



CORPUCIT Project

Provide tools to facilitate the citation of corpora or extracts from corpora.

Open Access Week

Université Paris Nanterre Wednesday 25 octobre 2023

Christophe Parisse¹, Driss Sadoun^{2 3}

¹ INSERM, MoDyCO (CNRS-Paris Nanterre University)

² ERTIM - INALCO

³ PostLab

Summary

1. Project presentation
2. Demo and test of CORPUCIT platform
3. Discussion

Context and objectives

Consortium HN Corpus, Langues et Interactions (CORLI)

A network of laboratories and researchers working on language corpora.

Objective: To offer tools, documentation and training on the scientific use of language corpora, following the FAIR principles (Findable, Accessible, Interoperable and Reusable).

CORLI is actively working on three projects :

- ▶ Collaborative annotation
- ▶ The French open corpus: corpus data and tools for the French language
- ▶ **Tools for citing corpora or extracts from corpora**

Principles for citing scientific data

- ▶ Since 2014, a joint statement of data citation principles (JDDCP) [Gro14] proposed by several organizations
- ▶ A set of guidelines and recommendations for citing scientific data
 - ▶ Data deposit in a permanent (long-term) archive
 - ▶ Use standard, widespread formats
 - ▶ Use a persistent identifier that leads to a landing page
 - ▶ Use standard citation structures

Long-term archives and citation formats



Figure: HAL citation export formats

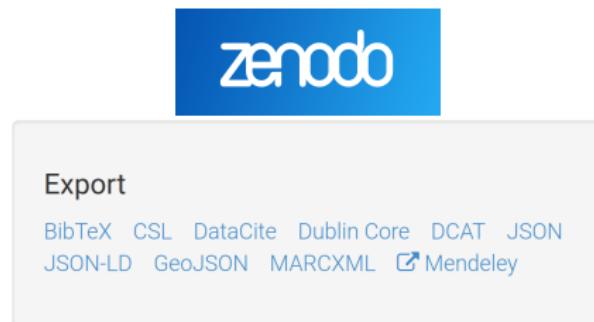


Figure: Zenodo citation export formats

Persistent identifier and landing page

A persistent identifier should link to a landing page rather than directly to the data. [CKG⁺18].

- ▶ Data are not always legally accessible
- ▶ Data can come in a variety of formats and versions
- ▶ Landing page metadata may be complementary
 - ⇒ Metadata must be both human-readable and machine-readable

Structure of a data citation

Recommendations for citing digitized data (corpus).

1. Author(s)
2. Title
3. Date of publication
4. Publisher (or archive where they are hosted)
5. Persistent identifier (PID) : *DOI* or *HANDLE*
6. Document type (optional)
7. Edition, volume or version (optional)
8. Date of last access (optional)

Structure of a citation on COCOON

Author(s) Date Title Version Editor PID

Danto Anatole 2017 Collection ApoliMer (*Anthropologie Politique de la Mer*)
Version 1 Anthropologie politique de la Mer
<https://doi.org/10.34847/COCOON.C7C3F56D-8CF4-3176-BD16-1DA83A8AFCF4>

Danto, Anatole. (2017). *Collection ApoliMer (Anthropologie Politique de la Mer)* (Version 1).
Anthropologie politique de la Mer.
<https://doi.org/10.34847/COCOON.C7C3F56D-8CF4-3176-BD16-1DA83A8AFCF4>

Anne Zribi-Hertz, Elena Soare, Sarra El Ayari 2016
Langues et Grammaires en (Ile-de-)France (LGIDF) Version 1
Structures formelles du langage
<https://doi.org/10.34847/COCOON.7A6A91B9-66ED-3099-BBED-FBB70361C226>

Zribi-Hertz, Anne, Soare, Elena, & El Ayari, Sarra. (2016). *Langues et Grammaires en (Ile-de-)France (LGIDF)* (Version 1). *Structures formelles du langage*.
<https://doi.org/10.34847/COCOON.7A6A91B9-66ED-3099-BBED-FBB70361C226>

Structure of a citation on NAKALA

Author(s) Date Title Type Editor PID

Pierre Fasula 2021 Séminaire Philosophie et psychanalyse 21-05-29 Sound
NAKALA <https://doi.org/10.34847/nkl.be9clr95>

Fasula, Pierre (2021) "Séminaire Philosophie et psychanalyse 21-05-29 - Pierre-Henri Castel"
[Sound] NAKALA. <https://doi.org/10.34847/nkl.be9clr95>

Heba Al Sakhel Amlou 2019 Hache 1 - 1 Image NAKALA
<https://doi.org/10.34847/nkl.ca935q8w>

Al Sakhel Amlou, Heba (2019) "Hache 1 - 1" [Image] NAKALA.
<https://doi.org/10.34847/nkl.ca935q8w>

Structure of a citation on ORTOLANG

Author(s) Title Editor Volume Date Page PID

Aliyah Morgenstern et Christophe Parisse The Paris Corpus
Journal of French Language Studies Volume 22 / Special Issue 01 March 2012 7-12
<https://hdl.handle.net/11403/colaje/v2.4>

*Aliyah Morgenstern and Christophe Parisse : The Paris Corpus, Journal of French Language Studies / Volume 22 / Special Issue 01 / March 2012, pp 7 - 12,
<https://hdl.handle.net/11403/colaje/v2.4>*

CORPUCIT Project

Stakeholders



PRISMES - Langues,
Textes, Arts et Cultures
du Monde Anglophone
- EA 4398



- ▶ Propose recommendations and tools for creating and citing extracts from language corpora.

Challenges

- ▶ Lift barriers to the reuse and emergence of new uses for scientific data
- ▶ Give corpora a clearer status as scientific deliverables
- ▶ Enable researchers to better value the design, collection and sharing of corpora as a fully-fledged scientific activity
- ▶ Promote the dynamics of open science and respect for the FAIR principles (Easy to Find, Accessible, Interoperable and Reusable)

CORPUCIT Objectives

Platform for linking scientific writings and language data extracts (text, sound, video, images), presented in context, facilitating scientific reflection and data reuse.

Usage:

1. Data archive : generate persistent identifiers and citations that can be inserted into scientific literature
2. Database collection : search, view, contextualize and manipulate corpus extracts

Corpus extracts definitions and examples

What is a corpus extract?

Definition of an extract in the scope of CORPUCIT

Section or portion taken from a document (written, audio, image, video, etc.) forming part of a language corpus.

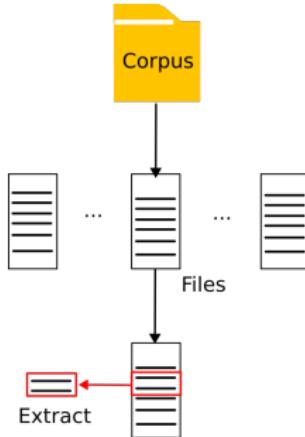


Figure: Illustration of a corpus extract

Extracts from a paper book

N°1 : Boule de Suif, Guy de Maupassant

Maupassant, dans ses contes et nouvelles, exerce son humour caricatural sur les femmes, particulièrement les vieilles-filles et les anglaises, et surtout sur les bourgeois. Il peint de préférence les petits bourgeois de province auxquels il reproche de trop bien vivre et de trop bien manger. Pour ce faire, il utilise la métaphore de l'homme-boule, métaphore utilisée dans *Boule de suif* à la fois pour la peinture des bourgeois que celle de la prostituée, ainsi que dans la peinture des prostituées de *La Maison Tellier*. D'autre part, son humour n'épargne pas les représentants de l'Eglise.

Fiche de lecture : L'humour dans les Contes et nouvelles

Extrait : Boule de Suif, p. 6, p.7 et p.7-8

Loiseau : « De taille exiguë, il présentait un ventre en ballon surmonté d'une face rougeaudé entre deux favoris grisonnants. »

Les deux religieuses : « L'une était vieille avec une face défoncée par la petite vérole comme si elle eût reçu à bout portant une bordée de mitraille en pleine figure. L'autre, très chétive, avait une tête jolie et maladive sur une poitrine de phthisique rongée par cette foi dévorante qui fait les martyrs et les illuminés. »

Boule de Suif : « Petite, ronde de partout, grasse à lard, avec des doigts bouffis, étranglés aux phalanges, pareils à des chapelets de courtes saucisses ; avec une peau luisante et tendue, une gorge énorme qui saillait sous sa robe, elle restait cependant appétissante et courue, tant sa fraîcheur faisait plaisir à voir. Sa figure était une pomme rouge, un bouton de pivoine prêt à fleurir ; et là-dedans s'ouvrailent, en haut, deux yeux noirs magnifiques, ombragés de grands cils épais qui mettaient une ombre dedans ; en bas, une bouche charmante, étroite, humide pour le baiser, meublée de quenottes luisantes et microscopiques. »

Figure: L'humour dans les Contes et nouvelles (Casden)

Extracts from a corpus of audios transcriptions

The screenshot shows the ORTOLANG interface. At the top, there's a navigation bar with icons for user profile, language, and login. Below it, a search bar and a 'Catalogue' dropdown. The main area has a 'DO' logo and a 'DOC-STL' title. A 'Retourner à la fiche' button is visible. The left sidebar shows a file tree: doc-id > CorpusOpinion > CorpusOpinion_FR > CorpusOpinion_FR_Natif > DOC_FR_2020_OPINION_1. The right side displays a table of files with columns: Nom, Type, Dernière modification, and Taille. One file, 'DOC_FR_2020_OPINION_1_Transcription.trs', is highlighted. On the far right, the full XML content of this transcription file is shown.

Nom	Type	Dernière modification	Taille
DOC_FR_2020_OPINION_1_Audio.TextGrid	text/plain	20/05/2021 14:27	56,6 Ko
DOC_FR_2020_OPINION_1_Audio.wav	audio/x-wav	24/02/2022 22:29	63 Mo
DOC_FR_2020_OPINION_1_Audio_non_anon.wav	audio/x-wav	14/01/2022 15:12	63 Mo
DOC_FR_2020_OPINION_1_Autorisation.pdf	application/pdf	20/05/2021 14:27	230,3 Ko
DOC_FR_2020_OPINION_1_CessionDroits.jpg	image/jpeg	20/05/2021 14:27	2,3 Mo
DOC_FR_2020_OPINION_1_Metadonnees.ods	application/vnd.oa...	02/11/2021 18:05	13,3 Ko
DOC_FR_2020_OPINION_1_Transcription.html	text/html	24/02/2022 22:19	28,3 Ko
DOC_FR_2020_OPINION_1_Transcription.trs	application/xml	20/05/2021 14:27	42,5 Ko
DOC_FR_2020_OPINION_1_Transcriptiontrs.textgrid	text/plain	20/05/2021 14:27	56,6 Ko
DOC_FR_2020_OPINION_1_Transcription.txt	text/plain	24/02/2022 22:19	13,6 Ko

DOC_FR_2020_OPINION_1_Transcription.trs

```
<Turn speaker="spk1 spk2" startTime="20.181" endTime="20.887">
<Sync time="20.181"/>
<Mo nb="1"/>
<Mo nb="1"/>
<Mo nb="2"/>
la date
</Turn>
<Turn speaker="spk1" startTime="20.887" endTime="21.896">
<Sync time="20.887"/>
la date de l'apocalypse en fait
</Turn>
<Turn speaker="spk1" startTime="21.896" endTime="23.384">
<Sync time="21.896"/>
ben c'est le vingt-neuf non c'est le vingt-sept je sais plus
</Turn>
<Turn speaker="spk1" startTime="23.384" endTime="23.963">
<Sync time="23.384"/>
ben oui le vingt-sept
</Turn>
<Turn speaker="spk1" startTime="23.963" endTime="24.916">
<Sync time="23.963"/>
ah oui le vingt-sept c'est vrai quand j'avais
</Turn>
<Turn speaker="spk2 spk1" startTime="24.916" endTime="25.353">
<Sync time="24.916"/>
<Mo nb="1"/>
en fait genre
<Mo nb="2"/>

<Event desc="" type="pronounce" extent="Instantaneous"/>
</Turn>
<Turn speaker="spk2" startTime="25.353" endTime="27.451">
<Sync time="25.353"/>
la date de l'apocalypse dans la série
<Sync time="26.635"/>
c'est le vingt-sept juin
</Turn>
<Text speaker="spk1" startTime="27.451" endTime="27.491">
```

Figure: Extracts (Transcriber format) from the corpus DOC-STL on ORTOLANG

Extracts from a corpus of video transcriptions



Anaé - sourire

Métadonnées

Projet : ColaJÉ

Enfant : Anaé
Âge : 0;03;00
Langue : fr
Activité : Activité langagière
Thèmes : Babillage
Mots-clé : Sourire ; Eveil
Durée : 12sec

Description

Dans cette vidéo, la maman d'Anaé essaie de lui faire imiter ces propres expressions du visage, ces sourires. Comme dans la célèbre expérience de Metzloff et Mounou (1983) Anaé imite le fait de tirer la langue.

Transcription

MÈRE : Oui c'est bien!
MÈRE : Tu tires la langue encore ?
MÈRE : Oui !
MÈRE : Tu fais comme maman regardes !



Citer cette vidéo

Si vous utilisez cette vidéo dans le cadre d'une présentation ou d'un article, veuillez utiliser la référence bibliographique appropriée.

Morganet, A. & Parisse, C. (Eds.), (2017). *Le langage de l'enfant. De l'édition à l'explosion*. Paris : Presses de la Sorbonne Nouvelle.
Morganet, A. & Parisse, C., (2012). *The Paris Corpus*. Journal of French Language Studies, Cambridge University Press (CUP), 22 (Special issue 1), pp.7-22.

Figure: Video extract and its transcription from VaLangE

Extracts from a collection of images



Figure: UCD Image Cropper (University College Dublin)

Citation of corpora extracts

Why citing a corpus extract

- ▶ A matter of scientific integrity
- ▶ Credit to those who contributed to the creation of the corpus
- ▶ Give the reader direct access to the extract
- ▶ Locate the extract within the original document or corpus
- ▶ Help with scientific validation
- ▶ Simplify viewing, manipulation and reuse of the extract

Elements related to a corpus extract

- ▶ Mentions of extract in scientific writing: example or illustration
- ▶ Extract citation: bibliographic reference



- ▶ Persistent identifier of the original corpus (PID)
- ▶ Path to the orginal document (file)
- ▶ Persistent identifier of the extract (PIDE)
⇒ URL (landing page) pointed by the PIDE

PID vs URL

PID

- ▶ Uniquely and permanently identifies a digital document
- ▶ Is invariant
- ▶ Points to a URL

URL

- ▶ Corresponds to the address of a digital document
- ▶ Can evolve over time (e.g. changes in corpus structure)
- ▶ Can be edited dynamically
- ▶ Can have parameters

Example of a PID that points to a URL :

<https://doi.org/10.34847/cocoon.ce8a4a03-5083-3144-88ce-7cc4606e8352>

Conclusion

Suggest best practices and develop tools for citing excerpts from corpora.

To promote the longevity and reusability of these tools :

- ▶ Choose standard citation formats and structures
- ▶ Work with open data
- ▶ Use and produce open source code

Roadmap :

- ▶ Project duration 2022-2025
- ▶ Review of citation practices + interviews with researchers
- ▶ A first a proof of concept (Demo)
- ▶ Incremental development of tools and best practices
- ▶ Building these tools for and with the scientific community

Tutorial

Demo CORPUCIT : <https://corpucit.postlab.fr/>

Access credentials

Login : corpucit

Password : demo

Tutorial steps

1. Explore extract examples
2. Create an account (register)
3. Create an extract
 - 3.1 leave the demo and select a corpus of your choice
 - 3.2 select a document from this corpus
 - 3.3 select an extract this document
 - 3.4 come back and log in to corpucit demo
 - 3.5 click on create an extract (Créer une fiche d'extrait)
4. View your extract (Mes extraits)
5. Edit your extracts

Explore

CORPUS

EXTRATS

PROJET

Liste des fiches d'extraits de corpus

Type	Titre	Auteur	Date	Action
IMAGE	TDM_1888_n56_209 "Vue de Schenstatt, Dessin de Lix, d'après une photographie de M. Meyer"	Dress Sadoun	23/06/2023	Voir
ANNOTATION AUDIO	DOC FR 2020 Conseil 1 "Extrait de transcription audio du corpus DOC"	Dress Sadoun	23/06/2023	Voir
TEXTE	A quoi sert de manger ? "Extrait d'un texte simplifié de Laurence Thivard"	Dress Sadoun	23/06/2023	Voir
	Corpus : Fonds A.-M. VURPAS - FRANÇOIS PROVENÇAL (963) "Extrait du corpus d'Orléans, réalisé dans le cadre de l'Enquête SocioLinguistique à Orléans à la fin des années 1960."			

Register

CORPUCIT

PROJET

Connexion

Inscription

First name: Demo

Last name: Corpucit

Email: demo@postlab.fr

Mot de passe:

Votre mot de passe doit comporter au moins 8 caractères, dont une majuscule, une minuscule, un chiffre et un caractère spécial.

[Afficher le mot de passe](#)

Répéter le mot de passe:

En cochant la case ci-dessous, vous acceptez nos [Conditions d'Utilisation](#)

Accepter les conditions

S'inscrire

Login

The screenshot shows a login interface for a platform named "PROJET". At the top, there is a red header bar with the word "CORPUCIT" on the left and a user icon with a dropdown arrow on the right. Below this, a blue navigation bar contains the word "PROJET" and a user icon with a dropdown arrow. A red oval highlights the "Connexion" option in the dropdown menu. The main content area has a light gray background and features a "Connexion" button at the top center. Below it is a form with fields for "Email" (containing "demo@postlab.fr") and "Mot de passe" (containing a masked password). To the right of the password field are two small icons: a blue one with a circular arrow and a blue one with a plus sign. Below the form is a blue "Me connecter" button. Underneath the button, there are two links: "Mot de passe oublié" and "Recevoir un nouveau lien d'activation". At the bottom of the page, there is a message: "Vous n'êtes pas encore inscrit sur Projet? Inscrivez-vous!" followed by a "Inscrivez-vous!" button. The footer of the page includes several small icons: a left arrow, a square, a right arrow, a double left arrow, a double right arrow, a magnifying glass, and a refresh symbol.

CORPUCIT

PROJET

Connexion

Email
demo@postlab.fr

Mot de passe

Me connecter

Mot de passe oublié
Recevoir un nouveau lien d'activation

Vous n'êtes pas encore inscrit sur Projet?
Inscrivez-vous!

Create an extract

1) From the *Home page* or *My extracts page* :

Click on the button "Créer une fiche d'extrait" to create an extract sheet.

The screenshot shows the CORPUCIT interface. At the top, there's a blue header bar with the word 'CORPUCIT' on the left, and 'EXTRAITS' and 'PROJET' with a dropdown arrow on the right. Below the header, the main content area has a title 'Liste des fiches d'extraits de corpus'. A blue button labeled 'Créer une fiche d'extrait' is visible. There are several extract cards listed. One card is highlighted, showing a thumbnail labeled 'IMAGE', the identifier 'TDM_1888_n56_209', the title 'Vue de Schestadt, Dessin de Lix, D'après une photographie de M. Meyer', the author 'Dress Sedoun', a small 'Image' button, the date '22/08/2023', and a 'Voir' button. Below this card, the text 'DOC FR 2020 Conseil 1' is partially visible. Other cards are partially visible below it.

Figure: Home page

The screenshot shows the CORPUCIT interface. The top navigation bar is identical to the one in the previous screenshot. The main content area has a title 'Mes fiches d'extraits de corpus' and a blue 'Créer une fiche d'extrait' button. On the right side, there's a user profile icon with a dropdown menu. The menu items are 'Mon Profil' (highlighted in light blue), 'Mes Extraits' (which is dark blue and has a white cursor arrow pointing to it), and 'Déconnexion'.

Figure: My extract page

2) Fill the form and save it.

Créer une nouvelle fiche d'extraits

[Revenir à la liste des fiches d'extraits](#)

3) On the newly created sheet, click on button "Ajouter un extrait".

The screenshot shows the CORPUCIT application interface. At the top, there is a blue header bar with the word "CORPUCIT" on the left and "EXTRAITS" and "PROJET" on the right. Below the header, the main content area has a title "Extract Demo". Underneath the title are three buttons: "Revenir à la liste des extraits" (yellow), "Éditer" (blue), and "supprimer" (red). A red circle highlights the "Ajouter un extrait texte" button, which is also blue. Below these buttons, a message says "Aucun extrait renseigné." In the bottom left corner of the main area, there is some descriptive text: "Créateur : Demo Corpucit" and "Date de création : 20/10/2023". Underneath this, there is a section titled "Description" containing the word "Demo". At the very bottom of the screen, there is a navigation bar with several icons.

4) Fill the form and save it to create an extract.

Créer un nouvel extrait "texte"

Titre de l'extrait*

Format*

Fichier d'origine*

Chemin du fichier d'origine*

Contenu textuel*

Styles | Normal | Bold | Italic | Underline | Align Left | Align Center | Align Right | Align Justify

5) Edit the extract sheet for additional metadata.

The screenshot shows the CORPUCIT software interface. At the top, there is a navigation bar with three tabs: "CORPUCIT" (highlighted in red), "EXTRAITS" (highlighted in blue), and "PROJET". Below the navigation bar, the main content area has a title "Extract Demo". Underneath the title are four buttons: "Revenir à la liste des extraits" (yellow), "Editer" (blue, circled in red), "supprimer" (red), and "Ajouter un extrait texte" (blue). A message box below these buttons says "Aucun extrait renseigné." In the lower part of the screen, there is a section titled "Créateur : Demo Corpucit" and "Date de création : 20/10/2023". Below this, there is a "Description" section containing the text "Demo".

6) Fill the form to add metadata and save.

Éditer un nouvel extrait

Titre de la fiche*

Description*

Image en avant

[Parcourir...](#)

Type de l'extrait*

Identifiant pérénne*

Identifiant du corpus d'origine

Comment citer

Bibtex

The extract has been created

You can find your new extract by navigating on the *home page* or on your *Mes extraits* (My extracts) page.

Discussion



References

-  Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Fiona Murphy, Patrick Polischuk, Maryann Martone, and Tim Clark, [A data citation roadmap for scientific publishers](#), bioRxiv **5** (2018).
-  Data Citation Synthesis Group, [Joint declaration of data citation principles](#), 2014.