



Projet CITATION : CORPUCIT
Fournir des outils pour faciliter la citation de corpus
ou d'extraits de corpus

Atelier " Citation des données : enjeux, pratiques, outils"
Université Paris Nanterre le vendredi 9 décembre 2022

Christophe Parisse (MoDyCo), **Driss Sadoun** (PostLab)

driss.sadoun@postlab.fr

Contexte et objectifs

Consortium HN Corpus, Langues et Interactions (CORLI)

Un réseau de laboratoires et de chercheurs travaillant sur les corpus de langue.

Objectif : Proposer des outils, de la documentation et des formations autour de l'utilisation scientifique des corpus de langue, en suivant les principes FAIR (Faciles à trouver, Accessibles, Interopérables et Réutilisables).

CORLI travaille activement sur trois projets :

- ▶ L'annotation collaborative
- ▶ Le Corpus Ouvert du Français: données de corpus et outils pour la langue française
- ▶ **La citation de corpus ou d'extraits de corpus**

Principes de citation des données scientifiques

- ▶ Depuis 2014, une déclaration commune des principes de citation des données (JDDCP) [Gro14] proposée plusieurs organisations
- ▶ Un ensemble de lignes directrices et des recommandations pour la citation des données scientifiques
 - ▶ Déposer ses données dans des archives pérennes
 - ▶ Utiliser des formats standards et répandus
 - ▶ Utiliser un identifiant pérenne qui dirige vers une page d'accueil
 - ▶ Utiliser des structures standards pour les citations

Archivages pérennes et Formats de citation

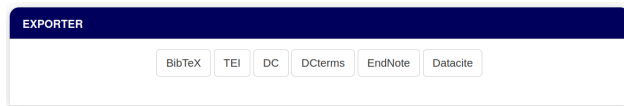


Figure: Formats d'export de citation sous HAL

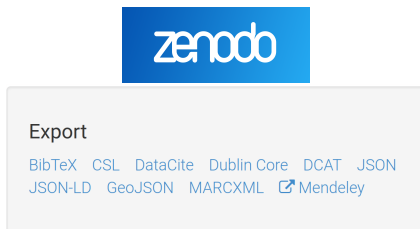


Figure: Formats d'export de citation sous Zenodo

Identifiant pérenne et page d'accueil

Un identifiant pérenne doit renvoyer vers une page d'accueil plutôt que directement vers les données [CKG⁺18].

- ▶ Les données ne sont pas toujours légalement accessibles
- ▶ Les données peuvent se présenter sous de multiples formats ou versions
- ▶ Les métadonnées de la page d'accueil peuvent être complémentaires à celles de la référence
⇒ Les métadonnées doivent être lisible par les humains et les machines

Structure d'une citation de données

Préconisation pour la citation des données (corpus) numérisés.

1. Auteur(s)
2. Titre
3. Date de publication
4. Editeur (ou archive où elles sont hébergées)
5. Identifiant pérenne (IDP) : *DOI* ou *HANDLE*
6. Type de document (optionnel)
7. Édition, volume ou version (optionnel)
8. Date du dernier accès (optionnel)

Structure d'une citation sur COCOON

Auteur(s) Date Titre Version Editeur IDP

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 d'une citation sur NAKALA

Auteur(s) Date Titre Type Editeur IDP

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 d'une citation sur ORTOLANG

Auteur(s) Titre Editeur Volume Date Page IDP

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>

Projet CORPUCIT

Ambition du projet CORPUCIT

- ▶ Proposer des recommandations et des outils pour la création et la citation d'extraits de corpus langagiers.

Enjeux

- ▶ Lever les freins à la réutilisation et à l'émergence de nouveaux usages autour des données scientifiques
- ▶ Offrir aux corpus un statut plus clair de livrable scientifique
- ▶ Permettre aux chercheurs de mieux valoriser la conception, la collecte et le partage de corpus comme une activité scientifique à part entière.
- ▶ Favoriser la dynamique de sciences ouvertes et de respect des principes FAIR (Faciles à trouver, Accessibles, Interopérables et Réutilisables)

Objectif du projet CORPUCIT

Plateforme permettant de lier finement écrits scientifiques et extraits de données de langage (écrits, sons, vidéos, images), présentées dans leur contexte, facilitant la réflexion scientifique et la réutilisation des données.

- ▶ Éditer des portions de corpus pour générer des extraits pouvant être insérer dans des écrits scientifiques.
- ▶ Générer des identifiants pérennes et des citations pouvant être insérer dans des écrits scientifiques.
- ▶ Visualiser, contextualiser et manipuler des extraits de corpus.

Parties prenantes



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



Extraits de corpus : définition et exemples

Qu'est qu'un extrait de corpus

Définition d'un extrait dans le contexte de CORPUCIT

Passage ou portion tiré d'un document (écrit, audio, image, vidéo etc) faisant partie d'un corpus langagier.

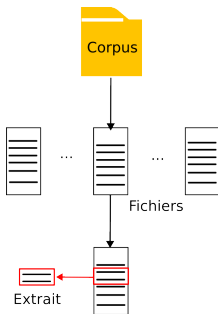


Figure: Illustration d'un extrait de corpus

Extraits issus d'un livre papier

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 exigüe, il présentait un ventre en ballon surmonté d'une face rougeaude 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 malade 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'ouvraient, 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)

Extraits issus d'un corpus journalistique numérisé

The screenshot shows the ORTOLANG interface. At the top, there is a navigation bar with 'ORTOLANG', 'Catalogue', 'Aide', 'Langue', and 'Se connecter'. Below this, the main header features the 'L'EST REPUBLICAIN' logo and the title 'Corpus journalistique issu de l'Est Républicain'. A 'Retourner à la fiche' button is present. The main content area shows a breadcrumb 'est_republicain > Année1999' and a search bar. A table lists several XML files from 1999-05, all with a date of 08/06/2020 16:33. The files are: 1999-05-23.xml (1,9 Mo), 1999-05-25.xml (2 Mo), 1999-05-26.xml (2,3 Mo), 1999-05-27.xml (2,2 Mo), 1999-05-29.xml (2,3 Mo), and 1999-05-30.xml (2,1 Mo).

Document	Type	Date	Taille
1999-05-23.xml	application/xml	08/06/2020 16:33	1,9 Mo
1999-05-25.xml	application/xml	08/06/2020 16:33	2 Mo
1999-05-26.xml	application/xml	08/06/2020 16:33	2,3 Mo
1999-05-27.xml	application/xml	08/06/2020 16:33	2,2 Mo
1999-05-29.xml	application/xml	08/06/2020 16:33	2,3 Mo
1999-05-30.xml	application/xml	08/06/2020 16:33	2,1 Mo

Un travail d'acrobate

Il n'a pas le vertige.

Une équipe spécialisée a été chargée par un propriétaire d'abattre deux énormes marronniers, d'une hauteur appréciable, en plein centre du village.

Ces faux châtaigniers, plus que centenaires, ne pouvaient pas être abattus d'une seule pièce, il a donc fallu les couper depuis le cime, bout par bout, branche par branche.

Un acrobate, le mot n'est pas de trop pour qualifier le travail de spécialiste, est monté jusqu'en haut, armé d'une tronçonneuse et l'a défilé. Un travail bien exécuté qui a nécessité beaucoup de réflexion et de sécurité. Bien entendu, ce travail a été suivi par de nombreux curieux.

```
<div type="article">
<head in travail d'acrobate </head>
<figure>
<head type="legende"> Il n'a pas le vertige. </head>
</figure>
<p> Une équipe spécialisée a été chargée par un propriétaire d'abattre deux énormes
marronniers, d'une hauteur appréciable, en plein centre du village. </p>
<p> Ces faux châtaigniers, plus que centenaires, ne pouvaient pas être abattus d'une
seule pièce, il a donc fallu les couper depuis le cime, bout par bout, branche par
branche. </p>
<p> Un acrobate, le mot n'est pas de trop pour qualifier le travail de spécialiste,
est monté jusqu'en haut, armé d'une tronçonneuse et l'a défilé. Un travail bien
exécuté qui a nécessité beaucoup de réflexion et de sécurité. Bien entendu, ce
travail a été suivi par de nombreux curieux. </p>
</div>
```

Figure: Extraits tirés du corpus de l'Est Républicain de ORTOLANG

Extraits issus d'un corpus de transcriptions audios

ORTOLANG Catalogue Aide Langues se connecter S'inscrire

DOC-STL
Retourner à la fiche

doc-stl > CorpusOpinion > CorpusOpinion_FR > CorpusOpinion_FR_Natif > DOC_FR_2020_OPINION_1

Nom	Type	Dernière modification	Taille
DOC_FR_2020_OPINION_1_AudioTestGrid	text/plain	20/05/2021 14:27	56,6 Ko
DOC_FR_2020_OPINION_1_Audio.wav	audio/wav	24/02/2022 22:29	63 Mo
DOC_FR_2020_OPINION_1_Audio_non_align.wav	audio/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.oas...	02/11/2021 18:05	13,3 Ko
DOC_FR_2020_OPINION_1_transcription.html	text/html	24/02/2022 22:19	28,9 Ko
DOC_FR_2020_OPINION_1_Transcription.trn	application/trn	20/05/2021 14:27	42,5 Ko
DOC_FR_2020_OPINION_1_Transcription.trn.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.trn
<Turn speaker="spk1 spk2" startTme="20.101" endTme="20.867">
<Sync time="20.101"/>
<Who sb="1"/>
& rebours 10
<Who sb="2"/>
la date
</Turn>
<Turn speaker="spk2" startTme="20.867" endTme="21.896">
<Sync time="20.867"/>
la date de l'apocalypse en fait
</Turn>
<Turn speaker="spk1" startTme="21.896" endTme="23.384">
<Sync time="21.896"/>
ben c'est le vingt-neuf non c'est le vingt-sept je sais plus
</Turn>
<Turn speaker="spk2" startTme="23.384" endTme="23.963">
<Sync time="23.384"/>
non c'est le vingt-sept
</Turn>
<Turn speaker="spk1" startTme="23.963" endTme="24.916">
<Sync time="23.963"/>
ah ouï le vingt-sept c'est vrai quand j'avais
</Turn>
<Turn speaker="spk2 spk1" startTme="24.916" endTme="25.353">
<Sync time="24.916"/>
<Who sb="1"/>
en fait genre
<Who sb="2"/>
<Event desc="" type="pronounce" extent="instantaneous"/>
</Turn>
<Turn speaker="spk2" startTme="25.353" endTme="27.451">
<Sync time="25.353"/>
la date de l'apocalypse dans la série
<Sync time="26.835"/>
c'est le vingt-sept juin
</Turn>
<Turn speaker="spk1" startTme="27.451" endTme="27.991">
```

DOC-STL
Retourner à la fiche

DOC_FR_2020_OPINION_1

Rechercher

Nom	Type	Dernière modification	Taille
DOC_FR_2020_OPINION_1_AudioTestGrid	text/plain	20/05/2021 14:27	56,6 Ko

Figure: Extrait au format Transcriber tiré du corpus DOC-STL de ORTOLANG

Extraits issus d'un corpus de transcription vidéos



Anaé - sourire

Métadonnées

Projet : COLAJE

Enfant : Anaé

Âge : 0:03:00

Langue : fr

Activité : Activité langagière

Thèmes : Babillage

Mot-clé : Sourire ; Excit

Date : 12sec

Description

Dans cette vidéo, la maman d'Anaé essaye de lui faire imiter ces propres expressions du visage, ses sourires. Comme dans la célèbre expérience de Meltzoff et Moore (1982) Anaé imite le fait de lever la langue.

Transcription

MÈRE : Ouh c'est bien!

MÈRE : Tu lèves la langue encore ?

MÈRE : Ouh !

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.

Morgenstern, A. & Parisse, C. (Eds.), (2017) *Le langage de l'enfant. De*

Ticksson à l'epulsion, Paris : Presses de la Sorbonne Nouvelle.

Morgenstern, A. & Parisse, C., (2012), *The Paris Corpus*, Journal of

French Language Studies, Cambridge University Press (CLUP), 22

(Special Issue 1), pp.7-12.

Figure: Extrait vidéo et sa transcription héberger sur VaLangE

Extraits issus d'un corpus d'images



Figure: UCD Image Cropper (University College Dublin)

Citation d'extraits de corpus

Pourquoi citer un extrait de corpus

- ▶ Question d'intégrité scientifique
- ▶ Crédit aux personnes ayant contribué à la constitution du corpus
- ▶ Donner au lecteur un accès direct à l'extrait
- ▶ Situer l'extrait dans le document ou le corpus original
- ▶ Aider à la validation scientifique
- ▶ Simplifier la visualisation, la manipulation et la réutilisation de l'extrait

Éléments d'une citation d'extrait de corpus

- ▶ Mention de l'extrait dans l'écrit scientifique : exemple ou illustration
- ▶ Citation de l'extrait : référence bibliographique



- ▶ Identifiant pérenne du document ou corpus d'origine (IDPC)
- ▶ Identifiant pérenne de l'extrait (IDPE)
- ▶ URL sur laquelle pointe l'IDPE

Différences entre un IDP et une URL

IDP

- ▶ Identifie de manière unique et pérenne un document numérique
- ▶ Est invariant
- ▶ Pointe vers une URL

URL

- ▶ Correspond à l'adresse d'un document numérique
- ▶ Peut évoluer dans le temps (ex évolution de la structure du corpus)
- ▶ Peut être éditée de manière dynamique
- ▶ Peut avoir des paramètres

Exemple d'un IDP qui pointe vers une URL :

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

Outils de création d'extraits de corpus

1. Visualiser le document d'origine
2. Sélectionner l'extrait à partir de paramètre(s) délimiteur(s)
3. Visualiser la sélection
4. Enregistrer et exporter la sélection
5. Créer un identifiant pérenne pour l'extrait
6. Créer une citation (référence bibliographique)

Exemples de sélection d'extraits d'images

Sélection via un outil en ligne

Sélection via les paramètres d'un URL

Conclusion

Proposer des bonnes pratiques et développer des outils pour la citation d'extraits de corpus.

Afin de favoriser la longévité et la réutilisation de ces outils :

- ▶ Choisir des formats et structures de citations standards
- ▶ Travailler avec des données ouvertes
- ▶ Utiliser et produire du code ouvert

Feuille de route :

- ▶ Lancement du projet
- ▶ Développement incrémental d'outils et bonnes pratiques
- ▶ Construire ces outils pour et avec la communauté scientifique

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).
-  Martin Fenner, Mercè Crosas, Jeffrey S. Grethe, David Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, Robin Berjon, Sebastian Karcher, Maryann Martone, and Tim Clark, [A data citation roadmap for scholarly data repositories](#), *Scientific Data* **6** (2019), no. 28.
-  Data Citation Synthesis Group, [Joint declaration of data citation principles](#), 2014.