

# *Utilitaire de passage d'un format XML à un format TXT ou WORD*

---

## **TRAVAUX PRATIQUES**

Enseignant

**Jean-Yves ANTOINE**

(Jean-Yves.Antoine AT univ-tours.fr)

# Passage du format XML au format TXT : XML2TXT

## 1. Présentation

Les transcriptions orthographiques fournies par Transcriber sont au format XML. Si ce format favorise l'interopérabilité des ressources, il peut ne pas être adapté à des besoins particuliers. Par exemple, certains utilisateurs peuvent désirer observer leurs corpus sous un éditeur de texte comme Microsoft Word. Là réside l'intérêt du format XML : étant donné qu'il sépare la structure du document de son contenu, on peut à loisir modifier le format de présentation des transcriptions sans changer le contenu. Il suffit pour cela de développer un petit utilitaire (un parseur XML) qui va assurer la transition entre un format de données et un autre.

A titre d'exemple, nous allons utiliser dans ce TP un petit utilitaire (XML2TXT) développé pour les besoins du programme PAROLE PUBLIQUE<sup>1</sup>. Celui-ci transforme une transcription Transcriber en un format ASCII, qui peut ensuite être compilé, par exemple sous format Microsoft Word (figures 1 à 3 ci-dessous). Bien entendu, cet utilitaire répond à un besoin particulier, il vous restera à en développer d'autres pour d'autres besoins.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="Foulon J" audio_filename="060_0000003d" version="1"
version_date="041124">
<Speakers>
<Speaker id="spk1" name="hotesse" check="no" dialect="native" accent=""
scope="local"/>
<Speaker id="spk2" name="client" check="no" dialect="native" accent=""
scope="local"/>
</Speakers>
<Episode>
<Section type="report" startTime="0" endTime="11.952">
<Turn startTime="0" endTime="1.812" speaker="spk1">
<Sync time="0"/>
U B S bonjour
</Turn>
<Turn speaker="spk2" startTime="1.812" endTime="10.032">
<Sync time="1.812"/>
oui bonjour madame j'aurais voulu avoir des renseignements pour e l' l'inscription
en A E S administration Économique et sociale
</Turn>
<Turn speaker="spk1" startTime="10.032" endTime="11.952">
<Sync time="10.032"/>
oui [pi] oui conserver je vais vous passer la personne
</Turn>
</Section>
</Episode>
</Trans>
```

**Figure 1 :** Extrait du corpus Accueil\_UBS : transcription orthographique (format XML)

```
<01> hotesse
h: U B S bonjour
<02> client
c: oui bonjour madame j'aurais voulu avoir des renseignements pour e l'
l'inscription en A E S administration économique et sociale
<03> hotesse
h: oui [pi] oui conserver je vais vous passer la personne
```

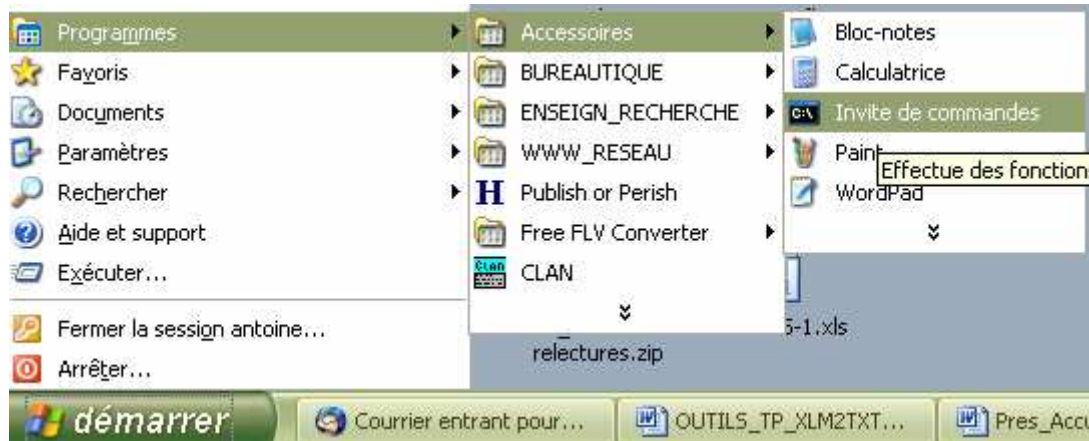
**Figure 2 :** Extrait du corpus Accueil\_UBS : transcription orthographique (format ASCII).

<sup>1</sup> [http://www.info.univ-tours.fr/~antoine/parole\\_publicue/index.html](http://www.info.univ-tours.fr/~antoine/parole_publicue/index.html)

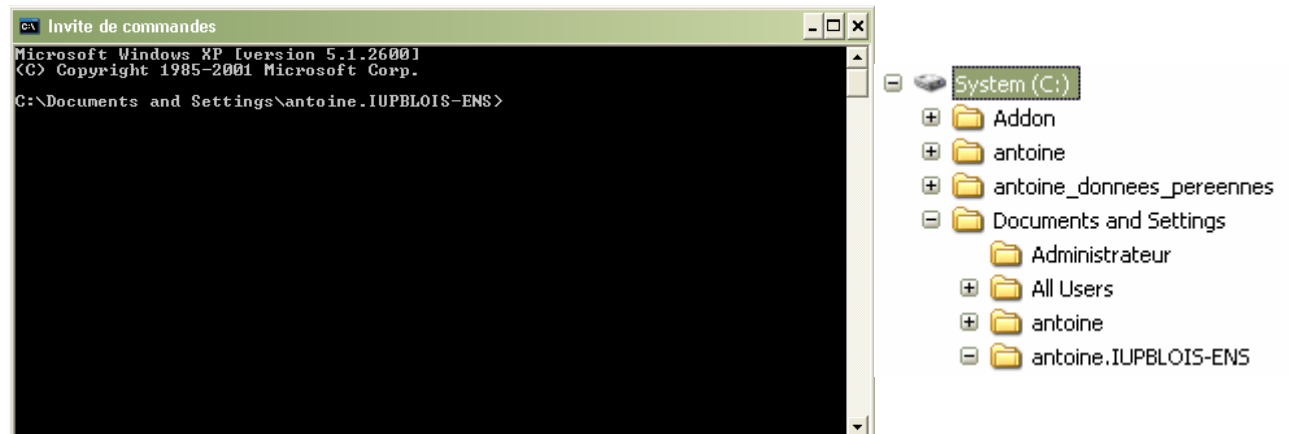
## 2. Passage de XML à ASCII : XML2TXT

Installez tout d'abord l'utilitaire *XML2TXT* dans un répertoire de votre ordinateur. Puis, placez l'ensemble des fichiers que vous désirez transformer en ASCII dans ce répertoire (ou un sous-répertoire si vous le préférez). Cet utilitaire est utilisable sous DOS, un système d'exploitation « caché sous Windows ». Pour l'utiliser, vous allez donc devoir goûter aux délices du contrôle de logiciel en ligne de commandes... comme avant l'invention de Windows ou MacOS.

Tout d'abord, ouvrez une fenêtre d'application DOS, en allant, à partir du menu *Démarrer* de Windows dans le sous-menu *Accessoires* pour choisir l'application *Invite de Commandes*.



Une antique fenêtre de commande s'affiche, dans laquelle vous allez désormais travailler. Sur l'exemple ci-dessous (à gauche), une ligne s'affiche : `C:\Documents and Settings\antoine.IUPBLOIS-ENS>`. Elle me précise où je me trouve dans l'arborescence de mon ordinateur : disque C :, puis sous-répertoire Documents and Settings et enfin sous-sous-répertoire antoine.IUPBLOIS-ENS. Cette présentation en ligne de l'information est totalement équivalente à celle que vous avez l'habitude de parcourir sous l'explorateur Windows (à droite).



Sous DOS, trois commandes (qui doivent être saisies après le chevron > de la ligne de commande dont nous venons de causer) vous permettent de parcourir l'arborescence de votre ordinateur :

- `dir/p` vous permet de voir le contenu du répertoire où vous vous trouvez
- `cd ssrep` vous permet de vous rendre dans le sous-répertoire `ssrep` dont vous venez de visualiser la présence par la commande `dir/p`
- `cd ..` vous permet de remonter au répertoire au-dessus du votre

Il y aurait bien d'autres commandes à étudier, mais celles-ci doivent suffire à vos besoins. En les utilisant à bon escient, rendez-vous dans le répertoire où se trouve l'utilitaire *XML2TXT* et vos fichiers de transcriptions.

Lancez alors l'utilitaire par la commande : `mxl2txt.exe nom_fichier_xml nom_fichier_txt`. Le premier élément de la commande est le nom du programme (constant), tandis que les deux autres éléments sont des éléments variables qui correspondent à vos besoins. Par exemple :

`mxl2txt.exe fichier1.xml fichier1.txt` transforme le fichier XML appelé `fichier1.xml` en fichier ASCII appelé `fichier1.txt`

`mxl2txt.exe *.xml *.txt` transforme tous les fichiers XML de votre répertoire (c'est le sens de l'étoile de la commande) en fichier ASCII du même nom, à la seule différence que l'extension `.xml` des noms de fichier est changée en `.txt`.

1. Lancez précisément cette dernière commande. Que remarquez-vous ?

L'erreur rencontrée par le programme est compréhensible : il ne connaît pas a priori la structure des documents XML que vous lui demandez de traiter. Cette structure est décrite par la DTD Transcriber (`trans-13.dtd`) que nous avons déjà rencontré.

2. Copiez donc ce fichier de DTD dans le répertoire de XML2TXT, puis relancez la commande : normalement, l'utilitaire vous crée autant de fichier ASCII que de fichier XML existants.

Vous pouvez ouvrir ces nouveaux fichiers pour vérifier que vous récupérez bien une écriture simplifiée de vos transcriptions.

### 3. Concaténation de plusieurs fichiers ASCII

Après cette opération de traduction en ASCII, votre corpus n'est peut-être pas encore utilisable comme vous le souhaitez. En effet, il y a autant de fichier TXT que de dialogue transcrits. Pour mettre bout à bout tous vos fichiers TXT dans un seul, il suffit de saisir, toujours sous DOS, la commande suivante :

```
copy *.txt corpus.txt
```

Cette commande copie à la suite tous les fichiers TXT dans un nouveau fichier de nom `corpus.txt`. Dès lors, vous pouvez par exemple ouvrir ce fichier avec Word ou OpenOffice pour en modifier la présentation comme vous le souhaitez.

L'erreur rencontrée par le programme est compréhensible : il ne connaît pas a priori la structure des documents XML que vous lui demandez de traiter. Cette structure est décrite par la DTD Transcriber (`trans-13.dtd`) que nous avons déjà rencontré.