Recueil et structuration de corpus – TD 3

Achille Falaise – Alexandre Roulois

Plan du TD

- Correction TD 2
- Balises à connaître
- Langue / script
- Pratique

Correction du TD 2

- Trouvez l’erreur

```
<?xml version="1.0" encoding="UTF-8?>"
<TEI>
    <teiheader xml:lang="fr">
    <titleStmt>
        <title> Interpellée pour une meilleure prise en compte des enfants transgenres, l’école tâtonne 03-02-2021</title>
          <authors> Mattea Battaglia et Solène Cordier </authors>
            <editor> satou</editor>
             </titleStmt>
             <publicationStmt>
               <availability statut: "limité">
            </availability>
               </publicationStmt>
               </titleStmt>
               </teiHeader>
                 <text>
                  <body>
                      <p>des enseignants aux conseillers principaux d’éducation (CPE), des infirmières scolaires aux assista
                         <p>es questionnements d’élèves, l’éducation nationale en a pris acte : « On perçoit, empiriquement,
                          <p>Une professeure confie, au détour d’une conversation, que deux de ses élèves de terminale sont
                          <p>Ce sont des jeunes gens qui, par milliers, s’affichent sur les réseaux sociaux sous les hashtag
                          <p>La multiplication de ces signaux a conduit l’éducation nationale, réputée plutôt frileuse à s’e
                             </body>
                             </text>
                                </TEI>

```

Correction du TD 2

- Trouvez les 3 erreurs

```
<?xml version="1.0" encoding="UTF-8?>
<TEI>
<teiheader xml:lang="fr">>
<titleStmt>
        <title> Le suicide d'une lycéenne transgenre à Lille interpelle la communauté éducative 19-12-2020 </title>
          <author>Laurie Moniez </author>
           <editor> satou</editor>
           </titleStmt>
             <publicationStmt>
              <availability statut= "limité">
                 </availability>
                 </publicationStmt>
                  </titleStmt>
                     </teiHeader>
                       <text>
                         <body>
                    </teiheader>
                      <p>Fouad, 17 ans, a mis fin à ses jours mardi. Deux semaines avant, elle s'était opposée à la
en jupe. Personne n'avait perçu sa souffrance ces dernières semaines.</p>
                   <p>A quelques jours des vacances scolaires de Noël, les élèves et les enseignants du lycée Fénelo
15 décembre, Fouad, élève transgenre de terminale de 17 ans, s'est donné la mort dans la chambre de son foyer, à Lam
en charge par l'Aide sociale à l'enfance (ASE). Le parquet de Lille a indiqué que des investigations médico-légales
                  <p>Depuis que ce décès est connu, l'émotion est allée crescendo, certains pointant rapidement la re
semaines auparavant, mercredi 2 décembre, Fouad avait eu un vif échange avec la direction de son établissement scola
habillée en jupe. Depuis jeudi, une vidéo, alors tournée par l'adolescente, circule sur les réseaux sociaux. On y en
conversation, la conseillère principale d'éducation expliquer : « Je comprends ton envie d'être toi-même. Ça, je le
justement, c'est fait pour t'accompagner au mieux. C'est ça que tu ne comprends pas ! Parce qu'encore une fois, il y
pas les mêmes. » « Mais c'est eux qu'il faut éduquer », répond Fouad.</p>
                    <p>Dans un courrier adressé au directeur académique des services de l'éducation nationale, la m
demandé que « toute la lumière soit faite sur ce drame qui a touché une lycéenne transgenre lilloise », précisant qu
combattue partout dans notre société ».</p>
                   <p>L'affaire est remontée, vendredi, jusqu'au ministre de l'éducation nationale. Jean-Michel Blan
Marly-la-Ville (Val-d'Oise), a défendu son action : « Nous avons fait énormément sur la lutte contre le harcèlement,
contre les élèves LGBT. » Il a néanmoins reconnu qu'il fallait « que nous réussissions beaucoup mieux à lutter contr
                    </text>
                    </TEI>
```

Correction du TD 2

- Trouvez les 3 erreurs

```
<?xml version="1.0" encoding="UTF-8?>
<TEI>
<teiheader xml:lang="fr">>
<titleStmt>
        <title> Le suicide d'une lycéenne transgenre à Lille interpelle la communauté éducative 19-12-2020 </title>
          <author>Laurie Moniez </author>
           <editor> satou</editor>
           </titleStmt>
             <publicationStmt>
              <availability statut= "limité">
                 </availability>
                 </publicationStmt>
                  </titleStmt>
                     </teiheader>
                      <text>
                        <body>
                    </teiheader>
                      <p>Fouad, 17 ans, a mis fin à ses jours mardi. Deux semaines avant, elle s'était opposée à la
en jupe. Personne n'avait perçu sa souffrance ces dernières semaines.</p>
                    <p>A quelques jours des vacances scolaires de Noël, les élèves et les enseignants du lycée Fénelo
15 décembre, Fouad, élève transgenre de terminale de 17 ans, s'est donné la mort dans la chambre de son foyer, à Lam
en charge par l'Aide sociale à l'enfance (ASE). Le parquet de Lille a indiqué que des investigations médico-légales
                  <p>Depuis que ce décès est connu, l'émotion est allée crescendo, certains pointant rapidement la re
semaines auparavant, mercredi 2 décembre, Fouad avait eu un vif échange avec la direction de son établissement scola
habillée en jupe. Depuis jeudi, une vidéo, alors tournée par l'adolescente, circule sur les réseaux sociaux. On y en
conversation, la conseillère principale d'éducation expliquer : « Je comprends ton envie d'être toi-même. Ça, je le
justement, c'est fait pour t'accompagner au mieux. C'est ça que tu ne comprends pas ! Parce qu'encore une fois, il y
pas les mêmes. » « Mais c'est eux qu'il faut éduquer », répond Fouad.</p>
                     <p>Dans un courrier adressé au directeur académique des services de l'éducation nationale, la m
demandé que « toute la lumière soit faite sur ce drame qui a touché une lycéenne transgenre lilloise », précisant qu
combattue partout dans notre société ».</p>
                    <p>L'affaire est remontée, vendredi, jusqu'au ministre de l'éducation nationale. Jean-Michel Blan
Marly-la-Ville (Val-d'Oise), a défendu son action : « Nous avons fait énormément sur la lutte contre le harcèlement,
contre les élèves LGBT. » Il a néanmoins reconnu qu'il fallait « que nous réussissions beaucoup mieux à lutter contr
                     </text>
                    </TEI>
```

Correction du TD 2

- XML correct, mais en TEI, un seul teiHeader par fichier.

```xml
<TEI>
  <teiHeader>
    <titleStmt>
      <title> Covid-19 </title>
      <author> L'Agora </author>
      <editor> Ruoxuan Li </editor>
    </titleStmt>
  <publicationStmt>
  <availability>
    <licence> © L'Agora </licence>
  </availability>
  </publicationStmt>
  </teiHeader>
  <text>
    <body>
      <p>La COVID-19 est la maladie causée par le coronavirus SARS-CoV-2, un nouveau virus détecté pour la première fois en décembre 2019. Le séquençage génétique du virus semble indiquer qu'il s'agit d'un bétacoronavirus étroitement lié au virus du SRAS. Les coronavirus forment une grande famille de virus. Certains d'entre eux peuvent infecter les animaux et d'autres, les humains.</p>
    </body>
  </text>
  <teiHeader>
    <titleStmt>
      <title> Madonna et son confinement - Covid-19 </title>
      <author> Madonna-world </author>
      <editor> Ruoxuan Li </editor>
    </titleStmt>
  <publicationStmt>
  <availability>
    <licence> © Madonna-world </licence>
  </availability>
  </publicationStmt>
  </teiHeader>
  <text>
    <body>
      <p>Le coronavirus poursuit sa propagation en Europe et en Amérique du Nord. Face à cette menace, les célébrités se plient aux mêmes règles de confinement que le commun des mortels. Madonna en témoigne depuis plusieurs jours sur les réseaux sociaux, en documentant son quotidien par le biais de vidéos sur Instagram. Mais l'isolement d'une reine de la pop multimillionnaire semble très éloigné de celui des internautes, et ces derniers font part de leur agacement.</p>
    </body>
  </text>
</TEI>
```

Correction du TD 2

- Le XML est correct, mais quels points de TEI on pourrait améliorer ? (→ sachant qu'on va donner le corpus à un ordinateur pour faire des stats)

```
  <text xml:lang='en'>
    <body>
      <p>1. INTRODUCTION. <lb/> In this paper we investigate irrealis mood in the Nyulnyulan
languages of the Dampier Land peninsula and adjoining parts of the Kimberley region
in the far northwest of Western Australia. As in many non-Pama-Nyungan languages,
in all Nyulnyulan languages with the possible exception of Yawuru -tense distinctions
are made in the irrealis. Correspondingly, irrealis categories are marked by the
combination of prefixes indicating mood and a sufix indicating tense (see Verstraete
2005, 2006; Lazard 2006). This is illustrated by examples (1) and (2), where the irrealis
prefix -la- cooccurs with the past tense sufix -na, as well as with a marked form of
the nominative pronominal prefix.</p>
      <p>(1) Marlu wi-la-rr-arli-na kinya mayi, marlu. warrwa <lb/>not 3:nom:irr-irr-aug-eat-pst
this food not <lb/> 'They didn't eat it.'(McGregor 1994:25) <lb/> </p>
      <p> Milarra oo-la-rli-na-ngayoo<lb/> almost 3:min:nom:fut/irr-irr-eat-pst-1:min:acc snake-
erg<lb/> 'The snake almost bit me.'' (Aklif 1991a)<lb/> </p>
    </body>
  </text>
```

Quelques balises XML-TEI

```
<p>On the one hand the <title>Nibelungenlied</title>
 is associated with the new rise of romance of twelfth-century France, the <foreign>romans
   d'antiquité</foreign>, the romances of Chrétien de Troyes, ...</p>
```

4.2 Headings and Closings

Every div may have a title or heading at its start, and (less commonly) a trailer such as 'End of Chapter 1' at its end. The following elements may be used to transcribe them:

<head> (heading) contains any type of heading, for example the title of a section, or the heading of a list, glossary, manuscript description, etc.
<trailer> contains a closing title or footer appearing at the end of a division of a text.

Some other elements which may be necessary at the beginning or ending of text divisions are discussed below in section 18.1.2 Prefatory Matter.

Whether or not headings and trailers are included in a transcription is a matter for the individual transcriber to decide. Where a heading is completely regular (for example 'Chapter 1') or may be automatically constructed from attribute values (e.g. `<div type="chapter" n="1">`), it may be omitted; where it contains otherwise unrecoverable text it should always be included. For example, the start of Hardy's *Under the Greenwood Tree* might be encoded as follows:

```
<div xml:id="UGT1" n="Winter" type="Part">
 <div xml:id="UGT11" n="1" type="Chapter">
  <head>Mellstock-Lane</head>
  <p>To dwellers in a wood almost every species of tree ... </p>
 </div>
</div>
```

Quelques balises XML-TEI

6.2 Quotations and Related Features

Like changes of typeface, quotation marks are conventionally used to denote several different features within a text, of which the most frequent is quotation. When possible, we recommend that the underlying feature be tagged, rather than the simple fact that quotation marks appear in the text, using the following elements:

<q> (quoted) contains material which is distinguished from the surrounding text using quotation marks or a similar method, for any one of a variety of reasons including, but not limited to: direct speech or thought, technical terms or jargon, authorial distance, quotations from elsewhere, and passages that are mentioned but not used.
<mentioned> marks words or phrases mentioned, not used.
<soCalled> contains a word or phrase for which the author or narrator indicates a disclaiming of responsibility, for example by the use of scare quotes or italics.
<gloss> identifies a phrase or word used to provide a gloss or definition for some other word or phrase.

Here is a simple example of a quotation:

```
<p>Few dictionary makers are likely to forget Dr. Johnson's description of the
 lexicographer as <q>a harmless drudge.</q>
</p>
```

Quelques balises XML-TEI

6.3 Foreign Words or Expressions

Words or phrases which are not in the main language of the texts may be tagged as such in one of two ways. If the word or phrase is already tagged for some reason, the element indicated should bear a value for the global @xml:lang attribute indicating the language used. Where there is no applicable element, the element foreign may be used, again using the @xml:lang attribute. For example:

```
<p>John has real <foreign xml:lang="fr">savoir-faire</foreign>.</p>
<p>Have you read <title xml:lang="de">Die
   Dreigroschenoper</title>?</p>
<p>
 <mentioned xml:lang="fr">Savoir-faire</mentioned> is French
 for know-how.
</p>
<p>The court issued a writ of <term xml:lang="la">mandamus</term>.</p>
```

As these examples show, the foreign element should not be used to tag foreign words if some other more specific element such as title, mentioned, or term applies. The global @xml:lang attribute may be attached to any element to show that it uses some other language than that of the surrounding text.

Quelques balises XML-TEI

16.1 Additional Elements for Technical Documents

The following elements may be used to mark particular features of technical documents:

<eg> (example) contains any kind of illustrative example.
<code> contains literal code from some formal language such as a programming language.
<ident> (identifier) contains an identifier or name for an object of some kind in a formal language. ident is used for tokens such as variable names, class names, type names, function names etc. in formal programming languages.
<gi> (element name) contains the name (generic identifier) of an element.
<att> (attribute) contains the name of an attribute appearing within running text.
<formula> contains a mathematical or other formula.
<val> (value) contains a single attribute value.

The following example shows how these elements might be used to encode a passage from a tutorial introducing the Fortran programming language:

```
<p>It is traditional to introduce a language with a
 program like the following: <eg xml:space="preserve"> CHAR*12 GRTG
          GRTG = 'HELLO WORLD'
          PRINT *, GRTG
          END
        </eg>
</p>
<p>This simple example first declares a variable <ident>GRTG</ident>, in the line
<code>CHAR*12 GRTG</code>, which identifies <ident>GRTG</ident> as consisting of 12 bytes
 of type <ident>CHAR</ident>. To this variable, the value <val>HELLO WORLD</val> is then
 assigned.</p>
```

Quelques balises XML-TEI

4.1 Text Division Elements

The body of a prose text may be just a series of paragraphs, or these paragraphs may be grouped together into chapters, sections, subsections, etc. Each paragraph is tagged using the p tag. The div element is used to represent any such grouping of paragraphs.

<p> (paragraph) marks paragraphs in prose.
** ** (text division) contains a subdivision of the front, body, or back of a text.

The @type attribute on the div element may be used to supply a conventional name for this category of text division, or otherwise distinguish them. Typical values might be ‘book’, ‘chapter’, ‘section’, ‘part’, ‘poem’, ‘song’, etc. For a given project, it will usually be advisable to define and adhere to a specific list of such values.

A div element may itself contain further, nested, divs, thus mimicking the traditional structure of a book, which can be decomposed hierarchically into units such as parts, containing chapters, containing sections, and so on. TEI texts in general conform to this simple hierarchic model.

Un autre exemple de XML-TEI

```xml
<text>
   <body>
      <div type="article">
         <head>
            <title type="Tetiere">83 VERDUN</title>
            <title type="SurTitre">Urbanisme</title>
            <title type="Titre">Quand la ville communique</title>
         </head>
         <div type="accroche">
            <p>L'opération de rénovation urbaine qui représente un investissement de 75 millions d'euros, incite la municipalité à communiquer et à informer le public sur ses choix.</p>
         </div>
         <div type="texte">
            <p>Arsène Lux a décidé. La grande barre de la Cité Verte sera détruite, vraisemblablement en 2011. Au total, dans le cadre de l'opération de rénovation urbaine (ORU), « 397 logements seront déconstruits, 262 reconstruits, 532 appartements réhabilités et 899 seront l'objet de travaux de résidentialisation consistant, par le traitement des halls d'entrée et des espaces privatifs en pied d'immeubles, à favoriser l'appropriation des lieux par les habitants. »</p>
            <div type="INTERTITRE">
               <head>Animations 3D</head>
               <p>Cette volonté de se rapprocher des habitants et de leurs préoccupations, du moins sur le plan de la communication, se traduira également par la création, courant 2010, d'un site internet consacré à l'ORU au sein duquel on pourra découvrir des animations 3D : « Ces animations permettront de voir comment les quartiers concernés évolueront et donc de mieux visualiser comment cela se passera », précise Coralie Batista, la chargée de communication de l'opération de rénovation urbaine. Des écrans de télévision présentant ces animations pourraient également être installés à l'entrée de la mairie.</p>
               <p>Enfin, pour être bien sûr que tout le monde soit informé du projet, des panneaux de quatre mètres de large pour trois de long seront installés aux entrées de ville et des quartiers concernés. Oyez, oyez, braves gens !</p>
            </div>
         </div>
      </div>
   </body>
</text>
```

Article de presse
Corpus de l'Est Républicain

Résumé

- Quelques balises XML-TEI à savoir utiliser
  - TEI, text, body
  - div, p
  - q, eg, foreign
  - head, title
  - lb
- Ne pas oublier l’attribut xm:lang="xxx" quand nécessaire (cf. documentation)
  - Codes ISO-639-2 ou ISO-639-3 pour la langue
  - Codes ISO-15924 pour le script [facultatif]
  - Codes ISO-3166 pour la région [facultatif]

Combinaisons code/script/région

sn
: Shona

zh-TW
: Taiwanese

zh-Hant-HK
: Chinese written in traditional script as used in Hong Kong

en-SL
: English as spoken in Sierra Leone

pl
: Polish

es-MX
: Spanish as spoken in Mexico

es-419
: Spanish as spoken in Latin America

Langue ≠ script

Légende

Code ISO 639-3 (langue)

Code ISO-15924 (script)

zho

注音符號 *Hant*

注音符号 *Hans*

ㄅㄆㄇㄈ *Hanb*

hbs

srpskohrvatski-hrvatskosrpski *Latn*

српскохрватски-хрватскосрпски *Cyril*

Bahasa Melayu *Latn*

بهاس ملايو *Arab*

zlm

Latn Türkçe

Arab الفبا

tur

Script ≠ transcription

Langue *japonais* (code *jpn*)

東京

Script *Han* (code *Han*)

Plein de transcriptions possibles

とうきょう

Script *Hiragana* (code *Hira*)

Tokyo

Tōkyō

Tôkyô

Tookyoo

Tohkyoh

Toukyou

toːkʲoː

Токио

Script *Cyrillique* (code *Cyril*)

La norme ISO-15924 ne distingue pas les méthodes de transcription latine (romanisation) : toutes ces graphies ont donc le même script : *Latin* (code *Latn*).

Remarque : dans la norme ISO-15924, même l'API est considéré comme une graphie latine.

Pratique

- Modifiez l’un de vos documents XML pour inclure un exemplaire de chacune des balises à savoir utiliser (diapo 14).
  - Pour l’exercice, vous avez le droit d’inventer du contenu, ou d’aller chercher un paragraphe dans un autre texte ! Mais la balise doit être employée de manière logique.
  - N’oubliez pas les codes langue, au moins sur les balises *text* et *foreign*.