Outil semi-automatique d'aide au balisage de texte pour le projet TMG

Florence Le Priol, Pôle des Systèmes d'Information,
Université Paris-Sorbonne

Plan

- Objectifs
- Choix techniques
- Fonctionnalités
- Démo
- Améliorations

Objectifs

- Construire un outil pour l'aide au balisage des textes
  - Automatiser ce qui peut l'être (nettoyage, balisage des termes connus)
  - Outil d'aide pour le reste (nouveaux termes)
- Répondre aux normes de la TEI.
- multiplateforme (windows, mac os)
- pas nécessairement disponible en ligne
- Installation et utilisation simples y compris pour des non-informaticiens.

Choix techniques

- Pourquoi choisir PYTHON
  - Python est bon dans tous les domaines
  - Il est excellent dans l'analyse des textes et les interfaces utilisateurs
  - Langage jamais bloquant quelque soit les besoins : nombreuses bibliothèques disponibles
  - Interfaçage natif avec de nombreux autres langages
  - Langage bien classé et en progression
  - Utilisé en production dans des développements commerciaux importants

Choix techniques

- Et pourquoi pas JAVA ?
  - Excellent langage
  - Le plus utilisé
- Néanmoins
  - Python est passé en têtes des langages d'apprentissage en Europe et aux Etats-Unis
  - La syntaxe de Python est plus simple et bien plus lisible que celle de Java
  - Python est plus simple à apprendre et à comprendre que Java

Choix techniques

- Distribution
  - Stand-alone
  - Version Mac os
  - Version Windows
  - Utilisation de cx_freeze

Fonctionnalités

- Nettoyage automatique :
  - Suppression des balises issues de la numérisation
  - Utilisation de listes de balises ou d'expressions régulières
- Entrée :
  - Texte tel que récupéré de la numérisation, au format txt (dossier TMG_sources)
  - Liste des balises (TMG_data/nettoyage.txt)
  - Liste des expressions régulières (TMG_data/nettoyageregex.txt)
- Sortie : texte nettoyé, au format txt (dossier TMG_textes)

Fonctionnalités

- Liste de balises

- Forme initiale fixe, Forme souhaitée
- Par exemple

  #.s#.s est

  remplacé

  par ſſ

```
%--B,B̲
%--A,A̲
%--a,a̲
#.sz,ſ
#.si,ſi
#.s#.s,ſſ
#.s,ſ
#.r,r
#.,,„
#;ou,ů
#;eæ,æ̊
{v_er_pu},ṕ
```

- Liste d'expressions régulières

- Forme initiale variable, Forme souhaitée

```
<SE[0-9][0-9]?[0-9]?>,
<filename:[a-zA-Z0-9_.]*/>,
```

- Par exemple

  <SE suivi de 1 à 3 chiffres > est supprimé (remplacé par rien)

Fonctionnalités

Extrait de texte
issu de la numérisation

```
<br type="page"/>
<p>@@0@<filename:TMG_Anonymus_1496_0005.tif/></p>
<p>@@1@<SE5></p>
<p><P>E{v_ee}erpta mu#.sice omnis</P></p>
<p><P>cantus G#.rego#.riani Pita/</P></p>
<p><P>go#.rici {et1$} contrap%-ucti #.simplicis</P></p>
<p/>
<p/>
<br type="page"/>
<p>@@0@<filename:TMG_Anonymus_1496_0006.tif/></p>
<p>@@1@<SE6></p>
<p><tspb>ela</p>
<p>b</p>
<p>dd dd dd ddla#.sol</p>
<p>cc#.solfa</p>
<p>b bbfahmi</p>
<p>b aalamire</p>
<p>g g g g#.solreut</p>
<p>ffaut</p>
<p>b elami</p>
<p>dla#.solre</p>
<p>c c c c#.solfaut</p>
<p>b bfahmi</p>
<p>b alamire</p>
<p>G#.solreut</p>
<p>f f f Effaut</p>
<p>b Elami</p>
<p>D#.solre</p>
<p>Cfaut</p>
<p>b Bmi</p>
```

nettoyé

```
@@0@
@@1@
Eꝛerpta muſice omnis
cantus Gregoriani Pita/
gorici ⁊ contrapūcti ſimplicis



@@0@
@@1@
<tspb>ela
b
dd dd dd ddlaſol
ccſolfa
b bbfahmi
b aalamire
g g g gſolreut
ffaut
b elami
dlaſolre
c c c cſolfaut
b bfahmi
b alamire
Gſolreut
f f f Effaut
b Elami
Dſolre
Cfaut
b Bmi
```

Fonctionnalités

- Balisage automatique :
  - Balisage des termes dont la correspondance est déjà connue
  - Utilisation de la liste des ponctuations
  - Utilisation du dictionnaire

- Entrée :
  - texte nettoyé, au format txt (dossier TMG_textes)
  - Liste des ponctuations (TMG_data/balises_ponctuations.txt)
  - Dictionnaire (TMG_dictionnaires/dico.dico → URL serveur)
- Sortie : texte balisé, au format xml (dossier TMG_textes)

Fonctionnalités

- Liste des ponctuations
- Ponctuation, Balise

```
/,<pc type="S2"> / </pc>
.,<pc type="S1">.</pc>
:,<pc type="S3">:</pc>
,,<pc type="S4">,</pc>
?,<pc type="S5">?</pc>
```

- Par exemple

/

est balisé

<pc type="S2"> / </pc>

- Dictionnaire
- Expression, Balise
- Par exemple

Componiſt

est balisé

<choice n = "gr"><orig>Componiſt</orig><reg>Componist</reg></choice>

Capellmeiſter,<choice n = "gr"><orig>Capellmeiſter</orig><reg>Capellmeister</reg></choice>
Componiſt,<choice n = "gr"><orig>Componiſt</orig><reg>Componist</reg></choice>
Componiſten,<choice n = "gr"><orig>Componiſten</orig><reg>Componisten</reg></choice>
Compoſition,<choice n = "gr"><orig>Compoſition</orig><reg>Composition</reg></choice>
ConcertGeſänge,<choice n = "gr"><orig>ConcertGeſänge</orig><reg>ConcertGesänge</reg></choice>
Conſequentz,<choice n = "gr"><orig>Conſequentz</orig><reg>Consequentz</reg></choice>
Conſonantia,<choice n = "gr"><orig>Conſonantia</orig><reg>Consonantia</reg></choice>
Conſonantiam,<choice n = "gr"><orig>Conſonantiam</orig><reg>Consonantiam</reg></choice>
Conſonantias,<choice n = "gr"><orig>Conſonantias</orig><reg>Consonantias</reg></choice>
Conſonantien,<choice n = "gr"><orig>Conſonantien</orig><reg>Consonantien</reg></choice>
Conſonatias,<choice n = "gr"><orig>Conſonatias</orig><reg>Consonatias</reg></choice>

Fonctionnalités

Extrait de texte nettoyé

balisé

Qm̄ magnos ſtudioſis afferat fructus antiquiſſima illa nobiliſſimaq;
muſicalis ſcientia. quantaue ſit autoritate probata. quātum deniq; lau=
dis honoriſue attigerit. haud facile dictu exiſtimem. Naz oēs fere ſancti
patres: prophete: philoſophi: ac romani pōtifices. aut muſici erant: aut
muſicis delectauere. Jd nimirū ꝑfecto cum nec creditum eſſet quēpiam
fatis eruditum fuiſſe: ni adipe perdulcis cātilene foret enutritus. Quod
et ipi lacedemonij maxima ope ſeruauere. dū apud eos Thaletas Cretē=
ſis gortinus magno precio accitus: pueros diſciplina muſice artis im=
bueret. Quoniā tanta fuit apud eos muſice diligentia: vt eam animos
q̄z obtinere arbitrarentur. ac ſane quidez rati. vulgatū em̄ eſt: q̄ſepe ira
cundias cantilena repreſſerit: q̄multa vel in corporum vel in animor um
affectioni[us)ſ miranda pfecerit. Legīt em̄ eam nonnūq; ſua dulcedine de
monia fugaſſe: Jnſanos quoq; ad rationis vſum: incontinētes ad caſti=
tatem: triſtes ad leticiā: egros ad ſanitatem. inordinatas imaginationes
ad conſtantiam deliberationēq;: pigros ⁊ inutiles ad agibilitatē reduxiſ=
ſe: nec nō etiam melanc oliam tyrannidemq; a magnati[us)ſ depuliſſe: Jdq;
non mō ſeſe in ſingulis vel ſtudijs vel etati[us)ſ tenet: verū ꝑ cuncta diffun
ditur ſtudia. vt infantes. inuenes. ſenes. ⁊ viri ⁊ mulieres. ita naturaliter
affectu quodā ſpontaneo modis muſicis adiunguntur: vt nulla omnino
ſit etas que a cantilene dulcis delectatione ſeiuncta ſit. Nonne etiā illud
manifeſtum eſt in bellum pugnantium animos tubarum carmine incēdi ⁊
ſopitos exitari: vigilantiū quoq; animis cum dulcior eſſet melodia ſopo
rem perſuaderi. Fuit em̄ pitagoricis id in morem diuq; permanſit. vt cū
diuturnas in ſōno reſoluerent curas quibuſdam cantilenis vterentur. vt
eis lenis ⁊ quietus eſſet ſopor. itaq; experecti alijs quibuſdā modis ſtu=
porem ſomni confuſionemq; purgabant Jd nimirū ſcientes ſiquidez eā

```
105 Qm̄ magnos ſtudioſis afferat fructus antiquiſſima illa nobiliſſimaq;
106 muſicalis ſcientia<pc type="S1">.</pc> quantaue ſit autoritate probata<pc type="S1">.</pc> quātum deniq; lau=
107 dis honoriſue attigerit<pc type="S1">.</pc> haud facile dictu exiſtimem<pc type="S1">.</pc> Naz oēs fere ſancti
108 patres<pc type="S3">:</pc> prophete<pc type="S3">:</pc> philoſophi<pc type="S3">:</pc> ac romani pōtifices<pc type="S1">.</pc> aut
      muſici erant<pc type="S3">:</pc> aut
109 muſicis delectauere<pc type="S1">.</pc> Jd nimirū ꝑfecto cum nec creditum eſſet quēpiam
110 fatis eruditum fuiſſe<pc type="S3">:</pc> ni adipe perdulcis cātilene foret enutritus<pc type="S1">.</pc> Quod
111 et ipi lacedemonij maxima ope ſeruauere<pc type="S1">.</pc> dū apud eos Thaletas Cretē=
112 ſis gortinus magno precio accitus<pc type="S3">:</pc> pueros diſciplina muſice artis im=
113 bueret<pc type="S1">.</pc> Quoniā tanta fuit apud eos muſice diligentia<pc type="S3">:</pc> vt eam animos
114 q̄z obtinere arbitrarentur<pc type="S1">.</pc> ac ſane quidez rati<pc type="S1">.</pc> vulgatū em̄ <choice n = "gr"><orig>eſt</
      orig><reg>est</reg></choice><pc type="S3">:</pc> q̄ſepe ira
115 cundias cantilena repreſſerit<pc type="S3">:</pc> q̄multa vel in corporum vel in animor um
116 affectioni[us)ſ miranda pfecerit<pc type="S1">.</pc> Legīt em̄ eam nonnūq; ſua dulcedine de
117 monia fugaſſe<pc type="S3">:</pc> Jnſanos quoq; ad rationis vſum<pc type="S3">:</pc> incontinētes ad caſti=
118 tatem<pc type="S3">:</pc> triſtes ad leticiā<pc type="S3">:</pc> egros ad ſanitatem<pc type="S1">.</pc> inordinatas imaginationes
119 ad conſtantiam deliberationēq;<pc type="S3">:</pc> pigros ⁊ inutiles ad agibilitatē reduxiſ=
120 ſe<pc type="S3">:</pc> nec nō etiam melanc oliam tyrannidemq; a magnati[us)ſ depuliſſe<pc type="S3">:</pc> Jdq;
121 non mō ſeſe in ſingulis vel ſtudijs vel etati[us)ſ tenet<pc type="S3">:</pc> verū ꝑ cuncta diffun
122 ditur ſtudia<pc type="S1">.</pc> vt infantes<pc type="S1">.</pc> inuenes<pc type="S1">.</pc> ſenes<pc type="S1">.</pc> ⁊ viri ⁊ mulieres<
      pc type="S1">.</pc> ita naturaliter
123 affectu quodā ſpontaneo modis muſicis adiunguntur<pc type="S3">:</pc> vt nulla omnino
124 ſit etas que a cantilene dulcis delectatione ſeiuncta ſit<pc type="S1">.</pc> Nonne etiā illud
125 manifeſtum <choice n = "gr"><orig>eſt</orig><reg>est</reg></choice> in bellum pugnantium animos tubarum carmine incēdi ⁊
126 ſopitos exitari<pc type="S3">:</pc> vigilantiū quoq; animis cum dulcior eſſet melodia ſopo
127 rem perſuaderi<pc type="S1">.</pc> Fuit em̄ pitagoricis id in morem diuq; permanſit<pc type="S1">.</pc> vt cū
```

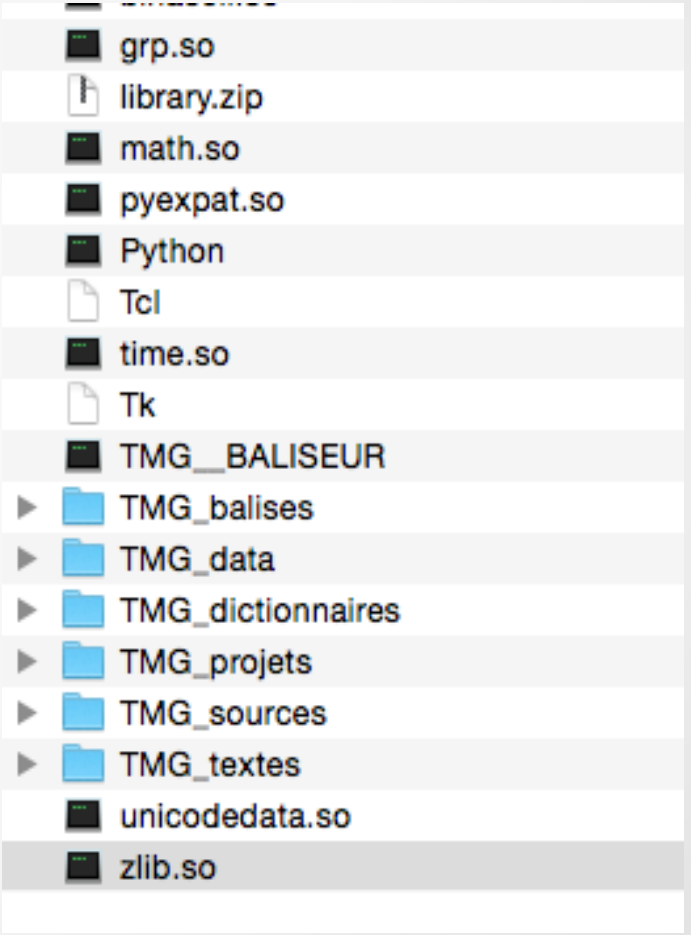
Fonctionnalités

- Balisage semi-automatique
  - Balisage des termes nouveaux (non présent dans le dictionnaire)
  - A l'aide de l'interface de balisage

- Entrée :
  - Texte balisé, au format xml (dossier TMG_textes)
  - Dictionnaire (TMG_dictionnaires/dico.dico → URL serveur)
  - Dictionnaire temporaire (TMG_dictionnaires/dico.dico.temp)
  - Schéma de balises (TMG_balises/balises.bal)
  - Liste des termes comportant des caractères spéciaux (à partir de TMG_data/caracteresspeciaux.txt)
- Sortie : Texte balisé, au format xml (dossier TMG_textes)

Démo

- Lancement : TMG__BALISEUR



Améliorations

- Traitement de tous les schémas de balise
- Possibilité d'individualiser le traitement d'un terme
- Navigation à travers les termes balisés / à baliser
- Rédiger les menus d'aide
- Visuel / ergonomie
- Version windows

- … et en fonction des retours des utilisateurs...