Traducció amb MTradumàtica

[Grup Tradumàtica]

Gökhan Dogru	Adrià Martín
Ramon Piqué	Pilar Sánchez-Gijón
Olga Torres-Hostench	

Facultat de Traducció i d'Interpretació, 5 de juliol de 2019

SESSIONS TRADUMÀTIQUES

Objectiu:

Entendre i gestionar MTradumàtica, un sistema de traducció automàtica estadística (TAE) basat en Moses.

1. En què consisteix un sistema de TAE?
2. Quines opcions tinc?
3. Si decideixo tenir el meu propi sistema, com ho he de fer?
   3.1 Procés d'entrenament d'un motor TAE
   3.2 Procés d'instal·lació d'MTradumàtica
   3.3 Treball amb MTradumàtica
4. Ho intentem?

1. En què consisteix un sistema TAE?

A partir d'una gran quantitat de text en la llengua de partida i en la llengua d'arribada, alineat i no alineat, el sistema cerca la millor traducció possible de cadascun dels elements d'una oració, partint de les agrupacions de paraules (*n-grams*) més grans possibles a les més petites (la paraula).

2. Quines opcions tinc?

- Utilitzar sistemes de TA (TAE o altres sistemes) ja existents
- Crear un motor dins d'un sistema que s'ofereix (generalment de pagament)
- Construir-se un sistema propi

2. Quines opcions tinc?

Construir-se un sistema propi

AVANTATGES:

- Adaptabilitat del corpus
- Confidencialitat i seguretat: accés restringit a arxius i corpus

2. Quines opcions tinc?

Construir-se un sistema propi:

INCONVENIENTS:

- Tecnologia
  Coneixements de programació? Linux? Gestió d'un servidor?
- Cost
  Servidor? Aspectes de seguretat?
- Interacció amb eines TAO
  Amb quina TAO? Com s'integren?
- Gestió d'arxius:
  Amb quins arxius treballo? On els puc obtenir? Com els gestiono?

3. Si decideixo tenir el meu propi sistema, com ho he de fer?

MTradumàtica és una plataforma web basada en Moses amb interfície gràfica d'usuari (GUI) per a la creació de motors de TAE personalitzats (Martín-Mor i Piqué, 2017).

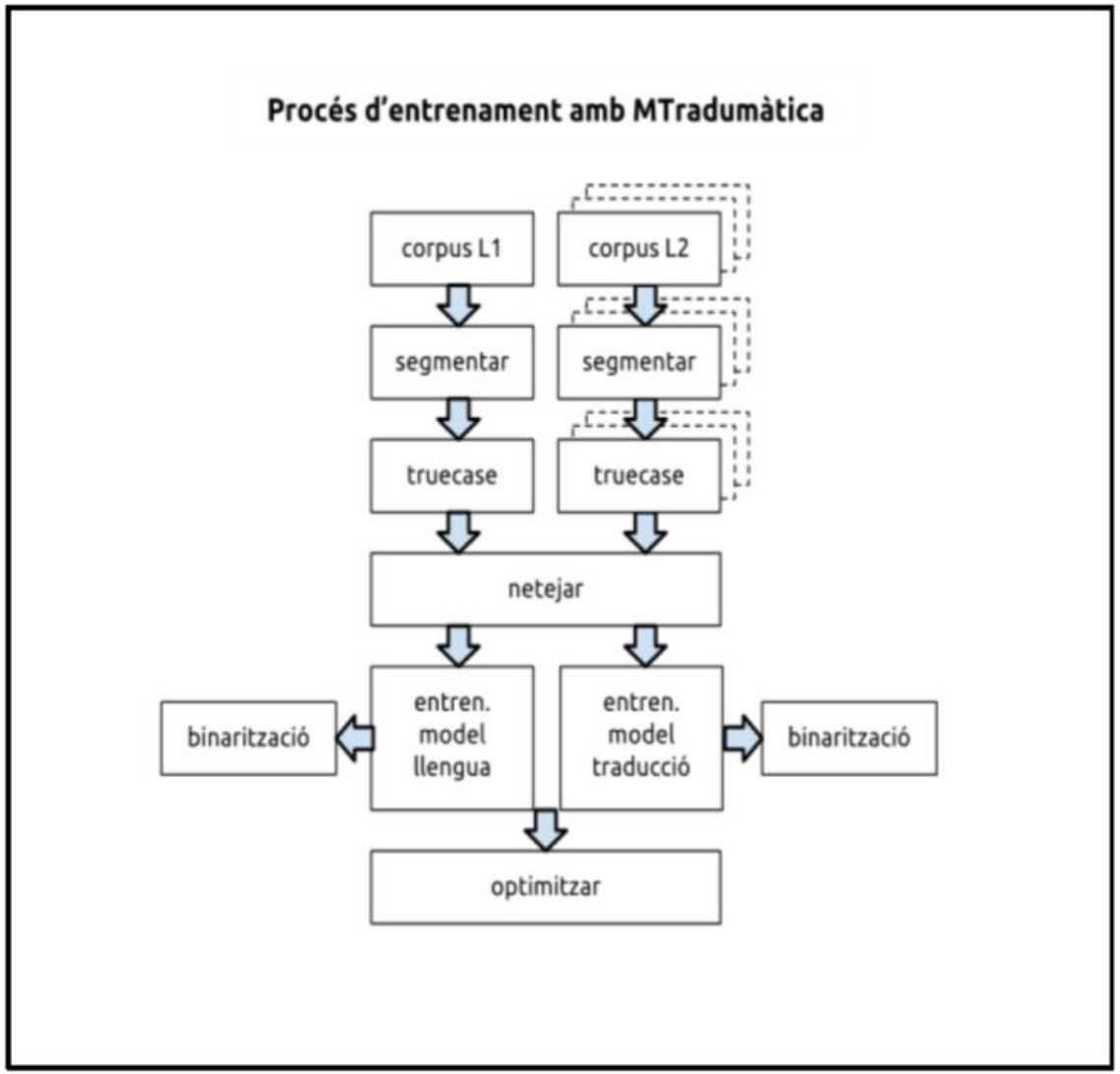
- Desenvolupat amb ProjecTA (FFI2013-46041-R) i ProjecTA-U (FFI2016-78612-R) del Ministeri d'Economia i Competitivitat.
- Llicència GPL (programari lliure), per a Linux; Windows i Mac mitjançant VirtualBox.

3. Si decideixo tenir el meu propi sistema, com ho he de fer?

Per treballar des d'un ordinador personal amb MTradumàtica, el procés és el següent:

1. Instal·lar el sistema MTradumàtica per crear i gestionar motors de TAE.
2. Crear un motor a partir de TMX propis de l'empresa o de corpus públics (Opus Corpus).
3. Incorporar la TA en el flux de treball:
    1. Obtenir un TMX amb la TA i incorporar-lo a l'eina TAO.
    2. Vincular MTradumàtica amb OmegaT.
4. Avaluar el resultat mitjançant mètriques automàtiques.

3.1 Procés d'entrenament d'un motor TAE



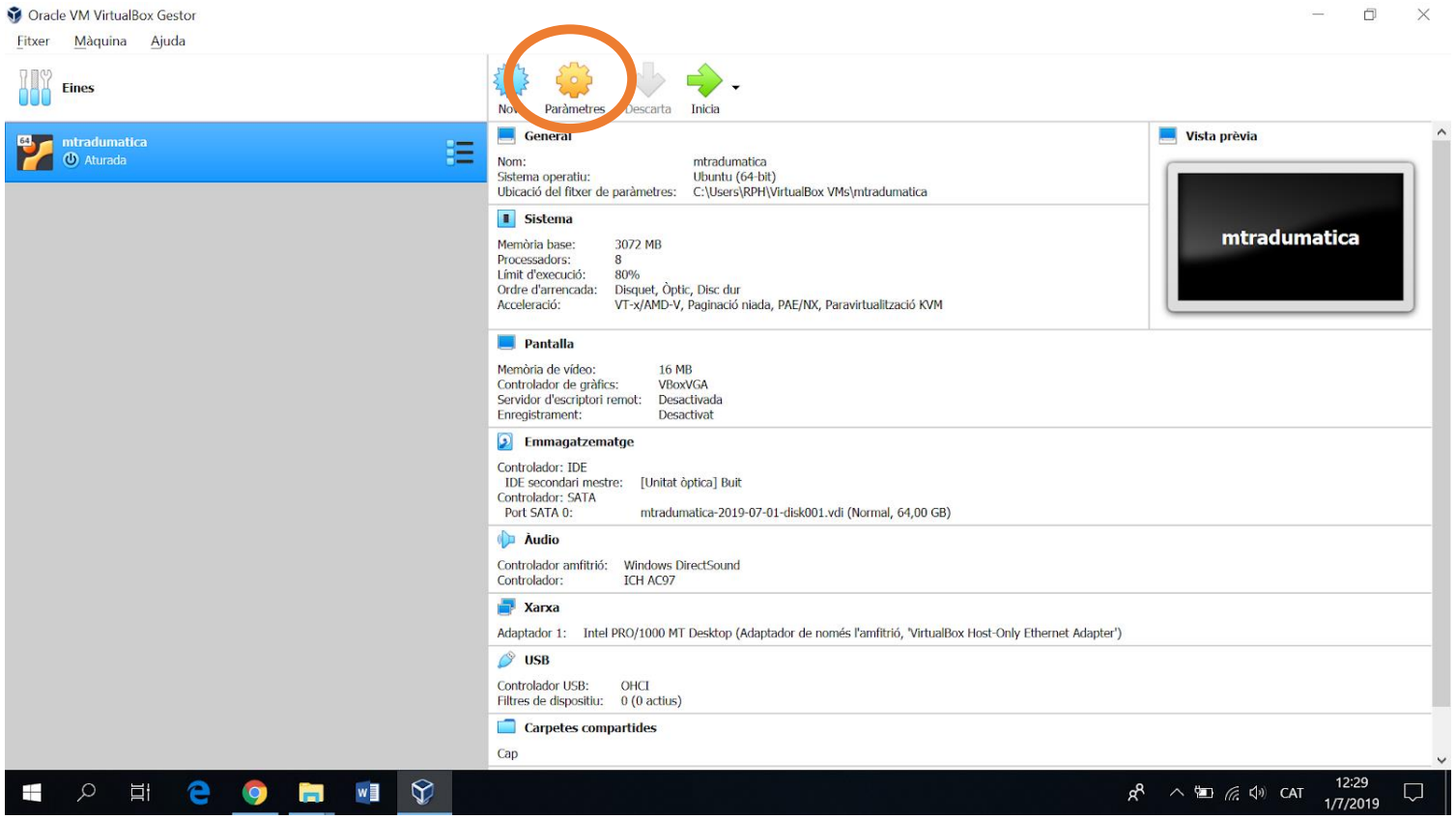
3.2. Procés d'instal·lació d'MTradumàtica

1. Instal·lar VirtualBox (adjunt al llapis de memòria)
2. Importar MTradumatica.ova (adjunt al llapis de memòria)
3. Canviar la configuració del motor (v. diapositives següents)
4. Iniciar el sistema (v. diapositives següents)
5. Generar la URL per treballar mitjançant el navegador

3.2. Procés d'instal·lació d'MTradumàtica

3. Canviar la configuració del motor:

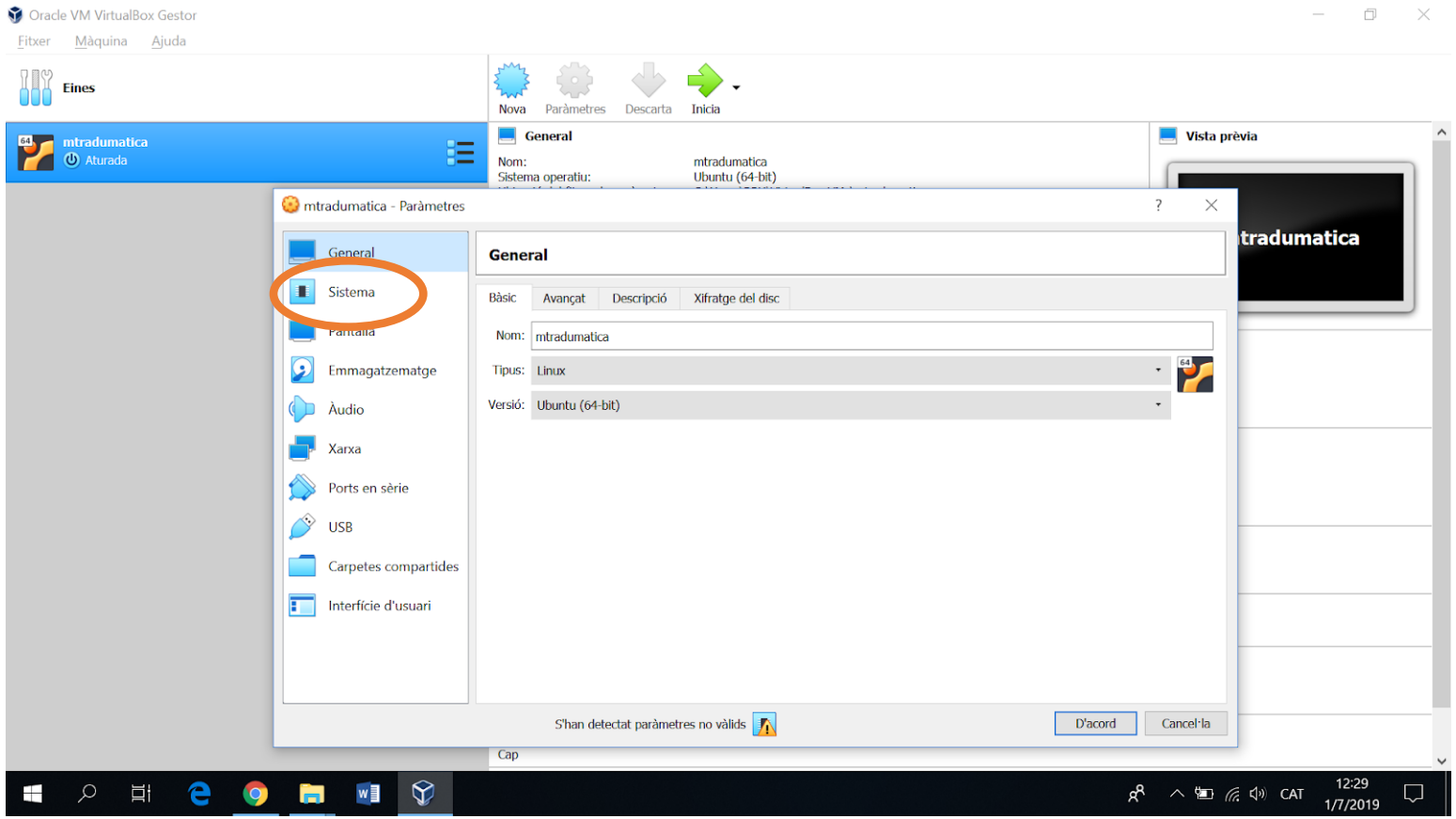
Paràmetres | Sistema:



3.2. Procés d'instal·lació d'MTradumàtica

3. Canviar la configuració del motor:

Paràmetres | Sistema:

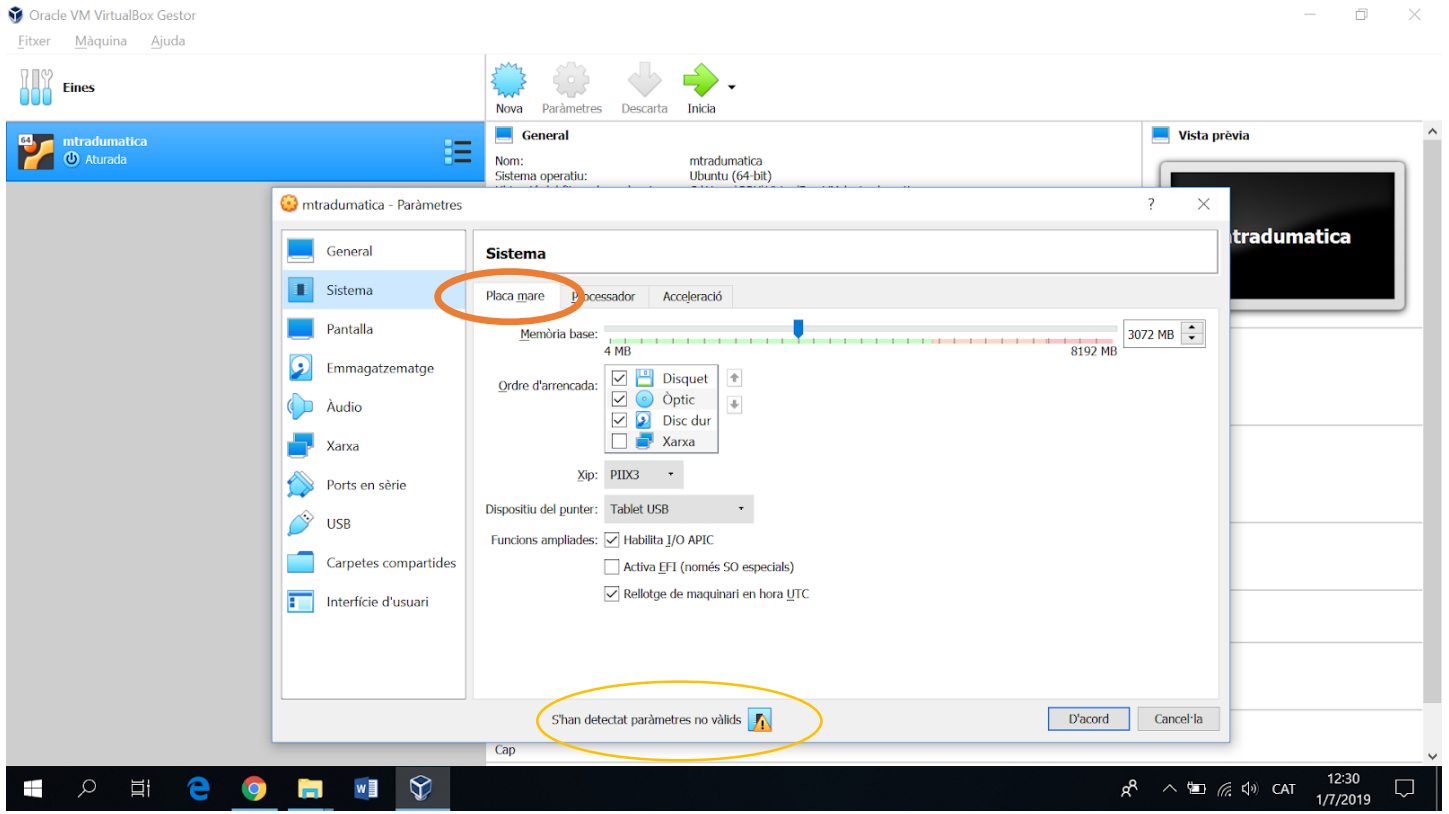


3.2. Procés d'instal·lació d'MTradumàtica

3. Canviar la configuració del motor:

Paràmetres | Sistema:

- **Placa mare**: a "Memòria base", indicar el valor màxim en franja verda que permeti l'ordinador.



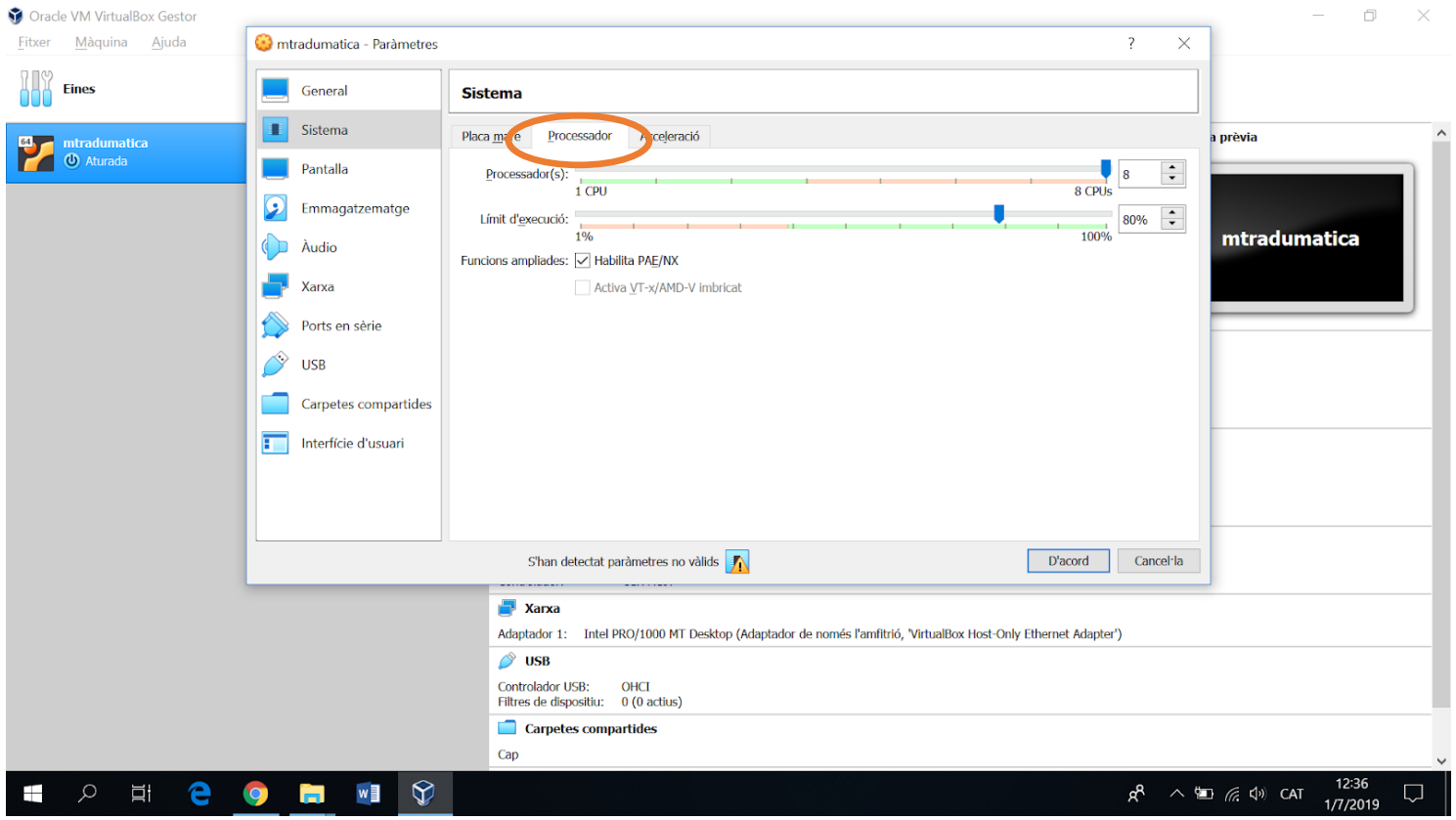
Evitar que aparegui el missatge de paràmetres no vàlids.

3.2. Procés d'instal·lació d'MTradumàtica

3. Canviar la configuració del motor:

Paràmetres | Sistema:

- **Processador**: a la barra "Processadors", indicar el nombre de CPU màxim possible. Si es posen tots els processadors disponibles, cal posar un límit d'execució de 80% o 90% per evitar bloquejar l'ordinador.

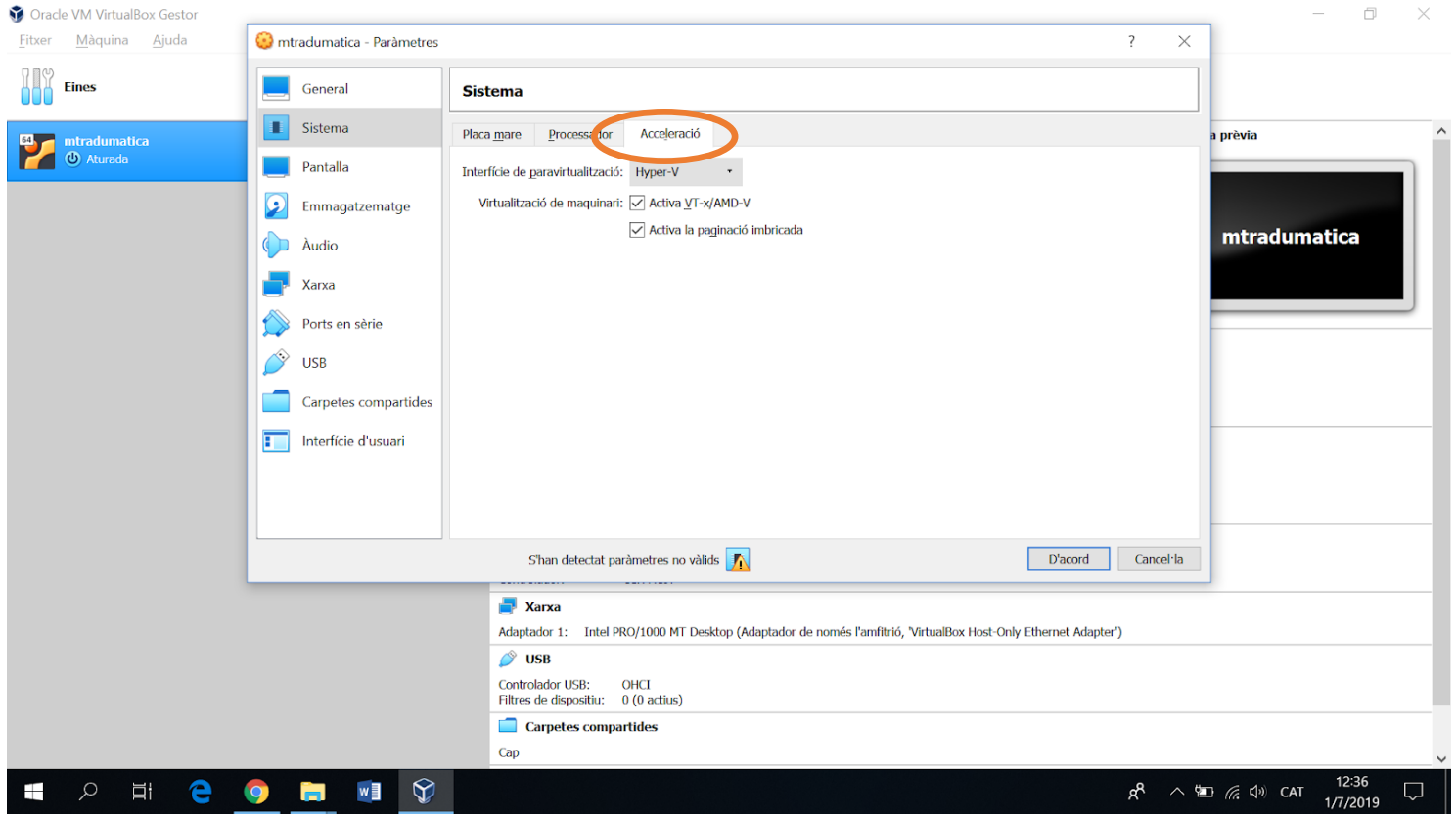


3.2. Procés d'instal·lació d'MTradumàtica

3. Canviar la configuració del motor:

Paràmetres | Sistema:

- **Acceleració**: a "Interfície d'acceleració", indicar Hyper-V si la configuració de l'ordinador ho permet. Si no, deixar l'opció per defecte.

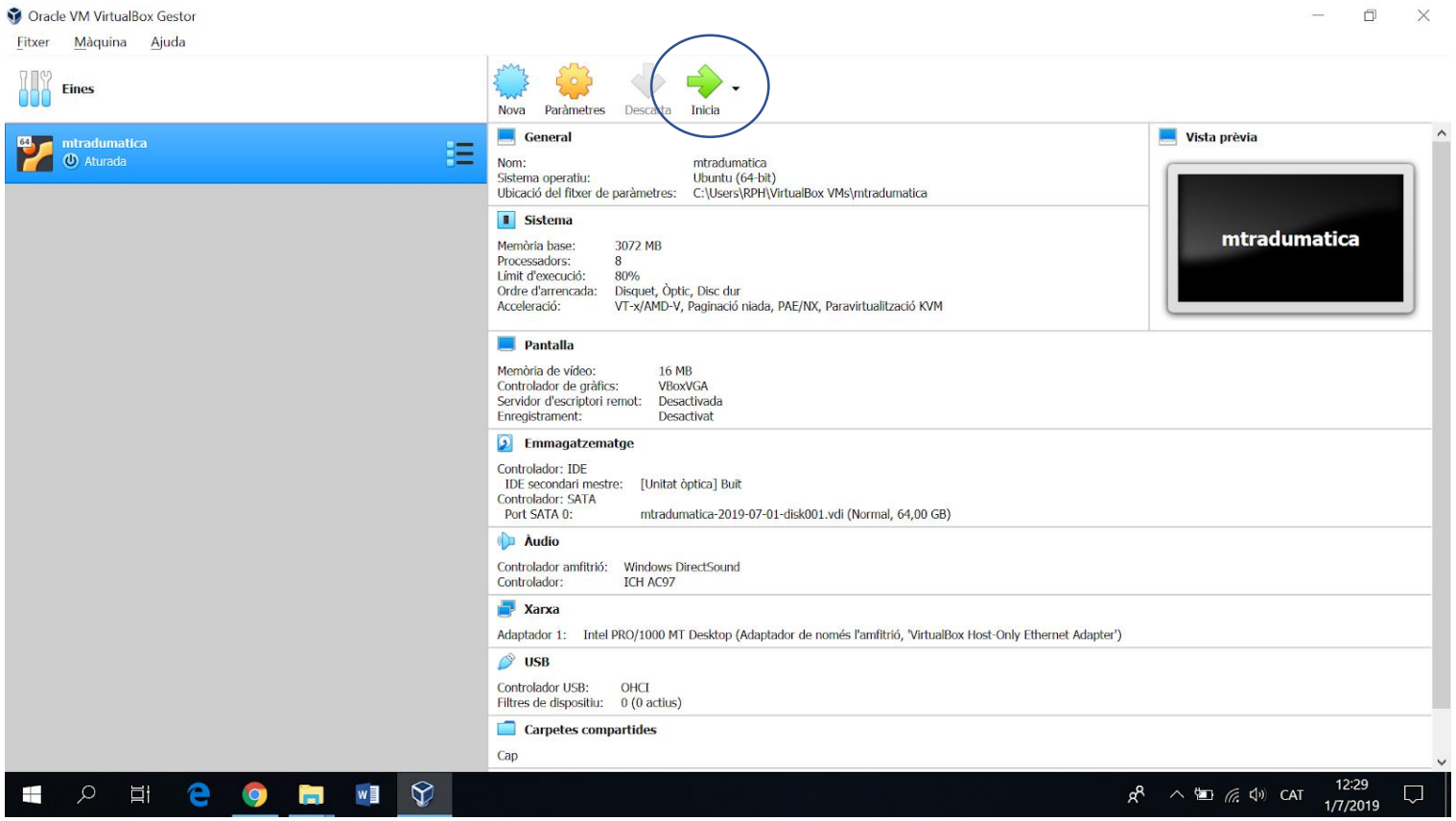


3.2. Procés d'instal·lació d'MTradumàtica

4. Iniciar el sistema:

Clic a la fletxa "Inicia".

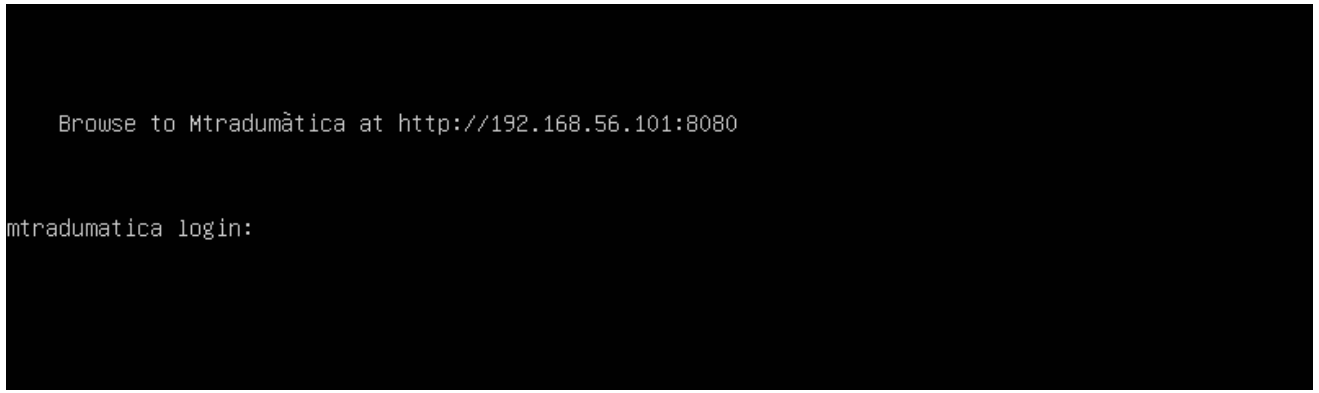
Si dona un error és probable que calgui introduir alguna modificació a la BIOS de l'ordinador. Les instruccions corresponents es troben al llapis de memòria que us hem lliurat, a la carpeta 01-Arxius d'instal·lació.



3.2. Procés d'instal·lació d'MTradumàtica

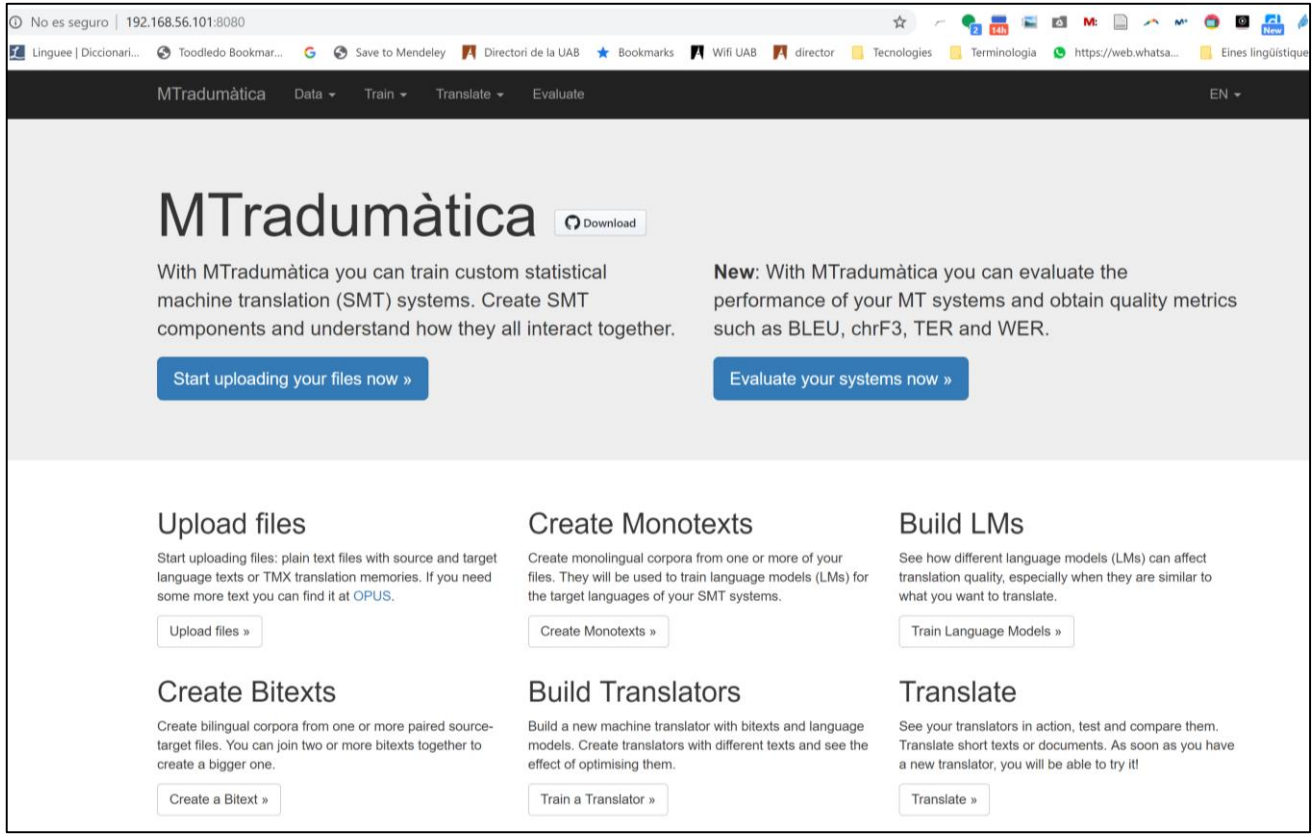
5. Generar la URL per treballar mitjançant el navegador

- Després de fer certes operacions que poden trigar un minut aproximadament, apareix la URL d'accés al sistema que cal copiar a la barra del navegador.
- No cal indicar cap usuari.
- Cada ordinador pot generar una URL diferent.

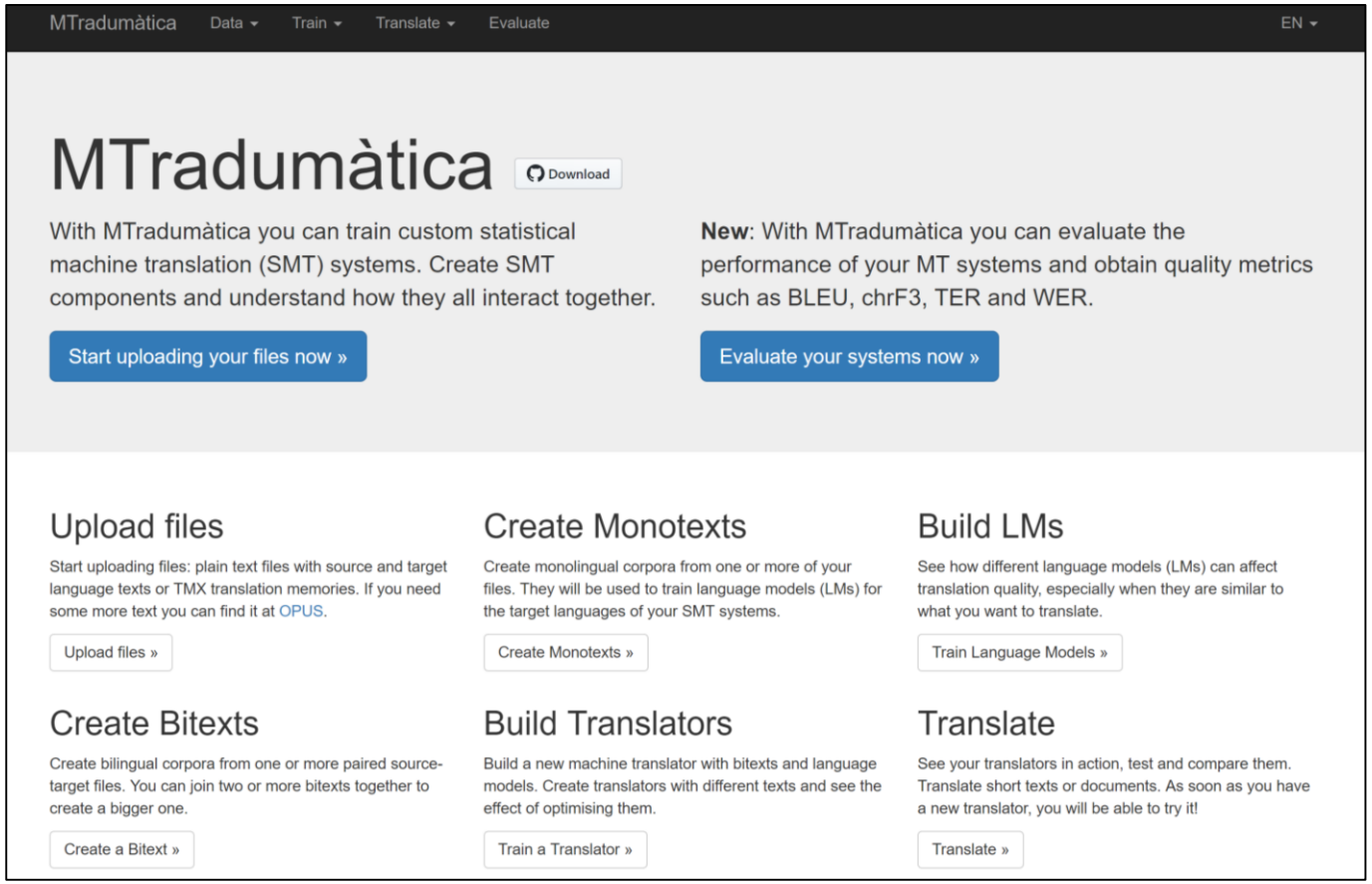


3.3 Treball amb MTradumàtica

Introduir la URL en el navegador:

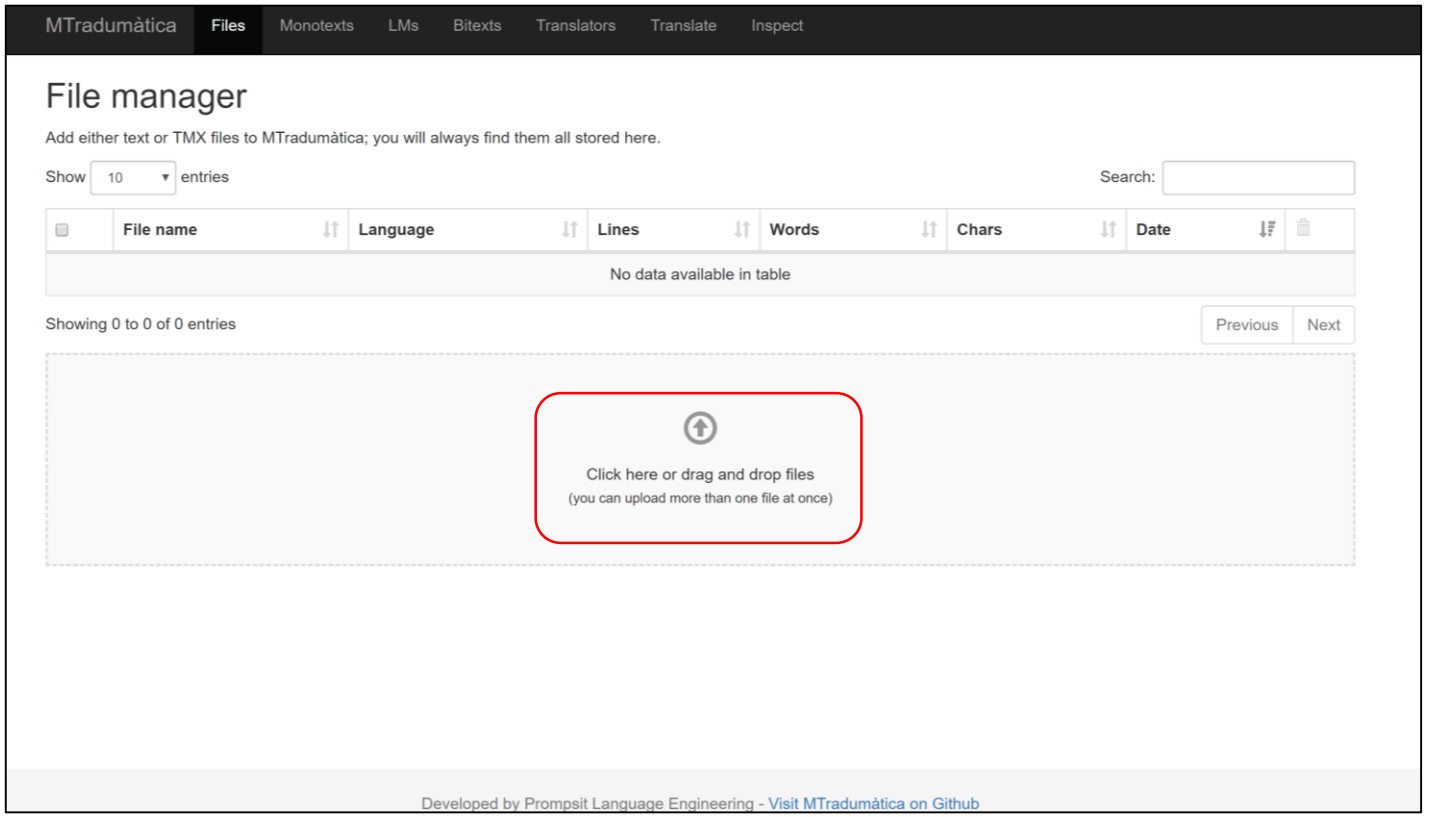


3.3 Treball amb MTradumàtica



1. Upload files: pujar arxius monolingües o bilingües per crear monotextos i bitextos.
2. Create Monotexts: crear conjunts de textos en una de les llengües de cada motor.
3. Build LMs: construir model de llengua en la llengua d'arribada del motor.
4. Create Bitexts: construir conjunts de textos paral·lels.
5. Build Translators: crear i gestionar motors.
6. Traduir.
7. Avaluar la traducció.

3.3.1 Upload files



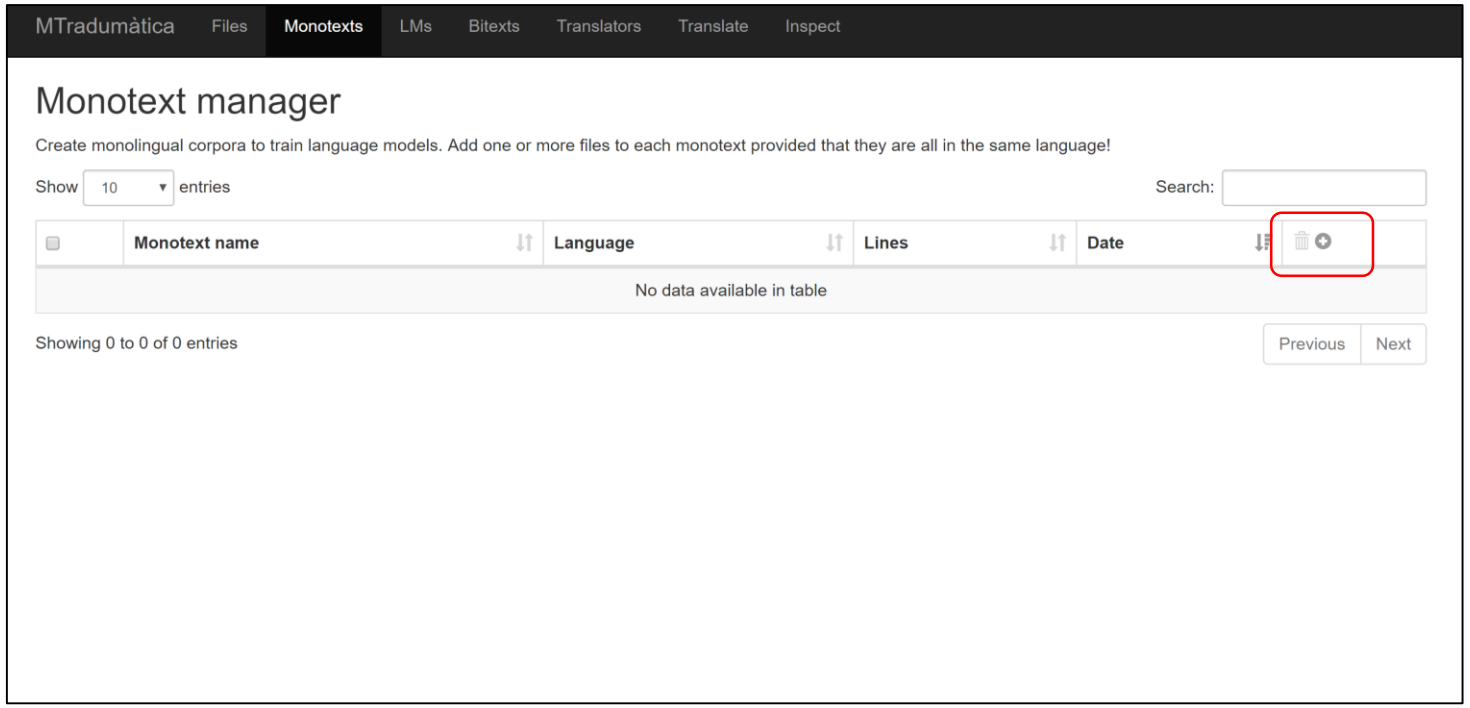
Formats dels arxius que es poden pujar:

- Txt (monolingües)
- Txt (alineats per Moses)
- TMX

Els arxius alineats se separen per llengües automàticament. Cada arxiu TMX o txt alineat dona lloc a dos arxius, un en la llengua de partida i un altre en la llengua d'arribada.

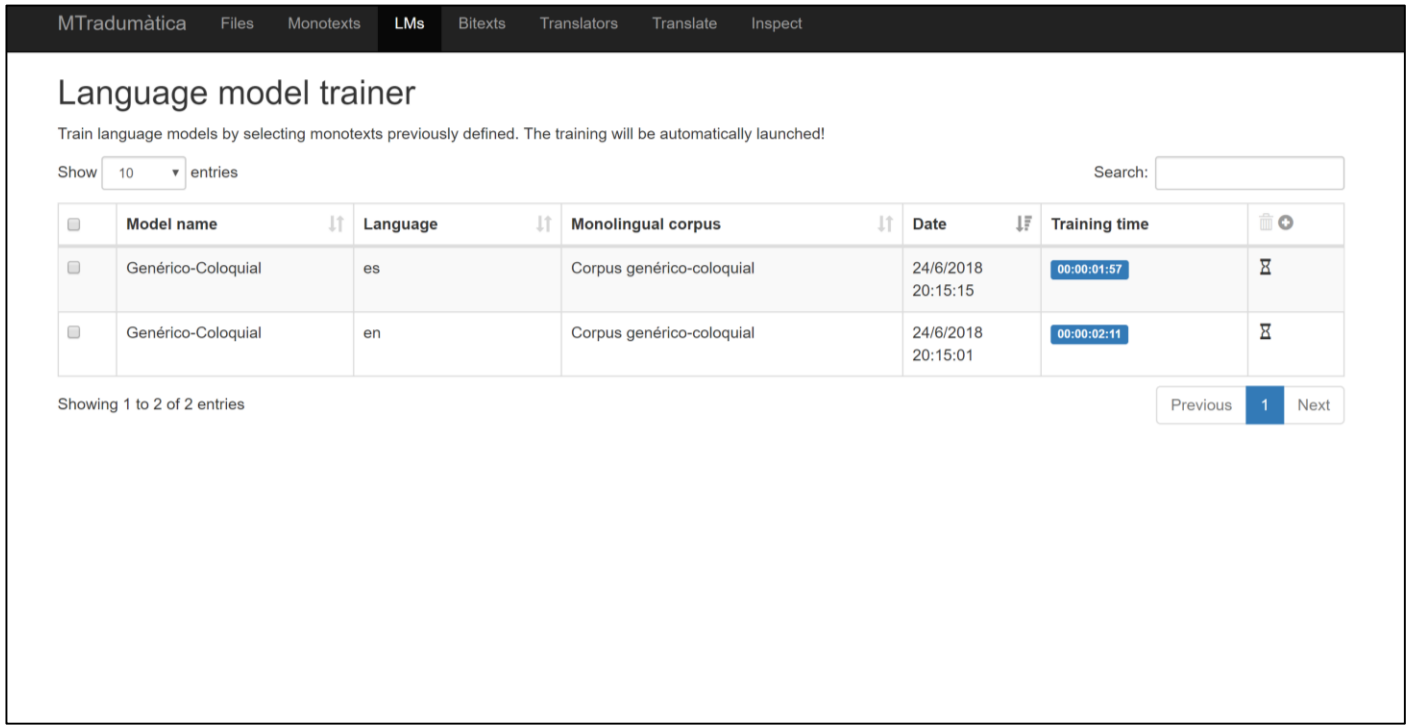
Atenció!!!! No pugeu arxius superiors a 2Gb. Es recomanable pujar arxius de no més de 0,5 Gb.

3.3.3 Monotext Manager



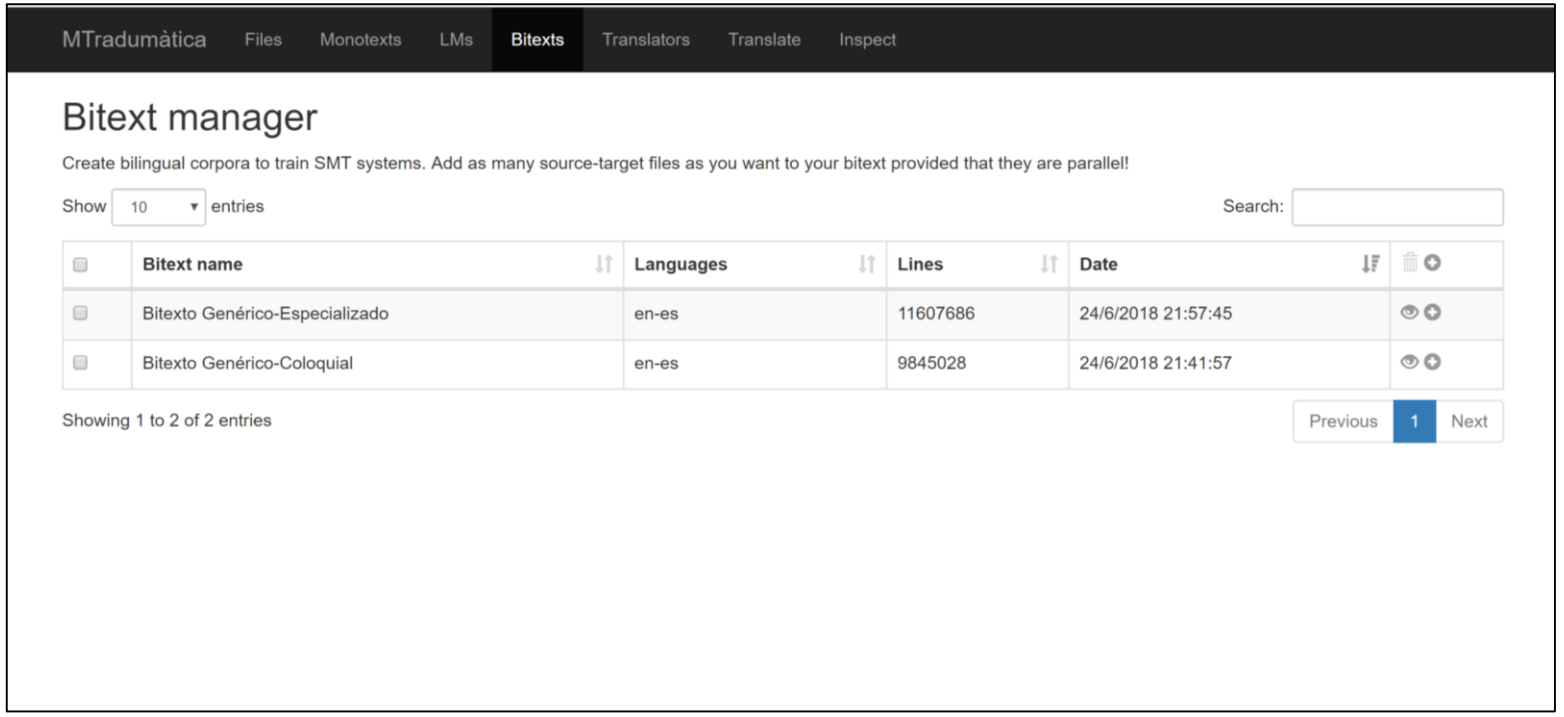
Crear corpus monolingües per generar els models de llengua. Només cal crear monotextos per a les llengües d'arribada dels motors, les que necessiten un model de llengua.
Primer es crea el monotext buit i després s'afegeixen tants textos com es vulguin (dels que hem carregat en la pantalla anterior).

3.3.4 Language Model trainer



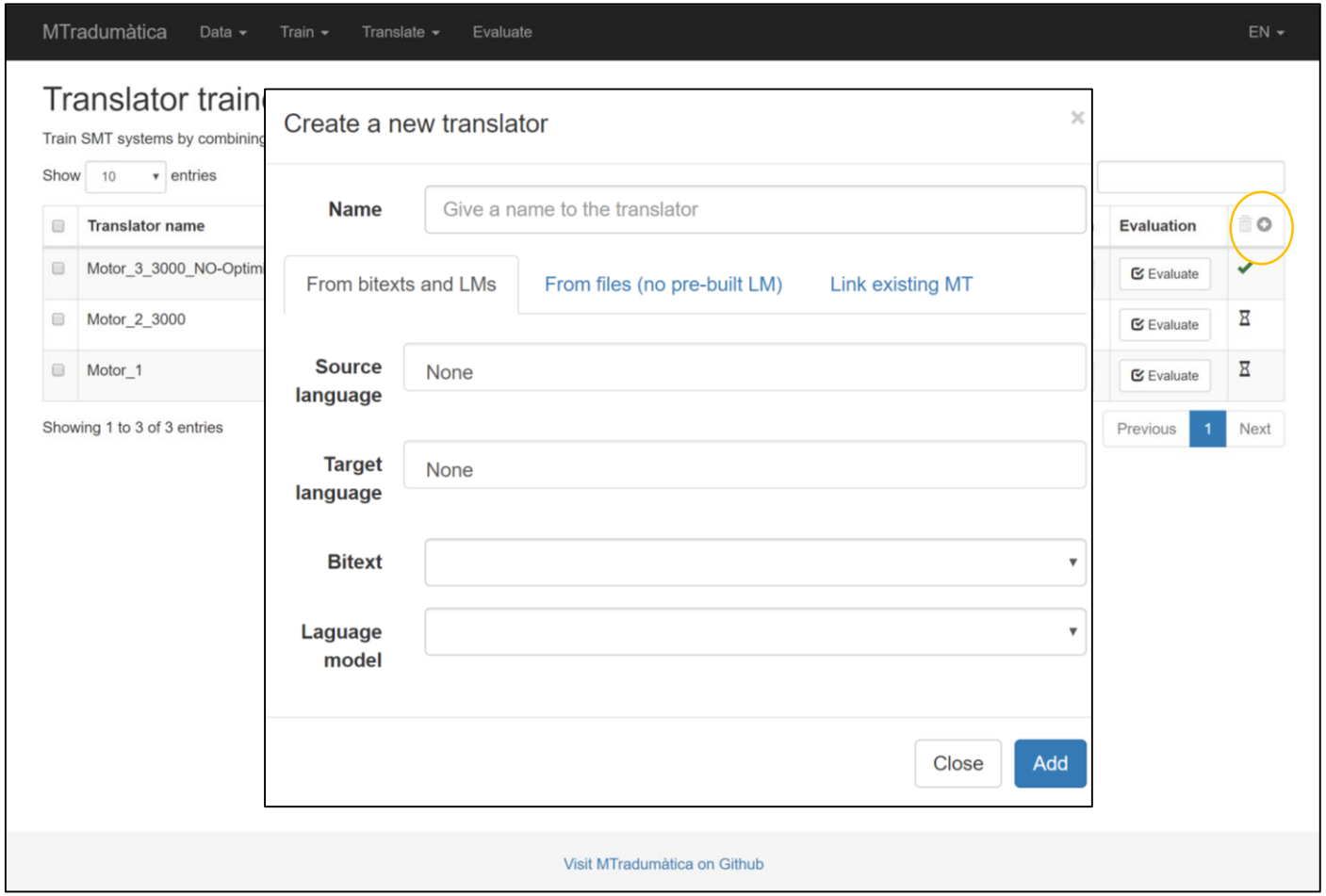
- Entrenament del model de llengua de la llengua d'arribada a partir d'un monotext.
- L'entrenament porta temps, depenent de la grandària de l'arxiu.

3.3.5 Bitext manager



- Construir el bitext a partir dels textos alineats que hem pujat.

3.3.6 Translator trainer



Espai d'entrenament i gestió de motors.

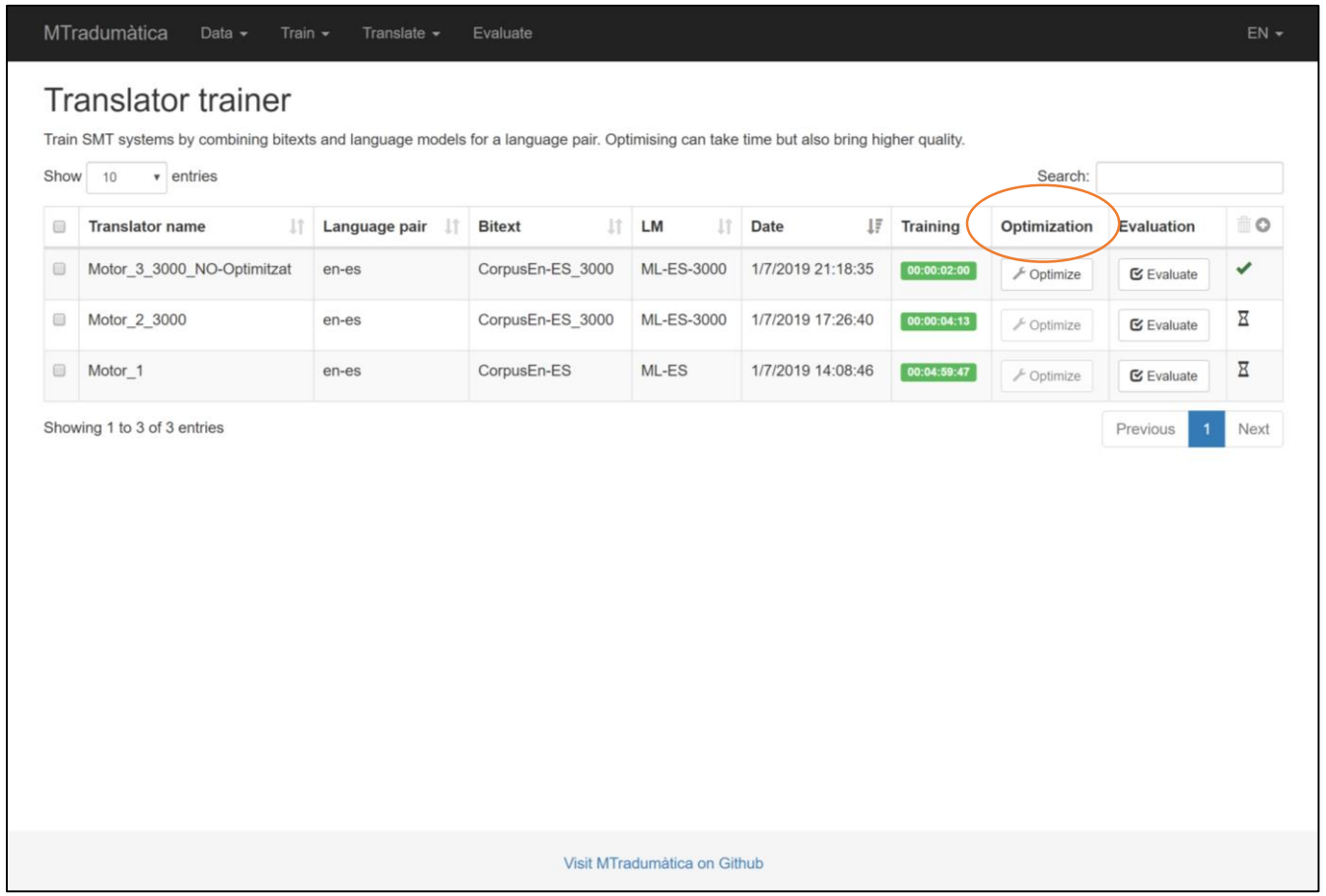
Per entrenar un motor, fem clic sobre el botó +.

Els elements que calen per obtenir la màxima qualitat possible:

- Un bitext (llengua de partida i d'arribada).
- Un model de llengua de la llengua d'arribada.

El procés d'entrenament requereix temps en funció de la grandària dels corpus.

3.3.6 Translator trainer



Un cop entrenat, per tal de configurar el motor de manera que proporcioni la màxima qualitat possible, s'ha d'optimitzar.

Per optimitzar un motor necessitem un corpus bilingüe petit en forma de bitext que no s'hagi introduït al motor.

El procés d'optimització pot ser tan lent o més que el procés de creació d'un motor.

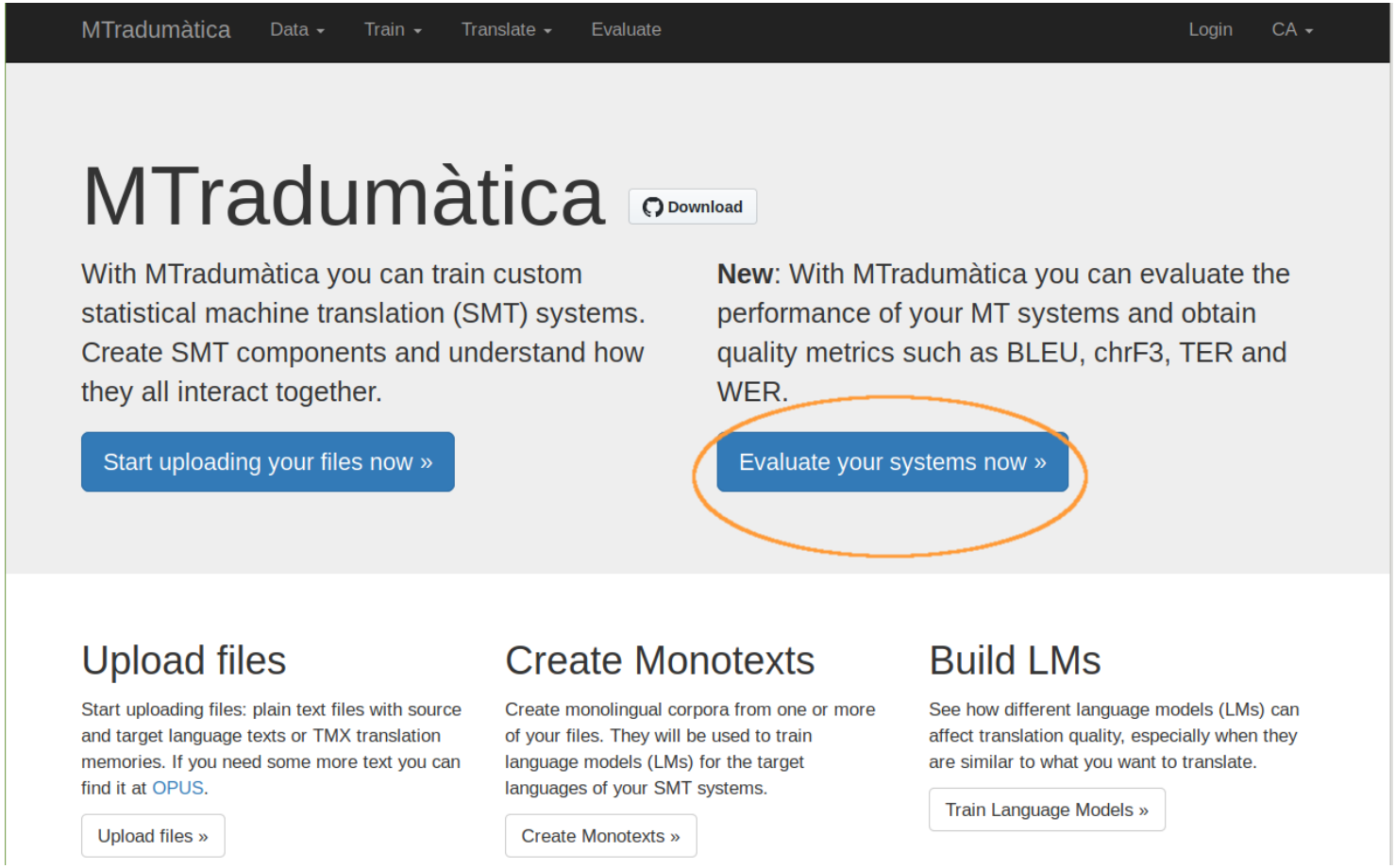
3.3.7 Avaluar motors

Dues possibilitats d'avaluació:

- Comparar MTradumàtica i motors externs amb un exemple de traducció humana.
- Avaluar un motor d'MTradumàtica amb un exemple de traducció humana.

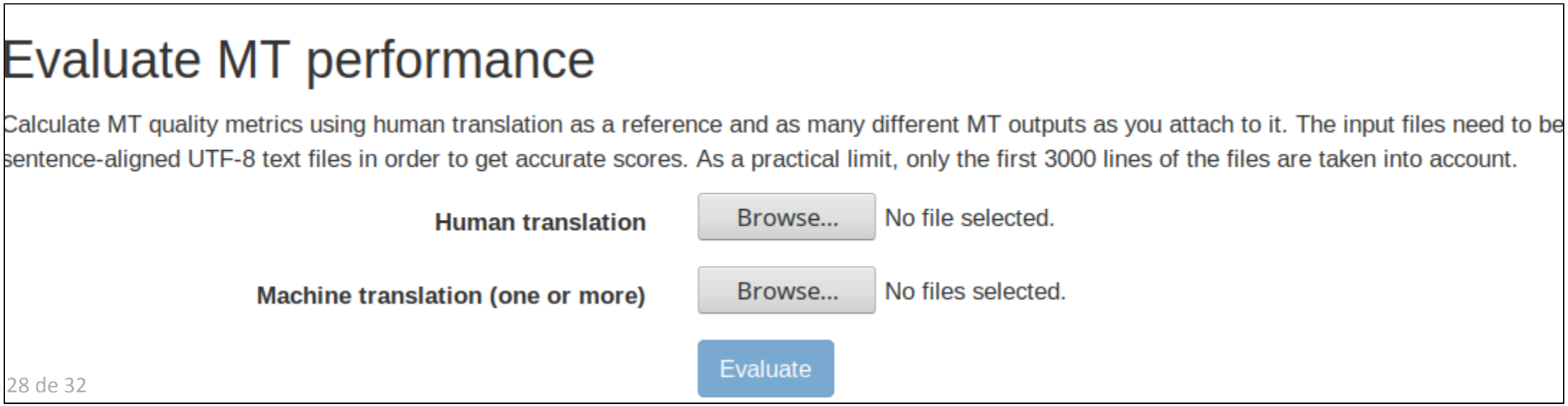
3.3.7.1 Comparar MTradumàtica i motors externs

Des de la pàgina inicial, accedir al mòdul d'avaluació.



3.3.7.1 Comparar MTradumàtica i motors externs

La interfície ens demana exclusivament traduccions (human i machine translation):



3.3.7.1 Comparar MTradumàtica i motors externs

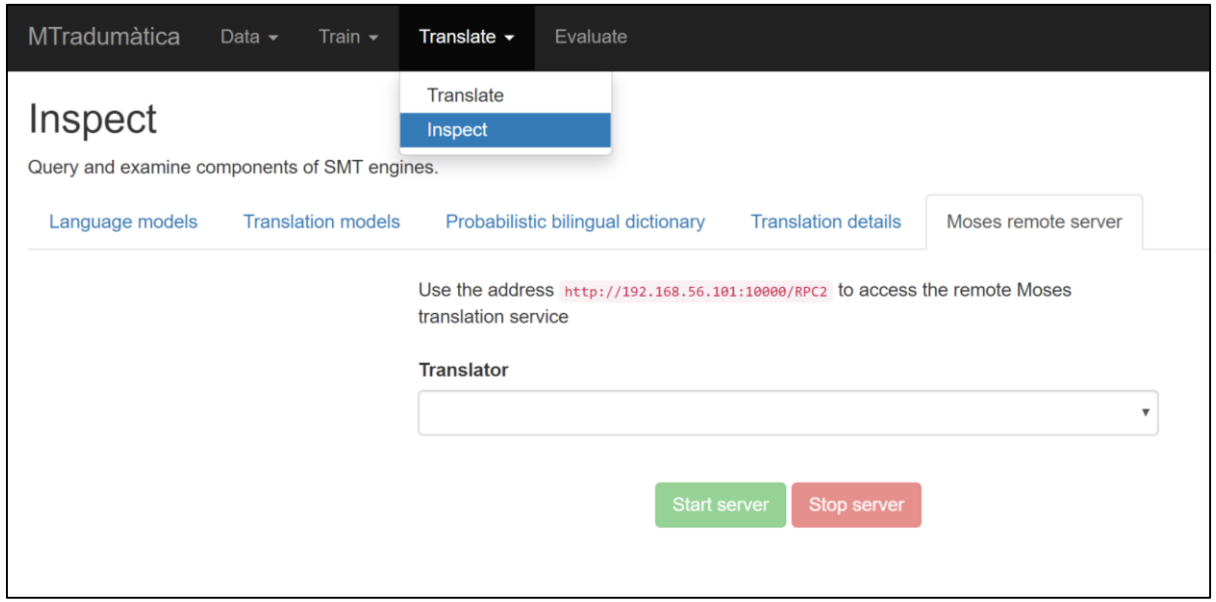
Podem carregar, doncs, la traducció humana i una o diverses traduccions automàtiques. Això ens serà útil sobretot per comparar mètriques entre motors diferents. Per exemple, si volem saber si per a un text determinat les mètriques d'avaluació són més o menys altes en dos motors (Apertium i Google), només ens cal fer-ne la TA (des del lloc web corresponent: www.apertium.org i translate.google.com) i carregar-ne els fitxers resultants a MTradumàtica. La pantalla resultant mostrarà una taula amb les mètriques d'avaluació i una llegenda amb una breu explicació de cada mètrica.

3.3.7.2 Avaluar un motor d'MTradumàtica

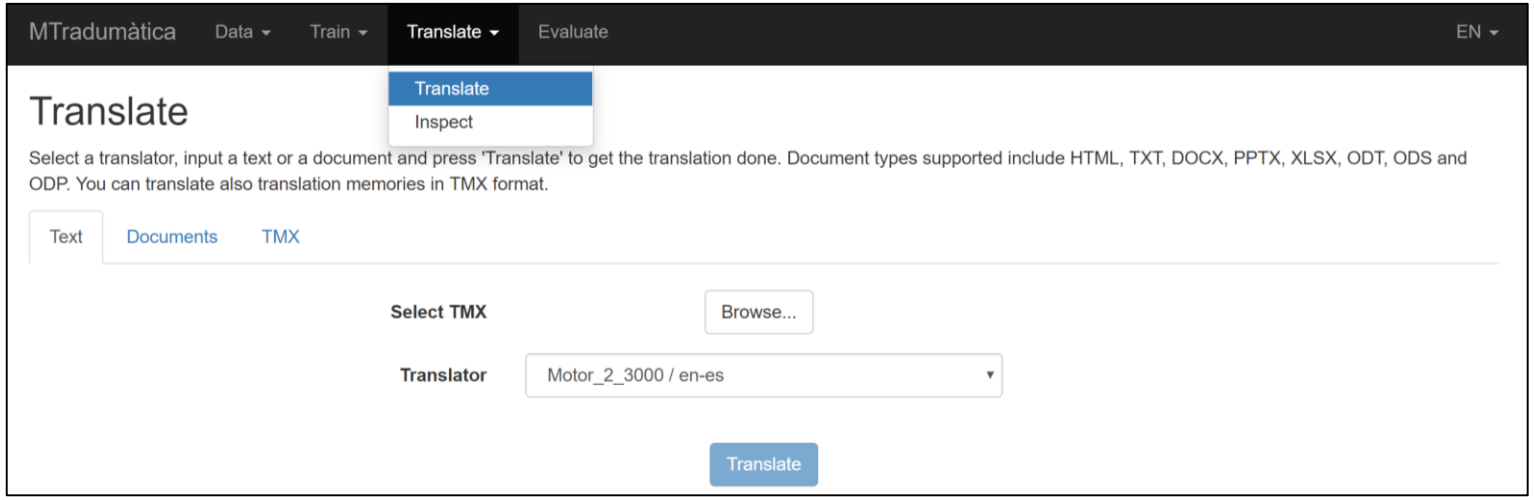
A la taula de motors, la penúltima columna serà Evaluation, i conté un botó Evaluate per a cada motor; des d'aquí, podrem carregar un text original i una traducció humana de referència (MTradumàtica s'encarregarà de fer-ne la TA amb el motor en concret):

Un cop carregats els textos, la mateixa taula anterior mostrarà el resultat de l'avaluació:

.....i tradueix!



Vincula el motor amb OmegaT mitjançant aquesta URL.



Tradueix frases, documents o TMX, i obté frases, documents o TMX.

4. Ho intentem?

Sessió de treball:

1. Instal·lació del sistema
    1. Instal·lar VirtualBox
    2. Importar i configurar MTradumàtica.ova
2. Creació d'un motor EN-ES a partir de l'arxiu Memoria_3000.tmx
    1. Pujar l'arxiu
    2. Crear el monotext ES_3000
    3. Crear el model de llengua ES_3000
    4. Crear el bitext EN-ES_3000
    5. Crear el motor EN-ES_3000
3. Optimització del motor a partir de l'arxiu Memoria_500.tmx
    1. Pujar l'arxiu
    2. Crear el bitext EN-ES_500
    3. Optimitzar el motor
4. Avaluar el motor a partir d'un arxiu original i la seva traducció humana. Utilitzarem:
    1. Original_EN.docx
    2. Human_Translation_ES.docx

5. Prova pilot a la teva empresa

Si us plau, contesta aquestes tres preguntes sobre el teu interès en fer una prova pilot amb la nostra col·laboració.

MOLTES GRÀCIES!!!