Plateforme TXM - Feature # 3506

Statut:	New	**Priorité:**	Normal
Auteur:	Serge Heiden	**Catégorie:**	Import
Créé :	17/11/2023	**Assigné à:**	
Mis-à-jour :	17/11/2023	**Echéance:**	
Sujet:	import, transcriber, ignore broken texts		

Description

Currently the TRS import module tries to process all the input files all along the different import steps even if they are impossible to process in the earlier steps.

First diagnostics are enough to understand the sources situation and further dummy diagnostics are useless and can confuse the user.

Here are an example session logs:

A) Converting TRS to TEI 162 files

```
% ..Unexpected error while parsing file
file:/C:/Users/ngamo/OneDrive/Documents/new%20ordinateur/documents/Bibiya/CFPRC/CFPRC-BZV07.1-.66.trjs.trs :
javax.xml.stream.XMLStreamException: ParseError at [row,col]:[1,1]
Message: Fin prématurée du fichier.
Location line: 1 character: 1
```

B) Tokenizing 162 files

```
% ..Error : C:\Users\ngamo\TXM-0.8.3\corpora\CFPRC\txm\CFPRC\CFPRC-BZV07.1-.66.trjs.xml
javax.xml.stream.XMLStreamException: ParseError at [row,col]:[2,1]
Message: Fin prématurée du fichier.
at
java.xml/com.sun.org.apache.xerces.internal.impl.XMLStreamReaderImpl.next(XMLStreamReaderImpl.java:652)
at org.txm.scripts.filters.Tokeniser.SimpleTokenizerXml.process(SimpleTokenizerXml.groovy:337)
at org.codehaus.groovy.vmplugin.v8.IndyInterface.fromCache(IndyInterface.java:318)
at org.txm.scripts.importer.transcriber.importer.run(importer.groovy:211)
at org.codehaus.groovy.vmplugin.v8.IndyInterface.fromCache(IndyInterface.java:318)
at org.txm.scripts.importer.transcriber.transcriberLoader.run(transcriberLoader.groovy:171)
at org.txm.groovy.core.GroovyScriptedImportEngine._build(GroovyScriptedImportEngine.java:130)
at org.txm.core.engines.ScriptedImportEngine.build(ScriptedImportEngine.java:57)
at org.txm.objects.Project._compute(Project.java:429)
at org.txm.core.results.TXMResult.compute(TXMResult.java:2540)
at org.txm.core.results.TXMResult.compute(TXMResult.java:2428)
at org.txm.rcp.handlers.scripts.ExecuteImportScript$2.run(ExecuteImportScript.java:176)
at org.eclipse.core.internal.jobs.Worker.run(Worker.java:63)
Failed to tokenize C:\Users\ngamo\TXM-0.8.3\corpora\CFPRC\txm\CFPRC\CFPRC-BZV07.1-.66.trjs.xml
```

C) Building 162 XML-TXM files

```
% ..Import terminé.
```

Analysis 1

From logs in step A) we understand that file CFPRC-BZV07.1-.66.trjs.xml is broken and that we cannot process it.

So in step B) we already know that we cannot tokenize the file: it is not useful to try to do it and misleading to the user.

Finally in step C) we can see that the total number of files processed is still the whole files (162) even if we know that at least one cannot

be processed.

Solution to analysis 1

Remove every input text, from the files to be imported in the next step, that cannot be processed at any import process step.

Analysis 2

From the following diagnostic:

> % ..Unexpected error while parsing file file:/C:/Users/ngamo/OneDrive/Documents/new%20ordinateur/documents/Bibiya/CFPRC/CFPRC-BZV07.1-.66.trjs.trs : javax.xml.stream.XMLStreamException: ParseError at [row,col]:[1,1]
> Message: Fin prématurée du fichier.
> Location line: 1 character: 1

We can understand that the file is probably empty.

This can happen when input files come from batch conversion processes.
So if an input file is empty we should:

- directly diagnose that it is empty
- not try to process it
- ignored it in further import steps

Solution to analysis 2

Test if each input file is empty and ignore it if it is the case.

Generalization to these Solutions

Apply Solution 1 & 2 to all import modules.

Historique