Cahier des charges pour l'importation dans la plateforme TXM du corpus Corptef (encodage XML-TEI-BFM)

Table de révision :

Date	Auteur	Commentaire
01/04/10	Alexei Lavrentiev (AL)	Création

Table des matières

But de ce document

Pour importer un corpus dans la plateforme TXM, il faut effectuer un certain nombre de traitements de transformation du format source vers le format cible d'intégration dans la plateforme. Ces traitements vont construire petit à petit les informations cibles, qui seront accessibles dans la plateforme pour les analyses, à partir des informations sources. Ce document spécifie les informations cibles à partir d'une description des informations sources. Il est structuré selon les catégories générales d'informations cibles. Pour chaque catégorie d'information, une spécification informelle (ou documentaire) et une spécification formelle (ou technique) est renseignée, si pertinent et si possible. Ces catégories ne sont pas étanches, il peut donc y avoir des redites. Pour chaque catégorie, une section de commentaires permet de faire un retour sur l'expression du besoin de sorte à faire évoluer le document vers une meilleure inter-compréhension de l'usage souhaité de la ressource importée dans la plateforme TXM. Pour faciliter la lecture, on peut saisir en rouge l'expression du besoin et en bleu les commentaires.

Commentaires généraux sur les spécifications

Dans ce document, les renseignements fournis par AL

Unité documentaire

L'unité documentaire sert dans un premier temps à porter les informations de métadonnées portant sur l'intégralité d'un document. Elle peut dans un second temps servir à construire des partitions sur le corpus (par exemple en créant un contraste diachronique entre textes selon leur date). Cette unité se décrit de façon documentaire (par exemple : roman, poème, pièce, œuvre, livre…) et de façon technique selon le format de la source (par exemple en TEI : <TEI>, <text>, <body>, …

Spécification documentaire de l'unité documentaire

L'unité documentaire est une édition d'une œuvre médiévale. Le plus souvent, il s'agit du « corps » de texte d'un livre, mais il peut s'agir :

- d'un volume dans une édition qui en content plusieurs ;
- d'un texte individuel publié dans une anthologie ou dans un article d'une revue...

Spécification technique de l'unité documentaire

Tous sont encodés en XML/TEI. Chaque unité documentaire est un fichier TEI contenant un entête <teiHeader> et un <text>. La typologie de chaque document est déclarée dans plusieurs éléments de l'entête.

Commentaires sur les spécifications

Métadonnées utiles

Les métadonnées servent à créer des sous-corpus et des partitions (métadonnées principales), mais aussi des références pour la localisation bibliographique dans les concordances par exemple (métadonnées secondaires) ou encore des informations qui participeront à l'établissement de la fiche bibliographique d'une unité documentaire.

Spécification des métadonnées principales

1. titre (identifiant)
2. auteur
3. date de composition formelle
4. sous-siècle
5. domaine
6. genre
7. forme
8. dialecte

Spécification technique des métadonnées principales

1. fileDesc/titleStmt/title[@type='reference']
2. fileDesc/titleStmt/
3. profileDesc/creation/date[type='compo']
4. profileDesc/creation/date[type='compo_sous_siecle']/@n
5. profileDesc/textDesc/domain/@type
6. profileDesc/textDesc/@n
7. profileDesc/textClass/catRef/@target[contains(.,'forme')]
8. profileDesc/creation/region[@type='dialecte_auteur']

Commentaires sur les spécifications

Spécification des métadonnées secondaires

- [à compléter... non essentiel pour l'intégration]

Spécification technique des métadonnées secondaires

1.

Commentaires sur les spécifications

Dimensions textuelles

Les dimensions textuelles servent à préciser quels seront les différents plans textuels à modéliser au sein de la plateforme ainsi que tout ce qui ne doit pas être pris en compte dans la source dans la construction de la surface du texte (hors-texte).

Source ou langue principale du document

Langue principale des unités lexicales qui composeront le vocabulaire et les index de recherche.

Spécification de la langue principale

Les textes sont écrits en français médiéval

Spécification technique de la langue principale

profileDesc/langUsage/language/@ident (code iso ‘fro’)

Commentaires sur les spécifications

Sources secondaires

Sources secondaires pouvant faire l'objet de vocabulaires et d'index de recherche spécifiques :

- Langues secondaires
  - latin : *[@xml:lang='lat'], en particulier <foreign>
    - construire un index de recherce spécifique
  - français moderne *[@xml:lang='fr']
    - exclure des indexes
- Titres de sections
  - <head>
    - exclure des indexes, si en français moderne (cf. ci-dessus)
- Notes
  - <note>
    - exclure des indexes
- Corrections orthographiques, Formes régularisées <choice> <sic><corr>; <orig><reg>; <abbr><expan>
  - En cas de <choice>
    - exclure <sic> et <orig> des indexes, afficher en rouge barré dans l’édition
    - inclure <corr> et <reg>, afficher en bleu dans l’édition
  - <supplied> :
    - inclure dans l’index le contenu sans garder trace de la balise
    - afficher des crochets dans l’édition
- Liste, Tableau
  - non
- etc.

Spécification des sources secondaires

Spécification technique des sources secondaires

Commentaires sur les spécifications

Hors-texte

Sources à exclure de la surface textuelle du corpus.

Par exemple, le contenu textuel du teiHeader.

Spécification du hors-texte

L'entête TEI

Spécification technique du hors-texte

teiHeader

Commentaires sur les spécifications

Unités

Unités lexicales

Les unités lexicales forment les unités de base du corpus (les feuilles de l'arborescence du modèle de corpus). Les unités peuvent être codées en XML, en XML-TEI ou pas. Elles peuvent se définir par des types de caractères Unicode et correspondre à des mots simples ou à des mots-composés. Elles peuvent porter des propriétés comme leur catégorie morpho-syntaxique ou leur lemme. Ces propriétés peuvent être enregistrées dans des fichiers annexes au corpus (annotation débarquée ou stand-off).

Spécification des unités lexicales

Spécification technique des unités lexicales

En général, les balises XML signifie une limite d'unité lexicale (voir tokenizer interne par défaut de TXM).

Les balises qui sont internes à un mot portent un attribut (soit @rend, soit @type) ayant la valeur 'word_part'.

Commentaires sur les spécifications

Unités de contextes

Les unités de contexte forment la structure du texte (les nœuds de l'arborescence du modèle de corpus). Ces unités correspondent aux limites de phrases, vers, chapitre, texte, etc. Elles peuvent porter des propriétés comme leur type ou leur numéro.

Spécification des unités de contextes

Spécification technique des unités de contextes

<TEI>, <text>, <body>, <head>, , <p>, <ab>, <q>, <speaker>, <stage>

toute balise qui porte un attribut dont la valeur est 'multi_s'

Commentaires sur les spécifications

Contrastes entre Unités

Au sein de la plateforme, les unités peuvent être comparées sur la base de contrastes. Ces contrastes peuvent être définis à partir des unités de contexte et des unités lexicales. On peut par exemple créer un contraste entre les chapitres d'une œuvre (progression infra-textuelle), entre les dates de discours (progression chronologique) ou entre genres de textes (contraste typologique). On peut également créer un contraste entre les forment ou les lemmes de verbes conjugués au passé, au présent et au futur.

Spécification des contrastes

texte, sous_siècle, domaine, genre, dialecte

Spécification technique des contrastes

voir les métadonnées principales

Commentaires sur les spécifications

Édition

Unité éditoriale

L'affichage des occurrences d'unités lexicales au sein des textes (le retour au texte) se réalise au sein d'une édition de chaque texte composant le corpus. Cette édition peut être segmentée en unités. Ces unités peuvent être définies selon différentes sources ou moyens :

– les pages physiques d'une édition à l'origine de la numérisation et leur disposition ;

– des pages définies par le dispositif d'affichage d'une édition de texte (par exemple un traitement de texte peut proposer une pagination) ;

– des pages composées d'un nombre fixe d'unités lexicales ;

– une unité logique comme une section de texte.

Spécification des unités éditoriales

Il faut générer des éditions monopage et multipages.

L'unité éditoriale est la page d'édition encodée avec '<pb/>'.

Spécification technique des unités éditoriales

– '<pb/>'

– '<lb/>'

Commentaires sur les spécifications

Rendu textuel

L'édition peut être formatée en fonction d'un dispositif d'affichage. Par exemple, une feuille de style CSS peut être associée à une version HTML de l'édition. Les unités lexicales peuvent faire l'objet de formatages particuliers comme l'affichage de leurs propriétés.

Spécification du rendu textuel

On peut sauter les lignes en cas de présence de '<lb/>'.

Les propriétés des unités lexicales sont à afficher en flyover.

Une XSL 1.0 expérimentale existe pour, à partir de la source, produire 1 page HTML incluant page de garde (à partir des métadonnées), mise en forme de discours direct, corrections éditoriales (crochets, italique...), foreign (italique), etc.

Une CSS 1.0 expérimentale existe pour l'édition des pages du Graal.

Spécification technique du rendu textuel

– '<lb/>', etc.

Commentaires sur les spécifications

Références

Les références servent à décrire la localisation des unités au sein des textes. Par exemple, « `[Queste del Saint Graal, folio 160, colonne d, ligne 3]` » peut exprimer la localisation d'une occurrence d'unité lexicale au sein d'un manuscrit médiéval :

– « Queste del Saint Graal » est le titre de l'œuvre ;

– « folio 10 » est la page du manuscrit ;

– « colonne d » est la colonne – a, b pour les deux colonnes du recto du folio ; c, d pour les deux colonnes du verso du folio ;

– « ligne 3 » est la ligne de texte dans la colonne où apparaît l'occurrence.

Les références peuvent être construites à partir de toute information se trouvant dans les sources du corpus. Elles peuvent varier selon le type de texte. Par exemple, on numérote souvent la prose en pages et la versification en n° de vers.

Les références apparaissent à différents endroits, et sous différentes formes, dans la plateforme :

– elles situent les pivots des concordances ;

– elles renseignent les pages d'édition ;

– etc.

Spécification des références

– pour les textes en prose et mixtes (forme) : texte_id, p. n° de page

– pour les textes en vers, indiquer en plus le numéro du vers

Spécification technique des références

– pb/@n

– lb/@n

Commentaires sur les spécifications

Alignement

Certaines unités des textes peuvent être en relation avec des unités d'autres textes - ces textes entretenant une relation de traduction par exemple. L'alignement décrit toutes les relations à prendre en charge dans la plateforme. Par exemple, deux textes peuvent être alignés au paragraphe ou à la phrase près.

Spécification de l'alignement

Spécification technique de l'alignement

Commentaires sur les spécifications

Participation des outils de Traitement Automatique de la Langue (TAL)

Un certain nombre d'outils de TAL externes peuvent participer à la construction automatique ou semi-automatique d'une partie des informations lors de l'importation du corpus. Par exemple :

- la segmentation en unités lexicales ;
- l'étiquetage morpho-syntaxique ;
- la segmentation en phrases ;
- la détection d'entités nommées (personnes, entreprises, lieux, dates, sommes, etc.) ;
- la construction d'unités syntaxiques ;
- etc.

Dans tous les cas d'importation, la plateforme peut offrir des services de segmentation en phrases et en unités lexicales, en tenant compte des différentes dimensions textuelles. Elle intègre déjà, par exemple, les étiqueteurs « TreeTagger » et « TnT ».

Spécification de l'appel d'outils de TAL

Les unités lexicales et les limites de phrases sont calculées par le tokenizer de la plateforme TXM.

Étiquetage des unités lexicales : voir spécifications techniques.

Spécification technique de l'appel d'outils de TAL

Les spécifications des unités lexicales et des phrases de la BFM sont intégrées à l'algorithme par défaut de TXM.

Les unités lexicales sont étiquetées morphosyntaxiquement et lemmatisées par TreeTagger avec le modèle AFR.par (propriété 'ttpos' et 'ttlem').

Les unités lexicales sont également étiquetées morphosyntaxiquement par TreeTagger avec le modèle RGAQCJ.par (propriété 'pos').

Commentaires sur les spécifications

Annexe A – liste des éléments XML du texte xxx