



UMR 5191 - CNRS / Université Lyon 2  
Interactions, Corpus, Apprentissages, Représentations

Paris PFC, 7.12.2011

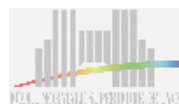
# ***Corpus oraux: multiplier les interfaces*** ***Présentation de la base de corpus CLAPI***

## **Lorenza Mondada**

laboratoire ICAR(CNRS, Lyon/France) & IUF

avec la collaboration du groupe ICOR

S. Bruxelles, C. Etienne, E. Jouin, J. Lascar, L. Mondada, V. Traverso, D. Valero



# Corpora of talk in interaction

- **Currently, only few big corpora integrate large naturally occurring interactions** (Talk Bank [McWhinney], Santa Barbara Corpus of Spoken American English [Du Bois], MOVIN corpus [Wagner], Institut für deutsche Sprache corpus [Depperman/Hartung])
- In traditional big corpora (e.g. BNC), the spoken part is largely less developed than the written part. Moreover, often, transcripts only are provided (no audio/video signal) and they are very superficial
- Spoken corpora raise specific technological and theoretical problems, challenging written norms and current standards
- **The available typology and varieties of naturalistic data of social interactions are very limited**
- Very few naturalistic corpora exist for French (most spoken corpora are constituted by sociolinguistic interviews or various kinds of experimental data)

*« The essential characteristics of corpus-based analysis are :*

- it is empirical, analyzing the actual patterns of use in natural texts ;*
- it utilizes a large and principled collection of natural texts, known as a « corpus », as the basis for analysis ;*
- it makes extensive use of computers for analysis, using both automatic and [human-computer] interactive techniques ;*
- it depends on both quantitative and qualitative analytical techniques »*

*(Biber et alii, 1998, 4)*

# A long tradition of studies on interaction at ICAR

- Research lab specialized on verbal interaction since the '80s
- Beginning of a research tradition on interaction in France in the 80s: Bange, de Gaulmyn, Cosnier, Kerbrat-Orecchioni
- Later on, international reference for interactional linguistics and conversation analysis (Mondada, Traverso), including the study of plurilingual interactions
- Also international reference for the use of video in science education (Tiberghien, since the 80s) and in France for the study of classroom interaction (Bouchard)
- Important projects on workplace interactions - within cognitive studies (Lund), activity theory (Grosjean), conversation analysis (Mondada, Traverso) - also taking into consideration technologically mediated interactions
- Tradition of work on gesture (Cosnier) and now international centre for video analysis and multimodal analysis (Mondada)
- Development of data bases, first for archiving purposes, and more and more for research purposes (TAL, computer linguistics, corpus linguistics)

**Interactional linguistics**  
**A multi-dimensional approach:**  
**prosody,**  
**phonetics**  
**syntax,**  
**lexis,**  
**pragmatics,**  
**gesture,**  
**Embodiment**  
→  
**Spoken data**  
**Multimodal resources**

# The CLAPI platform

- CLAPI is developed since 1999 at the ICAR research lab, by the ICOR group
- Internet acces: <http://clapi.univ-lyon2.fr>

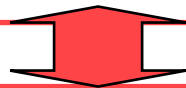
## Médiathèque

archive of original tapes and materials

Corpus data bank

## STORAGE

historical corpora	golden standard	private workspace
--------------------	-----------------	-------------------



Browsers, search tools

## TOOLS

**Interoperability** with other data bases archiving big corpora of spoken language in interaction



# CLAPI - a data base of corpora of French in interaction (Corpus de LAngue Parlée en Interaction)

- A openly accessible database for the study of spoken language interactions:
  - CLAPI is available on-line: <http://clapi.univ-lyon2.fr>
  - A related website for interactional research:  
CORINTE: <http://icar.univ-lyon2.fr/projets/corinte>
- Archive: **600h of audio-video data**
- Data base: **135h of stored data** (300 recordings, 500 transcripts)
- Browsable data: **50h = 500.000 words in XML format** (125 transcripts)
  - Aligned transcripts (text / sound-video) (streaming)
  - Original transcripts with different conventions
    - The original form is conserved but a XML version allows for searches
  - 75 descriptors as metadata
  - Diversity of documented social interactions
- Standards & compatibility
  - Proper XML Schema -> on-line browsable transcripts
  - Metadata: Dublin Core & OLAC compatible

# Website CLAPI : <http://clapi.univ-lyon2.fr>

## Bienvenue dans CLAPI banque de données et plateforme logicielle

Accueil

Présentation des corpus

Présentation des transcriptions

Outils de requêtes

Identification des formes

Gestion de vos collections

Changement de compte

Contact

### CLAPI



CLAPI est une banque de données outillée de Corpus de **L**Angue **P**arlée en Interaction enregistrés **en situation réelle**, dans des contextes variés : Interactions professionnelles, institutionnelles ou privées, commerciales, didactiques, médicales , ...

### VOLUME

45 corpus, 327 enregistrements (135 h), 514 transcriptions documentés par 75 descripteurs, 58 h balisées pour des traitements d'analyses et de requêtes

### PLATEFORME LOGICIELLE

--> Un ensemble d'outils d'**analyse automatique des données** : fréquences, co-occurrences, répétitions, ...  
--> Un outil de **requêtes complexes** pour mettre en évidence les corrélations entre tokens et phénomènes interactionnels  
--> Un concordancier aligné avec le signal par **streaming**

audio/vidéo  

### CONDITIONS D'ACCES

**==> 35 h de données, soit plus de 70 % de la base en accès libre pour les analyses et les requêtes**

**==> 100 transcriptions documentant une grande diversité de situations interactionnelles**

16 h de données téléchargeables (transcriptions, enregistrements)

Accès libre sur la totalité de la base aux éléments suivants : Descripteurs, Lexique, Fréquence des tokens, Synthèse des transcriptions

### CORPUS D'INTERACTIONS

Nous vous invitons à consulter [le site associé CORINTE](#), développé au sein du groupe ICOR, dédié à la recherche multimodale sur **CORpus** en **INTER**action

### CONVENTION ICOR

Vous pouvez télécharger [la convention ICOR](#) avec son formalisme et des exemples d'utilisation

### INTEROPERABILITE

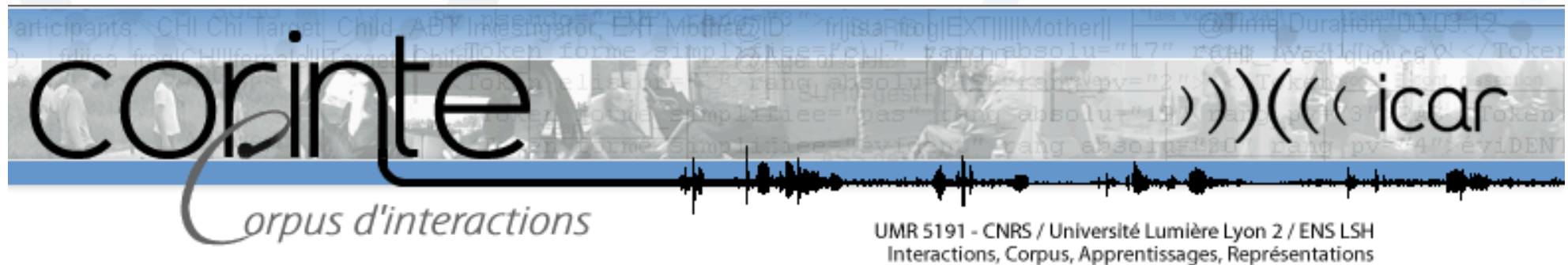


Un export des descripteurs des corpus, des enregistrements, des transcriptions est disponible en **Dublin Core- OLAC**

Un export des transcriptions (descripteurs et version textuelle de la transcription) est disponible en **TEI**

Le groupe ICOR, Interaction **CORpus**, est un collectif qui conçoit et développe la plateforme CLAPI et le site CORINTE. Actuellement, il est composé de Michel BERT, Sylvie BRUXELLES, Carole ETIENNE, Emilie JOUIN, Justine LASCAR, Lorenza MONDADA, Christian PLANTIN, Sandra TESTON, Véronique TRAVERSO, Daniel VALERO.

Website CORINTE : <http://icar.univ-lyon2.fr/projets/corinte>



[Recueil](#) ▶ [Confection](#) ▶ [Intégration](#) ▶ [Diffusion](#) ▶ [Analyse](#)

[Accueil](#)

Imprimer

## Présentation

CORINTE est un site dédié à la recherche sur les corpus de langue parlée en interaction. Il fonctionne en corrélation avec la base de données outillée CLAPI (Corpus de Langue Parlée en Interaction).

Le site CORINTE présente une *linguistique de l'interaction*, fondée sur l'enregistrement, la transcription et le traitement de données orales interactionnelles et multimodales.

Il comporte:

- des réflexions théoriques et méthodologiques sur la constitution de corpus d'interactions et sur leur analyse ainsi que sur la construction et l'exploitation de banques de données de corpus de langue parlée en interaction.

### Accueil

[Démarche scientifique](#)

[Programmes de recherche](#)

[Organigramme](#)

[CLAPI](#)

[Questions juridiques](#)

[Actualités](#)

[Conventions de transcription](#)

[Table des matières](#)

[FAQ](#)

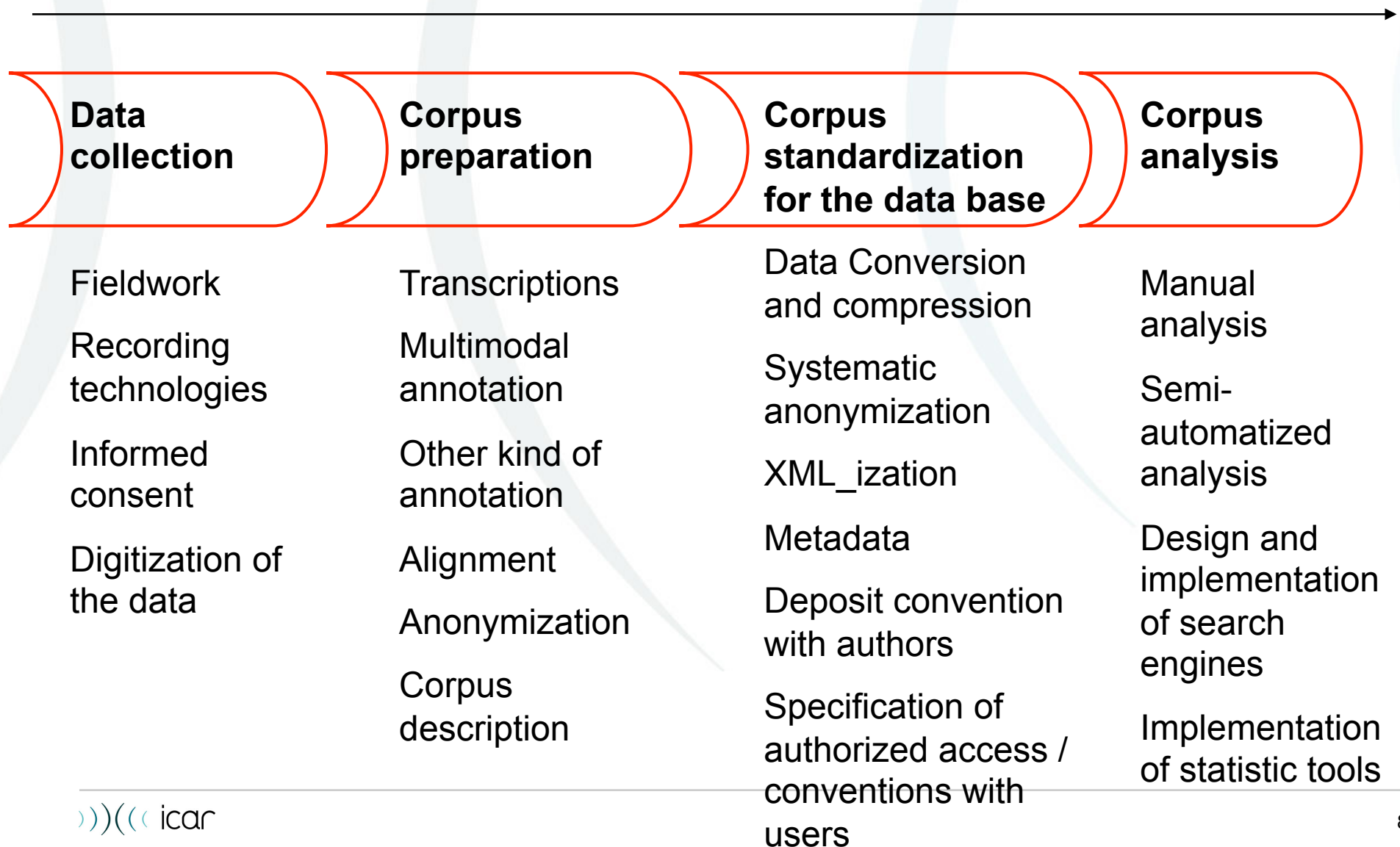
[Ressources](#)

[Glossaire](#)

[Liens](#)

[Contacts](#)

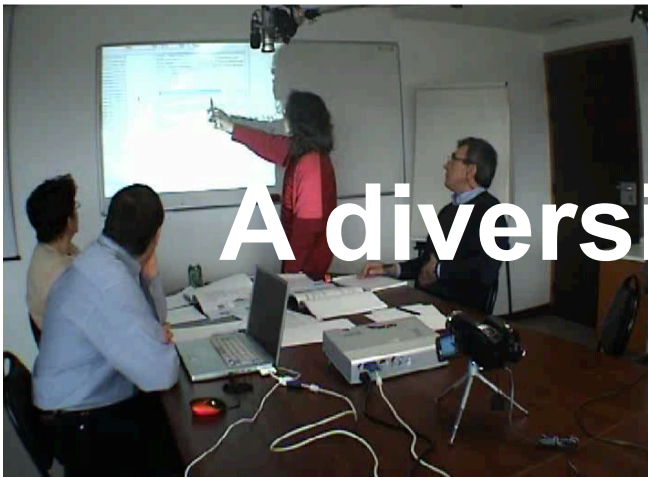
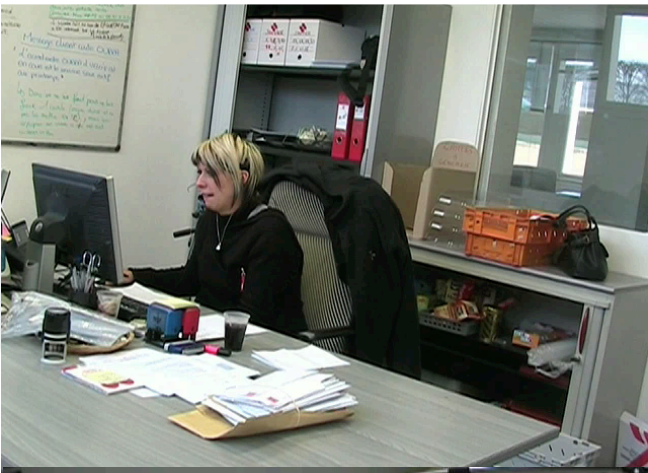
# The production of corpora: a long process





DATA





A diversity of data and settings

# Data in CLAPI

- Specificity: **naturalistic** data
  - vs. interviews, vs. experiments, ....
  - No media data (too easy to obtain + legal problems)
- Not only audio, but also **video** data
- **Diversity** of data and social settings
  - (vs. representativity)
  - Some 'rare' settings (e.g. dentist, divorce negotiations, workplace interactions)
  - Rich collections of ordinary data (e.g. dinner conversations)
- **Heritage data**
  - long tradition in Lyon, emeriti, ex doctoral students
  - donations from other labs, external researchers, other emeriti
- **Newest data**
  - Produced through sophisticated technologies
  - Multi-scope data (multi-sources, synchronization of multiple data)
  - Technological platform for video recordings
  - They constitute the Golden Standard Corpus

- Another data base at ICAR: VISA (specialized for classroom data)

# DATA collected within CIEL

- A different corpus design than CLAPI' s heritage archive
- Aims: comparison of resources, sequences and actions across different points within the Francophonie
- Two principles:
  - 1. Activity types
    - Dinner conversations
    - Work interactions (EITHER with customers, patients, etc., e.g. commercial interactions OR among professionals, e.g. meetings)
    - Radio debates and phone ins (local radios)
  - 2. Areas
    - Europe: F, B, CH
    - Americas: CAN, Antilles
    - Asia: Ile Maurice, La Réunion
    - Africa: Egypt, Algeria, Lebanon, Senegal, Ivory Coast, Cameroun



# ENRICHED DATA: the corpus as a complex object

# Corpus: a complex, multiple object

A complex object:

Signal (A/V/txt/image)

Primary data

Alignment (synchronization)

Secondary data

Transcripts

Annotations

+ Metadata

The screenshot shows a software interface with several panels:

- Top Panel:** Menus for 'Fichier', 'Edition', 'Annotation', 'Affichage', and 'Aide'. A file browser shows a folder named 'BEA' containing a list of items with numbers and phonetic transcriptions:
 

Nr	Transcription
1	[ah d'accord/
2	[super/
3	hm
4	hum
5	d'acco°rd°
- Video Panel:** A central window showing a fisheye view of a car's interior with two people, a driver and a passenger.
- Timeline Panel:** A horizontal timeline with a playhead at 00:00:01.410. Below it are two audio waveforms.
- Annotation Panel:** Multiple tracks showing linguistic and behavioral data:
  - RIC (Transcripts):** Text segments like 'le: bouton que vous (0.6)', 'qui est en p (0.7)', 'lui:/ bascu\*lera sur l'cé d (0.', 'aujourd'hui on a rien d'...'.
  - ricH (Annotations):** Segments like 'hand forth', 'points', 'h back'.
  - beaG (Annotations):** Segments like 'looks in front', 'looks at button', 'l. dashb', 'looks at button', 'looks at dashboard'.
  - ricG (Annotations):** Segments like 'looks at push\_button/steering', 'looks in front dashboard', '..... looks at bea', 'looks in fr'.
  - beaH (Annotations):** Segments like 'home position', 'h forth', '\*touches-pushes', 'maintains hand o'.
- Right Panel:** A table with columns 'Temps de dé...' and 'temps de fin' containing time intervals:
 

Temps de dé...	temps de fin
00:00:06....	00:00:06....
00:00:07....	00:00:07....
00:00:07....	00:00:08....
00:00:08....	00:00:08....
00:00:09....	00:00:09....



From the labile event...

... to primary data



.. to secondary data

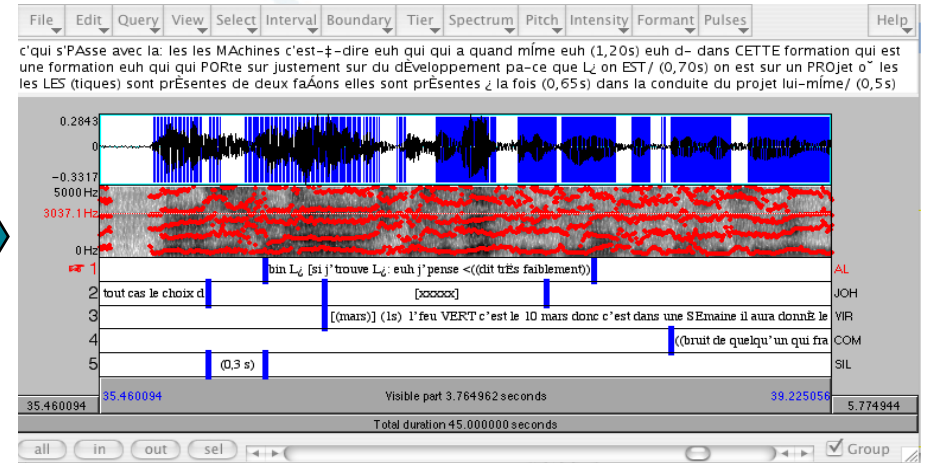
```

9 Hen ah mais c'est d'la coppa\
10 Val mhm
11 (1s)
12 Val tu [passes à papa/]
13 Yan [xxxx xxxxxxxx] à Rochefort/
14 je suis allé visiter le: ils^ont r'construit
15 le ba[teau:/ . de de de:: Lafayette/
16 Val [fais passer le plat . Henri:/]
17 Yan quand Lafayette est parti aux [Etats-Unis/
18 Hen [ah ouais]
19 (1s)
20 Yan c'est l'hermione .. j'ai ramené
21 des photos d'ailleu[rs\
22 Val [c'est papy qu'a été à:/
23 Hen il est chaud mon verre\

24 Val [ouais
25 Yan [et ils reconstruisent exactement le bateau/
26 . tel qu'il a été fait à l'époque\
  
```

# Primary and secondary data

- Transcribing involves much more than
  - audio/video data
  - and their textual representation
- OR
- **Primary data** (recordings)
- **And secondary data** (transcripts)



# Primary data



*EX: a meeting in an architecture office*

- Several views
- +
- manipulated plans
- drawn sketches
- customer's fax

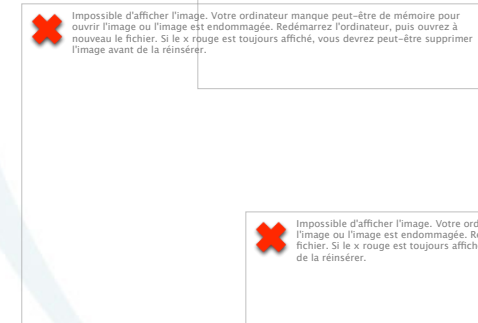
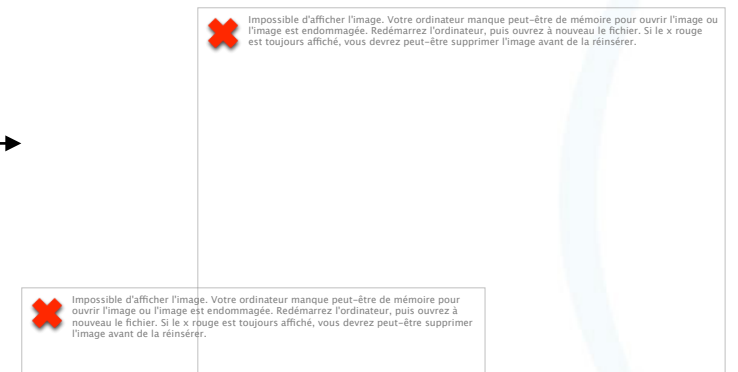
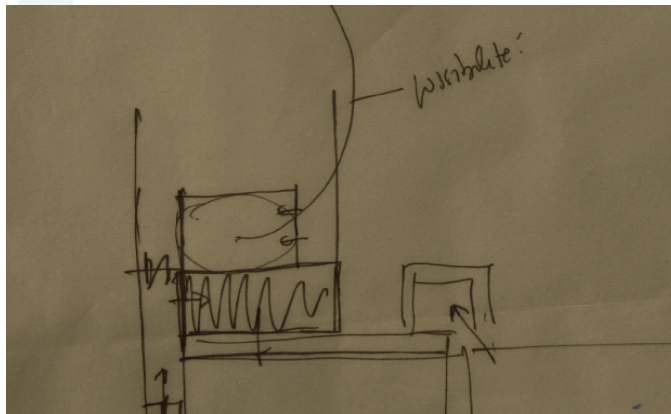
i.e.

## •recordings

- Audio source(s)
- Video source(s)
- Digitalization
- Compression
- Montage (multiscope)
- Anonymization

## •other relevant objects

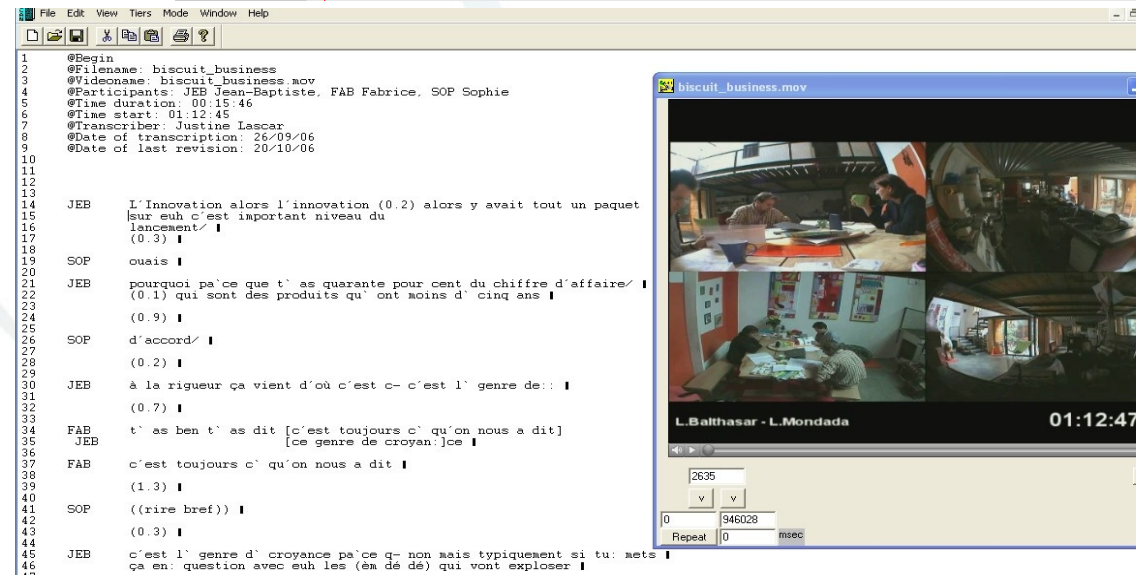
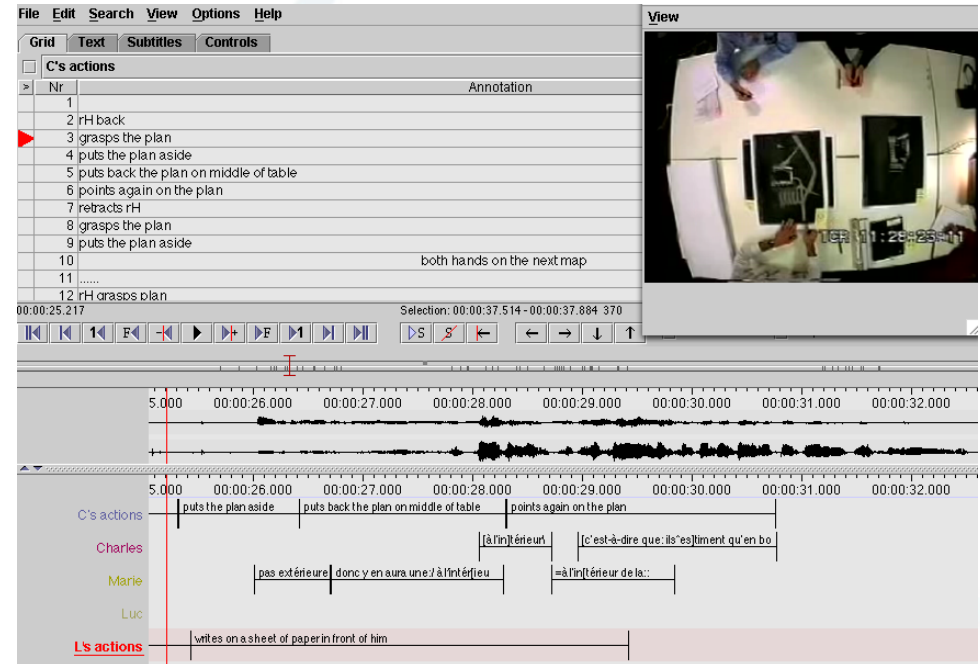
- Documents, plans, images...
- Read/changed/written...
- other objects (screens: dynamic objects)





# Secondary data

- Transcriptions: variants
  - Basis transcript
  - In depth transcript
  - Other versions corrected/deeped/changing with various focus of analysis
  - ± anonymized
  - ± aligned
  - ± annotated
- Transcript conventions
- Metadata
  - Ethnographic notes
  - Descriptions of speakers, setting, activity...
  - Description of recording conditions
- Authorizations & informed consent



# Data and metadata: CLAPI's 75 descriptors

CORPUS ==> NOTAIRES				
Collecté	entre le 01-09-1997 et le 01-04-1998	Durée	07:34:00	Langue Français
Présentation	-> AFFICHER	Nb Locuteurs	18 locuteurs : 6 notaires et 6 couples	
Responsable	Sylvie BRUXELLES, Info-clapi@univ-lyon2.fr			
13 Enregistrement(s)	Nb Etudes	7	-> AFFICHER LES ETUDES	

ENREGISTREMENT ==> BONNETIÈRE				
Durée	00:36:00	Lieu	région lyonnaise	
Recueil	visible	Support	audio, bipé, de bonne qualité	
Genres Interactionnels	Interactions de service Interactions en institution, négociation dans une étude de notaire (partage des biens dans un divorce)			
Accès	Soumis à la signature d'une convention de recherche		-> TELECHARGER LE SIGNAL	
1 Transcription(s)	Nb Locuteurs	4	-> AFFICHER LES LOCUTEURS	-> PLUS D'INFORMATIONS

TRANSCRIPTION ==> BONNETIÈRE -ADAPTÉE CLAPI				
Type	Totale, en orthographe adaptée	Traitement	Toilettée Minutée Anonymisée	Format Informatique
Convention	-> AFFICHER		Transcripteur(s)	Sylvie BRUXELLES
Accès	Emprunt, Analyses et Requêtes soumis à la signature d'une convention de recherche -> TELECHARGER LA TRANSCRIPTION			

ENREGISTREMENT ==> CHAUNE				
Durée	01:32:52	Lieu	région lyonnaise	
Recueil	visible	Support	audio, bipé, de qualité moyenne	
Genres Interactionnels	Interactions de service Interactions en institution, négociation dans une étude de notaire (partage des biens dans un divorce)			
Accès	Soumis à la signature d'une convention de recherche		-> TELECHARGER LE SIGNAL	
1 Transcription(s)	Nb Locuteurs	4	-> AFFICHER LES LOCUTEURS	-> PLUS D'INFORMATIONS

TRANSCRIPTION ==> CHAUNE				
Type	Totale, en orthographe adaptée	Traitement	Toilettée Minutée Anonymisée	Format Informatique
Convention	-> AFFICHER		Transcripteur(s)	Sylvie BRUXELLES, Séverine CHABOUT
Commentaire	Le timing de la transcription repart à zéro lors du changement de face de la cassette, à 00:46:24			
Accès	Emprunt soumis à la signature d'une convention de recherche -> TELECHARGER LA TRANSCRIPTION			



# TRANSCRIPTS

# CLAPI transcripts: variations

- A consequence of keeping heritage corpora: a diversity of transcript conventions
  - Different conventions for representing the same phenomenon
    - Same phenomenon, different notations
    - Different phenomena, same convention
    - Different levels of granularity: phenomena decomposed in different ways
      - e.g. pauses
  - Quality problems
    - Different levels of *coherence*: the same phenomenon transcribed in two different ways within the same transcript
    - *Exhaustivity*: a phenomenon might be transcribed in some parts of the transcript, but not in other parts
    - Different levels of *precision*
  - Different kinds of orthographic adaptations
    - « notre » / « not' »; « effectivement » / « -

## Various implementations:

- Heritage corpora with their original transcripts
- Golden corpora - ICOR advanced transcription + alignment tools (CLAN, ELAN, Praat)
- CIEL transcripts in Praat

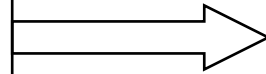
## -XMLization of various types of transcripts

- Adapted orthography -> automatic recognition of adapted spelling and transformation into standard orthography



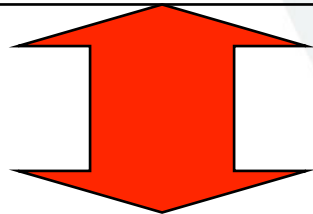
# CLAPI: not only DATA but also TOOLS

an **ARCHIVE**  
of corpora  
of language in interaction

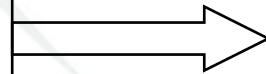


## Downloadable corpora

- whole/partial corpus,
- with/without signal,
- needing/or not a specific convention depending on the autorizations



a series of **TOOLS**  
for quantitative and  
automatized searches  
on corpora  
of spoken language



## Searchable corpora

- simple / complex lexical searches
- correlations between metadata and data
- search of forms in given sequential positions
- search of forms & interactional phenomena
- automatic identification of repetitions
- automatically generated frequencies

## Bienvenue dans CLAPI banque de données et plateforme logicielle

Vous avez choisi "Outils de requêtes", vous disposez depuis le menu horizontal des fonctions suivantes:

### . Concordancier

Afficher les attestations d'un token/mot donné, en contexte, alignées avec le signal audio/vidéo

### . Co-occurrences d'un token

Rechercher automatiquement les co-occurrences gauches et droites d'un token/mot donné, par fréquence décroissante, avec un accès direct au contexte (transcription et signal audio/vidéo)

### . Co-occurrences d'un phénomène

Rechercher automatiquement les co-occurrences gauches et droites d'un phénomène interactionnel, classées par fréquence, avec un accès direct au contexte (transcription et signal audio/vidéo)

### . Segments Répétés

Rechercher automatiquement toutes les répétitions dans une transcription donnée:

- . les répétitions d'un token/mot,
- . les répétitions d'une séquence de 2, 3 ou 4 tokens/mots dans la même production verbale,
- . les répétitions d'un token ou d'une séquence de la production verbale précédente (reprises)

### . Requêtes multi-critères

Effectuer des requêtes complexes pour combiner des séquences de mots/tokens avec des phénomènes interactionnels selon une cinquantaine de critères, organisés en deux niveaux suivant l'option "Productions ou Phénomènes".

Ces requêtes peuvent porter sur une ou plusieurs transcriptions données ou sur l'ensemble des transcriptions disponibles

### . Aperçu des traits contextuels d'un token

Etudier si un token/mot donné est utilisé

- . en corrélation avec une pause, un chevauchement (en début de segment chevauchant ou chevauché), sous une forme répétée,
- . à une certaine place dans le tour de parole (seul, au début ou en fin), dans des productions courtes,
- . dans des interactions de deux, trois ou plus locuteurs,
- . par des locuteurs de sexe féminin ou masculin

## Available tools

- search within the data base of single attested tokens
- concordances and co-occurrences of linguistic tokens and of interactional phenomena (pauses, overlaps...)
- repetitions
- combinatory searches, multi-criteria queries
- correlations between complex clusters of forms and metadata
- statistical informations regarding a corpus (distributions generated automatically)

# Automatic searches for interactional linguistics (1)

- How to go beyond word searches?
  - Most of the data bases of corpora provide for simple concordances and single word searches; sometimes for correlations with one external parameter
- The challenge of multi-criteria search engines

# Automatic searches for interactional linguistics (2)

- How to provide for a heuristic tool for helping to build « collections » of complex interactional phenomena?
  - Reminder: a collection in Schegloff's terms (1996) is a set of excerpts where the same formal features occur within a given sequential environment and perform a specific action
  - Thus, a collection covers aspects that are easily searchable (like forms) and aspects which are less searchable (sequential environments) or very difficult to search (actions)
  - How to 'translate' sequential features into searchable items?
  - Reverseely, how to build 'sequentially' sensitive tools ?
  - Sequentiality --->
    - Positions within the turn
    - Temporal feature (pauses, overlaps)
    - Formal resources
- Phenomena annotated:
  - Positions are expressed in terms of distance from between targeted items (e.g. pause + « well » in the next turn within the first words of the turn)
  - Pauses and overlaps are xml-tagged and allow for searches
  - For overlaps, overlapping and overlapped turns are automatically distinguished (possibility of searching for a particle in a turn-initial position in an overlapped turn)
  - For pauses, although they are transcribed in various ways (original conventions of the archived corpora are preserved - but new corpora are transcribed in the ICOR standard convention, which has a dtd) they are searchable as pauses, short pauses, long pauses
  - Formal resources are mainly reduced to 'words' (tokens) but they can be combined

RECHERCHER UNE CIBLE :

1) soit un token (un token seul sans apostrophe ni blanc) ==>

2) soit un phénomène interactionnel ==>  ▾

AJOUTER EVENTUELLEMENT LES CRITERES SUIVANTS

Choisir un ou plusieurs tokens autour de la cible

Suivi de	▾	<input type="text"/>
Suivi de	▾	<input type="text"/>
Suivi de	▾	<input type="text"/>
Inclure les formes éliées (b'jour, 'fin,...)	<input type="checkbox"/>	
Respecter les accents	<input type="checkbox"/>	
à une distance de	<input type="text" value="1"/>	token(s) MAXIMUM de la cible
Par défaut dans la même <a href="#">Production Verbale</a>		
ou bien dans une suite de	<input type="checkbox"/>	Productions verbales (PV) SUCCESSIVES

Combiner avec un phénomène interactionnel

Suivi de	▾	<input type="button" value="Choisir le phénomène"/> ▾	à une distance de	<input type="text" value="1"/>	token(s) MAXIMUM de la cible
----------	---	---	-------------------	--------------------------------	------------------------------

Préciser la longueur de la Production Verbale (PV)

==> dont la PV contient PLUS DE	<input type="checkbox"/>	tokens	et	▾	MOINS DE	<input type="checkbox"/>	tokens
---------------------------------	--------------------------	--------	----	---	----------	--------------------------	--------

Indiquer la localisation de la cible dans la Production Verbale (PV)

==> en début de PV, à	<input type="checkbox"/>	tokens du DEBUT de la PV	ou	▾	en fin de PV, à	<input type="checkbox"/>	tokens de la FIN de la PV
-----------------------	--------------------------	--------------------------	----	---	-----------------	--------------------------	---------------------------

Choisir **ou exclure** un ou plusieurs genres interactionnels

INCLURE les genres	<input type="text"/>	▾	<input type="text"/>	▾
# EXCLURE les genres #	<input type="text"/>	▾	<input type="text"/>	▾

Spécifier le nombre de locuteurs ou la nature de l'enregistrement

Nombre de locuteurs	AU MOINS	<input type="text" value="0"/>	et AU PLUS	<input type="text" value="0"/>	locuteurs
Source AUDIO ou VIDEO	<input type="button" value="audio ou vidéo"/> ▾				

Interface for searches with multiple criteria  
(note: it is possible to display more criteria)

# Searching interactional phenomena:

## the example of « attends »

- Possible searches:

- FORMS and their non-standard variations:



« -ttends », « attends »,  
« -ttendez », « attendez »

- POSITIONS

- Turn-initial, middle, turn-final
- Sensitive to TCU and TRP (under development)



In turn-initial position  
VS  
In the middle of the turn

- INTERACTIONAL DETAILS

- E.g. first or second word after an overlap / a pause



On the overlap onset /  
post-overlap resolution

- MULTIMODAL DETAILS (under development)

- E.g. simultaneous to a pointing gesture



Simultaneous with a  
« stopping » gesture



# « Attends »: some results

- Two sequential environments characterizing the distribution of « attends »
  - Turn/TCU-initial, in overlap, with « stopping » gestures, often in a serial form
  - In the middle of a long multi-unit turn, where an argumentation or a story telling is going on; where various stances are reported
- Two sets of practices accomplished through « attends »
  - Management of time, and of the ongoing projections (syntax as well as turn and action) in a context of disagreement about the next step
  - Argumentative moves, polyphonic stances
- Grammaticalization process
  - From an imperative verb
  - To a discourse particle
- In relation with metadata:
  - On a specific corpus (science classroom) Girls vs boys
  - Other forms in the same environment and doing the same job: « putain »



1 VER des petites réflexions que tu lui aurais fait toi y  
2 a dix ans en arrière/ elle aurait rien dit/ (.) ou  
3 elle aurait rigolé\ (.) là maintenant le MOINDRE  
4 truc/ (.) ATTENDS/ ce m- euh elle m'appelle ce  
5 matin à midi pour me dire qu'il y avait plus de  
6 gaz/ (.) j' lui dis bon/ (..) j' lui dis euh demain  
7 i' pourra pas parce qu'il va faire un squash avec  
8 NN/ oh ben i' vient quand i' veut/ i' vient quand  
9 i' veux\ (.) j'ui dis non mamy i' viendra cet après  
10 midi/ (.) et comme si je voulais pas qu'i' vienne/

# Future evolutions

Even more diversity:

- Integration of other languages
- Integration of code-switching, multilingual encounters, mixed lects
- Integration of linguistic variation in French (CIEL)
- Development of multimodal standard annotations (CLARIN)

Interoperability

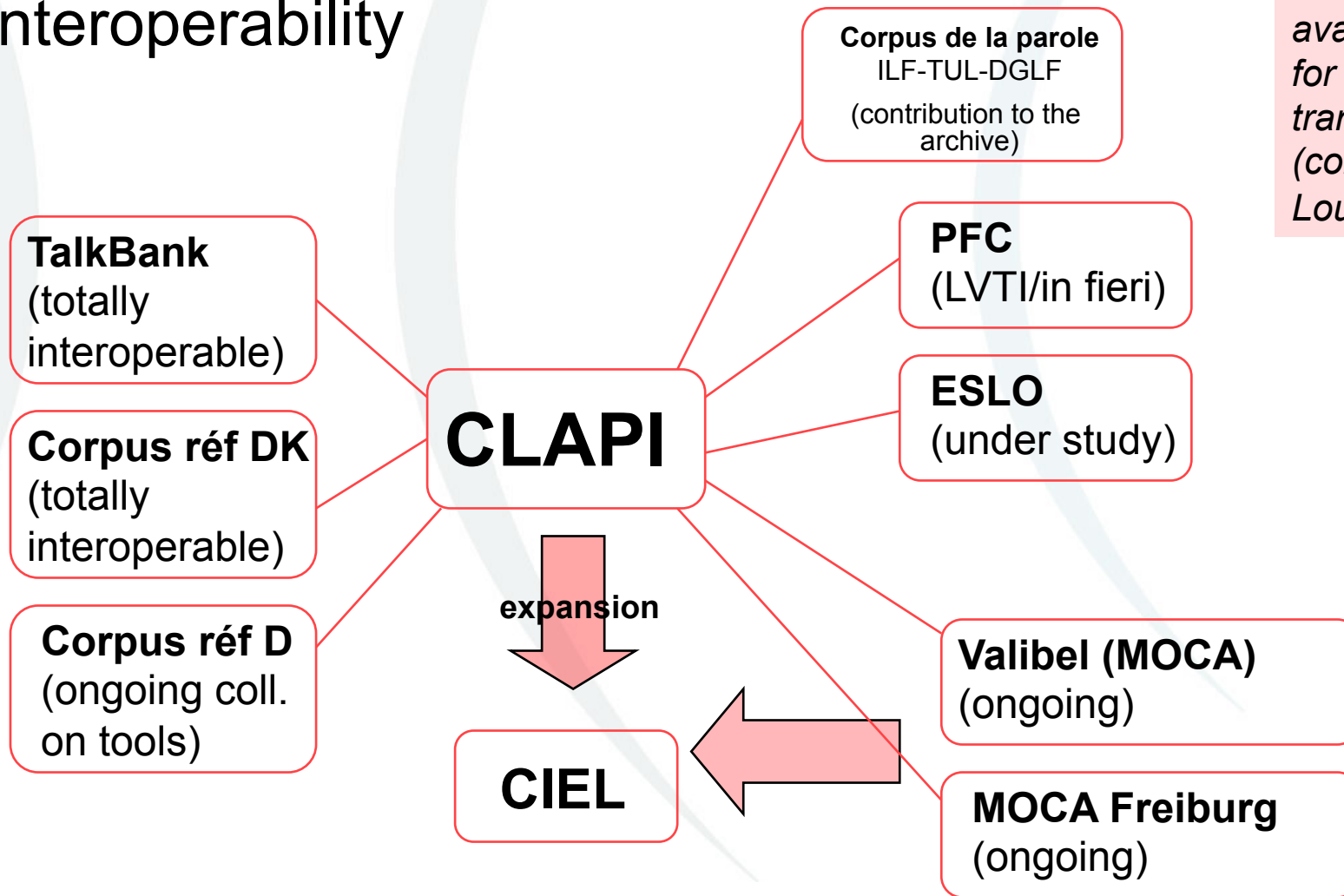
Perennity & sustainability



# Not only one data base: the need for interoperability

- Already within CLAPI we have different sub-data bases
- Importance of being connected to other data bases
- Standardization is not the only and ultimate solution
- Standards are only partially shared
  - Eg. Variety of XML formats
  - Eg. Variety of transcript conventions
  - Eg. Variations within the metadata
- Solution: pivot-formats and converters
- Importance of evolutive formats
  - Primary data formats change
  - Transcripts evolve, new demands emerge
  - Probl. of archive perennity and sustainability

# interoperability



*TEI (P5) export is available since 2006 for metadata and transcripts (collaboration with Lou Burnard)*

Ongoing collaborations with Brian MacWhinney (Carnegie Mellon U.), Johannes Wagner (SDU Odense, DK), Arnulf Deppner (IdS Mannheim, D), Jacques Durand (CLLE Toulouse), O. Baude (LLL Orléans), Anne-Catherine Simon (Valibel, Louvain, B), Stefan Pfänder (Freiburg, D).

# General conclusions

- CLAPI, data base of naturalistic corpora of social interaction, aims at
  - supporting data **sharing** and increasing the available downloadable data for the scientific community interested in the study of language in a diversity of contexts
  - Spreading « **best practices** » for data collection, data transcription, ethical guidelines for fieldwork and corpus constitution, qualitative and quantitative analysis of interactional corpora (see the CORINTE web site)
  - Thinking about the impact of **technology** on data processing and the new analytical opportunities opened up by possible technological choices
  - Using technologies for the **multimodal** and not only the linguistic study of social interaction, integrating new forms of alignment and annotation relevant for the study of gestures and other visual conducts
  - supporting **qualitative** research, by offering new data to the scientific community
  - supporting **quantitative** research, by offering new tools able to browse large amounts of data

# Merci beaucoup de votre attention !

- CLAPI - The platform of freely accessible corpora of interactions:
  - <http://clapi.univ-lyon2.fr>
- CORINTE - A related website for interactional research
  - <http://icar.univ-lyon2.fr/projets/corinte>