Plateforme TXM - Feature # 3520

Statut:	New	**Priorité:**	Normal
Auteur:	Serge Heiden	**Catégorie:**	TAL
Créé :	29/11/2023	**Assigné à:**	
Mis-à-jour :	29/11/2023	**Echéance:**	
Sujet:	Import, TreeTagger, upgrade TreeTagger options		

Description

Currently, TXM tokenises itself and calls tree-tagger directly. The TXM tokenizer separates clitics as the standard TreeTagger one, depending on language.
But TreeTagger parameter files often also depend on two additionnal lexicons related to tokenization to work properly (for example for **Spoken French**, **Old French** or **Spanish**):

- abbreviations: a list of abbreviations for a language
- mwls: a list of multi-tokens words for a language (called 'multi-words')

The full TreeTagger - Perl based - workflow is, for example for Spanish:

```
utf8-tokenize.perl -a spanish-abbreviations $* |
mwl-lookup.perl -f spanish-mwls |
tree-tagger -token -lemma -sgml spanish.par
```

Tools used

- utf8-tokenize.perl: standard TreeTagger tokenizer
- mwl-lookup.perl: merges multi-token words

New lexicons used

- spanish-abbreviations:

```
Ref.
Vol.
etc.
App.
Rec.
```

- spanish-mwls:

```
A diferencia de
A diferencia del
A fin de
A lo largo de
A medida que
A menudo
A partir de
A pesar de
...
```

To be compatible with certain parameter files (Spoken French, Old French, Spanish...) we need to implement the abbreviations and mwls algorithms and use their lexicons.

Solution 1

- add management between language names and abbreviations and mwls lexicons
- add abbreviations processing to TXM tokenizer
- implement mwl-lookup.perl multi-token processing

Solution 2

- add management between language names and abbreviations and mwls lexicons
- add a Perl processor
- call utf8-tokenize.perl and mwl-lookup.perl directly

Historique