



**HAL**  
open science

## A viewpoint on the place of CALL within the Digital Humanities: considering CALL journals, research data and the sharing of research results

Thierry Chanier

### ► To cite this version:

Thierry Chanier. A viewpoint on the place of CALL within the Digital Humanities: considering CALL journals, research data and the sharing of research results. EUROCALL 2013, Learning from the Past, Looking to the Future, Sep 2013, Evora, Portugal. edutice-00862024

**HAL Id: edutice-00862024**

**<https://edutice.hal.science/edutice-00862024>**

Submitted on 15 Sep 2013

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## A viewpoint on the place of CALL within the Digital Humanities: considering CALL journals, research data and the sharing of research results

Thierry Chanier, Université Blaise Pascal



Version 15th September 2013

*Download slides and all videos for this talk:  
link on <http://mulce.org>, main editorial article,*

Eurocall 2013, University of Évora , Portugal, 11-14 September, 2013



Download slides and videos, extended. version online

# Portugal & Clermont-Ferrand



Recent but strong relationships

Portugal and Clermont-Ferrand:  
Cultures and languages  
between the past and the future  
(3mn video)



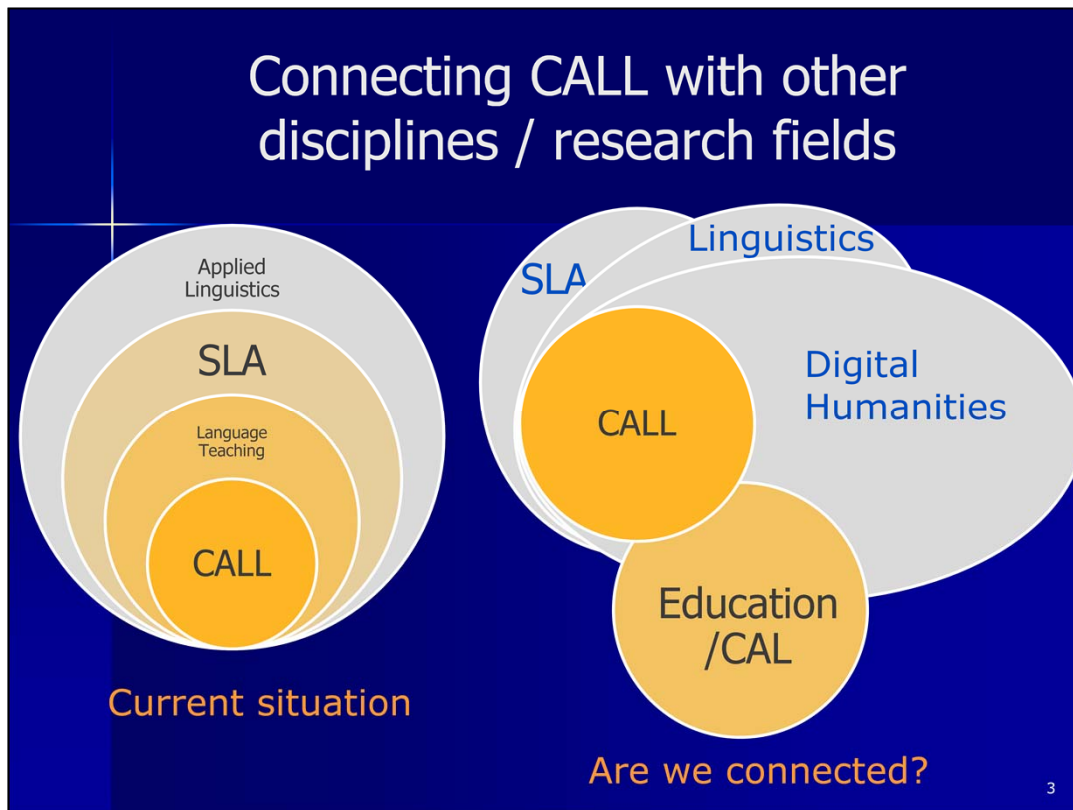
2

Caros colegas e amigos,

Não falo Português e nunca aprendi, mas antes de eu começar a minha apresentação sobre o tema referente as revistas e dados sobre as pesquisas, eu gostaria de fazer uma breve introdução sobre Portugal e minha cidade Clermont-Ferrand.

Esta cidade é a segunda maior cidade Portuguesa aqui na França, depois de Paris. Este pequeno vídeo mostra bem as relações que existem entre os nossos dois idiomas e culturas.

Obrigado pela vossa atenção.

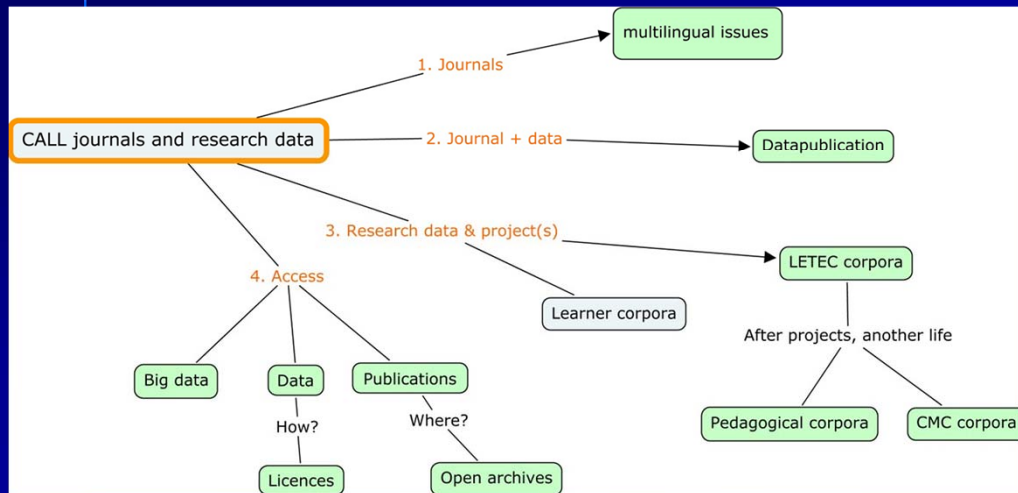


Generally now when you read CALL literature or attend CALL conferences, our domain is presented as being a subpart of fields such as Language teaching, Second Language Acquisition, not to say Applied Linguistics.

Without rejecting this standpoint, I would like here to turn our attention to other fields, more particularly to the so-called “Digital Humanities”.



# Overview



4

The term "Digital Humanities" (DH) made a buzz at the MLA (Modern Language Association) convention in 2009. The term is now in widespread use within the Humanities. CALL may be directly concerned: our field belongs to the Humanities and, from the outset, we have had a strong interest in computers and computing.

Although various meanings and interpretations can be attributed to this term, this presentation will address issues related to ways of promoting CALL research in order to meet what may soon become research standards within the Humanities.

We will talk about journals, data linked to publications, research data organized as corpora and of access to these publications and research data.

1

2

3

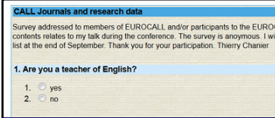
4



## **JOURNALS AND MULTILINGUAL ISSUES WITHIN THE CALL COMMUNITY**

5

First of all, let us start with CALL journals (and conferences) and (briefly) examine whether we have built a multilingual academic community



## Survey on CALL journals and research data

- Please participate in the online survey
- The survey is anonymous. I will publish the results on the EUROCALL mailing list at the end of September.
- Find the survey:
  - Link in the main editorial article on : <http://mulce.org>
  - Questions 1 to 5

6

About this first issue, as well as the other ones, I would be happy to collect, in an anonymous way (no possibility for me to know who answered what) you current standpoints.

If not already done before the conference you can access the survey. The link to the survey is available on the front page of the site [mulce.org](http://mulce.org).

Of course I will publish within one or 2 weeks results on Eurocall and CALICO memberlists.

Find the survey : <http://Mulce.org>

## History of ReCALL

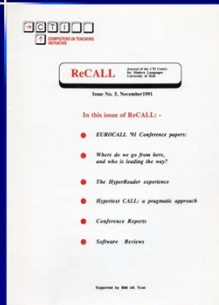


Since it is the 20<sup>th</sup> anniversary of Eurocall, let me first remind you that our journal ReCALL has got a long tradition. It first appeared in 89, and was at that time supported by the CTI centre for Modern Languages. At the end of the eighties, was launched a British national initiative to support the development of technology, CAL in education. The university of HULL, where Graham Chesters and June Thompson were working, was in charge of CALL development.

Besides publishing the Recall journal, Hull also provided a lot of services to language teachers. Teachers could visit the CTI and get software demos, or people from the CTI could come to their own institutions. The CTI also regularly published a software guide, and abroad we always were waiting for the last release.

# History of ReCALL

1989



1995?



2003?



## The EUROCALL Review

ISSN: 1695-2618

- Vol. 21 No. 1, March 2013
- Vol. 20 No. 2, September 2012
- Vol. 20 No. 1, March 2012
- No. 19, September 2011
- No. 18, March 2011
- No. 17, September 2010
- No. 16, March 2010
- No. 15, Mar-Sep 2009 special issue
- No. 14, Nov 2008
- No. 13, Mar 2008
- No. 12, Sep 2007
- No. 11, Mar 2007
- No. 10, Sep 2006
- No. 9, Mar 2006
- No. 8, Nov 2005
- No. 7, May 2005
- No. 6, Dec 2004
- No. 5, Aug 2004
- No. 4, Jan 2004
- No. 3, Sep 2003
- No. 2, Mar 2003
- No. 1, Nov 2002



June Thompson  
- there from  
the very  
beginning

Ana Gimeno  
(ed)

8

In 95, ReCALL became a joint publication of the CTI and the association Eurocall. In 2003, Eurocall and Hull who still had the editing responsibility gave the publishing task to Cambridge University Press. Of course I would not like to omit to cite the Eurocall review, edited par Ana Gimeno, review still directly published by Eurocall.

Since the beginning of Recall there has been one person who still is in charge of ReCALL, namely June Thompson. Unfortunately for the first time , June could not attend our conference, but we could here send her a real thanks for all the wonderful work she achieved !

# History of ReCALL

1995?

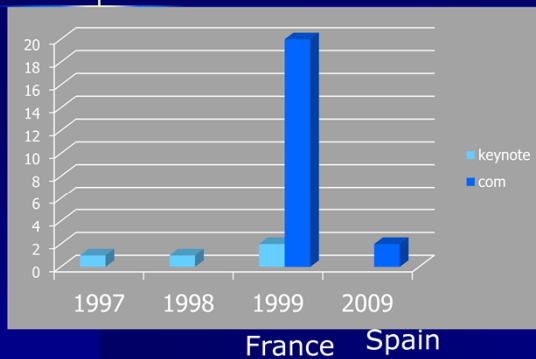
2003?



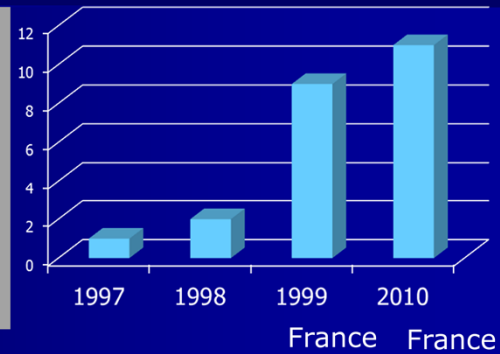
In 95, ReCALL became a joint publication of the CTI and the association Eurocall. In 2003, Eurocall and Hull who still had the editing responsibility gave the publishing task to Cambridge University Press. Of course I would not like to omit to cite the Eurocall review, edited par Ana Gimeno, review still directly published by Eurocall.

Since the beginning of Recall there has been one person who still is in charge of ReCALL, namely June Thompson. Unfortunately for the first time , June could not attend our conference, but we could here send her a real thanks for all the wonderful work she achieved !

## Does Eurocall support multi-languages?



Communications in languages other than English during Eurocall conferences  
(hard to be exhaustive, websites disappeared)



Publications not in English after Eurocall conferences

10

Now let us consider the multilingual issue. I made a quick enquiry among the discontinuous information available on Eurocall websites concerning communications not given in English during our conferences (starting in 93 up to 2012).

I could only note one keynote given in 97 and one in 98.

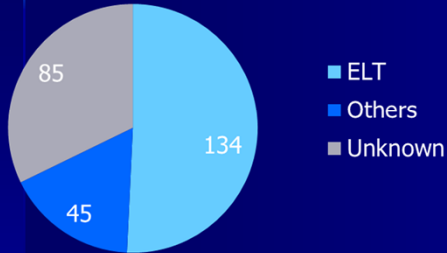
In 99 in Besançon we had one keynote and nearly 20 communications not in English, 2 in 2009 in Spain.

When looking at papers, full papers, published after Eurocall conferences, I found one in 97 in Recall, 2 in 98, but 8 and 11, respectively after the 99 and 2010 conferences in France.

As you can see, the whole amount of communications and articles not in English is fairly reduced.

## WorldCALL and multi-languages

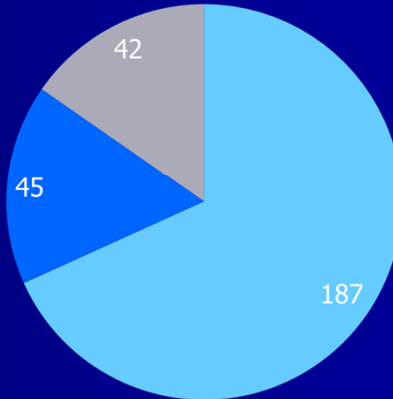
Target languages in WorldCALL13, v1



Sum of papers, posters, courseware.

- When tandems involving ELT, count 1 for ELT and 1 for others
- More than half of Unknown from Asia (English as L2)

Target languages in WorldCALL13, v2



Half of « unknown » counted as ELT.

Let us have a look outside Eurocall. In July this year we had the WorldCALL conference in Glasgow. There were more than 250 events, whether papers, posters, courseware. 134 of them were given by English language teachers and, naturally, only related to this language. Only 45 events were produced by colleagues teaching other languages. 85 authors did not mention any language in their abstracts.

When looking more closely to this latter 85 events, a majority of them were organized by colleagues from East Asia. A very large majority of them are teachers of English. So I can reasonably count half of this 85 events as being connected to ELT. This gives us around 190 ELT events. In other words 70% of the conference related to ELT, versus only 17% clearly not referring to the English language.

Without denying a good scientific quality to WorldCALL conferences, one may wonder whether we are not building another kind of TESOL conference.



## Unpleasant situations for (Euro)CALL

- Nothing against English (cf. my position on French-FLE) (ReCALL accepts submissions in other languages)
- Language is culture and politics
- The humanities generally a multilingual domain: cf. pedagogy ≠ didactique ≠ didaktik
- Can we be trusted by learners when we assert that other languages are used for academic / scientific purposes?

12

Please do not misunderstand my viewpoint. I have nothing against English. When we created the first French speaking CALL journal, namely *Alsic*, we did it in full collaboration with the other English speaking journals. We did jointly organized *Eucall99* in Besançon.

Neither can we complain against *ReCALL*. We always have welcomed submissions of papers in other languages.

But we all know that language issues are closely connected to cultural and political ones. Hence we may be on the way of losing our autonomy.

Moreover, Humanities is generally considered by other scientific fields as a place where researchers can publish good quality papers in vernacular languages. What would it mean for example to be obliged to translate the words “didactique” in French, or “didaktik” in German into pedagogy when knowing these 3 terms have each a distinctive meaning representative of different perspectives in the Educational Research field?

Eventually, can we really be trusted by our learners when we try to explain them “Oh yes please learn Arab, or French, or Spanish, Portuguese, etc. because these are important scientific languages” !

## What can we do?

- Raise awareness through conferences:
  - Specify language taught when submitting
  - Conference organizers build statistics
  - Organize national events during conferences (cf. Portugal this year, Spain, Belgium, France,...) and encourage com. in vernacular language
- Publish in several languages (cf. telecollaboration projects)
- Develop international CALL journals in other languages

13

What can we do in order to develop multilingual spaces in our community:

## Develop international CALL journals



Apprentissage des Langues et Systèmes d'Information et de Communication  
alsic.org ou alsic.u-strasbg.fr

Volume 3  
Numéro 1  
juin 2000

Revue internet francophone pour chercheurs et praticiens

● **Éditorial**  
pages 1 à 3

● **Numéro spécial**

**Sélection d'articles du congrès EUROCALL'99**

Systèmes d'information et de communication (SIC) dans des situations diversifiées d'apprentissage des langues

**eurocall'99**  
coordonné par  
Maguy Pothier et Thierry Chanier

pages 3 à 18 PDF  
La conception des environnements d'apprentissage : de la théorie à la pratique / de la pratique à la théorie.  
de Christian Depover, Jean-Jacques Quintin & Bruno De Lièvre

pages 19 à 47 PDF  
Projet TECHNE : vers un apprentissage collaboratif dans une classe virtuelle bilingue.  
de Françoise Blin & Roisim Donohoe

pages 49 à 59 PDF  
La persévérance dans l'enseignement à distance - Une étude de cas.  
de Lise Desmarais

pages 61 à 76 PDF  
Présentation d'un logiciel de visualisation pour l'apprentissage de l'oral en langue seconde.  
de Aline Germain & Philippe Martin

pages 77 à 98 PDF  
Français, autoformation et ELAO à l'université : didactique du texte et pratique de l'hypertexte.  
de Guy Achard-Bayle & Michèle Redon-Dilax

pages 99 à 108 PDF  
La Dictée Interactive.  
de Simona Ruggia

pages 109 à 123 PDF  
L'utilisation des stratégies d'apprentissage d'une langue dans un environnement des TIC.  
de Janet Altan



After Eurocall2010 (Bordeaux) publications in ReCALL and in another journal

After Eurocall99 (Besançon) publications in ReCALL and in Alsic

14

I mention the larger number of papers published in other languages than English after these two Eurocall conferences organized in France. How did this happen?

In Besançon we offered the possibility to authors to submit papers either to ReCALL or to Alsic . In Bordeaux, proceedings have been published with another publisher.

These opportunities have been jointly organized with EuroCALL. ReCALL lost nothing.

How can we develop publications in other languages?

# Exemples from other disciplines

**SciELO 15 anos**  
Scientific Electronic Library Online

**Sobre o SciELO**  
Sobre o SciELO  
Indicadores Bibliométricos  
Acesso via OAI e RSS

**Rede SciELO**  
**colecções de Livros**  
Brasil  
**colecções de Periódicos**  
África do Sul  
Argentina  
Brasil  
Chile  
Colômbia  
Costa Rica  
Cuba  
Espanha  
México  
Portugal  
Venezuela  
Saúde Pública  
Social Sciences

**Key words:** Protein electrophoresis; Overflow proteinúria; Mioglobinúria; Rhabdomyolysis.

**Resumo**  
A destruição do músculo esquelético na condição patológica conhecida como rabdomiólise resulta na liberação ao torrente sanguíneo de elevadas concentrações da proteína mioglobina de 17 kDa a qual filtra livremente através do glomérulo ultrapassando frequentemente a capacidade de reabsorção do túbulo proximal. Portanto, a identificação de mioglobina em urina é uma ferramenta essencial que complementa outros parâmetros bioquímicos no diagnóstico da doença. No presente trabalho, mediante a combinação de eletroforese em géis de agarose e imunofixação empregando anticorpos específicos, é fornecida evidência direta da presença de mioglobina intacta na urina de um paciente com insuficiência renal aguda associada a rabdomiólise desencadeada por efeito secundário de uma terapia redutora de lipídeos. Os dados eletroforéticos e imunquímicos foram confirmados mediante sequência N-terminal de aminoácidos, immunoblot e espectrometria de massa. A simples combinação de eletroforese e imunofixação fornece uma estratégia flexível que pode se estender à identificação de diversas proteínas envolvidas em proteinúrias de sobrecarga.

**Palavras chave:** Proteinograma; Proteinúrias de sobrecarga; Mioglobinúria; Rhabdomyolise.

**Introducción**  
El proteinograma es una técnica versátil del laboratorio de análisis clínicos que se utiliza de forma rutinaria en el estudio de mezclas complejas de proteínas en fluidos biológicos -como suero, orina y líquido cefalorraquídeo- y que se basa en la carga eléctrica diferencial exhibida por los distintos componentes a un determinado pH. En términos generales, el perfil electroforético de proteínas plasmáticas se encuentra delineado por la distribución diferencial de 14 componentes principales<sup>(1)</sup>. En condiciones normales, este perfil electroforético permanece relativamente constante; cambios en el número o concentración de los componentes se encuentran típicamente asociados a procesos patológicos, aunque en ciertos casos simplemente reflejan diferencias genéticas sin connotaciones patológicas. El análisis electroforético de proteínas es utilizado más frecuentemente cuando se investiga la presencia de componentes adicionales, típicamente ausentes en condiciones normales. En estos casos, la técnica se utiliza generalmente en

In fact, it already exists in other disciplines.

Here is the example of the very large scientific publisher named Scielo. It publishes scientific journals in Spanish and Portuguese with authors belonging to America and Europe. Here is an illustration of a journal in biochemistry published in Portuguese.

## European publishing structures exist

OpenEdition : OpenEdition Books Revues.org Calenda Hypotheses Newsletters and alerts OpenEdition Freemium Search

Three platforms for electronic resources in the humanities and social sciences: Revues.org, Hypotheses.org, Calenda

Home

**openedition**  
REVUES.ORG CALENDAL HYPOTHESES.ORG

About OpenEdition  
Freemium programme  
Membership and training  
Subscribe to the newsletter  
OpenEdition RSS feeds

**CATALOGUES**

800 BOOKS	22723 EVENTS
392 JOURNALS	668 BLOGS

SEARCH

REVUES.ORG NEWS CALENDAL NEWS HYPOTHESES NEWS

OpenEdition is the umbrella portal for OpenEdition Books, Revues.org, Hypotheses and Calenda, four platforms dedicated to electronic resources in the humanities and social sciences. If you wish your university to subscribe to this service and give you access to articles in downloadable formats (PDF, ePub), please visit [OpenEdition Freemium presentation page](#).

**USERS**

To users, OpenEdition offers a vast catalogue of academic publications in the humanities and social sciences, mostly in Open Access. Additional services are available through libraries and institutions subscribing to the OpenEdition Freemium program: detachable formats, search tools, alert service, etc.

**PUBLISHERS**

OpenEdition is a complete infrastructure for electronic publishing dedicated to academic communication in the humanities and social sciences. To editorial teams and publishers, it offers a range of solutions adapted to the publication of books, journals, research blogs and announcements of scientific events.

**LIBRARIES**

To librarians, OpenEdition Freemium program provides a full range of services designed to assist them in managing their OpenEdition subscription and enhancing their users' research experience.

16

You could think “well this is only an example for natural science and technology”. Indeed in Humanities we have a great opportunity with the public publisher OpenEdition. It is managed by academic people, it already publishes 400 journals in the academic field. This is where, for example, the Alsic editorial board moved 5 years ago in order to publish its journal. It has now set up offices in Spain and Portugal, publishes in these languages (here a journal in Sociology) and start discussing with German editorial teams.

So, why not soon start publishing an international CALL journal in Spanish & Portuguese (only one for America & Europe) to ensure a strong scientific basis from the beginning?

## European publishing structures exist

The screenshot displays the Alsic journal website. The header includes navigation links for OpenEdition, OpenEdition Books, Revues.org, Calenda, Hypotheses, Newsletters and alerts, and OpenEdition Freemium. The main content area features the Alsic logo and the subtitle "Apprentissage des Langues et Systèmes d'Information et de Communication". A central text block describes the journal as a federated platform for research in linguistics, didactics, psychology, and communication. It mentions that the journal is referenced in the MLA International Bibliography and the ERIH index. A sidebar on the left contains a search bar, an index, and a list of volumes from 2011 to 2013. A "Dernières nouvelles" section on the right lists recent news items.

17

You could think “well this is only an example for natural science and technology”. Indeed in Humanities we have a great opportunity with the public publisher OpenEdition. It is managed by academic people, it already publishes 400 journals in the academic field. This is where, for exemple, the Alsic editorial board moved 5 years ago in order to publish its journal. It has now set up offices in Spaine and Portugal, publishes in these languages (here a journal in Sociology) and start discussing with German editorial teams.

So, why not soon start publishing an international CALL journal in Spanish & Portuguese (only one for America & Europe) to ensure a strong scientific basis from the beginning?

## European publishing structures exist

OpenEdition Books Revues.org Calenda Hypotheses Notícias e alertas OpenEdition Freemium

# SOCIOLOGIA

PROBLEMAS E PRÁTICAS

Pesquisa

...então. É uma revista  
... é publicar artigos de análise  
... estantes de investigação original, elaboração  
teórica ou balanço temático. Está também aberta  
interdisciplinarmente a trabalhos provenientes de outras áreas  
das ciências sociais. Dirige-se ao espaço internacional, publicando  
artigos em português, inglês, espanhol e francês, de autores de  
variados países. A orientação editorial da revista pauta-se por  
princípios de qualidade científica, pluralismo paradigmático e  
relevância social. Procura que os artigos publicados constituam  
contributos significativos para o avanço do conhecimento. Os  
artigos propostos são submetidos a avaliação independente de  
pelo menos dois especialistas reconhecidos de diversos países, em  
regime de duplo anonimato.

último número online  
**72 | 2013**  
**Varia 72**

**Números em texto integral**

- 72 | 2013  
Varia 72
- 71 | 2013  
Varia 71
- 70 | 2012  
Varia 70
- 69 | 2012  
Varia 69
- 68 | 2012  
Varia 68
- 67 | 2011  
Varia 67
- 66 | 2011  
Varia 66

18

You could think “well this is only an example for natural science and technology”. Indeed in Humanities we have a great opportunity with the public publisher OpenEdition. It is managed by academic people, it already publishes 400 journals in the academic field. This is where, for example, the Alsic editorial board moved 5 years ago in order to publish its journal. It has now set up offices in Spain and Portugal, publishes in these languages (here a journal in Sociology) and start discussing with German editorial teams.

So, why not soon start publishing an international CALL journal in Spanish & Portuguese (only one for America & Europe) to ensure a strong scientific basis from the beginning?



1 2 3 4

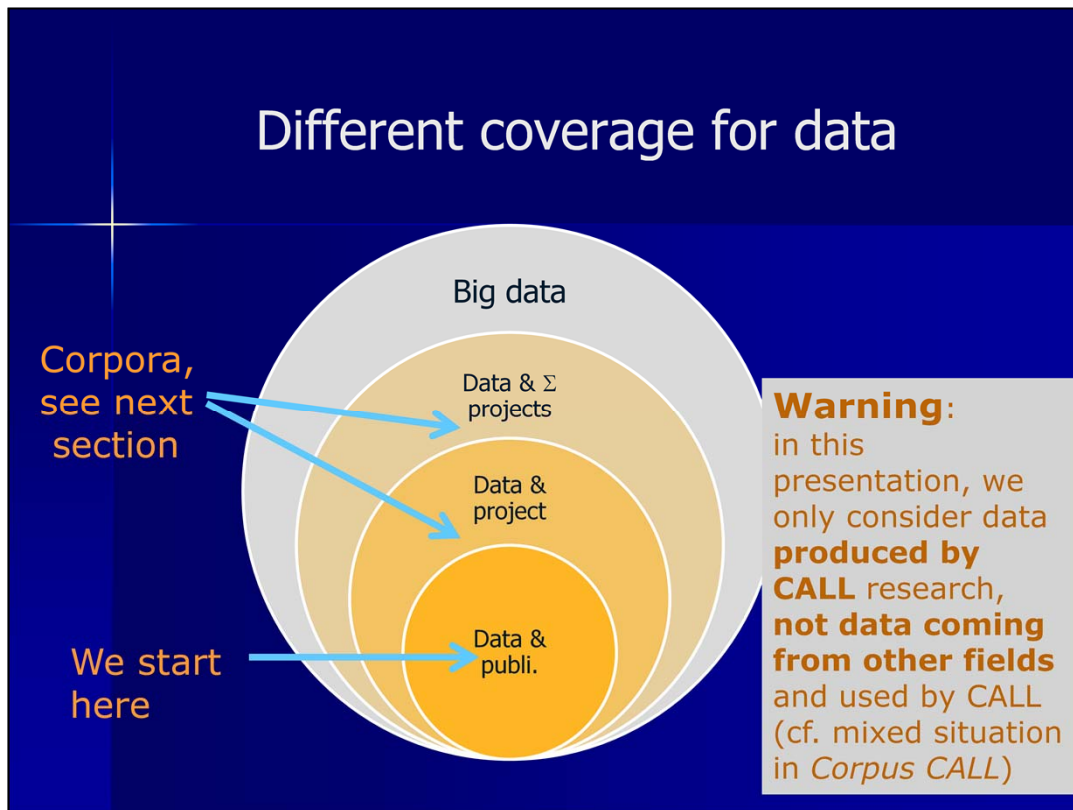
Enhance research quality in CALL

**ORGANIZE AND PUBLISH  
RESEARCH DATA**

19

After having considered publications, let us turn our attention to research data





What types of data can we find in research?

- Data linked to publications
- Data coming out of research projects (one or several CALL experiments)
- A recent new type the so-called Big Data (see the extended version)

Let us start with data linked to publications

## Current situation in CALL (and many, but not all, fields in Humanities)

- Some (not all) of our papers are based on research data
- These data (empty forms, forms filled, spreadsheets, transcriptions, language data and their computation, audio, video, etc.) are not accessible to reviewers, nor to the readers once papers are published

## What other disciplines say

“**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 King (2007). "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," *Sociological Methods and Research*, Vol. 32, No. 2

22

What other disciplines in Humanities say about this?

Gary King, a sociologist, explains why we should make data, he called « replication data sets » available with publications.

## What Europe says

### 8. What does the Commission propose to do about open access to data in Horizon 2020?

The Commission proposed in its Communication 'Towards better access to scientific information' to develop a pilot on open access to data, primarily those data underlying (open access) scientific publications. The areas covered by the pilot should be discussed together with the thematic Units within the Commission and their stakeholders.

COMMISSION RECOMMENDATION of 17.7.2012 on access to and preservation of scientific information : [http://ec.europa.eu/research/science-society/document\\_library/pdf\\_06/recommendation-access-and-preservation-scientific-information\\_en.pdf](http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf)

23

### What does Europe say?

A recent text published by the commission with respect to the new 2020 European framework proposes, in order to improve scientific communication, to develop a pilot on open access to data, primarily those data underlying scientific publications.



**Data publication for CALL journals:  
proposal for a joint project**

25th July 2013

In CALL, since we are not waiting for the last train to arrive, we started discussing a joint project among 5 journals, European and North American ones, which have the habits of already cooperating, where part of the people here in the room are reviewers or members of the editorial boards.

## Contents of the proposal

- Reviewers will access data when reading the paper (strengthen the review process)
- Once the paper is accepted, data are published
- The reader (researcher) can access these data in order to replicate, join them to her/his own data, etc.), cf. Opendata
- The author is the great winner! Two references to her/his work: data will have an individual reference (but linked to) the paper's reference



25

Here is the contents of the proposal made to these journals.  
I am pleased to announce that yesterday the ReCALL editorial board, supported by CUP, accepted to joint the project !

# Link between publication & data: example from earth sciences

Arason, P et al. (2011): Plume-top altitude time-series during 2010 volcanic eruption of Eyjafjallajaj??. Icelandic Meteorological Office, Reykjavik, [doi:10.1594/PANGAEA.760690](https://doi.org/10.1594/PANGAEA.760690),  
Supplement to: Arason, Pordur; Petersen, G N; Bjornsson, H (2011): Observations of the altitude of the volcanic plume during the eruption of Eyjafjallajaj, April-May 2010. *Earth System Science Data*, 3, 9-17, [doi:10.5194/essd-3-9-2011](https://doi.org/10.5194/essd-3-9-2011)

**Observations of the altitude of the volcanic plume during the eruption of Eyjafjallajajkull, April–May 2010**  
P. Arason, G. N. Petersen, and H. Björnsson  
Icelandic Meteorological Office – Vindurstafræði Íslands, Bustaðavegur 9, 15-150 Reykjavík, Iceland

**Abstract.** The eruption of Eyjafjallajajkull volcano in 2010 lasted for 39 days, 14 April–23 May. The eruption had two explosive phases separated by a phreatic eruption and reduced explosive activity. The height of the plume was monitored every 5 min with a C-band weather radar located in Keflavík International Airport, 155 km distance from the volcano. Furthermore, several web cameras were mounted with a view of the volcano, and their images saved every 5 min. Time series of the plume-top altitude were constructed from the radar observations and images from a web camera located in the village Hvolsvöllur at 24 km distance from the volcano. This paper presents the independent radar and web camera time series and performs cross validation. The echo top radar sea level altitude of the volcanic plume are publicly available from the PANGAEA Publishing Network (<https://doi.org/10.1594/PANGAEA.760690>).

Discussion Paper (PDF, 1592 KB) • Supplement (44 KB) • Interactive Discussion (Closed, 6 Comments) • Final Revised Paper (4550)

**Citation:** Arason, P., Petersen, G. N., and Björnsson, H.: Observations of the altitude of the volcanic plume during the eruption of Eyjafjallajajkull, April–May 2010, *Earth Syst. Sci. Data Discuss.*, 4, 1–23, [doi:10.5194/essd-4-1-2011](https://doi.org/10.5194/essd-4-1-2011), 2011. • [BibTeX](#) • [EndNote](#) • [Reference Manager](#) • [RSS](#)

Journal site  
Data site

The screenshot shows the PANGAEA data repository page for the dataset 'Plume-top altitude time-series during 2010 volcanic eruption of Eyjafjallajajkull'. The page includes a citation, abstract, and download options. The citation is: Arason, P. et al. (2011): Plume-top altitude time-series during 2010 volcanic eruption of Eyjafjallajajkull. Icelandic Meteorological Office, Reykjavik, doi:10.1594/PANGAEA.760690. The abstract describes the data collection methods and the time series of the plume-top altitude. The download options include a PDF of the discussion paper, a supplement, and the final revised paper.

# WHAT WE STARTED TO DO IN FRANCE

27



## Datapublication (French project)

- With the help of TGE-Adonis (national infrastructure for humanities)
  - Now part of Huma-Num
- For Alsic and Sticef journals (as a starting point)
- Every journal has its entries, have an internal review process (cf. OJS) for data
- Reviewers can look at data when reading the paper (data are not open at this stage)
- When the paper is accepted data are published



<http://datapublication.tge-adonis.fr>

http://sticef.univ-lemans.fr/num/vol2012/05-guichon/sticef\_2012\_guichon\_05.htm  
http://datapublication.tge-adonis.fr/data/d-001-102

## An exemple

**Data publication** - publish articles and research data -

Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation

version pleine page  
version à télécharger (pdf)

Volume 19, 2012  
Article de recherche

### Données de l'enquête sur les usages des TIC par les lycéens

#### Les usages des TIC par les lycéens - déconnexion entre usages personnels et usages scolaires

Nicolas GUICHON (ICAR, Lyon 2)

**Information**  
Year: 2012  
Author(s): Nicolas Guichon  
Provide by: Nicolas Guichon  
Keywords: usages numériques, fracture numérique, compétences, Technologies de l'Information et de la Communication (ICT), ICT usage, digital divide, compétences

**Description**  
Un questionnaire concernant l'origine (profession des parents, lycée, équipement informatisé administré au printemps 2011 à 1300 élèves de première en France métropolitaine. L'enquête concerne les lycéens de l'enseignement secondaire général en France. Elle a concerné 10 lycées issus de périphérie urbaine ou centre d'une grande ville.

**Restrictions**  
Les données issues de ce questionnaire sont destinées à la communauté scientifique. To diffusion doit faire l'objet d'une demande particulière à l'auteur. L'utilisation commerciale n'est pas autorisée.

**Data file**  
To download the files click on their name. The downloadable files are the formats provided by a request must be do from the author.

<a href="#">Questionnaire_FINAL-Guichon.pdf</a> application/pdf - 25.4 ko	Other
<a href="#">Enquête Guichon2011.xls</a> application/vnd.ms-excel - 1.23 Mo	Other
<a href="#">Codage1.csv</a> application/vnd.ms-excel - 3.74 ko	Data
<a href="#">Codage2.csv</a> application/vnd.ms-excel - 6.07 ko	Other
<a href="#">Etablissement.csv</a>	

**RÉSUMÉ** : Cette recherche, qui adopte la perspective de la sociologie des usages, s'appuie sur une enquête à la fois par questionnaire et par entretiens pour sonder les usages numériques des lycéens de l'enseignement secondaire général en France. Deux objectifs sont visés : d'une part, grâce aux données empiriques obtenues, un état des lieux des usages numériques des jeunes est conduit. D'autre part, cette étude investigate de quelles façons les outils numériques sont utilisés pour le travail à la maison et pour l'apprentissage des langues étrangères. Les résultats mettent au jour une déconnexion entre usages des Technologies de l'Information et de la Communication (TIC) entre la sphère privée et la sphère scolaire.

**MOTS CLÉS** : usages numériques, fracture numérique, compétences, Technologies de l'Information et de la Communication (TIC)

**DONNÉES ASSOCIÉES** : Accessibles via le lien permanent (Datapublication.org : Tge-adonis.fr.)  
<http://hdl.datapublication.org/11107/d-001-102>

**IRIS IS NOT THE PROJECT  
WE ARE LOOKING AT**

30

http://www.iris-database.org



IRIS

A digital repository of data collection instruments for research into second language learning and teaching

THE UNIVERSITY of York  
GEORGETOWN UNIVERSITY

[Home](#) [Submit](#) [Search and Download](#)

[Login to IRIS](#)

### Welcome to IRIS

IRIS is a free and public resource. If you are submitting to IRIS, we recommend you [Login](#) first so that you can come back and edit your information at a later date.

[Submit  
Instrument / Materials](#)

[Search and Download](#)

*The IRIS team*  
*Project Directors:*  
Emma Marsden  
Alison Mackey

*Development & Administration:*  
Julie Allinson  
Frank Feng  
Julia Key

*Advisory Board:*  
Rod Ellis  
Susan Gass  
Jan Hulstijn  
Lourdes Ortega  
Leah Roberts  
Norman Segalowitz  
Peter Skehan

*Repository Expertise:*  
Michael Day  
Judith Klavans  
David Martin

IRIS is funded by the ESRC and is an Academy Research Project, funded by the British Academy

### About IRIS

IRIS is a collection of instruments, materials and stimuli used to elicit data for research into second and foreign languages. Materials are freely accessible and searchable, easy to upload (for contributions) and download (for use).

- [What people are saying about IRIS](#)
- [Find out more about IRIS](#)
- [Download the IRIS Flyer](#)

[Tweet](#) (1)

[Help](#)

[Links](#)

[Conference](#)

[Statistics](#)

[Journals that support IRIS](#)



## Why not IRIS?

- Iris is an interesting OpenData project with links to journals from UK and USA universities, sponsored by UK, but ...
- Data are not part of the review process
- Once a paper is accepted authors do as they pleased, e.g: some put the form of a questionnaire, not the data collected (answers), nor the computation (spreadsheet)
- Metadata are not standard (just for search on the site, like Merlot)
- They are local and cannot be harvested
- No reference to the data (cf. DataCite) , no permalink
- No crosslink between data and publication (which would not make sense because data are not exhaustive) and have not been part of the evaluation process

# CALL Datapublication project

- Make a common proposal at the European Union level (Research agency) via DARIAH
- Get logistical and official scientific support in order to design and open a website site (Datapublication)
- Where our 5 journals will have separate access for their editorial board in order to manage distinct review process
- Manage a joint design for the workflow of the review process
- Metadata format will be standard, permalink given, full reference with link and full reference of papers
- When the web site is open, for every journal author's guidelines need to be changed (when authors submit papers which rely on data) and links be implemented in order to point from the journal to the data site
- Then the Datapublication website may be open for other journals in humanities (best to get EU support) whether they are based in or out of EU



1 2 3 4

With extracts from Wigham & Chanier (2013)

# DATA & PROJECT(S), LETEC CORPORA

34

In this 3<sup>rd</sup> section we will now consider data stemming from research projects.

## First corpora in CALL : learner corpora



- Building corpora : collecting learners' production (essais), structuring, annotating, processing
- Using corpora
  - To enhance learning (DDL: data driven learning) under some circumstances
  - To enhance research
- Thinking about : Eurocall SIG, conferences, special issues, etc.

35

Generally when one talks in CALL about data assemble into corpora, we immediately think of learner corpora.

As you know, a learner corpus is made out of learners' productions. It is studied as a way to enhance language learning or the understanding of the learning process. Since there are plenty of opportunities on this issue in this conference, I will not say more on learner corpora, except when mentioning the question of access to data.

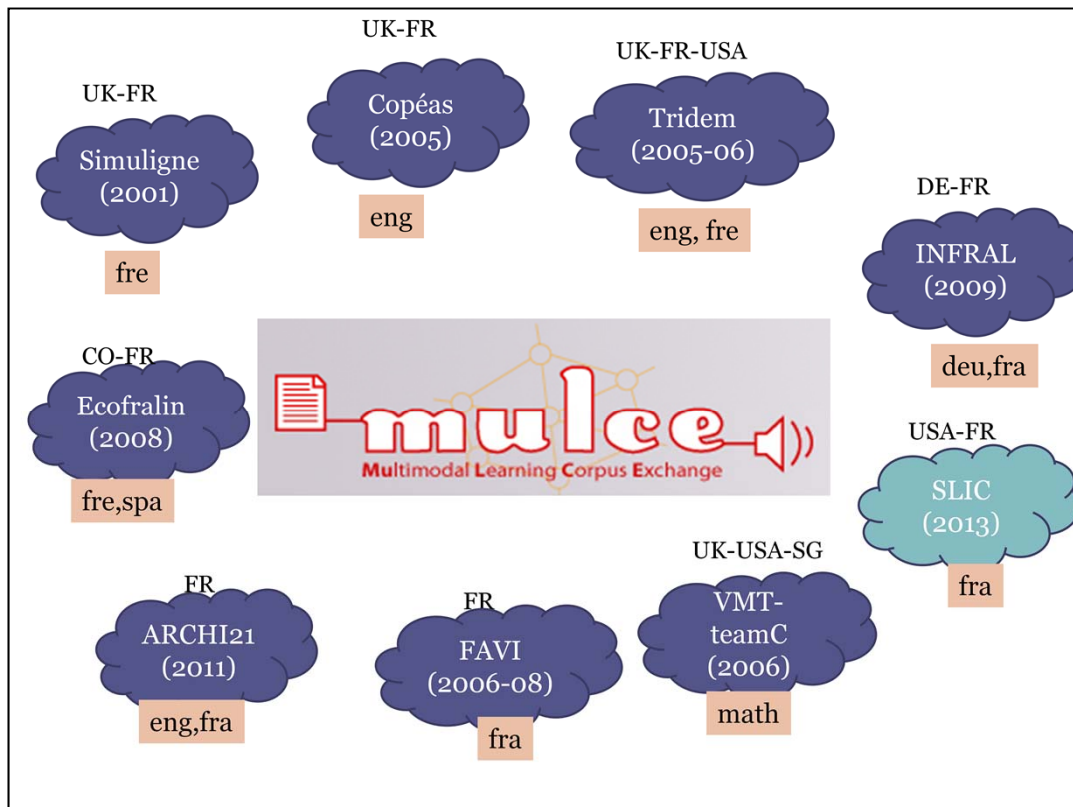


## New type of corpora

- LEarning and TEaching Corpora (LETEC) (*corpus d'apprentissage*)
- data-sharing and repository for research on multimodal interactions

36

Here I would like to introduce another type of corpus, the Learning & teaching corpus (LETEC). Examples will be taken from the field of online learning situations in a multimodal context.



Over the past 12 years a community of researchers have been involved in online language learning projects, either for designing pedagogical scenarios, research protocols, for online tutoring, for collecting, analysing data or/and for publishing.

Projects started in 2001, with the global simulation Simuligne. From 2007 some of us decided that it was time to adopt a coherent and systematic way to organize data in order to improve our research methodology for reasons I will soon mention. It gave birth to the Mulce project.

It is impossible for me to cite all my colleagues here. Let me just mention Christophe Reffay who was there from the beginning, Marie-Noelle Lamy with whom we co-develop the Simuligne project and the Mulce project. Marie-Laure Betbeder, Maud Ciekanski and Ciara Wigham came later on and made a great deal of work. Chris Jones, present here,

co-constructed the Tridem project with Mirjam Hauck, Tim Lewis,, and Bonie Youngs.

As you can see here, every cloud corresponds to an online learning situation, a research project. Above clouds you have IDs of the country involved (Colombia, Germany, UK, USA, etc.), most of the situations correspond to what is now called a telecollaborative project. Under the clouds you have the languages that were at stake.

## Data validity & reliability in CALL research?

- Questions related to validity and reliability
- Problems in Humanities, Social Sciences and CALL:
  - Visibility, accessibility of research data
  - Data representative / anecdotal?
  - Publication (already mentioned)
- CALL data is often:
  - not contextualised – pedagogical & technological situations (Kern *et al.*, 2004)
  - tangled in specific software using proprietary formats
- Replication for interaction analysis in online learning near impossible:
  - variables that are difficult to control
  - replication does not imply that phenomenon previously observed will reoccur (Reffay *et al.*, 2012)



Mulce researchers were concerned with questions related to validity and reliability. When we set up a learning situation, study it and publish to which extent what we say is anecdotal or can be generalized? Did we actually studied what we pretended to study or did we neglected hidden factors which may open the way to other explanations, conclusions?

When we want to give some insurance about these issues and discuss them with other researchers, we are in a real trouble, for many reasons.

Among others:

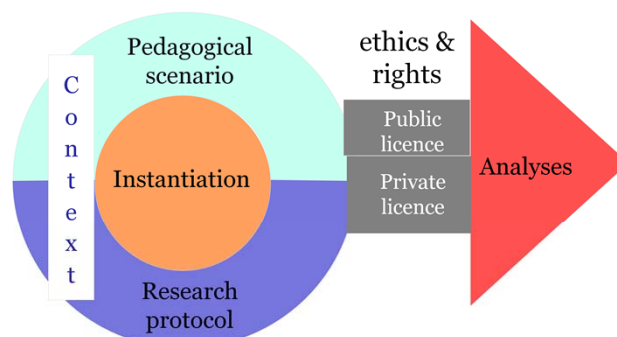
- As mentioned no data associated with publications
- Almost no research data accessible nor visible
- When you do have slices of data they often are not contextualized (what were the precise technological and pedagogical situations?)
- They are tangled in specific software using proprietary formats

## Research data quality: Mulce project

- Interoperability:
  - Structured and coherent data sets
  - => analyses can be completed by researchers who did not participate in the course
- Sustainability:
  - Independent from online platforms
  - Stored in independent formalisms
- Open access to research data & appropriate licences 
- Accessibility:
  - Finding the research data thanks to harvesting protocols based on standard metadata — 
  - OLAC* (Open Language Archives Community)

Hence in Mulce we decided to design corpora with these criteria in minds:

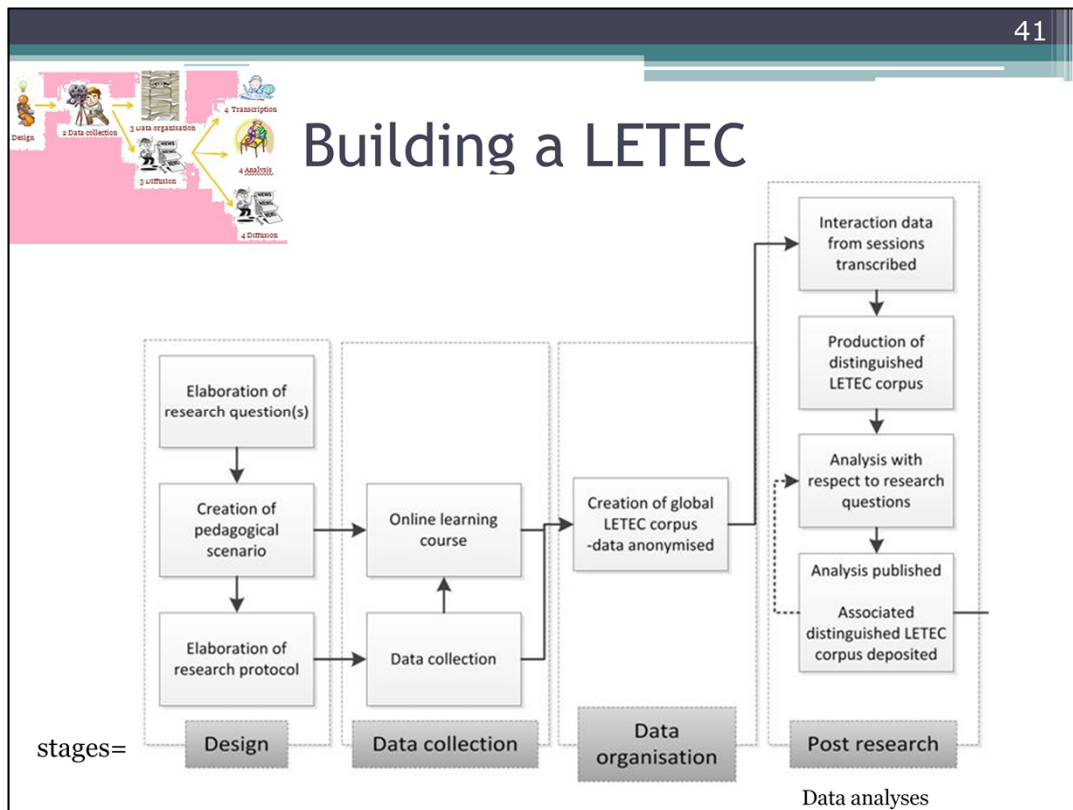
## LETEC Components



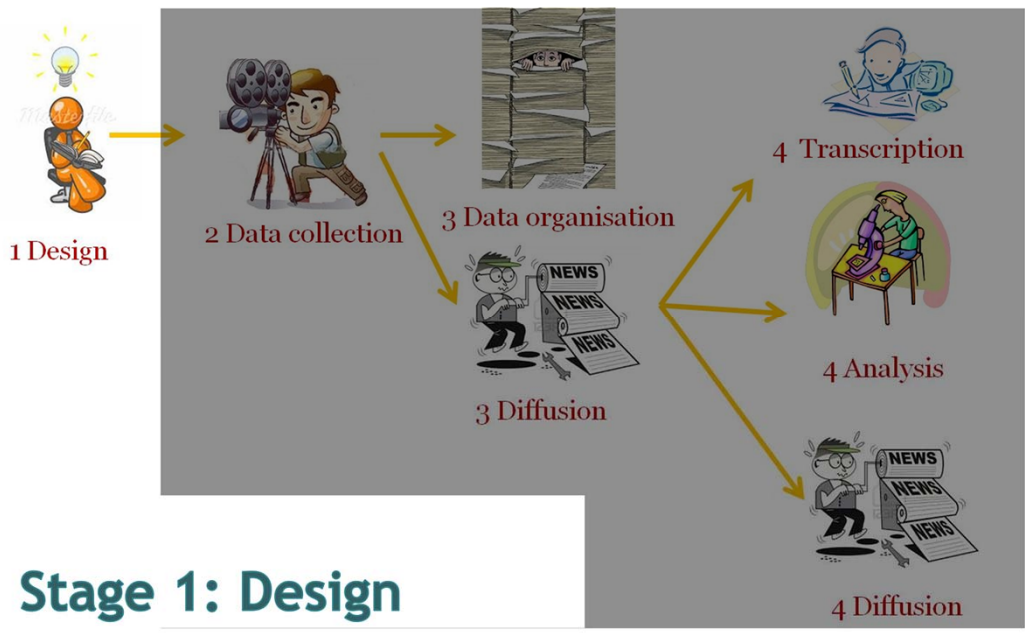
"A LETEC corpus collects in a **systematic and structured way** all the data from **interactions** which occur during a course which is **partially or entirely online**. These data are enriched by technical, pedagogical and scientific information as well as information about the participants and are organized to allow **contextualized analyses** to be performed." (Mulce-documentation, 2013)

Here is the definition of what we mean by a LETEC corpus.

A learning and teaching corpus is made of several parts. The research protocol and the pedagogical scenario describe the context of the learning situation. The technical term "instantiation" (coming from the IMS consortium) refers to interactions and participants' productions collected during the learning situations. We also assemble forms and licences related to ethics and right. Forthcoming analysis may be attached to the first version of the corpus or come later on.



How can a LETEC be build? Here is a schema detailing the process. We presented it at WorldCALL. I will quickly skip into it just for the sake of understanding the rest of my presentation.





## Illustration of methodology-


- European project KA2 Languages
  - CLIL approach (Content and Language Integrated Learning)
    - Architecture + French / English L2
  - Hybrid course "Building Fragile Spaces" : 5-day studio Feb. 2011
  - 17 students, 2 architecture tutors, 1 EFL tutor, 1 FFL tutor
- Working with external partners: exchanges

To contextualized things let me choose one of our latest project, Archi21.

We designed a CLIL scenario jointly with teachers of architecture and language teachers. Learners had an intensive course in order to develop an architectural project, simultaneously F2F and online in Second Life. Language teachers were only at a distance. We had 4 groups of learners either in French or English as a foreign language.

## Elaboration of research areas

- Interplay between verbal and non verbal modes
- Role of nonverbal in identity construction
- Interplay between textchat & voicechat modalities



Support for L2 verbal participation and production

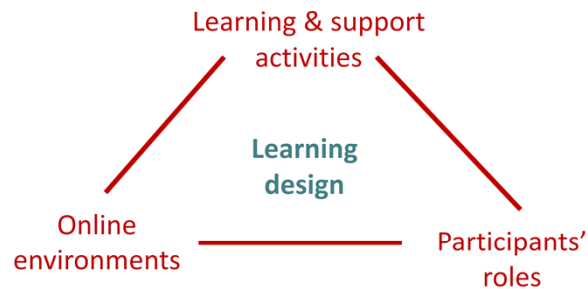
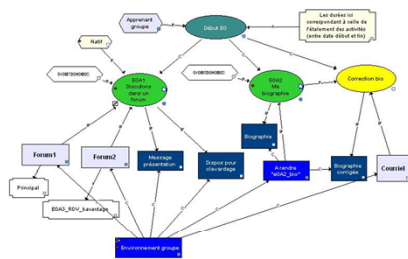
Wigham (2012) – PhD Thesis <http://tel.archives-ouvertes.fr/tel-00762382>

During this first stage, the design stage, research questions are fixed, here mainly:

- relationships between verbal and non verbal modes (by “verbal” please understand it as being the antonym of non-verbal. No exclusive relation to speech)

- And Interplay between textchat & voicechat modalities

## Pedagogical Design



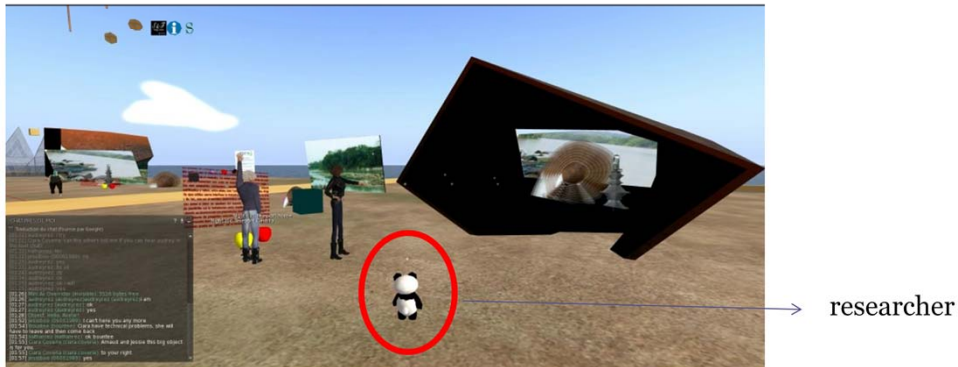
- Macro-task– collaboratively elaborate a model in a synthetic world (*Second Life*) as a response to an architectural problem brief
- Architectural studio, hybrid CLIL approach
- 4 workgroups

The design process of a learning scenario encompasses the description of the 3 main elements : online environments, learning activities and participant's roles.

*Generally we present our learning design in a pretty formal way (see the graph on the left) in order to let people clearly understand it and to relate every step to pedagogical documents such as the guidelines and resources given to the learners. But there is no obligation for using such description format. A corpus compiler can just describes the pedagogical design as a simple text.*

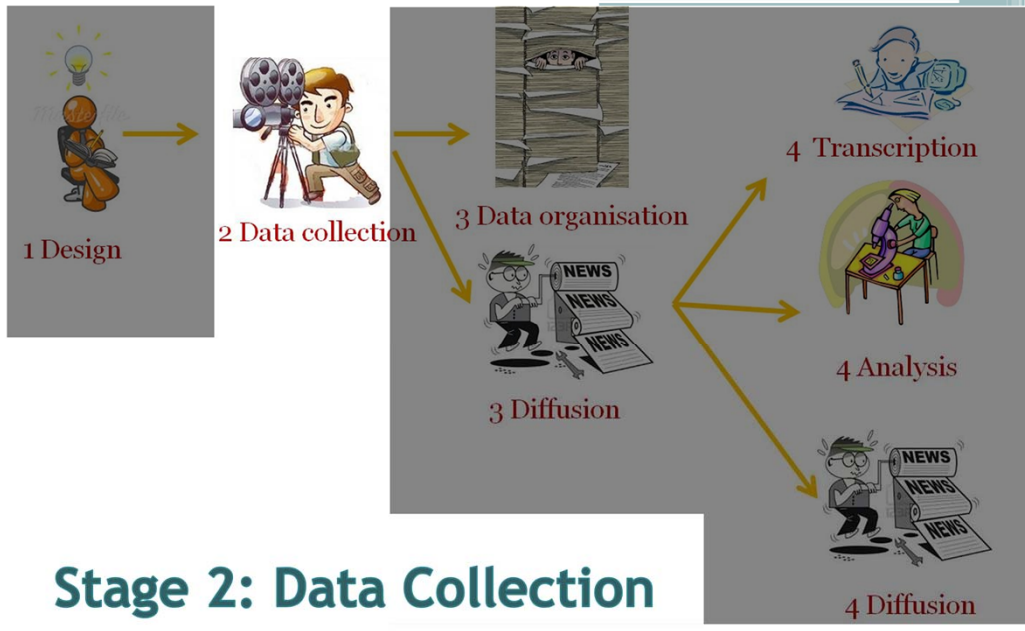
## Research protocol

- Research protocol design
  - Protocol for data collection
  - Researchers' roles
  - Timetable of research activities



Wigham & Chanier, 2013 *ReCALL*

The design of the research protocol includes definition of researchers' role besides the teacher's one, the protocol for data collection, for questionnaires, ethics agreement, etc.

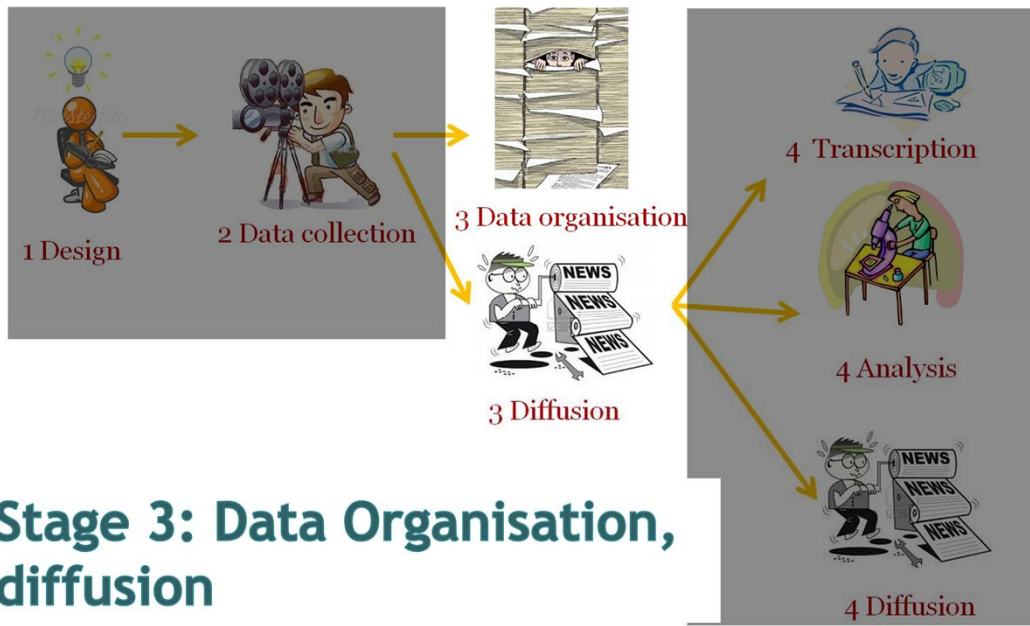


## Stage 2: Data Collection

## Data collection & coverage for Archi21

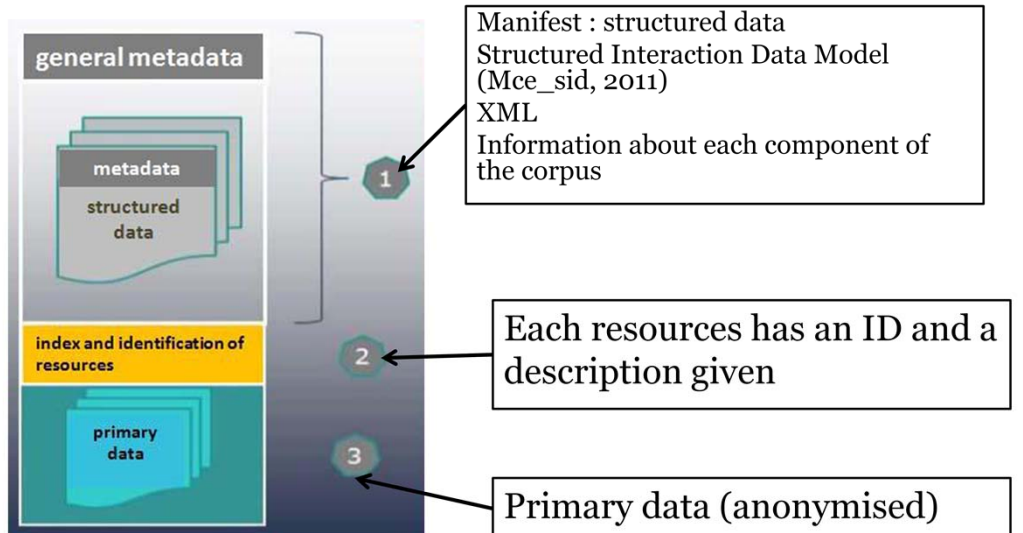
	pre-course	during course		post-course	
Data collected	Pre-questionnaires	Session data		Post questionnaires	Semi-directive interviews
Environment	Kwiksurveys	<i>Second Life</i>	<i>VoiceForum</i>	Kwiksurveys	Skype
Data type	Spreadsheet file	Video screen captures	Audio recordings	Spreadsheet file	Audio recordings
Quantity & coverage of data	17 student questionnaires	20 group sessions & 2 presentation sessions 19h40m	64 forum messages	16 student questionnaires	5 student interviews 2h30

Here is an overview of data collected during the Archi21 course: pre and post questionnaires, 20hours of videoscreen captures, post-interviews, etc.



### Stage 3: Data Organisation, diffusion

## LETEC global corpus: IMS content packaging



In order to organize data, we use the IMS-CP format (a standard coming from the IMS consortium already mentioned. This international consortium which gathers academic institutions and companies is concerned with establishing interoperability for learning systems and learning content) , which I have no time to develop here.

*The bottom part of the corpus assemble the primary data (after they have been anonymised). In the second part, a set of IDs make the link between the top description and the corresponding files of the primary data.*

*The top part, called the “manifest” (another technical term) is made of one XML file and give information about each component of the corpus: metadata, technological environments used in the course, bio information and IDS about participants (teachers, learners, groups).*

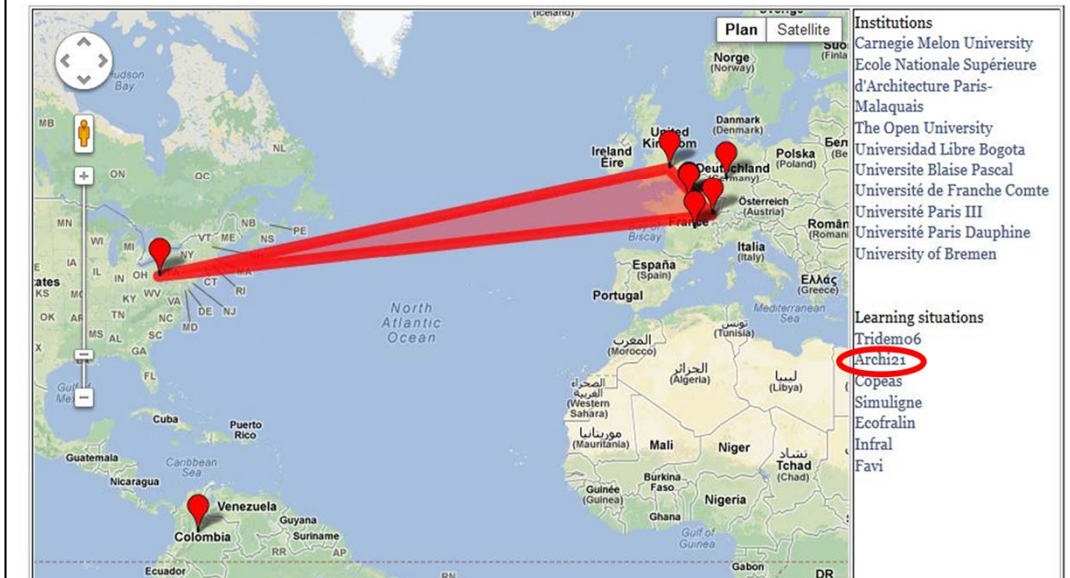
*Then in a structure called workspaces are the interactions, links between what participants did with respect to the pedagogical*



*scenario.*

## Corpus deposit

- *Mulce* corpus repository : <http://repository.mulce.org>



All these 3 components are assembled in an archive which is deposited into the corpus databank, “Mulce repository”.

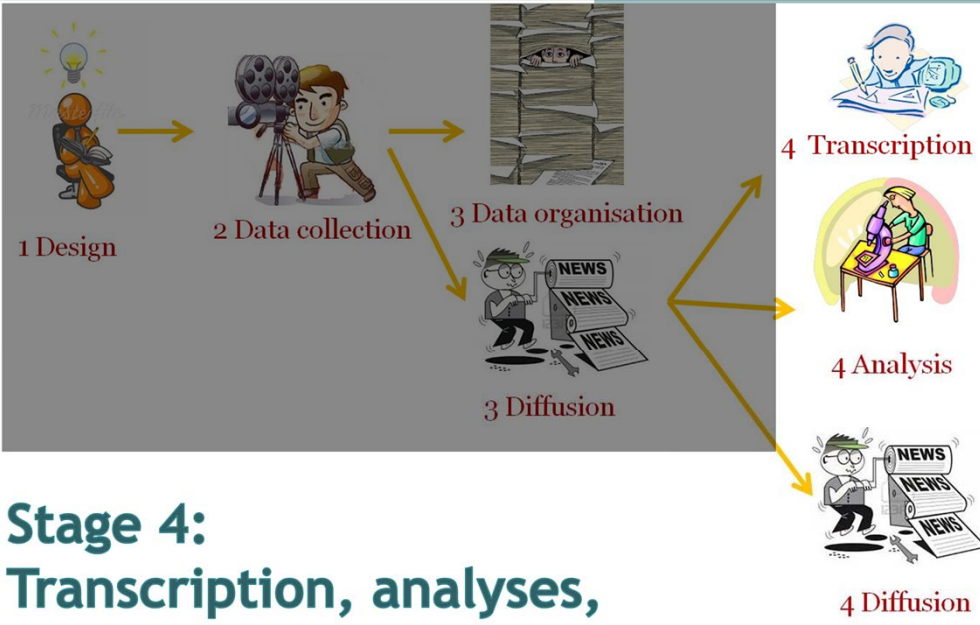
Here is part of the interface with the locations of the different experiments already mentioned. It is one way to access the archives / corpora.

## Corpus diffusion

- Description of corpus; interface to browse structure; zip file to download

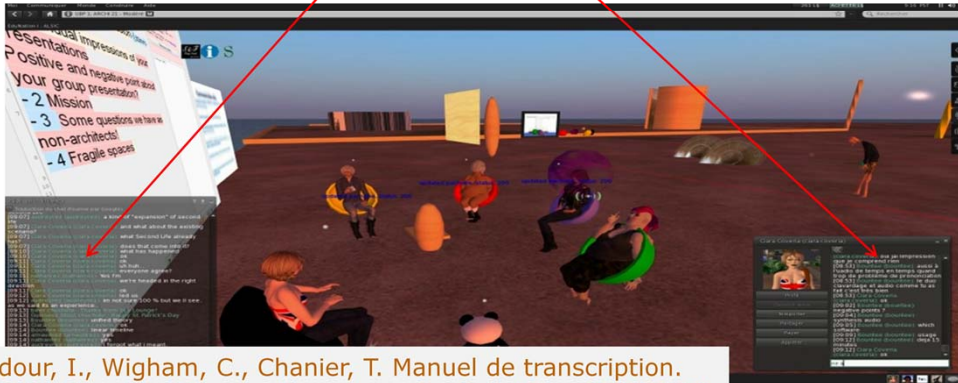
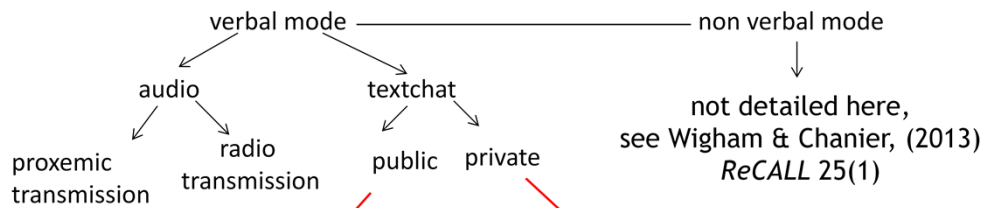
	Recherche	Réinitialisation
Situations d'apprentissage	Archi21, Copeas, Ecofralin, Favi, Infral, Letec, Simuligne, Tridem06, VMT	
Objets	Scénario pédagogique, Corpus global, Corpus distinguable, Protocole de recherche	
Options		
Technologies	Plateforme textuelle (WebCT), Synchron multimodal (Lyceum), 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	

Another way of selecting and downloading corpora is offered through this second interface which details corpora criteria such as, learning situations, technological environments used, languages, types of interactions, pedagogical approaches, etc.



**Stage 4:  
Transcription, analyses,  
publications**

## Multimodal data transcription



Saddour, I., Wigham, C., Chanier, T. Manuel de transcription. (2011) - <http://edutice.archives-ouvertes.fr/edutice-00676230>

We started transcribing multimodal data in 2005 thanks to Lyceum, the Open University environment used in the Copéas experiment. Here is a simplified view of the kinds of interactions we transcribe in Second Life (it is detailed in one of our recent paper in ReCALL).

Whilst teaching and researching in various environments, we elaborated our transcription methodology. The latest version of our manual for transcription is online and, as all our publication, open access.

## Production & deposit of LETEC distinguished corpus

- Particular analysis of a selected part of the global LETEC corpus

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

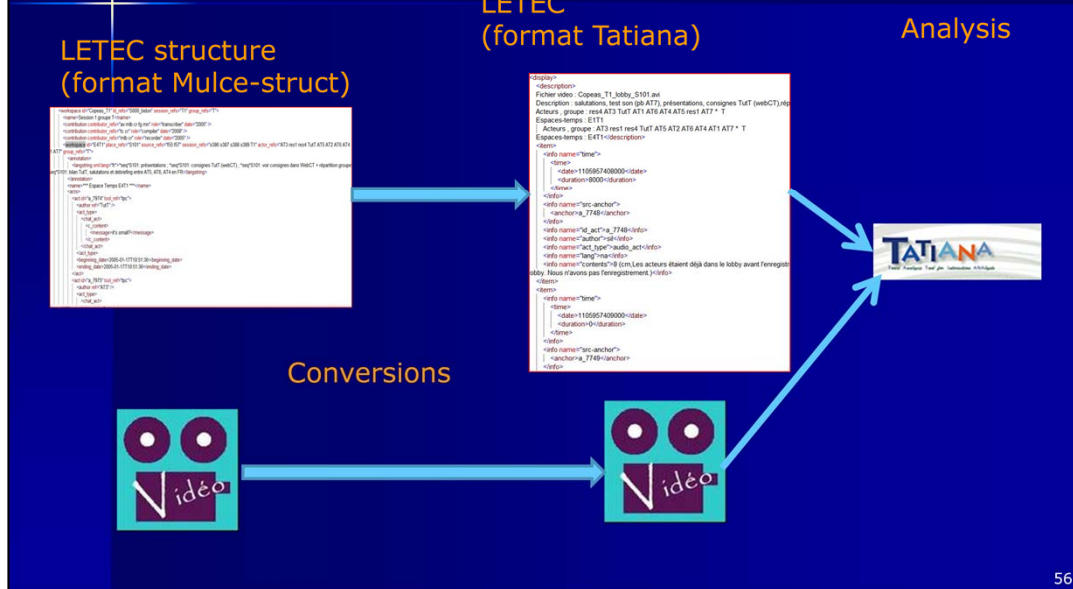
- Only contains transformed data (=the transcriptions)
- Refers to a selection of the original data in global corpus (=videos)
- Software used for transcription cited (=ELAN)

Once you achieve your transcription and analysis, you compile them into what we called a distinguished corpus (the one before being called the global corpus) and you deposit this second corpus.

As you can see here every corpus receive its own reference on the repository.

The new corpus only contains transformed data. It gives links to data already described into the global corpus, and add description on the tools used during the analysis and the transcription step.

## Simple conversions from LETEC to analysis tools

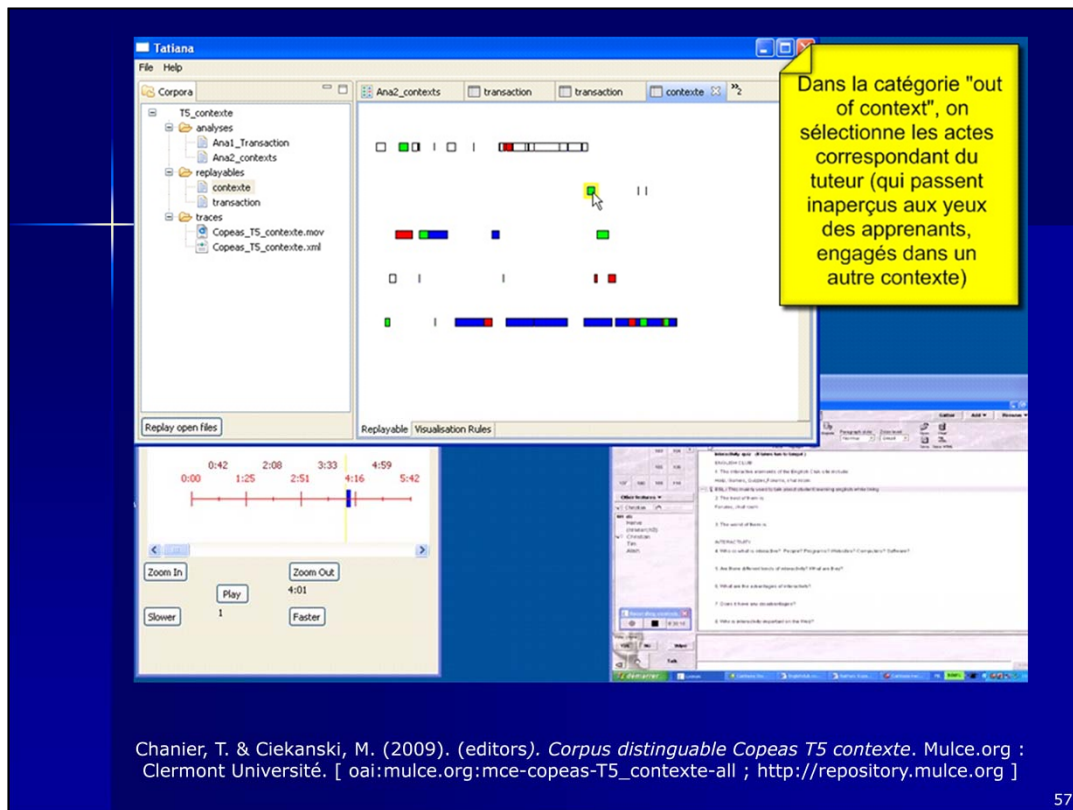


56

Some colleagues often wonder why spending so much time on organizing data. Well research is not a one shot process. Once you have your learning situation and your data, you may want to study it from different perspectives. News ideas may come and new analysis tools help you build these new ideas, it is an interactive process.

For example here, on the left you have transcriptions coming from a Copeas corpus, transcriptions in our XML format, linked to a video capture in a given format. We have been interested in using the Tatiana software. It has been a straightforward process to transform Mulce XML format to Tatiana XML one and convert videos formats..





As an example, you can see an extract of a session in Lyceum, displayed in Tatiana : colors refers to modalities (audio, textchat, word processor) and layers to participants. 3 learners are collaboratively writing into a document, whereas the tutor tries several times intervening but is completely ignore by the learners.

Maud Ciekanski gave an interpretation to this surprising phenomena through an analysis of the Context. She used the Goodwin & Duranti 92 model to explain why the tutor was out of context. We published this analysis as a LETEC corpus.

Funnily I recently discovered in a paper published by Lamy in 2012, this quotation.

It is a general paper were she explains why Social semiotics can be a good theoretical framework to study multimodality.

As far as I understood, Marie-Noelle was not referring to a specific example.

We can feel here how it could be interesting, around pieces of data, to raises different theoretical issues, or explanations.



Dans la catégorie "out of context", on sélectionne les actes correspondant du tuteur (qui passent inaperçus aux yeux des apprenants, engagés dans un autre contexte)

Various interpretations on data :  
 - (Ciekanski & Chanier, 2007) Context (Goodwin & Duranti, 1992 )  
 - (Lamy, 2012) Social semiotics (Kress & Leeuwen, 2001)

“imagine that the tutor led his tutorial via postings in the text-chat while students talked about other topics in the audio channel. It is unlikely that the group would accept such a position for the tutor, and we draw from multimodal social semiotics to help explain why.”

Chanier, T. & Ciekanski, M. (2009). (editors). *Corpus distinguishable Copeas T5 contexte*. Mulce.org : Clermont Université. [ oai:mulce.org:mce-copeas-T5\_contexte-all ; http://repository.mulce.org ]

As an example, you can see an extract of a session in Lyceum, displayed in Tatiana : colors refers to modalities (audio, textchat, word processor) and layers to participants. 3 learners are collaboratively writing into a document, whereas the tutor tries several times intervening but is completely ignore by the learners.

Maud Ciekanski gave an interpretation to this surprising phenomena through an analysis of the Context. She used the Goodwin & Duranti 92 model to explain why the tutor was out of context. We published this analysis as a LETEC corpus.

Funnily I recently discovered in a paper published by Lamy in 2012, this quotation.

It is a general paper were she explains why Social semiotics can be a good theoretical framework to study multimodality.

As far as I understood, Marie-Noelle was not referring to a specific example.

We can feel here how it could be interesting, around pieces of data, to raises different theoretical issues, or explanations.

## What providing access to data means

- Go in depth into discussions about models, what they explained
- Carefully compare previous and new situations
- Limit research cycles which may not be so interesting:
  - Re-inventing the wheel: new techno. environments, new affordances, but..
  - Back to the endless comparison with F2F, with the standpoint that when online you lose things (cf. current papers on webcams, presence, anxiety, etc.)
  - Could we at last reason on new possibilities to discuss and learn in L2 online?

(De Los Arcos, Coleman, Hampel, 2009)

59

What can we gain when giving access to research data?

De Los Arcos paper demonstrate for example that anxiety is not present in the audiographic environment and consequently this type environment needs to be studied as specific topic.

1 2 3 4

Reference corpus & Pedagogical copora

**ANOTHER LIFE FOR LETEC  
DATA (AFTER REUSE FOR CALL RESEARCH)**

60

Organising and publishing data may be useful to CALL research. But we can extend our perspective and see whether it may be of interest more generally in linguistics or for pedagogical motivations.

scientific network:  
<Building & Annotating  
CMC Corpora />  
wiki.itmc.tu-dortmund.de/cmc/

DEUTSCHES REFERENZKORPUS  
DeRiK  
D W I D S  
berlin-brandenburgische  
AKADEMIE DER WISSENSCHAFTEN  
TU  
Technische Universität  
Dortmund  
KOMMUNIKATION

CoMeRe

Salut s que  
COM\_4> c dcd à  
d pr sa cop  
rain?

ETANDEM-NOUS BIEN !

Linguistic perspective: reference corpus

**CORPORA WHICH MAY  
INCLUDE CALL CMC (COMPUTER  
MEDIATED COMMUNICATION)**

61

Let us firstly consider the linguistic perspective, and keep in mind that we all here in CALL have got very rich CMC data.

## Reference corpora of different languages

- **Corpus in German, DWDS**  
*Digitales Wörterbuch der deutschen Sprache*, <http://www.dwds.de/>
- **Corpus in Flemish / Dutch, SoNaR**  
*STEVIN Nederlandstalig Referentiecorpus*
- **Corpus in French (in progress)**
- **Common aims:**
  - Billions of tokens, 500 M structured & annotated (POS), access for linguistic research
  - Extension to Internet communication



62

After the historical project on the English language which led to the creation of the BNC corpus (British National Corpus), during the last 10 years linguists started to develop reference corpora of other European languages started :

- There is a reference corpus of German
- A second one for Flemish Dutch
- Another reference corpus for French is in progress.

Their common features are :

- Large coverage, Billions of tokens, 500 M structured & annotated (POS),
- Provide access for linguistic research
- They seek to develop extensions to Internet communication

# CMC macro and micro structures

**Bud-Spencer-Tunnel**

Die Benennung eines Tunnels in [Schwäbisch Gmünd](#), bei der auch für Bud Spencer-Tunnel gestimmt werden kann, zieht aktuell einige Kreise: [2], genereller [3]. Könnte irgendwo eingebaut werden. Zeugt ja von größerer Beliebtheit unter den Netzbewohnern. --Gormo 12.21, 22. Jul. 2011 (CEST)

Der Abschnitt ist (noch) totaler Käse. Der grundlegende Punkt ist noch nicht mal Fakt: der Name wurde nur vorgeschlagen und das finden viele lustig. Der Abschnitt gehört so wie er ist zum Thema "Glaskugel". [deeleres](#) ansicht 11:43, 24. Jul. 2011 (CEST)

Original data (Wikipedia discussion)

Encoding

```

<listPerson>
  <person xml:id="A01">
    <persName>Gormo</persName>
    <sig>
      <div type="text">
        Gormo
      </div>
    </sig>
  </person>
  <person x...
    <sig>
      <div type="text">
        dee
      </div>
    </sig>
  </person>
  <person>
    <sig>
      <div type="text">
        Ber
      </div>
    </sig>
  </person>
  ...
</listPerson>
<front>
  <timeline>
    <swi...
    <wh...
  </timeline>
  ...
</front>
<body>
  <div type="...
  <div type="...
    <posting synch="#01" who...
    <p>Die Benennung ei...
    </p>
  </div>
  ...
</body>

```

**Three-paper panel:**

**Computer-mediated communication in TEI: What lies ahead**

Proposal for: **The Linked TEI: Text Encoding in the Web.**

2013 Annual Conference and Members' Meeting of the TEI Consortium, 2–5 Oct 2013, Rome.

three paper-panels = 1.5 hours with 3 papers on the same or related topics

**TEI < Text Encoding Initiative >**

CMC macrostructure (type thread)

CMC microstructure (= internal structure of the posting)

elements of microstructure

auto-generated user signature

When colleagues in Linguistics build corpora they systematically structure their data.

Hence when considering CMC data, one of the first thing to do is study their specific micro and macro structures (here Wikipedia forums).

Very often they use the TEI (Text Encoding Initiative) , a standard previously designed for text, then extend to speech. They now want to propose another CMC extension to the TEI.

## Multimodality and CMC ?

The element <posting> is the basic CMC-specific element in our schema. In CMC documents it represents the largest structural unit that can be assigned to one author and one point in time. The category *posting* is defined as a content unit that has been sent to the server "en bloc".

TEI and CMC,  
(Beißwenger et al., 2012)

```
(3) aud, tingrabu [07:20-08:48]: ok hm for me this presentation was hm + become <anno id="an18">too fast</anno> because it's always the same in our architecture school euh we have not time and hm + <anno id="an21" function="form" ntl="gram" type="cf-rpt cf-ack" ref="an19">too quickly sorry</anno> and hm + we cant do good images because euh + euh it's xtime I don't know ++ and euh of course we whole project ++ is about motion and hm we make just some pictures hm statics pictures and hm it's + and it's it's a big matter because hm we always brought about teleportation our + motion is and hm +++ and <anno id="an27" function="form" ntl="lex" type="rpt ack" ref="an29">everyday lack of time ok thank you</anno> xxx and hm this is + this is hm really difficult for us because hm <anno id="an28">we have not enough time</anno> to do good presentation euh in + one night and I hope so tues wednesday could be better + it should be + may be I don't know <anno id="an32" function="form" type="ack" ref="an31">[_chuckles]</anno>
tc, <form> tfrez2, [07:32-07:33]: <anno id="an19" function="form" ntl="gram" type="cf-con" author="tut" ref="an18">it went too quickly?</anno>
tc, <form> tfrez2, [07:38-07:38]: <anno id="an20" function="task" type="cf-con" author="tut" ref="an18">or it was too early in the week?</anno>
tc, <task> romeorez [07:54-07:55]: <anno id="an22" ref="an20">i think it was too early</anno>
tc, <form> romeorez [07:59-07:59]: <anno id="an23" function="form" ntl="typ" type="cf-sr" author="st" ref="an22">too</anno>
tc, <form> tfrez2 [07:59-07:59]: <anno id="an24" function="form" ntl="gram" type="cf-rec" author="tut" ref="an22">too early</anno> <anno id="an25" function="form" type="cf-ref" author="tut" ref="an23"> ok</anno>
tc, <form> tfrez2 [08:08-08:10]: <anno id="an26" function="form" ntl="gram" type="cf-ml" author="tut" ref="an21">too quickly means that you didn't have enough time to speak</anno>
tc, <task form> quentinrez [08:16-08:16]: <anno id="an29" type="cf-pr" author="pr" ref="an28">yes, it's an everyday lack of time</anno>
tc, <task> romeorez [08:43-08:43]: <anno id="an30" ref="an28">that more that we have to show something that we don't really know </anno>
tc, <form> tfrez2 [08:08-08:10]: <anno id="an31" function="form" ntl="gram" type="cf-rec" author="tut" ref="an28 an29">you didn't have enough time</anno>
tc, <task> romeorez [08:43-08:44]: <anno id="an27" ref="an28">fore the shape</anno>
```

(Chanier, Saddour & Wigham, 2012) LETEC corpus





My German linguist colleague, Michael Beißwenger, after studying discussion forums and textchats recently characterized CMC structures.


He sais ...

But in CALL, when we study multimodal CMC the "en bloc" nature does not apply anymore. Here is a transcription of a LETEC corpus which shows interplay between modalities. Since it s not very readable, let us see it differently.



## Modality interplay









1.5 mn video

tingrabu	tfrez2	romeorez	quentinrez
<p>ok for me this presentaiion was become too fast because it's always the same in our architectural school we have not time and too quickly sorry and we can't do good images because it's less time <del>and</del> I don't know [...] and it's a big matter because we always talk about teleportation [...] an everyday lack of time ok thank you quentinrez and this is very difficult [...]</p>	<p>it went too quickly or it was too early in the week? ok too quickly means you didn't have enough time</p>	<p>i think it was to early too</p>	<p>yes, it's an everyday lack of time</p>

\* Paper: (Wigham & Chanier, 2013) CALL journal  
 \* Data: (Chanier, Saddour & Wigham, 2012) LETEC corpus



In one of our paper, which will appear in the CALL journal, and the corresponding data are already online in Mulce, Ciara Wigham discusses the interplay between audio and textchat.

Here is an extract from Archi21. In the left column you have the transcription of the audio of one learner, who presents his feeling related to the on-going process of his architectural project. He is a French native and speaks in English as his L2. In the 3 other columns on the right, you find textchats turns coming from the tutor and two other learners belonging to the same architectural project group.

Let me show you a short video.

\*\*\*\* In this example of conversation doubling, the acts in the text chat respond to the voice chat (blue arrows) but equally acts in the voice chat respond to the text chat (orange arrows) and text chat acts respond to interaction in both voice chat and text chat modalities and prompt interaction in both modalities



Salut s que <NOM\_4> c dcd à ht 1 dvd pr sa cop ki e pa la 2main?

ETANDEM-NOUS BIEN !

Paix historique

SMS / texts  
Tweets  
Blogs  
Forums  
Text chat  
Etc.

CoMeRe.org: CMC corpus in French

CoMeRe: Communication Médinée par les Réseaux)

Forum	Topics	Posts	Last Post
TOPIC #1			
Announcements	Read me first before posting anywhere!	179	264
TOPIC #2			
phpBB Support	Get help with installation and running phpBB 2.0.x here. Please do not post bug reports, feature requests or MOD-related questions here.	236249	1162001
Convertors	Converting from other board software? Good decision! Need help? Have a question about a convertor? Wish to offer a convertor package? Post here. Please post language pack questions to the support forum.	2455	20707
phpBB Discussion	Do not post support requests or bug reports or feature requests. Discuss phpBB here. Non-phpBB related discussion goes in General Discussion!	18931	92214

Thanks to the rich kind of data we find in CALL, we have been able to create with colleagues belonging to a dozen of different linguistic research labs the CoMeRe project.

CoMeRe is the French name given to CMC.

We aim at building a CMC corpus in French and to participate at the same time at a European level, with Michael Beisswenger, to the extension of the TEI to CMC.

1 2 3 4

Example from sports science

# PEDAGOGICAL CORPORA

67

Let us now consider pedagogical applications of LETEC corpora

## Training the pre-service teacher in sport

- Step1: course on building a lesson
- Step2: personal live experience in a school ; record interaction (video) ; reflexion (document)
- Step3: back at university: share experience and reflection (**process not deep enough**)
- Step4 : teacher uses selected data from previous research for cross confrontation

(Researcher in physical activity: N. Gal-Petitfaux, Université Blaise Pascal)<sub>68</sub>

The idea of pedagogical corpora stemmed out of discussions with colleagues who are simultaneously researchers in physical activity and teacher trainers in sport. Here is how they design new procedures to train pre-service sport teacher.

- step1 : the teacher trainer explains how a sport lesson should be designed
- Step 2 : students, organized by couples, have a live experience : one teaches a lesson in a school ; the second one records the lesson
- Step3 : students come back at the university, they share their experience, use their videos. But the reflection process is not deep enough
- hence step 4 : the teacher trainer uses selected data from previous research situation for cross confrontation.

1

2

3

4

Authors : Ciara Wigham, Thierry Chanier

# **PEDAGOGICAL CORPORA CREATED OUT OF LETEC CORPORA**

69

## Starting from a distinguished corpus

Lewis, T. (2006) *When Teaching is Learning: A Personal Account of Learning to Teach Online*. CALICO, Vol 23, No. 3, May 2006. pp 581-600 [http://calico.org/html/article\\_110.pdf](http://calico.org/html/article_110.pdf)

### *When Teaching is Learning: A Personal Account of Learning to Teach Online*

TIM LEWIS  
*The Open University*

---

#### ABSTRACT

This article is aimed at educators who find themselves facing the need to develop their e-teaching skills, with little or no formal training or institutional support.

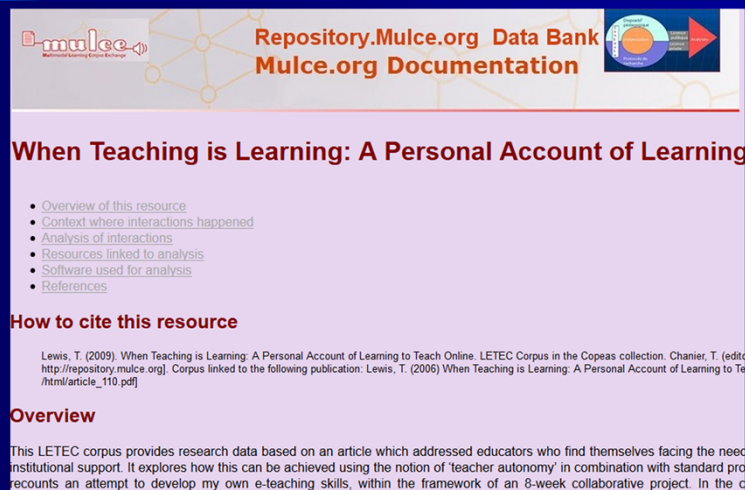
70

To understand the flavour of pedagogical corpora, let us consider a specific situation. In 2006, our colleague Tim Lewis, from the Open University, who have had his first experience as online tutor in a multimodal environment during the Copeas experiment, published afterwards a paper in the CALICO journal. We assembled in a LETEC corpus all the data he gave us, related to this paper : personal diary, videos interviews of learners. We added other data coming from learners reports, discussion forums. Tim was quite pessimistic about the nature of the collaborative process among the learners, and also between himself and the learners. But when you closely look at the data different perspectives appear.

Directly using a LETEC corpus in a training situation is not that easy.

Hence we extracted data, and with Ciara Wigham, we imagine a pedagogical corpus.

## Starting from a distinguished corpus




The screenshot displays the Repository.Mulce.org Data Bank interface. At the top, there is a header with the Mulce.org logo and the text 'Repository.Mulce.org Data Bank' and 'Mulce.org Documentation'. Below the header, the title of the resource is 'When Teaching is Learning: A Personal Account of Learning'. A list of links is provided: 'Overview of this resource', 'Context where interactions happened', 'Analysis of interactions', 'Resources linked to analysis', 'Software used for analysis', and 'References'. Under the heading 'How to cite this resource', a citation is shown: 'Lewis, T. (2009). When Teaching is Learning: A Personal Account of Learning to Teach Online. LETEC Corpus in the Copeas collection. Chanier, T. (ed). <http://repository.mulce.org/>. Corpus linked to the following publication: Lewis, T. (2006) When Teaching is Learning: A Personal Account of Learning to Teach Online. [http://repository.mulce.org/html/article\\_110.pdf](http://repository.mulce.org/html/article_110.pdf)'. The 'Overview' section begins with the text: 'This LETEC corpus provides research data based on an article which addressed educators who find themselves facing the need for institutional support. It explores how this can be achieved using the notion of 'teacher autonomy' in combination with standard pedagogical practices. It recounts an attempt to develop my own e-teaching skills, within the framework of an 8-week collaborative project. In the c...

71



To understand the flavour of pedagogical corpora, let us consider a specific situation. In 2006, our colleague Tim Lewis, from the Open University, who has had his first experience as online tutor in a multimodal environment during the Copeas experiment, published afterwards a paper in the CALICO journal. We assembled in a LETEC corpus all the data he gave us, related to this paper : personal diary, videos interviews of learners. We added other data coming from learners reports, discussion forums. Tim was quite pessimistic about the nature of the collaborative process among the learners, and also between himself and the learners. But when you closely look at the data different perspectives appear.

Directly using a LETEC corpus in a training situation is not that easy.

Hence we extracted data, and with Ciara Wigham, we imagine a pedagogical corpus.

5 mn video 

# Lead-in document


**PEDAGOGICAL CORPUS**  
Reflective teaching journals

Original project (distinguished corpus) : mce-copeas-reflexive-tutor


**Reflective teaching journals**

- Learning and Teaching Corpus, Copeas, EFL tutor, environment : *Lyceum*
- Description: This pedagogical corpus questions whether, in a reflective approach, tutors should base their reflections on teaching journals and how we can define whether an activity is collaborative or not.
- Sources: interaction data from Lyceum, selected text from an article published in Calico (Lewis, 2006), notes from tutor's teaching journal, tutor post-course interview data

Date of recording: 2013-04-25      Length : 5m03s



©(Wigham, C.R. & Chanier, T. 2013)

Licence <http://mulce.org> 

72

Here is an extract of a video which has been created as a lead-in document for the new pedagogical corpus.

The screenshot shows the Mulce website interface. At the top, there are logos for 'mulce Multimodal Learning Corpus Exchange', 'Pedagogical Corpora', and 'UNIVERSITE BLAISE PASCAL CLERMONT UNIVERSITE BP'. Below the logos is a navigation menu with four items: 'Overview', 'Reflection introduction', 'Reflective tasks - collaboration', and 'Reflective tasks - journals'. The main content area is titled 'LETEC Pedagogical Corpus: Reflective Teaching Journals' and contains a list of links: 'Objectives', 'Learning outcomes', 'What are you going to do with this pedagogical corpus?', 'Context of online language course from which this pedagogical corpus is built', 'Technical requirements', and 'LETEC corpora references'. Below this is a section 'How to cite this pedagogical corpus' with a citation: 'Wigham, C.R. & Chanier, T. (2013) Pedagogical corpus: Reflective Teaching Journals. Mulce.org : Clermont Université. [oai : mulce.org:mce-peda-rtjournals ; <http://repository.mulce.org>].'. The next section is 'Target users' with the text: 'This pedagogical corpus, its resources and tasks have been designed for pre-service language teachers and teacher trainers. Whilst the activities are in English, in this version of the corpus (July 2013) the resources are in French. Target users must thus be proficient in the French language.' At the bottom of the page, the same citation is repeated, and the number '73' is visible in the bottom right corner.

The pedagogical corpus can be downloaded out of Mulce repository.

It uses data from the Copeas experiment and offer new activities around the perception of collaborative process as a language tutor.

*It aims at*

- *identify language tutors' and students' differing views of successful collaboration*
- *summarize the characteristics of successful collaboration and produce a list of implications for practice*
- *appraise the advantages of keeping a teaching journal*
- *compare and contrast reflections from a teaching journal with naturally occurring data (interaction tracks) and researcher-provoked data (student feedback) to analyse whether teachers should base reflections about teaching practice solely on journal entries and personal reactions*



**mulce** Multimodal Learning Corpus Exchange

**Pedagogical Corpora**

Overview Reflection introduction Reflective tasks - collaboration

## Collaboration - Introduction

### Activity 1: Personal definitions

(Allow 10 minutes for this activity. Individual activity)

In your notebook write down the first five words that spring to mind when you think about each of 'collaboration', 'cooperation' and 'community'.

### Activity 2: Personal interpretations of successful collaboration

(Allow 10 minutes for this activity. Individual activity)

Think back and choose a time when you had to collaborate with other people during a learning activity organised an activity in which you asked students to collaborate. What elements meant that the collaboration was successful, or not?

Use your notebook to write down the characteristics of a collaborative learning activity that you would consider successful. If your situation is not an online situation, do you imagine that the same characteristics would apply to an online situation?

Wigham, C.R. & Chanier, T. (2013) Pedagogical corpus: Reflective Teaching Journals. Mulce.org : Clermont Université. [oai : mulce.org:mce-peda-rtjournals ; <http://repository.mulce.org>]

74

The pedagogical corpus can be downloaded out of Mulce repository.

It uses data from the Copeas experiment and offer new activities around the perception of collaborative process as a language tutor.

*It aims at*

- *identify language tutors' and students' differing views of successful collaboration*
- *summarize the characteristics of successful collaboration and produce a list of implications for practice*
- *appraise the advantages of keeping a teaching journal*
- *compare and contrast reflections from a teaching journal with naturally occurring data (interaction tracks) and researcher-provoked data (student feedback) to analyse whether teachers should base reflections about teaching practice solely on journal entries and personal reactions*

1 2 3 4

**CALL journals and research data**  
Survey addressed to members of EUROCALL and/or participants to the EURO contents relates to my talk during the conference. The survey is anonymous. I will list at the end of September. Thank you for your participation. Thierry Chanier

1. Are you a teacher of English?  
1.  yes  
2.  no

Survey on CALL journals and research data :  
- Link in the main editorial article on :  
<http://mulce.org>  
- Questions 10 to 17

OpenData  
**OPEN ACCESS TO  
PUBLICATIONS & DATA**

75

4<sup>th</sup> section , last subject

## Enclosing the Commons of the Mind

- I seriously doubt that we would create the Web today—at least if policy makers and market incumbents understood what the technology might become early enough to stop it. (p.278)
  - Almost everything on the Internet is copyrighted, even if its creators do not know that and would prefer it to be in the public domain. (p. 26)
- (Boyle, J.2008, *The Public Domain: Enclosing the Commons of the Mind*) Boyle is one of the creators of the Creative Common – CC project

76

Let us start with a warning coming from James Boyle, one of creator of Creative Common project. Many people would like to enclose the Commons of the Mind as he says.

I seriously ...

And he gave a second warning about the necessity of paying attention to licences even when looking at public domain issues.

1

2

3

4

Chanier, T. "Commentary: Open Access to Research and the Individual Responsibility of Researchers". *Language Learning & Technology*, vol. 11, 2 (2007).

Open archives

**FREE AND IMMEDIATE  
ACCESS TO PUBLICATIONS  
(ONCE ACCEPTED BY REVIEWERS)**

77

Let us briefly look at open access to publications, subject I already detailed in LLT in 2007..

## Guidelines for researchers (EU level)

- “The Commission proposes to make open access to scientific publications a general principle of Horizon 2020, building on the already existing activities in FP7 (e.g. eligibility of open access publishing costs, embargo for 'Green' open access of six to twelve months).

- there should be open access to publications resulting from publicly funded research as soon as possible, preferably immediately and in any case no later than six months after the date of publication, and twelve months for social sciences and humanities;
- licensing systems contribute to open access to scientific publications resulting from publicly-funded research in a balanced way, in accordance with and without prejudice to the applicable copyright legislation, and encourage researchers to retain their copyright while granting licences to publishers;

[http://ec.europa.eu/research/science-society/document\\_library/pdf\\_06/background-paper-open-access-october-2012\\_en.pdf](http://ec.europa.eu/research/science-society/document_library/pdf_06/background-paper-open-access-october-2012_en.pdf)

78

Here is a recent statement from the European commission which give guidelines to researchers for the “Horizon 2020” framework (recently opened).

The report says : there should be open access to publications resulting from publicly founded research as soon as possible, and forms should be changed in order to let researchers retain their copyright while granting licences to publishers. These are 2 different issues. Open access needs not wait for the new wording of copyright forms.

The screenshot shows the website interface for Tematic FMSH. At the top, there are navigation tabs: Accueil, Mon espace, Déposer, Consulter, Rechercher, and Services. The main content area displays search results for 'Chanier'. A red box labeled 'National repository' highlights the search path. Another red box labeled 'Insitutional repository' highlights the DOI link and the abstract text. A red arrow points from the 'Insitutional repository' box to the 'Click here to request a copy from the OU Author.' link. The abstract text reads: 'This paper addresses the lack of formalised methodology for analysing learner interaction data created in conversation on audiographic platforms. First the author shows the importance of conversations in language learning and the need for researchers to understand how users learn from these interactions. Then the author establishes that appropriate methodologies for investigating interaction data collected from online platforms have as yet emerged neither from the...'

There exist a huge literature (open access) on this topic, which details the main ways of providing open access to publications, the so-called Green and Gold roads. As regards CALL journals and the Gold Road, Language Learning & Technology (LLT) and Alsic were the first to provide open access to their articles. Recently CALICO reduced its moving wall to one year (one year after their publication, articles become open access).

Let me just recall here, it is firstly the author responsibility to deposit her/his article once it has been accepted by reviewers in open archives (the green road), whichever kind of archive it is, national or institutional.

1 2 3 4

OpenData

**OPEN ACCESS TO RESEARCH  
DATA**

*OER : Open Educational Ressources are important, but not  
considered here*

80

Let us spend a little more time on access to research data,  
to the new concept of OpenData

# Opendata

- Term which is starting to be widely used with different aims in mind, among other things:
  - 1) Academic world: share research results
  - 2) Government and public institutions: open their data to the public
- Here we mainly consider the 1<sup>st</sup> perspective

81

This term starts being widely used with different aims in mind:


- In the academic world it refers to the way of sharing research data
- For government and public institutions, they started opening their data to the public.

This is of course the first case we will here consider.



## Opendata def

- "Open data is data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike." OpenDefinition.org



# Opendata criteria

- **"Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Reuse and Redistribution:** the data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets. The data must be machine-readable.
- **Universal Participation:** everyone must be able to use, reuse and redistribute – there should be no discrimination against fields of endeavor or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed. "OpenDefinition.org

83

There exist 3 main criteria that research data should follow in order to be considered OpenData.

Besides being obviously available, the interesting perspective is the fact that data can be access in order to be reuse and mix with other data.

Second interesting point is that the constraints for reuse should be reduced to a minimum, then the definition stipulate that non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes are not allowed


## Why should we use licences?

- "In most jurisdictions there are intellectual property rights in data that prevent third-parties from using, reusing and redistributing data without explicit permission. Even in places where the existence of rights is uncertain, it is important to apply a license simply for the sake of clarity. Thus, **if you are planning to make your data available you should put a license on it** — and if you want your data to be open this is even more important."  
OpenDefinition.org

84

Another interesting point is also the fact that authors should always put a licence on data, when they plan to make their data available.

## Example of licences on learner corpora: ICLE



**International Corpus of Learner English V2**  
**(Handbook + CD-ROM -single user)**  
Sylviane GRANGER, Estelle DAGNEAUX, Fanny MEUNIER, Magali PAQUOT  
Presses universitaires de Louvain • Hors collection (Presses universitaires de Louvain)

**Paperback + CD-Rom - In English 272.25 €** [Buy](#)

The International Corpus of Learner English (Version 2) is a corpus of writing by higher intermediate to advanced learners of English. It contains 3.7 million words of EFL writing from learners representing 16 different mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana). It differs from the first version published in 2002 not only by increased size and range of learner populations, but also by its interface, which contains two functionalities: built-in concordancer allowing users to search for word forms, lemmas and/o

[Licence Agreement \(PDF 137 KB\)](#)  
comment

- No access given on the website, except « pay to look at »
- Nothing about reuse, mixing, etc.

85

What about access to data coming from the language learning fields. Here is the example of the most well known learner corpus, namely the International Corpus of Learner English.

On the website, no access is given, except the “pay to look at”. Nothing is mentioned about reuse and mixing.

## Example of licences on learner corpora: ELFA

LICENSE AGREEMENT

<https://elomake.helsinki.fi/lomakkeet/43518/lomake.html>

### CLARIN RESTRICTED END-USER LICENCE (RES)

**Copyright holder:** Anna Mauranen; The Faculty of Arts, PO Box 3, 00014 University of Helsinki; anna.mauranen@helsinki.fi

**Resource:** ELFA – English as a Lingua Franca in Academic Settings

The Copyright holder grants the End-User a free, non-exclusive and perpetual (for the duration of the copyright) right to **use and make copies of the Resource for personal use** as such, as modified, or as part of a combined work. The permission applies to all known modes and means of communication and includes a right to make modifications enabling the use of the Resource on other devices and in other formats. The purpose of the Resource use must be outlined in the research plan. Distribution of copies is not allowed.

The Resource may be used for publishing articles and analyses in scientific journals, conference presentations or other similar forums. In such publications, proper reference to the resource should be given. [INF: If the Resource is used as material for a scientific publication, the Copyright holder is to be informed.]

[PD: Due to the nature of the material, the Resource should be handled with care in order to respect the privacy of the personal data. If samples of the data are published, they must be anonymized according to best practices.]

[NC: The Resource may not be used for gaining economic benefit. Government-funded or non-profit research projects, e.g. projects funded by Academy of Finland or TEKES, are not regarded as gaining economic benefit even if a portion of the financing is contributed by companies.]

[ReD: A revised or enhanced version of the Resource may be redeposited with the CLARIN Service. The conditions for redepositing will be agreed separately.]

- Open access, but for personal use (hence not for research)
- Important restriction (NC), where are the sound files?

86

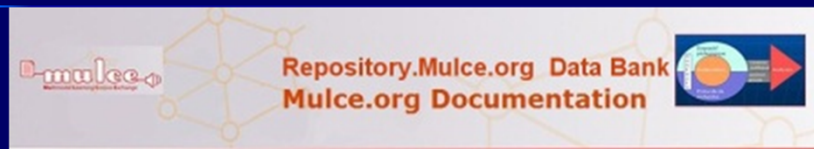
Here is another learner corpus, the ELFA (English as a Lingua Franca in Academic Settings).

Creators made some progress.

In the earlier license, users were being charged over 100 euros for a mere six-month license (and just for the text corpus, not the audio), with instructions to "destroy" the files at the end of this period or purchase a new license!

Now text data are open access but for personal use only (i.e. not here for example). They are important restrictions in the license and still no access is given to the audio.

# Open access, ethics and licence



Open Data:  
<http://opendefinition.org/guide/>



For usage:  
licence




For participants:  
Informed  
consent form  
+  
Anonymization  
process

Note : Incoherence between licences on our sites: changes are not yet achieved

87

As regards the LETEC corpora and Mulce repository, all our corpora are of course open access , without any registration. In each corpus we included the informed consent form signed by participants. We started to change our licence in order to become fully compliant with OpenData criteria.

# Usual CC (open but not necessarily compliant with OpenData)

 <p><b>Attribution</b> CC BY</p> <p>Cette licence permet aux autres de distribuer, remixer, arranger, et adapter votre œuvre, même à des fins commerciales, tant qu'on vous accorde le mérite de la création originale en citant votre nom. C'est le contrat le plus souple proposé. Recommandé pour la diffusion et l'utilisation maximales d'œuvres licenciées sous CC.</p> <p><a href="#">Voir le Résumé Explicatif</a>   <a href="#">Voir le Code Juridique</a></p>	 <p><b>Attribution - Partage dans les Mêmes Conditions</b> CC BY-SA</p> <p>Cette licence permet aux autres de remixer, arranger, et adapter votre œuvre, même à des fins commerciales, tant qu'on vous accorde le mérite en citant votre nom et qu'on diffuse les nouvelles créations selon des conditions identiques. Cette licence est souvent comparée aux licences de logiciels libres, "open source" ou "copyleft". Toutes les nouvelles œuvres basées sur les vôtres auront la même licence, et toute œuvre dérivée pourra être utilisée même à des fins commerciales. C'est la licence utilisée par Wikipédia ; elle est recommandée pour des œuvres qui pourraient bénéficier de l'incorporation de contenu depuis Wikipédia et d'autres projets sous licence similaire.</p> <p><a href="#">Voir le Résumé Explicatif</a>   <a href="#">Voir le Code Juridique</a></p>
 <p><del><b>Attribution - Pas de Modification</b> CC BY-ND</del></p> <p><del>Cette licence autorise la redistribution, à des fins commerciales ou non, tant que l'œuvre est diffusée sans modification et dans son intégralité, avec attribution et citation de votre nom.</del></p> <p><del><a href="#">Voir le Résumé Explicatif</a>   <a href="#">Voir le Code Juridique</a></del></p>	 <p><b>Attribution - Pas d'Utilisation Commerciale</b> CC BY-NC</p> <p>Cette licence permet aux autres de remixer, arranger, et adapter votre œuvre à des fins non commerciales et, bien que les nouvelles œuvres doivent vous créditer en citant votre nom et ne pas constituer une utilisation commerciale, elles n'ont pas à être diffusées selon les mêmes conditions.</p> <p><a href="#">Voir le Résumé Explicatif</a>   <a href="#">Voir le Code Juridique</a></p>
 <p><b>Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions</b> CC BY-NC-SA</p> <p>Cette licence permet aux autres de remixer, arranger, et adapter votre œuvre à des fins non commerciales tant qu'on vous crédite en citant votre nom et que les nouvelles œuvres sont diffusées selon les mêmes conditions.</p> <p><a href="#">Voir le Résumé Explicatif</a>   <a href="#">Voir le Code Juridique</a></p>	 <p><del><b>Attribution - Pas d'Utilisation Commerciale - Pas de Modification</b> CC BY-NC-ND</del></p> <p><del>Cette licence est la plus restrictive de nos six licences principales, autorisant les autres à télécharger vos œuvres et à les partager tant qu'on vous crédite en citant votre nom, mais on ne peut les modifier de quelque façon que ce soit ni les utiliser à des fins commerciales.</del></p>

Here is the listing of the Creative Common licences, created by James Boyle and his colleagues. 2 of them are not OpenData compliant, because they forbid direct use for commercial purposes. In Mulce, we started to switch from the BY-NC-SA to the BY only, i.e. the only obligation of to refer to the authors and the original work.

## 2 licences on data fully compliant with OpenData

- CC0 : As creators, I may have had some rights (rights on models, rights on data, etc.) on the work and I waive them (permanent , irrevocable)
- PDDL : I do not even mention the fact that I may have had rights over something

### About CC0 — “No Rights Reserved”



CC0 enables scientists, educators, artists and other creators and owners protected content to waive those interests in their works and thereby place possible in the public domain, so that others may freely build upon, enhance and reuse the

Conformant Data Licenses

License	Domain	By	SA	Comments
Open Data Commons Public Domain Dedication and Licence (PDDL)	Data	N	N	Dedicate to the Public Domain (all rights waived)
Open Data Commons Attribution License	Data	Y	N	Attribution for data(bases)
Open Data Commons Open Database License (ODbL)	Data	Y	Y	Attribution-ShareAlike for data(bases)
Creative Commons CCZero	Content, Data	N	N	Dedicate to the Public Domain (all rights waived)

The CC-BY licence is an improvement. But it still restricts possibilities for mixing and reusing data. There exist 2 licences which are fully compliant: the PDDL and CC0 ones. These licences make a significant step forward. They waive intellectual property rights, data then become part of public domain.



- What will happen if the attribution licence is not there anymore?
- I may not be cited?

90

When hearing this you may be afraid or at least sceptical:

- What will happen if the attribution licence is not there anymore?
- I may not be cited?

## No confusion between attribution(IPR) and citation-references

- We give users the way to refer to our work (metadata : OLAC – bibliographicCitation) and will use this in our list of publication & works. For exemple:
  - 1) creator of the corpus
    - Wigham, C.R. (2013). *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>]
  - 2) creator and editor
    - Stahl, Gerry ; Weimar, Steve ; Shumar, Wes (2009). *LETEC Corpus Virtual Math Team*. Reffay, C. (editor). Mulce.org : Clermont Université. [oai : mulce.org:mce-vmt-letec-teamc ; <http://repository.mulce.org>]

91

But we should not be afraid.

We have the habits of confusing 2 very different issues IPS and citation-references:

- The first one only refer to legal issues : “you did not cite me, I am going to take you before the court!”
- In the second one, we have our academic procedure. We need to refer to previous work and when authors do not do it properly, their work is rejected by peers-reviewers.

Hence we need only worry about correctly referencing our work and making this reference clearly available.

Here are for example two references toLETEC corpora, the first one to the author, only creator of the corpus ; the second one where creators and editors are distinguished (like a chapter of a book).

Moreover these references are tagged as such, included in metadata which are harvested on Internet thanks to the OLAC harvesting protocol.



## Recommendations

If we want to be connected to Digital Humanities

### ■ Actions

- Open our data (provided that ethics is OK – anonymisation)
- Choose licences with the fewest restrictions
- Cite others and your data as bibliographic references
- List them in your work

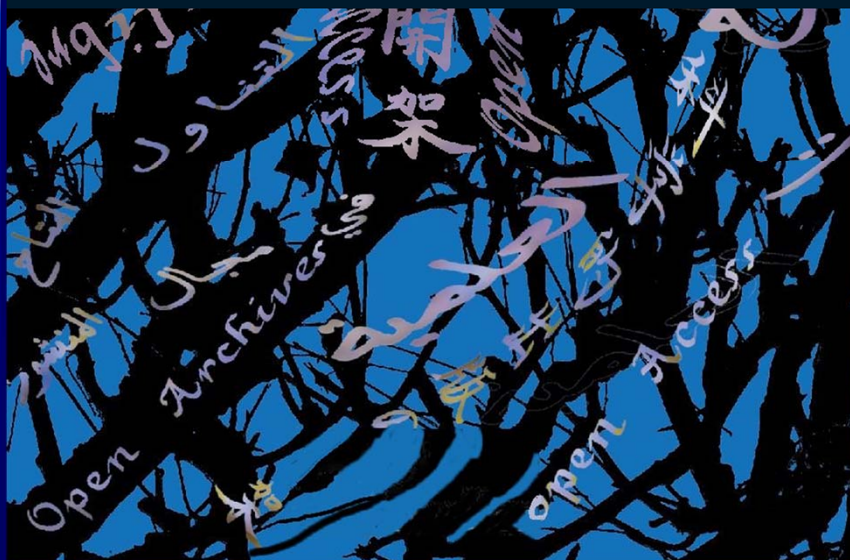
### ■ Implications

- Acknowledgement will come (from institutions, other colleagues)
- CALL research will progress (re-analysis, coverage extended with mixing)
- CALL data will be reused by other fields

Thank you for your attention!

Thierry.chanier at univ-bpclermont.fr

<http://lri.univ-bpclermont.fr/spip.php?rubrique98>



93