# RJCP 2023
# Formation sur les corpus oraux pour la science ouverte


Christelle Dodane – CLESTHIA – Université Sorbonne Nouvelle

christelle.dodane@sorbonne-nouvelle.fr

avec les précieux conseils de Delphine Charuau, Solaine Evain et Mathias Quillot

# Plan de la présentation

1. Les corpus oraux
2. Avant de collecter mon corpus
3. Sur le terrain : recueil des données et enregistrements
4. Après la collecte : transcrire les données
5. Diffuser et publier les corpus
6. Ressources sur les corpus

# 1) Les corpus oraux

# Qu’est-ce qu’un corpus oral ?

- Collection ordonnée de données langagières orales et/ou multimodales sélectionnées et organisées selon des critères linguistiques pour servir d’échantillon de langage (Sinclair, 1996)
- Un corpus oral est constitué de plusieurs éléments :
  - **Données primaires** (enregistrements) > terrain
  - **Données secondaires** (transcriptions), dérivées des données primaires
  - **Métadonnées** (descripteurs)

Tout corpus est une **construction**, au sens où il est toujours le produit des analyses du chercheur (E. Ochs, 1979)

# Pourquoi constituer un corpus oral ?

- Décrire et formaliser des faits linguistiques dans une population donnée
- Enrichir le patrimoine sur les langues et les pratiques linguistiques (UNESCO)
- Constituer de grands corpus de référence (ex. British National Corpus : http://www.natcorp.ox.ac.uk/)
- Développer le traitement automatique de la parole (reconnaissance et synthèse de parole, dialogues humain-machine) > ELRA : http://www.elra.info/

# Réemployer un autre corpus

- Utiliser les données d'autres chercheurs pour des fins de recherche : mutualisation > science ouverte (ex. BD CHILDES, Ortolang)
- Données réutilisables si :
  - anonymisées (données personnelles des participants masquées)
  - les participants ont donné leur consentement
- Précautions à prendre (métadonnées suffisantes, qualité des transcriptions > fichiers audio et/ou vidéo ; accord du chercheur)

Guide pratique de la publication en ligne et de la réutilisation des données publiques (CNIL : p. 19) :
https://www.cnil.fr/sites/cnil/files/atoms/files/guide_open_data.pdf

# 2) Avant de collecter mon corpus

# Constitution d'un corpus oral

- Objectifs de recherche visés (dimension développementale, sociale, clinique, etc.)
- Type de corpus (transversal, longitudinal, etc.)
- Taille du corpus
- Choix des locuteurs en fonction des objectifs de départ (ex. BD CLAPI), mode de recrutement des locuteurs (flyers, assoc., listes, etc.)
- Modalités d'enregistrement des données (données sollicitées vs données de parole continue)
- Tâches

**Protocole de collecte des données + Documentation**

# Exemple de constitution d'un corpus

- Projet AADI (« *Aphasie et Analyse du Discours* », Nowakowska, Praxiling, UPV) :
  - Locuteurs : personnes aphasiques et non aphasiques
  - Signature du consentement (patient, aidant)
  - 5 tâches différentes (cf. Aphasia Bank) :
    1. Série de question sur son quotidien et son AVC
    2. Description d'images
    3. Narration d'histoire (Cendrillon)
    4. Lecture de 10 phrases
    5. Entretien avec un proche (parole spontanée)
  - Document associé : bilan orthophonique

https://cnrs.hal.science/hal-03913435

# Quelle réglementation respecter ?

- Dans nos travaux en SHS, implications de la personne humaine et utilisation de données personnelles
- Mise en application de deux cadres s'appliquant à tous les secteurs faisant intervenir la personne humaine :
  - **Loi Jardé** (2012)
  - **RGPD** (2018)

Lalain et al. (2021) : De la protection des données à la protection de la personne.

https://journals.openedition.org/corpus/5895

# RGPD

- Règlement Général sur la Protection des Données (amplification du pouvoir coercitif de la CNIL)
- Cadre européen concernant le traitement et la circulation des données à caractère personnel (depuis le 25 mai 2018)
- Réponses à de nombreuses questions > documentation
- Enregistrée auprès du **Délégué à la Protection des Données** (**DPO**) de votre structure : registre dédié au traitement des données

Lalain et al. (2021) : De la protection des données à la protection de la personne.

https://journals.openedition.org/corpus/5895

https://www.cnil.fr/fr/donnees-personnelles

# Loi Jardé (2012)

- Simplification de la loi de protection des personnes applicable aux personnes se prêtant à la recherche biomédicale
- Définir le type de recherche pour respecter des règles adaptées :
  - Implications de la personne humaine (**Loi Jardé**)
  - Recherche interventionnelle / interventionnelles à risque et contraintes minimes / non interventionnelle (**RIPH**) > **CPP** (Comité de Protection des Personnes)
- Dans tous les cas :
  - Information / consentement du participant
  - Justification du respect de la protection des données des participants
  - **Délégué à la Protection des Données** (DPO) de votre structure

https://soepidemio.com/2020/01/22/la-loi-jarde-en-4-minutes/

# CER ou CPP ?

- Avant toute collecte, qualification du protocole (CER vs CPP)
- Toute recherche portant sur la personne humaine visant à augmenter les connaissances médicales ou biologiques est considérée comme une RIPH (Recherche Impliquant la Personne Humaine) et doit passer devant un **CPP**
- Si ce n'est pas le cas > Comité d'Ethique de la Recherche (CER)
- Formulaire d'auto Qualification (fédération des CER) :

https://www.federation-cer.fr/cer-ou-cpp-pas-simple.../comment-savoir-si-on-doit-demander-un-avis-ethique-a-un-cer-ou-a-un-cpp,24565,40663.html

# Lettre d'information + Consentement éclairé

- Consentement (écrit et/ou oral) : il implique qu'une « *personne physique affirme clairement qu'elle accepte que les données personnelles la concernant fassent l'objet d'un traitement* » (Baude, 2006 : 113)
- Cette demande d'autorisation dépend de « l'information préalable » (Baude, 2006 : 113) qui implique que le participant soit mis au courant des raisons de la collecte des données ainsi que des différents traitements qui en seront faits

Fiche juridique : « Le consentement » (Baude et al., 2006, Guide des Bonnes Pratiques, p.113)

https://hal.science/hal-00357706/file/Corpus_Oraux_guide_des_bonnes_pratiques_2006.pdf

# Plan de gestion des données / Cycle de vie

- Plan de gestion des données (document formalisé qui explique la manière dont seront obtenues et traitées les données tout au long de leur cycle de vie, de leur collecte à l'archivage)
- À l'issue du traitement, le RGPD oblige à ce que les données personnelles soient détruites, anonymisées ou archivées, mais elles peuvent être conservées plus longtemps à des fins de recherche
- Dépôt des données « anonymisées » dans des entrepôts nationaux (Nakala, Ortolang, etc.)

Lalain et al. (2021) : De la protection des données à la protection de la personne.

https://journals.openedition.org/corpus/5895

https://www.cnil.fr/fr/donnees-personnelles

# Principales obligations liées à la collecte et au traitement des données personnelles

- Sécuriser les fichiers (sécurité des locaux et des systèmes d'information).
- S'assurer de la confidentialité des données.
- Indiquer, avec précision, le but de la collecte et du traitement des données (*principe de finalité*).
- Fixer la quantité des données personnelles à collecter et leur durée de conservation en fonction de l'objectif poursuivi (*principe de proportionnalité*).
- Permettre l'information des personnes concernées par l'étude.
- Soumettre à l'autorisation de la CNIL les traitements informatiques de données personnelles qui présentent des risques particuliers d'atteinte aux droits et aux libertés.
- S'assurer de la cohérence des informations exploitées dans un fichier par rapport aux objectifs.

https://comite-ethique.cnrs.fr/wp-content/uploads/2019/10/GUIDE-2017-FR.pdf

CNIL

# 3) Sur le terrain : recueil des données et enregistrements

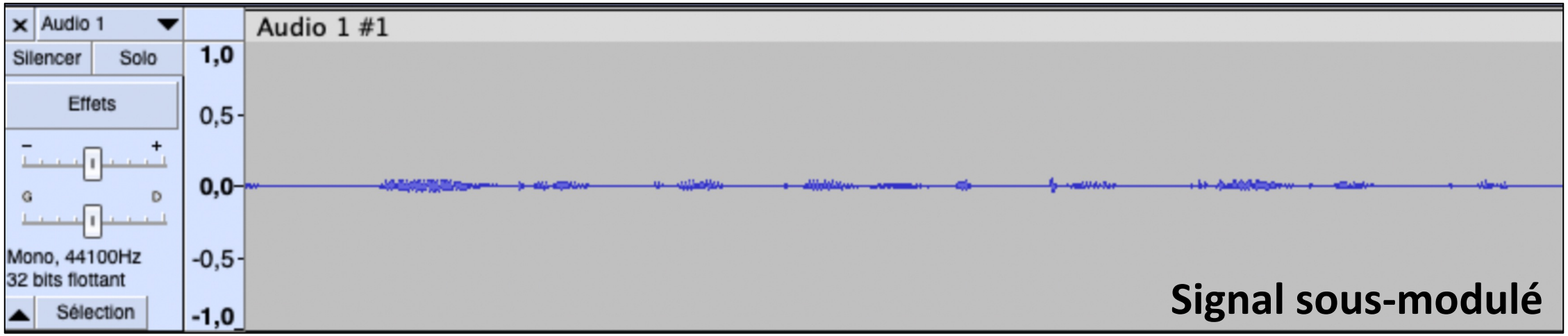
# Enregistrements audio

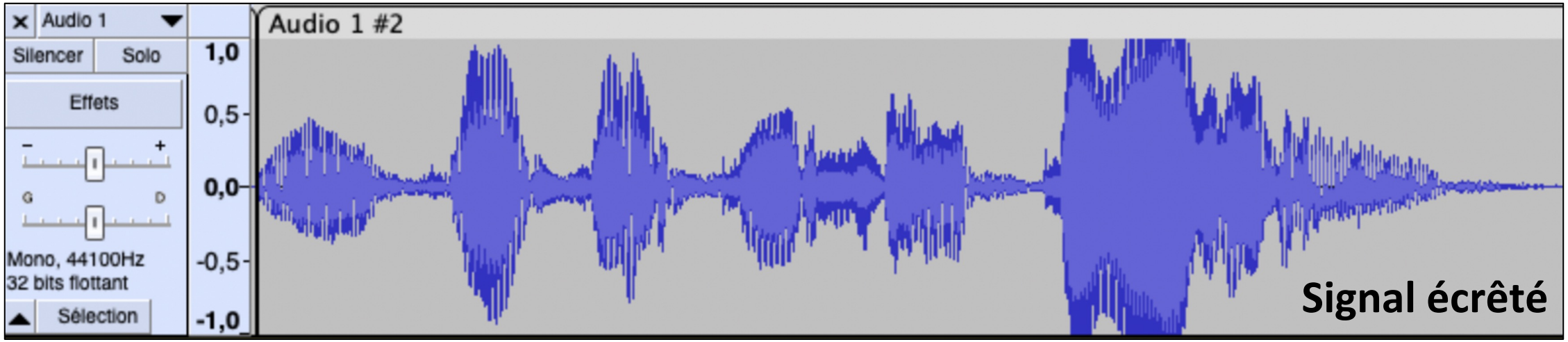
- Fréquence d'échantillonnage (Hz) : recommandations IASA : WAV ou AIFF ou PCM 96 KHz 24 bits (minimum 44.1 KHz, 16 bits)
- Ne jamais enregistrer un fichier son dans un format compressé (aac, mp3, etc.)
- Utiliser une phrase avec niveaux élevés pour régler les gains (du type : « *Papa est parti pour Paris* »)
- Canaux (mono, stéréo, multipiste)
- Traitement : Audacity (https://www.audacityteam.org/), sox ou ffmpeg (Linux)

Vincent (2017), L'acquisition et le traitement de données multimodales en linguistique

https://hal.science/hal-01225952v1/file/COLDOC-VINCENT.pdf

# Enregistrements audio



# Précautions à prendre

- Vérifier que les équipements sont adaptés et que le micro soit bien placé
- Éteindre les sources de bruit
- Plusieurs participants > plusieurs micro (cravate) en plus du micro global
- Eviter de mettre l'enregistreur sur la table où est posée le micro
- Pop : Filtre anti-pop ou décaler le micro sur le côté
- Enregistreurs portables (plutôt sur batterie que sur secteur)
- Cartes SD non pleines ☺ !
- **Tester le matériel et le protocole avant !!**

Vincent (2017), L'acquisition et le traitement de données multimodales en linguistique

https://hal.science/hal-01225952v1/file/COLDOC-VINCENT.pdf

# Enregistrements vidéo

- **Fond bleu :** qualité d'image optimale
- **Normes pour la vidéo :**
  - pour la numérisation et la conservation : format DV-PAL
  - pour la diffusion: MPEG-4 (choix opportuniste qui peut changer)
  - pour convertir un fichier vidéo en fichier son : ShareDocs, Handbrake, Ffmpeg
- **Synchronisation plusieurs sources :** clap
- **Editeur vidéo** (libre d'accès) : Avidemux
- **Annotation vidéo** : kinovea

Vincent (2017), L'acquisition et le traitement de données multimodales en linguistique

https://hal.science/hal-01225952v1/file/COLDOC-VINCENT.pdf


ANR DINLANG

# ANR STACCATO

00:30:06:00

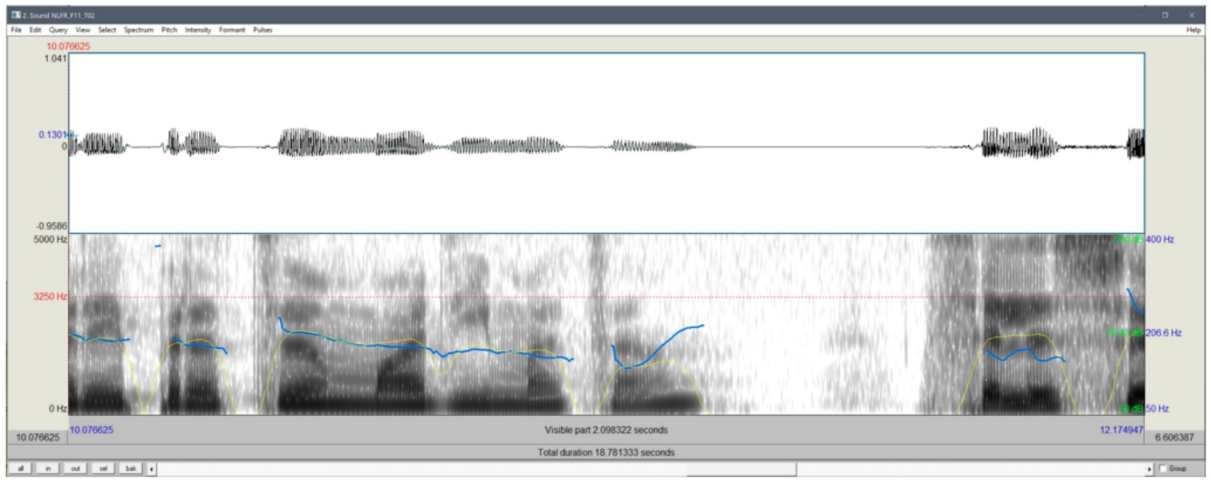
# Bonnes et mauvaises pratiques lors de la constitution d'un corpus

| **Doit** | Ne **doit pas** |
|---|---|
| ✓ **Tester** divers type de matériel. | ✗ **Ne pas** connaître le matériel approprié à sa tâche. |
| ✓ Essayer le matériel choisi. | ✗ **Se rendre sans préparation** sur le lieu d'enregistrement en utilisant l'appareil pour la première fois. |
| ✓ Penser à avoir des piles neuves ou charger son appareil s'il fonctionne sur batterie. | ✗ Oublier de se munir de piles neuves pour l'appareil d'enregistrement ou oublier de le charger. (Voir anecdote) |
| ✓ **Avoir** ses documents de collecte de données. | ✗ **Oublier** les documents importants pour la tâche. |

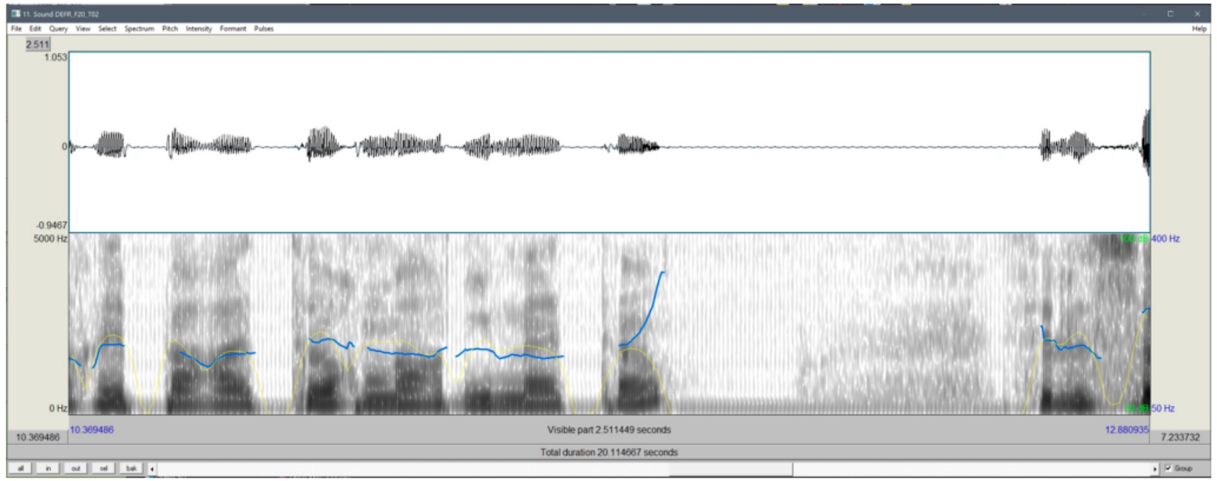
https://corli.huma-num.fr/bonnes-pratiques-pour-la-constitution-de-corpus/

# Bonnes et mauvaises pratiques lors de la constitution d'un corpus

- **Exemple** : Branchement de l'enregistreur sur secteur : buzz à 50 Hz




Spectrogramme normal




Spectrogramme courant électrique

https://corli.huma-num.fr/bonnes-pratiques-pour-la-constitution-de-corpus/

# 4) Après la collecte : Transcrire les données

# Normes / standards de transcription

- Ensemble de choix à faire : quels phénomènes transcrire ? (*granularité de la transcription*, Mondada, 2008)
- Dépendent de la finalité de la recherche
- Découlent du cadre théorique et de la problématique de la recherche (Mondada, 2008)
- Problèmes de lisibilité ?
- Fins d'analyse et/ou de publication ?
  - Cf. Corpus COLAJE (Ortolang : https://ct3.ortolang.fr/data/colaje/)

# Normes / standards de transcription

- Choix du type de représentation graphique pour transcrire l'oral
- Trois possibilités :
  - Orthographe adapté, proche de la production orale
  - Orthographe non adapté (standard), facile à lire
  - Transcription phonétique fidèle à la production orale mais chronophage et difficile à lire
- Deux types de format : texte (CLAN) ou partition (ELAN)
- Guide d'annotation

https://lidilem.univ-grenoble-alpes.fr/sites/lidilem/files/Mediatheque/Documents/Corpus/anrmultimodalite-manueldecodage.pdf

https://corli.huma-num.fr/guides-dannotation/

# Normes / standards de transcription

| Normes | Lien | Domaine | BD associée |
|---|---|---|---|
| API | https://www.internationalphoneticassociation.org/ | Phonétique | |
| TOS | https://www.projet-pfc.net/ | Phonologie du Français Contemporain | PFC : https://www.projet-pfc.net/ |
| CHAT | https://talkbank.org/manuals/CHAT.pdf | Acquisition du langage, données cliniques, etc. | https://childes.talkbank.org/ https://www.talkbank.org/ |
| ICOR | http://icar.cnrs.fr/ecole_thematique/tranal_i/documents/Mosaic/ICAR_Conventions_ICOR.pdf | Parole spontanée | CLAPI : http://clapi.icar.cnrs.fr/ |
| TCOF | https://www.cnrtl.fr/corpus/tcof/TCOFConventions2017.pdf | Corpus oraux en français (ATILF) | TCOF : https://tcof.atilf.fr/ |
| GAT / GAT2 | https://kops.uni-konstanz.de/handle/123456789/38351 | Analyse conversationnelle, analyse de discours | |

# Exemple 1 : Convention GAT / GAT 2

Pause de 200 ms

```
08    JEFF:    i an' i(.) (0.2) jus' talked to this (0.3) Asian guy,
               h°
09             u:m
10             °hh who:'s twenny: h° six years OLD;
11             and
→12            (0.9) ((click)) he's a !VE:R!y: sweet guy;=
13             =he jus' moved to laGUna.
```

Inspiration/expiration (entre 200 et 500 ms)

Allongement

Accent de focalisation très fort

Suivi immédiatement d'un tour de parole ou d'un segment

Action non verbale

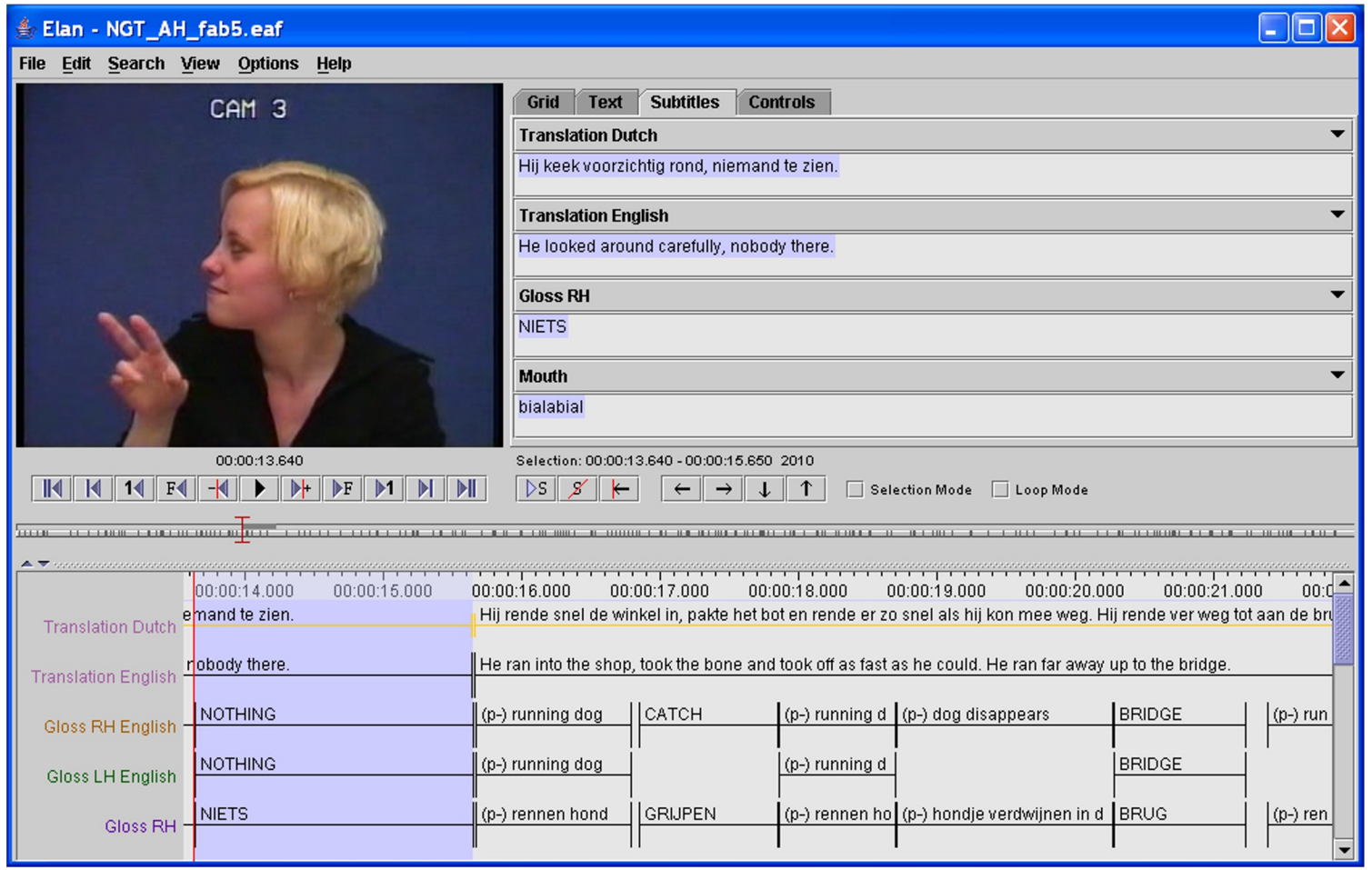
# Exemple 2 : Convention CHAT

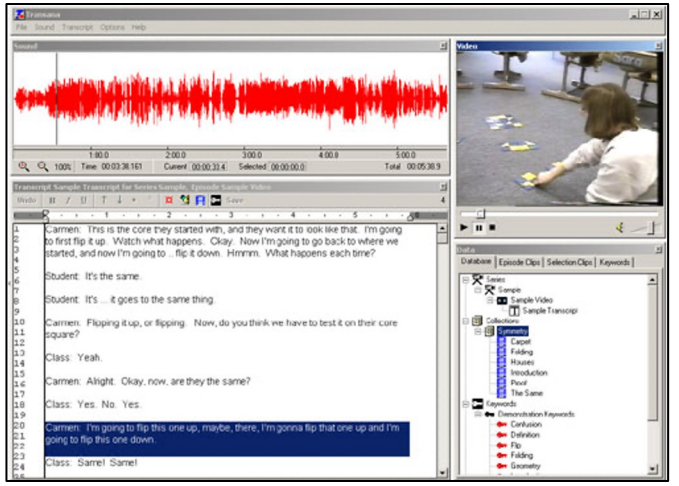
*MOT: +, les garçons sont très [>] contents !
*CHI: yy [<] .
*CHI: yy (.) 0 [=! rit] !
*OBS: oh c'est quoi ?
*OBS: un p(e)tit bout ?
*CHI: yy [>] .
*MOT: <tu> [<] le manges ?
*CHI: ah œ@fs yy (.) 0 [=! rit] !
*MOT: tu le manges ?
*MOT: mange le [/] le petit bout tu vois parce+qu' après +...
*MOT: +, on le voit plus !
*MOT: on sait plus où il est !
*CHI: ah !
*CHI: ah !
*MOT: tu l' as trouvé ?
*MOT: Anaé tu veux bien aller fermer la porte regarde y+a (.) Omer qui a ouvert la porte .

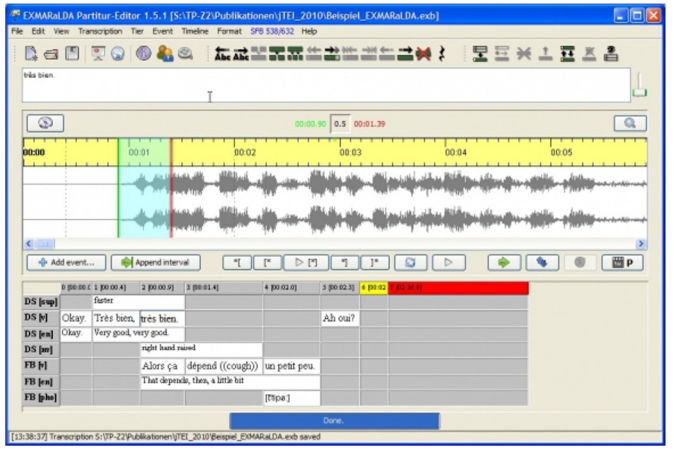
# Importance de la transcription

- Transcription > étape clé dans l'exploitation des données puisqu'elle conditionne l'analyse et doit être pensée en fonction de la recherche
- Compromis entre la lisibilité et le degré de granularité du système
- Choix du système de transcription en fonction de ses objectifs de recherche :
  - Si je travaille sur les enfants > CHAT
  - Si je veux réaliser une analyse conversationnelle > GAT
  - Si je veux transcrire la prosodie, les gestes, etc.
- Importance du choix du logiciel d'aide à la transcription

# Logiciels d’aide à la transcription manuelle





# Logiciels d'aide à la transcription manuelle

| Logiciel | Lien | Domaine | Données |
|---|---|---|---|
| CLAN | https://dali.talkbank.org/clan/ | Large (acquisition, clinique, etc.) | audio/vidéo |
| PRAAT | https://www.fon.hum.uva.nl/praat/ | Phonétique, phonologie | audio |
| PHON | https://www.phon.ca/phon-manual/getting_started.html | Phonétique, phonologie | audio |
| ELAN | https://archive.mpi.nl/tla/elan/download | Études gestuelles | vidéo |
| TRANSANA | https://www.transana.com/ | Analyse de discours et conversation | audio/vidéo |
| EXMARaLDA | https://exmaralda.org/en/ | Analyse du discours, dialectologie, sociolinguistique, etc. | audio/vidéo |
| ANVIL | http://www.anvil-software.de/ | Études gestuelles, capture de mouvements | audio/vidéo |

# Interopérabilité : TEI Convert (Parisse & Madjoub)

**1) Choisir le Format Destination**

- (•) TEI (xml / tei_corpo.xml / teiml / trjs)
- ( ) TRS (transcriber)
- ( ) CHA (chat - childes)
- ( ) TXT (texte - utf8)
- ( ) DOCX (microsoft word)
- ( ) XLSX (microsoft excel)
- ( ) CSV (tableurs)
- ( ) TEXTGRID (praat)
- ( ) EAF (elan)
- ( ) TXM (xml/w)
- ( ) Lexico/Le Trameur (.txt)

https://ct3.ortolang.fr/teiconvert/index-fr.html

# Transcription automatique (1/2)

- Outils comme « **whisper** » (ShareDocs / Humanum)
- Grâce à **TEI CONVERT** -> transformation du fichier au format cha, eaf, etc.
- **Alignement phonétique** : Easy Align, Train&Align, SPPAS, etc.
- **Attention : tout traitement automatique implique d'être repris à la main !!**


https://www.huma-num.fr/

http://icar.cnrs.fr/tutoriel-retranscrire-des-entretiens-avec-whisper-via-huma-num/

# Transcription automatique (2/2)

- **Noota** (https://fr.noota.io/solutions/noota-for-academic-researchers) qui permet la transcription de 5h de données gratuitement par mois
- **Speechbrain** -> systèmes d'ASR déjà entraînés https://huggingface.co/speechbrain/asr-wav2vec2-commonvoice-fr)
- **Etranslation** (https://commission.europa.eu/resources-partners/etranslation_fr)

# Exploitation des corpus

- CLAN (acquisition du langage) : ensemble de commandes pour le traitement automatique des corpus (MLU, TTR, VOCD, DSS, etc.)
- PRAAT (visualisation du signal) : traitement automatique des analyses acoustiques (scripts)
- EXMARaLDA : logiciels annexes pour traiter et éditer des corpus entiers
- ELAN (https://ct3.ortolang.fr/toolselan/statselan/html/statselan.html)
- PHON
- R
- etc.

# 5) Diffuser et publier les corpus

# Evaluer l'annotation manuelle

- Accord inter-annotateurs (ou accord « *entre juges* ») : degré d'adéquation entre plusieurs annotateurs (de 2 à n)
  - Comparaison entre deux annotateurs (S. de Bennett, $\pi$ de Scott, 1855 ; Kappa de Cohen, 1960)
  - Comparaison entre plusieurs annotateurs (Multi-$\pi$ de Fleiss, 1971 ; Multi-k de Davies et Fleiss, 1982)
  - Accord de catégorisation ($\alpha$ de Kripeendorf, 2013 ; kw de Cohen, 1968)

Baledent (2017), De la complexité de l'annotation manuelle

https://www.theses.fr/2022NORMC253

# Anonymiser les données

- L'anonymisation est « *l'opération par laquelle se trouve supprimé d'un ensemble de données recueillies auprès d'un individu ou d'un groupe tout lien permettant l'identification de ces derniers* » (Baude, 2006 : 193).
- Elle doit opérer à la fois sur les données primaires (bipper, brouiller, déformer le son, flouter, inverser les couleurs, etc.) et secondaires
- Données anonymisées (non soumis à RGPD) vs pseudonymisées (table de correspondance > soumis à RGPD)

Fiche juridique : « Anonymisation » (Guide des Bonnes Pratiques, 2006)

https://hal.science/hal-00357706/file/Corpus_Oraux_guide_des_bonnes_pratiques_2006.pdf

# Finaliser son corpus - Principes FAIR (Union européenne, H2020)

Gestion des données de la recherche qui doivent être FAIR càd :

- Faciles à trouver (métadonnées riches, HAL, DOI)
- Accessibles (stockage pérenne, accès libre et gratuit)
- Interopérables (langages et formats ouverts, cf. TEI)
- Réutilisables (métadonnées)

https://www.ccsd.cnrs.fr/principes-fair/

# Diffuser son corpus sur une base de données

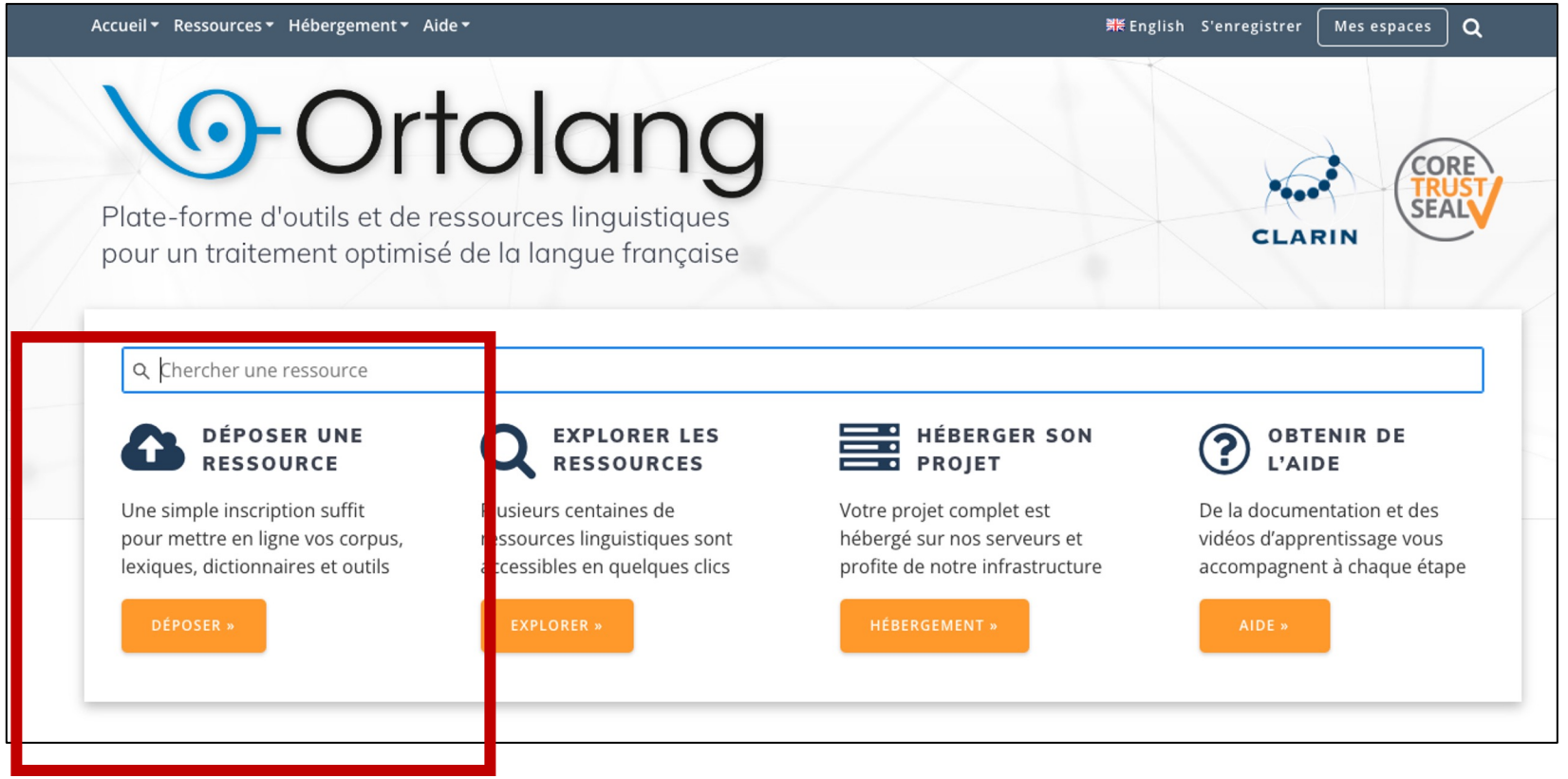
Ortolang

https://www.ortolang.fr/fr/accueil/

CENTRE K CLARIN

CLARIN

Huma-Num IR*




https://www.ortolang.fr/fr/deposer/

# Diffuser son corpus sur une base de données

- https://ct3.ortolang.fr/data/colaje/
- https://ct3.ortolang.fr/data/colaje/madeleine/MADELEINE-04-1_02_14/

Télécharger les transcriptions et les medias - Download the transcriptions and the medias
Fichier/File: MADELEINE-04-1_02_14-480p.mp4 (641 Mo)
Fichier/File: MADELEINE-04-1_02_14.cha (71 ko)
Fichier/File: MADELEINE-04-1_02_14.tei_corpo.xml (263 ko)
Visualisation/Playing: MADELEINE-04-1_02_14.tei_corpo.xml (vue de toute la transcription/full transcription view)
Fichier/File: MADELEINE-04-1_02_14.wav (607 Mo)

# Diffuser son corpus sur une base de données

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<TEI xmlns="http://www.tei-c.org/ns/1.0">
  ▼<teiHeader>
    ▼<fileDesc>
      ▼<titleStmt>
        ▼<publicationStmt>
          <distributor>tei_corpo</distributor>
        </publicationStmt>
        ▼<title>
          <desc>Fichier TEI obtenu à partir du fichier CLAN /applis/data/colaje/madeleine/MADELEINE-04-1_02_14/MADELEINE-04-
          1_02_14.cha</desc>
        </title>
      </titleStmt>
      ▼<notesStmt>
        ▼<note type="COMMENTS_DESC">
          <note type="other">@Birth of CHI: 14-APR-2005</note>
          <note type="scribe">Mélanie Dumez, Françoise Bourdoux (July 2008, revised February 2009, revised June 2009, revised June
          2012), Naomi Yamaguchi (%pho)</note>
        </note>
        ▼<note type="TEMPLATE_DESC">
          ▼<note>
            <note type="type">-</note>
            <note type="parent">-</note>
            <note type="code">CHI</note>
          </note>
          ▼<note>
            <note type="type">-</note>
            <note type="parent">-</note>
            <note type="code">MOT</note>
          </note>
          ▼<note>
            <note type="type">-</note>
            <note type="parent">-</note>
            <note type="code">OBS</note>
          </note>
          ▼<note>
            <note type="code">com</note>
            <note type="type">Symbolic_Association</note>
            <note type="parent">annotationBlock</note>
          </note>
          ▼<note>
            <note type="code">act</note>
            <note type="type">Symbolic_Association</note>
            <note type="parent">annotationBlock</note>
          </note>
          ▼<note>
            <note type="code">mod</note>
            <note type="type">Symbolic_Association</note>
            <note type="parent">annotationBlock</note>
          </note>
          ▼<note>
            <note type="code">ximi</note>
            <note type="type">Symbolic_Association</note>
            <note type="parent">annotationBlock</note>
          </note>
```

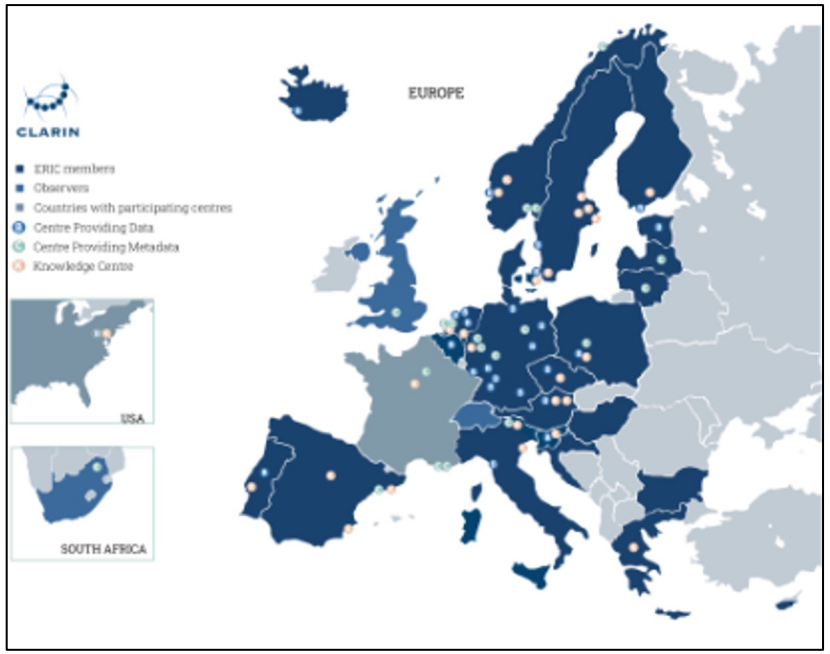

**Format TEI** https://tei-c.org/tools/


| Loc | Ts | Te | Transcription L: 0 - 0 - 16 T: ( - 0:52:42 ) P: - |
|---|---|---|---|
| +div+ | 0:00:00 | 0:52:42 | Situation [+] CHI est assise sur l'escalier puis va dans le salon avec MOT pour ranger des cubes en bois |
| +div+ | 0:00:00 | 0:01:41 | G [+] arrivée de OBS puis jeu de cubes dans le salon |
| MOT | 0:00:00 | 0:00:01 | tu peux ranger les cubes Madeleine ? |
| CHI | 0:00:01 | 0:00:01 | yyy [=! chuchote] . |
| *pho* | | | je |
| *xpnt* | | | show la caméra de l'index |
| MOT | 0:00:01 | 0:00:04 | oui elle a eu ses nouveaux cubes . |
| MOT | 0:00:04 | 0:00:08 | +, que sa gentille marraine lui a donnés . |
| OBS | 0:00:08 | 0:00:10 | +< xxx . |
| MOT | 0:00:10 | 0:00:12 | ils sont beaux hein ? |
| CHI | 0:00:12 | 0:00:12 | 0 . |
| *act* | | | CHI rejoint MOT dans le salon . |
| MOT | 0:00:12 | 0:00:14 | on les met dans la boîte ? |
| CHI | 0:00:14 | 0:00:15 | 0 . |
| *act* | | | CHI met un cube dans la boîte . |
| MOT | 0:00:15 | 0:00:17 | ouais ! |
| CHI | 0:00:17 | 0:00:19 | yyy |

# CLARIN

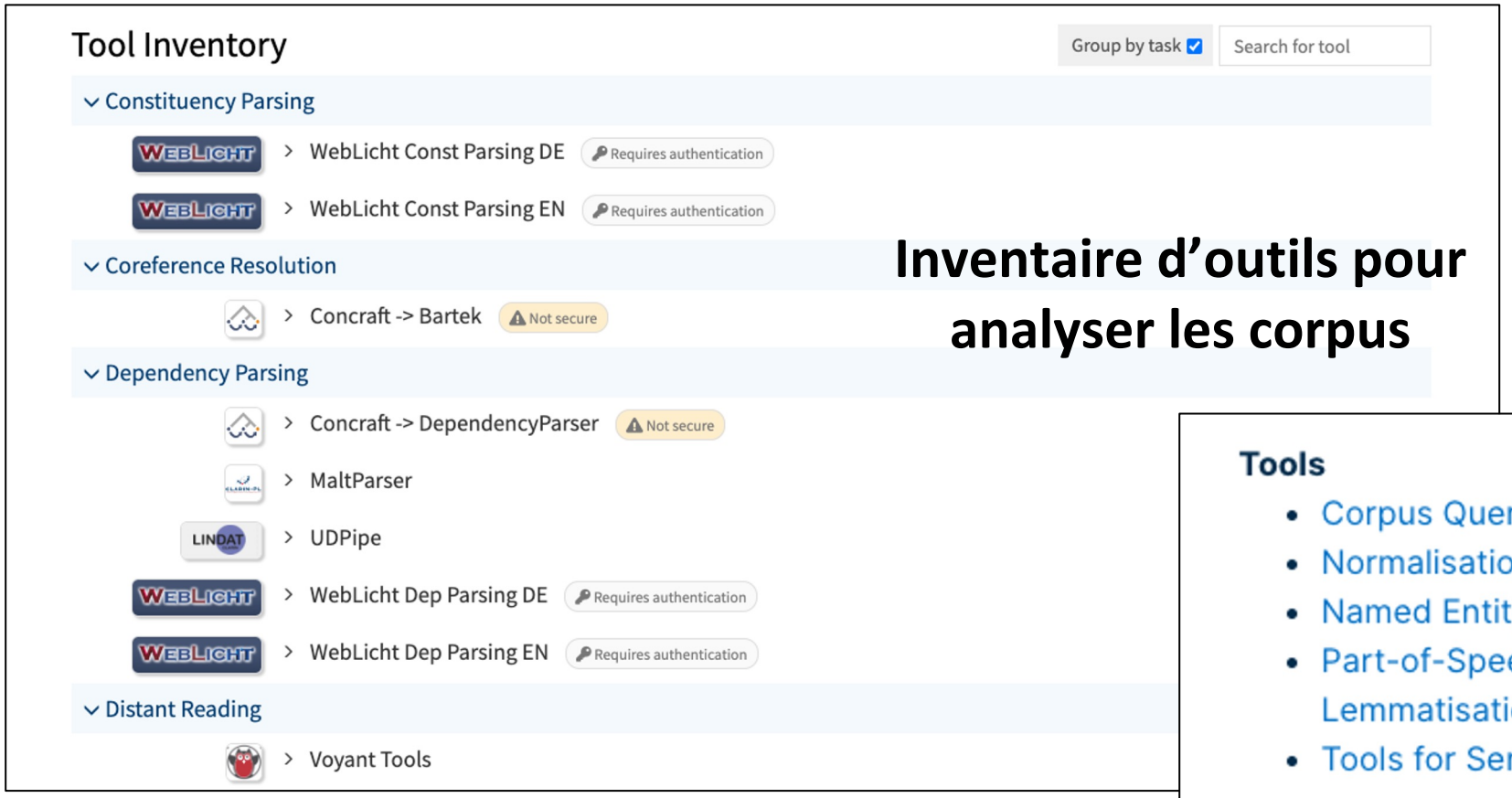
https://clarin-fr.fr/index.html.fr

- Infrastructure fournissant des données, des outils et des services sur les ressources langagières




# CLARIN

https://clarin-fr.fr/index.html.fr




**Inventaire d'outils pour analyser les corpus**

**Tools**

- Corpus Query Tools
- Normalisation
- Named Entity Recognition
- Part-of-Speech Tagging and Lemmatisation
- Tools for Sentiment Analysis

# Bases de données / documentation (1/2)

CHILDES

https://childes.talkbank.org/

TalkBank

https://www.talkbank.org/

clapi
corpus de langues parlées en interaction

http://clapi.icar.cnrs.fr/

COCOON

https://cocoon.huma-num.fr/exist/crdo/

CORLI

https://corli.huma-num.fr/

# Bases de données / documentation (2/2)



https://live.european-language-grid.eu/

8013 corpus

https://www.nakala.fr/

Partage des scripts (Praat, R, etc.)

http://phonetics.linguistics.ucla.edu/facilities/acoustic/praat.html

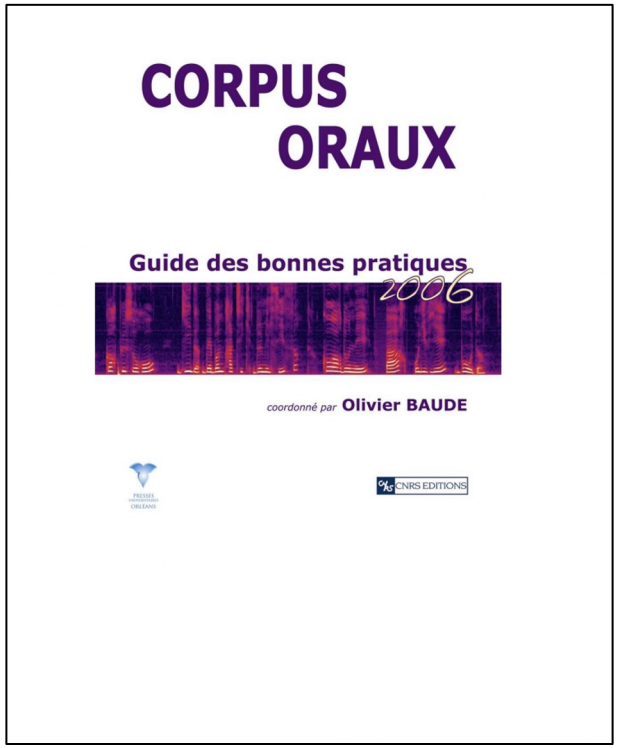
# Data.gouv.fr

- Plateforme ouverte des données publiques françaises : https://www.data.gouv.fr/fr/
- Centralise et structure les données ouvertes en France (open data)
- Données en libre et gratuit

# 6) Ressources sur les corpus

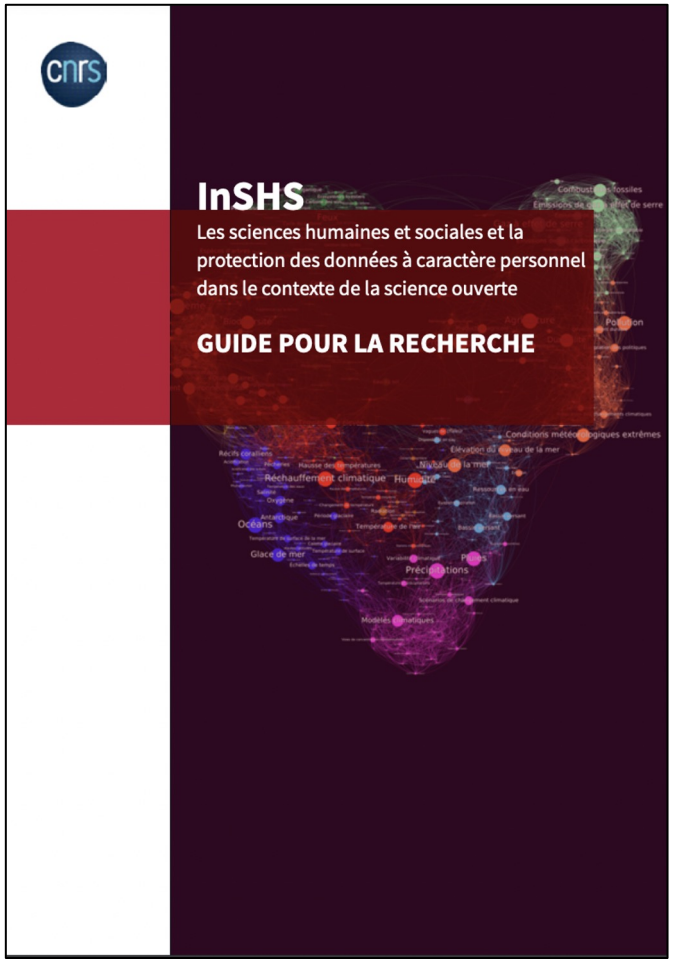
# Guide des bonnes pratiques

- Baude (2006). Corpus oraux : guide des bonnes pratiques. CNRS Editions : https://hal.science/hal-00357706/file/Corpus_Oraux_guide_des_bonnes_pratiques_2006.pdf



# Questions d’éthique : ressources

- RGPD, guide pour la recherche en sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte – Emilie Masson (2019) : https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/guide-rgpd_2.pdf



# Questions d'éthique : ressources

- Diffusion numérique des données SHS : Guide des bonnes pratiques éthiques et juridiques – Véronique Ginouvès et Isabelle Gras (2018) : https://amu.hal.science/hal-01903040/document
- Protection des données personnelles, de la vie privée et de l'image - Nathalie Mallet-Poujol (2014) : https://corli.huma-num.fr/wp-content/uploads/2022/08/ProtectionDonneesPersonnelles_MalletPoujol.pdf
- Droit de la propriété intellectuelle – Agnès Robin : https://corli.huma-num.fr/wp-content/uploads/2022/08/DroitProprieteIntellectuelle_Robin.pdf
- https://corli.huma-num.fr/les-groupes-reseaux/gp4/

# Questions d'éthique : ressources

- COMETS, Comité d'éthique du CNRS : https://comite-ethique.cnrs.fr
- Guide « *Pratiquer une recherche intègre et responsable* » (2017) : https://comite-ethique.cnrs.fr/wp-content/uploads/2019/10/GUIDE-2017-FR.pdf

# Questions d’éthique : ressources

- RGPD (mai 2018) : Règlement général européen sur la protection des données personnelles : https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre1#Article4

# Merci de votre attention !

Et également un grand merci à Delphine, Solaine et Mathias
pour leurs précieux conseils
pendant la préparation de cette présentation !!