

# Création d'une mémoire TMX

- Objet : mélanger (merge) une source et sa traduction
- Niveau requis :  
[débutant](#), [avisé](#)
- Commentaires : *la débrouille sans le web*
- Débutant, à savoir : [Utiliser GNU/Linux en ligne de commande, tout commence là !](#) 😊

## Introduction

Le but est de créer une mémoire TMX, utilisée par OmegaT, à partir d'un fichier source.txt (chez phlinux des textes en anglais) et de sa traduction obtenue par un traducteur automatique (genre Google) et mise dans un fichier trad.txt.

## Installation

Paquets hors du commun requis : ssed

## Utilisation

Un pack avec 1 script bash et 2 fichiers de règles de sentences utilisés par ssed. Les fichiers source.txt et trad.txt sont placés dans le même répertoire et dans lequel on lance le script (omegat\_tmx.sh).

La commande de lancement sera:

```
omegat_tmx.sh source.txt trad.txt
```

Le script principal :

[omegat\\_tmx.sh](#)

```
#!/bin/bash

# Penser à adapter les chemins des fichiers des règles de sentence :
~/.script/sentEN.sed et ~/.script/sentFR.sed

source=$1
trad=$2

nom=`sed 1q $source | tr -d " "`

# application règles de sentence sur la source + génère fichier
ssed -R -f ~/.script/sentEN.sed < $source > /tmp/source
```

```
sed -i -e '/^$\|^s$/d' /tmp/source

# application règles de sentence sur la trad + génère fichier
ssed -R -f ~/.script/sentFR.sed < $trad > /tmp/trad
sed -i '/^$/d' /tmp/trad

# compte le nombre de lignes
nb_source=`awk 'END{print NR}' /tmp/source`
nb_trad=`awk 'END{print NR}' /tmp/trad`

:> $PWD/omegat_merge_dif.txt
echo "sentences de la source : "$nb_source >> $PWD/omegat_merge_dif.txt
echo "sentences de la traduction : "$nb_trad >>
$PWD/omegat_merge_dif.txt

# boucle 1
ext=1

# lecture source ( apparemment inutile)
#sentEN=`sed r /tmp/source`

# création du fichier tmx
mergetmx()
{
oldifs=$IFS
IFS='
'
# création fichier de numération des sentences trad
for i in `seq 1 $nb_trad`
do
echo tab$i="\$(cat /tmp/trad | awk 'NR == "$i" {print}')) ;
echo \$"tab$i
done > /tmp/sent_var.txt
while read line; do eval $line; done < /tmp/sent_var.txt > /dev/null
# création des balises d'en tête
echo "<?xml version=\"1.0\" encoding=\"UTF-8\"?>"
echo "<!DOCTYPE tmx SYSTEM \"tmx11.dtd\""
echo "<tmx version=\"1.1\""
echo "<header creationtool=\"OmegaT\" o-tmf=\"OmegaT TMX\"
adminlang=\"EN-US\" datatype=\"plaintext\"
creationtoolversion=\"3.1.0\" segtype=\"sentence\" srclang=\"EN-US\"/>"
echo "<body>"
echo "<!-- Default translations -->"
# création des segments
for elmt_source in $(cat /tmp/source); do
echo "<tu>"
echo "<tuv lang=\"EN-US\""
echo "<seg>$elmt_source</seg>"
echo "</tuv>"
echo "<tuv lang=\"FR-FR\" creationid=\"ph\" creationdate=\"\">"
```

```

        for n in $nb_trad
        do
            sentFR="echo \${tab}$ext"
            echo "<seg>`eval $sentFR`</seg>"
        done
        echo "</tuv>"
        echo "</tu>"
        let "ext+=1"
    done
    # création des balises de fin
    echo "<!-- Alternative translations -->"
    echo "</body>"
    echo "</tmx>"

IFS=$oldifs
}

mergetmx > $nom.tmx

if [ $nb_source != $nb_trad ]; then
    echo "Différence du nombre de sentences"
    echo "source: $nb_source"
    echo "traduction: $nb_trad"
else
    echo "Source et Traduction correspondent"
    tail $nom.tmx
fi

exit

```

Le fichier des règles de sentence de la source (ici de l'anglais):

[sentEN.sed](#)

```

#!/bin/sed -f

## utilisé par ssed -R
## Omegat: regex pour texte en anglais

s/(?<![A-Z])\.\s/\.\n/g
s/\?\s/\?\n/g
s/\?'s/\?'n/g
s/\?"s/\?"n/g
s/\b\.\s/\.\n/g
s/\b\.'s/\.'n/g
s/\b\."s/\."n/g
s/\b\!s/\!\n/g
s/\b\.'s/\.'n/g
s/\b\.\s/\.\n/g

```

```
s/\b\.\]\s/\.\]\n/g
```

Le fichier des règles de sentence de la traduction (ici du français):

[sentFR.sed](#)

```
#!/bin/sed -f

## utilisé par ssed -R
## Omevat: regex pour texte en français

s/(?<![A-Z])\.\s/\.\]\n/g
s/\?\s/\?\n/g
s/\?''\s/\?''\n/g
s/\?"'\s/\?"'\n/g
s/\b\.\.\s/\.\.\n/g
s/\b\.\.'\s/\.\.'\n/g
s/\b\.\!''\s/\.\!''\n/g
s/\b\.\!''\s/\.\!''\n/g
s/\b\.\!''\s/\.\!''\n/g
s/\b\.\.'\s/\.\.'\n/g
s/\b\.\.\.\.'\s/\.\.\.\.'\n/g
s/\b\.\.)\s/\.\.)\n/g
s/\b\.\.]\s/\.\.]\n/g
```

Pour que la mémoire TMX soit “parfaite” il faudra obtenir le message suivant : Source et Traduction correspondent. Sinon ce sont les nombres de sentences qui seront affichés. Il faudra alors faire des corrections dans les fichiers source.txt et/ou trad.txt, mais aussi peut être dans les règles de sentences (selon les langues utilisées).

From:  
<http://debian-facile.org/> - **Documentation - Wiki**

Permanent link:  
<http://debian-facile.org/utilisateurs:phlinux:tutos:omegat-merge-de-la-source-avec-la-traduction>

Last update: **02/01/2016 19:29**

