# Guide et format de transcription E-Calm – version 1.5

| Version | Date | Rédacteur.trice.s | Contributeur.trice.s |
|---|---|---|---|
| 1.2 | 09/02/2018 | Claude Ponton | *Myriam El Helou, Xiaoshu Xie, Mai Ho-Dac, Serge Fleury, Claude Ponton* |
| | *Cette version du guide de transcription ne s'intéresse qu'au temps 1 de la production à savoir le premier jet. Les temps suivants (commentaires de l'enseignant, corrections par l'élève…) seront traités ultérieurement. La définition et le codage des métadonnées sont à l'étude et ne sont pas intégrées à cette version.* | | |
| 1.3 | 04/12/2019 | Claude Ponton | *Claude Ponton* |
| | *Révisions de la version précédente : structure générale, structure de la production*<br>*Ajout de la structure titre : <head>*<br>*Suppression de la balise <head> pour reprise du nom de fichier*<br>*Ajout de la balise <surplus>* | | |
| 1.4 | 13/03/2020 | Claude Ponton | |
| | *Corrections et compléments*<br>*Ajout chapitre Opérations de révision*<br>*Ajout chapitre Éléments textuels (caractères spéciaux, césure)*<br>*Ajout paragraphe Mise en forme* | | |
| 1.5 | 03/04/2020 Claude Ponton | | |
| | *Proposition de transcription des commentaires généraux des enseignants* | | |
| | O6/04/2020 | | |
| | *Proposition de transcription des commentaires spécifiques* | | |
| 1.6 | *Ajout du renseignement des méta-données dans le teiHeader* | | |

## Table des matières

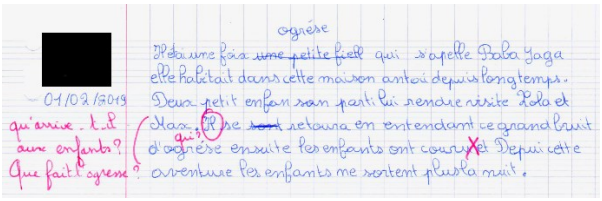

Table des matières ..... 1

1. Notre approche de la transcription ..... 3
   a. Que transcrire ? ..... 3
   b. Comment transcrire ? ..... 4
2. Nommage des fichiers ..... 4
3. Structure générale du fichier XML-TEI ..... 4
4. Éléments de structure du <teiHeader> contenant les méta-données ..... 5
5. Éléments de structure du <text > contenant le texte de la copie ..... 7
   a. Paragraphe ..... 8
   c. Lignes et pages ..... 8
   d. Titre ..... 9
6. Éléments textuels ..... 9
   a. Typographie ..... 10

b. Césure ..... 10
7. Difficultés de lecture ..... 11
8. Présence d'éléments extratextuels ..... 12
a. Mise en forme ..... 12
b. Présence de dessins ..... 12
e. Présence du prénom ..... 12
f. Présence d'entête ..... 13
g. Présence de pied-de-page ..... 14
h. Autres éléments métatextuels ..... 14
9. Opérations de révision ..... 15
a. Suppression de texte ..... 16
b. Insertion de texte ..... 16
c. Remplacement de texte ..... 17
10. Commentaires enseignants ..... 18
a. Commentaires généraux ..... 18
b. Commentaires spécifiques ..... 19
Quelques exemples de commentaires ..... 20
L'exemple de Littéracie avancée ..... 21
Application au projet E-Calm ..... 21

# 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 (cf. exemple figure 1). 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 ? ».



```
<p>
Il étai une foix <mod type="subst"><del>une petite fiell</del><add>
ogrése</add></mod> qui s'apelle BabaYaga <lb/>elle habitait dans cette
maison antai depuis longtemps. <lb/>Deux petit enfan son parti lui
rendre visite Lola et <lb/>Max. Il se <mod type="subst"><del>sont
</del></mod> retoura en entendant ce grand bruit <lb/>d'ogrése ensuite
les enfants ont couru et Depuis cette <lb/>aventure les enfants ne
sortent plus la nuit.
</p>
```

Fig.1 Exemple d'une copie et de sa transcription

## 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.
- Le recours à un outil de transcription permet d'éviter les erreurs d'encodage XML.

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

### 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 pour 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. Structure générale du fichier XML-TEI

Dans le respect de la norme TEI P5, chaque fichier aura la structure générale suivante :

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE TEI SYSTEM 'TeiP5.dtd'>
<TEI>
	<teiHeader>
		<!-- description et métadonnées à définir -->
	</teiHeader>
	<text>
		<front>
			<!-- facultatif -->
			<!-- éléments situés avant la production -->
		</front>
		<body>
			<!-- contenu de la production -->
		</body>
		<back>
			<!-- facultatif -->
			<!-- éléments situés après la production -->
		</back>
	</text>
</TEI>
```

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

## 4. Éléments de structure du <teiHeader> contenant les méta-données

La partie **<teiHeader>** d'un fichier XML encodé selon la TEI contient toutes les métadonnées c'est-à-dire toutes les informations relatives à la description du fichier et aux caractéristiques de l'encodage selon la TEI-P5 du document.

**Parties principales :**

**fileDesc** qui décrit le fichier XML : son nom, sa licence de diffusion, le projet à l'origine du fichier (considéré comme l'éditeur) et les différentes étapes de sa génère

**profileDesc** qui contient les méta-données associées à la copie :

- **textDesc** : le type de texte produit (brouillon, réécriture, rédaction, etc.),
- **particDesc** : les auteurs des différentes traces d'écriture (élève, enseignant)
- **settingDesc** : le contexte d'écriture (information sur l'établissement scolaire)

```
<teiHeader>
   <fileDesc>
      <titleStmt>
        <title><!-- nom du fichier XX-XXX-20XX-XXX-RX-VX.xml --></title>
        <respStmt>
          <resp>collecte et scan de la copie</resp>
          <name><!-- nom de la / des personne(s) ayant récolté le fichier .xml. --></name>
        </respStmt>
        <respStmt>
          <resp>transcription et encodage au format XML - TEIP5</resp>
          <name><!-- nom des personnes ayant réalisé la transcription du scan au format XML --></name>
        </respStmt>
      </titleStmt>
      <publicationStmt>
        <publisher>
          <orgName><!--laboratoire--></orgName>
          <address>
            <street></street>
            <postCode></postCode>
            <settlement></settlement>
          </address>
        </publisher>
        <availability status="free">
          <licence target="https://creativecommons.org/licenses/by-sa/3.0/legalcode.fr">
            <p>Ce fichier numérique est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage à l'Identique 3.0 France (http://creativecommons.org/licenses/by-nc-sa/3.0/fr)</p>
                  <p>Autorisation parentale signée le xx/xx/xxxx</p>
          </licence>
        </availability>
      </publicationStmt>
      <sourceDesc>
```

```
        <bibl>Copie d'élève récoltée dans l'académie de XXX dans le cadre du projet É:calm : "Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques" (ANR- 17- CE28- 0004- 04)
          <extent><!-- nb de pages de la copie--></extent>
        </bibl>
      </sourceDesc>
</fileDesc>

<encodingDesc>
    <projectDesc>
Ce fichier a été produit dans le cadre du projet É:calm : "Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques" (ANR- 17- CE28- 0004- 04) démarré en 2018 et achevé en 2022· http://e-calm·huma-num·fr/
    </projectDesc>
</encodingDesc>

<profileDesc>
 <langUsage>
  <language ident="fr">French</language>
 </langUsage>

 <textDesc n="schoolwork">
  <channel mode="w">manuscript</channel>
  <constitution type="single">chaque texte correspond à une copie produite en classe par un élève ou un étudiant· Pour certaines copies, des interventions de l'enseignant sont également présentes·</constitution>
  <derivation type="(original/revision)"/>
  <domain type="education"/>
  <factuality type="(fiction/mixte/scientifique/…)"/><!--insérer ici tout commentaire sur la factualité supposée du texte produit· Par ex : "La majorité du texte est imaginaire· Certains éléments factuels peuvent cependant apparaître·" On peut également insérer ici des informations sur la consigne données· Ex : "La consigne donnée aux enfants est la suivante : "--></factuality>
  <interaction type="none"><!-- Indiquer ici si le travail a donné lieu à une évaluation noté ou si c'était une activité libre--></interaction>
  <preparedness type="(none/revised/scripted)"><!-- indiquer ici le type de travail préparatoire qu'il y a eu avant la réalisation du devoir
"none" pour les écrits spontannés et non préparés
"scripted" pour les écrits qui se basent sur une étape de brouillon
"formulaic" pour les écrits qui suivent des règles précises d'écriture e·g· répondre à un formulaire
"revised" pour les écrits résultant d'une réécriture sur la base d'une première version
-->
</preparedness>
  <purpose type="entertain" degree="medium"></purpose>
  <purpose type="express" degree="high"/>
  <!-- parmi les visée discursives possibles, la TEI-P5 distingue :  "persuade", "express", "inform" et "entertain" -->
 </textDesc>

 <particDesc>
    <listPerson>
      <person age="##ans##mois" xml:id="W1" sex="(F|G)">élève né·e en 20## au mois ##</person>
```

```
      <!-- Pour l'indication des années et des mois, mettre systématiquement deux chiffres. Ex : age="09ans02mois" et élève né·e en 2002 au moi 06 -->
      <person xml:id="W2" sex="G">professeur·e ayant plus de X années d'expérience / professeur·e stagiaire</person>
      </listPerson>
 </particDesc>

 <settingDesc>
     <name type="region"></name>
     <date><!-- Indiquer ici la date de production. Si la date de production n'est pas connue, indiquer la date de signature de l'autorisation --></date>
     <locale>(école/collège/lycée/université) publique (en éducation prioritaire REP(+)/hors REP) en zone (rurale/urbaine/mixte). (Moins de 100 élèves/Entre 100 et 200 élèves/…/plus de 500 élèves) </locale>
    <activity>(Rédaction/Dictée/xxxx)</activity>
    <p><!-- indiquer ici toute information complémentaire sur la situation de production --></p>
 </settingDesc>
</profileDesc>
</teiHeader>
```

vérifier les méta-données en vous basant sur les autorisations et autres informations fournies

Il faut absolument que la date de rédaction soit indiquée. Si cette l'information est manquante, mettre la date de signature de l'autorisation.
Une fois la date modifiée, mettre à jour l'âge de l'enfant dans les <particDesc>

Caractéristiques de l'établissement (<locale> dans l'élément <settingDesc>) :

- la zone éducation prioritaire, se baser sur les informations fournies ici : https://www.reseau-canope.fr/education-prioritaire/sinformer/annuaires/academie/grenoble.html
- la zone, chercher la localisation sur Google Map, le nb d'habitants de la ville et également des recherches sur le site de l'académie, du ministère de l'éducation, du rectorat, etc.

ex : https://mobile.education.gouv.fr/pid24301/annuaire-accueil-recherche.html
Note : il n'y a pas de nomenclature nationale et aucun collègue n'est d'accord sur les critères. Donc le mieux est de trancher nous mêmes, même si on n'est pas acteur de l'éducation nationale !

- l'effectif doit également être cherché dans les documents founies par l'éducation nationale

## 5. Éléments de structure du <text > contenant le texte de la copie

La partie <text> d'un fichier XML encodé selon le TEI-P5 correspond au corps du texte qui est lui-même subdivisé en 3 parties principales :

```
<text>
<front>
      <!-- Si des éléments ont été insérés avant le début à proprement parlé du texte (ex : la consigne, la date, etc.), les insérer ici -->
</front>
```

```
<body>
<head><!-- Si un titre a été donné au texte, l'indiquer ici --></head>


</body>
<back>
        <!-- Si des éléments ont été insérés après ou en marge du texte, les insérer ici -->
        <metamark who="transcripteur"><!-- indiquer ici tout commentaire relatif à la transcription--></metamark>
        <metamark who="R2 "><!-- indiquer ici les commentaires du professeur présents sur la copie --></metamark>
</back>
</text>
```

une partie <body> correspondant au corps de texte et deux parties facultatives contenant dans une partie <front> les éléments textuels précédent le corps de texte en lui-même (ex : la page de titre, les préfaces, la tables des matières, etc.) et dans une partie <back> les éléments textuels suivant le corps de texte en lui-même (ex : la bibliographie, les notes de  n, les annexes, les postface, etc.)

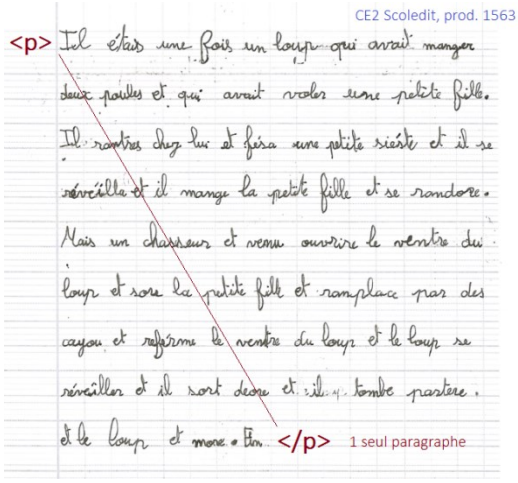
## a. Paragraphe

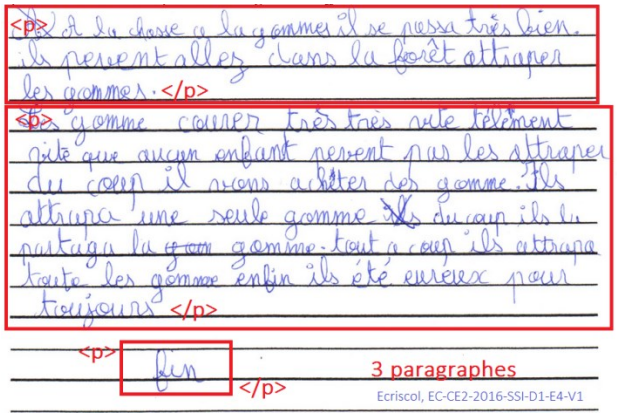
Chaque production est vue comme une suite de paragraphes. On considère comme un nouveau paragraphe, le texte qui suit un retour à ligne volontaire de la part de l'enfant. Chaque paragraphe est encadré des balises … .

**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

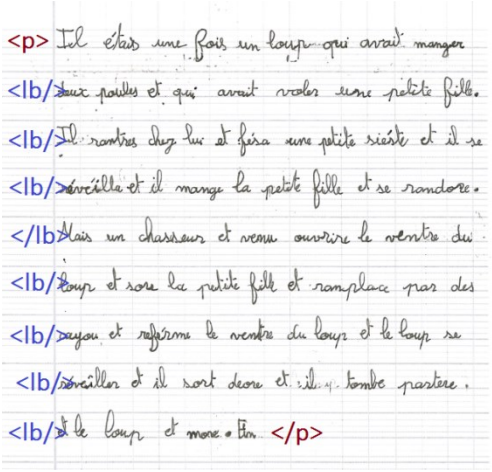
Afin de permettre un alignement scan/transcription à l'affichage, nous avons opté pour une transcription pseudo-diplomatique [réf.]. Les fins « physiques » de lignes et de pages sont notées respectivement par `<lb/>` et `<pb/>`.

**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.*

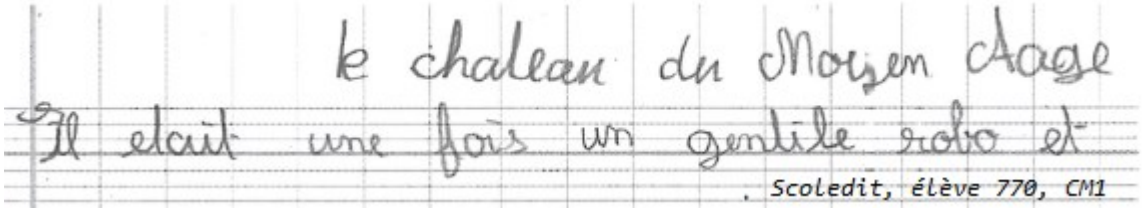
### d. Titre

La présence d'un titre définit par le scripteur dans la production fera l'objet d'un marquage spécifique à l'aide de la balise **<head>**.

**Références TEI :**

- **<head>** *(en-tête) contient tout type d'en-tête, par exemple le titre d'une section, ou l'intitulé d'une liste, d'un glossaire, d'une description de manuscrit, etc.*

**Exemples**




Dans l'exemple ci-dessus (Scoledit CM1, 770), la première ligne indique clairement un titre pour l'histoire.

**Transcription** : `<head>le château du Moyen Aage</head> Il etait une fois un gentile robo et <lb/>`

## 6. Éléments textuels

Le principe général de transcription du texte est de rester au plus proche de la graphie de l'enfant. Ainsi, dans la majorité des cas (majuscules accentuées, inversion d'accent…), on respectera fidèlement les caractères utilisés. Toutefois, certains cas (ambiguïté, nécessité de traitement ultérieur…) obligent à déroger à ce principe.

## a. Typographie

Certains caractères posent problème lors des phases d'analyse. La transcription de ces caractères est indiquée dans le tableau suivant :

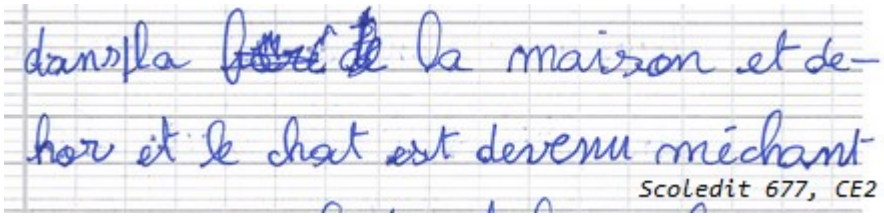
| Caractère | Transcription |
|---|---|
| œ *ou* Œ | oe, OE |
| æ | ae |
| « | « |
| " *ou* ” | " |
| ' *ou* ‘ *ou* ´ | ' |
| € | € |

## b. Césure

Le tiret est utilisé par les enfants pour décrire aussi un séparateur de mot-composé (*ie*. Moyen-Âge) qu'une césure. A ce titre, il représente une ambiguïté lors des phases d'analyse. Pour lever cette ambiguïté, nous reprenons le procédé utilisé à Tours par l'équipe des Bibliothèques Virtuelles Humanistes pour encoder les textes de la Renaissance et temps modernes[6].

Le signe de césure[7] est conservé mais la marque de fin ligne (<lb/>) est typée à l'aide de l'attribut rend="hyphen".

**Exemple**




**Transcription** : [...] `la maison et de-<lb rend="hyphen"/>hor et le` [...]

Ici, le tiret indique une césure. Lors de l'analyse la forme "**dehor**" sera reconstituée et le tiret supprimé contrairement à l'exemple ci-dessous où **Moyen-<lb/>Âge** sera reconstitué en **Moyen-Âge** avec conservation du tiret.

[6] http://www.bvh.univ-tours.fr/XML-TEI/manuelTEIrenaissance3_2012.pdf
[7] Le signe de césure conventionnel en français est le '-' mais on trouve parfois d'autres symboles comme le '='.


## 7. Difficultés de lecture

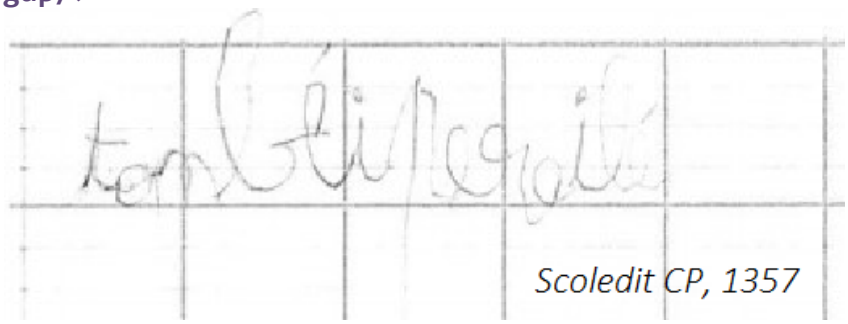
Face à un passage mal écrit, la TEI offre plusieurs possibilités. Nous ne retenons que deux cas :

- 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 :**

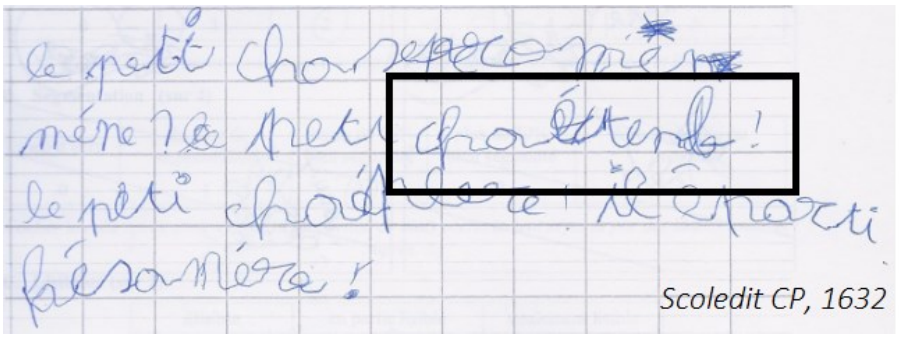
- `<gap>` *(omission) indique une omission dans une transcription, soit pour des raisons éditoriales décrites dans l'en-tête TEI au cours d'un échantillonnage, soit parce que le matériau est illisible ou inaudible.* [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-gap.html]
- `<unclear>` *(incertain) 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]

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.



# 8. Présence d'éléments extratextuels

## a. Mise en forme

Certaines productions présentent de marques de mise en forme comme des soulignements ou des passages colorés. La balise **<hi>** associée à l'attribut rend (https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-att.global.rendition.html) est alors utilisée pour les marquer. Les valeurs de rend ne sont pas contraintes.

**Références TEI**

- **<hi>** *(mis en évidence) distingue un mot ou une expression comme graphiquement distincte du texte environnant, sans en donner la raison.* [https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-hi.html]

Les valeurs de rend ne sont pas contraintes par la TEI, voici la liste de celles utilisées dans le projet E-Calm :

- underline : élément souligné

**Exemple** : à trouver

**Transcription** :

## b. 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.

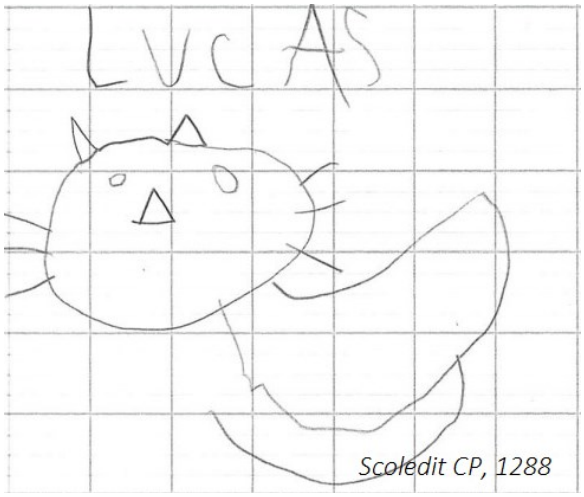
## e. Présence du prénom

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.*

## f. Présence d'entête

Certaines copies comportent des éléments situés avant le texte comme la date, des consignes, des interventions générales de l'enseignant, 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**
Mettre une copie Resolco où la consigne a été collée ou copiée en entête de la copie

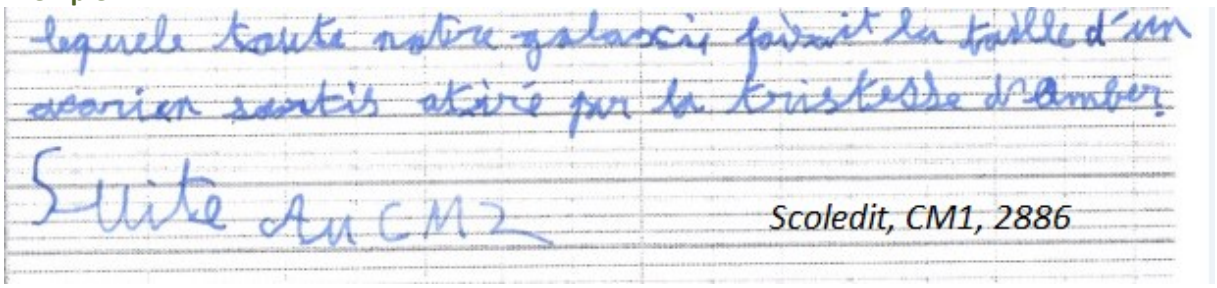
## g. Présence de pied-de-page

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]

**Exemple**




**Transcription :** `<body> […] atiré par la tristesse d'Omber. </p></body> <back>Suite Au CM2</back>`

## h. Autres éléments métatextuels

Dans le texte même des productions tous les éléments qui ne participent pas au contenu (ie. qui ne rentrent pas en compte dans l'analyse linguistique) mais qui doivent être transcrits, sont encadrés avec la balise **<surplus>**.

**Référence TEI**

- **<surplus>** *(Texte superflu) permet d'encoder une partie de texte présente dans la source lorsque l'éditeur la considère superflue ou redondante.*

**Exemple**


**Transcription :** `<head>Les meilleures amies</head>` **<surplus>**`1`**</surplus>** `Un jour une sorcière qui se promenée<lb/>` **<surplus>**`2`**</surplus>** `dans la forêt […]`

**Explication** : la numérotation des lignes n'entre pas en compte dans l'énoncé mais doit être transcrite pour rester fidèle au fac-similé. Chaque numéro est donc encadré de la balise **<surplus>**.

# 9. Opérations de révision

Comme le précisent Claire Doquet *et al*. (2017), "*Au-delà des textes eux-mêmes, nous souhaitons accéder à la reconstitution de l'écriture telle qu'elle se donne à lire à partir des ratures et de l'ensemble des interventions, verbales ou non, opérées sur les différents états du texte*". Il s'agit donc de transcrire les différents états des textes : écriture et révision de l'enfant, interventions de l'enseignant (révision et commentaires), réécriture de l'enfant... Ce chapitre ne traitent que des opérations de révision; les opérations de commentaires seront décrites dans le chapitre XXX.

Chaque opération de révision est notée entre les balises **<mod>** et **</mod>.**

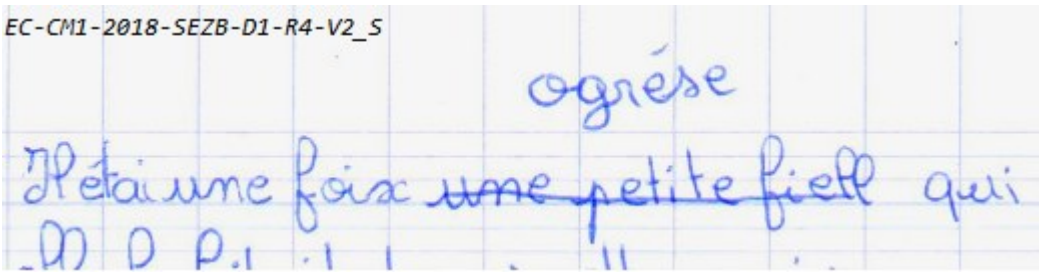
**Référence TEI**

- **<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]

Ces opérations de révision sont caractérisées par les éléments suivants :

- une temporalité : premier jet de l'enfant, correction enseignante...
  - Attribut : **stage**
  - Valeurs possibles de **stage** : T1, T2, T3, T4...
- un scripteur : celui qui intervient
  - Attribut : **who**
  - Valeurs possibles de **who**
    - E : enfant
    - P : enseignant
- un type de révision : effacement, ajout, substitution
  - Attribut : **type**
  - Valeurs possibles de **type**
    - add : ajout/insertion
    - del : suppression
    - subst : remplacement
    -

**Exemple**




**Transcription** : *Il étai une foix <mod type="subst" who="E" stage="T1"><del>une petite fiell</del><add>ogrése</add></mod> qui…*

Les éléments modifiés sont ensuite décrits précisément.

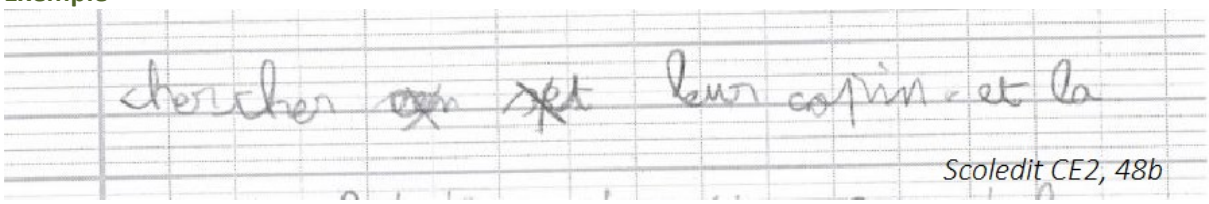
## a. Suppression de texte

En cas de suppression de texte via une rature, un gommage… l'élément supprimé est encadré par les balises **<del>** et **</del>**. 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" stage ="T1" who ="E"><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>.*

## b. 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**

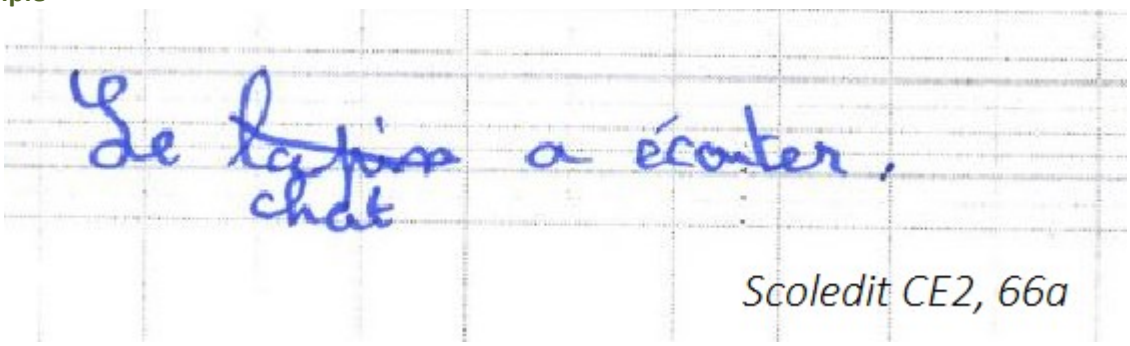
```
toi et le <mod type="add" stage ="T1" who ="E"><add>chat</add></mod> a faim
```

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

## c. Remplacement de texte

En cas de remplacement d'un texte par un autre, l'élément supprimé est indiqué entre les balises **<del>** et **</del>** (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" stage ="T1" who ="E">
<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.*

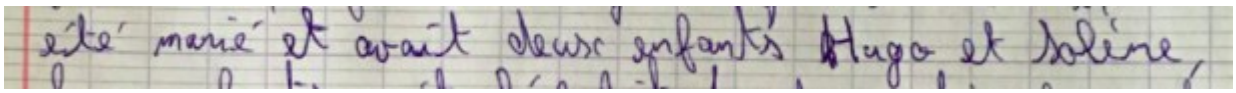
## d. En cas d'incertitude sur les révisions

Il n'est pas rare de se retrouver face à une incertitude quant à l'interprétation des révisions. Par exemple, lorsque il y a une superposition de caractères et que rien ne permet de savoir avec certitude quelle chaîne de caractère s'est substitué à quelle chaîne.
Fac à cette situation, l'annotateur pourra indiquer son degré de certitude en insérant l'attribut @cert aux balises <mod> et en indiquant comme valeur son degré de certitude qui peut aller de "low" (certitude faible) à "mid" (certitude moyenne). lorsque l'annotateur est certain de son annotation (cert="high"), il est inutile de l'indiquer afin de ne pas surcharger la balise.
Exemples:


**Transcription proposée**

**<lb/>**le grenier et ils entendérent un autre **<unclear>**bu**</unclear>**it.**<mod** type="subst" cert="low" stage ="T1" who ="E"**><del>**e**</del><add>**E**</add></mod>**t ils vi**<lb rend="hyphen"/>**

L'attribut @cert porte sur l'action d'écriture que transcrivent les balises (le fait de faire une rature, <mod type="del " cert="low">) et non la reconnaissance des caractères qui est encodé grâce à la balise <unclear>. Les deux éléments peuvent donc se combiner comme dans l'exemple :



**Transcription proposée**

```
<lb/>été marié et avait deux enfants <mod type="subst" cert="low" stage="T1" who="E"><del><unclear>h</unclear></del><add>H</add></mod>ugo et Solène
```

***Explication** :*

# 10. Commentaires enseignants

## a. Commentaires généraux

On considère comme commentaires généraux, les interventions de l'enseignant ne portant pas sur un passage spécifique mais sur l'intégralité de la production. Selon leur emplacement sur la copie, ils seront transcrits dans la partie <front> (ie. pour les commentaires placés plutôt en début de copie) ou dans la partie <back> (ie. pour les commentaires placés plutôt en fin de copie).

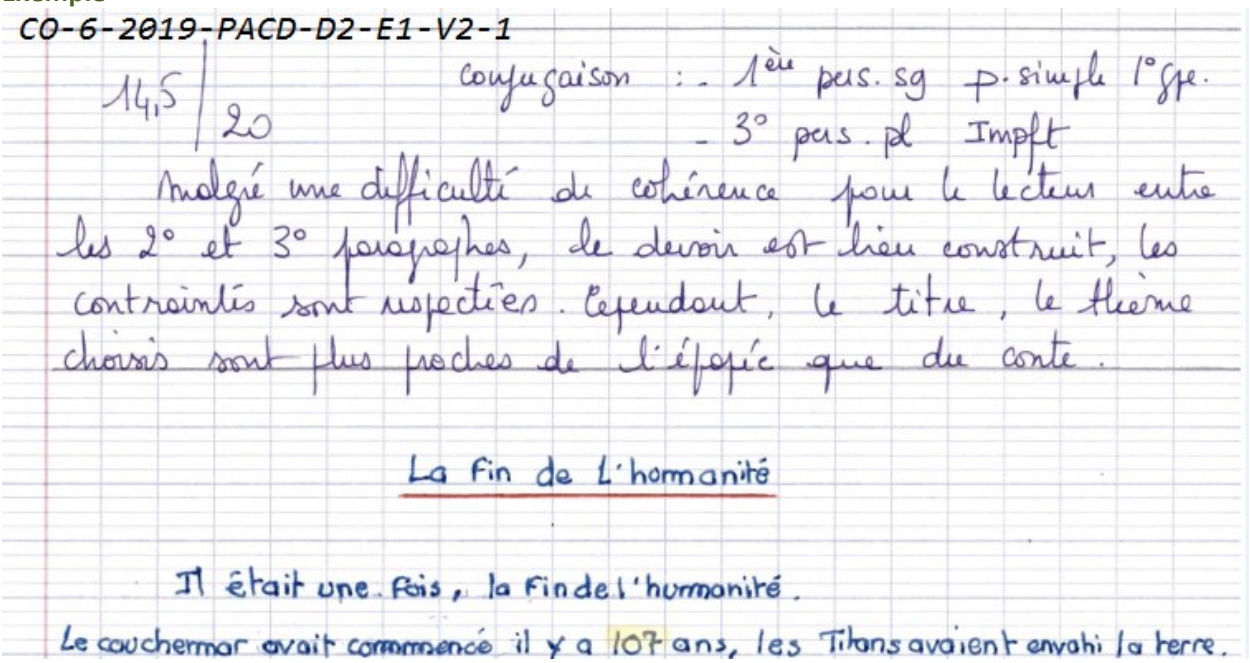
Chaque commentaire est transcrit entre les balises `<metamark>` et `</metamark>.`

**Référence TEI**

- `<metamark>` : *contains or describes any kind of graphic or written signal within a document the function of which is to determine how it should be read rather than forming part of the actual content of the document* [https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-metamark.html]

Cette balise <metamark> peut être complétée avec les attributs ***stage***, ***who*** et ***rend*** (cf. § précédents).

**Exemple**


CO-6-2019-PACD-D2-E1-V2-1

14,5/20

conjugaison : - 1ère pers. sg p. simple 1°gpe.
- 3° pers. pl Impft

Malgré une difficulté de cohérence pour le lecteur entre les 2° et 3° paragraphes, le devoir est bien construit, les contraintes sont respectées. Cependant, le titre, le thème choisis sont plus proches de l'épopée que du conte.

La fin de L'humanité

Il était une fois, la Fin de l'humanité.
Le cauchemar avait commencé il y a 107 ans, les Titans avaient envahi la terre.

**Transcription proposée**

```
<text>
<front>
      <metamark stage="T2" who="P">14,5/20 </metamark>
      <metamark>conjugaison : - 1ère pers. sg p. simple 1e gpe. <lb/> - 3e pers. pl.
      impft</metamark>
      <metamark stage="T2" who="P">malgré une difficulté de cohérence pour le lecteur
      entre<lb/> les 2e et 3e paragraphes, le devoir est bien construit, les<lb/> contraintes sont
      respectées. Cependant, le titre, le thème<lb/> choisis sont plus proches de l'épopée que du
      conte.</metamark>
</front>
<body>
   <head><hi rend="underline" stage="T1" who="E">La fin de L'humanité</hi>
   </head>
```

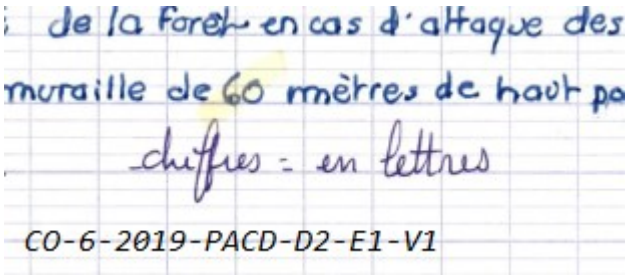
***Explication** : le commentaire de l'enseignant est situé au début de la copie. Sa transcription est donc placée dans la partie <front>. Les 3 commentaires distincts (ie. note, remarque sur la conjugaison et commentaire général sur le travail) sont transcrits dans 3 balises <metamark> séparées. Cela permettra par la suite de typer différemment ces commentaires.*

## b. Commentaires spécifiques

On considère comme « commentaires spécifiques » les commentaires portant sur une portion de texte délimitée. Plusieurs éléments décrivent ce type de commentaire :

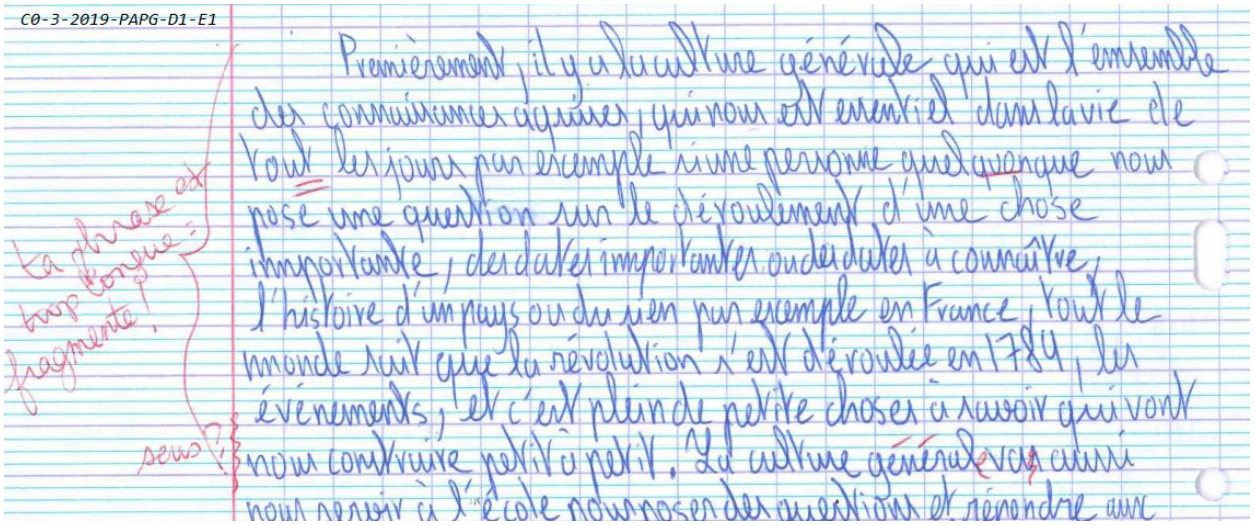
- La délimitation de la portion
- Le lien entre la portion et le commentaire
- Le commentaire lui-même

## Quelques exemples de commentaires




Dans l'exemple ci-dessus, on a :

- Délimitation implicite à savoir le nombre 60
- Lien implicite : pas de trace directe mais juste un placement proche
- Commentaire : *chiffres = en lettres*




Dans l'exemple ci-dessus, on trouve, en plus des révisions directes du texte (« … générale va… ») traitées précédemment, plusieurs annotations de l'enseignant : des soulignements de parties de mots et deux commentaires dans la marge portant sur des portions de texte.

**Soulignement des mots**

- Délimitation directe par le soulignement
- Lien implicite : comme le commentaire est vide, il n'y a pas de lien
- Commentaire implicite (faute d'orthographe)

**Commentaires dans la marge**

- Délimitation directe une accolade ou un trait ondulé
- Lien implicite : le commentaire est collé au délimiteur
- Commentaires : « *ta phrase est trop longue = fragmente !* » et « *sens ?* »

Chaque enseignant commente les copies à sa façon. La délimitation des zones et le marquage des liens sont donc extrêmement variées. Toutefois, comme les tâches d'analyse du projet E-Calm ne s'intéressent qu'au type d'intervention des enseignants, nous ne transcriront pour chaque commentaire spécifique que son contenu et la partie du texte ciblé. Nous reprenons pour cela l'approche TEI adoptée par M.P. Jacques (réf) pour la constitution du corpus de Littéracie avancée.

## L'exemple de Littéracie avancée

Avant toute chose, rappelons que le corpus Littéracie avancée est composé de copies numériques d'étudiants commentées via un logiciel de traitement de texte. Il n'y a donc pas de phase de transcription. La problématique consistait ici à passer ces copies annotées dans le format TEI.

**Principes généraux**

- Les liens entre sélections et commentaires ne seront pas décrits
- Les différentes interventions des enseignants pouvant potentiellement se chevaucher, les concepteurs du corpus ont choisi de n'utiliser que des balises autofermantes pour identifier les zones sélectionnées.
- Le contenu des commentaires est décrit dans des balises **<note>** dans le **<back>** avec un lien vers la zone sur lequel il s'applique.
- Il est possible de décrire le positionnement approximatif du commentaire dans la copie avec l'attribut **place**.

**Exemple**

- Sélection de la zone
  - *Nous tenterons donc en* *****s'***<anchor xml:id="endC5"/>** *appuyant principalement sur les réflexions de Cisel* ;
  - Le « s' » est donc repéré entre les deux balises autofermantes  et <anchor/> avec l'identifiant « c5 ».
- Contenu du commentaire (dans le <back>)
  - **<note place="margin-right" target="#c5">** « nous » **</note>**
  - Le texte du commentaire est mis en <note> avec une référence vers la cible (target="#c5").
  - Le commentaire est situé dans la marge droite de la copie (place="margin-right").

## Application au projet E-Calm

Pour le projet E-Calm, nous avons besoin de décrire également le temps de l'intervention et l'intervenant. Nous ajouterons donc les attributs **who** et **stage** utilisés précédemment. Pour ne pas répéter inutilement l'information, ces deux attributs seront uniquement attachés au commentaire lui-même. Pour garder une cohérence avec l'ensemble des descripteurs, les commentaires seront indiqués dans des balises **<metamark>** placées dans le **<back>.** Pour l'exemple précédent, cela donnerait :

- Sélection de la zone
  - *Nous tenterons donc en* *****s'***<anchor xml:id="endC5"/>** *appuyant principalement sur les réflexions de Cisel* ;
- Contenu du commentaire (dans le <back>)
  - **<metamark who="P" stage="T2" place="margin-right" target="#c5">** « nous » **</metamark>**
  - Le commentaire a été fait par l'enseignant dans le temps T2.
  - *A discuter*
    - *L'attribut peut être facultatif*
    - *On peut ajouter l'attribut rend s'il l'on veut préciser la mise en forme*