CCdC - Le Corpus Canopé de Corse

Laurent Kevers

CCdC – Le Corpus Canopé de Corse

Laurent Kevers – UMR CNRS 6240 LISA, Université de Corse Pascal Paoli
(kevers_l@univ-corse.fr)

Version de ce document : 1.0

Version concernée du corpus : 1.0

Date de révision : 13/01/2023 (première version le 12/12/2022)

Introduction

Le corpus CCdC est un ensemble de textes en corse, une langue peu dotée principalement parlée en Corse et dans le nord de la Sardaigne. Le corse est une langue régionale de France.

Le corpus CCdC a été développé dans le cadre du projet « Un outil linguistique au service de la Corse et des Corses : la Banque de Données Langue Corse (BDLC) » (financé par le programme CPER), et plus spécifiquement des développements en TAL (Traitement Automatique des Langues) qui y sont menés. Ce projet a entre autres pour objectif de constituer et mettre à disposition des ressources linguistiques, dont ce corpus est un exemple. Plus d'informations sur ce projet sont disponibles dans la bibliographie ci-dessous ou à l'adresse https://bdlc.univ-corse.fr/tal/.

Propriété intellectuelle et licence

Le corpus a été créé par **Laurent Kevers** (ingénieur de recherche, UMR CNRS 6240 LISA), avec l'aide de **Connor MacLean** (étudiant en master 1 à l'Université de Strasbourg, stagiaire durant deux mois à l'Université de Corse), à partir de documents en corse édités par Canopé de Corse[1] et disponibles au format PDF sur le site « EduCorsica[2] ».

Le fonds documentaire a été mis à disposition par Canopé de Corse par l'intermédiaire d'une convention passée avec l'Université de Corse (convention 2021-262). Celle-ci nous permet d'extraire le texte des documents PDF qui présentent une mise en page plus ou moins complexe, afin de les rendre plus directement accessibles, en particulier pour les travaux en TAL.

Le corpus est publié et disponible sur la plateforme Ortolang : https://www.ortolang.fr/. Cette ressource dispose d'un identifiant pérenne qui facilite son identification et sa citation.

Citation

Si vous utilisez cette ressource, merci de citer ce rapport, ainsi que la ressource en elle-même.

1 Voir https://www.reseau-canope.fr/canope-academie-corse/
2 Voir https://www.educorsica.fr/

Laurent Kevers (2022). CCdC - Le Corpus Canopé de Corse. Rapport, UMR 6240 CNRS LISA - Université de Corse, 12/12/2022.

Laurent Kevers, Connor MacLean (2022). Corpus Canopé de Corse (CCdC) [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/corpus-canope-de-corse.

Description du corpus

Le corpus est constitué d'œuvres littéraires (adulte, jeunesse, enfants) ainsi que de documents relatifs au patrimoine et à l'histoire corses.

La majorité du contenu est en corse, mais d'autres langues – le français en particulier – peuvent apparaître ponctuellement.

La taille du corpus, pour cette version initiale, est de 507 215 mots (comptage effectué à l'aide de l'outil *wc* sous Linux) répartis dans 40 documents.

Différentes métadonnées sont disponibles directement dans le fichier XML, et un résumé en est donné en annexe 1.

La répartition selon les différentes catégories est donnée ci-dessous, au tableau 1 ainsi qu'aux figures 2 et 1.

	Nombre de documents	Nombre de mots	Proportion (nb. Docs)	Proportion (nb. mots)
Littérature	8	293173	20,00 %	57,80 %
Patrimoine et Histoire	10	116544	25,00 %	22,98 %
Littérature jeunesse	5	84707	12,50 %	16,70 %
Littérature enfantine	17	12791	42,50 %	2,52 %
TOTAUX	**40**	**507215**	**100,00 %**	**100,00 %**

Tableau 1: Contenu du corpus CCdC par catégories.

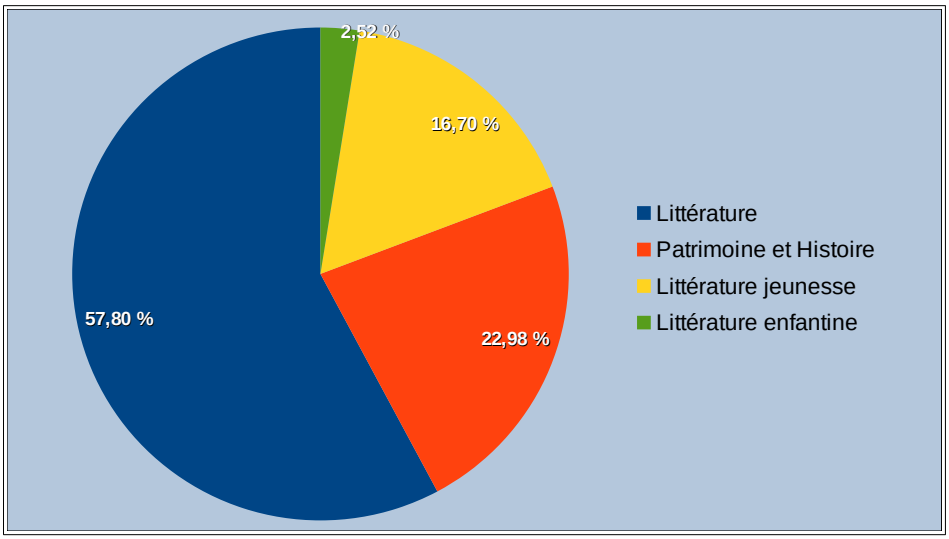

Figure 1: Répartition du contenu du corpus CCdC dans les différentes catégories (nombre de mots).

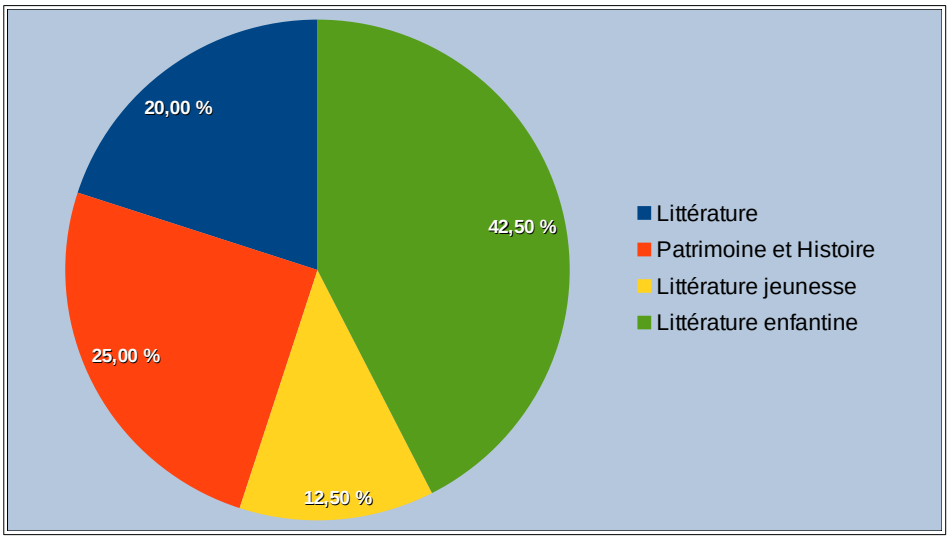

Figure 2: Répartition du contenu du corpus CCdC dans les différentes catégories (nombre de documents).

Le corpus a été transformé au format XML TEI P5[3] à partir des documents PDF. Les sections suivantes expliquent brièvement le processus de traitement, ainsi que le format XML TEI adopté.

Description du processus de traitement

Les versions converties au format TEI XML contiennent normalement le texte intégral, bien que certains éléments aient pu être supprimés (notamment, le cas échéant : préface, introduction, légendes d'illustrations, encadrés, notes de bas de page, etc.)

Les documents ont été segmentés en « chapitres » et en « paragraphes ». Nous avons fait le choix de conserver cette structuration de manière uniforme pour l'ensemble des documents, même si celle-ci convient naturellement mieux à la littérature qu'aux autres types de documents.

Les chapitres ont été découpés manuellement à partir du fichier PDF à l'aide de *pdftk* (*cf.* exemple ci-dessous).

```
pdftk culomba.pdf cat 22-35odd output culomba_CO_Ch02.pdf
```

Pour la majorité des documents, le texte a été obtenu à partir du fichier PDF, en utilisant le convertisseur *pdftotext* (version 0.86.1 ; *cf.* exemple ci-dessous).

```
pdftotext -bbox-layout culomba_FR_Ch01.pdf culomba_FR_Ch01.xml
```

Le fichier XML résultant est structuré en différents niveaux (doc, page, flow, block, line, word), le plus bas étant "word". Pour chaque élément "word" les coordonnées au sein de la page sont renseignées (xMin, yMin, xMax, yMax). Ces informations peuvent être utilisées pour détecter les paragraphes, mais aussi filtrer les éléments tels que les entêtes, les notes de bas de page, les notes disposées dans la marge, etc. Ce filtrage est implémenté dans un script python qui a parfois nécessité des adaptations spécifiques au document traité (en fonction de la mise en page).

3 Voir https://tei-c.org/

Dans certains cas, lorsque la mise en page s'est avérée trop complexe, le fichier PDF a été édité à l'aide de *LibreOffice Draw* afin de supprimer certains contenus. La version modifiée a ensuite été réexportée au format PDF pour suivre le traitement habituel.

Un paragraphe a été défini comme étant un bloc de texte survenant après une indentation horizontale ou un espacement vertical supérieur à l'interligne habituel. Pour des mises en page particulières (encadrés, contenu centré, etc.), la segmentation a été effectuée au cas par cas.

Pour quelques documents, la qualité du texte extrait à l'aide de *pdftotext* était très mauvaise, probablement en raison d'une océrisation peu performante à l'époque de la création du fichier PDF. Pour ces documents, le texte a été obtenu à partir du document PDF, en réexécutant une reconnaissance optique de caractères (OCR), en l'occurrence celle fournie par *Tesseract*[4] avec le profil italien (version 4.1.1-rc2-20-g01fb). Cette procédure particulière a permis d'améliorer la qualité du texte obtenu, mais a également généré quelques problèmes récurrents qu'il a fallu traiter. Les documents concernés par cette procédure sont les suivants :

- *À umbria è à sulia*
- *Filidatu è Filimonda*
- *Ricordi*

Pour l'ensemble des documents, des normalisations ont été effectuées. Les apostrophes et les traits d'union ont été uniformisés, vers les caractères UTF-8 '27' et '2D'. Les points de suspensions (caractère Unicode U+2026) ont été remplacés par trois points (trois caractères UTF-8 '2E' successifs). Les mots contenant une césure ont également été reconstruits.

Avertissement : Malgré tous nos efforts, les multiples vérifications et corrections effectuées, il pourrait subsister des coquilles par rapport à l'œuvre originale. Celles-ci peuvent nous être signalées et des corrections seront intégrées, dans la mesure du possible, lors d'une révision du corpus.

Description de l'encodage XML TEI

Liste des métadonnées

- Métadonnées du document XML TEI `(<publicationStmt>)`
  - Titre du document XML TEI
  - Auteur(s) (de l'œuvre)
  - Traducteur(s) (de l'œuvre)
  - Autres intervenants sur l'œuvre (par exemple *Maestru*, *Aiutu*, etc.)
  - Chercheur qui est principalement responsable de la création du document XML TEI
  - Organisation éditrice du document XML TEI
  - Intervenants sur la compilation et l'encodage XML TEI
  - Date et numéro de version du document XML TEI
  - Taille en nombre de mots (évalué selon une méthode constante via l'outil *wc* sous linux)
  - Organisation responsable de la publication du document XML TEI
  - Licence d'utilisation du document XML TEI

4 Nous avons utilisé ce logiciel au travers du service ShareDocs proposé par HumaNum (https://www.huma-num.fr/).

- Métadonnées du document source `(<sourceDesc>)`
  - Titre (et sous-titre éventuel)
  - Auteur(s)
  - Traducteur(s)
  - Autres intervenants sur l'œuvre (par exemple *Maestru*, *Aiutu*, etc.)
  - Éditeur
  - Date (dépôt légal) de la version exploitée et, en cas de réédition, de la première version
  - ISBN
  - Lien vers la version numérique en ligne
  - Lien(s) vers une ou plusieurs notices d'autorités (BnF, Sudoc)
  - Licence de la version numérique en ligne (fichier PDF source)

- Métadonnées relatives aux aspects méthodologiques de la création du corpus `(<encodingDesc>)`
  - Objectif du projet et méthodes de traitement
  - Principaux logiciels utilisés pour la conversion en texte
  - Choix de traitement (normalisation, définition des unités du texte…)

- Métadonnées liées à la caractérisation du contenu `(<classDecl>)`
  - Définition de la nomenclature « Niveau de lecture » (catégorie d'âge) : `École – cycle 1 (4–6 ans), École – cycle 2 (7–9 ans), École – cycle 3 (10–11 ans), Collège (12–15 ans), Lycée (16–18 ans), Université, Tout public.`
  - Définition de la nomenclature des catégories de la collection Canopé : `Littérature enfantine, Littérature jeunesse, Littérature, Poésie, Patrimoine et Histoire, Manuel scolaire.`
  - Niveau(x) de lecture applicable(s) au document.
  - Niveau(x) de lecture normalisé(s) (échelle globale du Cadre Européen Commun de Référence[5]) applicable(s) au document
  - Catégorie(s) « Canopé » applicable(s) au document
  - Langue(s) utilisée(s) dans le document[6]

- Métadonnées liées à l'historique de révision du document XML TEI `(<revisionDesc>)`
  - Liste des versions successives (date, commentaire incluant idéalement un numéro de version et les changements intervenus depuis la version précédente)

Structure XML TEI

Chaque document du corpus a été enregistré dans un fichier XML séparé. Un exemple général d'encodage qui reprend les métadonnées décrites ci-dessus est disponible en annexe 2.

5 Voir https://www.coe.int/fr/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

6 Actuellement seule la mention de la langue majoritairement présente est reprise. Une liste plus complète des langues détectées, renseignant leur importance respective, pourra être fournie dans le futur.

Évolutions futures

- Le corpus devrait prochainement faire l'objet d'une publication dans une interface de consultation (concordancier).
- Ajout de documents supplémentaires.
- Ajout d'annotations (identification de langue, parties du discours).
- Ajout des traductions françaises lorsqu'elles sont disponibles.

Remerciements

Nous tenons à remercier Marie-Luce Massa (Canopé) de s'être intéressée à notre démarche et d'avoir été un contact privilégié auprès de Canopé, ainsi que Marie-Françoise Saliceti (Université de Corse) de nous avoir accompagné dans l'élaboration de la convention.

Merci à Alice Millour (Université Paris 8), d'avoir contribué à l'encadrement du stage de Connor MacLean à l'Université de Corse.

Bibliographie

Kevers L., Millour A. (2022). Réalisations, obstacles et perspectives pour l'outillage du corse. *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, Nov 2022, Marseille, France. pp.154-161. ⟨hal-03846829⟩

Kevers L., Retali-Medori S. (2020). Towards a Corsican Basic Language Resource Kit. In : *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, p.2726-2735. 11–16 May 2020, Marseille, France. ⟨https://www.aclweb.org/anthology/2020.lrec-1.332/⟩

Kevers L., Retali-Medori S., Tognotti A.G. (2021). A Survey of Language Technologies Resources and Tools for Corsican. Research Report. UMR 6240 CNRS LISA - Université de Corse. ⟨hal-03228733⟩

Annexe 1 : Résumé du contenu du corpus et principales métadonnées associées

Titre	Auteur	Traducteur	Date	ISBN	Nb. mots	Nb. chars	Collection	Niveau	Niveau Normalisé
Abdelcader Camal	Jacqueline Favreau	Michele Frassati	12/2008	2 86 620 219 4	943	5022	Littérature enfantine	Ecole - Cycle1,Ecole - Cycle2,Ecole – Cycle3	inconnu
Animali in puesia	Maria Antonia Salini	n/a	09/2007	2 86 620 203 3	537	2701	Littérature enfantine	Ecole - Cycle1,Ecole - Cycle2,Ecole – Cycle3	inconnu
A Balagna	Antoine Marchini	Ghjuvan Battistu Paoli	04/2010	978 2 86 620 243 9	10118	56508	Patrimoine et Histoire	Tout public	inconnu
A fola di topa pinnutella	Sonia Moretti	n/a	09/2017	978 2 240 04427 3	728	4079	Littérature enfantine	École – cycle 1, École – cycle 2	inconnu
A magacciula sdenticata	Rosa Maria Ottavi	n/a	11/2010	978 286 620 254 5	635	3257	Littérature enfantine	École – cycle 1, École – cycle 2, École – cycle 3	inconnu
A magia di natale	Maria Antonia Salini	n/a	11/2017	978 2 240 04429 7	1050	5295	Littérature enfantine	École – cycle 1, École – cycle 2	inconnu
A Signora di a furesta	Jean André Alesandri	Ghjuvan Battistu Paoli	05/2000	2 86 620 135 3	613	3305	Littérature enfantine	Ecole – Cycle3	inconnu
À umbria è à sulia (U cimiteriu di l'elefanti). Cronaca di u vita corsa d'en è d'oghje	Michele Poli	n/a	11/2002	2 86620 157 4	84817	475249	Littérature	Collège, Lycée, Université	inconnu
Bastia	Antoine Marchini	Ghjuvan Battistu Paoli	02/2012	978 2 86 620 284 2	13167	74285	Patrimoine et Histoire	Tout public	inconnu
Capi corsu	Antoine Marchini	Ghjuvan Battistu Paoli	09/2010	978 2 86 620 252 1	10721	59275	Patrimoine et Histoire	Tout public	inconnu
Carulina, l'ultima tribbiera	Ghjuvan Micheli Weber	n/a	12/2009	978 286 620 241 5	875	4677	Littérature enfantine	Ecole - Cycle1,Ecole - Cycle2,Ecole – Cycle3	inconnu
Carulina, u paese di i strumenti	Ghjuvan Micheli Weber	n/a	12/2001	2 86 620 153 1	643	3378	Littérature enfantine	Ecole - Cycle1,Ecole - Cycle2,Ecole – Cycle3	inconnu
Castagniccia	Antoine Marchini	Ghjuvan Battistu Paoli, Ghjuvan Micheli Weber	06/2012	978 2 86 620 293 4	23424	132937	Patrimoine et Histoire	Tout public	inconnu
Cavallaria paisana	Natale Rochiccioli	n/a	12/1999	2 86 620 133 7	31424	172720	Littérature	Collège, Lycée, Université	inconnu
Culomba	Prosper Mérimé	Ghjuvan Battistu Paoli	12/2016	978 2 240 03948 4	48559	260308	Littérature	Collège, Lycée, Université	inconnu
E pieve di marina	Antoine Marchini	Ghjuvan Micheli Weber	03/2011	978 2 86 620 269 9	10080	57136	Patrimoine et Histoire	Tout public	inconnu
Filidatu è Filimonda (o A filastrocca di Maccu-Mahò)	Sebastianu Dalzeto	n/a	02/1995	2 86 620 085 5	38842	224969	Littérature	Collège, Lycée, Université	inconnu
Fiumorbu	Gilles Guerrini, Antoine Marchini	Ghjuvan Micheli Weber	04/2011	978 2 86 620 270 5	10004	56873	Patrimoine et Histoire	Tout public	inconnu
Golu	Antoine Marchini	Ghjuvan Battistu Paoli	12/2010	978 2 86 620 263 7	12610	68746	Patrimoine et Histoire	Tout public	inconnu
I setti mulini	Jean Alesandri	Ghjuvan Micheli Weber	05/2010	9 78 286 620 249 1	9777	54340	Littérature jeunesse	Collège, Lycée	A2-B1

Titre	Auteur	Traducteur	Date	ISBN	Nb. mots	Nb. chars	Collection	Niveau	Niveau Normalisé
Nantu à i passi di Pasquale Paoli	Antoine Marchini	Ghjuvan Battistu Paoli	10/2009	978 286 620 227 9	6872	37456	Patrimoine et Histoire	Collège, Lycée, Université	inconnu
Nebbiu	Antoine Marchini	Ghjuvan Battistu Paoli	11/2011	978 2 86 620 273 6	9091	50342	Patrimoine et Histoire	Tout public	inconnu
Oghje hè Carnavale!	Maria Antonia Salini	n/a	10/2019	978 2 240 04948 3	786	3813	Littérature enfantine	École – cycle 1, École – cycle 2	inconnu
Prosa sculare	Auteurs divers	n/a	07/1999	2 86 620 125 6	17771	94014	Littérature	Collège, Lycée, Université	inconnu
Raconti	Ghjuvan Micheli Weber	n/a	02/2011	9 78 286 620 266 8	10704	57379	Littérature jeunesse	Ecole – Cycle3, Collège	A2
Ricciulellu è u gricciu	Jean Alesandri	n/a	12/2017	978-2-240-04426-6	642	3424	Littérature enfantine	École – cycle 1, École – cycle 2	inconnu
Ricordi	Ignaziu Colombani	n/a	10/1996	2 86 620 097 7	34719	189620	Littérature	Collège, Lycée, Université	inconnu
Sant'Andria	Scola di Vighjaneddu	Rinatu Coti	02/2010	9 78 286 620 244 6	3560	18993	Littérature jeunesse	Collège, Lycée	A2
Scopra l'Udissea	Homère	Ghjuvan Micheli Weber, Ghjuvan Battistu Paoli, Maria Dumenica Predali	2018	9 782240 047175	15135	80012	Littérature	Collège, Lycée	inconnu
Spichjittinu u topu mascaratu	Marilena Menozzi, Rosamaria Ottavi, Pasqualina Pergola, Mariateresa Tomasi	n/a	11/2018	978-2-240-04941-4	382	1949	Littérature enfantine	École – cycle 1, École – cycle 2	inconnu
Tavignanu	Antoine Marchini	Ghjuvan Battistu Paoli	12/2009	978 2 86 620 240 8	10457	57645	Patrimoine et Histoire	Tout public	inconnu
Ulisse è u ciclopu	Marilena Menozzi, Rosamaria Ottavi, Pasqualina Pergola, Mariateresa Tomasi	n/a	11/2018	978-2-240-04942-1	595	3030	Littérature enfantine	École – cycle 1, École – cycle 2	inconnu
U cinquantottesimu	Scola di Santa Riparata di Balagna	n/a	05/2009	9 78 286 620 226 2	12080	63717	Littérature jeunesse	Collège, Lycée	A2-B1
U lionu è a lefra	Marilena Menozzi, Rosamaria Ottavi, Pasqualina Pergola, Marie-Thé Tomasi	n/a	11/2018	978-2-240-04940-7	244	1239	Littérature enfantine	École – cycle 1, École – cycle 2	inconnu
U pane azimu	Ghjaseppu Maria Bonavita	n/a	02/2001	2 86620 144 2	21906	117776	Littérature	Collège, Lycée, Université	inconnu
U sicretu di l'Acciaccila	Scola bislingua di Sandreshi	n/a	11/2018	978-2-240-04943-8	544	3123	Littérature enfantine	École – cycle 1, École – cycle 2	inconnu
U sumere è u porcu	Rosa Maria Ottavi	n/a	12/2005	2 86 620 185 X	378	1939	Littérature enfantine	École – cycle 2	inconnu
U Sumere Marchese	Hélène Suzzoni	Marianghjula Antonetti Orsoni	12/2017	978-2-240-04425-9	2578	13688	Littérature enfantine	Ecole - Cycle2,Ecole – Cycle3, Collège	inconnu
U Viaghjone di Vannina	Gabriel-Xavier Culioli	Ghjuvan Battistu Paoli	12/2017	978 2 240 04432 7	48586	262871	Littérature jeunesse	École – cycle 3	inconnu
Zhù Xiù và à a scola	Marilena Menozzi, Rosamaria Ottavi, Pasqualina Pergola, Mariateresa Tomasi	n/a	11/2018	978 2 240 04716 8	618	3085	Littérature enfantine	École – cycle 1, École – cycle 2	inconnu

Annexe 2 : Exemple de structure XML TEI

La structure XML ci-dessous constitue une généralisation des structures produites pour chaque document XML. En fonction de l'oeuvre concernée, il est possible que la structure soit légèrement différente ou que certaines métadonnées n'apparaissent pas.

La coloration indique le niveau d'imbrication des principales balises qui permettent d'appréhender la structure XML : **Orange** > **Vert** > **Bleu** > **Rouge**.

Les deux principales sections **teiHeader** et **text** sont mis en évidence par des caractères gras noir. Les segments colorés en **gris** correspondent aux (méta)données en tant que telles.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" type="application/xml" schematypens="http://purl.oclc.org/dsdl/schematron"?>

<TEI xmlns="http://www.tei-c.org/ns/1.0">

  <teiHeader>

    <fileDesc>

      <titleStmt>
        <title>Titre du document XML TEI</title>
        <author>
             <orgName>Organisation considérée comme auteur</orgName>
             <persName>Nom de l'auteur</persName>
        </author>
        <editor role="translator" xml:lang="cos">
            <persName>Nom du traducteur (si pertinent)</persName>
        </editor>
        <editor role="Maestru">
            <persName>Nom du « maître »</persName>
        </editor>
        <editor role="Aiutu">
            <persName>Nom de l'« aide »</persName>
        </editor>
        <principal>
            <persName>Laurent Kevers</persName>
            <ptr target="https://orcid.org/0000-0001-5058-6706"/>
        </principal>
        <editor>
```

```
      <orgName>Università di Corsica Pasquale Paoli</orgName>
      <orgName>UMR CNRS 6240 LISA</orgName>
  </editor>
  <respStmt>
      <resp>Compiled and TEI encoded by </resp>
      <persName>Laurent Kevers</persName>
  </respStmt>
  <respStmt>
      <resp>Compiled and TEI encoded by (intern)</resp>
      <persName>Connor Maclean</persName>
  </respStmt>
</titleStmt>

<editionStmt>
    <edition n="x.y">
        <date>AAAA-MM-JJ</date>
    </edition>
</editionStmt>

<extent>
    <measure unit="words">XXX (estimated by linux word count utility 'wc')</measure>
</extent>

<publicationStmt>
    <publisher>Università di Corsica Pasquale Paoli</publisher>
    <availability>
        <licence>
            This XML TEI corpus is available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
            <ptr target="https://creativecommons.org/licenses/by-nc-sa/4.0/" />
        </licence>
    </availability>
</publicationStmt>

<sourceDesc>
  <p>
    This work is derived from 'Titre de l'oeuvre' (Auteur de l'oeuvre) edited by Canopé de Corse (<ptr target="http://www.cndp.fr/crdp-corse/"/>)
    <bibl>
      <title>Titre de l'oeuvre</title>
      <title type="sub">Sous-titre de l'oeuvre</title>
      <author>
        <persName>Auteur de l'oeuvre</persName>
      </author>
      <editor role="translator" xml:lang="cos">
        <persName>Traducteur de l'oeuvre</persName>
      </editor>
```

```
        <editor role="Maestru">
          <persName>Nom du « maître »</persName>
        </editor>
        <editor role="Aiutu">
          <persName>Nom de l'« aide »</persName>
        </editor>
        <date>MM/YYYY</date>
        <idno type="ISBN">Numéro ISBN</idno>
        <publisher>CANOPÉ - CANOPÉ de CORSE</publisher>
        <ptr target="Adresse Internet où peut être téléchargé le fichier PDF original"/>
        <ptr target="Notice d'autorité BnF"/>
        <ptr target="Notice d'autorité Sudoc"/>
        <availability>
          <licence>
            CC BY-NC-ND 2.0 FR : Les documents peuvent être téléchargés et sont placés sous licence creative commons pour un usage non commercial.
            Canopé has granted the University of Corsica an authorisation to use, transform and redistribute this work (convention 2021-262).
          </licence>
        </availability>
      </bibl>
    </p>
  </sourceDesc>

</fileDesc>

<encodingDesc>
  <classDecl>
    <taxonomy xml:id="reading_level">
      <category xml:id="rl-1">
        <catDesc xml:lang="eng">Primary School - 1st cycle (4-6 years)</catDesc>
        <catDesc xml:lang="fra">École - cycle 1 (4-6 ans)</catDesc>
      </category>
      <category xml:id="rl-2">
        <catDesc xml:lang="eng">Primary School - 2nd cycle (7-9 years)</catDesc>
        <catDesc xml:lang="fra">École - cycle 2 (7-9 ans)</catDesc>
      </category>
      <category xml:id="rl-3">
        <catDesc xml:lang="eng">Primary School - 3rd cycle (10-11 years)</catDesc>
        <catDesc xml:lang="fra">École - cycle 3 (10-11 ans)</catDesc>
      </category>
      <category xml:id="rl-4">
        <catDesc xml:lang="eng">Secondary School (12-15 years)</catDesc>
        <catDesc xml:lang="fra">Collège (12-15 ans)</catDesc>
      </category>
      <category xml:id="rl-5">
        <catDesc xml:lang="eng">Further Education (16-18 years)</catDesc>
```

```
        <catDesc xml:lang="fra">Lycée (16-18 ans)</catDesc>
      </category>
      <category xml:id="rl-6">
        <catDesc xml:lang="eng">Higher Education</catDesc>
        <catDesc xml:lang="fra">Université</catDesc>
      </category>
      <category xml:id="rl-0">
        <catDesc xml:lang="eng">All audiences</catDesc>
        <catDesc xml:lang="fra">Tout public</catDesc>
      </category>
    </taxonomy>
    <taxonomy xml:id="canope_collection">
      <category xml:id="col-1">
        <catDesc xml:lang="eng">Children's Literature</catDesc>
        <catDesc xml:lang="fra">Littérature enfantine</catDesc>
      </category>
      <category xml:id="col-2">
        <catDesc xml:lang="eng">Young Adult Literature</catDesc>
        <catDesc xml:lang="fra">Littérature jeunesse</catDesc>
      </category>
      <category xml:id="col-3">
        <catDesc xml:lang="eng">Literature</catDesc>
        <catDesc xml:lang="fra">Littérature</catDesc>
      </category>
      <category xml:id="col-4">
        <catDesc xml:lang="eng">Poetry</catDesc>
        <catDesc xml:lang="fra">Poésie</catDesc>
      </category>
      <category xml:id="col-5">
        <catDesc xml:lang="eng">Heritage</catDesc>
        <catDesc xml:lang="fra">Patrimoine</catDesc>
      </category>
      <category xml:id="col-6">
        <catDesc xml:lang="eng">History</catDesc>
        <catDesc xml:lang="fra">Histoire</catDesc>
      </category>
      <category xml:id="col-7">
        <catDesc xml:lang="eng">Textbook</catDesc>
        <catDesc xml:lang="fra">Manuel scolaire</catDesc>
      </category>
    </taxonomy>
  </classDecl>
</encodingDesc>

<profileDesc>
```

```
      <langUsage>
        <language ident="cos" usage="100">Corsican</language>
      </langUsage>

      <textClass>
        <keywords scheme="#reading_level">
            <list>
              <item>rl-4</item>
              <item>rl-5</item>
              <item>rl-6</item>
            </list>
        </keywords>
        <keywords scheme="https://www.coe.int/fr/web/common-european-framework-reference-languages">
          <list>
            <item>A2</item>
            <item>B1</item>
          </list>
        </keywords>
        <keywords scheme="#canope_collection">
            <list>
              <item>col-3</item>
            </list>
        </keywords>
      </textClass>
    </profileDesc>

    <revisionDesc>
       <change when="2022-07-06">Original version of this XML TEI file, compiled and TEI encoded by <persName>Laurent Kevers</persName>, <persName>Connor Maclean</persName>,
<orgName>Università di Corsica Pasquale Paoli</orgName></change>
    </revisionDesc>

  </teiHeader>

  <text>
    <body>
      <div type="chapter" n="1">
        <div type="paragraph" n="1.1">
          <p>Du texte...</p>
        </div>
      </div>
    </body>
  </text>

</TEI>
```