# Corpus Inter Langue (CIL)

*Plan de gestion de données créé à l'aide de DMP OPIDoR*

**Créateurs du PGD :** Leonardo Contreras Roa, Thomas Gaillat

**Affiliation du créateur principal :** Université Rennes 2

**Modèle du PGD :** Science Europe - DMP template (english)

**Dernière modification du PGD :** 02/07/2021

**Financeur :** LIDILE

**Résumé du projet :**

The Corpus InterLangue (CIL) project is a collection of spoken and written productions from learners of **English and French as second languages (L2)**. The corpus provides various sources of learner input completing different tasks (Ellis 2003). Learner data have been a source for evidence-based research in Second Language Acquisition for over two decades (Granger, Gilquin, and Meunier 2015). This type of data gives insights into learners' language features which can be analysed in the light of the interlanguage (IL) hypothesis (Selinker 1972).

The CIL data have been collected since 2008 as part of a research programme conducted by the LIDILE research team. The data sources have been stored digitally in non-public spaces. The LIDILE team now wishes to make this data available to the community.

This DMP can be reused freely.

**Chercheur Principal :** Thomas Gaillat

**Identifiant ORCID :** 0000-0003-3433-6533

**Contact pour les Données :** Thomas Gaillat

# Corpus Inter Langue (CIL)

## 1. Data description and collection or re-use of existing data

The corpus CIL corresponds to newly created data. All data elements are created as part of the project. The first corpus elements were collected in 2008. The project is on-going as data are collected on a yearly basis.
The project includes three types of data:

1. learner-related data such as recordings, transcription, writings and authorisations
2. corpus documentation, information notice and technical documents
3. data processing scripts

1. The Corpus CIL is a compilation of audio recordings and written productions of learners (hereinafter 'the Learners') of French and English as a foreign language (L2). Recordings and written productions are supervised by students (hereinafter 'the Researchers') at the *Didactique des Langues* masters program at the University of Rennes 2, as part of their training in corpus linguistics.

Data are annotated with the ELAN application for transcription tasks.

When re-used, the data must be anonymized and their source must be quoted.

Data provenance is documented at the time of recording via an authorization form provided by the Researcher and filled in by the Learner. The form logs the following information:

- Date of the recording
- Information about the learner:
    - Gender
    - Year of birth
    - Country of birth
    - Native language (L1)
    - Regional variety of L1
    - Current country of residence
    - Previous countries of residence
    - Level of education
    - Occupation
    - Number of years studying the L2
    - Self-assessed L2 oral and written proficiency (according to CEFR levels)
    - Other languages spoken (L3)
    - Self-assessed L3 oral and written proficiency (according to CEFR levels)

The documentation includes details on the metadata vocabulary.

2. Corpus documentation, information notice and technical documents are part of the distribution.

The documentation covers how the corpus is collected and the data organised and structured. A consent form and a metadata collection form are part of the project. A metadata collection tool is also part of the project. It consists in a MOODLE database activity.

3. Data processing scripts are available to allow users to upload data to the nakala server as well as applying queries to the corpus. The available scripts are written in R and are distributed under the CC Share-alike licence.

The data is stored in several formats corresponding to open standards (TEI XML, CSV, WAV and UTF-8 txt). The total size of the corpus is c. 25 GB.

The **collected data** corresponds to:

1. An audio recording of a conversation (i.e. a semi-guided interview) between the Researcher and the Learner, saved in wav format.
2. A handwritten text produced by the Learner, scanned and stored in pdf format.
3. An authorization form filled in and signed by the Learner, scanned and stored in pdf format.

Based on these two files, **further data** is produced and stocked with the above-mentioned data:

1. Three orthographic annotation files: one in eaf format [ELAN annotation], one in cha format [CLAN annotation], and one in txt format.
2. One phonetic annotation file in TextGrid format [Praat annotations]
3. One transcription of the handwritten text, in txt format.
4. One metadata file with information about the Learner's demographic and educational profile, in csv format.

## 2. Documentation and data quality

The data is documented in the *Corpus InterLangue* documentation available on the project's website. It includes a description of the collected metadata and file formats.

Date of recording, licence type and data type are stored following the Dublin Core standard.

A csv files includes the learner-specific metadata (not standardized):

- Date of the recording
- Information about the learner:
  - Gender
  - Year of birth
  - Country of birth
  - Native language (L1)
  - Regional variety of L1
  - Current country of residence
  - Previous countries of residence
  - Level of education
  - Occupation
  - Number of years studying the L2
  - Self-assessed L2 oral and written proficiency (according to CEFR levels)
  - Other languages spoken (L3)
  - Self-assessed L3 oral and written proficiency (according to CEFR levels)
  - CEFR levels assessed by linguists

The corpus documentation includes details on the variable encoding conventions.

- Version control strategy versioning on nakala

The data are organised around the concept of learner. Each learner has a unique ID. All the files corresponding to one learner are stored under the same ID. Each data element corresponding to a learner has a DOI.

Data are collected by LIDILE research team's Master's students (First year). The recordings are transcribed with ELAN. The files are verified in their structure and annotation by a lecturer of the research team specialised in linguistics.

The same protocol was followed for L2 French and L2 English learners, adapted to their respective target languages. Learners were asked to complete three tasks:

1. A semi-guided interview (15-25 minutes), aimed at obtaining spontaneous speech samples from learners. The interview is not scripted, but investigators are instructed to ask questions which elicit four different types of speech productions from the learner: 1. A description of themselves, 2. Talking about past events, 3. Talking about future plans, 4. Arguing for or against a given topic.
2. A read aloud task of a page-long text (1-2 minutes), aimed at obtaining a controlled speech sample suitable for comparative phonetic analyses.
3. A writing task elicited by the read aloud task, aimed at obtaining a handwritten production of a 1 or 2 page-long text.

Prior to the beginning of all recording sessions, learners are asked to read and sign a **consent form** and to fill in a **metadata questionnaire**. The consent form and metadata questionnaire were created as part of the project. A new version of the consent form is based on the documents available from https://corli.huma-num.fr/bonnes-pratiques-juridiques/

Master students use a University of Rennes 2 Moodle database to save the metadata referring to a learner.

## 3. Storage and backup during the research process

Each data item has its specific DOI. Data files attached to this data item have their own URL too.

The data and metadata are stored on servers belonging to the Human-Num TGIR French public programme in a Nakala database. The service includes backup on a daily basis.

Data are stored under unique ID numbers to which several files are attached. Files are accessible with persistent identifiers.

Data protection relies on pseudonymization as described in the documentation. Metadata collected by master's students are stored in a Rennes 2 Moodle database. Access is restricted to the students in charge of the tasks. These metadata are then pseudonymized.

The data includes no identification of persons. Only the first two letters of their first name and family names are retained as well as the sex and the year of birth. An example ID is: fra_ca_de_90_f_15. This protects the owner from identification.

Raw data are stored on two hard drives of the research team. Only two researchers (Thomas Gaillat and Leonardo Contreras) have access to these drives. These drives are backed up on a weekly basis.

Pseudonymized data are accessible on the nakala.fr database. Each data item has a persistent URL.

Huma-num data protection policy applies to the data stored on Nakala. Backup solutions are provided as part of the service. In case of data loss on the server, the data are also stored as in files and directories archived on Huma-Num sharedoc service. Access is restricted to the PI and project members.

## 4. Legal and ethical requirements, codes of conduct

Subjects are given information regarding the pseudonymization use of the data and they sign a consent form. They can contact the project leader for access to their data. Public access to their data can be removed upon their request.

The Corpus InterLangue **information notice** is given to subjects prior to the recording session.

The data is held by the LIDILE research team of Rennes 2 university. Access is controlled by head researchers of the team, i.e. the Director and the PI.

The data on Huma-Num Nakala is available publicly via persistent URLs. It is protected under a Creative Commons share-alike non commercial licence.

Only data necessary for studying language acquisition processes are stored with public access on Nakala. Full data (authorisation, personal information) are available on Huma-Num shared doc with restricted access. The identity of Learners will only be accessible to the principal investigators (PI) in disconnected hard drives.

| Role | Type of access | Location |
|---|---|---|
| PI | Full | Hard drives, sharedoc, Nakala |
| Master students | One learner's files | Student's hard drives |
| LIDILE researchers | On request | Nakala and sharedoc |

## 5. Data sharing and long-term preservation

Data are shared publicly through query interfaces (Nakal UI and API) linked to data stored in the Nakala database. These interfaces will only give access to pseudonymized information.

Scripts provide access to data sets mixing metadata and linguistic data.

Data are shared under the Creative commons share alike non-commercial licence.

Data will remain in Nakala servers for as long as the project will be active. This is done because accumulated data will provide insights in learner acquisition processes. All metadata, recordings and transcriptions will be deposited on the Huma-Num Nakala repository.

Data are publicly available to the research community. The data are stored and available online in the long run in order to support longitudinal research in Second Language Acquisition. The data could be reused for different types of analyses focused on different linguistic dimensions.

Following recommendations from the "Association des archivistes français": the data will be kept indefinitevely: "Conservation définitive et intégrale des documents dont l'intérêt historiqueou scientifiquele justifie, dans le service public d'archives territorialement compétent." (Source document)

Nakala APIs are available. R scripts are available on the project website to provide access to the pseudonymized data. Queries on the database will be handled directly.

Nakala provides persistent identification of the data via a unique **handle** managed by the Corporation for National Research Initiatives (CNRI).

## 6. Data management responsibilities and resources

Data management, quality, storage and backup as well as DMP implementation will be overseen by the PI.

The DMP was verified with the help of the "Service d'accompagnement à la gestion des données" from the university of Rennes 2 (date 22/06/2021).

Researchers (doctoral and master's students) at the LIDILE research unit are responsible for regular data collection, storage and curation tasks. This work time is be part of their academic tasks.

1. Master students collect the metadata (2 months)
2. PhD students and researchers verify the data (2 months)
3. PI is in charge of uploading the data on Nakala (2 months)

A postdoctoral researcher recruited in October 2020 by LIDILE was in charge of data workflow structuring and database architecture design.

A computer specialist (Computer science L3 student) was in charge of developing scripts for uploading and querying the corpus.

The cost of data management and technical support is entirely covered by the partnership with Huma-Num and the Maison des Sciences de l'Homme en Bretagne.