Documentation

Stage IA et encodage TEI de gros volumes de données textuelles

ICAR (ENS de Lyon) & LIRIS (INSA)

Encadrants : Denis Vigier (ICAR) et Ludovic Moncla (LIRIS)

Maritza Beatriz García Rodríguez
M2 Humanités Numériques
Université Lumière Lyon 2 (ICOM)

10 février 2026-24 juillet 2026

Table des matières

I Motivation du choix de standard pour l'encodage en TEI : limites et possibilités de trois standards

Cette première section présente la réflexion autour de la comparaison entre les trois standards TEI dans lesquels notre schéma d'encodage aurait pu s'inscrire : la TEI P5, le module *dictionaries* de la TEI P5 et la TEI Lex-0.

TEI P5

Apport :

L'apport le plus important de la TEI P5 est la possibilité d'utiliser l'élément <p> pour rendre compte de la division en paragraphes des articles du dictionnaire.

Limites importantes :

À l'intérieur de l'élément <p>, la TEI P5 n'admet pas les éléments spécifiques à la structuration de dictionnaires : <form>, <gram>, <sens>, <def>, <cit>, etc. Inversement, la TEI Dictionnaires et la TEI Lex-0, qui prescrivent l'utilisation de ces éléments, n'admettent pas la balise <p> avec ces dernières.

Utilisation dans d'autres projets :

Une solution à cette incompatibilité a été apportée par Alice Brenon dans le projet DISCO-LGE, portant sur l'encodage de *La Grande Encyclopédie*. En raison de l'inadaptabilité du module *dictionaires* aux besoins spécifiques des encyclopédies, détaillés par l'autrice dans « Encoding the Specificities of Encyclopedias[1] », elle a opté par l'utilisation du *core module* de la TEI P5, avec un balisage basé sur les éléments , <head> et <p> et l'utilisation de l'attribut @type, contenant une valeur qui rappelle le nom des éléments utilisées dans la spécification de la TEI pour l'encodage de dictionnaires ("sense", "header", etc.). Cet emploi de l'attribut @type s'inscrit dans le sillage d'un article qui montrait depuis 1995 les limites du chapitre *dictionaries* de la TEI. Nancy Ide et Jean Véronis écrivaient dans « Encoding Dictionaries » :

[1] Alice Brenon, « Encoding the Specificities of Encyclopedias », in *Structuring Lexical Data and Digitising Dictionaries*, eds. Javier Martín Arista and Ana Elvira Ojanguren López, Leinden and Boston, Brill, p. 36-62, 2025.

> The Dictionary Working Group recognized a parallel between dictionary divisions and the nested document divisions (volume, chapter, section, sub-section, etc.), which are treated in the TEI by providing a generic tag which nests recursively and takes a type attribute describing the division type, rather than providing specific « hard-coded » tags such as <chapter>, <section>, etc. (TEI P3, p. 2019)[2].

TEI P5 *dictionaries* au miroir de la TEI Lex-0

Apports de la TEI P5 *dictionaries*

Dans un article qui dessinait les lignes directrices de la TEI Lex-0, tout en expliquait sa nécessité dans le contexte de l'encodage de dictionnaires, Piotr Bański, Jack Bowers et Tomaz Erjavec écrivent :

> While the TEI has defined a number of specialized elements within **gramGrp**, TEI-Lex0 takes a more generic route in this respect, for reasons of uniformity and sustainability. The former criterion makes it possible to simplify the processing tools and unify the representation. The latter makes the format more resilient to future modifications of the TEI: if, for example, at some point in the future, the TEI defines an element **voice** for grammatical voice, the TEI-Lex0 guidelines will not need to be adjusted – all that will be necessary will be another mapping between, say, **<voice>active</voice>** in the target dictionary and **<gram type="voice">active</gram>** in TEI-Lex0[3].

Si cela n'était pas encore le cas en 2017, année où a eu lieu la conférence dont est issu l'article, actuellement on retrouve dans les guidelines du chapitre 10 de la TEI P5 le choix entre deux méthodes pour encoder l'information grammaticale. L'une selon laquelle l'information grammaticale adopte la forme d'un élément (<pos>v</pos>), et l'autre qui apparaît en tant que valeur de l'attribut @type contenu dans l'élément <gram> (<gram type="pos">v</gram>)[4]. Cette introduction, postérieure à 2017, semble être une réponse de la TEI Dictionnaires aux demandes de la communauté qui développerait la Lex-0.

Du fait de compter sur un catalogue de balises plus large que celui de la Lex-0, la TEI *dictionaries* permet de pallier l'impossibilité de la Lex-0 de rendre compte des paragraphes tout

[2] Nancy Ide, Jean Véronis, « Encoding Dicitonaries », *Computers and the Humanities*, n° 29, p. 172, 1995.
[3] Piotr Bański, Jack Bowers, Tomaz Erjavec, « TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms », *Electronic Lexicography in the 21st Century: Proceedings of ELex 2017 Conference*, Sept. 2017, Leiden, Netherlands, 2017, p. 4. En ligne sur : hal-01757108.
[4] Les deux exemples sont disponibles sur : https://tei-c.org/release/doc/tei-p5-doc/fr/html/DI.html#DITPGR.

en respectant la structure caractéristique des dictionnaires[5]. Pour cela, j'ai proposé l'utilisation de l'élément <milestone/>, dont l'attribut @unit prend la valeur "paragraph" :

```
<entry type="relatedEntry">
   <form type="lemmaGrp"><form type="lemma"><orth>Abandonner</orth></form>
      <metamark function="lemmaDelimiter">,</metamark>
   </form>
   <sense n="1" xml:id="DUFLT1743.abandonner.VII.1">
      <def>signifie aussi simplement, Quitter, laisser, renoncer à quelque profession ou à quelque
personne.</def>
                <!--[…]-->
      <milestone unit="paragraph" rendition="#leftShift"/>
      <def>Dans ce sens il se dit de la résignation que nous faisons à Dieu de nous-mêmes &amp;amp; de tout ce
         qui nous touche.</def> <!--[…]-->
   </sense> <!--[…]-->
</entry>
```

Figure 1. Proposition d'encodage TEI du début de l'article ABANDONNER, DUFLT 1743, tome 1, p. 11-12.

La structure autofermante de la balise <milestone unit="paragraph" rendition="#leftShift"/> permet d'éviter les problèmes d'*overlapping* que nous aurions retrouvés avec des éléments comme <p>, admettant qu'il est valide dans le module *dictionaries* – ce qui n'est pas le cas –, ou avec d'autres éléments comme <seg> accompagné d'un attribut @type="paragraph". Bien que l'attribut @unit soit obligatoire dans la TEI P5, le choix de sa valeur reste libre au-delà des propositions qui se déploient sur Oxygen :

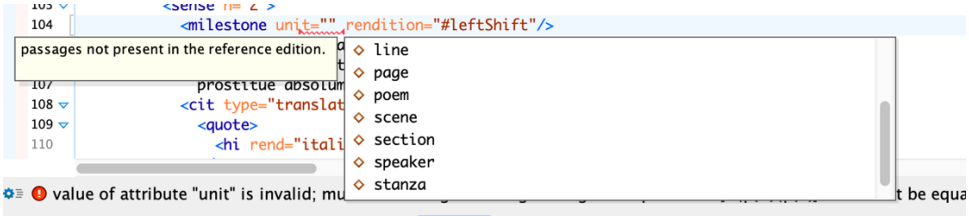


Figure 2. Aperçu de la liste déroulante des valeurs proposées pour l'attribut @unit.

Le deuxième attribut, @rendition, permet de rendre compte des spécificités de la mise ne page (un alinéa négatif dans l'édition de 1743). Aussi, puisque les entrées et sous-entrées supposent un nouveau paragraphe, la présence de <milestone/> reste superflue devant les vedettes et sous-vedettes.

D'autres éléments admis par le module *dictionaries* de la P5 et qui sont exclus de la Lex-0 sont <oRef> (pour indiquer les étymons au sein des étymologies) et <foreign> (pour indiquer les expressions en langue étrangère qui ne relèvent pas de traductions ni d'étymons).

[5] En effet, la TEI Dictionnaires et la Lex-0 admettent la balise <p>, mais celle-ci ne peut contenir ni être contenus par les balise spécifiques aux dictionnaires : <sense>, <def>, <form>, etc.

Ces deux éléments avaient été utilisés dans le projet BASNUM[6], dont nous nous inspirons dans leur utilisation.

Un autre avantage, de la P5 à l'égard de la Lex-0 concerne la possibilité de retrouver l'élément <hi> à l'intérieur de <gloss>, introduisant des remarque typographiques dans les glose. Dans l'entrée CHARPENTIER : <gloss>On l'a nommé autrefois <hi rend="italic">Chapuis</hi>.</gloss>.

Alors que la TEI Dictionnaires admet du texte à l'intérieur de l'élément <sense>, la Lex-0 nécessite d'autres éléments intermédiaires, comme <usg>, <def> ou <gloss>, pour introduire du texte dans <sense>.

Enfin, la Lex-0 restreint le choix parmi les valeurs de l'attribut @type dans les éléments <cit> et <xr> :

- <cit type=" "></cit> : Les valeurs de /cit[@type] que l'on impose sont « cognate », « cognateSet », « etymon », « example », « translation » et « translationEquivalent ».
- <xr type=" "></xr> : Les valeurs de /xr[@type] que l'on impose sont « antonymy », « hyperonymy », « hyponymy », « meronymy », « related » et « synonymy ».

Limites partagés par la TEI P5 *dictionaries* et la Lex-0

Les limites du chapitre 10 de la TEI P5 rejoignent celles de la TEI Lex-0. On ne peut pas introduire du texte contenu directement dans l'élément <cit> ni par <xr> (renvoi). En conséquence, j'ai utilisé le contournement proposé par la Lex-0, c'est-à-dire les éléments <quote> et <bibl> pour contenir le texte dans <cit>, et les éléments <lbl> et <ref> comme intermédiaires du texte dans <xr>.

Choix de schéma

Le choix de la TEI *dictionaries* a été motivé principalement par les limites dans le catalogue d'éléments disponibles par la Lex-0, notamment <milestone>, nous permettant de traiter les débuts de paragraphe, mais aussi <oRef> et <foreign> qui rendent compte de notions qu'on ne saurait pas traiter autrement dans la Lex-0.

L'intérêt de la TEI *dictionaries* et de la Lex-0 face au *core module* de la TEI P5 réside dans l'inscription d'un cadre disponible pour la communauté travaillant sur des dictionnaires,

[6] Voir notamment sur <oRef> l'article de Ioana Galleron, « L'entropie informationnelle dans le *Dictionnaire Universel* (1701) », dans Robert Alessi, Marcello Vitali-Rosati (dir.), *Les éditions critiques numériques* (édition augmentée), Les Presses de l'Université de Montréal, Montréal, 2023. En ligne sur : https://parcoursnumeriques-pum.ca/12-editionscritiques/chapitre8.html.

malgré l'hétérogénéité de cette catégorie d'objet textuel, dont les standards TEI rendent compte avec difficulté et pas tout le temps de manière satisfaisante, particulièrement lorsqu'il s'agit de dictionnaires historiques. Avec la TEI *dictionaries*, nous nous inscrivons notamment dans le sillage de BASNUM.

II Description des éléments TEI mobilisés lors de l'encodage

Dans un premier temps, vous trouverez la liste d'éléments TEI qui ont eu une place lors de l'encodage de la vérité de terrain, leur spécificité et fonctionnement dans la globalité de l'encodage. Dans un deuxième temps, seront présentés des les éléments qui n'ont pas vu le jour lors de l'encodage mais qui restent dans notre catalogue comme étant pertinents à la réflexion autour du *Trévoux*.

Liste d'éléments utilisés

<milestone unit="paragraph" rendition="#leftShift"/>

Indication de début de paragraphe. La balise autofermante <milestone/> est utilisée pour « marque[r] un point permettant délimiter les sections d'un texte selon un autre système que les éléments de structure ; une balise de ce type marque une frontière[7]. » Bien que le paragraphe soit un élément de structure avec une balise associée, l'incompatibilité de la balise <p> avec la spécificité de la structure des dictionnaires, voir *supra*, nous oblige à considérer le paragraphe comme une unité de <milestone/>.

L'attribut @unit peut contenir invariablement la valeur "paragraph", tandis que la valeur de @rendition peut osciller, par exemple, entre « #leftShift » et « #rightShift » selon l'alinéa négatif (#leftShift) ou positif (#rightShift) identifié dans l'océrisation.

Comme indiqué précédemment, la présence de <milestone/> reste superflue devant les vedettes et sous-vedettes, dont l'encodage signale un début de paragraphe.

<entry>

Élément qui « contient une entrée structurée de dictionnaire[8] ». Parmi les attributs qu'elle peut adopter :

- **@xml:id** (identifiant unique),
- **@type** (parmi les valeurs suggérées : "mainEntry" pour les entrées principales introduites par la vedette, "relatedEntry" pour les sous-entrées),
- **@xml:lang** (identifie la langue de l'objet auquel il est associé, son contenu sera expliqué *infra*).

---

7 Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-milestone.html. Dernière consultation le 22 juillet 2026.

8 Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-entry.html. Dernière consultation le 22 juillet 2026.

Marque : Le marquage qui permet d'identifier le commencement d'une entrée est la vedette en capitales et la sous-vedette en petites capitales pour les sous-entrées, elles apparaissent au début de la ligne, généralement avec un alinéa négatif, positif ou une manicule, indiquant le début d'un paragraphe. La sous-vedette peut avoir la même forme orthographique que la vedette ou être formée par une flexion de celle-ci. L'entrée est composée par la vedette ou la sous-vedette et le texte qui la suit jusqu'à la prochaine vedette ou sous-vedette, qui fera partie de l'entrée suivante.

S'il s'agit de l'entrée principale (type="mainEntry"), la vedette qui suit démarque la fin de l'entrée dont il est question. L'entrée principale peut contenir d'autres sous-entrées. S'il s'agit d'une sous-entrée (type="relatedEntry"), elle sera délimitée par la sous-vedette suivante ou par la vedette de l'entrée qui suit. Les sous-entrées sont contenues par une entrée principale. Aussi, j'opte pour un encodage dans lequel l'entrée principale englobe les sous-entrées :

```
<entry type="mainEntry">
    <entry type="relatedEntry"></entry>
    <entry type="relatedEntry"></entry>
</entry>
```

Figure 3. Structure de l'imbrication des entrées et des sous-entrées.

<form>

Élément qui « regroupe toutes les informations relatives à la morphologie et à la prononciation d'une entrée[9] ». Il peut contenir les attributs :

- **@type** permet d'identifier l'ensemble d'information qui entoure le lemma ("lemmaGrp"), le lemme ("lemma"), les formes fléchies ("inflected"), les synonymes ("synonym"), les formes composées ("compound") et les variantes ("variant").
- **@subtype** pour définir un sous-ensemble de @type lorsque celui-ci indique une variante ("variant"). Parmi les trois @subtype des variantes : orthographiques ("orthographic"), diachroniques ("diachronic") et diatopiques ("diatopic").

Marque : Le lemme ou la flexion contenus par l'élément <form> apparaissent généralement en capitales (vedette) ou en petites capitales (sous-vedettes), si l'OCR le

[9] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-form.html. Dernière consultation le 22 juillet 2026.

permet. Bien que l'on retrouve la balise à l'intérieur de <cit> ou de <sense>, un balisage de surface pourrait se limiter aux <form> autour des vedettes et des sous-vedettes en début d'entrée et comme enfant direct de <entry>.

Tandis que la forme dans laquelle les mots-vedettes et sous-vedettes apparaissent est généralement celle du lemme (<form type="lemma">), la balise <form> peut aussi englober les flexions grammaticales du terme (<form type="inflected">). Celles-ci peuvent adopter, par exemple, la forme de la terminaison du participe « ée » (par ex., dans la sous-entrée « Abandonné, ée »).

<form type="lemmaGrp"> recouvre toutes les informations concernant la forme orthographique et phonétique de la vedette ou sous-vedette avant l'ouverture du sens (<sense>). Cet élément contient non seulement d'autres <form> comme enfants directs, mais aussi des <metamark>, des <pc> et des <gramGrp>. À l'intérieur des <form> qui contiennent le lemme, des flexions ou des variantes, on retrouve des éléments <orth> et <pron>.

```
<entry type="mainEntry">
        <form type="lemmaGrp">
            <form type="lemma"><orth>PRÉSIDENTAL</orth></form>
            <pc>,</pc>
            <form type="inflected"><orth>ALE</orth></form>
            <metamark function="lemmaDelimiter">.</metamark>
            <gramGrp>
                <gram type="pos" expand="adjectif" norm="ADJ">adj.</gram>
            </gramGrp>
        </form>
```

Figure 4. Proposition d'encodage pour l'entrée PRÉSIDENTAL, DUFLT 1743, tome 5, p. 489.

<orth> et <pron>

Éléments qui englobent la forme orthographique (<orth>) et phonétique (<pron>) du mot-vedette[10].

Marque : Dans un balisage de surface, on trouve <orth> principalement à l'intérieur de <form>, contenant la forme orthographique de la vedette ou de la sous-vedette, ainsi que leur forme fléchie ou leur variante.

[10] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». Pour <orth> : https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-orth.html. Pour <pron> : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-pron.html. Dernière consultation le 22 juillet 2026.

<metamark>

Élément qui « contient ou décrit tout type de signal graphique ou écrit dans un document, dont la fonction est de déterminer comment celui-ci doit être lu plutôt que de faire partie intégrante du contenu même du document[11] ».

Dans un article de 2024 portant sur la numérisation et l'encodage du *Diccionario de Morais*, Ana Salgado, *et al*. proposent l'utilisation de cet élément comme délimiteur dont la fonction est indiquée par l'attribut **@function**[12]. Parmi les différentes fonctions proposées par les auteurs, nous retiendrons celle de "lemmaDelimiter".

Marque : Dans notre balisage de surface, nous l'utiliserons pour baliser le signe de ponctuation qui suit la vedette ou la sous-vedette. Il s'agit en effet d'une virgule ou d'un point. Lorsque la vedette est composée de plusieurs mots, par exemple, un lemme suivi de la flexion, le <metamark> apparaît après l'inflexion.

<gram>

Cet élément « contient de l'information grammaticale relative à un terme, un mot ou une forme dans une entrée de dictionnaire ou dans un fichier de données terminologiques[13]. »

Marque : L'information grammaticale apparaît souvent sous forme d'abréviation. Voici une liste pas exhaustive : « adj. » pour adjectif, « s. » et « sub. » pour substantif, « v. » ou « verb. » pour verbe, « adv. » pour adverbe, « m. » pour masculin, « f. » pour féminin, « act. » pour la voix active, « part. » pour le participe, « pass. » pour la voix passive, etc.

Les attributs **@type**, **@expand** et **@norm** fournissent des informations supplémentaires sur ces abréviations. L'attribut **@type** peut adopter la valeurs "pos" (*part of speech*), lorsqu'il s'agit d'un substantif, adjectif, adverbe, verbe, pronom, article, préposition, conjonction et interjection ; la valeur "gen" lorsqu'il s'agit du genre grammatical, "num" pour le nombre, et la valeur "voice" pour la voix passive et active. Les valeurs de **@type** ne sont pas restreints aux listes proposées par le module *dictionaries*.

La valeur de **@expand** contient la forme développée des abréviations.

---

[11] Nous traduisons. Voir : « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-metamark.html. Dernière consultation le 22 juillet 2026.

[12] Ana Salgado, Laurent Romary, Rute Costa, *et al*., « Following Best Practices in a Retro-Digitized Dictionary Project », *International Journal of Humanities and Arts Computing*, 2024, n° 18, vol. 1, p. 132.

[13] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-gram.html. Dernière consultation le 22 juillet 2026.

La valeur de **@norm** fournit des formes normalisées de ces éléments. Nous proposons de suivre les **Universal Dependencies**, parmi lesquelles on trouve les **Universal POS tags** pour les *part of speech*, et les **Universal Features**, proposant des formes normalisées pour les propriétés lexicales et grammaticales qui ne sont pas couvertes par les POS tags[14].

<gramGrp>

Élément qui « regroupe de l'information morphosyntaxique sur un item lexical[15] ».

Marque : Dans un balisage de surface, on retrouve la balise <gramGrp> englobant les balises <gram> qui contiennent des éléments offrant une information morphosyntaxique.

```
<entry type="mainEntry">
        <form type="lemmaGrp">
            <form type="lemma"><orth>PRÉSIDIAL</orth></form>
            <pc>,</pc>
            <form type="inflected"><orth>ALE</orth></form>
            <metamark function="lemmaDelimiter">.</metamark>
            <gramGrp>
                <gram type="pos" expand="substantif" norm="NOUN">s.</gram>
                <gram type="gen" expand="masculin" norm="Masc">m.</gram> &amp;
                <gram type="pos" expand="adjectif" norm="ADJ">adj.</gram>
                <gram type="gen" expand="masculin" norm="Masc">m.</gram> &amp;
                <gram type="gen" expand="feminin" norm="Fem">f.</gram>
            </gramGrp>
        </form>
```

Figure 5. Proposition d'encodage pour l'entrée PRÉSIDIAL, DUFLT 1743, tome 5, p. 490.

<sense>

Élément qui « regroupe toutes les informations relatives à un sens d'un mot dans une entrée de dictionnaire (définitions, exemples, etc.)[16] ».

Marque : Cet élément commence après la balise fermante </form> (dont la valeur de @type est "lemmaGrp") et se ferme à la fin du sens, qui peut coïncider avec un retour chariot et donc un nouveau paragraphe vers un autre sens, mais aussi avec la fin de l'entrée ou de la sous-entrée.

[14] Voir Universal POS tags : https://universaldependencies.org/u/pos/index.html ; et Universal Features : https://universaldependencies.org/u/feat/index.html. Dernière consultation le 22 juillet 2026.

[15] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-gramGrp.html. Dernière consultation le 22 juillet 2026.

[16] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-sense.html. Dernière consultation le 22 juillet 2026.

À l'intérieur de <sense> on retrouve les définitions (<def>) ; les marques d'usage (<usg>) ; les exemples (<cit type="example"><quote>) ; les traductions (<cit type="translation"><quote>) ; les remarques ou développements encyclopédiques (<seg>) ; les gloses qui commentent des éléments qui sont des enfants directs de <sense> ; les renvois vers des références externes (<xr type="cf">) et vers d'autres entrées (<xr type="related">), lorsque ces dernières ne sont pas le seul contenu de l'entrée ; et les étymologies (<etym>).

<usg>

Élément qui « contient, dans une entrée de dictionnaire, les informations sur son usage ». Nous le retrouverons comme enfant direct de <sense>.

Cet élément contient un attribut **@type** qui peut adopter la valeur "geo" pour les usages spécifiques à certains lieux (« en Bretagne »), "dom" pour les usages relevant de domaines (« en termes de fleuriste »), "time" pour des usages diachroniques (« vieux »), "style" pour des usages qui relèvent de styles de langage (« se dit proverbialement », « se dit figurément »). Son deuxième attribut, **@ana**, récupère le terme qui concerne la marque d'usage et pointe vers des futures taxonomies[17] qui listeraient les éléments géographiques, temporels, des domaines et de style. Par ex., <usg @type="dom" ana="#fleuriste"> En termes de fleuriste</usg>.

Marque : Cet élément est introduit par des périphrases telles que « en termes de », « se dit », etc. Les projets autour du *Basnage* et du *Morais* avaient déjà inclus dans l'attribut <usg> non seulement le terme qui correspond à l'usage, mais aussi la périphrase qui l'introduit. Dans le *Morais* : <usg type="domain">usa-se na<hi rend="italics">Architect.</hi></usg>[18].

<def>

Élément qui « contient le texte de la définition dans une entrée de dictionnaire[19] ». Bien qu'il puisse aussi être contenu dans <cit>, <entry>, <etym>, on le verra principalement dans <sense> avec la définition qui correspond au sens dont il est question.

[17] Pour un exemple de l'utilisation de l'élément <taxonomy>, voir Ana Salgado, Laurent Romary, Rute Costa, Toma Tasovac, Anas Fahad Khan, *et al*., art. cit., p. 140-143.
[18] *Ibidem*, p. 142.
[19] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-def.html. Dernière consultation le 22 juillet 2026.

<cit>

Élément qui contient une « citation provenant d'un autre document comprenant la référence bibliographique de sa source. Dans un dictionnaire il peut contenir un exemple avec au moins une occurrence du mot employé dans l'acception qui est décrite, ou une traduction du mot-clé, ou un exemple[20] ». Bien qu'il puisse être contenu par <def>, <entry>, <form>, <etym>, <gram>, <gramGrp>, <lang>, <lbl>, <orth>, <pron>, <usg>, <xr>, on l'utilisera surtout à l'intérieur de <sense>, <etym> et <seg>. Cet élément ne peut pas contenir du texte directement, c'est pourquoi il adopte l'élément <quote> comme intermédiaire entre <cit> et le texte.

Notre usage de <cit> se limite aux i) exemples apportés d'utilisation du terme, ii) à des exemples de discours rapportés et iii) aux traductions. C'est dans la valeur de l'attribut **@type** où la fonction de <cit> est définie.

D'une part, **<cit type="example">** contient des exemples avec des occurrences du mot-vedette mis en contexte – qui apparaît généralement en italiques – et des discours rapportés avec ou sans le terme dont il est question. On trouve souvent dans les exemples l'élément <bibl>, lorsque l'auteur est mentionné.

D'autre part, **<cit type="translation" xml:lang="la">** contient la traduction du terme sur lequel porte l'entrée (en latin pour le *Trévoux*) ou d'un autre terme ou expression dont on nous fait parvenir une traduction. Selon les Guidelines, la valeur de **@xml:lang** doit être conforme au BCP 47, utilisant les codes RFC 3066, qui suivent les recommandations ISO 639. Tandis que ISO 639-1 se concentre sur les codes à 2 lettres, ISO 639-2 a pour objet les codes à 3 lettres. Selon les recommandations du RFC 3066 : « When a language has both an ISO 639-1 2-character code and an ISO 639-2 3-character code, you MUST use the tag derived from the ISO 639-1 2-character code[21] ». Une liste des codes ISO 639-1 et des ISO 639-2 est disponible sur les Standards de la Library of Congress[22]. On y retrouve notamment "la" pour le latin, "grc" pour le grec, "fr" pour le français, "en" pour l'anglais, "de" pour l'allemand et "he" pour l'hébreu.

[20] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-cit.html. Dernière consultation le 22 juillet 2026.
[21] Voir sur : https://www.ietf.org/rfc/rfc3066.txt. Dernière consultation le 22 juillet 2026.
[22] Voir sur : https://www.loc.gov/standards/iso639-2/php/code_list.php. Dernière consultation le 22 juillet 2026.

L'usage de l'élément <cit> pour encoder les exemples et les traductions a été documenté par Bowers et Romary (2017)[23] ; et son usage pour les traductions a été repris par Ioana Galleron (2023) dans sa proposition d'encodage pour le projet BasNum[24] :

```
<note>Les Espagnols <lang xml:lang="es"/> disent, <foreign xml:id="idiom1" xml:lang="es">Morir sin fabla</foreign> ou <foreign xml:id="idiom2">fabuls</foreign> ; pour dire, <cit type="translation" corresp="#idiom1 #idiom2"><quote>Mourir intestat</quote></cit><pc>.</pc></note>
```

Figure 6. Encodage de l'entrée FABLE proposé par I. Galleron, dans le *Dictionnaire Universel* (1701).

<quote>

Élément qui « contient une expression ou un passage que le narrateur ou l'auteur attribue à une origine extérieure au texte[25] ». Nous le retrouvons notamment contenu dans <cit>. Il peut contenir <lang>, <oRef> et <pRef> (ce dernier élément n'a pas encore été mobilisé dans notre encodage).

<xr> et <ref>

<xr> : élément qui contient un renvoi, « une expression, une phrase ou une icône qui invite le lecteur à se référer à un autre endroit, dans le même texte ou dans un autre texte[26] ».

<ref> : élément qui « définit une référence vers un autre emplacement, la référence étant éventuellement modifiée ou complétée par un texte ou un commentaire[27] ».

Le renvoi (**<xr>)** est introduit par « Voyez », « V. » ou « On dit aussi », expression contenue dans un élément <lbl> et suivi de la référence (<ref>). Le renvoi peut être interne (vers un autre endroit dans le même texte, souvent le nom d'une autre entrée du dictionnaire) ou externe (vers un autre texte), ce qui sera défini par la valeur de l'attribut **@type**. Lorsque le renvoi est interne on retrouve : <xr type="related"> et lorsqu'il s'agit d'un renvoi externe : <xr type="cf">. Les renvois peuvent être contenus par <entry>, <etym> et <sense>, mais jamais par <seg> – cela forcera parfois la fermeture de <seg> pour ouvrir un <xr>. Lorsqu'une entrée

[23] Jack Bowers et Laurent Romary, « Deep encoding of etymological information in TEI », *Journal of the Text Encoding Initiative*, n° 10, 2017, p. 3. Les auteurs écrivent : « A variety of further descriptors are available in <sense> to provide such information as a definition (<def>), examples or translations (<cit>), various grammatical (<gramGrp>) or usage (<usg>) constraints, and of course etymological information (<etym>) ».

[24] Ioana Galleron, art. cit.

[25] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-quote.html. Dernière consultation le 22 juillet 2026.

[26] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-xr.html. Dernière consultation le 22 juillet 2026.

[27] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-ref.html. Dernière consultation le 22 juillet 2026.

est composée uniquement par un renvoi, celui-ci se retrouve comme enfant direct de <entry>, mais si l'entrée est accompagnée par d'autres éléments qui invitent à ouvrir un <sense> comme la définition, l'exemple, la traduction, etc., le renvoi s'inscrit à l'intérieur de <sense>.

Quant aux **références internes**, elles ont l'attribut **@type** avec la valeur <ref type="entry">. Certaines références, notamment les externes, peuvent apparaître sans renvoi. Bien que cela n'ait pas été inclus dans notre encodage, des attributs comme **@target** (« précise la cible de la référence en donnant une ou plusieurs références URI[28] ») ou **@corresp** (« pointe vers des éléments qui ont une correspondance avec l'élément en question[29] ») pourraient être intéressants ultérieurement pour renvoyer vers un identifiant unique ou des liens externes.

<lbl>

Élément contenant une « étiquette pour la forme d'un mot, pour un exemple, pour une traduction, ou pour tout autre type d'information, par exemple 'abréviation pour', 'contraction de', 'littéralement', 'approximativement', 'synonymes', etc.[30] ». On le retrouve à l'intérieur de <etym> contenant l'expression qui marque la couleur étymologique : « du » + langue, « ce mot est » + langue, « ce mot vient de » ; et à l'intérieur de <xr>, introduisant le renvoi : « Voyez ».

<bibl> et <author>

<bibl> : élément qui « contient une référence bibliographique faiblement structurée dans laquelle les sous-composants peuvent ou non être explicitement balisés[31] ». On le retrouve notamment à l'intérieur de <cit> et de <ref>. Bien que pour le moment nous ayons limité son contenu à l'élément <author>, cet élément peut également contenir <title>, avec le titre d'un ouvrage, et <biblScope>, avec des extensions de l'information bibliographique concernant des pages, volumes, tomes, etc.

<author> : élément qui « dans une référence bibliographique contient le nom de la (des) personne(s) physique(s) ou du collectif, auteur(s) d'une œuvre ; par exemple dans la même

[28] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.pointing.html. Dernière consultation le 22 juillet 2026.

[29] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-att.global.linking.html. Dernière consultation le 22 juillet 2026.

[30] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-lbl.html. Dernière consultation le 22 juillet 2026.

[31] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-bibl.html. Dernière consultation le 22 juillet 2026.

forme que celle utilisée par une référence bibliographique reconnue[32] ». Les auteurs apparaissent généralement en petites capitales et souvent abrégés. Dans notre proposition d'encodage concentrée sur l'élément <body>, <author> apparaît uniquement à l'intérieur de <bibl>.

<seg>

La TEI décrit cet élément comme contenant « une unité de texte quelconque de niveau 'segment'[33] ». Notre utilisation du marquage des segments se concentre sur du contenu à caractère encyclopédique, des digressions du lexicographe à caractère érudit, parfois anecdotique, et concernant l'histoire culturelle. Son manque sémantique peut être pallié par la valeur de l'attribut @type="encyclopedic".

Dans le sillage des commentaires du relecteur ou de la relectrice de l'article soumis à la revue *Corpus*, nous avons modifié notre choix de <note> (inspiré du BasNum), le remplaçant par <seg>. Cette modification a eu comme conséquence que <seg> doive être contenu par <sense> – et qu'il n'apparaisse jamais comme enfant direct de <entry> –, ce qui a motivé également l'emplacement des <etym> sous <sense>, en raison des entrées où une étymologie est suivie d'un développement encyclopédique, pour éviter de fermer le sens avant l'étymologie et le rouvrir avant <seg>.

On retrouve aussi l'élément <seg> à l'intérieur des étymologies (<etym>).

<etym>

Élément qui « contient les informations étymologiques de l'entrée[34] ». Elle est généralement introduite par des marques linguistiques englobées par <lbl> telles que « du » + langue, « ce mot est » + langue, « ce mot vient de » + étymon. Au sein de l'information étymologique on retrouve aussi la langue d'origine du mot (<lang>), l'étymon (<oRef type="etymon") ou mot d'origine, généralement en langue étrangère, et la signification en français de l'étymon présentée dans un élément <gloss>. On peut également y retrouver des développement encyclopédiques (<seg>) et des références (<ref>). Pour les raisons expliquées

[32] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-author.html. Dernière consultation le 22 juillet 2026.

[33] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-seg.html. Dernière consultation le 22 juillet 2026.

[34] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-etym.html. Dernière consultation le 22 juillet 2026.

supra, nous avons encodé <etym> à l'intérieur de <sense>. Si l'étymologie apparaît entre deux sens distincts, elle sera incluse dans le premier élément <sense>.

<oRef>

« [D]ans un dictionnaire, [cet élément] indique une référence à la forme orthographique de la vedette ou sous-vedette[35]. » <oRef> a été inclus dans notre encodage pour signaler la présence de l'étymon (@type="etymon") à l'intérieur de <etym>. Son emploi pour baliser les étymons est inspiré par l'article d'Ioana Galleron, « L'entropie informationnelle dans le Dictionnaire Universel (1701) »[36].

<lang>

Élément qui contient le « nom de la langue mentionnée [dans] des informations de nature linguistique (étymologique ou autre)[37] ». On le retrouve principalement à l'intérieur de <etym>, englobant la langue de l'étymon si celle-ci est énoncée, mais d'autres éléments peuvent le contenir : <def>, <seg>, <sense>, <usg>, <xr>, etc. Cette langue est standardisée grâce à l'attribut **@xml:lang**, dont la valeur dépend de l'ISO 639, mentionné *supra*.

<foreign>

Élément qui « reconnaît un mot ou une expression comme appartenant à une langue différente de celle du contexte[38] ». La langue de son contenu apparaît standardisée dans l'attribut **@xml:lang**, dont la valeur dépend toujours de l'ISO 639 (voir *supra*). Pour éviter des redondances, on ne l'utilisera pas lors des traductions en latin propres au *Trévoux*, ni avec les étymons. On peut le retrouver dans des éléments comme <def>, <etym>, <seg>, <sense>, <usg>, <xr>, etc.

---

[35] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-oRef.html. Dernière consultation le 22 juillet 2026.
[36] Ioana Galleron, art. cit.
[37] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-lang.html. Dernière consultation le 22 juillet 2026.
[38] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-foreign.html. Dernière consultation le 22 juillet 2026.

<gloss>

Élément qui « identifie une expression ou un mot utilisé pour fournir une glose ou une définition à quelque autre mot ou expression[39] ». Les gloses peuvent être introduites par des marqueurs introducteur de commentaires comme « c'est-à-dire » ou « pour dire ». À l'intérieur de <sense>, il s'agit généralement de commentaires explicatifs sur les exemples ou les usages, et, à l'intérieur des étymologies, <gloss> introduit l'équivalent en français des étymons.

Encodage idéal présenté de manière schématique

```
<entry xml:id=" " type="mainEntry" xml:lang=" ">
        <form type="lemmaGrp">
              <form type="lemma"><orth></orth><pron></pron></form>
              <metamark function="lemmaDelimmiter"></metamark>
              <gramGrp><gram type=" " expand=" " norm=" "></gram></gramGrp>
        </form>
        <sense>
               <usg></usg>
               <def></def>
               <cit type="translation" xml:lang=" "><quote></quote></cit>
               <cit type="example"><quote></quote><bibl><author></author></bibl></cit>
                <milestone unit="paragraph" rend="leftShift"/>
                <etym><lbl></lbl><lang xml:lang=" "></lang><oRef type="etymon">
                 </oRef> <gloss></gloss>
                </etym>
                <xr><lbl></lbl><ref></ref></xr>
                <milestone unit="paragraph" rend="leftShift"/>
                 <seg></seg>
                <xr><lbl></lbl><ref><bibl><author></author></bibl></ref></xr>
         </sense>
         <entry xml:id="" type="relatedEntry" xml:lang="">
                  <form type="lemma">
                          <form type="lemma"><orth></orth></form>
                          <metamark function="lemmaDelimmiter"></metamark>
                          <gramGrp><gram type=" " expand=" " norm=" "></gram></gramGrp>
                  </form>
                  <sense>
                          <def></def>
                         <cit type="example"><quote></quote></cit>
                 </sense>
         </entry>
</entry>
```

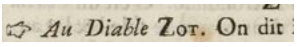
Figure 7. Proposition schématique de l'encodage idéal.

[39] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-gloss.html. Dernière consultation le 22 juillet 2026.

Éléments susceptibles d'être utilisés dans la suite du projet

Manicules

<g> : « représente un glyphe, ou un caractère non standard[40] ». Parce que l'élément <g> ne peut pas être un enfant direct de <entry>, on ne peut que l'insérer dans <form> au début d'entrée.




```
<entry type="mainEntry">
        <form type="lemmaGrp">
            <g ref="#manicula"/>
            <form type="compound"><orth>Au Diable Zot</orth></form>
            <metamark function="lemmaDelimiter">.</metamark>
        </form>
        <sense n="1">On dit […]</sense>
</entry>
```

Figure 8. Capture d'écran et proposition d'encodage de la manicule de l'entrée AU DIABLE ZOT, DUFLT 1743, tome 6, p. 1006.

Mise en évidence <hi>

Une fois l'OCR sera capable d'identifier certains trait typographiques comme les petites majuscules ou les italiques, <hi> permettra de faire la distinction graphique[41]. L'attribut @rend peut apporter l'information concernant la mise en évidence avec des valeurs comme "italic" ou "small-caps".

Normalisation de l'orthographe

<choice> : « regroupe un certain nombre de balisages alternatifs possibles pour un même endroit dans un texte[42] ».

<orig> : « contient une partie notée comme étant fidèle à l'original et non pas normalisée ou corrigée[43] ».

[40] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-g.html. Dernière consultation le 22 juillet 2026.
[41] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release//doc/tei-p5-doc/fr/html/ref-hi.html. Dernière consultation le 22 juillet 2026.
[42] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-choice.html. Dernière consultation le 22 juillet 2026.
[43] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-orig.html. Dernière consultation le 22 juillet 2026.

<reg> : « contient une partie qui a été régularisée ou normalisée de façon quelconque[44] ».

Ex. d'usage : <choice><orig></orig><reg></reg></choice>

Correction de l'orthographe

<sic> : « contient du texte reproduit quoiqu'il [soit] apparemment incorrect ou inexact ».

<corr> : « contient la forme correcte d'un passage qui est considéré erroné dans la copie du texte[45] ».

Ex. d'usage : <choice><sic></sic><corr></corr></choice>

<teiHeader> et la responsabilité des LLMs

Le questionnement sur la responsabilité dans le cadre de l'édition numérique semble central à l'ère des IAs génératives. L'élément **<appInfo>** préexistait à la vague des grands modèles de langue, permettant d'annoter dans le **<encodingDesc>** « des informations sur l'application qui a été utilisée pour traiter le fichier[46] ». Cependant, ces derniers temps, la discussion sur l'emploi de cette balise a été relancée tout en mobilisant la notion de responsabilité[47].

En admettant que l'on peut considérer un LLM comme « responsable » du balisage textuel, au-delà du débat éthique qui mobilise la propriété intellectuelle et la responsabilité juridique, la mention du modèle dans le **<respStmt>** semblerait appropriée, d'autant plus que ses descendants directs permettent de détailler l'action dont le modèle serait le responsable. En 2016, dans la discussion, encore ouverte sur GitHub, du groupe de travail qui réfléchit à ce sujet, on a proposé la piste de la communication des deux balises par leurs attributs :

> Pointing is surely much cleaner than changing the schema to allow the element to appear in multiple places, surely?
> <respStmt><resp ref="#ImageMarkupTool1.8"/> ...</respStmt>
> [...]

[44] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-reg.html. Dernière consultation le 22 juillet 2026.

[45] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-corr.html. Dernière consultation le 22 juillet 2026.

[46] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-appInfo.html. Dernière consultation le 22 juillet 2026.

[47] Voir notamment le fil de mails de la liste TEI-L disponible sur TEI-L Archives : https://lists.psu.edu/cgi-bin/wa?A2=TEI-L;8549786.2605&S=. Voir aussi les dernières réflexions sur l'utilisation de <appInfo> sur GitHub : https://github.com/TEIC/TEI/issues/1516. Dernière consultation le 22 juillet 2026.

<appInfo xml:id="ImageMarkupTool1.8">[...]</appInfo>[48]

Pour le moment deux voies se dessinent, celle qui accorde de la responsabilité au LLM (<respStmt> pouvant communiquer avec <appInfo>) et celle plus exigeante vis-à-vis de la notion de responsabilité (<appInfo>).

```
<teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Title</title>
      </titleStmt>
      <editionStmt>
        <edition></edition>
        <respStmt ref="#gpt-5-mini"><resp>TEI encoding</resp><name>gpt-5-mini</name></respStmt>
      </editionStmt>
      <publicationStmt>
        <p>Publication Information</p>
      </publicationStmt>
      <sourceDesc>
        <p>Information about the source</p>
      </sourceDesc>
    </fileDesc>
   <encodingDesc>
      <appInfo xml:id="gpt-5-mini">
        <application ident="gpt" version="5"><label>gpt-5-mini</label></application>
      </appInfo>
    </encodingDesc>
</teiHeader>
```

Figure 9. Proposition d'encodage du <teiHeader> prêtant attention aux éléments <respStmt> et <appInfo>.

En admettant que l'on utilise l'attribut **@resp** dans le <body> pour désigner vers quelle entité repose la responsabilité (on songe notamment à le combiner avec l'attribut **@cert** indiquant un seuil de certitude), la valeur de **@resp** doit pointer vers **<respStmt** : « To reduce the ambiguity of a resp pointing directly to a person or organization, we recommend that resp be used to point not to an agent (person or org) but to a respStmt, author, editor or similar element which clarifies the exact role played by the agent[49] ». Le fait que <respStmt> pointe vers <appInfo> aurait comme conséquence que @resp pointe aussi vers l'identifiant indiqué dans <appInfo>, bien que cet élément ne mentionne pas la fonction de l'agent numérique.

[48] Voir : https://github.com/TEIC/TEI/issues/1516. Dernière consultation le 22 juillet 2026.
[49] Voir « TEI : Recommandations pour l'encodage et l'échange de textes électroniques ». En ligne sur : https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.global.responsibility.html. Dernière consultation le 22 juillet 2026.

III Table de correspondances entre les balises du *Trévoux* 1743 (Le Robert) et notre proposition

Balisage du Trévoux 1743 par Le Robert	**Notre proposition de balisage**
<article ID="…">…</article>	<entry xml:id="…" type="mainEntry" xml:lang="…">
<Nat_Art>Trévoux 1743</Nat_Art>	-
<G><vedette>VEDETTE,</vedette></G>	<form type="lemma"> <orth>VEDETTE</orth> </form> <metamark function="lemmaDelimiter">,</metamark>
<cat_gra>…</cat_gra>	<gramGrp> <gram type="…" expand="…" norm="…">…</gram> <gramGrp>
<S><svedet>Sous vedette,</svedet></S>	<entry xml:id="…" type="relatedEntry" xml:lang="…"> <form type="lemma"> <orth>VEDETTE</orth> </form> <metamark function="lemmaDelimiter">,</metamark> … </entry>
<I>…</I>	<hi rend="italic">…</hi>
<S><œuvre>…</œuvre></S>	<bibl>…</bibl>
<S><auteur>…</auteur></S>	<bibl><author><hi rend="small-caps">…</hi></author></bibl>
<dom>…</dom>	<usg type="dom" value="…">…</usg>
{12}Saut de page	<pb n="…"> ou -
<grec></grec>	@xml:lang <foreign xlm:lang="…">…</foreign> <cit type="transation" xlm:lang="…"><quote>…<quote><cit> <lang xlm:lang="…">…</lang> Dans le <teiHeader>, ex. : <profileDesc><langUsage> <language ident="la"/> </langUsage></profiledesc>
<cit>…<cit>	<cit><quote>…<quote><cit>
-	<milestone/>
-	<sense>…</sense>
-	<def>…</def>
-	<etym>…</etym>
-	<xr>…</xr>

IV Méthodologie[50]

Données d'entrée

Le corpus est constitué par des données textuelles issues du DUFLT 1743, en version numérisée, corrigée et structurée en XML, que l'on a découpée par page, et dont on a supprimé le balisage existant. Ces fichiers constituent les données d'entrée utilisées pour le premier niveau des expérimentations au format .txt. Il s'agit également des mêmes pages sur lesquelles on a construit la vérité de terrain.

Configurations

Modèles

Quatre modèles de langue ont été à mobilisés tout au long de nos expérimentations : **gemini-3.1-flash-lite-preview**, **gemma-3-27b-it**, **gpt-5-mini** et **gpt-5.4-mini**. Ils ont été interrogés via l'API OpenRouter.

Stratégies

Chacun des niveaux d'encodage présenté *infra* se décline en trois stratégies de prompt engineering différentes : *few-shot* (**FS**), *few-shot* avec des instructions réduites (**FS-R)** et *zero-shot* (**ZS**). Le fichier de chaque prompt est présenté au format .txt. Ils comptent avec une stratégie de *persona prompting* (ou *role prompting*), suivie des règles absolues demandant de la fidélité au texte source et des fichiers de sortie au format XML valide, puis de la structure XML attendue. La stratégie *few-shot* inclut une explication détaillée de la logique d'analyse et d'encodage, ainsi que quelques exemples avec l'état de l'entrée (input) et le balisage de sortie (output). Quant à la stratégie *few-shot* avec des instructions réduites, elle reprend les élément de la stratégie *few-shot*, tout en éliminant les explications concernant la logique d'analyse et d'encodage. Enfin, la stratégie *zero-shot* garde lesdites explications, mais élimine les exemples.

Traitement

Comme nous le détaillerons *infra*, on a exploré, d'un côté, un traitement par niveaux avec un prompt unique contenant tout l'encodage du niveau et, d'un autre côté, un traitement

[50] Pour plus de détails sur la méthodologie, voir Maritza Beatriz García Rodríguez, Ludovic Moncla, Denis Vigier, Tatiana Lesnova, « L'encodage en XML-TEI du Dictionnaire Universel François et Latin de Trévoux. Propositions et Perspectives », *Corpus*, n° 28, 2026. À paraître.

par *prompt chaining*, qui divise les tâches en prompts successifs, de sorte que les sorties d'un modèle pour une stratégie sont les entrées du même modèle et de la même stratégie interrogés avec le prompt suivant.

Retrieval Augmented Generation (RAG)

Les explications des prompts (FS et ZS) qui concernent la normalisation de l'information grammaticale (<gram @norm="">) suivant les Universal Dependencies, et la standardisation des codes ISO des langues dans @xml:lang mobilisent des techniques propres à la *retrieval augmented generation* ou génération augmentée par récupération. Ce processus permet d'obtenir de l'information issue d'une source externe, en l'occurrence d'un URL qui offre de l'information qui permettra d'améliorer la performance des LLMs.

Niveaux d'encodage explorés

Le choix de niveaux d'encodages fait partie d'une démarche expérimentale cherchant un équilibre entre la complexité des prompts et le nombre de prompts nécessaires pour obtenir le balisage au kilomètre le plus proche possible de la vérité de terrain contenant un encodage idéal. La séquence des niveaux explorés se base sur l'accumulation des éléments introduits dans les niveaux qui précèdent (voir figure 10).

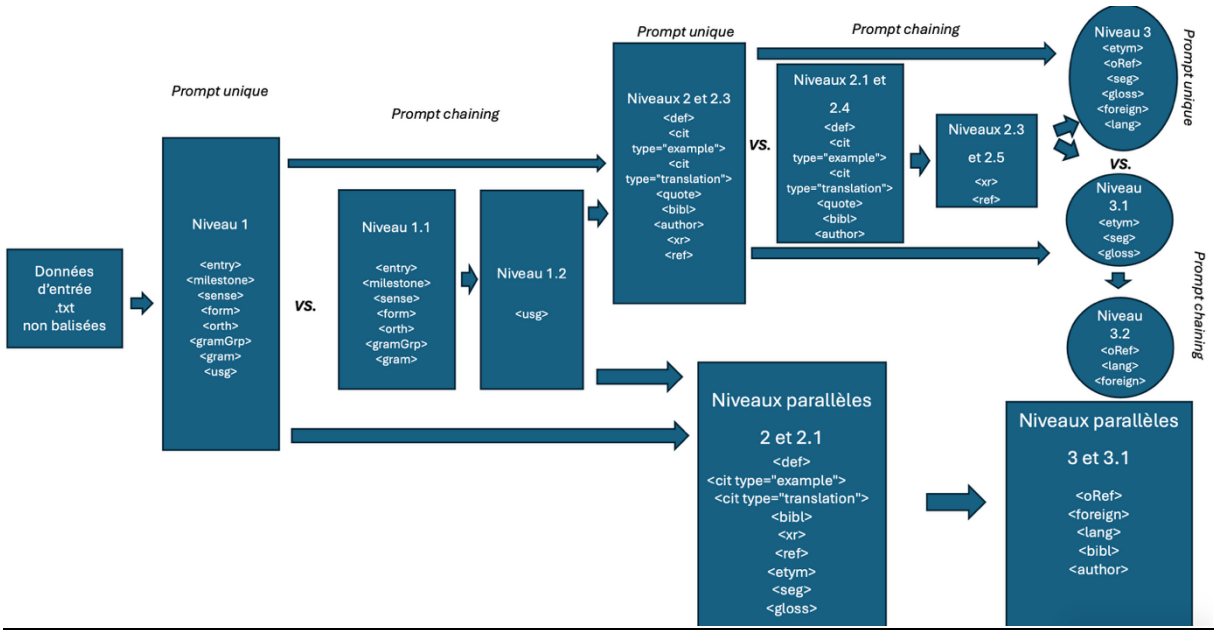


Figure 10. Structuration de l'encodage selon les éléments introduits à chaque niveau.

Premier niveau (1, 1.1 et 1.2)

Le premier niveau porte sur les éléments qui permettent de baliser les entrées et sous-entrées (<entry>), les paragraphes (<milestone>), les formes orthographiques (<form><orth>), les informations grammaticales (<gramGrp> et <gram>) et les marques d'usages. Deux traitements ont déterminé ce premier niveau : celui du prompt unique (niveau 1), dans lequel tous les éléments ont été exigés en même temps aux LLMs, et celui du *prompt chaining*, dans lequel les éléments ont été exigés de manière progressive dans deux prompts. Dans le niveau 1.1 on retrouve <entry>, <milestone>, <form>, <orth>, <gramGrp>, <gram> et <sense>, puis on introduit dans le niveau 1.2 l'élément <usg>.

Tandis que pour le prompt unique (1) et pour le premier prompt du *prompt chaining* (1.1) on explore toutes les combinaisons possibles de modèles et de stratégies, dans le deuxième prompt du *prompt chaining* (1.2), les modèles et les stratégies dont sont issues les entrées (sorties de 1.1) restent les mêmes modèles et stratégies qu'on applique pour obtenir les sorties 1.2.

Deuxième et troisième niveaux (2, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 3, 3.1, 3.2)

Le deuxième et le troisième niveau ont été conçus pour alléger les éléments demandés aux LLMs. Les niveaux 2, 2.3 et 2.6 appellent, avec un prompt unique, les définitions (<def>), les examples (<cit type="example">), les traductions (<cit type="translation">), avec l'élément <quote> nécessaire à <cit>, les renvois internes ou externes (<xr>), les références (<ref>), les auteurs (<author>) et les références bibliographiques (<bibl>). Les niveaux 2.1, 2.4 et 2.7 correspondent avec le premier prompt du *prompt chaining*, où l'on retrouve <def>, <cit type="example">, <cit type="translation">, <quote>, <author>, <bibl> ; et les niveaux 2.2, 2.5 et 2.8 correspondent au deuxième prompt du *prompt chaining*, qui interroge les mêmes modèles et stratégies des entrées appartenant au premier prompt du *prompt chaining*.

Les niveaux 2, 2.1 et 2.2 correspondent à un modèle de prompts améliorés dans les niveaux 2.3, 2.4 et 2.5. Les niveaux 2.6, 2.7, 2.8 partagent les mêmes prompts que 2.3, 2.4 et 2.5, respectivement, mais leurs entrées correspondent à la vérité de terrain du niveau 1.

Le troisième niveau, qui résulterait des meilleurs modèles et configurations du deuxième niveau existe seulement en tant que prompts, en raison des résultats insatisfaisants du deuxième niveau. On y retrouve donc les prompts uniques du niveau 3 avec les étymologies (<etym>) et leurs étymons (<oRef>), des explications ou traductions de l'étymon (<gloss>), la mention des langues (<lang>), des développements encyclopédiques (<seg>), et des mots ou expressions en

langues étrangères (<foreign>). Le premier prompt du *prompt chaining* (3.1) contient <etym>, <seg> et <gloss>, complété par le deuxième prompt avec <oRef>, <lang> et <foreign>.

Deuxième et troisième niveaux parallèles (2, 2.1, 3 et 3.1)

En réponse aux résultats insatisfaisants issus du deuxième niveau de balisage, ce que j'ai appelé le deuxième et le troisième niveau parallèle répondent à l'hypothèse selon laquelle les LLMs seraient plus performants si l'on approche le deuxième niveau selon un axe « syntagmatique », puis « paradigmatique » (voir figure 11). Il s'agirait donc de prioriser les éléments adelphes dans un même niveau et, dans le niveau suivant, d'inclure leurs enfants.

Les niveaux parallèles 2 et 2.1, chacun dans un prompts unique, introduisent les définitions (<def>), les examples (<cit type="example">), les traductions (<cit type="translation">), avec l'élément <quote> nécessaire à <cit> et parfois <bibl> — parce que <cit> ne peut pas avoir directement du texte —, les renvois internes ou externes (<xr>), les références (<ref>), les étymologies (<etym>), les explications ou traductions de l'étymon (<gloss>), et les développements encyclopédiques (<seg>). Les niveaux parallèles 3 et 3.1 ajoutent les étymons (<oRef>), la mention des langues (<lang>), les mots ou expressions en langues étrangères (<foreign>), les références bibliographiques <bibl> et la mention des auteurs (<author>). Les niveaux parallèles 2.1 et 3.1 améliorent les prompts des niveaux 2 et 3, respectivement.

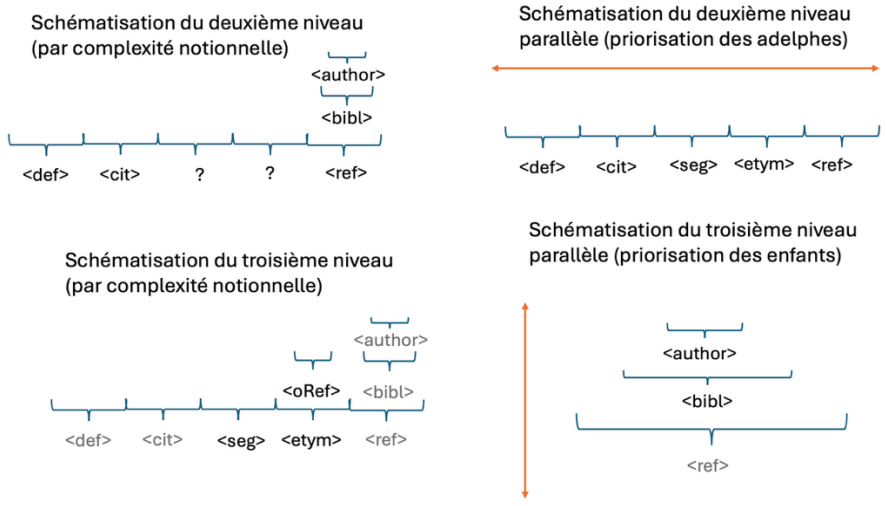


Figure 11. Schématisation du choix d'encodage des niveaux 2 et 3 (à gauche) et 2 et 3 parallèles (à droite).

Vérité de terrain

Les dix pages qui constituent la vérité de terrain ont été extraites du DUFLT 1743 et choisies de manière aléatoire parmi ses six tomes. Nous avons veillé pour garder une variété de configurations textuelles et typographiques.

volume 1 : pages 453-454 et 2001-2002 ;

volume 2 : pages 37-38 et 1785-1786 ;

volume 3 : pages 5-6 et 7-8 ;

volume 4 : pages 131-132 et 489-490 ;

volume 5 : page 505-506 ;

volume 6 : page 1003-1004.

Elles ont été annotées manuellement selon le niveau 1, le niveau 2 et le niveau 3 d'encodage, et servent de référence pour l'évaluation quantitative et qualitative.

Évaluation

Une évaluation quantitative et qualtitative a été faite par niveau d'encodage, comparant la performance des trois stratégies, des quatre modèles et des deux traitements – prompt unique et *prompt chaining*. Comme nous l'avons expliqué dans l'article soumis à la revue *Corpus*, pour l'évaluation quantitative nous avons calculé le rappel[51], la précision[52] et la micro F-mesure[53]. Pour le premier niveau d'encodage, on a prêté attention à l'identification des entrées (<entry>), tout en faisant la distinction entre les entrées (@type="mainEntry") et les sous-entrées (@type="relatedEntry"). Nous nous sommes aussi intéressés aux éléments <usg> et <milestone>, pour lesquels les mesures ne portent pas sur un appariement structurel fin des éléments, mais sur l'adéquation entre le nombre d'éléments prédits et le nombre d'éléments attendus par document.

L'évaluation quantitative du deuxième niveau et des niveaux parallèles s'est concentrée uniquement sur la quantité de balises observées, ce qui exige en contrepartie une analyse qualitative spécialement attentive. Le deuxième niveau se concentre sur les définitions (<def>), les exemples (<cit type="example">), les traductions (<cit type="translation">), les renvois

[51] Le rappel est le résultat de la division des vrais positifs entre la somme des vrais positifs et des faux négatifs, permettant d'évaluer la performance des modèles à l'égard des cas pertinents. Elle est aussi appelée sensibilité.
[52] La précision est le résultat obtenu de la division des vrais positifs entre la somme des vrais positifs et des faux positifs. Elle offre une perspective sur le taux d'erreurs de ce qui a été reconnu comme positif.
[53] La F-mesure est une métrique qui calcule la moyenne harmonique de la précision et le rappel. Elle permet de les évaluer, tout en proposant une vue globale sur la performance du modèle. La micro F-mesure est calculée à partir de la somme de tous les vrais positifs, les faux positifs et les faux négatifs.

(<xr>), les références (<ref>), les références bibliographiques (<bibl>) et les noms d'auteurs (<author>). Si l'on avait exécuté les prompts du troisième niveau – ce qui n'a pas été le cas –, on aurait calculé les métriques des étymologies (<etym>), des développements encyclopédiques (<seg>), des gloses (<gloss>), des étymons (<oRef>), des mots ou expressions en langues étrangères (<lang>) et des mentions de langues (<lang>). Le deuxième niveau parallèle se concentre sur les définitions (<def>), les exemples (<cit type="example">), les traductions (<cit type="translation">), les renvois (<xr>), les références (<ref>), les étymologies (<etym>), les développements encyclopédiques (<seg>) et les gloses (<gloss>). Finalement, le troisième niveau parallèle évalue le nombre de références bibliographiques (<bibl>), de noms d'auteurs (<author>), d'étymons (<oRef>), de mots ou expressions en langues étrangères (<lang>) et de mentions de langues (<lang>) retrouvés par les modèles.

En raison de la quantité importante de fichiers en sortie par niveau, modèle, stratégie et pages, l'évaluation qualitative a été biaisée par les résultats qualitatifs observés. Pour chaque niveau, nous avons étudié plus attentivement le ou les modèles, la ou les stratégies qui ont eu les meilleurs résultats. Nos résultats partiels (voir *infra*) portent sur ces observations.

V Contenu du dépôt GitHub

Correspondance des dossiers avec leur fonction

Vérité de terrain (./internship-data/gt)

Le dossier **gt** contient trois sous-dossiers (**niveau1**, **niveau2**, **niveau3**), chacun avec la vérité de terrain de dix pages, toujours les mêmes, dont le balisage correspond pour chaque sous-dossier aux trois niveaux d'encodage prévus dans un premier temps.

Fichiers de départ .txt (./internship-data/input)

Le dossier **input** contient les dix fichiers en format .txt que l'on donne en entrée aux modèles au début des expérimentation. Il s'agit des mêmes dix pages qui nous ont permis de constituer la vérité de terrain.

Prompts (./internship-data/prompts)

Le dossier **prompts** contient autant de sous-dossiers qu'il y a eu de niveaux d'encodage explorés (ou à explorer, comme c'est le cas des niveau3, niveau3.1 et niveau3.2). Chaque sous-dossier (**niveau1**, **niveau1.1, niveau1.2**, **niveau2**, **niveau2.1**, **niveau2.2**, **niveau2.3**, **niveau2.4**, **niveau2.5**, **niveau2.6**, **niveau2.7**, **niveau2.8**, **niveau3**, **niveau3.1**, **niveau3.2**, **niveau_parallele_2** et **niveau_parallele_3**) contient trois fichiers .txt avec les prompts selon la configuration *few-shot* (**FS**), *few-shot* réduit (**FS-R)** et *zero-shot* (**ZS**). Bien que pour des raisons expliquées *infra* on n'ait pas exploré les niveau3, niveau3.1 et niveau3.2, on a fait le choix de documenter les prompts qu'on a conçus pour eux.

Output (./internship-data/output)

Le dossier **output** contient autant de sous-dossiers qu'il y a eu de niveaux d'encodage explorés. Chaque sous-dossier (**niveau1**, **niveau1.1, niveau1.2**, **niveau2**, **niveau2.1**, **niveau2.2**, **niveau2.3**, **niveau2.4**, **niveau2.5**, **niveau2.6**, **niveau2.7**, **niveau2.8**, **niveau_parallele_2** et **niveau_parallele_3**) contient quatre autres sous-dossiers qui correspondent aux sorties obtenues par modèle (**gemini-3.1-flash-lite-preview**, **gemma-3-27b-it**, **gpt-5-mini**, **gpt-5.4-mini**). Chacun de ces quatre sous-dossiers contient trois sous-dossiers qui correspondent aux stratégies de prompt qui ont interrogé chaque modèle : *few-shot* (**FS**), *few-shot* réduit (**FS-R)** et *zero-shot* (**ZS**).

Évaluation (./internship-data/evaluation)

Le dossier **evaluation** contient autant de sous-dossiers qu'il y a eu de niveaux d'encodage évalués. Chaque sous-dossier (**niveau1**, **niveau1.2**, **niveau2**, **niveau2.2**, **niveau2.3**, **niveau2.5**, **niveau2.6**, **niveau2.8**, **niveau_parallele_2** et **niveau_parallele_3**) contient autant de fichiers en format .json qu'il y a eu de combinaison de modèle (4 modèles), de stratégie de prompts (3 stratégies) et de fichiers d'entrée (10 fichiers) par niveau, c'est-à-dire 120 fichiers par niveau qui seront à la base de l'évaluation quantitative. Les dossiers **niveau2.3**, **niveau2.5**, **niveau2.6**, **niveau2.8**, **niveau_parallele_2** et **niveau_parallele_3** comptent également avec un fichier .csv qui contient les valeurs de la micro F-mesure par élément, modèle, stratégie et niveau.

Premiers résultats (./internship-data/results)

Le dossier **results** contient deux sous-dossiers qui correspondent aux métriques (précision, rappel, F-mesure, micro F-mesure) appliquées aux niveaux 1 et 1.2 (**niveau1**, **niveau1.2**). Chaque sous-dossier a deux sous-dossiers, **figures** et **tables**, l'un avec des visualisations .png, et l'autre avec des tableaux .csv contenant les valeurs obtenues par élément selon différentes configurations. Le script permettant d'obtenir ces résultats ont été conçu rapidement, de sorte qu'il mériterait d'être révisé et amélioré avant d'être adapté pour les niveaux subséquents.

Correspondance des notebooks avec leur fonction

./ internship-data/few_shot_encoding_from_raw_text_refairexmlnonvalide.ipynb

Notebook permettant d'interroger les modèles de langue stockés dans la variable « **models** », via l'API OpenRouter. Il contient deux programmes A et B, qui permettent d'obtenir des fichiers à partir d'un chemin spécifique donné en entrée (**ocr_path_name**) et le chemin vers le niveau des prompts qui nous intéresse (**prompts**). Si les fichiers obtenus en sortie ne sont pas valides, le programme retente sa création deux fois.

Pour le programme A, le point de départ est une même configuration (ex. fichiers.txt d'origine, ou les résultats de gpt-5-mini, FS et *prompt chaining*). Le programme produit un fichier en sortie – stocké dans le répertoire donné dans **output_path** – pour chaque prompt, chaque modèle et chaque fichier en entrée.

En revanche, le programme B s'applique aux deuxièmes prompts du *prompt chaining*, qui héritent les résultats des configurations précédentes du *prompt chaining*, c'est-à-dire le même modèle et la même stratégie des fichiers produits en sortie seront le modèle et la stratégie du fichier donné ensuite en entrée. Pour chaque niveau, modèle, stratégie et chaque fichier en entrée, on obtiendra un nouveau fichier.

./internship-data/formNote_toSeg.ipynb

Notebook permettant de modifier automatiquement un élément par un autre, comme a été le cas de l'utilisation de <note></note> remplacé par <seg></seg>.

./ internship-data/Eval_niveau1.ipynb

Notebook qui crée un fichier .json contenant des métriques par page pour les niveaux 1 et 1.2. Dans chaque fichier .json, on trouve les éléments TEI évalués selon la précision, le rappel, la F-mesure, la quantité d'éléments retrouvés dans la page évaluée, ainsi que la quantité d'éléments retrouvés dans son équivalent de la vérité de terrain. Les entrées principales et les sous-entrées sont calculées selon la fidélité dans l'alignement entre la vérité de terrain et la sortie, tandis que les retours chariots et les marques d'usage sont comptabilisées par nombre d'occurrence.

./ internship-data/Eval_niveaux2et3.ipynb

Notebook qui crée un fichier .json contenant des métriques par page pour les différentes niveaux 2 et 3. Dans chaque fichier .json, on trouve les éléments TEI évalués selon la précision, le rappel, la F-mesure, la quantité d'éléments retrouvés dans la page évaluée, ainsi que la quantité d'éléments retrouvés dans son équivalent de la vérité de terrain. Cette évaluation dépend du nombre d'occurrence de chaque élément, qu'il soit repéré au bon endroit ou pas (ce point de fragilité exige un travail d'attention supplémentaire lors de l'évaluation qualitative). Dans le notebook, on décommente ou commente les éléments selon le niveau que l'on évalue.

Ce notebook produit, sous forme de tableau .csv, un premier aperçu des résultats (micro-rappel, micro-précision et micro F-mesure) comparés entre deux niveaux de notre choix indiqués dans les variables config_niveau1 et config_niveau2. Cela permet notamment de mieux comparer les sorties issues d'un même modèle et d'une même stratégie de prompting, selon un traitement avec un seul prompt ou selon le *prompt chaining*.

./ internship-data/Eval_niveauxParall.ipynb

Comme le précédent, ce notebook crée un fichier .json contenant des métriques par page, cette fois-ci pour les niveaux 2 parallèle et 3 parallèle (dont le balisage est différent de celui du notebook précédent). Dans chaque fichier .json, on trouve les éléments TEI évalués selon la précision, le rappel, la F-mesure, la quantité d'éléments retrouvés dans la page évaluée, ainsi que la quantité d'éléments retrouvés dans son équivalent de la vérité de terrain. Cette évaluation dépend du nombre d'occurrence de chaque élément, qu'il soit repéré au bon endroit ou pas (ce point de fragilité exige un travail d'attention supplémentaire lors de l'évaluation qualitative).

Ce notebook produit des tableaux .csv avec les valeurs « micro » obtenues par modèle, stratégie, niveau et élément.

./ internship-data/generate_article_results.py

Script permettant l'obtention du dossier **results** avec les visualisations et les tables décrites *supra*. Il se base sur les fichiers .json produits par le notebook Eval_niveau_1.

VI Résultats partiels et pistes à explorer

Premier niveau

Après les premiers résultats issus du premier niveau d'encodage expérimental, inclus dans l'article soumis à la revue *Corpus*[54], le *prompt chaining* est apparu comme l'une des pistes à explorer. En effet, on a vu une amélioration importante lors de la division des tâches du premier niveaux divisé en deux prompts différents (« niveau1.1 » et « niveau1.2 »). Ces résultats ont été présentés lors de la Journée d'étude « Le Trévoux à l'ère des humanités numériques et de l'IA[55] ». Ainsi, par rapport au prompt unique, le traitement en *prompt chaining* a permis d'augmenter la performance de la configuration gpt-5-mini et une stratégie *few-shot*, lorsqu'il s'agit notamment du repérage des élément compris dans le deuxième prompt (<usg type="dom"> et <usg> avec un autre @type que "dom").

traitement	**mean_f1 mainEntry**	**mean_f1 relatedEntry**	**mean_f1 milestone**	**mean_f1 usg_dom**	**mean_f1 usg_other**
prompt unique	0,94	0,20	0,46	0,10	0,20
prompt chaining	0,95	0,24	0,57	0,61	0,51

Figure 12. Moyenne de la f-mesure par type d'élément et (mainEntry, relatedEntry, milestone, usg_dom, usg_other).

Avec un traitement en *prompt chaining* et utilisant comme modèle gpt-5-mini, les résultats de la stratégie few-shot sont supérieure à ceux des strategies *zero-shot* et de la version réduite du *few-shot*, notamment lorsqu'il s'agit des <milestone>, des <usg type="dom"> et des <usg> avec un autre @type que "dom" :

stratégie	**mean_f1 mainEntry**	**mean_f1 relatedEntry**	**mean_f1 milestone**	**mean_f1 usg_dom**	**mean_f1 usg_other**
FS	0,95	0,24	0,57	0,61	0,51
FS-R	0,94	0,23	0,07	0,15	0,21
ZS	0,95	0,20	0,33	0,15	0,00

[54] Maritza Beatriz García Rodríguez, Ludovic Moncla, Denis Vigier, Tatiana Lesnova, « L'encodage en XML-TEI du Dictionnaire Universel François et Latin de Trévoux. Propositions et Perspectives », art . cit.

[55] Maritza Beatriz García Rodríguez, Denis Vigier, Ludovic Moncla, « IA générative pour l'encodage de grands corpus lexicographiques en XML-TEI », Communication Journée d'étude « Le Trévoux à l'ère des humanités numériques et de l'IA », ENS de Lyon, le 25 juin 2026.

Figure 13. Moyenne de la f-mesure par type d'élément et par stratégie de *prompting*.

Finalement, si l'on oriente la comparaison des résultats obtenus avec une configuration de stratégie *few-shot* et de traitement *prompt chaining*, on observe une plus grande variabilité dans des modèles qui se concurrencent selon le type d'élément analysé. Gpt-5-mini a la plus haute F-mesure dans les entrées, les <usg type="dom"> et les <usg> avec un autre @type que "dom" :

model	mean_f1 mainEntry	mean_f1 relatedEntry	mean_f1 milestone	mean_f1 usg_dom	mean_f1 usg_other
gemma-3-27b-it	0,76	0,26	0,44	0,00	0,00
gemini-3.1-flash-lite-preview	0,86	0,23	0,67	0,46	0,27
gpt-5-mini	0,95	0,24	0,57	0,61	0,51
gpt-5.4-mini	0,93	0,14	0.17	0,52	0,17

Figure 14. Moyenne de la f-mesure par type d'élément et par modèle.

Malgré ces résultats favorables pour la configuration gpt-5-mini, *few-shot* et *prompt chaining*, quelques problèmes majeurs ont été observés, tels que la suppression de texte (ponctuation et conjonctions entre les différentes formes orthographiques des entrées) et la duplication de certaines vedettes, lorsque celles-ci s'inscrivent dans la syntaxe de la phrase.

Deuxième et troisième niveaux

En raison des contraintes temporelles liées à la fin du stage, l'analyse qualitative des résultats du niveau 2, ainsi que des niveaux dits « parallèles » ont été biaisés par les résultats quantitatifs affichés. Ainsi, pour les différents niveaux 2 (prompt unique et *prompt chaining*, avec l'entrée issue du niveau 1 et avec l'entrée issue de la vérité de terrain) je me suis concentrée sur les sorties des trois stratégies de prompts produites par gemini-3.1-flash-lite-preview, et pour les niveaux 2 et 3 « parallèles », j'ai seulement analysé les sorties de gemini-3.1-flash-lite-preview selon la stratégie FS.

Deuxième niveau

Après une première tentative avec les prompts des niveaux 2, 2.1 et 2.2, on a amélioré les résultats obtenus dans les prompts 2.3, 2.4 et 2.5 grâce à l'indication qui distinguait les

définition des étymologies et des développements encyclopédiques (voir figure 15). Néanmoins, cela est le principal problème du deuxième niveau. Malgré les bons résultats quantitatifs, les définitions et les exemples apparaissent comme des faux positifs, là où on devrait trouver de l'information étymologique, des indications linguistiques, ou des développements encyclopédiques. Selon la stratégie, selon le traitement et selon la page, la quantité de <def> ou de <cit type="example"> explose dans gemini-3.1-flash-lite-preview. C'est pourquoi il n'a pas été possible de choisir une stratégie et un traitement plus performant que les autres. Il en a été de même pour les sorties des niveaux 2.6, 2.7 et 2.8 (voir figure 16), dont les fichiers d'entrée étaient issus de la vérité de terrain du premier niveau.

		tag	auteur	bibliographie	definition	exemple	reference	renvoi	traduction
modele	type_prompt	niveau							
gemini-3.1-flash-lite-preview	FS	niveau2.3	0.9186	0.6230	0.8871	0.6817	0.6552	0.8545	0.7786
		niveau2.5	0.8276	0.4387	0.8551	0.7623	0.6182	0.7800	0.7829
	FS-R	niveau2.3	0.7609	0.7210	0.8302	0.7346	0.6809	0.8624	0.7564
		niveau2.5	0.6825	0.6557	0.7902	0.7273	0.6522	0.7158	0.7845
	ZS	niveau2.3	0.6255	0.3262	0.8622	0.6648	0.6415	0.8244	0.7518
		niveau2.5	0.6693	0.0656	0.9000	0.7583	0.6957	0.5714	0.7609

Figure 15. Micro F-mesure par type d'élément selon le modèle gemini-3.1, selon les trois stratégies de prompting (*few-shot*, *few-shot* réduit et *zero-shot*) et selon un traitement à prompt unique (niveau2.3) ou avec du *prompt chaining* (niveau2.5). Les fichiers d'entrée sont issus des résultats produits au niveau 1 (gpt-5-mini, FS et *prompt chaining*)

		tag	auteur	bibliographie	definition	exemple	reference	renvoi	traduction
modele	type_prompt	niveau							
gemini-3.1-flash-lite-preview	FS	niveau2.6	0.8425	0.5522	0.7054	0.7682	0.3622	0.7483	0.7518
		niveau2.8	0.8185	0.3094	0.9027	0.7422	0.4706	0.7719	0.7333
	FS-R	niveau2.6	0.5823	0.7961	0.8098	0.7377	0.5484	0.7717	0.7698
		niveau2.8	0.6393	0.6494	0.7843	0.7265	0.8000	0.7347	0.7698
	ZS	niveau2.6	0.7259	0.4756	0.8847	0.7330	0.6296	0.6923	0.7872
		niveau2.8	0.5897	0.2411	0.8777	0.6737	0.4324	0.3836	0.7094

Figure 16. Micro F-mesure par type d'élément selon le modèle gemini-3.1, selon les trois stratégies de prompting (*few-shot*, *few-shot* réduit et *zero-shot*) et selon un traitement à prompt unique (niveau2.6) ou avec du *prompt chaining* (niveau2.8). Les fichiers d'entrée sont issus du premier niveau de la vérité de terrain.

Les résultats qualitatifs insatisfaisants, ainsi que la légère amélioration apportée par l'indication de ce qui n'est pas une définition, ont motivé le choix d'exploration de niveaux parallèles avec une stratégie focalisée d'abord sur les adelphes de <def> et <cit type="example">, pour le deuxième niveau parallèle, puis sur les enfants de ces éléments, pour le troisième niveau parallèle.

Restructuration des niveaux : Niveau 2 parallèle et stratégie FS

J'ai retenu les résultats de gemini-3.1-flash-lite-preview, à configuration *few-shot* en raison de l'excellente performance à l'égard des définitions, des citations et des développements encyclopédiques. Lorsqu'un développement encyclopédique englobe plusieurs paragraphes, le modèle ferme l'élément <seg> avant le <milestone> et rouvre un nouveau <seg> après. Malgré la mauvaise évaluation quantitative des <etym>, les étymologies non reconnues sont principalement des commentaires étymologiques concernant les étymons, à l'intérieur d'une autre étymologie, de sorte qu'il s'agit d'étymologies cachées par les <etym> qui les englobent. D'autres <etym> non reconnus sont souvent considérés comme des <seg>, peut-être en raison de leur présentation contenant des considérations érudites et des commentaires savants.

Certains exemples sont considérés comme des <seg>, notamment lorsqu'il s'agit de discours rapportés, mais même lors d'un encodage fait par des humains la distinction n'est pas toujours évidente. Les <seg> qui ne sont pas reconnus se retrouvent notamment à l'intérieur des <etym>.

L'élément <gloss> est très mal reconnu, de manière irrégulière, il y a des étymologies où <gloss> apparaît, et d'autres où celui-ci est absent. Cela dépend notamment des pages, au sein desquelles il y a des régularités qui disparaissent dans le passage d'une page à l'autre.

modele	type_prompt	tag / niveau	definition	encycl	etym	exemple	glose	reference	renvoi	traduction
gemini-3.1-flash-lite-preview	FS	niveau_parallele_2	0.9638	0.7633	0.4727	0.8301	0.4667	0.4444	0.8504	0.7786
	FS-R	niveau_parallele_2	0.9231	0.7548	0.3421	0.7277	0.1224	0.6737	0.8387	0.7191
	ZS	niveau_parallele_2	0.8414	0.6125	0.1871	0.6185	0.0833	0.6857	0.8545	0.7286
gemma-3-27b-it	FS	niveau_parallele_2	0.2010	0.3019	0.0000	0.0569	0.0000	0.1538	0.0000	0.0787
	FS-R	niveau_parallele_2	0.4071	0.3797	0.0000	0.0956	0.1224	0.2128	0.0000	0.1514
	ZS	niveau_parallele_2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
gpt-5-mini	FS	niveau_parallele_2	0.7088	0.4279	0.2500	0.5260	0.1600	0.4062	0.1356	0.5066
	FS-R	niveau_parallele_2	0.3061	0.0546	0.0465	0.1500	0.0000	0.0000	0.0000	0.1566
	ZS	niveau_parallele_2	0.1515	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
gpt-5.4-mini	FS	niveau_parallele_2	0.6879	0.2524	0.2000	0.7179	0.0816	0.5000	0.5526	0.6693
	FS-R	niveau_parallele_2	0.7948	0.2775	0.2378	0.3754	0.2963	0.4494	0.3889	0.6718
	ZS	niveau_parallele_2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Figure 17. Micro F-mesure par type d'élément selon le modèle et les trois stratégies de prompting (*few-shot*, *few-shot* réduit et *zero-shot*) pour le niveau_parallele_2.

Niveau 3 parallèle et stratégie FS

À partir des sorties issues du modèle gemini-3.1-flash-lite-preview avec la stratégie few-shot, nous avons exécuté le niveau parallèle 3, dont les meilleurs résultats quantitatifs ont été produit par gemini-3.1-flash-lite-preview et stratégie FS, malgré les chiffres basses pour l'élément <foreign> (voir la figure 18).

		tag	auteur	bibl	foreign	lang	oRef
modele	type_prompt	niveau					
gemini-3.1-flash-lite-preview	FS	niveau_parallele_3	0.9160	0.6667	0.4889	0.6250	0.7200
	FS-R	niveau_parallele_3	0.9064	0.6909	0.1765	0.6667	0.7826
	ZS	niveau_parallele_3	0.8342	0.3922	0.5882	0.3333	0.6809

Figure 18. Micro F-mesure par type d'élément selon gemini-3.1-flash-lite-preview et les trois stratégies de prompting (*few-shot*, *few-shot* réduit et *zero-shot*) pour le niveau_parallele_3.

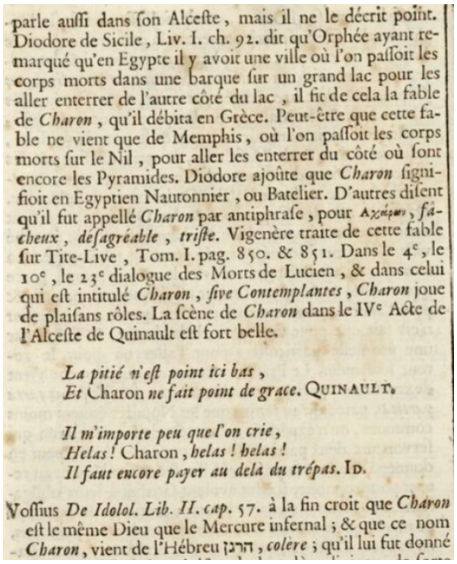
Lors de l'évaluation qualitative du niveau parallèle 3, on a constaté que l'élément <oRef> est facilement confondu avec <foreign> et que gemini-3.1-flash-lite-preview avec une configuration FS a introduit des caractères en grec et en hébreux, à la page 2001 du tome 1, là où l'OCR n'avait pas reconnu ces caractères. Il ne s'agit pas du mot exact que l'on retrouve dans le *Trévoux*, mais les graphies restent proches.

```
<seg>parle aussi dans son Alceste, mais il ne le décrit point. <ref><bibl><author>Diodore
       de Sicile</author>, Liv. I. ch. 92.</bibl></ref> dit qu'Orphée ayant remarqué
  qu'en Egypte il y avoit une ville où l'on passoit les corps morts dans une barque sur un
  grand lac pour les aller enterrer de l'autre côté du lac, il fit de cela la fable de
  Charon, qu'il débita en Grèce. Peut-être que cette fable ne vient que de Memphis, où
  l'on passoit les corps morts sur le Nil, pour aller les enterrer du côté où sont encore
  les Pyramides. Diodore ajoûte que Charon signifioit en Egyptien Nautonnier, ou Batelier.
  D'autres disent qu'il fut appellé Charon par antiphrase, pour <foreign xml:lang="grc"
    >χαίρων</foreign>, fâcheux, désagréable, triste.
```

Figure 19. Caractères en grec, ajoutés en sortie par le modèle gemini-3.1-flash-lite-preview, avec une stratégie *few-shot*, et selon le niveau_parallele_3, pour la page 2001, du tome 1 du DUFLT 1743, entrée CHARON.

```
<milestone unit="paragraph" rendition="#leftShift"/>
<etym><ref><bibl><author>Vossius</author> De Idolol. Lib. II. cap. 57.</bibl></ref> à la
  fin croit que Charon est le même Dieu que le Mercure infernal ; &amp;amp; que ce nom Charon,
  vient de l'Hébreu <foreign xml:lang="he">חרון</foreign>, colère ; qu'il lui fut donné
  parce qu'il étoit le ministre de la colère divine ; de sorte que Charon signifie
  proprement un mauvais Ange, dont l'office est de conduire les ames criminelles au lieu
  du supplice.</etym>
```

Figure 20. Caractères en hébreu, ajoutés en sortie par le modèle gemini-3.1-flash-lite-preview, avec une stratégie *few-shot*, et selon le niveau_parallele_3, pour la page 2001, du tome 1 du DUFLT 1743, entrée CHARON.



parle auſſi dans ſon Alceſte, mais il ne le décrit point. Diodore de Sicile, Liv. I. ch. 92. dit qu'Orphée ayant remarqué qu'en Egypte il y avoit une ville où l'on paſſoit les corps morts dans une barque ſur un grand lac pour les aller enterrer de l'autre côté du lac, il fit de cela la fable de *Charon*, qu'il débita en Grèce. Peut-être que cette fable ne vient que de Memphis, où l'on paſſoit les corps morts ſur le Nil, pour aller les enterrer du côté où ſont encore les Pyramides. Diodore ajoûte que *Charon* ſignifioit en Egyptien Nautonnier, ou Batelier. D'autres diſent qu'il fut appellé *Charon* par antiphraſe, pour Αχαίρων, *fâcheux, deſagréable, triſte.* Vigenère traite de cette fable ſur Tite-Live, Tom. I. pag. 850. & 851. Dans le 4^e^, le 10^e^, le 23^e^ dialogue des Morts de Lucien, & dans celui qui eſt intitulé *Charon, ſive Contemplantes*, *Charon* joue de plaiſans rôles. La ſcène de *Charon* dans le IV^e^ Acte de l'Alceſte de Quinault eſt fort belle.

La pitié n'eſt point ici bas,
Et Charon *ne ſait point de grace.* QUINAULT.

Il m'importe peu que l'on crie,
Helas! Charon, *helas! helas!*
Il faut encore payer au dela du trépas. ID.

Voſſius *De Idolol. Lib. II. cap.* 57. à la fin croit que *Charon* eſt le même Dieu que le Mercure infernal; & que ce nom *Charon*, vient de l'Hébreu חרון, *colere*; qu'il lui fut donné

Figure 21. Page 2001, du tome 1 du DUFLT 1743, entrée CHARON, où l'on retrouve les termes en grec (ligne 11) et en hébreu (dernière ligne) reproduits par gemini-3.1-flash-lite-preview, *few-shot* et niveau_parallele_3.

Niveaux parallèles 2.1 et 3.1 et stratégie FS

Le niveau parallèle 2.1 est une version des prompts du niveau parallèle 2 notamment avec des légers changements, notamment ceux qui concernent la possibilité de retrouver l'élément <cit> à l'intérieur de <seg> et l'élément <seg> à l'intérieur de <etym>. En conséquence, les discours rapportés sont mieux repérés lorsqu'ils se trouvent à l'intérieur des <seg>. Bien que le modèle trouve plus d'étymologies que le niveau parallèle 2, les étymologies des étymons ne sont toujours pas retrouvées.

Malgré ces améliorations, les définitions qui sont repérées sont moins fines, parfois celles-ci recouvrent des variantes de la vedette bien identifiées préalablement et respectées dans le niveau parallèle 2. Le modèle a également modifié certains éléments <sense>, les fusionnant avec d'autres, ce qui explique une partie des <def> non reconnues.

		tag	definition	encycl	etym	exemple	glose	reference	renvoi	traduction
modele	type_prompt	niveau								
gemini-3.1-flash-lite-preview	FS	niveau_parallele_2.1	0.9504	0.7855	0.5434	0.7920	0.6667	0.4848	0.8281	0.7609
	FS-R	niveau_parallele_2.1	0.9208	0.5746	0.3636	0.8198	0.3793	0.5872	0.9000	0.7786
	ZS	niveau_parallele_2.1	0.6723	0.6174	0.3333	0.5474	0.3137	0.7606	0.8850	0.6996

Figure 22. Micro F-mesure par type d'élément selon gemini-3.1-flash-lite-preview et les trois stratégies de prompting (*few-shot*, *few-shot* réduit et *zero-shot*) pour le niveau_parallele_2.1.

Le niveau parallèle 3.1 a comme entrée le niveau 2.1 avec une stratégie FS, et son prompt modifie légèrement les instructions de <oRef>, <foreign> et <lang>, ce qui permet d'obtenir des légères améliorations dans les résultats quantitatifs de chacun, bien que cela ait comporté une légère baisse de <author>. Les sorties de la stratégie *few-shot* et *few-shot* réduit montrent certains <foreign> là où devrait apparaître <oRef>. FS génère des mots grecs ajoutés par le modèle à l'intérieur de <oRef> (p. 1003-1004, tome 6) et FS-R, du grec et de l'hébreu (p. 2001, tome 1). Ces ajouts ne correspondent pas systématiquement aux étymons qui apparaissent dans le DUFLT 1743.

		tag	auteur	bibl	foreign	lang	oRef
modele	type_prompt	niveau					
gemini-3.1-flash-lite-preview	FS	niveau_parallele_3.1	0.8707	0.7579	0.5581	0.7826	0.7654
	FS-R	niveau_parallele_3.1	0.8707	0.7447	0.5000	0.6364	0.8788
	ZS	niveau_parallele_3.1	0.0000	0.5169	0.2759	0.0000	0.0000

Figure 23. Micro F-mesure par type d'élément selon gemini-3.1-flash-lite-preview et les trois stratégies de prompting (*few-shot*, *few-shot* réduit et *zero-shot*) pour le niveau_parallele_3.1.

Pistes à explorer

Dans le deuxième et troisième niveaux (parallèles et non parallèles), le fait de consacrer plus d'attention aux sorties dont la micro F-mesure a des valeurs basses permettrait d'éviter le biais lié à l'évaluation quantitative. Cela n'a pas eu lieu en raison des contraintes temporelles. C'est aussi en raison de la fin du stage que je n'ai pas pu explorer de traitement *prompt chaining* pour les niveaux parallèles. Cependant, le *prompt chaining* risquerait d'aller à l'encontre de la sélection des adelphes de <def> et de <cit type="example"> utilisés pour éviter des faux positifs. Une possibilité serait de déplacer dans un deuxième prompt les renvois et les références, dont la structure semble facilement repérable et, à mon sens, ne risque pas d'être confondue par des définitions ou des exemples.

Bien que la configuration la plus performante du premier niveau d'encodage semble claire (gpt-5-mini, FS et *prompt chaining*) et que le deuxième niveau s'incline vers les sorties des niveaux parallèles produites par gemini-3.1-flash-lite-preview, le choix entre les niveaux 2 et 3 parallèles ou 2.1 et 3.1 parallèles est encore à faire. Les résultats au niveau des définitions sont plus intéressants dans le niveau 2 parallèle – bien que rien n'ait changé autour de cet élément dans le prompt –, tandis que les étymologies et les gloses sont mieux reconnues dans 2.1 parallèle – ce qui s'explique par les améliorations du prompt concernant ces éléments. Cet endroit d'indécision entre 2 et 2.1 (et en conséquence entre 3 et 3.1) pourrait se résoudre avec l'exploration des niveaux parallèles traités avec du *prompt chaining*.