

Corpus oraux : les « bonnes pratiques » des linguistes

OLIVIER BAUDE

Laboratoire Ligérien de Linguistique UMR 7270
DGLFLF Ministère de la Culture olivier.baude@univ-orleans.fr

0. Introduction

- ▶ Retour sur une expérience de savoirs et savoir-faire scientifiques:
 - ▶ **Le programme *Corpus de la parole* et le *Guide des bonnes pratiques***
 - *Depuis 2004, la Délégation générale à la langue française et aux langues de France développe un programme en faveur de la conservation, la mise à disposition, la diffusion et la valorisation des corpus oraux produits par des linguistes : le programme *Corpus de la parole*.*
 - **Au cœur de ce programme il y a une démarche réflexive : La conservation est nécessairement l'occasion, pour le linguiste de rendre explicite l'ensemble de ses pratiques.**
 - **Initiative originale autour d'une réflexion sur les aspects juridiques liés à la conservation et la diffusion des corpus oraux.**

1. Les linguistes et la mémoire de l'oral

► Brefs repères historiques :

- Début du XXe : La voix peut être enregistrée. Naissance des Archives de la Parole "d'étude, d'archivage et d'analyse de parlers d'hommes et de femmes parlant comme à l'auberge ou à la fontaine" [Brunot 1911]
- 1938, Création de la phonothèque nationale, qui devint le département des archives sonores puis en 1995 le département de l'audiovisuel de la BnF-François Mitterrand
- Deuxième partie du XXe : Réalisation des atlas linguistiques et début des corpus en sociolinguistique, linguistique descriptive, acquisition du langage, français parlé...
- Fin du XXe : corpus pour l'ingénierie linguistique, reconnaissance et synthèse de la parole, traduction...

Pourtant aujourd'hui, il y a très peu de données sonores disponibles pour la linguistique. Pourquoi ?

1. Les linguistes et la mémoire de l'oral

- ▶ Des données dignes d'un savoir ?

1. Les linguistes et la mémoire de l'oral

► Données et "technologies"

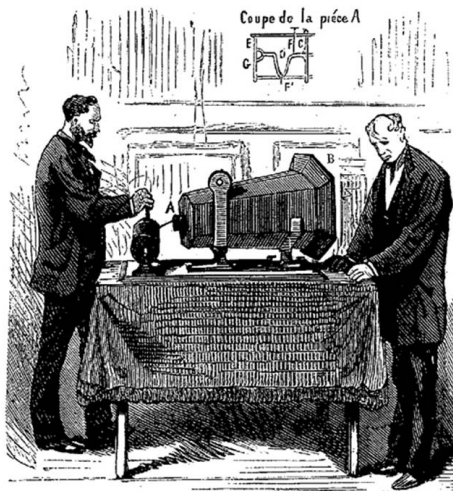


Fig. 466. — Le phonautographe.

1856



Fig. 474. — Le nouveau phonographe d'Edison.

1878

1. Les linguistes et la mémoire de l'oral

► Les Archives de la parole

- *Début du XXe : enregistrement de la voix : Archives de la Parole "d'étude, d'archivage et d'analyse de parlers d'hommes et de femmes parlant comme à l'auberge ou à la fontaine" [Brunot 1911]*



1. Les linguistes et la mémoire de l'oral

► Les Archives de la parole

- *La mission dans les Ardennes est la première pierre de l'atlas linguistique phonographique rêvé par Ferdinand Brunot avec les Archives de la parole. Elle a lieu en juin-juillet 1912. De leur mission dans les Ardennes, Ferdinand Brunot et Charles Bruneau ramènent 166 documents sonores enregistrés sur disques plats Pathé Saabir de 25 cm de diamètre grâce à un matériel d'enregistrement spécial*



<http://gallica.bnf.fr/dossiers/html/dossiers/VoyagesEnFrance/themes/Chansons3.htm>

1. Les linguistes et la mémoire de l'oral

- Discours inaugural par Ferdinand Brunot  :

Il suffit qu'une voix s'éteigne pour que nous en soyons séparé par un espace infranchissable, mais nous sommes au siècle des merveilles, si le monde a comme on le dit des réalités que la science n'atteint pas, en échange la science donne sans cesse au monde des réalités qu'il n'avait pas.

C'en est fait des lieux communs que depuis l'antiquité on répétait sur la parole ailée ou sur l'homme attaché à la terre, voici qu'à peu près en même temps l'homme commence à faire son chemin vers le ciel, la parole se grave dans la matière pour toujours. Brunot 1911

1. Les linguistes et la mémoire de l'oral

► L'oral, un savoir peu légitime...

- *"Longtemps, le marché de la linguistique en France, centralisé par le quasi monopole de Paris sur la circulation des biens symboliques et dominés par les études littéraires dans l'organisation universitaire, favorise objectivement la recherche abstraite, universalisante ou formalisante, **au détriment des pratiques de recension descriptives, autrement dit, ceux qui travaillent en bibliothèque plutôt que sur le terrain.**"*

Bergounioux, G, 1992, "les enquêtes de terrain en France", Langue française n°93, Enquêtes corpus et témoins en France, hier et aujourd'hui, Larousse, Paris, p18.

- *"Mais qui s'intéresse au français parlé? (...) **peu de gens y voient un objet légitime d'étude (même chez les linguistes)** pour bon nombre de ceux-ci la langue parlée c'est bon pour l'exotisme; la description de la langue parlée vaut pour les dialectes et les patois du français ; elle vaut aussi pour les langues sans écritures dites "exotiques" ; mais pas pour une langue de culture comme le français".*

Blanche-Benveniste et JeanJean, *Le français parlé, transcription et édition*, Didier

Erudition, 1987

1. Les linguistes et la mémoire de l'oral

L'oral un savoir sans mémoire ?

- ***L'exemple du français fondamental*** : "Nous ne nous soucions pas non plus de la conservation des disques. Nous profitons largement des possibilités qu'offrent les disques en papier d'être effacés et de servir ainsi à plusieurs enregistrements successifs. Il aurait été beaucoup trop coûteux de conserver tous les enregistrements comme beaucoup de bons esprits nous le suggéraient" [1].
[1] Gougenheim et al. 1964, p.63.

- ***L'exemple du corpus d'Orléans :***



- **Une tentative : la convention des atlas linguistiques (1979).** Une bande enregistrée versée à la BnF contre 2 bandes magnétiques vierges.

2. Le programme *Corpus de la parole*

2. Le programme *Corpus de la parole*

▶ Naissance d'un programme

- ▶ Depuis plus d'un siècle les linguistes collectent des enregistrements sonores (corpus oraux) afin de décrire les langues et de réaliser des applications diverses (des méthodes de langues à l'ingénierie linguistique et au Traitement Automatique du Langage en passant par la description et les études linguistiques.).

▶ **Constats :**

- Ces ressources sont en nombre insuffisant en France.
- Depuis Ferdinand Brunot et les archives de la parole créées en 1911, le ministère de la Culture et celui de la recherche ont régulièrement conjugués leurs efforts pour considérer ces objets comme un patrimoine.
- Malgré ces efforts des documents précieux n'ont pas été sauvegardés.
- Les technologies actuelles permettent d'apporter des solutions innovantes.

2. Le programme *Corpus de la parole*

- ▶ Actuellement ces enregistrements sonores numérisés peuvent :
 - ▶ Être conservés à moindre coût (et catalogués et indexés),
 - ▶ Être facilement diffusés par l'Internet,
 - ▶ Être exploité selon les méthodes de la recherche en linguistique et en informatique :
 - transcription avec synchronisation entre la transcription et le signal,
 - annotations (traductions, gloses, étiquetage morphosyntaxique),
 - applications (reconnaissance et synthèse de la parole, outils didactiques, etc.)

2. Le programme *Corpus de la parole*

- ▶ Objectifs d'un programme:
 - ▶ Plan de numérisation
 - ▶ Partenariat avec le CNRS pour une conservation et une diffusion d'archives scientifiques devenues objets de patrimoine.
 - ▶ Accessibilité des corpus et valorisation.
 - Aspects juridiques : Guide des bonnes pratiques
 - l'établissement de normes et de pratiques d'interopérabilité,
 - la gestion d'un réservoir de corpus numérisés,
 - la création d'un site de valorisation

2. Le programme *Corpus de la parole*

► Difficultés repérées :

- Hétérogénéité des objets "corpus", enregistrements, transcriptions, documentation papier, corpus enrichis (annotations, TAL).
- Problèmes juridiques.
- Pratiques des chercheurs (pas de politique ni même de "pratiques sauvages" d'archivage et de mise à disposition).
- Impossibilité du recours à une externalisation des opérations de numérisation. Méthodologie exclusive des chercheurs (choix théoriques, description scientifique, transcription, traduction,...) à croiser avec méthodologie des institutions de conservation (techniques de numérisation, catalogage, indexation)
- Nécessité d'élaboration de bonnes pratiques par une communauté.

2. Le programme *Corpus de la parole*

► **Un conseil scientifique**

► **Ministère de la Culture et de la Communication**

- DGLFLF
- BnF,
- INA,
- DREST, SIAF

► **CNRS**

- DSA section 34,
- Fédérations ILF et TUL
- TGIR ADONIS-CRDO,
- Juristes (Cecoji) Isabelle De Lamberterie
- Linguistes.

2. Le programme *Corpus de la parole*

- ▶ **Un contexte favorable (technique, scientifique, politique)**

- ▶ **Contextes techniques**

- données sonores numérisées,
 - manipulation et diffusion aisées,
 - transcriptions synchronisées,
 - outils d'aide à la transcription,
 - avancées intéressantes reco/synthèse,
 - outils favorisant l'interopérabilité,
 - ...

2. Le programme *Corpus de la parole*

► Un contexte favorable (technique, scientifique, politique)

• Contextes scientifiques

- **Programme des fédérations de recherche en linguistique**
- initiative ASILA-EPML50 sur les corpus oraux, ELCO, ..
- constitution de corpus (ESLO, PFC, Coral-Rom...)
- appel d'offre ANR "corpus"
- création des centres de ressources numériques (CRDO),
- normalisation et interopérabilité (OLAC, TEI, ...),

2. Le programme *Corpus de la parole*

- ▶ Un contexte favorable (technique, scientifique, politique)

- Contextes politiques d'une diffusion de la recherche et des ressources

- « Les organismes publics doivent avoir le souci constant de faire bénéficier au mieux la collectivité nationale des fruits de leurs travaux... ».
« La politique de la recherche et du développement technologique vise à l'accroissement des connaissances, à la valorisation des résultats de la recherche, à la diffusion de l'information scientifique et technique et à la promotion du français comme langue scientifique »

Art 5 de la Loi n°82-610 du 15 juillet 1982 modifiée d'orientation et de programmation pour la recherche et le développement technologique de la France, aujourd'hui art. L 111-1 du code de la recherche. JO du 16-07-1982, p. 2273 et ss.

2. Le programme *Corpus de la parole*

- ▶ **Un contexte favorable (technique, scientifique, politique)**

- **Contextes politiques d'une diffusion de la recherche et des ressources**

- Le 22 octobre 2003, à Berlin, la plupart des Directeurs Généraux des Établissements Publics à caractère Scientifique et Technologique (EPST) ont signé la *Déclaration de Berlin sur le Libre Accès à la Connaissance en Sciences exactes, Sciences de la vie, Sciences humaines et sociales*, dont l'objectif est de promouvoir Internet « comme instrument fonctionnel au service d'une base de connaissance globale de la pensée humaine ».

2. Le programme *Corpus de la parole*

► Un contexte favorable (technique, scientifique, politique)

- Contextes politiques d'une diffusion de la recherche et des ressources

- Les programmes de numérisation patrimoniale

En 2001 à Lund, en Suède, un groupe de représentants nationaux des États membres de l'Union européenne, intéressés par les problèmes de numérisation, a élaboré un texte qui prône notamment : la mise en place de standards d'interopérabilité ; la diffusion de bonnes pratiques dont la gestion des droits de propriété intellectuelle ; l'organisation de centres de compétences sur la numérisation dont les professionnels de l'information ont la responsabilité.

- En France : plan de numérisation du ministère de la culture, en 2007 les langues sont pour la première fois l'une des thématiques.

3. Groupe de travail sur les aspects juridiques

- ▶ **Composition du groupe de travail**

- ▶ Experts linguistes
- ▶ Représentants des fédérations de recherche
- ▶ Représentants de projets en cours
- ▶ Conservateurs
- ▶ Juristes (CECOJI)
- ▶ Informaticiens – gestion de corpus

▶

3. Groupe de travail sur les aspects juridiques

► Objectifs

- recenser les pratiques actuelles et définir en priorité les contraintes méthodologiques et théoriques liées à la recherche ;
- diffuser une synthèse sur la législation existante ;
- établir des recommandations ; *et, le cas échéant, en cas de vide ou de flou, formuler des propositions pour l'élaboration de normes et règles juridiques (notamment européennes).*

3. Groupe de travail sur les aspects juridiques

► Un regard sur une expérience

- Objectifs et méthodes du groupe de travail Ministère de la Culture/CNRS

Méthode

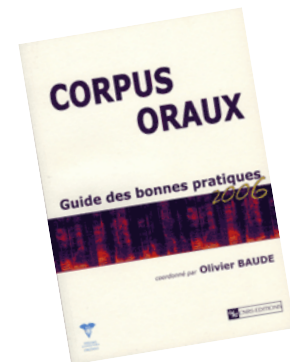
- pas de réponses toutes faites
- ne pas "brider" la recherche
- respect de la méthodologie du chercheur
- mettre en commun les compétences requises
- respect de la loi et de l'éthique
- outil d'expertise des risques
- prise en compte des cadres institutionnels

⇒ **Respect des méthodes des chercheurs**

⇒ **Respect des données et "des sources"**

⇒ **Interopérabilité, diffusion et accès**

⇒ **Une pratique élaborée par une communauté scientifique**



3. Groupe de travail sur les aspects juridiques

- ▶ Aspects juridiques :

3 grands domaines :

- propriété matérielle et intellectuelle
- traitement des données personnelles
- responsabilité des hébergeurs et des diffuseurs

3. Groupe de travail sur les aspects juridiques

- ▶ Aspects juridiques :

- 1. définir juridiquement ce qu'est un "corpus » (qu'est-ce que « figurer dans un corpus » ?)**
- 2. définir les principes et les modalités de la gestion contractuelle des droits des personnes concernées**
- 3. Déterminer droits et responsabilités dans les étapes du process (créateurs, hébergeurs, diffuseurs...)**

4. Aspects juridiques : **La propriété intellectuelle**

- ▶ Contient-il une création protégée par le droit d'auteur ?

4. Aspects juridiques : **La propriété intellectuelle**

- ▶ Contient-il une création protégée par le droit d'auteur ?

- ▶ Soit le corpus est constitué d'œuvre du domaine public :
 - *Le domaine public recouvre non seulement les idées de liberté d'accès et de gratuité d'utilisation des données, mais aussi la possibilité pour chacun de les exploiter.*
 - Il est libre de droits : Liberté d'accès, gratuité d'utilisation, liberté d'exploitation, de fait, ou après 70 ans...

- ▶ Soit il est protégé par le droit d'auteur : sous quelles conditions?

4. Aspects juridiques : **La propriété intellectuelle**

- ▶ Contient-il une création protégée par le droit d'auteur ?

3 conditions pour qu'un corpus soit protégé :

- ▶ - Il faut en premier lieu qu'il corresponde à l'exigence d'une **activité créatrice** : un travail de compilation d'informations n'est pas protégé en soi.
- ▶ - Pour être protégé, il est par ailleurs indispensable que le corpus ait une **forme définie**. Ce qui est protégé, ce n'est pas le contenu du corpus mais son enveloppe, son architecture.
- ▶ - Enfin, la forme du corpus doit répondre à la condition d'être **originale**. Que signifie l'originalité d'un corpus ? L'originalité de nombreuses créations de l'ère du numérique, comme les logiciels ou les bases de données, ne peut être appréciée que d'après des critères objectifs. Il semble qu'il en soit de même des corpus oraux, ceux-ci pouvant le plus souvent être assimilés à une base de données. C'est alors, le plus souvent, le fait que le corpus soit ou non copié et révèle un minimum d'activité créative qui servira de critère pour déterminer s'il est ou non original (et non pas uniquement la prise en compte de l'empreinte de la personnalité de son auteur).

4. Aspects juridiques : **La propriété intellectuelle**

► Qui est l'auteur ?

L'auteur est en principe la (ou les) personne(s) physique(s) sous le nom de laquelle (ou desquelles) l'œuvre est divulguée.

Si plusieurs auteurs, 2 cas :

- co auteurs (mêmes droits pour chaque auteur)
- **œuvre collective** peuvent être qualifiées d'œuvre collective lorsqu'elles sont créées « sur l'initiative d'une personne physique ou morale qui l'édite, la publie et la divulgue sous sa direction et sous son nom, et dans laquelle la contribution personnelle des divers auteurs se fond dans l'ensemble »^[1].

Dans ce dernier cas, c'est la personne physique ou morale qui a pris l'initiative de l'œuvre qui dispose des droits d'auteur

^[1]Art. L. 113-2 du CPI.

4. Aspects juridiques : **La propriété intellectuelle**

► **Quels droits pour les corpus protégés ?**

- Droits patrimoniaux (autoriser ou interdire la reproduction ou la communication au public).
- Prérogatives du droit moral (divulgation, retrait, paternité, respect de l'œuvre).

Le chercheur auteur qui refuse de divulguer le corpus qu'il a créé est dans son droit (au titre du droit d'auteur), même si par ailleurs il peut être sanctionné administrativement pour ne pas avoir exécuté sa mission de service public qui est de communiquer les résultats de sa recherche.

4. Aspects juridiques : **La propriété intellectuelle**

▶ **Quels droits pour les corpus protégés ?**

▶ **Domaine public / droit d'auteur :**

- Une solution intermédiaire : liberté d'accès et d'exploitation

Sans être dans le domaine public, ces corpus sont – de par la volonté de leurs créateurs – libres d'accès et d'utilisation. Néanmoins, si les créateurs peuvent renoncer à exercer leurs droits patrimoniaux, il ne leur est pas possible de renoncer à leur droit moral, qui reste imprescriptible.

4. Aspects juridiques : **les données personnelles**

▶ **Le corpus contient-il des données personnelles ?**

▶ **Enjeux :**

- Si données personnelles : respect de la loi Informatique et libertés
- Les données anonymisées sortent du champ de la loi

4. Aspects juridiques : **les données personnelles**

- ▶ **Le corpus contient-il des données personnelles ?**

- ▶ **Qu'est-ce qu'une "donnée personnelle"**
 - ▶ **La loi du 6 janvier 1978 s'articule autour de la notion de donnée nominative. La Convention 108 du Conseil de l'Europe de 1981 lui préfère celle de données personnelles et la directive 95/46/CE choisit l'expression « donnée à caractère personnel »**

4. Aspects juridiques : **les données personnelles**

- ▶ **Le corpus contient-il des données personnelles ?**

- ▶ **Qu'est-ce qu'une "donnée personnelle"**
 - ▶ **Données "identifiantes" :**
 - « ...pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne ».

4. Aspects juridiques : **les données personnelles**

► **Traitement des données personnelles :**

- L'article 2b de la directive européenne définit ainsi le traitement des données :

« toute opération ou ensemble d'opérations portant sur de telles données, quel que soit le procédé utilisé, et notamment la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction. »

4. Aspects juridiques : **les données personnelles**

► **Traitement des données personnelles :**

- Tout traitement des données doit avoir un responsable

La directive européenne 95/46/CE ("Flux de données transfrontières") dans son article 2d, repris pour la refonte de loi "informatique et libertés" donne la définition suivante :

"Le responsable d'un traitement de données à caractère personnel est, sauf désignation expresse par les dispositions législatives ou réglementaires relatives à ce traitement, la personne, l'autorité publique, le service ou l'organisme qui détermine ses finalités et ses moyens."

4. Aspects juridiques : **les données personnelles**

▶ **Traitement des données personnelles :**

▶ Responsable de quoi ?

▶ Le responsable du traitement se doit donc de veiller à :

1. la qualité des données ;
2. à leur(s) finalité(s) ;
3. au recueil du consentement.

▶ Il est responsable de la déclaration à la CNIL

"Tout traitement de données personnelles à caractère direct ou indirect doit être déclaré à la CNIL".

5. Que le linguiste sache ce qu'il fait

5. Que le linguiste sache ce qu'il fait

- ▶ Un regard sur une expérience

- ▶ Pour définir le statut juridique il faut **connaître les conditions d'élaboration et d'exploitation** du corpus.

- Seule la description des "conditions d'élaboration et d'exploitation" permettra de répondre aux questions suivantes :

- Q1 : *contient-il une création protégée par le droit d'auteur ?*

- Q2 : *contient-il des données personnelles ?*

- C'est une nécessité juridique et éthique indissociable du travail scientifique. **La constitution et l'exploitation de corpus relèvent d'une démarche que le chercheur doit expliciter.**

5.1 Un corpus oral ?

- Des enregistrements sonores sur des *supports*



5.1 Un corpus oral ?

► *De l'écrit* (traductions, annotations,...)

Le mouvement : action catholique, le
Re : messe, mariage, promenade en voi-
diner, vélo + triest
ite (profession de mari) donc sports o
nt, colonies de vacances
lecture
responsabilités

lecture

en société, prendre responsabilité, être trop livresque, enseigner techniquement, exemple de sa fille aînée

ma math, les filles laissent RR pro

a) difficile de trouver situation d

de l'espit - culture générale

les enfants, 16 ans en minimum

à pousser au maximum, filles pas

survivant leur personnalité

les jours

de sports pour les jeunes
principal; gros industriels du fait d'
assez calme; certaines personnes
mécontentement; syndicats débordés

[illegible]

01 Depuis 1955.
02 Changement de métier.
03 Oui, mais se trouve plus souvent en dehors d'Orléans que dans
(5 jours mis).
04 Probablement, 3 mois de circonstances imprévues.

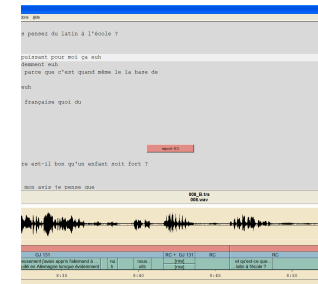
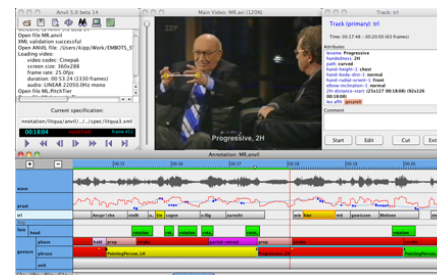
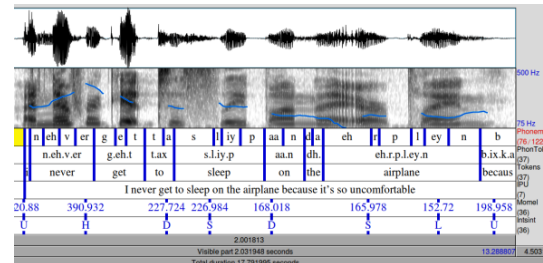
T1 R R représentant de café. "Vente placier-livreur".
T2 R Travail fatigant, absorbant. (60 heures par semaines).
T3 Premier atelier : Charbon + bon métier bien diné.
Aurait voulu être louangeur.
T5 Difficile à dire - lui dans un sens, mais non car impôts plus
Hais dans l'ensemble, une force doit travailler.
T(4) A toujours travaillé.
T(7) Deux enfants : un fils dans l'armée
une fille mariée.
T(3) Non
T(9) des commerçants (épiciers, louangeurs, restaurateurs)
T(10) Du commerce

[illegible]

```

1 <?xml version="1.0" encoding="iso-8859-1"?>
2 <?xml-stylesheet type="text/xsl" href="q2.xml"?>
3 <TEI>
4   <teiHeader>
5     <fileDesc>
6       <titleStmt>
7         <title>Poésies</title>
8         <author>Arthur Rimbaud</author>
9       </titleStmt>
10    </fileDesc>
11  </teiHeader>
12  <text>
13    <group>
14      <text>
15        <front>
16          <head>Le Dormeur du Val</head>
17          <dateLine>7 octobre 1870</dateLine>
18        </front>
19        <body>
20          <lg>
21            <l>C'est un trou de verdure où chante une rivière</l>
22            <l>À accrochant follement aux herbes des haillons</l>
23            <l>D'argent : où le soleil, de la montagne fière,</l>
24            <l>Dit : c'est un petit val qui mousse de rayons.</l>
25          </lg>
26          <lg>
27            <l>Un soldat jeune, bouche ouverte, tête nue,</l>
28            <l>Et la nuque baignant dans le frais cresson bleu,</l>
29            <l>Dort : il est étendu dans l'herbe, sous la nue,</l>
30            <l>Pâle dans son lit vert où la lumière pleut.</l>
31          </lg>
32          <lg>
33            <l>Les pieds dans les glaçes, il dort. Souriant comme</l>
34            <l>Souriait un enfant malade, il fait un somme :</l>
35            <l>Mature, bécote le chaudement : il a froid.</l>
36          </lg>
37          <lg>
38            <l>Les parfums ne font pas frissonner sa narine</l>
39            <l>Il dort dans le soleil, la main sur sa poitrine</l>
40            <l>Tranquille. Il a deux trous rouges au côté droit.</l>
41          </lg>
42        </body>
43      </text>
44    </group>
45  </text>
46 </TEI>

```



► Des métadonnées

020 témoin UQ 417
 chercheuse CS

UQ 417 né 1911 à Orléans

enseignement primaire public
diplôme : CEP
fin d'études : 14 ans

profession chef de service banque

INSEE : 34

échelle AM : C

famille femme sans activité

fille, 30 ans, BEPC, Secrétaire
 fille, 25 ans, Bac, licence d'anglais
 Professeur

politique Gauche

remarque intérêt linguistique : emploi très fréquent des constructions impersonnelles - surtout avec ça ; terminologie financière.

sommaire

né

28

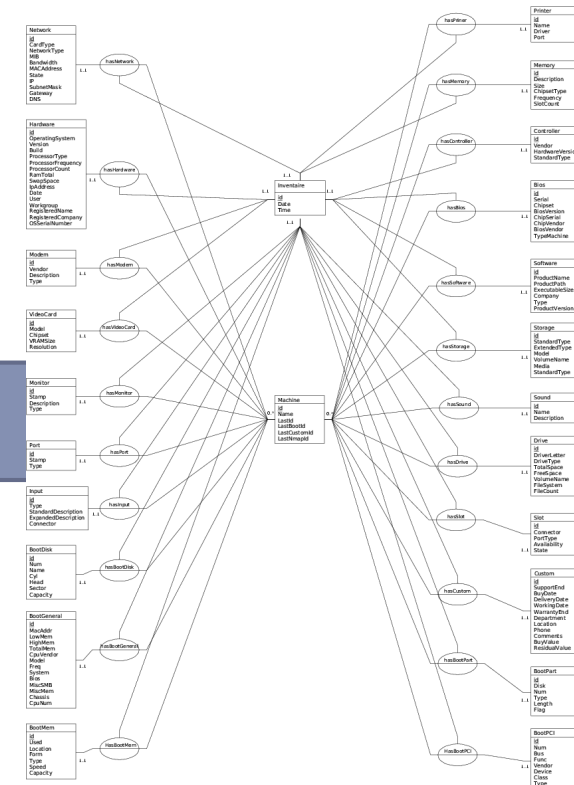
date 11 avril 1969

lieu chez témoin

durée 47'

questionnaires enregistrés ouvert, onelette,
sociolinguistique
transcription extraits : travail
enseignement

enregistrements reliés
234 reprise de contact



5.1 Un corpus oral ?

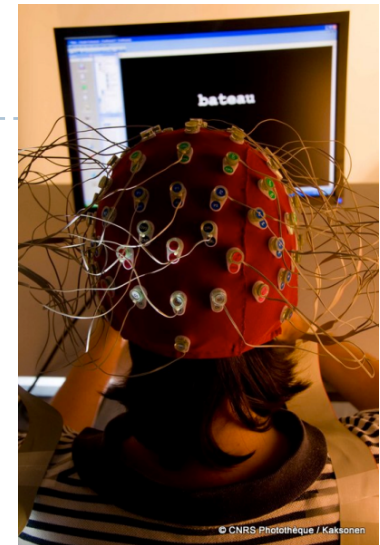
- ▶ Des enregistrements sonores sur des *supports*
- ▶ Le support détermine un objet *différent* à l'origine ou après un ou des traitement(s) (numérisation) :
 - ▶ *conservation*
 - ▶ *manipulation*
 - ▶ *original et copie*
 - ▶ *diffusion* = Transformation d'un "objet" scientifique



5.2 Un corpus oral : comment ?

► Des techniques d'enquêtes

- Enregistrement en laboratoire
- Questionnaire
- Entretien
- Recueil de contes, chants,...
- Récit de vie
- Activités dans leur contexte ordinaire
- Reprise d'enregistrements



- *Les données enregistrées ne sont pas des données préexistantes et recueillies, elles sont le produit de la situation d'enquête*

5.2 Un corpus oral : comment ?

- ▶ Une catégorisation des participants :
 - ▶ Les techniques d'enquêtes produisent (ou valident) une **catégorisation** des participants : *informateurs, témoins, sujets, cobayes, natifs, enquêtés, collaborateurs, observés, participants, acteurs sociaux, ...*
 - ▶ Cette catégorisation résulte de statuts mais aussi de rôles construits par la technique d'enquête : *témoin VS observateur-participant*

5.2 Un corpus oral : comment ?

- ▶ Une construction d'une relation sociale : le mode d'approche
 - ▶ *La façon dont les personnes sont approchées sur le terrain construit une relation sociale aux implications éthiques et personnelles déterminantes.*

5.2 Un corpus oral : comment ?

- ▶ Une construction d'une relation sociale : le mode d'approche
 - ▶ *La façon dont les personnes sont approchées sur le terrain construit une relation sociale aux implications éthiques et personnelles déterminantes.*

Eslo > Accueil



Bienvenue sur le site du portrait sonore de la ville d'Orléans

Une suite 40 ans après... En 1969, des universitaires britanniques ont réalisé un premier portrait sonore de la ville en enregistrant plusieurs centaines d'Orléanais dans la vie de tous les jours. Il s'agit du plus important témoignage sur le français des années soixante-dix. En 2013, quarante ans après cette première étude, l'université d'Orléans, en partenariat avec le CNRS, le Ministère de la Culture et la Région Centre, renouvelle l'expérience en procédant à des enregistrements avec des habitants de toute l'agglomération.

Des paroles de la vie quotidienne. Les enregistrements réalisés, que ce soient des interviews ou des paroles captées dans la rue, les transports publics, les commerces, les lieux de travail ou chez les habitants, forment un formidable témoignage sur la ville et sur le français et les langues parlées quotidiennement dans toutes leurs variétés et leur diversité.

Pour faire quoi ? Ces enregistrements transcrits, rendus anonymes et informatisés constituent une très riche ressource pour les chercheurs en tout genre : historiens, sociologues, linguistes, etc. Les spécialistes du langage et des langues l'utilisent pour décrire le français dans ses usages les plus divers, afin de mieux le connaître, d'élaborer des dictionnaires, des grammaires, des méthodes de langues, des outils de traduction et même des applications telles que la synthèse ou la reconnaissance automatique de la parole.

Tous les Orléanais ont la parole : Participez au portrait sonore de la ville d'Orléans par ses habitants !

Tous les Orléanais ont la parole : Participez au portrait sonore de la ville d'Orléans par ses habitants !

Un portrait sonore ? En 2013, des chercheurs et des étudiants de l'université d'Orléans réalisent 150 interviews d'Orléanais afin de dresser le portrait sonore de la ville par les paroles de ses habitants. Ces interviews ont pour thème les Orléanais et leur ville (leurs quartiers, leurs loisirs, leur vie quotidienne,...).

Qui peut participer ? Tout le monde ! Tous ceux qui habitent Orléans ou son agglomération depuis plus d'un an, quels que soient le quartier, l'âge, la profession... La plus grande diversité est recherchée !

Comment ? Si vous acceptez de consacrer un peu de temps libre à une interview, il suffit de contacter l'équipe par mail ou par téléphone :

5.2 Un corpus oral : comment ?

- ▶ Une construction d'une relation sociale : le mode d'approche
 - ▶ *La façon dont les personnes sont approchées sur le terrain construit une relation sociale aux implications éthiques et personnelles déterminantes.*
 - *Fieldwork (travail de terrain) : relations personnelles, nécessité de confiance, dette de*



<http://hiatus.typepad.com/hiatusbienvenue/travail-du-terrain/>

5.2 Un corpus oral : comment ?

- ▶ Une construction d'une relation sociale : le mode d'approche
 - ▶ *La façon dont les personnes sont approchées sur le terrain construit une relation sociale aux implications éthiques et personnelles déterminantes.*
 - *Micro-trottoir : participants choisis plus ou moins étroitement à cause de leur présence*



5.2 Un corpus oral : comment ?

- ▶ Un premier traitement : le dispositif d'enregistrement
 - ▶ *Le dispositif d'enregistrement conditionne le traitement des données :*
 - *Modalité : audio ou video*
 - *Dispositif voyant ou discret*
 - *Complet ou partiel (angles morts, microphone uni/omnidirectionnel)*
 - *Permettant un montage (pauses, coupes ultérieures)*
 - *Manipulés par les chercheurs, par des techniciens, par les participants*
 - *Dans un lieu dédié/utilisé/"interdit-tabou"*

5.3 Un corpus oral : quels traitements ?

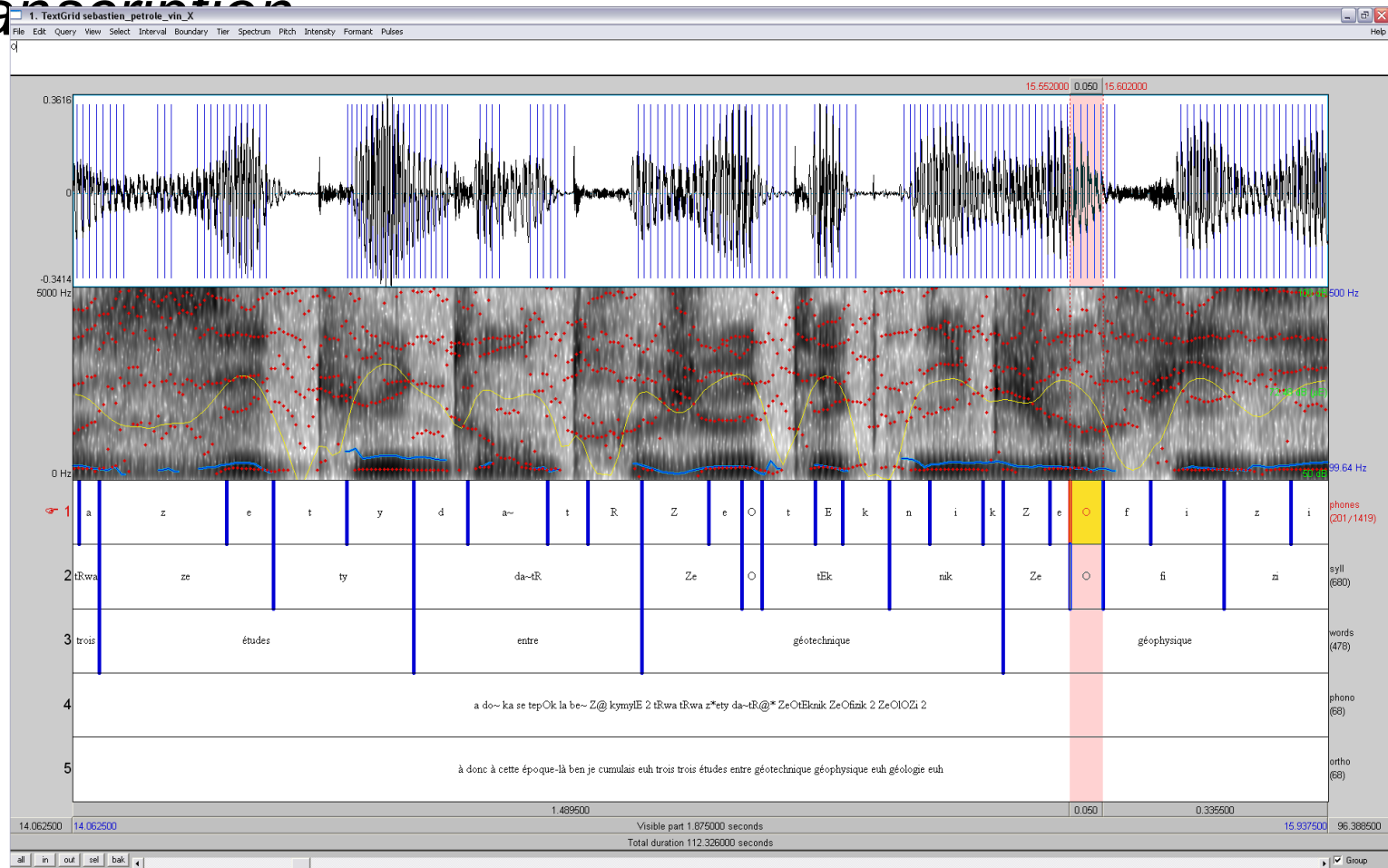
- ▶ Traitements des données : *transformation du terrain en un espace domestiqué conforme aux ordres des phénomènes recherchés et des analyses qu'ils subiront* [Mondada 18].

Pour cela le chercheur utilise des techniques qui :

- "tamisent" les données, en éliminant les bruits qui les entourent sur le terrain,
- rendent les données "compatibles" avec les analyses, les calculs, la formalisation dont elles feront l'objet,
- assurent la "comparabilité" des données.

5.3 Un corpus oral : quels traitements ?

- Traitements des données : *l'exemple de la transcription*



5.3 Un corpus oral : quels traitements ?

► Traitements des données : *l'exemple de la transcription*

The screenshot displays the Transcriber 1.5.1 application window. The main text area contains a transcription of an audio recording, with words highlighted in green. The transcription is as follows:

raport
Seb
géotechnique géologie géophysique
voilà j'ai touché un petit peu à tout ce qui a rapport avec euh le terroir le sous-sol l'étude géologique
tout ce qui est ben interprétation du du faciès géologique de surface et en profondeur
à donc à cette époque-là ben je cumulais euh trois trois études entre géotechnique géophysique euh géologie euh
bah pour faire son son trou un petit peu parce que
ce type de métier il est lié au prix du baril de pétrole
et à l'époque le prix du baril de pétrole était très faible nos nos professeurs nous disaient
tant que le baril fera pas vingt-cinq dollars vous ne trouverez pas de boulot
et donc ben j'ai joué un petit peu les mercenaires en faisant monter les enchères euh je suis allé à Orléans Paris Lille
Perpignan Toulon quand je suis remonté à Lille
je voulais acquérir un maximum d'expérience et puis ben je je me suis un petit aussi euh
carbonisé en voulant griller un griller un peu un peu les étapes
et puis voilà l'envie de se mettre à son compte j'ai commencé commerçant ambulant dans le Nord-Pas-de-Calais dans le domaine du vin j'ai
j'ai exposé des vin j'ai vendu
en deux-mille-quatre je me suis rapproché de la région
et les choses se sont faites comme ça
mes parents me l'ont pas donné j'ai j'ai acheté tout ce qu'il y a ici
simplement parfois avec un petit peu de conflit générationnel et conflit d'idées hein euh mais
ben on travaille avec des petits récoltants
donc quand j'ai repris la structure il fallait direct aller dans le bain
et non pas aller s'échapper faire des études pour pour compléter les connaissances
euh j'ai appris sur le tas et puis euh tout ce qui est lié au vin ben j'ai j'ai beaucoup lu euh
goûté appris et puis finalement vous savez c'est comme dans tout métier y a un certain nombre de mots de vocabulaire
qu'il faut mettre dans l'ordre qui sont les mêmes mots de français pour utiliser dans un autre contexte
on est une toute petite PE même TPE donc on est on est un petit peu fragile dans tout ça donc faut
donc faut voilà faut être vigilant faut bien gérer euh faut pas rouler en quatre-quatre euh
mais on se développe oui oui depuis huit ans on n'a jamais eu de de d'années en régression on est toujours en progression et
sachez que j'habite à Cepoy
que le mot terme de Cepoy vient du mot cep
que le cep désigne un pied de vigne
que la commune de Cepoy était à une époque productrice de vin
moi j'ai dans l'idée peut-être quelque part
de éventuellement peut-être réhabiliter une parcelle à Cepoy
ben vous allez rire mais j'habite rue des vignes à Cepoy

sebastien_petrole_vin_Cats
sebastien_petrole_vin.mp3

The bottom of the window shows a waveform of the audio file and a timeline with a cursor at 0.0.

5.3 Un corpus oral : quels traitements ?

► Traitements des données : *l'exemple de la transcription*

Seb:

Ben euh gé-géotechnique géologie géophysique voilà j'ai touché un petit peu à tout ce qui a rapport avec euh le terroir le sous-sol l'étude géologique tout ce qui est ben interprétation du du faciès géologique de surface et en profondeur à donc à cette époque-là ben je cumulais euh trois trois études entre géotechnique géophysique euh géologie euh bah pour faire son son trou un petit peu parce que ce type de métier il est lié au prix du baril de pétrole

5.3 Un corpus oral : quels traitements ?

► Traitements des données : *l'exemple de la transcription*

Une expérience ordinaire...

Voilà j'ai touché un peu à tout ce qui a rapport avec le terroir, le sous-sol, l'étude géologique. Tout ce qui est interprétation du faciès géologique de surface et en profondeur.

Donc à cette époque-là, je cumulais trois études entre géotechnique, géophysique, géologie. Pour faire son trou, car ce type de métier est lié au prix du baril de pétrole et à l'époque le prix du baril de pétrole était très faible., nos professeurs nous disaient : « *tant que le baril fera pas vingt-cinq dollars vous ne trouverez pas de boulot* » et donc j'ai joué un petit peu les mercenaires en faisant monter les enchères.

-6-

5.3 Un corpus oral : quels traitements ?

- Traitements des données : *l'exemple de la trans*



5.3 Un corpus oral : quels traitements ?

- ▶ Traitements des données :
 - ▶ Enrichissement (annotations)
 - ▶ "Appauvrissement" (destruction, caviardage pour anonymisation, etc.)
 - ▶ Codage
 - ▶ Catalogage (catégorisation)
 - ▶ Formatage
 - ▶ Standardisation/Normalisation
 - ▶ Structuration (Base de données)

5.4 Un corpus oral : quelles finalités ?

5.4 Un corpus oral : quelles finalités ?

- ▶ Un corpus est aussi défini par les buts du projet de recherche :
 - ▶ Analyses (diffusion des analyse : rapports, colloques, articles, ...)
 - ▶ Diffusion des données primaires
 - ▶ Diffusion des données enrichies (avec des outils de fouilles : concordancier, langage de requêtes, etc.)
 - ▶ Exploitation (ingénierie, didactique,...)
 - ▶ Conservation (stockage, stockage sécurisé, archivage pérenne)

6. Pratiques éthiques et juridiques

- ▶ **Le travail réflexif sur la démarche permet d'élaborer des pratiques qui tiennent compte des aspects juridiques et éthiques :**
 - ▶ **L'anonymisation**
 - ▶ **Le recueil de consentement**

6.1 L'anonymisation

- ▶ **Que doit-on anonymiser ? (Cf. CNIL)**

"...pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne"

- ▶ Identification directe ou indirecte
- ▶ Probabilité suffisante de rapprochement
- ▶ Le récepteur constitue un élément important

6.1 L'anonymisation

- ▶ **Que doit-on anonymiser ?**
 - ▶ repérer les éléments à traiter
 - ▶ mais aussi les opérations que vont subir les données
 - ▶ **Ce qui peut permettre d'identifier directement**
 - ▶ **Ce qui peut porter préjudice**
 - ▶ **Tout ce qui peut permettre indirectement l'identification**

6.1 L'anonymisation

- ▶ **Que doit-on anonymiser ?**
 - ▶ **Tout ce qui peut permettre indirectement l'identification**
 - ☐ formes nominatives
 - ☐ données personnelles
 - ☐ profession, statut, titres,
 - ☐ activités sociales,
 - ☐ parenté, réseaux,
 - ☐ référence à des lieux
 - ☐ référence à des caractéristiques de la personne
 - ☐ etc.
 - ▶ **Quelles données ?**
 - ☐ les données premières vidéo / audio,
 - ☐ les données premières textuelles
 - ☐ les données secondaires (transcription, métadonnées, analyses
 - ☐ les données secondaires visuelles (copies d'écran, spectrogrammes)

6.I L'anonymisation

- ▶ **Quand ?**
 - ▶ **préservation des données originales**
 - ▶ **préservation des données pour la conservation**
 - ▶ **accès restreint aux données non anonymisées**
 - ▶ **diffusion de données anonymisées**

6.1 L'anonymisation

- ▶ **Comment ?**
 - ▶ **Audio/video**
 - ☐ **Silence**
 - ☐ **Bruit**
 - ☐ **Signal + ou - déformé**
 - ▶ **Texte**
 - ☐ **Caviardage**
 - ☐ **Code**
 - ☐ **Hyperonyme**
 - ☐ **Pseudonyme,...**

6.2 Le recueil de consentement

- ▶ **Les titulaires des droits se sont-ils prononcés sur les conditions de mise à disposition et de réutilisation des corpus ?**

6.2 Le recueil de consentement

► Qui consent ?

Toute personne physique. Lorsqu'il s'agit d'une personne déclarée incapable (qu'elle soit majeure ou mineure), l'information doit parvenir au représentant légal. Pour le cas d'enfants, il faut l'autorisation des parents ou du dépositaire de l'autorité parentale.

6.2 Le recueil de consentement

- ▶ **Consentir à quoi ?**

Consentement éclairé: sans informations la demande d'autorisation n'a pas de sens

6.2 Le recueil de consentement

- ▶ **Consentir à quoi ?**
- ▶ **Le consentement "éclairé" selon la CNIL :**
 - ▶ l'identité du responsable du traitement et, le cas échéant, de celle de son représentant ;
 - ▶ la finalité poursuivie par le traitement auquel les données sont destinées ;
 - ▶ le caractère obligatoire ou facultatif des réponses ;
 - ▶ les conséquences éventuelles, à son égard, d'un défaut de réponse ;
 - ▶ les destinataires ou catégories de destinataires des données ;
 - ▶ l'existence d'un droit d'accès, de rectification voire d'opposition à la collecte.

6.2 Le recueil de consentement

- ▶ **Propositions pour un consentement éclairé :**
 - ▶ **Adéquation au destinataire**
 - ▶ **Finalités de l'enquête**
 - ▶ **Responsables de l'enquête**
 - ▶ **Adresse de contact,**
 - ▶ **Personnes qui auront accès aux données**
 - ▶ **Façon dont les données seront anonymisées,**
 - ▶ **Conventions particulières de transcription**
 - ▶ **L'archivage**
 - ▶ **Les modalités d'accès et de diffusion**
 - ▶ **Les risques éventuels**

6.2 Le recueil de consentement

- ▶ **Consentir à quoi ?**
 - ▶ les actions
 - ▶ les formats
 - ▶ les conditions de diffusion
 - ▶ les contextes de diffusion
 - ▶ les applications (manuel, méthode de langue,...)

6.2 Le recueil de consentement

- ▶ **Consentir quand ?**
 - ▶ **préparée à l'avance**
 - ▶ **juste avant**
 - ▶ **juste après**
 - ▶ **demande rapide avant et formalisée après**

6.2 Le recueil de consentement

- ▶ **Consentir comment ?**
 - ▶ - demande écrite et signée
 - ▶ - ou demande orale
 - ▶ - ou ...

Même si le législateur ne l'impose pas, le consentement écrit est considéré comme une bonne pratique. Dans des situations particulières, d'autres formes peuvent être choisies, l'important est de pouvoir faire la preuve de la volonté de la personne concernée (ex. : enregistrement d'un accord verbal).

[exemple]

7. *Corpus de la parole* : la réalisation



7. *Corpus de la parole* : la réalisation

1. Des données

État au 1^{er} mars 2013

- **Numérisation** : + de 1000 heures (français, franco-provençal, dialectes occitans, ndyuca (langue de Guyane), maoré (langues des Comores), berbère, langues kanaks, créole, judéo-espagnol), + de 40 langues.
- **Transcription** : + de 150 heures transcrites / traduites
- **Catalogage** Plusieurs milliers de documents décrits et catalogués

7. *Corpus de la parole* : la réalisation

2. Partager des bonnes pratiques de numérisation

1. Les Enregistrements

a) Normes pour l'audio

- pour la numérisation et la conservation: recommandations IASA*.
(choix validés par le conseil scientifique),
- pour la diffusion: WAV/PCM, 22KHz, 16bits, mono (choix opportuniste qui peut changer), MP3, streaming.

b) Normes pour la vidéo

- pour la numérisation et la conservation: format DV-PAL.
- pour la diffusion: MPEG-4 (choix opportuniste qui peut changer)

* WAV ou BWF PCM 96KHz 24bits (minimum 44.1KHz 16 bits)

7. *Corpus de la parole : la réalisation*

2. Partager des bonnes pratiques de codage

2. Les transcriptions

a) Normes

- Conservation: XML/Unicode avec une DTD spécifique -> TEI.
- Diffusion:
 - . animation multimédia: XHTML + realPlayer
 - . SMIL (à l'étude pour la vidéo)

Ce modèle, exprimé dans une DTD XML, définit une structure minimale permettant de coder la transcription d'un enregistrement ; d'ajouter une traduction en français (ce qui été demandé pour les langues autres que le français ou pour des transcriptions non orthographiques du français) ; de découper la transcription en segments (phrase ou groupe de souffle) ; de noter les repères temporel de début et de fin des segments.

Quelques raffinements du modèle permettent également d'indiquer le locuteur (utile pour les dialogues), le type de transcription (orthographique, phonétique, phonologique).

7. *Corpus de la parole* : la réalisation

3. Partager des bonnes pratiques de catalogage

3. Les Métadonnées

a) Stockage dans une Base de données

- Chaque ressource est décrite par un jeu de métadonnées.
- Les métadonnées sont stockées dans un format XML dans une Base de données native XML (eXist).
- Leur codage respecte un schéma explicite validé par le conseil scientifique. Le schéma reprend les recommandations d'OLAC (Dublin-Core-Spécifié enrichies de 5 éléments spécifiques à la communauté des linguistes [languageCode, linguisticType, linguisticField, discourseType, rôle])

7. *Corpus de la parole* : la réalisation

3. Partager des bonnes pratiques de catalogage

3. Les Métadonnées

dc:title. [1, 1]. Le titre.

dc:subject. [1, n]. Description du sujet du contenu de la ressource. Trois structures sont possibles :

dc:description. [0, n].

dcterms:tableOfContents. [0, n]. Description de type table des matières

dcterms:abstract. [0, n]. Description de type résumé

dc:type. [0, n]. Nature ou genre du contenu de la ressource.

dc:source. [0, n]. Référence à une ressource à partir de laquelle la ressource actuelle a été dérivée. Codages

dcterms:spatial. [0, n]. Couverture spatiale. En général le point d'enquête, sauf s'il ne représente absolument pas la

dcterms:temporal. [0, n]. Couverture temporelle de la ressource. A ne pas confondre avec les dates d'enregistrement,

dc:creator. [0, n]. Entité responsable de l'élaboration du contenu de la ressource (individu, institution, organisation)

dc:publisher. [0, n]. Entité responsable de la mise à disposition de la ressource, dans sa forme actuelle.

dc:contributor. [0, n]. Entités ayant contribué à la création du contenu de la ressource. Préciser leurs rôles par choix

dc:rights. [0, 1]. Indiquer la mention de copyright

dcterms:license. [1, 1]. URL de la licence creative-commons choisie

dcterms:created. [1, 1]. : date de création de la ressource.

dcterms:modified. [1, 1]. : date de dernière modification de la ressource.

dc:language. [0, n]. Langue du contenu intellectuel de la ressource. Langue de l'enquêteur dans un document sonore,

7. *Corpus de la parole* : la réalisation

4. Prendre en compte les aspects juridiques

- **Consentement**
- **Gestion des données personnelles**
- **Licences Creative Commons**

Selon les conditions suivantes :



Attribution — Vous devez attribuer l'œuvre de la manière indiquée par l'auteur de l'œuvre ou le titulaire des droits (mais pas d'une manière qui suggérerait qu'ils vous approuvent, vous ou votre utilisation de l'œuvre).



Pas d'Utilisation Commerciale — Vous n'avez pas le droit d'utiliser cette œuvre à des fins commerciales.



Partage dans les Mêmes Conditions — Si vous modifiez, transformez ou adaptez cette œuvre, vous n'avez le droit de distribuer votre création que sous une licence identique ou similaire à celle-ci.

7. *Corpus de la parole* : la réalisation

4. Prendre en compte les aspects juridiques

✓ Protection de la vie privée et données personnelles

- procédures d'anonymisation
- consentement éclairé
- gestion des BDD et Cnil

✓ Propriété intellectuelle et droits d'auteur

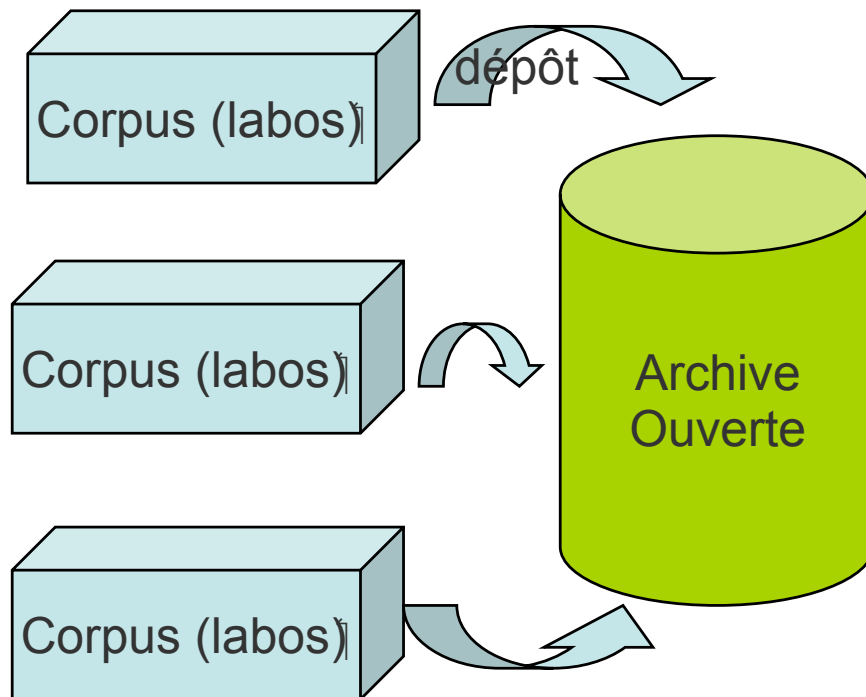
- statut de l'œuvre et de l'auteur
- droits moraux et droits patrimoniaux

Nécessité d'être en mesure d'explicitier la démarche du chercheur

7. Corpus de la parole : la réalisation

5. Architecture et services

L'Entrepôt



Le CRDO (2005-2011) recueille les corpus et les organise dans une "**archive ouverte**" (au sens OAI).

Dans cette archive, les données sont **normalisées**:

- les annotations en **xml**
- les enregistrements audio en wav/pcm

Les ressources sur le "français et les langues de France" constituent une collection

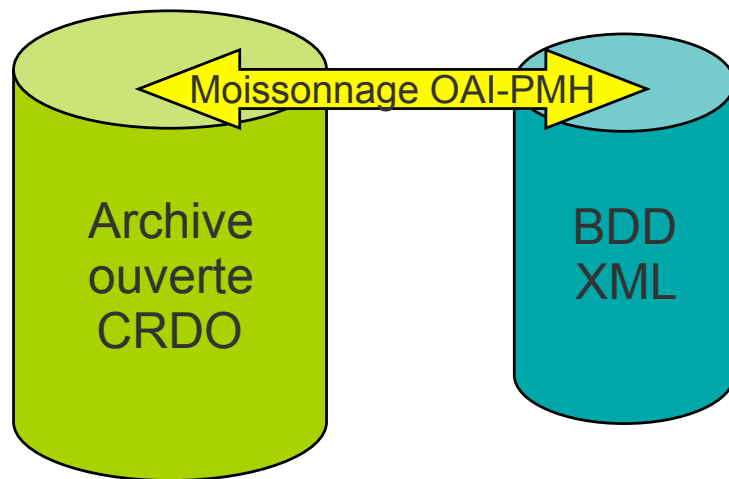
Chaque ressource est décrite par **des métadonnées codées en XML (Dublin-Core / OLAC)**

7. *Corpus de la parole* : la réalisation

Diffusion

- L'ensemble des métadonnées publiques de l'entrepôt est accessible via le protocole défini par l'OAI.
- Toute les ressources à destination du portail forment, de manière non-exclusive, une collection « langues de France » (correspond à un *set* OAI).

Architecture OAI



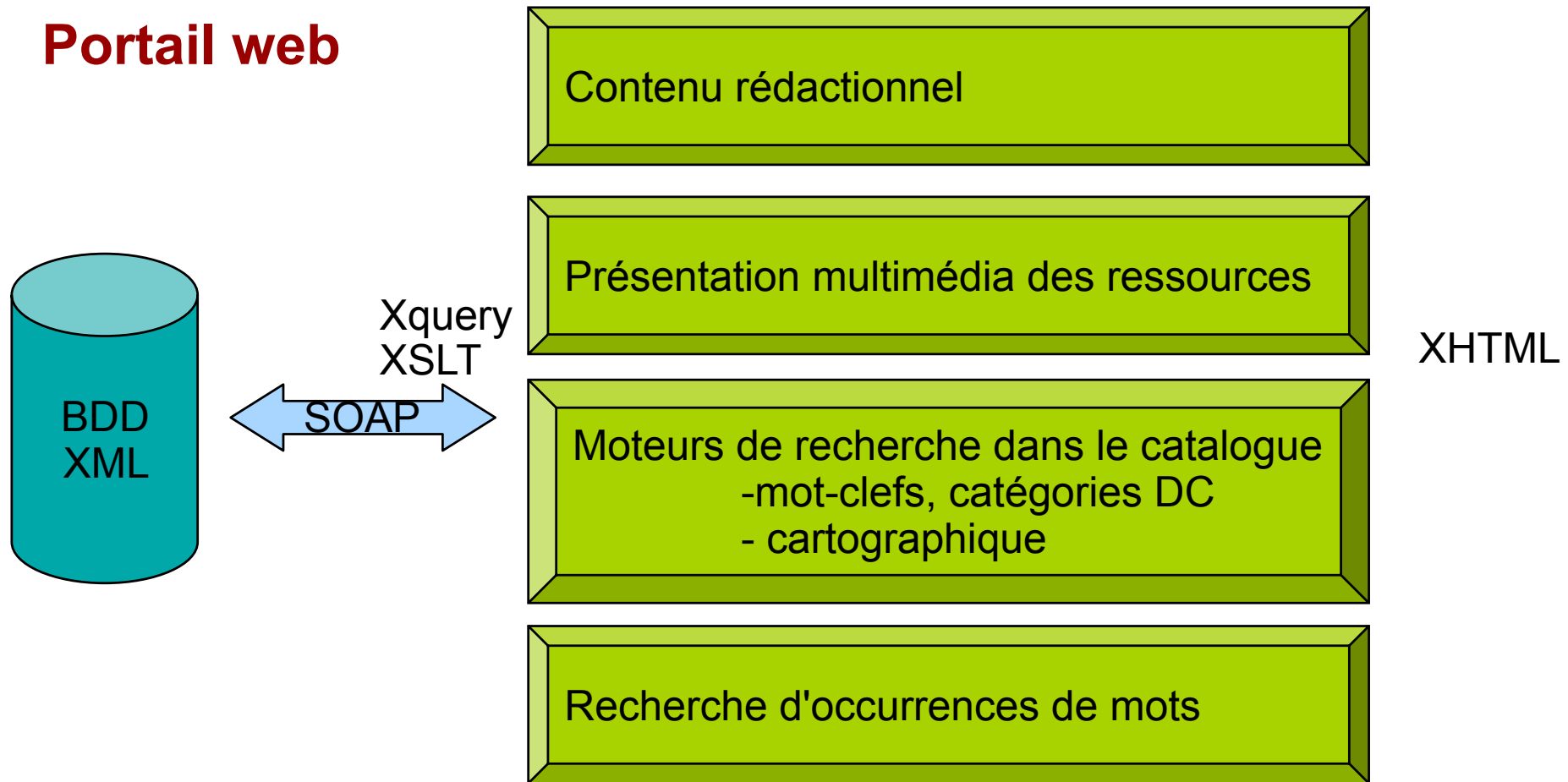
Moissonnage de l'archive ouverte de l'entrepôt:

- *set* "français et langues de France".
- *metadataPrefix*: olac
- tous les jours les nouveaux enregistrements ainsi que les enregistrements modifiés viennent mettre à jour la base de données XML

7. *Corpus de la parole* : la réalisation

Le portail *Corpus de la parole*

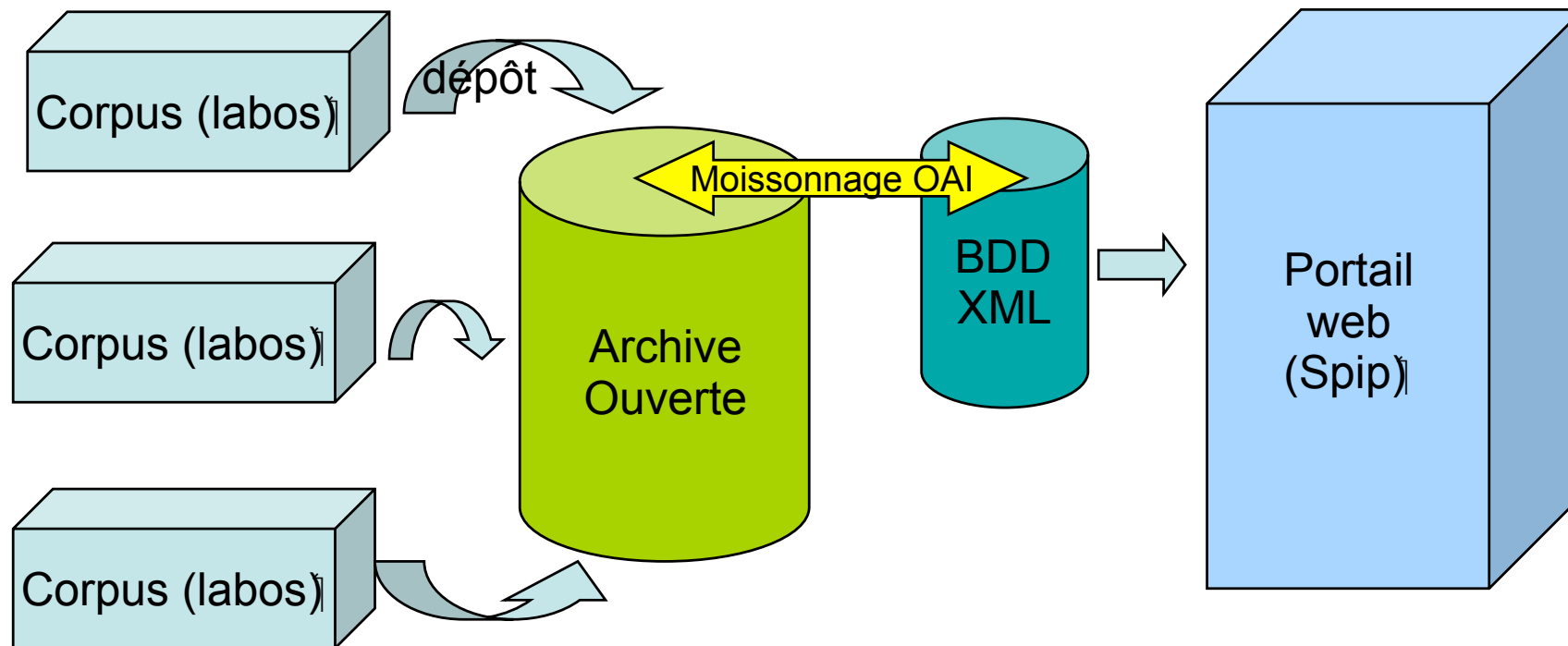
Portail web



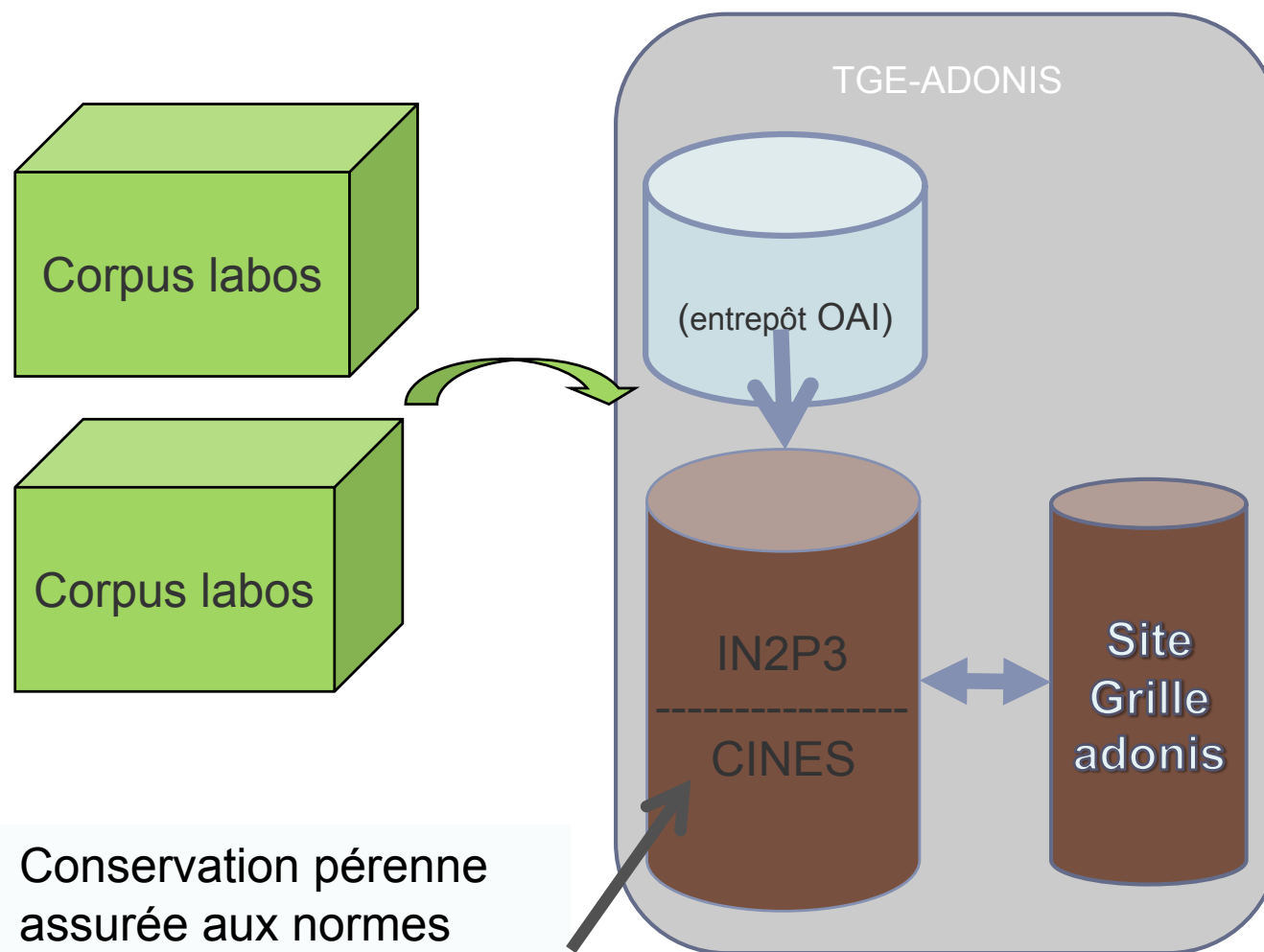
7. *Corpus de la parole* : la réalisation

Le portail *Corpus de la parole*

Architecture générale 2006-2011

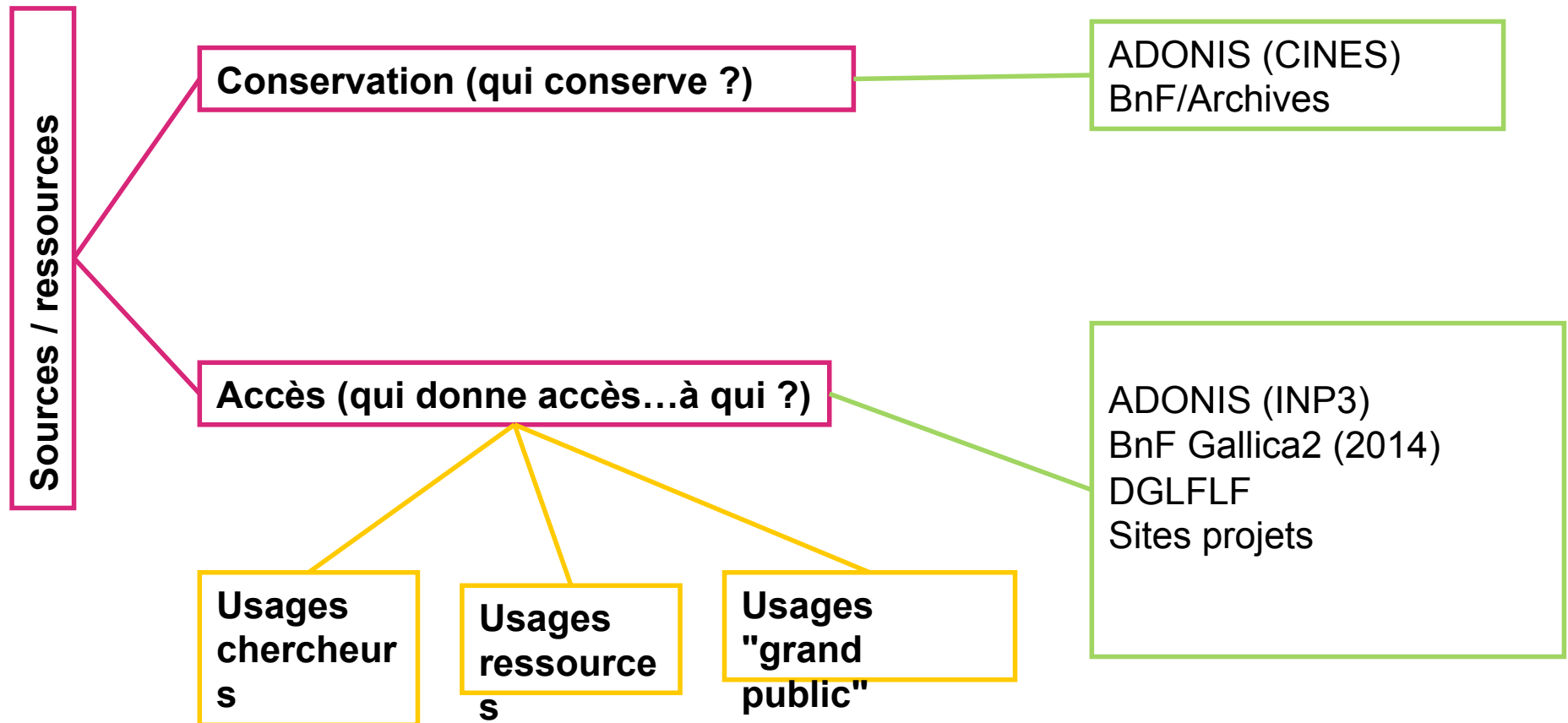


7. *Corpus de la parole* : la réalisation



7. *Corpus de la parole* : la réalisation

Travail commun CNRS Culture



7. Corpus de la parole : la réalisation

4. Le site "Corpus de la parole"

Le programme Corpus de la parole du ministère de la culture et de la communication a pour but de valoriser le patrimoine linguistique de la France. Il donne accès en ligne à des fonds sonores transcrits et numérisés, en français et dans différentes langues parlées sur le territoire national, en métropole et outremer. Ces langues sont considérées comme "Langues de France".

Ces corpus offerts à tous permettront de mieux appréhender la richesse de ce patrimoine linguistique.

On pourra :

- découvrir ces langues à partir d'un parcours sonore ;
- découvrir comment ces données ont été produites et comment on peut les exploiter.

Ce site est destiné aux curieux, aux amateurs avertis, aux chercheurs.

NOTA BENE

Les acteurs du projet:

Ce site a été réalisé dans le cadre d'un partenariat entre les Fédérations "Typologie et Universaux Linguistiques" (<http://www.typologie.cnrs.fr/>) et "Institut de Linguistique Française" (<http://www.ilf.cnrs.fr/>) du Centre National de la Recherche Scientifique - CNRS (<http://www.cnrs.fr/>) et la "Délégation générale à la langue française et aux langues de France" - DGLFLF (<http://www.dglflf.culture.gouv.fr/>) ainsi que la "Mission pour la recherche et la Technologie" du Ministère de la Culture et de la Communication (<http://www.culture.gouv.fr/>). La coordination de ce projet a été assurée par Benoît Habert et Stéphane Robert, pour le CNRS, et Olivier Baude et Jean Sibille, pour la DGLFLF. La réalisation a été effectuée par Stéphanie Girault et Michel Jacobson dans le cadre du Centre de Ressources pour la Description de l'Oral - CRDO-Paris (<http://crdo.risc.cnrs.fr/>) du CNRS, ainsi que par Julie Remfort pour la DGLFLF. Le "Relai d'Informations des Sciences Cognitives" - RISC (<http://www.risc.cnrs.fr/>) du CNRS assure l'hébergement du site.

Une quarantaine de chercheurs ont participé à ce projet en fournissant les données que vous allez découvrir. [Voir la liste des participants](#)

Evénements récents :

1911-2011 : Les Archives de la Parole ont 100 ans

Journée d'étude organisée par la BnF en collaboration avec la Délégation générale à la langue française et aux langues de France et le Laboratoire Ligérien de Linguistique

Vendredi 17 juin de 10h à 18h, BnF Site François-Mitterrand-Tolbiac, Paris, Hall Est- petit Auditorium. Entrée libre.

[voir plus de détails sur cette manifestation](#)

DES SITES QUI PARLENT DE LA PAROLE

Patrimoine numérique.

Patrimoine numérique, Catalogue des collections (...)

CLAPI

Projet "Corpus des Langues Parlées en Interaction"

CRDO

Le « Centre de Ressources pour la Description de l'Oral (...) »

[tous les liens](#)

[RSS](#) | mentions légales | contact | ©

7. *Corpus de la parole* : la réalisation

2. Le site "Corpus de la parole"

Le français, dans toute sa diversité

Les langues régionales (alsacien, basque, breton, occitan,...)

Les langues d'outre-mer (palikur, iaai, ndjuka,...)

Les langues non territoriales (romani, berbère, judéo-espagnol,...)

La Langue des Signes Française

Soit plus de **70 langues** "de France" (parlées par des citoyens français sur le territoire de la République depuis assez longtemps pour faire partie du patrimoine culturel national et qui ne sont langue officielle d'aucun État).

7. Corpus de la parole : la réalisation

2. Le site "Corpus de la parole"

Valoriser
des projets :

corpusdelap parole.in2p3.fr/spip.php?rubrique9

Eslo http://tools.google.c... Google CHEZ BAUDE (429) Nouvel onglet Banque de dépanna... Colloques Afficher les tickets - T... Connecter son ordina... théories des signes_f...

CORPUS DE LA PAROLE

Comment on parle en France aujourd'hui

→ accueil → pourquoi des corpus ? → les langues de France → rechercher ↓ projets et liens

Projets

Les corpus de langues parlées sont collectés dans le cadre de projets scientifiques d'envergure développés par les linguistes. Ces projets répondent à différents objectifs comme la description d'une langue, l'analyse linguistique, l'expertise ou la politique linguistique et sont tous fondés sur des corpus consultables sur ce site. Ils sont, parmi d'autres, représentatifs de la diversité des pratiques de chercheurs reconnus au meilleur niveau international. Voici quelques exemples de projets de recherche dans le cadre desquels ont été collectés les données mises en ligne sur ce site.

Le projet CLAPI (Corpus de Langue Parlée en Interaction)

Le laboratoire ICAR (ex-GRIC) (UMR 5191 du CNRS) mène à Lyon des recherches sur les interactions depuis une trentaine d'années et est reconnu internationalement dans ce domaine. Depuis quelques années, une banque de données de Corpus de Langue Parlée en Interaction, CLAPI, est développée dans le but d'assurer la sauvegarde de corpus anciens, à valeur patrimoniale et historique, et de stimuler la production et l'exploitation informatique de nouveaux corpus.

La base CLAPI compte en Janvier 2009 :

- ▶ 43 corpus, 315 enregistrements soit 133h, 510 transcriptions documentés par 75 descripteurs ;
- ▶ 50 h balisées pour des traitements d'analyses et de requêtes ;
- ▶ 35 h de données, soit plus de 70 % de la base en accès libre pour les analyses et les requêtes ;
- ▶ 10h d'enregistrements téléchargeables librement parmi 100 transcriptions documentant une grande diversité de situations interactionnelles (conversations quotidiennes, activités de travail, situations institutionnelles, interactions dans des commerces, en classe, chez le médecin, ...).

La base CLAPI ne se limite pas à gérer des corpus ; elle est avant tout fondée sur un savoir-faire développé par une équipe, cf site dédié à l'étude des interactions [CORINTE](#), qui concerne :

- ▶ le terrain : le recueil de données en situation « naturelle » repose sur une approche ethnographique qui permet d'identifier les lieux et les moments les plus pertinents et propices à la constitution du corpus.

7. Corpus de la parole : la réalisation

Cette rubrique vous permet de rechercher et de consulter des ressources dans le corpus de la parole selon les trois approches :

1) Recherche par langue
Vous pouvez soit :
- sélectionner une langue avec l'option liste (par défaut),
- choisir une région avec l'option carte, les corpus étant localisés géographiquement
Dans les deux cas, une liste des ressources correspondant à votre critère s'affiche avec :
- le nom de la ressource,
- une colonne écouter qui permet d'accéder à la fiche ressource (a),
- le nom du contributeur et la langue (utile si vous avez choisi 'toutes les langues' comme critère).
(a) La fiche ressource comprend l'enregistrement complet avec son player, la totalité de la transcription suivie de sa traduction et les métadonnées au format OLAC.

2) Recherche par mot
Vous pouvez rechercher un mot ou un phonème soit :
- dans la transcription donc dans la langue de celle-ci (option 1)
- dans la traduction donc en français (option 2)
Vous pouvez entrer le mot complet ou une expression régulière (une expression comprenant des caractères génériques pour traduire des propriétés comme commence/contient/et/ou/sauf/...)
Dans les deux cas, le nombre de résultats est affiché, avec pour chacun d'eux (10 résultats par page) :
- le nom de la ressource surlignée en première colonne, avec un lien à sa fiche (cf (a) ci-dessus)
- une icône représentant la touche play pour écouter la phrase où apparaît le mot recherché
- la transcription de la phrase, suivie de sa traduction (en gris), le mot trouvé est affiché en rouge soit dans la transcription (option 1), soit dans la traduction (option 2)

3) Recherche par les informations sur l'enregistrement (date, titre, sujet...)
Vous recherchez des ressources avec des critères portant sur les métadonnées c'est à dire les informations suivantes :
Titre, Editeur, Contributeur, Description, Date, Type, Format, Identifiant, Source, Langue, Relation, Couverture, Droits, Créateur, Sujet dans une langue donnée ou dans toutes les langues.
L'affichage est identique à la recherche simple par langue.

Recherche par langue

toutes les langues liste carte

Le symbole "+" sous la rubrique écouter vous renseigne quant à la disponibilité d'une transcription pour l'extrait sonore.

Corpus de la parole - chercher - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://www.corpusdelap parole.culture.fr/spip.php?rubrique3

Google

CORPUS DE LA PAROLE

Comment on parle en France aujourd'hui

→ accueil → entendre ↓ chercher → comprendre → en savoir plus

Vous trouverez dans cette rubrique plusieurs moteurs de recherche permettant de consulter sous plusieurs angles la base de donnée des corpus oraux. Il est par exemple possible de visualiser sur une carte les corpus géolocalisés. Il est également possible de chercher un mot ou un phonème dans le contenu de la transcription des corpus.

Recherche par langue

toutes les langues → Rechercher

liste carte

Recherche sur un mot

transcription → Rechercher

Recherche par les informations sur l'enregistrement (date, titre, sujet...)

* toutes catégories mots clefs toutes les langues → Rechercher 5 Lignes/Page

Terminé 93

Démarrer 3 Mozilla Th... Calendrier - ... 6 Microsoft ... site corpus d... Lecteur Wind... 2 Microsoft ... Acrobat Rea... Corpus de L... Rechercher sur l'ordi...

23:27

→ accueil

→ entendre

↓ chercher

→ comprendre

→ en savoir plus

Recherche par langue

toutes les langues

→ Rechercher

liste

carte



Terminé

94

Démarrer

3 Mozilla Th...

Calendrier - ...

6 Microsoft ...

site corpus d...

Lecteur Wind...

2 Microsoft ...

Acrobat Rea...

Corpus de L...

Rechercher s

→ accueil

→ entendre

↓ chercher

→ comprendre

→ en savoir plus

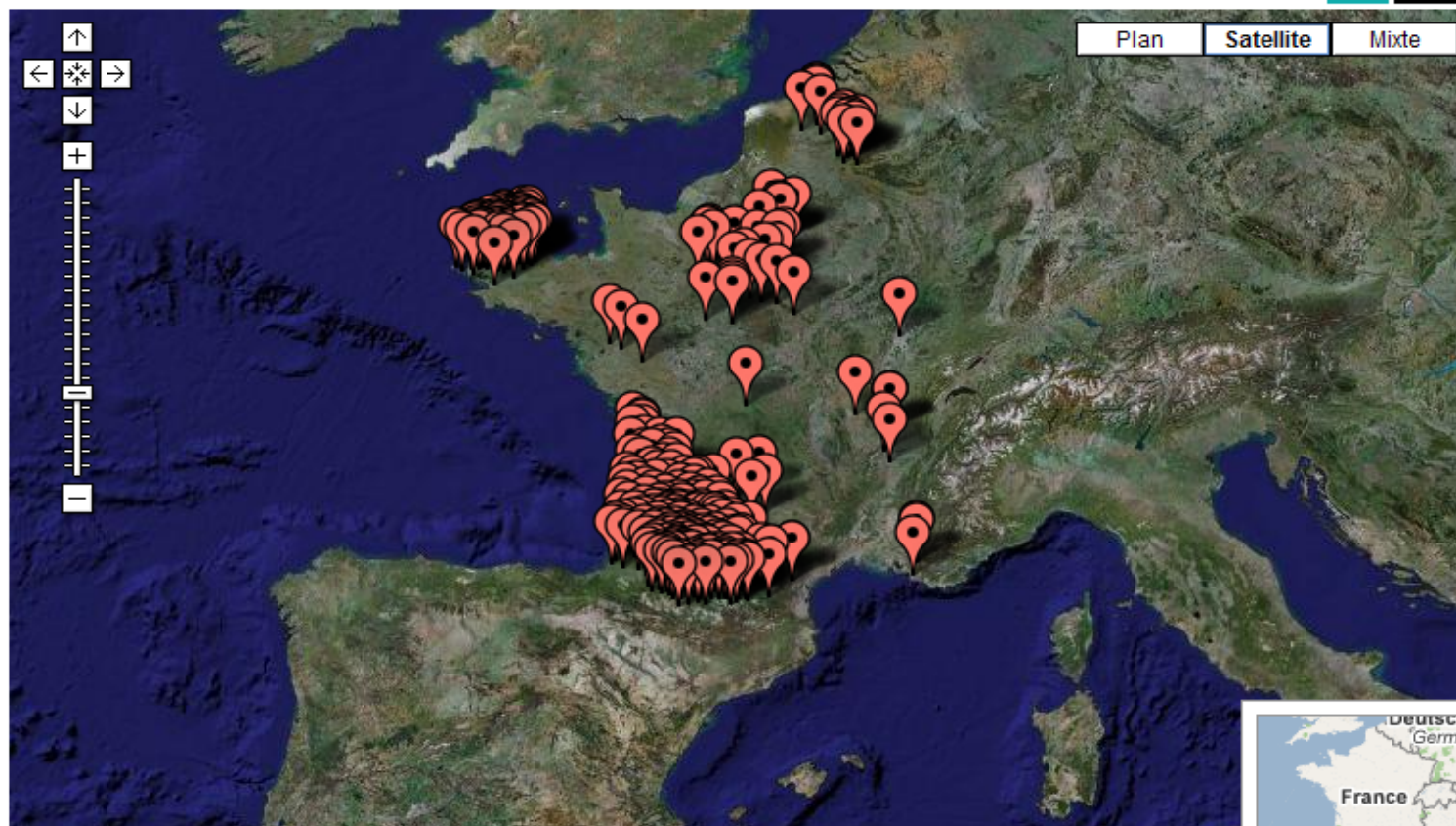
Recherche par langue

toutes les langues

→ Rechercher

liste

carte



CORPUS DE LA PAROLE

Comment on parle en France aujourd'hui

[→ accueil](#)[→ entendre](#)[↓ chercher](#)[→ comprendre](#)[→ en savoir plus](#)

Recherche par langue

[→ Rechercher](#)[liste](#)[carte](#)

titre	écouter	contributeur	langue
Corpus: CONVERSATIONS LYON SAXE (l...	 +	Lukas Balthasar et...	Français
ESLO - Omelette bande 131	 +	Université d'Orléans/CORAL	Français
ESLO - Omelette bande 115	 +	Université d'Orléans/CORAL	Français
ESLO - Omelette bande 120	 +	Université d'Orléans/CORAL	Français
ESLO - Omelette bande 055	 +	Université d'Orléans/CORAL	Français

page 1 sur 42 

ESLO 2006

CORPUS DE LA PAROLE

Comment on parle en France aujourd'hui

[→ accueil](#)[→ entendre](#)[↓ chercher](#)[→ comprendre](#)[→ en savoir plus](#)

Recherche par langue

[→ Rechercher](#)[liste](#)[carte](#)

titre	écouter	contributeur	langue
Corpus: CONVERSATIONS LYON SAXE (L...	 +	Lukas Balthasar et...	Français
ESLO - Omelette bande 131	 +	Université d'Orléans/CORAL	Français
ESLO - Omelette bande 115	 +	Université d'Orléans/CORAL	Français
ESLO - Omelette bande 120	 +	Université d'Orléans/CORAL	Français
ESLO - Omelette bande 055	 +	Université d'Orléans/CORAL	Français

CORPUS DE LA PAROLE

Comment on parle en France aujourd'hui

→ accueil

→ entendre

↓ chercher

→ comprendre

→ en savoir plus

Consulter une ressource

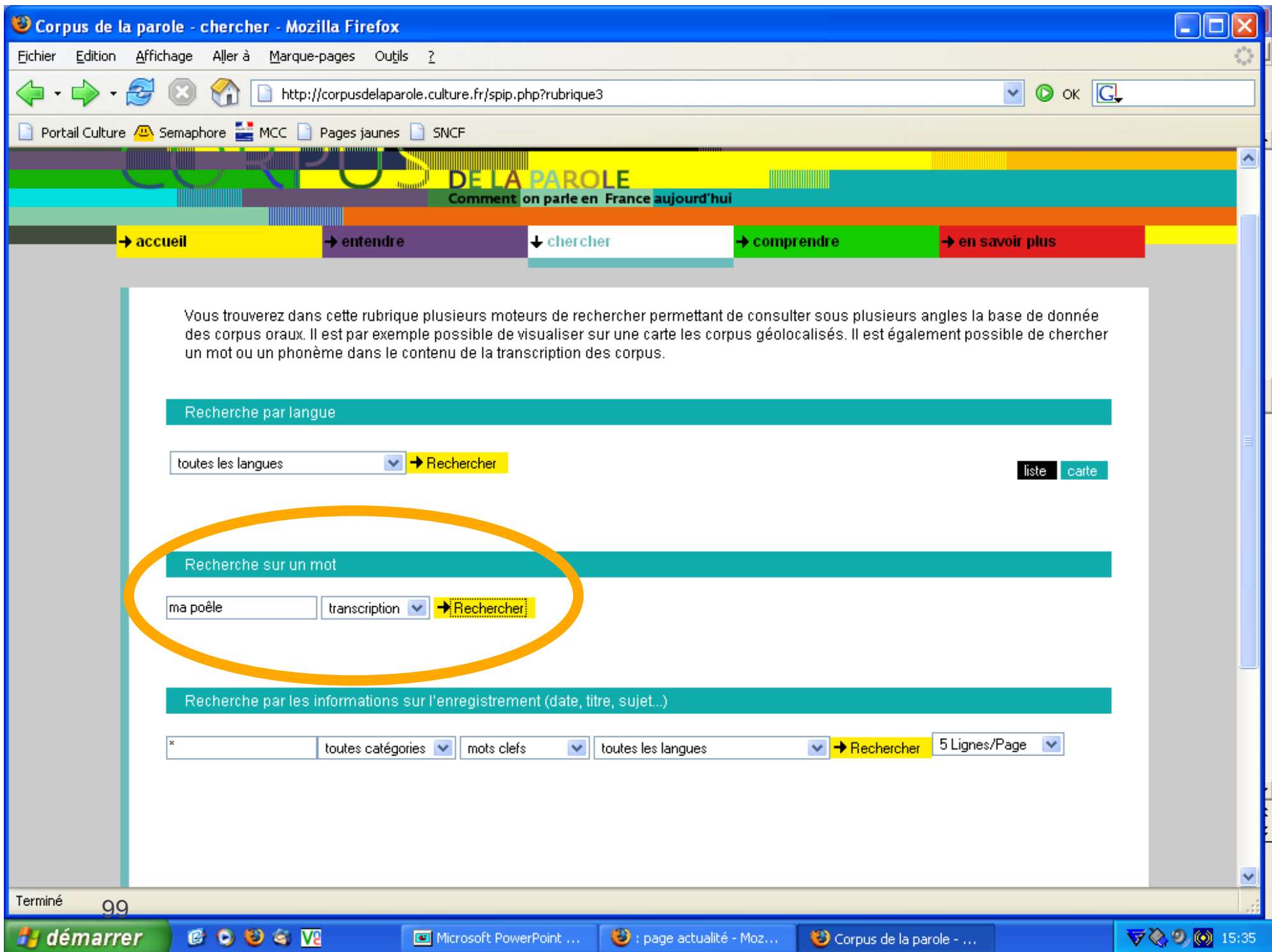


- JEB: ▶ L'Innovation alors l'innovation (0.2) alors y avait tout un paquet
- ▶ sur euh c'est important niveau du lancement/
- ▶ (0.3)
- SOP: ▶ ouais
- JEB: ▶ pourquoi pa'ce que t' as quarante pour cent du chiffre d'affaire/
- ▶ (0.1) qui sont des produits qu' ont moins d' cinq ans
- ▶ (0.9)

Informations

Corpus: CONVERSATIONS LYON SAXE (LSG35): EXTRAIT LSG 35: BISCUIT BUSINESS

javascript:playFrom("SAXE_BISCUITS6")



Corpus de la parole - Recherche sur un mot - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

http://corpusdelap parole.culture.fr/spip.php?article8

Portail Culture Semaphore MCC Pages jaunes SNCF

CORPUS DE LA PAROLE

Comment on parle en France aujourd'hui

→ accueil → entendre ↓ chercher → comprendre → en savoir plus

Recherche sur un mot

ma poêle transcription → Rechercher 15 élément(s) trouvé(s)

ESLO - Omelett... ▶ **ML 533:** des pommes de terre euh des pommes de terre je les coupe [pi:pronounce:instantaneous]je fais l'omelette un peu à **ma** manière à moi je coupe les() les pommes de terre petits en petits carrés je les coupe en petits carrés je bats les oeufs je laisse frire les pommes de terre après je fous les oeufs dans dans la **poêle** p(u)is je coupe du jambon je mets les tranches des petits bouts de jambon à l'intérieur

ESLO - Omelett... ▶ **NV 558:** bon bah j(e) mets [rire:noise:previous]mon beurre dans **ma poêle** [i:noise:instantaneous]je mets je prends des oeufs je les casse je les tourne avec une fourchette puis quand je vois qu'elle est bien faite et bien j(e) la mets d(e) dans p(u)is je la fait retourner (de) temps en temps avec une fourchette je je gratte un p(e)tit peu c'est tout

ESLO - Omelett... ▶ **1268:** j(e) prépare **ma poêle** je mets du beurre

ESLO - Omelett... ▶ **CW 739:** comment je fais une omelette [rire:noise:instantaneous]eh bien **ma** foi je prends une **poêle** je mets du beurre dans **ma poêle** et [b:noise:instantaneous]je casse des oeufs je les bats dans séparément naturellement pas dans la **poêle** dans un saladier à côté p(u)is quand mon beurre est fondu bah [rire:noise:instantaneous]j(e) fais mon omelette j(e) la fais cuire

ESLO - Omelett... ▶ **ES 187:** je prépare **ma poêle** je mets de la matière grasse j(e) la j(e) laisse chauffer assez bien et p(u)is je verse mes oeufs

http://crdo.risc.cnrs.fr/crdo_servlet/wave-cut?start=31.592&end=51.247&file=eslo/014.wav

démarrer Microsoft PowerPoint ... : page actualité - Moz... Corpus de la parole - ... 15:38

CORPUS DE LA PAROLE

Comment on parle en France aujourd'hui

→ accueil → entendre ↓ chercher → comprendre → en savoir plus

Vous trouverez dans cette rubrique plusieurs moteurs de recherche permettant de consulter sous plusieurs angles la base de donnée des corpus oraux. Il est par exemple possible de visualiser sur une carte les corpus géolocalisés. Il est également possible de chercher un mot ou un phonème dans le contenu de la transcription des corpus.

Recherche par langue

 → Rechercher[liste](#) [carte](#)

Recherche sur un mot

 transcription → Rechercher

Recherche par les informations sur l'enregistrement (date, titre, sujet...)

 Date toutes les langues → Rechercher

CORPUS DE LA PAROLE

Comment on parle en France aujourd'hui

[→ accueil](#)[→ entendre](#)[↓ chercher](#)[→ comprendre](#)[→ en savoir plus](#)

Recherche par les informations sur l'enregistrement (date, titre, sujet...)

 [→ Rechercher](#)

titre	écouter	contributeur	langue
ALIFO: Saint-Arnoult-des-Bois (1)		Simoni-Aurembou, Marie-Rose	Français, variation d'Île de France
ALIFO: Saint-Denis-les-Ponts (1)		Simoni-Aurembou, Marie-Rose	Français, variation d'Île de France
ALIFO: Saint-Denis-les-Ponts (2)		Simoni-Aurembou, Marie-Rose	Français, variation d'Île de France
Story of Mâgénéliné	+	Rivierre, Jean-Claude	Kwényii
ALLOc : Mouillac (1)		Ravier, Xavier	Occitan/Languedocien

 [▶](#)

8. Conclusion

- ▶ La mémoire scientifique : un tournant ?
 - ▶ TGIR ADONIS-CORPUS
 - ▶ Les initiatives diverses "aspects juridiques / éthiques"



<http://ethiquedroit.hypotheses.org/>

- ▶ Une nécessité : **"que le chercheur sache ce qu'il fait"**

olivier.baude@univ-orleans.fr

OLIVIER BAUDE

Laboratoire Ligérien de Linguistique UMR 7270
DGLFLF Ministère de la Culture *olivier.baude@univ-orleans.fr*