier, T. (2004). *Archives ouvertes et publication scientifique. Comment mettre en place l'accès libre aux résultats de la recherche ?* L'Harmattan. http://archivesic.ccsd.cnrs.fr/sic_00001103/fr/

Chanier, T. & Cartier, J (2006). "Communauté d'apprentissage et communauté de pratique en ligne : le processus réflexif dans la formation des formateurs", *Revue internationale des technologies en pédagogie universitaire (RITPU)*, vol. 3, n°3. pp. 64-82. http://www.profetic.org/revue/IMG/pdf/RITPU-Vol_3_3.pdf

Chanier, T.& Vetter. A.(2006). "Multimodalité et expression en langue étrangère dans une plate-forme audio-synchrone". *Apprentissage des langues et Système d'Information et de Communication (Alsic)*, vol. 9. pp 61-101. http://alsic.revues.org/index270.html

Ciekanski, M. & Chanier, T. (2009a). *Manifeste du corpus distinguable fournissant des données prêtes à l'analyse pour la notion de contexte dans Copéas* [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/copeas-T5_contexte.xml

Ciekanski, M.& Chanier, T. (2009b). *Manifeste du corpus distinguable fournissant des données prêtes à l'analyse pour la notion d'écriture multimodale dans Copéas* [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/copeas-T8_ecriture_multimodale_s101.xml

Ciekanski, M.& Chanier, T.(2009c). *Manifeste du corpus distinguable fournissant des données prêtes à l'analyse pour la notion d'écriture multimodale dans Copéas* [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/copeas-T8_ecriture_multimodale_s102.xml

Ciekanski, M.& Chanier, T (2008). "Developing online multimodal verbal communication to enhance the writing process in an audio-graphic conferencing environment". *Recall*, vol. 20, n° 2. pp. 162-182. doi:10.1017/S0958344008000426 http://edutice.archives-ouvertes.fr/edutice-00200851

Clapi (2009). Site *de la banque de corpus sur les interactions verbales* [site Internet]. Université Lyon 2 / Cnrs. http://clapi.univ-lyon2.fr

Crapel (2007). Site du colloque *"Des documents authentiques oraux aux corpus : questions d'apprentissage en didactique des langues",* décembre, Nancy [site Internet]. Nancy : ATILF. http://www.atilf.fr/atilf/evenement/Colloques/Crapel2007/crapel2007.htm

Dataverse (2009). *Site de la communauté Dataverse Network développant un réseaux de banques de données de recherche associées aux publications*. [site Internet]. Harvard University. http://thedata.org/

Dyke, G., Girardot, J-J., Lund, K. & Corbel, A. (2007). "Analysing face-to-face computer-mediated interactions". *EARLI'07*. Budapest, Hongrie.

Dyke, G. Lund, K. & Girardot, J-J. (2008) "Managing, synchronising, visualising, analysing and sharing multimodal computer-mediated human interaction data: introducing Tatiana (A Trace Analysis Tool for Interaction Analysts)", ICLS *2008 Workshop: A Common Framework for CSCL Interaction Analysis*, Utrecht, Pays Bas, 23-28 Juin.

Francis, W.N. & Kucera, H. (1964). *Brown Corpus: A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Brown University Providence, Rhode Island [corpus] http://www.hit.uib.no/icame/brown/bcm.html

Gary K. (2007). "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," Sociological Methods and Research, vol. 32, n0 2. pp. 173-199. http://gking.harvard.edu/files/dvn.pdf

Halliday, M.A.K. (1989). "Part A". In Halliday, M.A.K. & Hasan, R. (dirs.). *Language, Context, and Text : Aspects of language in a social-semiotic perspective.* Oxford University Press. pp. 55-79.

Harrer, A., Zeini, S., Kahrimanis, G., Avouris, N., Marcos, J-A., Martinez-Mones, A., Meier, A., Rummel, N. & Spada, H (2007). "Towards a flexible model for computer-based analysis and visualisation of collaborative learning activities". *Proceedings CSCL 2007*, 16-27 july 2007, New Jersey. USA.

IMS (2008). *Site du IMS Global Learning Consortium, Inc* [site Internet]. http://www.imsglobal.org/ .

Jepson, K., (2005). "Conversations-and negotiated interaction- in text and voice chat rooms". *Language Learning and Technology,* vol.9, n°3. pp. 79-98. http://llt.msu.edu/vol9num3/pdf/jepson.pdf

Jones, R. (2004). "The problem of context in computer-mediated communication". In Levine, P., Scollon, R. (dirs). *Discourse and technology: multimodal discourse analysis*. Georgetown University Press. pp. 20-33.

Kern, R. (2000). *Literacy and Language Teaching*. Oxford University Press.

Kern, R., Ware, P. & Warshauer, M. (2004). "Crossing frontiers: new directions in online pedagogy and research". *Annual Review of Applied Linguistics*, vol. 24. pp. 243-260.

Kramsch, C., Thorne, S. L. (2001). "Foreign language learning as global communicative practice" . In D. Block & D. Cameron (dir.), *Globalization and language teaching*, pp. 83-100. Routledge : Londres.

Kuhn, T.S. (1962/1983). *La structure des révolutions scientifiques* (traduction de *The structure of scientific revolutions*). Flammarion.

Lamy, M-N. (2007). "Multimodality in online language learning environments: looking for a methodology". In Baldry, A., Montagna, E. (dirs). *Interdisciplinary perspectives on multimodality : theory and practice*. Proceedings of the third international conference on multimodality. Campobasso: Palladino. pp. 237-254.

Laks, B. (2008). "Pour une phonologie de corpus". In *Le français à la lumière des corpus,* Durand, J. (dir.), numéro thématique de *Journal of French Language Studies*, vol. 18 , n°1. pp. 3-32.

Lewis, T. (2009). *Manifeste du corpus distinguable fournissant les données liées à l'article* (Lewis, 2006) [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/copeas-reflexive-tutor.xml

Lewis, T. (2006). "When Teaching is Learning: A Personal Account of Learning to Teach Online". *Calico*, vol. 23, n°3. pp 581-600. http://calico.org/html/article_110.pdf

Lund, K. & Mille, A. (2009). "Traces, traces d'interactions, traces d'apprentissages : définitions, modèles informatiques, structurations, traitements et usages". In Lund, K., Mille, A. (dirs). *Analyse de traces et personnalisation des environnements informatiques pour l'apprentissage humain.* Hermès. pp. 21-66.

McEnery T. & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh University Press.

MICASE (2009). *Site de la banque de corpus "Michigan Corpus of Academic Spoken English"* [site Internet]. The University of Michigan. http://quod.lib.umich.edu/m/micase/

Mulce (2010a). *Site de documentation du projet Multimodal Learning Corpus Exchange* (2007-2010) [site Internet]. http://mulce.org

Mulce (2010b). *Site de la banque de corpus Mulce* [site Internet]. Université de Franche-Comté. http://mulce-pf.univ-fcomte.fr/PlateFormeMulce/

Norris, S. (2004). "Multimodal discourse analysis: a conceptual framework". In Levine, P., Scollon, R. (dirs). *Discourse and technology: multimodal discourse analysis*. Georgetown University Press. pp. 101-115.

Oates, J. (2006). "Ethical frameworks for research with human participants". In Stephen Potter (dir) *Doing Postgraduate Research*. Londres : Sage. pp. 200-228.

O'keefe, A. Mccarthy, M. & Carter, R. (2007). *From corpus to classroom. Language use and language teaching*. Cambridge University Press.

OLAC (2008). "Best Practice Recommendations for Language Resource Description". In *Site de l'Open Language Archives Community*. University of Pennsylvania. http://www.language-archives.org/REC/bpr.html

Parpette, C. (2007). "Les discours universitaires oraux : questions de recherche, questions d'enseignement en FLE". *Colloque (Crapel, 2007).* Résumé en ligne à http://www.atilf.fr/atilf/evenement/Colloques/Crapel2007/Resume_PARPETTE.pdf

Pérez-Llantada, C (2009). "Textual, Genre and Social Features of Spoken Grammar: A Corpus-Based Approach", *Language Learning & Technology*, vol. 13, n°1. pp. 40-58. http://llt.msu.edu/vol13num1/perezllantada.pdf

Rastier F. (2005). "Enjeux épistémologiques de la linguistique de corpus". In Williams, G. (dir.) *La linguistique de corpus*. Presses Universitaires de Grenoble. pp. 31-46.

Reffay C. (2009). *Manifeste du corpus distinguable fournissant les données liées à l'article* (Reffay & Chanier, 2003) [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/mce-simu-sna.xml

Reffay, C. & Betbeder, M-L. (2009). "Sharing corpora and tools to improve interaction analysis. *EC-TEL 2009*, *Fourth European Conference on Technology Enhanced Learning*, Nice, septembre-octobre. http://edutice.archives-ouvertes.fr/edutice-00399841/fr/

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),* vol. 15, http://sticef.univ-lemans.fr/num/vol2008/01-reffay/sticef_2008_reffay_01p.pdf

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

SACODEYL (2008). *Chaîne de traitements développée par le projet Sacodeyl*, corpus oraux d'adolescents européens composé à des fins pédagogiques [site Internet]. Espagne : Universidad de Murcia http://www.um.es/sacodeyl/en/pages/software.htm

Settouti, L-S., Prié, Y., Mille, A. & Marty, J-C. (2006). "Systèmes à base de traces pour l'apprentissage humain". *Actes de TICE 2006*. Toulouse, octobre.

Sinclair, J. (dir.) (1987). *Collins COBUILD English Language Dictionary* (1ère édition). Londres : Collins.

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

Thorne, S. L. (2003). "Artifacts and cultures-of-use in intercultural communication", *Language Learning & Technology*, vol. 7, n°2. pp.38-67. . http://llt.msu.edu/vol7num2/thorne/

Vetter, A. & Chanier, T. (2006). "Supporting oral production for professional purpose, in synchronous communication with heterogeneous learners". *ReCALL*, vol.18, n°1. pp 5-23. http://edutice.archives-ouvertes.fr/edutice-00080316 doi:10.1017/S0958344006000218

Wang, Y. (2004). "Internet-based desktop videoconferencing in supporting synchronous distance language learning". *Language Learning and Technology*, vol. 8, n°3. pp. 90-121. http://llt.msu.edu/vol8num3/pdf/wang.pdf

Zimmerman, R. (2000). "L2 Writing: subprocesses, a model of formulating and empirical findings". *Learning and Instruction*, vol. 10. pp. 73-99.

7 Remerciements

Mulce est un projet soutenu par l'ANR Corpus et Outils en SHS (ANR-06-CORP-006). Il rassemble des membres des laboratoires LRL (Université Blaise Pascal), LIFC (Université de Franche-Comté) et CREET (The Open University), coordonnées respectivement par Thierry Chanier, Christophe Reffay et Marie-Noëlle Lamy, auxquels s'ajoutent Marie-Laure Betbeder et Maud Ciekanski. Nous remercions F. Tajariol pour sa participation à la transformation des données du format Mulce au format Tatiana.

****

Thierry Chanier est professeur des universités. Ses domaines d'enseignement et de recherche portent sur l'apprentissage des langues et les systèmes d'information et de communication, sur l'ingénierie de formation. Il étudie plus particulièrement les systèmes de formation à distance et les interactions en ligne sur des sujets tels que l'interculturel, le processus réflexif dans la formation des enseignants, le dialogue dans les environnements multimodaux. Il s'intéresse également à la structuration et aux modalités d'échanges entre chercheurs des corpus d'apprentissage.

Maud Ciekanski est maître de conférences à l'Université Vincennes-Saint-Denis Paris 8. Ses travaux relèvent de la didactique des langues et de l'analyse des interactions en situation d'apprentissage, dans les dispositifs d'autoformation et à distance. Depuis 2006, elle s'intéresse aux environnements d'apprentissage multimodaux et aux pratiques qui en découlent.

thierry.chanier@univ-bpclermont.fr , Université Blaise Pascal

http://lrlweb.univ-bpclermont.fr/spip.php?rubrique98

LRL , Maison des Sciences de l'Homme
4 rue Ledru - 63057 Clermont-Ferrand Cedex 1

maud.ciekanski@univ-paris8.fr, Université Paris 8

UFR SEPF, Département Com/FLE
2 rue de la Liberté- 93526 Saint-Denis Cedex.