Guide et format de transcription E:Calm – version 1.1

Rédaction: Claude Ponton
Proposition issue du travail collectif: Myriam El Helou, Xiaoshu Xie, Serge Fleury et Claude Ponton

Rappel : cette version du guide de transcription ne s'intéresse qu'à **la version 1 de la production**. Les versions suivantes (commentaires de l'enseignant, corrections par l'élève…) seront traitées dans un deuxième temps. La définition et le codage des métadonnées sont à l'étude et ne sont pas intégrées à cette version.

Sommaire

1. Notre approche de la transcription

L'étape de transcription consiste, à partir d'une copie manuscrite ou d'un scan de cette copie, à proposer une version numérique du texte de la production. Il s'agit donc pour le transcripteur de reproduire numériquement le texte de l'auteur. C'est sur ces transcriptions que s'effectueront l'ensemble des traitements d'analyse ultérieurs. Il est donc crucial de maîtriser au mieux ce processus qui soulève deux problèmes principaux : « que transcrire ? » et « comment transcrire ? ».

a. Que transcrire ?

Sachant qu'une copie contient bien autre chose que du simple texte, la première difficulté consiste à sélectionner les éléments à transcrire (le texte, sa mise en page, sa mise en forme, ses évolutions...). En fonction des objectifs du projet E:Calm (dépôt du corpus sur Ortolang, mise en ligne scans/transcriptions, analyses linguistiques), nous avons opté pour une approche pseudo-diplomatique de la transcription. Contrairement à l'approche diplomatique qui vise à produire une « photographie » du document « *en rapportant, avec les outils qui le permettent, malgré leurs limites, tous les événements du manuscrit* »[1], nous adoptons une approche pseudo-diplomatique qui « *reproduit autant que possible la graphie et la mise en page ligne par ligne, fournissant ainsi une aide à la lecture* »[2]. Il sera ainsi possible d'afficher un alignement ligne à ligne entre le scan et la transcription de chaque production.

La transcription est donc une approximation de la production originale. L'humain transcripteur doit déchiffrer le texte puis le décrire dans un format donné. Cette opération est donc source d'erreur. De plus, certains passages peuvent poser des problèmes de déchiffrement allant d'un doute du transcripteur jusqu'à l'impossibilité de proposer une transcription (texte illisible). Afin de proposer des transcriptions d'une qualité relativement contrôlée, nous proposons la méthodologie suivante :

- Proposer un guide clair et unique de transcription testé et validé par l'ensemble des membres du projet.
- Former et faire travailler les transcripteurs par deux sur même type de corpus. Par expérience, l'échange entre les transcripteurs permet des transcriptions homogènes et de meilleure qualité, de débloquer des doutes et de maintenir la motivation. L'idéal serait d'avoir une double transcription de chaque copie et de développer un outillage de comparaison...
- Par expérience également, il est nécessaire que les transcriptions soient validées par un expert.

---

1 http://www.item.ens.fr/articles-en-ligne/structuration-des-manuscrits-du-corpus-a-la-region/#ftn1

2 Hélène de JACQUELOT, *Les Manuscrits de Stendhal et l'édition des « Journaux et Papiers » en ligne et sur papier*, La Francesistica italiana à l'ère du numérique, Publifarum, n. 25, pubblicato il 25/04/2016, consultato il 09/02/2018, url: http://www.farum.it/publifarum/ezine_articles.php?id=332

- L’outil choisi pour les transcriptions devrait être la plateforme PHuN (http://www.espace-transcription.org/). Cette plateforme de transcription permet d’accueillir plusieurs projets et propose une centralisation des transcriptions réalisées. Son interface (boutons, menus) est spécifique à chaque projet puisqu’elle est calculée en fonction de la DTD fournie.

b. Comment transcrire ?

L'un des objectifs du projet E:Calm est de « *structurer et mettre à disposition de la communauté scientifique un vaste corpus d'écrits d'élèves et d'étudiants permettant des analyses quantitatives et des traitements automatiques* »[3]. Dans cette optique, nous avons opté pour le format TEI[4] qui constitue le format standard actuel le partage de corpus. Ainsi, l'ensemble des transcriptions réalisées respectent cette norme suivant les recommandations fournies dans la version P5 des directives TEI[5].

2. Nommage des fichiers

Chaque production sera stockée dans un fichier XML-TEI dont le nom devra respecter la norme de nommage suivante : `ETAB-NIV-ANNEE-CLASSE-DEVOIR-ELEVE-VERSION`

Où :

- ETAB : désigne le type d'établissement EC (école), CO (collège), LY (lycée), UN (université)
- NIV : le niveau scolaire : CP, CE1…
- ANNEE : l'année scolaire de production (un texte écrit le 3 mars 2015 compte pour l'année scolaire 2014)
- CLASSE : l'identifiant de la classe (XX si pas d'identifiant de classe) et/ou de l'établissement
- DEVOIR : identifiant du devoir : D1, D2….
- ELEVE : identifiant élève précédé d'une indication sur la provenance du corpus (E : Ecriscol, S : Scoledit…). Par exemple, S138 désigne l'élève 138 du corpus Scoledit.
- VERSION : V1, V2…

3. Saut de paragraphe *vs* saut de ligne

a. Paragraphe

Chaque production est vue comme une suite de paragraphes. On considère comme un nouveau paragraphe, le texte qui suit un retour à ligne volontaire de la part de l'enfant. Chaque paragraphe est encadré des balises … .

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE TEI SYSTEM 'TeiP5.dtd'>
<TEI>
     <teiHeader>
          … description et métadonnées…
     </teiHeader>
     <text>
          <front></front>
          <body>
               <head> si le devoir présente un titre </head>
               <!– production = une suite de paragraphes -->
```

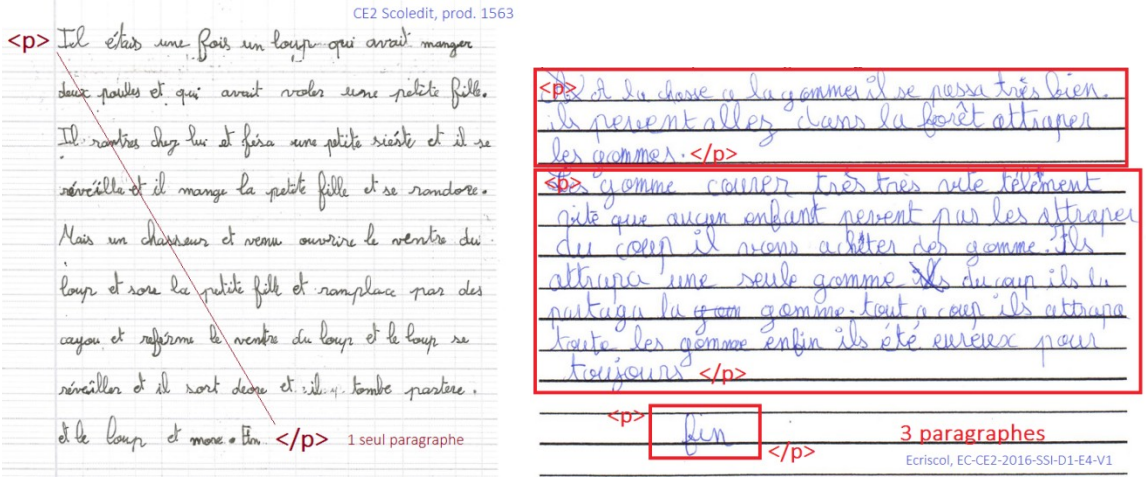
3 Proposition détaillée, Projet E:CALM, AAPG ANR 2017
4 http://www.tei-c.org
5 http://www.tei-c.org/Guidelines/P5/

```
                <p>  ...  </p>
                <p>  ...  </p>
            </body>
            <back></back>
        </text>
    </TEI>
```

Référence TEI

- **<p>** *(paragraphe) marque les paragraphes dans un texte en prose.* [www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-p.html]

Exemples



Explication *: le premier exemple comporte un seul et unique paragraphe alors que le deuxième comporte 3 paragraphes distincts.*

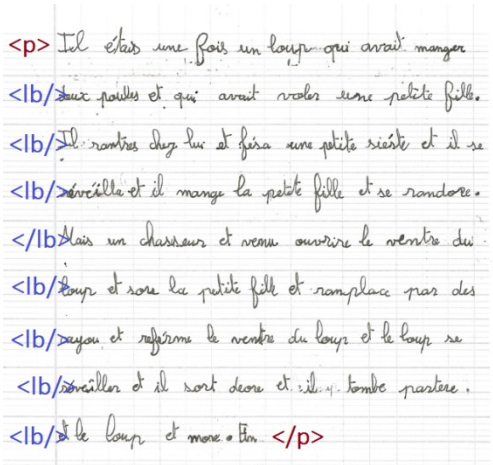
c. Lignes et pages

Afin de permettre un alignement scan/transcription à l'affichage, nous avons opté pour une transcription pseudo-diplomatique [réf.]. Les fins « physiques » de lignes et de pages sont notées respectivement par **<lb/>** **<pb/>**.

Références TEI

- **<lb>** *(début de ligne) marque le début d'une nouvelle ligne (typographique) dans une édition ou dans une version d'un texte.* [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-lb.html]
- **<pb>** *(saut de page) marque le début d'une page de texte dans un document paginé.* [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-pb.html]

Exemple



Explication*: cette production comporte un seul et unique paragraphe composé de 9 lignes. Le début de la première ligne coïncidant forcément avec le début du paragraphe n'est pas indiqué.*

4. Difficultés de lecture

Face à un passage mal écrit, la TEI offre plusieurs possibilités. Nous ne retenons que deux cas :

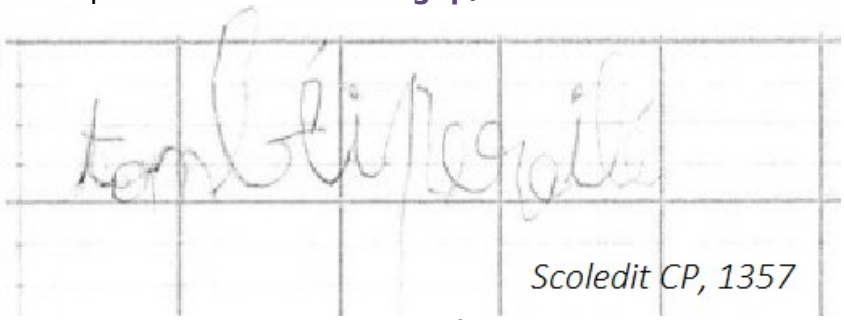
- Passage complètement illisible : **<gap/>**
- Passage lisible mais le transcripteur n'est pas sûr de sa transcription (ceci permettra une seconde relecture) : **<unclear> … </unclear>**

Références TEI :

- **<gap>** *(omission) indique une omission dans une transcription, soit pour des raisons éditoriales décrites dans l'en-tête TEI au cours d'un échantillonnage, soit parce que le matériau est illisible ou inaudible.* [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-gap.html]
- **<unclear>** *(incertain) contient un mot, une expression ou bien un passage qui ne peut être transcrit avec certitude parce qu'il est illisible ou inaudible dans la source.* [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-unclear.html]

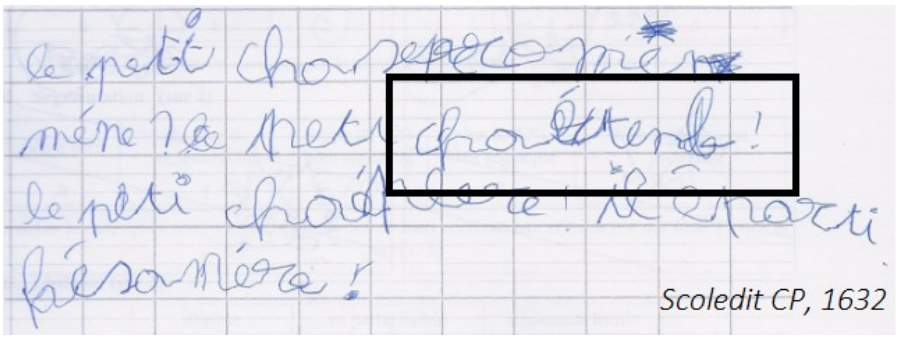
Exemples

Dans l'exemple ci-dessous (Scoledit CP, 1357), une large partie du texte est illisible. La transcription serait : `tonbéi`**<gap/>**



***Explication**: le texte compris entre `tonbéi` et le `i` final n'est pas déchiffrable.*

Dans l'exemple ci-dessous (Scoledit CP, 1632), le transcripteur n'est pas sûr de la transcription du passage encadré. Il notera donc **<unclear>**chaétenb**</unclear>** ce qui facilitera la révision de tous ces passages incertains.



5. Présence d'éléments extratextuels

a. Présence de dessins

Dans certaines productions, notamment au CP, on trouve parfois des dessins qui remplacent en partie ou complètement la production écrite. Les dessins ne seront pas décrits mais la balise **<figure/>** indiquera simplement leur existence.

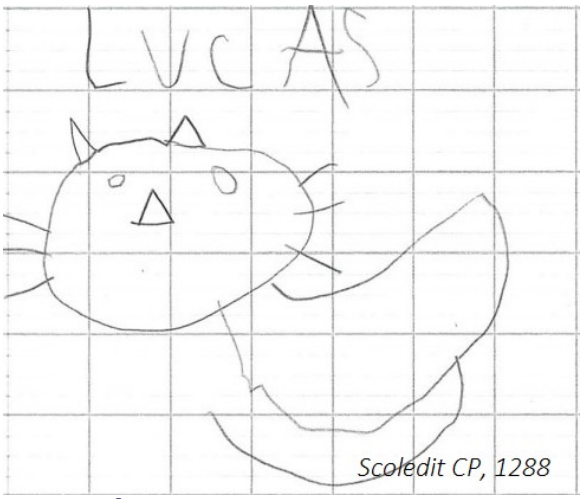
d. Présence du prénom

Les scans sont normalement anonymisés et ne devraient pas contenir d'indications sur l'identité de l'élève. Toutefois, dans certains cas, le prénom peut figurer sur le scan. Celui-ci n'est pas transcrit mais il est remplacé par la balise **<name/>**.

Références TEI

- **<name>** *(nom, nom propre) contient un nom propre ou un syntagme nominal.* [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-name.html]
- **<figure>** *(figure) regroupe des éléments représentant ou contenant une information graphique comme une illustration ou une figure.* [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-figure.html]

Exemple



Transcription : **<name/><figure/>**
Explication *: la transcription comporte une référence à un prénom et ou un nom suivi d'un dessin.*

e. Présence d'entête

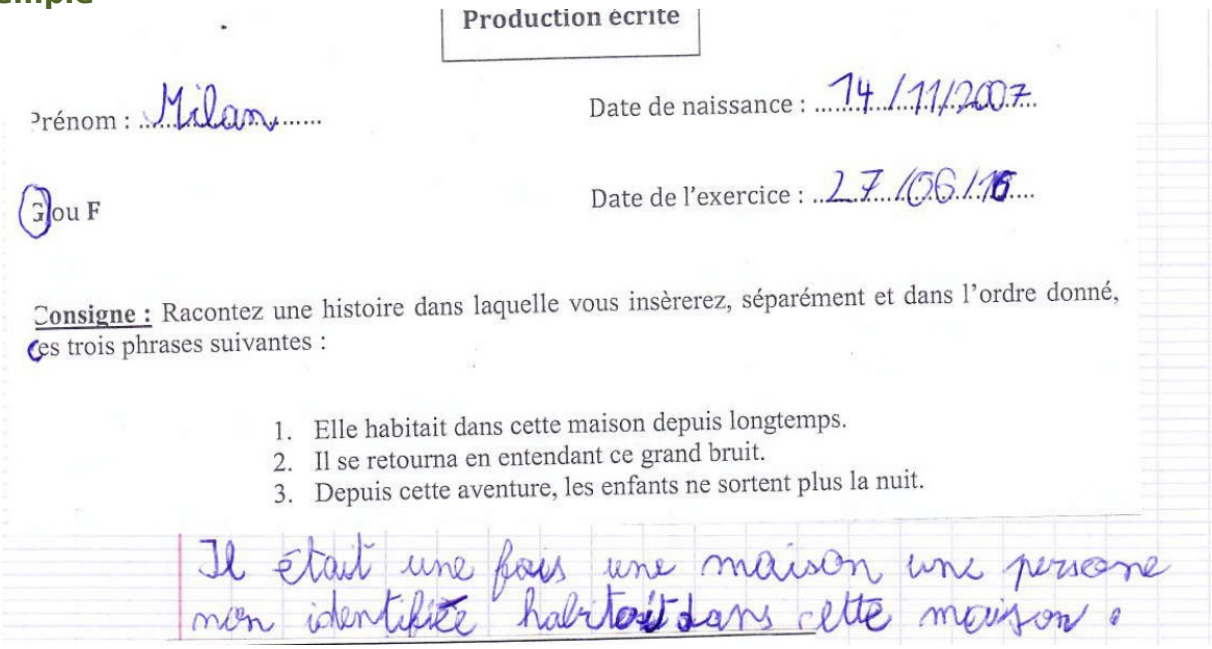
Certaines copies comportent des éléments situés avant le texte comme la date, des consignes, etc. Si l'on désire conserver ces informations, elles doivent être transcrites avant la zone **<text>** proprement dite et encadrées par les balises **<front>** et **</front>**.

Référence TEI

- **<front>** : *texte préliminaire) contient tout ce qui est au début du document, avant le corps du texte : page de titre, dédicaces, préfaces, etc.*

[http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-front.html]

Exemple



Production écrite

Prénom : Milan

Date de naissance : 14/11/2007

G ou F

Date de l'exercice : 27/06/15

Consigne : Racontez une histoire dans laquelle vous insèrerez, séparément et dans l'ordre donné, ces trois phrases suivantes :

1. Elle habitait dans cette maison depuis longtemps.
2. Il se retourna en entendant ce grand bruit.
3. Depuis cette aventure, les enfants ne sortent plus la nuit.

Il était une fois une maison une personne non identifié habitait dans cette maison.

Transcription : **<front>** Consigne ResolCo collée / copiée en en-tête **</front>**
Explication : *La consigne ResolCo apparaît collée en en-tête, précédant ainsi la production écrite de l'élève.*

f. Présence de pied-de-page

Certaines copies comportent des éléments situés après le texte. Si l'on désire conserver ces informations, elles doivent être transcrites après la zone **<text>** proprement dite et encadrées par les balises **<back>** et **</back>**.

Référence TEI

- **<back>** : *(texte annexe) contient tout supplément placé après la partie principale d'un texte : appendice, etc.*
  [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-back.html]

Ce champ sera utile si l'enseignant a rédigé un commentaire au bas de la copie par exemple.

6. Révisions de l'élève

Globalement, toutes les révisions du texte seront notées entre les balises **<mod>** et **</mod>**. Le type de la modification sera précisé avec l'attribut **type**.

Les valeurs possibles pour l'attribut **type** sont : **add**, **del**, **subst**.

- add : ajout/insertion
- del : suppression
- subst : remplacement

Référence TEI

- **<mod>** *represents any kind of modification identified within a single document.* [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-mod.html]

Les éléments modifiés sont ensuite décrits précisément.

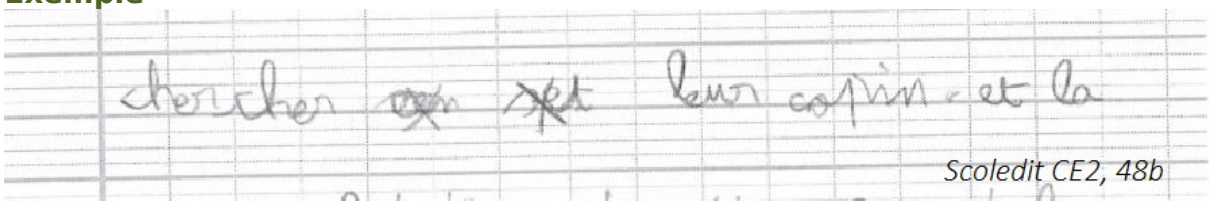
a. Suppression de texte

En cas de suppression de texte via une rature, un gommage… l'élément supprimé est encadré par les balises **~~** et **~~**. Si l'élément est lisible, il est alors transcrit sinon la balise **<gap/>** est utilisée.

Référence TEI

- **<del>** *(suppression) contient une lettre, un mot ou un passage supprimé, marqué comme supprimé, sinon indiqué comme superflu ou erroné dans le texte par un auteur, un copiste, un annotateur ou un correcteur.* [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-del.html]

Exemple



Transcription proposée

```
chercher <mod type="del"><del><gap/><del></mod> <mod type="del">
<del>set</del></mod> leur copin et la
```

Explication *: la production comporte deux éléments supprimés. Le premier n'est pas déchiffrable ; il est donc transcrit <gap/>. Le deuxième est déchiffrable ; il est donc explicitement transcrit entre les balises <del></del>.*

g. Insertion de texte

En cas d'insertion, le texte ajouté est indiqué entre les balises **<add>** et **</add>**.

Référence TEI

- **<add>** *(ajout) contient des lettres, des mots ou des phrases insérés dans le texte par un auteur, un copiste, un annotateur ou un correcteur.* [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-add.html]

Exemple

Transcription proposée

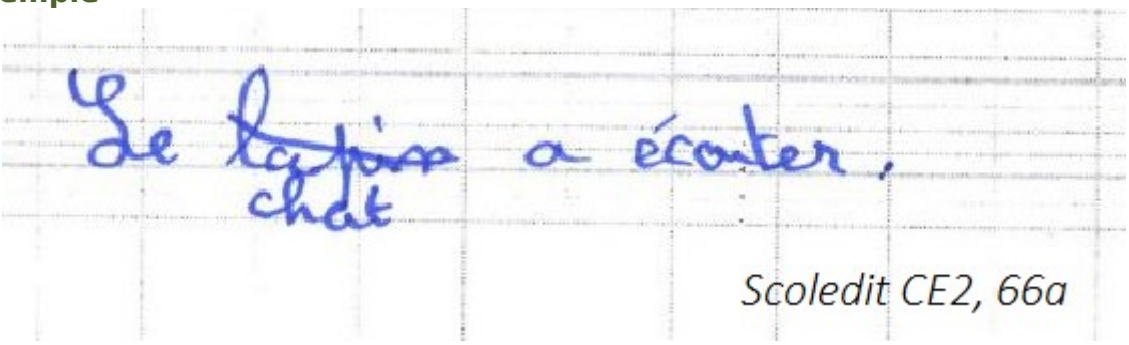
```
toi et le <mod type="add"><add>chat</add></mod> a faim
```

Explication *: l'enfant a inséré le mot* ***chat*** *dans sa production. La transcription est réalisée à l'endroit où le mot doit s'insérer.*

h. Remplacement de texte

En cas de remplacement d'un texte par un autre, l'élément supprimé est indiqué entre les balises **~~** et **~~** (cf.6.a) et l'élément inséré est placé entre les balises **<add>** et **</add>** (cf.6.b).

Exemple



Transcription proposée

```
Le <mod type="subst"><del>lapin<del><add>chat</add></mod> a écouter.
```

Explication *: l'enfant a remplacé le mot* ***lapin*** *par le mot* ***chat*** *dans sa production. Pour cela, il a supprimé le mot* ***lapin*** *en le rayant et a ajouté le mot* ***chat****. La transcription est réalisée à l'endroit de la substitution dans le texte.*