"Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité (CELVA.Sp)" project DMP

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "Science Europe - DMP template (english)" fourni par Science Europe.

Plan Details

Plan title	"Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité (CELVA.Sp)" project DMP
Version	First version
Fields of science and technology (from OECD classification)	Languages and literature
Language	eng
Creation date	2020-09-04
Last modification date	2023-05-09
Identifier type	DOI
License	**Name** Creative Commons Attribution 4.0 International **URL** http://spdx.org/licenses/CC-BY-4.0.json
Management plans related to the project	• A language-learning analytics system :

Project Details

Project title	Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité (CELVA.Sp)
Acronym	CELVA.Sp
Abstract	The Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité (CELVA.Sp) project is a collection of spoken and written productions from learners of **a second language (L2)**. The corpus provides transcripts of learner input completing a writing task (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 CELVA.Sp data have been collected since 2008 as part of a research programme conducted by the LIDILE research team. The data sources are stored digitally in non-public spaces. This DMP can be reused freely.
Start date	2019-09-10
End date	2025-06-27

Research outputs :

1. Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité (Collection)

Contributors

Name	Affiliation	Roles
Gaillat Thomas - https://orcid.org/0000-0003-3433-6533	LIDILE	• DMP manager • Personne contact pour les données • Project coordinator

"Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité (CELVA.Sp)" project DMP

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

1a. How will new data be collected or produced and/or how will existing data be re-used?

The Corpus d'Etude des Langues Vivantes Appliquées à un Spécialité (CELVA.Sp) corresponds to newly created data. All data elements are created as part of the project. The first corpus elements were collected in 2018. The project is on-going as data are collected on a yearly basis.

The project includes three types of data:

1. learner-related data including writings, socio-educational metadata and authorisations
2. corpus documentation, information notice and technical documents
3. data processing scripts

1. The Corpus d'Etude des Langues Vivantes Appliquées à un Spécialité (CELVA.Sp) is a collection of written productions of learners (hereinafter 'the Learners') of French, English, Spanish, German and Swedish as foreign languages (L2). Written productions are supervised by language teachers and researchers (hereinafter 'the Researchers') at the Centre de Langues at the University of Rennes 2 and the language department at the University of Rennes 1.

Data are annotated with the CEFR levels.

When re-used, the data must be pseudonymised 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:
    - ID_etudiant:
    - date_soumission_text:
    - Texte_etudiant_
    - nb_annees_L2:
    - L1:
    - Domaine_de_specialite:
    - Acceptation_donnees
    - Sejours_duree_semaines
    - Sejours_duree_mois
    - Sejours_frequence
    - Lang_exposition:
    - Section_renforcee:
    - Annee_naissance:
    - Niveau_etudes_actuel:
    - Age:
    - L2:
    - Note_dialang_ecrit:
    - Keylog_information traces clavier seulement pour la saisie du texte "Texte étudiant".

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 form is also part of the project. It consists in a MOODLE database activity. All learner texts are part of one csv file which supports the creation of several files per learner. Learner Identifiers will be created on the basis of a method that remains to be determined.

3. Data processing scripts will be part of the distribution. These scripts apply linguistic processing techniques to the textual data in order to add linguistic annotation to words. they aslo format the data to prepare dataset of linguistic features related to metatada elements such as the CEFR level of the learners.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

The data are stored in several formats corresponding to open standards (CSV, UTF-8 txt). The total size of the corpus is

circa 1,000 texts.
The **collected data** correspond to:

1. A handwritten text produced by the Learner
2. metadata corresponding to socio-educative information

---

2. Documentation and data quality

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

The data is documented in the *CELVA.Sp* 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):

- Information about the learner:
  - Date of the recording
  - Information about the learner:
    - ID_etudiant:
    - date_soumission_text:
    - Texte_etudiant_
    - nb_annees_L2:
    - L1:
    - Domaine_de_specialite:
    - Acceptation_donnees
    - Sejours_duree_semaines
    - Sejours_duree_mois
    - Sejours_frequence
    - Lang_exposition:
    - Section_renforcee:
    - Annee_naissance:
    - Niveau_etudes_actuel:
    - Age:
    - L2:
    - Note_dialang_ecrit:
    - Keylog information for the writing task
  -

The corpus documentation includes details on the variable encoding conventions.

2b. What data quality control measures will be used?

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.
The same protocol is followed for all L2 learners, adapted to their respective target languages. Learners are asked to complete a writing task aimed at obtaining a digital writing production of a 50 to 300 words.
Prior to the beginning of all recording sessions, learners are asked to read and digitally accept 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/
Prior to being shared on Nakala the data and writings are verified by the Pi and his team.

---

3. Storage and backup during the research process

3a. How will data and metadata be stored and backed up during the research?

Each data item will have its specific DOI. Data files attached to this data item have their own URL too.
The data and metadata will be stored on servers belonging to the Human-Num TGIR French public programme.

- Identifiable data are stored on a Sharedoc server only accessible to the researchers.
- Pseudonymised data will be available from 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.

3b. How will data security and protection of sensitive data be taken care during the research?

Data protection relies on secure access and pseudonymization as described in the documentation. The collected metadata are stored in a Rennes 2 Moodle database. Access is restricted to the researchers with full rights given by the PI. Students have access to their own data.
the data will be available on Huma-Num Sharedoc service located in France. Secure access will guarantee that only the PI and specific researchers have access to the resource.
Pseudonymized data will be accessible on the Huma-Num nakala.fr database. Each data item will have 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

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

Subjects are given information regarding the access, availability and pseudonymization of the data. Subjects tick an online box to 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 CELVA.Sp **information notice** is available to subjects prior to the recording session.
Rennes 2 University's DPO has validated the protocol.
Rennes 2 university has also issued an ethics certificate (Reference number 2023-009)

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

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 will be available publicly via persistent URLs. It will be protected under a Creative Commons share-alike non commercial licence.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

Only data necessary for studying language acquisition processes are stored with public access on Nakala. Full data (authorisation, personal information) are available on Rennes Uiversity MOODLE platform. The identity of Learners will only be accessible to the principal investigator (PI) and specific researchers.

Role	Type of access	Location
PI	Full	University's MOODLE
LIDILE researchers	On request	Nakala and MOODLE
Researchers	On request	Nakala

---

5. Data sharing and long-term preservation

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

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

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

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

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Data will remain in Nakala servers for as long as the project is 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 will be publicly available to the research community. The data will be stored and available online until the completion of the study 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 until the completion of the study: "Conservation définitive et intégrale des documents dont l’intérêt historique ou scientifique le justifie, dans le service public d’archives territorialement compétent." (Source document)

5c. What methods or software tools are needed to access and use data?

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

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

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

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

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).

6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

Researchers at the LIDILE research unit and English teachers at the Universities of Rennes are responsible for regular data collection, storage and curation tasks. This work time is be part of their academic tasks.

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

A computer specialist will be in charge of developing scripts for pseudonymising, 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.