Glasgow
WorldCALL
2013

Global perspectives on Computer-Assisted Language Learning

Glasgow, 10-13 July 2013

LEarning and TEaching Corpora (LETEC): data-sharing and repository for research on multimodal interactions

Ciara R. Wigham
Clermont Université
Clermont-Ferrand, France

Thierry Chanier
Clermont Université
Clermont-Ferrand, France

The number of online environments language teachers can employ is constantly growing, offering increased potential for multimodal L2 interaction analysis. This paper introduces the LEarning and TEaching Corpora (LETEC) methodology that links, following international standards, all elements resulting from an online learning situation, whose context is described by a pedagogical scenario and a research protocol. The corpus components include the learning design, the research protocol, the interaction data, all participants' productions and licences relating to ethics and access rights. An XML schema allows interactions from different tools and environments to be stored and described in a standardized way, facilitating data analysis.

We explore the stages for building LETEC; from the design of an online course to analysis phases and the diffusion of results, and describe the *Mulce* repository developed for sharing these corpora. We then focus on ways LETEC methodology may contribute to sustaining CALL research beyond the hype of the latest online environment. Firstly, the methodology's successive research phases allow for data to be examined from both research and pedagogical angles. Secondly, the decomposition of online environments by their communication modes and modalities offers a systematic approach to studying a range of online learning environments. Thirdly, the open-data that LETEC corpora generate are shared via a corpus repository and can be reused by researchers, not necessarily involved in the learning event, to perform cumulative or contrastive analyses. The paper concludes by discussing our current perspectives: the development of pedagogical corpora to train pre-service language teachers out of in-world situations, built upon multimodal materials from global LETEC corpora.

Keywords: learning and teaching corpus (LETEC); data sharing; research data repository; online multimodal interactions

1. Introduction

The number of online environments language teachers can employ is constantly growing, offering increased potential for L2 interaction analysis. However, research cannot necessarily keep up with technology innovation. One danger is that CALL research will reinvent the wheel each time a new technology emerges. Sharing research situations in formats that allow comparisons between interactions in different online environments will help CALL research to better understand L2 interaction across different multimodal environments.

This paper introduces the *LEarning and TEaching Corpora* (LETEC) methodology that links, following international standards, all elements resulting from an online learning situation. We introduce the methodology for building LETEC, contrasting it with that of *learner corpora*, and describe the *Mulce* repository (*Mulce*-repository, 2011) developed for corpora sharing. We then suggest ways in which LETEC contribute to sustaining CALL research.

2. LEarning and TEaching Corpora

In the language-learning domain, *learner corpora* (Granger, 2002; Meunier *et al.*, 2011) are exploited for Second Language Acquisition research. Frequently comprising data from test situations (Reffay *et al.*, 2008) and used in learner-native speaker comparative studies (Boulton *et al.*, 2012), *learner corpora* focus on learners' productions and consider neither other course participants (tutors, native speakers...) nor the learning context.

LETEC "collect in a systematic and structured way all the data from interactions which occur during a course that is partially or entirely online" (Chanier & Ciekanski, 2010:para59[1]). All course participations and their productions and interactions are considered. The pedagogical scenario forms an integrated component, to inform studies into learning environments' affordances or pedagogical design.

LETEC comprise a XML "manifest" describing the corpus' components: the learning design (pedagogical scenario), the research protocol, the interaction data and participants' productions (instantiation of the pedagogical scenario), and licences relating to ethics and access rights (Fig.1). The structure allows for subparts of components to be linked. For example, a researcher examining interaction data from a textchat session can understand, from the pedagogical scenario, the session's objectives and technical context and, by referring to the research protocol, the data collection and anonymization methodologies.

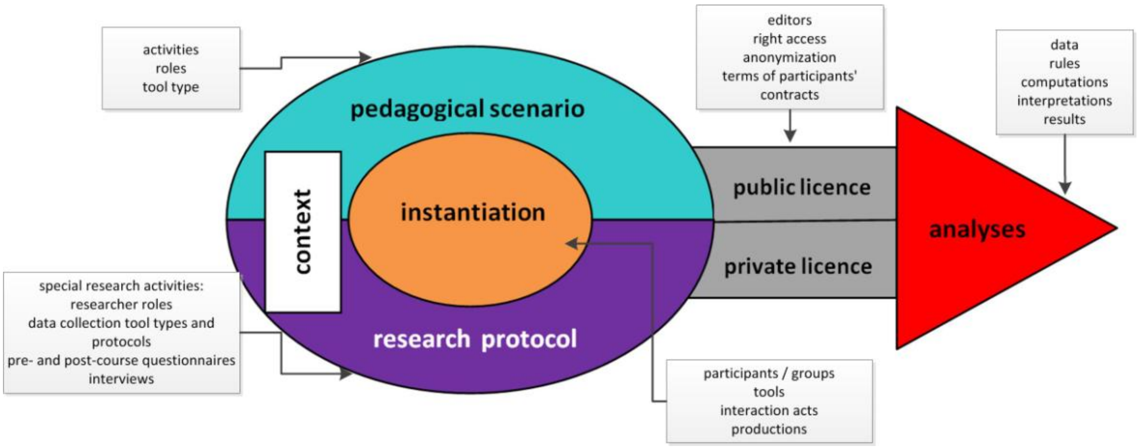


Fig1. LETEC components

The notion of LETEC was developed within a French national research project (*Mulce*-documentation, 2011) and conforms to criteria suggested by Chanier & Ciekanski (2010) if the term 'corpus' be employed (Fig2).

[1] Authors' translation.

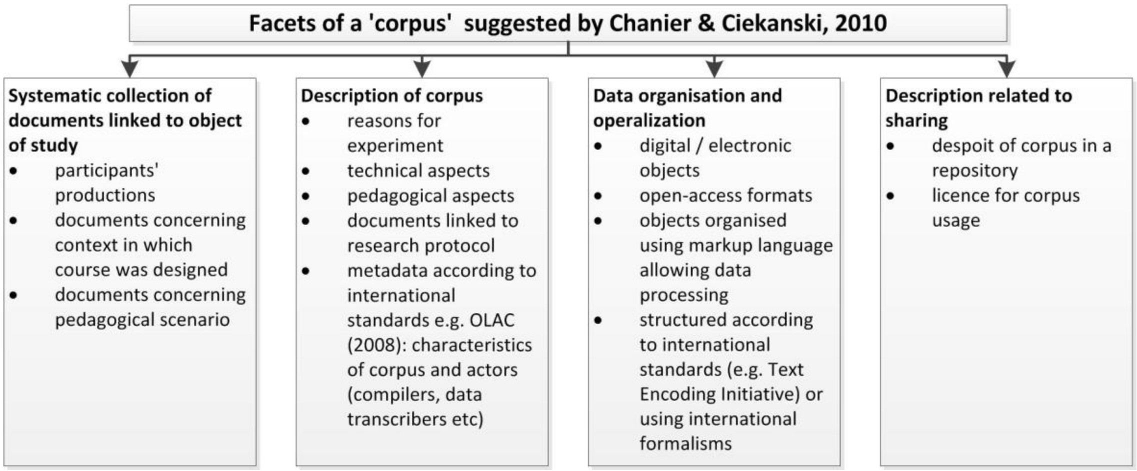

Facets of a 'corpus' suggested by Chanier & Ciekanski, 2010
Systematic collection of documents linked to object of study
• participants' productions
• documents concerning context in which course was designed
• documents concerning pedagogical scenario
Description of corpus
• reasons for experiment
• technical aspects
• pedagogical aspects
• documents linked to research protocol
• metadata according to international standards e.g. OLAC (2008): characteristics of corpus and actors (compilers, data transcribers etc)
Data organisation and operalization
• digital / electronic objects
• open-access formats
• objects organised using markup language allowing data processing
• structured according to international standards (e.g. Text Encoding Initiative) or using international formalisms
Description related to sharing
• despoit of corpus in a repository
• licence for corpus usage

Fig2. Facets of a corpus

3. LETEC contributions to sustaining CALL research

This section presents leads offered by LETEC for helping sustain CALL research.

3.1. Successive research phases

The LETEC approach to data collection, structuring and analysis is composed of successive phases (Fig3).

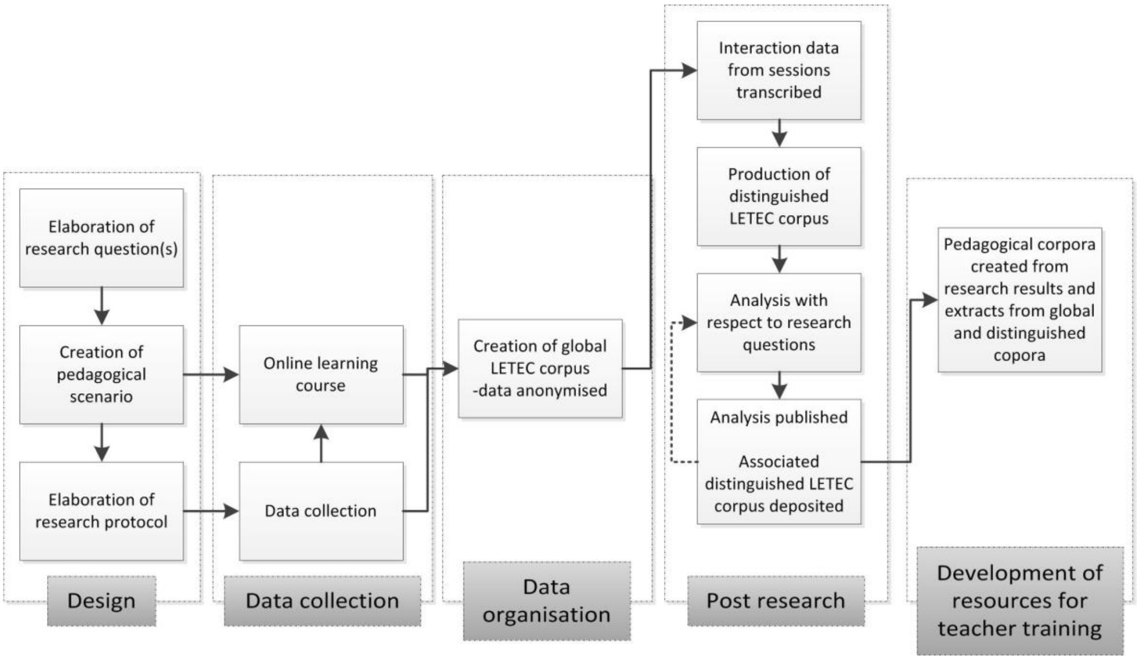
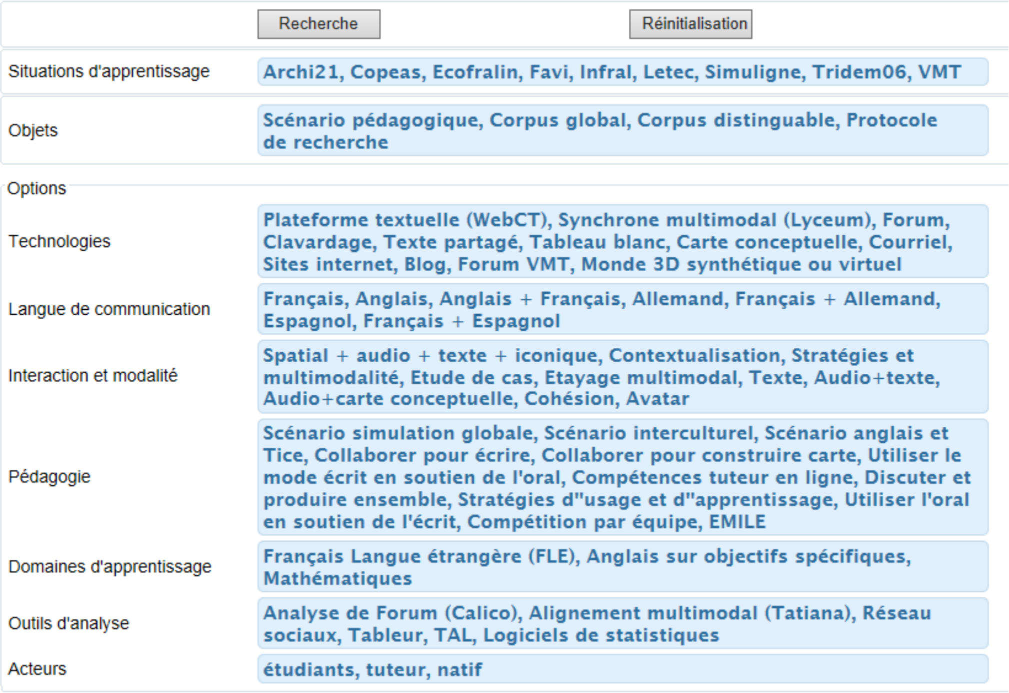

Elaboration of research question(s)
Creation of pedagogical scenario
Elaboration of research protocol
Design
Online learning course
Data collection
Data collection
Creation of global LETEC corpus -data anonymised
Data organisation
Interaction data from sessions transcribed
Production of distinguished LETEC corpus
Analysis with respect to research questions
Analysis published
Associated distinguished LETEC corpus deposited
Post research
Pedagogical corpora created from research results and extracts from global and distinguished copora
Development of resources for teacher training

Fig3. LETEC methodology phases

Prior to the site of experiment, research questions are formulated. A pedagogical scenario is elaborated and the conditions under which to conduct the experiment prepared. In parallel, a research protocol is designed around the experiment. During the online course, data is collected according to this. In the data organisation phase, a process of data display is performed on the pedagogical scenario and research protocol, and the data collected are anonymised before being structured into a *global* LETEC corpus and deposited on the *Mulce* repository (*Mulce*-repository, 2011). Structured using an XML schema, interactions from different tools and environments are described in a standardized way. They are described by metadata, allowing on the corpora repository, researchers to search for corpora according to different criteria (Fig4).



Situations d'apprentissage	Archi21, Copeas, Ecofralin, Favi, Infral, Letec, Simuligne, Tridem06, VMT
Objets	Scénario pédagogique, Corpus global, Corpus distinguable, Protocole de recherche
Technologies	Plateforme textuelle (WebCT), Synchrone multimodal (Lyceum), Forum, Clavardage, Texte partagé, Tableau blanc, Carte conceptuelle, Courriel, Sites internet, Blog, Forum VMT, Monde 3D synthétique ou virtuel
Langue de communication	Français, Anglais, Anglais + Français, Allemand, Français + Allemand, Espagnol, Français + Espagnol
Interaction et modalité	Spatial + audio + texte + iconique, Contextualisation, Stratégies et multimodalité, Etude de cas, Etayage multimodal, Texte, Audio+texte, Audio+carte conceptuelle, Cohésion, Avatar
Pédagogie	Scénario simulation globale, Scénario interculturel, Scénario anglais et Tice, Collaborer pour écrire, Collaborer pour construire carte, Utiliser le mode écrit en soutien de l'oral, Compétences tuteur en ligne, Discuter et produire ensemble, Stratégies d"usage et d"apprentissage, Utiliser l'oral en soutien de l'écrit, Compétition par équipe, EMILE
Domaines d'apprentissage	Français Langue étrangère (FLE), Anglais sur objectifs spécifiques, Mathématiques
Outils d'analyse	Analyse de Forum (Calico), Alignement multimodal (Tatiana), Réseau sociaux, Tableur, TAL, Logiciels de statistiques
Acteurs	étudiants, tuteur, natif

Fig4. *Mulce*-repository search criteria

Post research includes data transcription. In the *Mulce* LETEC methodology, transcriptions are characterized by communication modes and modalities. This allows a systematic approach to studying online environments. If new environments present new modalities these are added to the transcription methodology and *Mulce* metadata, rather than new methodologies being invented each time new technologies emerge. For example, Wigham & Chanier (2013), details how new nonverbal modalities present in the synthetic world *Second Life* were accounted for in relation to verbal modalities (audio and textchat) identified during earlier studies which used the audiographic conferencing environment *Lyceum* (Ciekanski & Chanier, 2008).

The interaction data's multimodal transcription leads to the production of a *distinguished* LETEC. A *distinguished corpus* includes a particular transformation of a selected part of the *global corpus*. For example, the transformation of a video file into an XML/text file of the transcribed interaction data and its associated metadata. Following transcription, data analyses are performed. The data transformations conducted during these (e.g. data annotation or coding) are structured into other *distinguished corpora*. These do not copy the structured data available in the *global corpus* upon which the post research was performed, but refer to the latter data and only add the transformed data for the specific analysis.

Distinguished corpora help sustain CALL research by giving value to the researcher's analyses. The analysed data can be presented in parallel with results and distinguished corpora can be cited in papers/articles. Explicit connexions and open access (Open Data, 2012) between data and publications enhance CALL research quality with possibilities offered to the research community for validity and reliability checks and reuse of data (see 3.2).

In a final phase, extracts of LETEC are currently being developed into teacher-training resources. *Pedagogical corpora* are based on a reflective approach to teacher training and offer the possibility to observe, examine and explore selected parts of a LETEC with reference to a lead that has been identified within the research analyses performed. Reusing research data in pedagogical contexts aids widen CALL research's applicability.

3.2. Reusability of Open Data

In online learning situations, the replication of the ecological context is practically impossible to obtain: "collaborative online learning situations have a number of variables difficult to control" (Reffay *et al.*, 2012). If the same learning design is reused with different participants, the observable phenomenon will not necessarily be the same. Structuring online interactions as LETEC allows researchers, not necessarily involved in the learning event, to be able to reuse the data for further cumulative or contrastive analyses.

For example, the *distinguished corpus* 'mce-archi21-modality-textchat' (Wigham, 2013) contains data from Content and Language Integrated Learning sessions in the synthetic world *Second Life*, annotated with respect to a study concerning textchat usage for feedback (Wigham & Chanier, 2012). In Rodrigues & Wigham (in print) this annotated data was reused: a further annotation layer of XML was added to study problematic vocabulary points' resolution. Structured corpora present advantages for research teams: Each researcher has his individual research questions but each researcher's analysis enriches the corpus. Analyses can thus be cumulated. This is facilitated by the fact that LETEC are structured using a set of documented structured XML formalisms (Reffay *et al.*, 2012) rendering online interaction data autonomous from any platform, in a tool agnostic form and thus increasing data longevity. For example, data format can be adapted to become input for annotation tools (lexicometric, multimodal tools (Ciekanski & Chanier, ibid)). Natural Language Processing techniques can also be applied to interactions and new annotation layers added (see examples in Mulce-Repository (2013) and explanations in Mulce-Documentation (2013).

LETEC open data are also being reused in projects related to fields other than that of language learning. The CoMéRé project (2013) aims to gather CMC corpora that will be integrated into the future reference corpus of French Language. Within a European context, CoMéRé will propose a TEI extension adapted for CMC communication.

4. Conclusions and Perspectives

After having gathered data from online language learning situations, which occurred between 2001 and 2011 (the Mulce-repository is still open for new deposits), defined the LETEC structure, and applied it to CALL research purposes, we now face the challenge of developing *pedagogical corpora*. By "developing" we not only mean defining their structure (i.e. ways of extracting data – video, audio, transcripts of interactions, views of the pedagogical scenario, linking them to tasks for trainee language teachers), but also integrating them into teacher-training classrooms. Our first pedagogical corpora will be online this summer. Finding a rational path for their integration into teacher training courses is another issue.

Language teacher trainers are used to training pre-service teachers in using software, websites, and learning material repositories (e.g. *MERLOT world languages* (1997)). When trainers want their students to gain skills in developing online learning situations based on interactive, multimodal environments, they generally have recourse to the reading of CALL literature disconnected from actual data and/or participation in experiments integrated into the academic year where their students act as learners or tutors. In the latter case, these pre-service teachers may fall into the trap of only considering their current individual experience. If the teacher trainers introduce a *reflection-in-practice* process around the online experiment to share experiences, it becomes difficult to manage: students' materials are often heterogeneous and quickly extracted from the on-going experiment. Carefully documented and selected materials put into their original context would be very helpful. This will come from pedagogical corpora.

Training pre-service teachers out of in-world situations, built upon multimodal materials (carefully analysed with respect to theoretical viewpoints) is not simply a concern of the language-learning field. There is extensive experience coming from teacher training in physical education (Roche & Gal-Petitfaux, 2012). Our current interest in pedagogical corpora is thus now becoming an inter-disciplinary project.

5. References

Boulton, A., Carter-Thomas S. & Rowley-Jolivet, E. (2012). (Eds.), *Corpus-informed Research and Teaching in ESP*. Amsterdam/Philadelphia: Benjamins Publishing Company.

Chanier, T. & Ciekanski, M. (2010). Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage : un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage. *ALSIC,* 13, [doi: 10.4000/alsic.1666].

Ciekanski, M., Chanier, T (2008). Developing online multimodal verbal communication to enhance the writing process in an audio-graphic conferencing environment. *ReCALL,*vol. 20 (2), 162-182. [doi:10.1017/S0958344008000426].

CoMéRé (2013) *CoMéRé (Communication médiée par les réseaux) project website* [http://corpuscomere.wordpress.com/apropos/].

Granger, S. (2002). A Bird's-eye View of Computer Learner Corpus Research. In S. Granger., J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3-33). Amsterdam/Philadelphia: Benjamins Publishing Company.

Mulce-Documentation (2013). *Website explaining the Mulce methodology and commenting scientific events around the project* [http://Mulce.org].

Mulce-Repository (2013). *Open Access Repository where LETEC Corpora may be downloaded*. *Mulce*.org: Clermont Université. [http://repository.Mulce.org].

Merlot World Languages (1997) *Merlot World Languages Portal* [http://worldlanguages.merlot.org/].

Meunier, F., De Cock, S., Gilquin, G. & Magali, P. (2011). (Eds.), *A Taste for Corpora. In Honour of Sylvaine Granger.* Amsterdam/Philidelphia: Benjamins Publishing Company.

Open Data (2012). *Definition of Open Data*. Open Knowledge Foundation. [http://opendefinition.org/okd/].

Reffay, C., Betbeder, M.-L. & 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.

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

Roche, L. & Gal-Petitfaux, N. (2013). La médiation audio-visuelle pour former à l’expeérience de l’enseignant d’EPS en situation de class. *STAPS* (98), 95-111.

Rodrigues, C. & Wigham, C.R. (in print). Les apports d’un monde synthétique pour l’apprentissage du vocabulaire en langue étrangère. *Recherches en didactique des langues et des cultures : les Cahiers de l’Acedle.* [http://edutice.archives-ouvertes.fr/edutice-00785802].

Wigham, C.R. (2013). (Ed.), *Distinguished Corpus : Interplay between textchat and audio modalities during the Second Life Reflective Sessions*. *Mulce*.org: Clermont Université. [oai : *Mulce*.org:mce-archi21-modality-textchat; http://repository.*Mulce*.org].

Wigham, C.R. & Chanier, T. (2012). Interactions between text chat and audio modalities for L2 communication in the synthetic world Second Life. In J. Colpaert, A. Aerts, W.-C. V. Wu, & Y.-C. J. Chao (Eds.). *Fifteenth International CALL Conference, The Medium Matters, Proceedings, 24-27 May 2012*. Taichung, Taiwan: Providence University. [http://hal.archives-ouvertes.fr/hal-00660865/].

Wigham, C.R. & Chanier, T. (2013). A study of verbal and nonverbal communication in Second Life – the ARCHI21 experience. *ReCALL*, 25(1), 63-84. [doi:10.1017/S0958344012000250].