Stratégie d'insertion des balises (V1.0)

Initialement le texte que l'on charge dans l'éditeur est dépourvu de balise, c'est l'annotateur qui les introduit une à une dans l'ordre qu'il souhaite. Il est néanmoins souhaitable d'effectuer un certain nombre de contrôles ou d'opérations automatiques afin de faciliter la tâche de l'utilisateur et lui éviter des erreurs.

Contrôles du balisage

Dans la version actuelle de l'éditeur la seule manière de vérifier que le balisage effectué par l'utilisateur est correct vis-à-vis de la DTD est d'utiliser « a posteriori » un validateur. Or certaines situations devraient être évitables « a priori ». Par exemple, si l'utilisateur effectue deux balisages successifs d'une partie de texte avec un recouvrement, on obtient aujourd'hui un code XML incohérent :

1) Tag A : « un exemple de phrase » : un exemple <A> de phrase </A>
2) Tag B : « un exemple de phrase » : <B> un exemple <A> de </B> phrase </A>

Plusieurs solutions sont possibles pour éviter/corriger une telle situation :

- soit en décalant les étiquettes précédemment existantes (ici <A>) de manière à faire disparaître le chevauchement des balises,
- soit en décalant une des étiquettes en conflit de manière à obtenir une inclusion stricte (A dans B ou l'inverse).
- soit en dupliquant le texte (ici le mot « de » apparaîtrait dans les deux zones),
- soit plus simplement en signalant une erreur

Le comportement à adopter pourrait être, au moins dans les deux premiers cas, partiellement automatisé en prenant en compte la structure de la DTD :

- Si A est plus spécifique[1] que B <B> un exemple <A> de phrase </A> </B>
- Si B est plus spécifique que A <A> <B> un exemple de phrase </A> </B>
- Si A et B sont sans relation <B> un exemple de </B> <A> phrase </A>

Dans ce dernier cas (A et B sans relation) si les zones sélectionnées étaient exactement les mêmes (mêmes positions de début et fin) on pourrait éventuellement supprimer le balisage A et le remplacer entièrement par celui de B.

Autre point, il serait très souhaitable que l'éditeur adopte plus largement lorsque cela à un sens *le comportement classique des TdT*. Ainsi, deux zones contiguës ayant le même balisage devraient être automatiquement fusionnées. Par exemple:

1) Tag A : « un exemple de phrase » : un exemple <A> de phrase </A>
2) Tag A : « un exemple de phrase » : un <A> exemple de phrase </A>
~~un <A> exemple </A><A> de phrase </A>~~

[1] Plus spécifique signifiant que dans la structure de la DTD la balise A peut apparaître dans une balise de type B a une profondeur quelconque.

De même lorsque la balise qui est appliquée sur la zone sélectionnée est la même que la balise que l'on trouvait antérieurement (et que cette balise ne peut être appliquée récursivement), on peut modifier simple ment la taille de la zone concernée.

1) Tag A : « un exemple de phrase » : un exemple <A> de phrase </A>
2) Tag A : « un exemple de phrase » : <A> un exemple de phrase </A>
~~<A> un <A> exemple de phrase </A></A>~~

Enfin, une autre manière de faciliter l'introduction des balises serait de n'activer[2] à tout instant que les balises qui sont « potentiellement » acceptables, c-à-d celles qui sont « plus spécifiques[3] » dans la DTD que la balise « courante » qui est explicitée par la position du curseur ou celle de la zone sélectionnée (en prenant en compte l'exception précédemment mentionnée ou les zones sélectionnées et balisées sont exactement les mêmes). Pour chaque balise, cette liste de dépendances peut être facilement calculée une fois pour toute lors de la lecture de la DTD …

Ces options sont évidemment à discuter plus amplement en fonction de la complexité de leur mise en place. Pour les implémenter il faut probablement disposer d'une API minimale du genre associée à une liste (ou arbre) triée des positions de balises :

GetGlobalBalise (selection) → Balise englobant immédiatement la sélection
GetBeginningBalise (selection) → Liste des balises commencant dans la zone sélection et se finissant à l'extérieure de celle-ci
GetEndingBalise (selection) → Liste des balises finissant dans la zone sélection et se commencant à l'extérieure de celle-ci
GetIncludedBalise (selection) → Liste des balises incluses dans la sélection
GetMostSpecific(balise) → Renvoie les balises plus spécifiques de la DTD

Il est clair que lorsque la zone sélectionnée est importante ces méthodes peuvent conduire à des temps de calcul importants. Une manière de limiter ce problème serait par exemple de borner la taille des listes renvoyées ce qui ne devrait pas créer de différence notable de fonctionnement du système du point de vue de l'utilisateur.

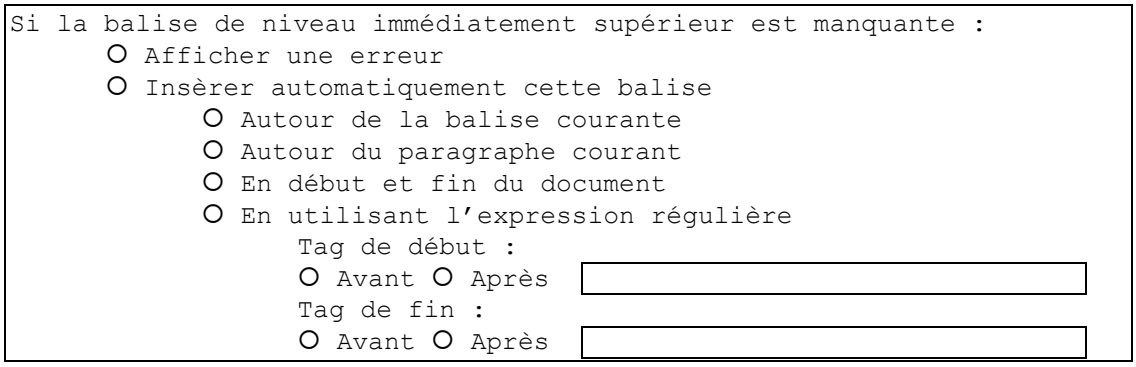
Insertion automatique

Dans la DTD que nous utilisons (mais ce n'est pas spécifique), le marquage d'un agent d'interaction, d'une cible, d'une interaction … sont assez simples à effectuer pour l'utilisateur (biologiste) puisqu'elles consistent à affecter un sens biologique aux éléments du texte. Par contre, l'introduction des balises de structuration plus générales (par exemple ABSTRACT, SENTENCE, …) est nettement moins intuitive et elles risquent d'être assez systématiquement oubliées ce qui conduira à devoir effectuer de nombreuses corrections. Par ailleurs, on constate que certaines balises sont

[2] Deux solution sont possibles : soit clairement de désactiver les balises non autorisées, soit de les tagger (icône) de manière particulière de manière à signaler qu'une incohérence existe entre les deux balisages.
[3] La mise en place d'une optio complémentaire qui consisterait à n'activer que des balises « plus générales » que l'ensemble des balises contenue dans la zone sélectionnée est également envisageable mais risque d'être plus couteuse en temps.

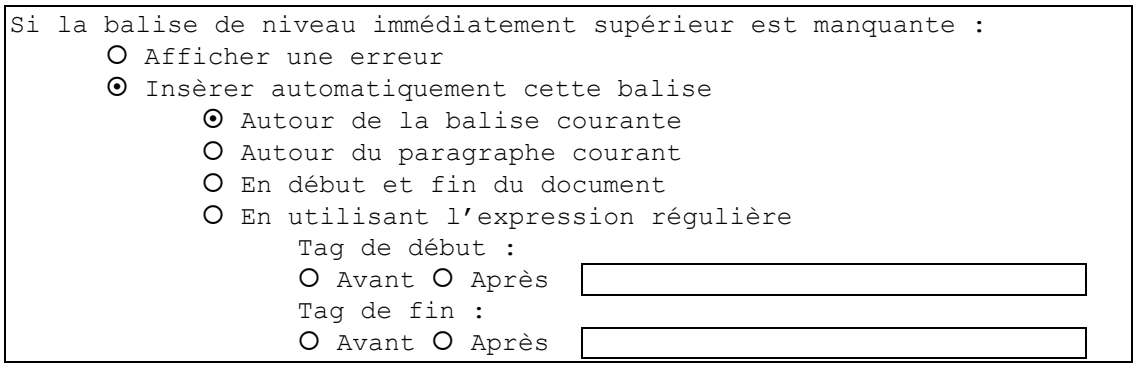
sémantiquement « couplées » dans la DTD : par exemple, une balise d'agent d'interaction <A> est forcément englobée dans un fragment <AF>.

Pour résoudre ces différents problèmes, il semble nécessaire de pouvoir associer aux feuilles de style des directives permettant d'insérer automatiquement les balises de niveau supérieur (englobantes) lorsqu'elles sont absentes dans le document. Or, cette information n'est « qu'en partie » déductible de la DTD, car si celle-ci permet de connaître les balises englobantes, elle ne permet pas de savoir OU elles doivent être introduite dans le document. Pour y parvenir voici les options que l'on pourrait placer dans la feuille de style associée à une balise.



On trouve deux séries d'options imbriquées (radio-boutons) dans la boîte de dialogue. Tout d'abord pour décider si l'absence de l'étiquette de niveau supérieur déclenche l'affichage d'un dialogue d'erreur ou non, puis si ce n'est pas le cas pour décider où sera insérée la balise supérieure manquante. La dernière option permet d'effectuer une recherche des points d'insertion à l'aide d'une expression régulière (la recherche du point d'insertion de la balise ouvrante s'effectuant vers le début du document).

Ainsi, l'exemple ci-dessus est la configuration de la feuille de style associée à une balise <An> de manière à ce qu'on ajoute automatiquement les balises <Afn> autour de la balise <An> introduite en premier.



Dès lors, grâce à un tel mécanisme si l'annotateur sélectionne la partie "centrale" de l'agent (balise <An>) d'une interaction avant d'avoir décrit la partie "large" (balise <Afn>) il serait possible d'ajouter automatiquement cette dernière dans le texte.

En pratique, cette gestion de l'insertion des balises manquantes doit être effectuée de manière récursive jusqu'à ce qu'on arrive à la racine de la DTD ou jusqu'à ce que l'on retombe sur une balise qui était prévue par la DTD. Ainsi si l'utilisateur charge un

document et commence directement à introduire une balise <A1> dans une phrase quelconque, cela déclenchera les insertions suivantes en cascade (pour peu que les feuilles de styles aient été correctement configurées) :

- Ajout de la balise <AF1> autour de <A1>
- Ajout de la balise <INTERACTION> autour du paragraphe courant
- Ajout de la balise <SECTION> autour de <INTERACTION>
- Ajout de la balise <ANNOTATED-DOCUMENT> en début et fin de document

On pourra utiliser une boîte de dialogue de ce type pour associer les feuille de style aux balises et modifier les paramètres graphiques et sémantiques. Eventuellement les deux types de paramètres seront placés dans des onglets différents de manière à clarifier l'affichage.