16 – Interprétation et analyse

On dit souvent que le balisage constitue une forme d'interprétation ou d'analyse. Tandis qu'il est certainement difficile, voire impossible, de distinguer de façon certaine entre les informations « objectives » et « subjectives » d'une manière universelle, il reste vrai que les jugements concernant les dernières sont plus sujets à controverse que les premières. De nombreux chercheurs préfèrent donc présenter de telles interprétations uniquement lorsqu'il est possible d'informer le lecteur qu'il s'agit de parties plus sujettes à caution que le reste du balisage. Cette section décrit certains des éléments fournis par la TEI pour répondre à ce besoin.

16.1. Phrases orthographiques

Typiquement, l'interprétation porte sur l'ensemble d'un texte, sans prendre particulièrement en compte les autres unités structurales. Un préliminaire utile avant toute interprétation intensive consiste donc à segmenter le texte dans des unités discrètes et identifiables, chacune pouvant ensuite recevoir une étiquette servant de « référence canonique ». Afin de faciliter ce genre d'application, ces unités ne doivent pas se croiser, ni être imbriquées les unes dans les autres. Il est possible de les représenter de façon commode en utilisant l'élément suivant :

<s> identifie un *s–unit* au sein d'un document, dans le but d'établir un système de référence s'appliquant au texte entier ; parmi les attributs possibles, citons :

- *type* qui catégorise l'unité (par exemple comme étant `déclaratif` `interrogatif`, etc.).

Comme le nom suggère, l'élément *<s>* est le plus souvent employé (au moins dans les applications linguistiques) pour marquer des phrases orthographiques, c'est-à-dire des unités définies par des caractéristiques orthographiques telles que la ponctuation. Par exemple, le passage pris dans *Jane Eyre* cité dans le premier chapitre pourrait être divisé en *s–units* comme suit :

```
<pb n='474'>
<div1 type=chapter n='38'>
<p><s n=001>Reader, I married him.</s>
<s n=002>A quiet wedding we had:</s>
<s n=003>he and I, the parson and clerk, were
         alone present.</s>
<s n=004>When we got back from church, I went
into the kitchen of the manor-house, where Mary
was cooking the dinner,
and John cleaning the knives, and I said &dash;</s>
<p><q><s n=005>Mary, I have been married
to Mr Rochester this morning.</s></q>...}
```

Les balises de fin utilisées ci-dessus ne sont pas strictement nécessaires, car les éléments *<s>* ne peuvent pas s'emboîter : le début d'un élément *<s>* implique que le précédent est terminé. Quand les *s–units* sont balisés comme indiqué ci-dessus, il est conseillé de baliser le texte tout entier, afin que chaque mot du texte analysé soit contenu dans exactement un élément *<s>*, dont l'identifiant peut alors servir de référence unique. Si les identifiants employés sont uniques au sein du document, alors l'attribut *id* pourrait être utilisé de préférence à l'attribut *n* employé dans l'exemple ci-dessus.

16.2. Eléments d'interprétation polyvalents

Un élément de segmentation plus polyvalent, le *<seg>* a déjà a été présenté : il sert à identifier les cibles de références croisées et de liens hypertexte qui ne disposent pas par ailleurs d'éléments d'identification – se reporter à la section 8 « Liens et références croisées ». Cet élément identifie une quelconque expression de type «phrase» à laquelle le codeur peut attribuer un type spécifié par l'utilisateur, ainsi qu'un identifiant unique ; il peut ainsi servir à baliser les caractéristiques textuelles pour lesquelles rien n'a été prévu dans la version publiée des recommandations de la TEI.

Par exemple, les Recommandations ne fournissent aucun élément *<apostrophe>* pour marquer les parties d'un texte littéraire où le narrateur s'adresse au lecteur (ou à l'auditeur) directement. Une approche possible serait de considérer ces cas comme des exemples de l'élément *<q>*, distingués des autres par une valeur appropriée de l'attribut *who*. Cependant, une solution peut-être plus simple (et certainement plus générique) serait d'employer l'élément *<seg>* comme suit :

```
<div1 type=chapter n='38'>
<p><seg type='apostrophe'>Reader, I married him.</seg>
A quiet wedding we had:...
```

L'attribut *type* de l'élément <*seg*> peut prendre n'importe quelle valeur, donc peut servir à enregistrer les phénomènes du genre «expression» de toute sorte ; il est souhaitable d'enregistrer les valeurs employées ainsi que leur signification dans l'en-tête.

Un élément <*seg*> d'un type donné (à la différence de l'élément <*s*> auquel il ressemble superficiellement) peut s'emboîter à l'intérieur d'un élément <*seg*> du même type ou non. Cela permet la représentation de structures relativement complexes ; certains exemples ont été donnés dans la section 8.3 « Attributs de liaison » ci-dessus. Cependant, puisqu'il doit respecter la contrainte imposée par SGML qui est que les éléments doivent être emboîtés correctement et ne doivent pas se chevaucher, cet élément ne permet pas d'associer une interprétation à des segments arbitraires d'un texte éventuellement sans tenir compte de la hiérarchie du document. Il nécessite également que l'interprétation elle-même soit représentée par une valeur unique codée au sein de l'attribut *type*.

Aucune de ces contraintes ne s'applique à l'élément <*interp*>, qui comporte de puissantes caractéristiques permettant de coder d'une manière assez simple des interprétations relativement complexes.

<*interp*>
: fournit une annotation interprétative qui peut être reliée à un passage de texte. Les attributs sont :

  valeur identifie le phénomène spécifique en cours d'annotation ;

  resp identifie le responsable à l'origine de l'interprétation ;

  type indique quelle sorte de phénomène est notée dans ce passage ; des valeurs possibles de cet attribut sont :
  `image`, `caractère`, `thème`, `allusion`,
  ou le nom d'un type particulier de discours dont des exemples sont identifiés;

  inst pointe vers des exemples de l'analyse ou de l'interprétation représentée par l'élément courant.

<*InterpGrp*>
: regroupe un ensemble de balises <*interp*>.

Ces éléments permettent au codeur de préciser à la fois la classe d'une interprétation, et l'instance particulière de cette classe impliquée par l'interprétation. Ainsi, tandis qu'avec <*seg*> nous pouvons dire simplement que quelque chose est une apostrophe, avec <*interp*> nous pouvons dire qu'il s'agit d'une instance (apostrophe) d'une plus grande classe (de figures rhétoriques).

De plus, *<interp>* est un élément vide qui doit être relié au passage auquel il s'applique, soit au moyen de l'attribut *ana* présenté dans la section 8.3 «Attributs de liaison», ci-dessus, soit au moyen de son propre attribut *inst*. Cela signifie que toute sorte d'analyse peut être représentée, sans que l'on soit obligé de respecter la hiérarchie SGML du document ; cette approche facilite également le regroupement d'analyses d'un type particulier. Un élément spécifique *<interpGrp>* est fourni à cette fin.

Par exemple, supposons que l'on désire baliser des caractéristiques textuelles variées telles des thèmes, des sujets ou des figures rhétoriques, ainsi que les emplacements des diverses scènes d'un récit. Par exemple, diverses parties de notre passage spécimen puisé dans *Jane Eyre* pourraient être associées à des figures rhétoriques comme l'apostrophe, l'hyperbole et la métaphore, accompagnées de références à des églises, des serviteurs, la cuisine, la poste ou des lunes de miel, et illustrées par des scènes à l'église, dans la cuisine, ou dans un endroit non spécifié (le salon ?).

Ces interprétations pourraient être placées n'importe où dans l'élément *<texte>* ; il est cependant souhaitable de les inclure au même endroit (par exemple dans une section séparée des textes liminaires ou annexes), comme dans l'exemple suivant :

```
<back>
<div1 type='Interpretations'>
<interp id='fig-apos'  resp='LB, MSM'
     type='figure of speech' value='apostrophe'>
<interp id='fig-hyp'   resp='LB, MSM'
     type='figure of speech' value='hyperbole'>
<!--... -->
<interp id='set-church'  resp='LB, MSM'
     type='setting' value='church'>
<!--... -->
<interp id='ref-church'  resp='LB, MSM'
     type='reference' value='church'>
<interp id='ref-serv'    resp='LB, MSM'
     type='reference' value='servants'>
<!--... -->
</p></div>
```

La redondance évidente de ce type de codage peut être considérablement réduite en utilisant l'élément *<interp>* pour regrouper tous ces éléments *<interp>* qui partagent des valeurs d'attribut communes, comme suit :

```
<back>
<div1 type='Interpretations'>
<interpGrp type='figure of speech' resp='LB, MSM'>
<interp id='fig-apos' value='apostrophe'>
<interp id='fig-hyp'  value='hyperbole'>
<interp id='fig-meta' value='metaphor'>
<!--... -->
</interpGrp>
<interpGrp type='scene-setting' resp='LB, MSM'>
<interp id='set-church'  value='church'>
<interp id='set-kitch'   value='kitchen'>
<interp id='set-unspec'  value='unspecified'>
<!--... -->
</interpGrp>
<interpGrp type='reference' resp='LB, MSM'>
<interp id='ref-church' value='church'>
<interp id='ref-serv'   value='servants'>
<interp id='ref-cook'   value='cooking'>
<!--... -->
</interpGrp>
</p></div>
```

Une fois que ces éléments d'interprétation ont été définis, ils peuvent être reliés aux parties du texte auquel ils s'appliquent, d'une ou deux manières. L'attribut *ana* peut servir sur tout élément approprié :

```
<div1 type=chapter n='38'>
<p id='P38.1' ana='set-church set-kitch'>
<s id=P38.1.1 ana='fig-apos'>Reader, I married him.</s>
...
```

Notons dans cet exemple que puisque le paragraphe a deux scènes (dans l'église et dans la cuisine), les identifiants des deux sont fournies.

Alternativement, les éléments *<interp>* peuvent pointer sur toutes les parties du texte auquel ils s'appliquent, en utilisant leur attribut *inst* :

```
<interp id='fig-apos' type='figure of speech' resp='LB, MSM'
   value='apostrophe' inst='P38.1.1'>
<!--... -->
<interp id='set-church'  type='scene-setting' value='church'
   inst='P38.1' resp='LB, MSM'>
<interp id='set-kitchen' type='scene-setting' value='kitchen'
   inst='P38.1' resp='LB, MSM'>
<!--... -->
```

L'élément <*interp*> n'est pas limité à un quelconque type d'analyse ; l'analyse littéraire montrée ci-dessus n'est qu'une possibilité et l'on pourrait aussi bien employer <*interp*> afin de donner une analyse linguistique des parties du discours. Par exemple, la phrase d'exemple donnée dans la section 8.3 « Attributs de liaison » suppose une analyse linguistique qui pourrait être représentée comme suit :

```
<interp id=NP1 type=pos value='noun phrase, singular'>
<interp id=VV1 type=pos value='inflected verb,
             present-tense singular'>
...
```