

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

TRADUMÀTICA

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 treballa? 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](#)).

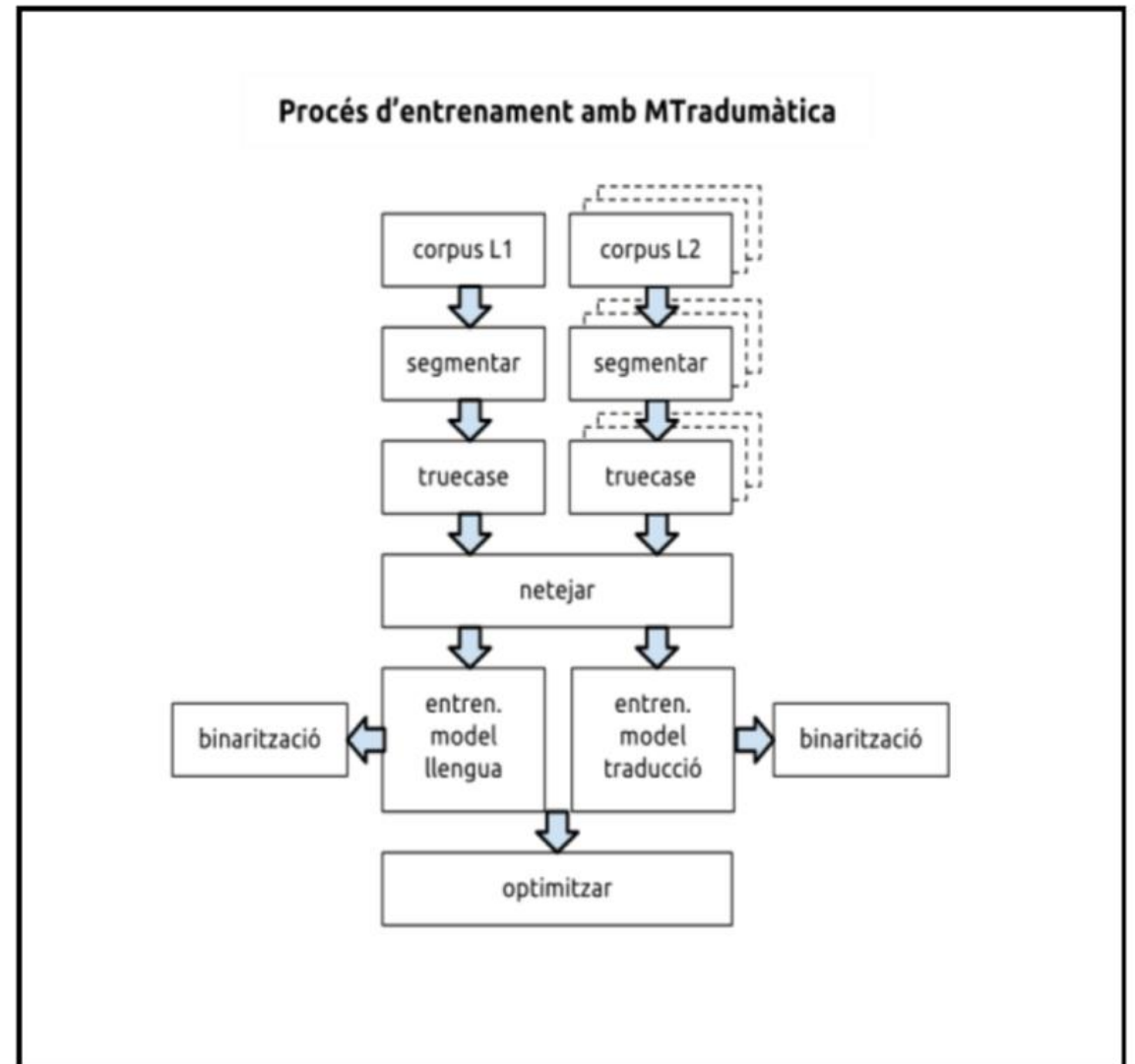
- Desenvolupat amb ProjectA (FFI2013-46041-R) i ProjectA-U (FFI2016-78612-R) del Ministeri d'Economia i Competitivitat.
- Llicència GPL (programari lliure), per a Linux; Windows i Mac mitjançant VirtualBox.

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

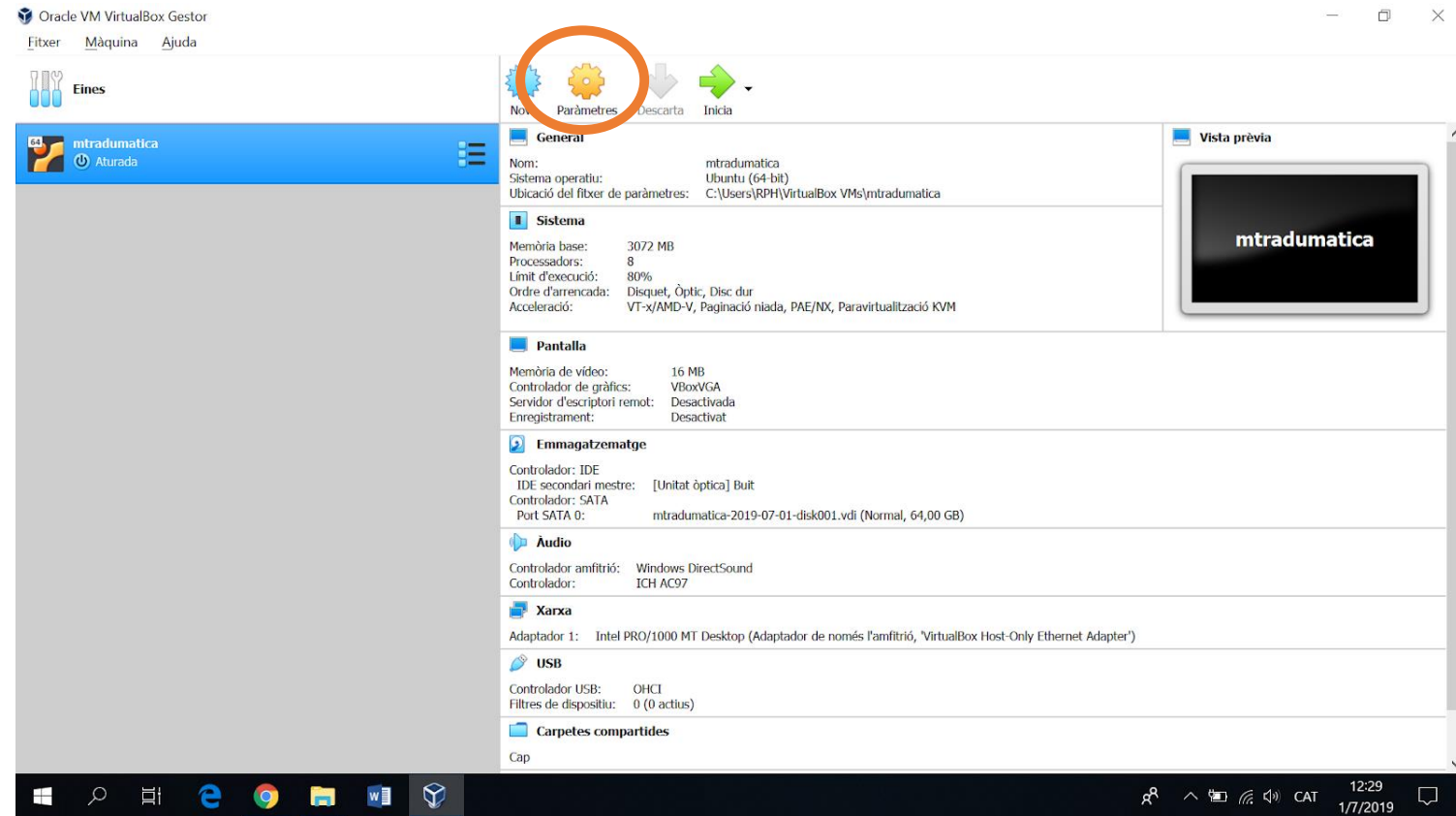


3.2. Procés d'instal·lació d'MTradumàtica

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

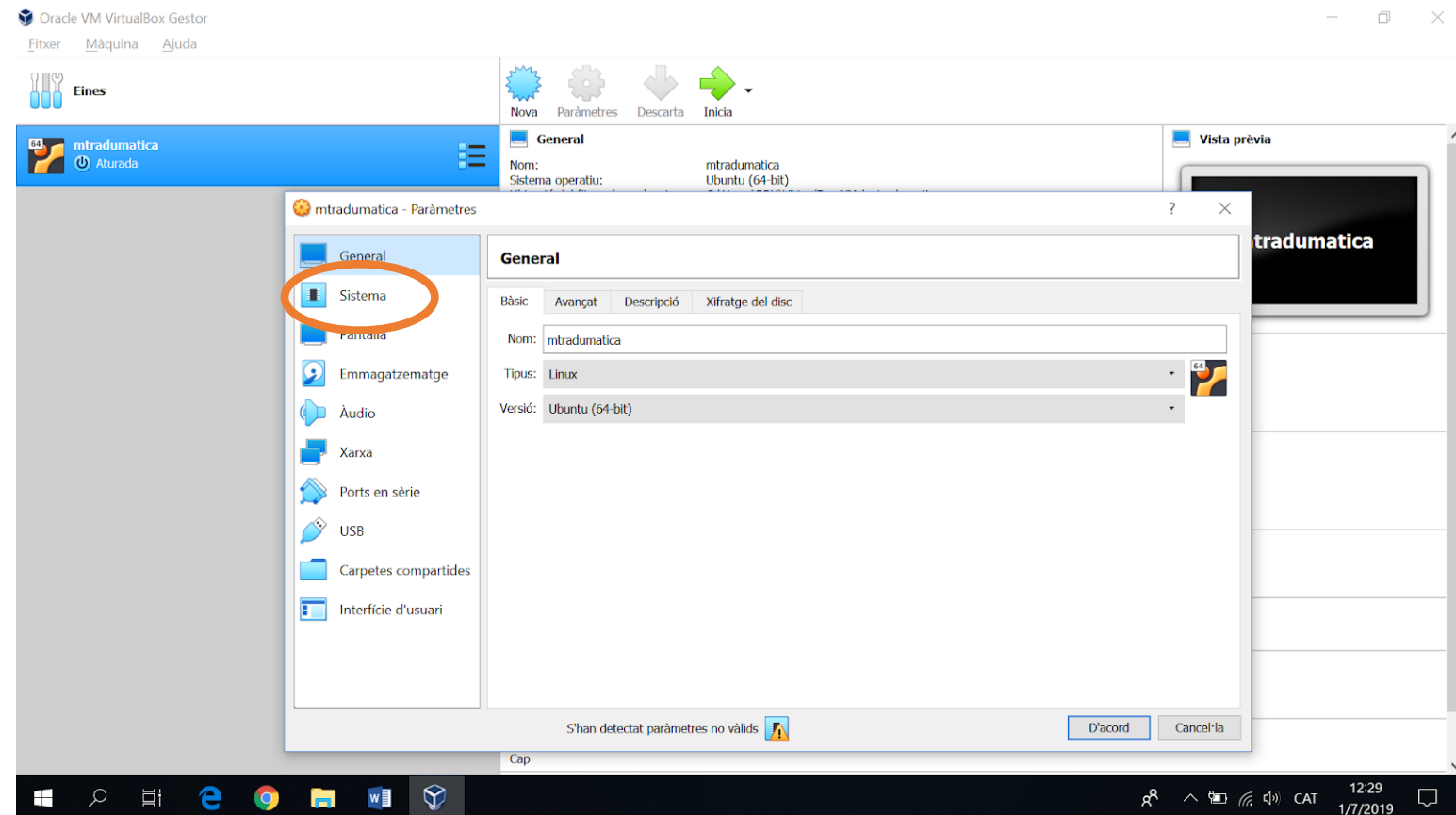
3. Canviar la configuració del motor:
Paràmetres | Sistema:



3.2. Procés d'instal·lació d'MTradumàtica

3. Canviar la configuració del motor:

Paràmetres | Sistema:

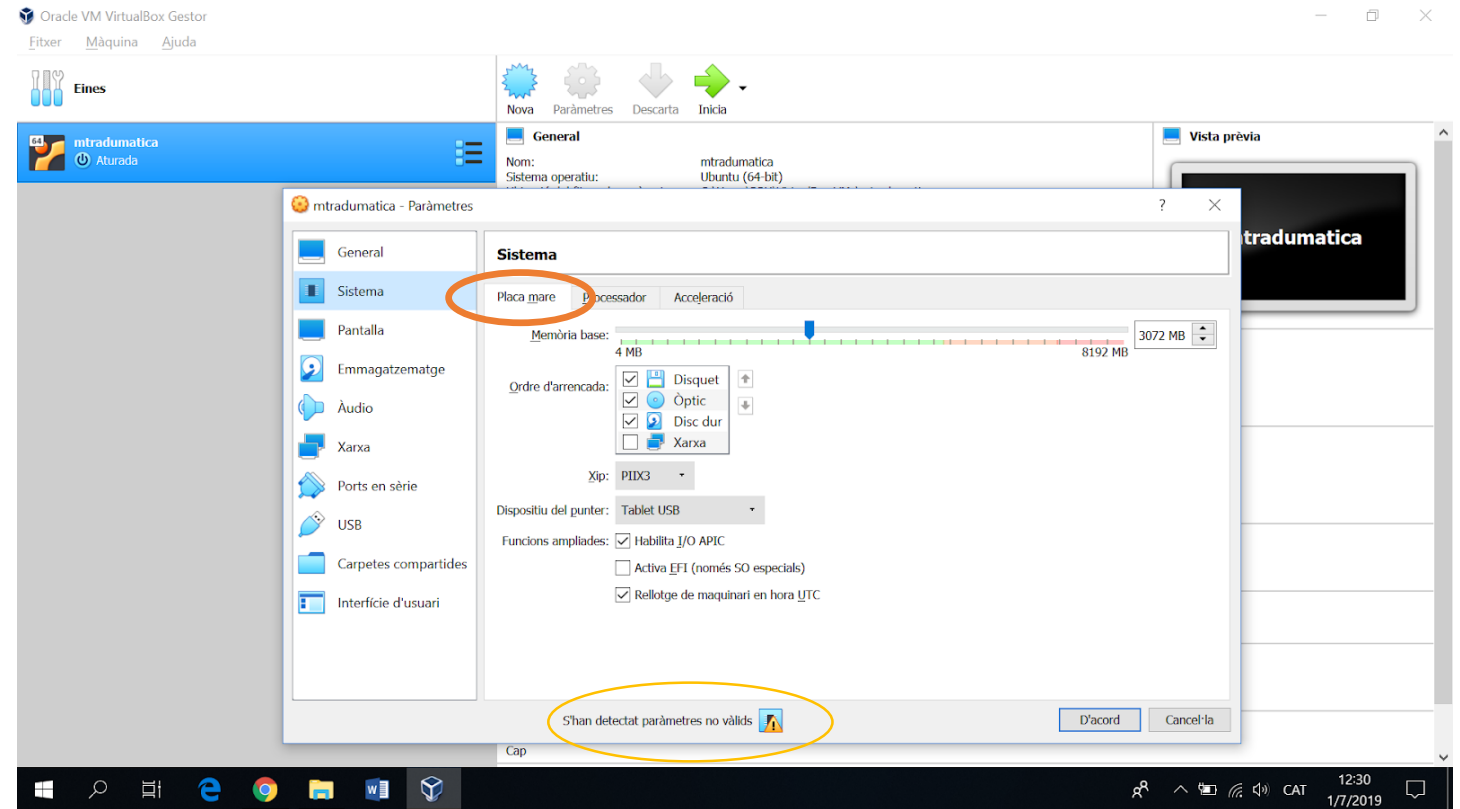


3.2. Procés d'instal·lació d'MTradumàtica

3. Canviar la configuració del motor:

Paràmetres | Sistema:

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



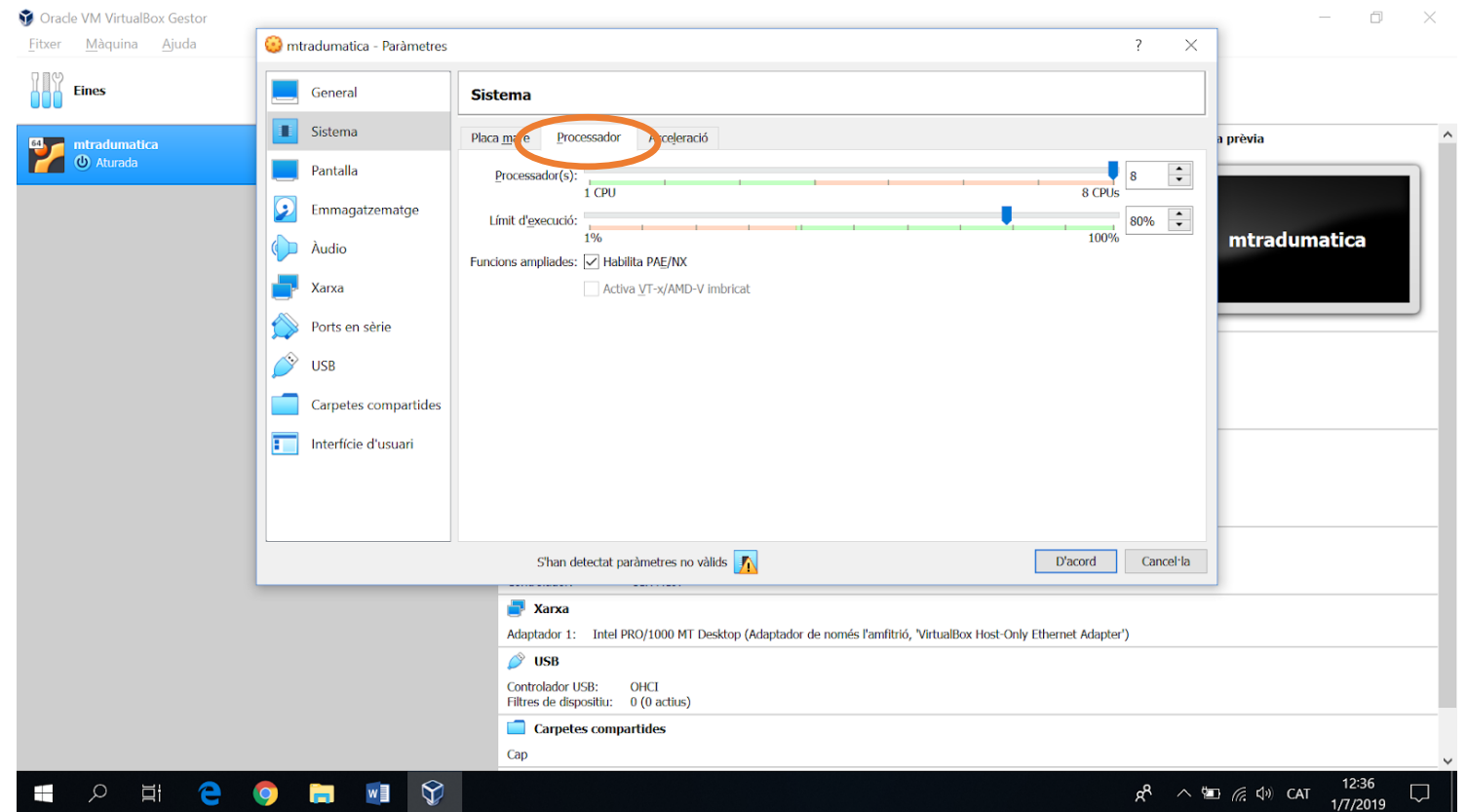
Evitar que aparegui el missatge de paràmetres no vàlids.

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.

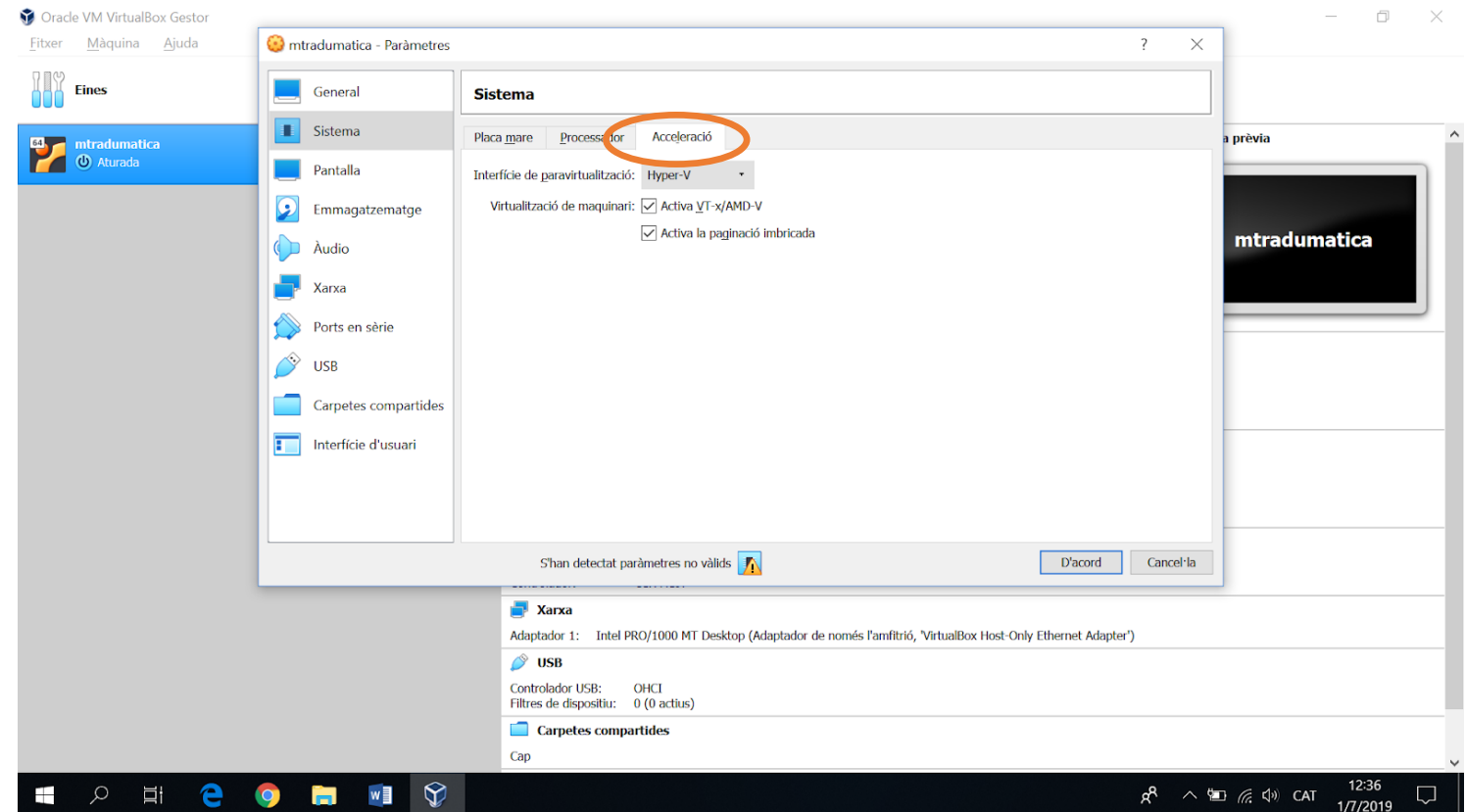


3.2. Procés d'instal·lació d'MTradumàtica

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.

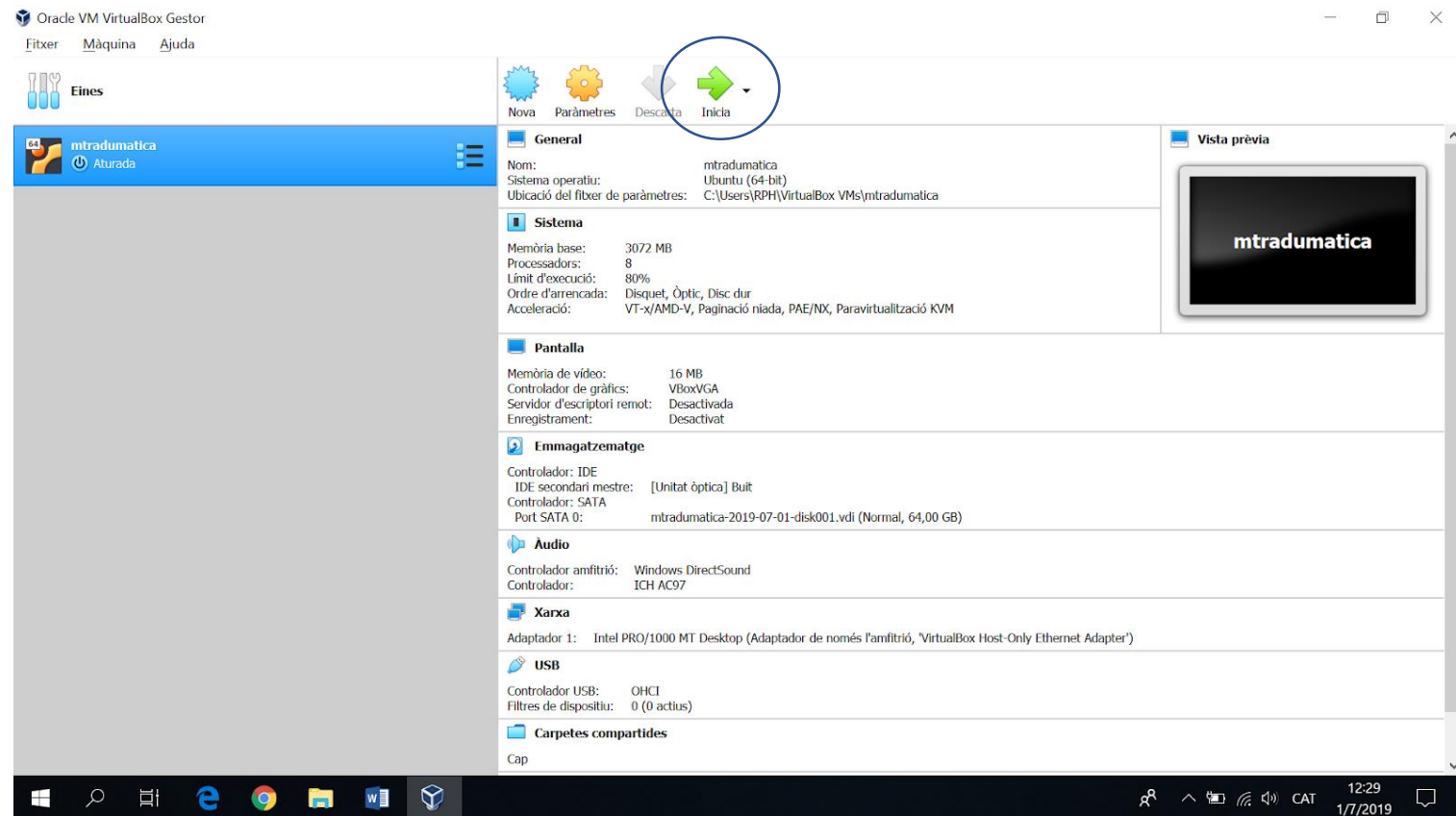


3.2. Procés d'instal·lació d'MTradumàtica

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-Arxiu d'instal·lació.



3.2. Procés d'instal·lació d'MTradumàtica

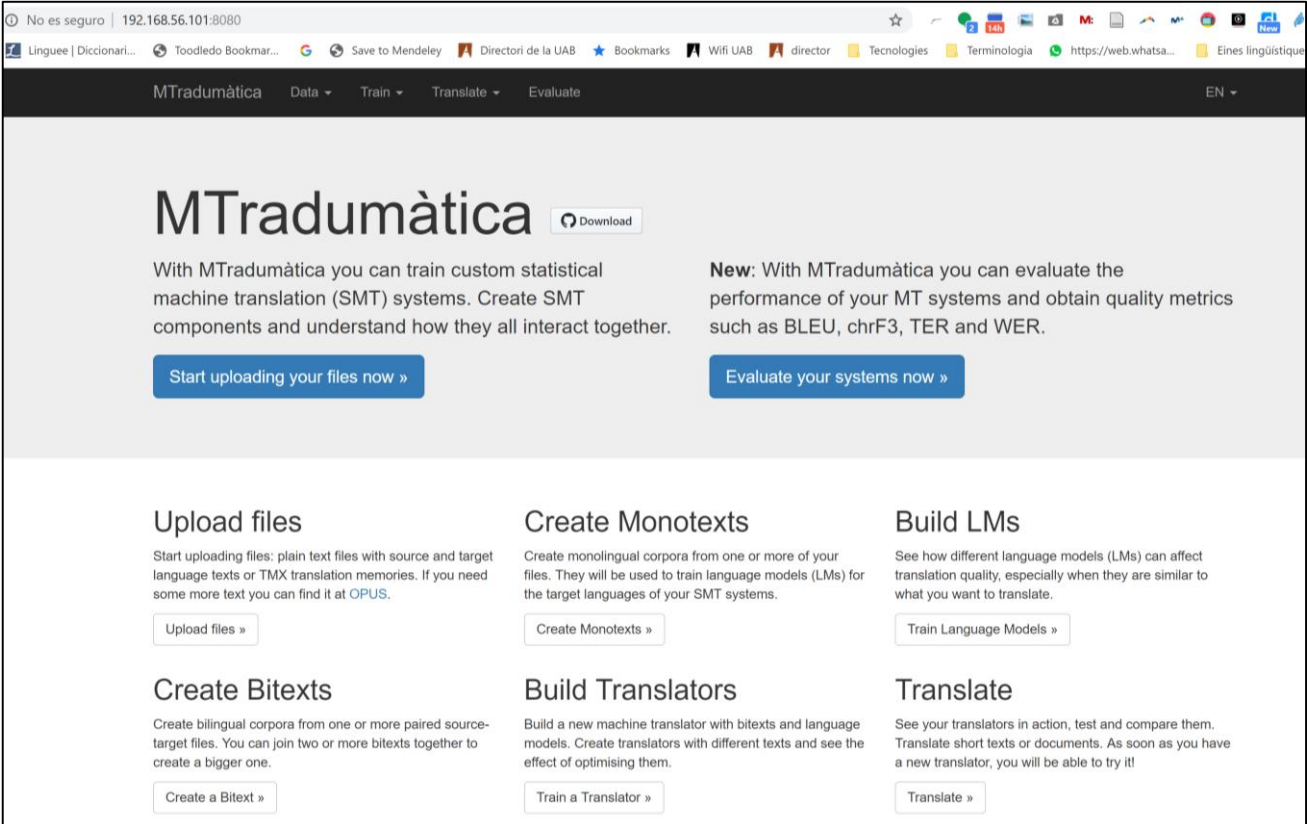
5. Generar la URL per treballar mitjançant el navegador

- 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.

```
Browse to Mtradumàtica at http://192.168.56.101:8080  
  
mtradumatica login:
```

3.3 Treball amb MTradumàtica

Introduir la URL en el navegador:



The screenshot shows the MTradumàtica website interface. The browser address bar displays "No es seguro | 192.168.56.101:8080". The website has a dark navigation bar with the following menu items: "MTradumàtica", "Data", "Train", "Translate", "Evaluate", and "EN". The main content area features a large heading "MTradumàtica" with a "Download" button. Below this, there are two columns of text. The left column describes the tool's capabilities: "With MTradumàtica you can train custom statistical machine translation (SMT) systems. Create SMT components and understand how they all interact together." and includes a blue button "Start uploading your files now >". The right column highlights a new feature: "New: With MTradumàtica you can evaluate the performance of your MT systems and obtain quality metrics such as BLEU, chrF3, TER and WER." and includes a blue button "Evaluate your systems now >". Below these are six sections, each with a title, a brief description, and a button:

- Upload files**: Start uploading files: plain text files with source and target language texts or TMX translation memories. If you need some more text you can find it at [OPUS](#). Button: "Upload files >"
- Create Monotexts**: Create monolingual corpora from one or more of your files. They will be used to train language models (LMs) for the target languages of your SMT systems. Button: "Create Monotexts >"
- Build LMs**: See how different language models (LMs) can affect translation quality, especially when they are similar to what you want to translate. Button: "Train Language Models >"
- Create Bitexts**: Create bilingual corpora from one or more paired source-target files. You can join two or more bitexts together to create a bigger one. Button: "Create a Bitext >"
- Build Translators**: Build a new machine translator with bitexts and language models. Create translators with different texts and see the effect of optimising them. Button: "Train a Translator >"
- Translate**: See your translators in action, test and compare them. Translate short texts or documents. As soon as you have a new translator, you will be able to try it! Button: "Translate >"

3.3 Treball amb MTradumàtica

MTradumàtica Data Train Translate Evaluate EN

MTradumàtica

Download

With MTradumàtica you can train custom statistical machine translation (SMT) systems. Create SMT components and understand how they all interact together.

New: With MTradumàtica you can evaluate the performance of your MT systems and obtain quality metrics such as BLEU, chrF3, TER and WER.

Start uploading your files now » Evaluate your systems now »

Upload files

Start uploading files: plain text files with source and target language texts or TMX translation memories. If you need some more text you can find it at [OPUS](#).

Upload files »

Create Monotexts

Create monolingual corpora from one or more of your files. They will be used to train language models (LMs) for the target languages of your SMT systems.

Create Monotexts »

Build LMs

See how different language models (LMs) can affect translation quality, especially when they are similar to what you want to translate.

Train Language Models »

Create Bitexts

Create bilingual corpora from one or more paired source-target files. You can join two or more bitexts together to create a bigger one.

Create a Bitext »

Build Translators

Build a new machine translator with bitexts and language models. Create translators with different texts and see the effect of optimising them.

Train a Translator »

Translate

See your translators in action, test and compare them. Translate short texts or documents. As soon as you have a new translator, you will be able to try it!

Translate »

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

MTradumàtica **Files** Monotexts LMs Bitexts Translators Translate Inspect


File manager

Add either text or TMX files to MTradumàtica; you will always find them all stored here.

Show entries Search:

| File name | Language | Lines | Words | Chars | Date |
|----------------------------|----------|-------|-------|-------|------|
| No data available in table | | | | | |

Showing 0 to 0 of 0 entries Previous Next


Click here or drag and drop files
(you can upload more than one file at once)

Developed by Prompsit Language Engineering - Visit [MTradumàtica on Github](#)

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

MTradumàtica Files **Monotexts** LMs Bitexts Translators Translate Inspect

Monotext manager

Create monolingual corpora to train language models. Add one or more files to each monotext provided that they are all in the same language!

Show 10 entries Search:

| <input type="checkbox"/> | Monotext name | Language | Lines | Date |   |
|----------------------------|---------------|----------|-------|------|---|
| No data available in table | | | | | |

Showing 0 to 0 of 0 entries Previous Next

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

MTradumàtica Files Monotexts **LMs** Bitexts Translators Translate Inspect

Language model trainer

Train language models by selecting monotexts previously defined. The training will be automatically launched!

Show entries Search:

| <input type="checkbox"/> | Model name | Language | Monolingual corpus | Date | Training time | <input type="checkbox"/> |
|--------------------------|--------------------|----------|---------------------------|-----------------------|---------------|--------------------------|
| <input type="checkbox"/> | Genérico-Coloquial | es | Corpus genérico-coloquial | 24/6/2018 20:15:15 | 00:00:01:57 | <input type="checkbox"/> |
| <input type="checkbox"/> | Genérico-Coloquial | en | Corpus genérico-coloquial | 24/6/2018 20:15:01 | 00:00:02:11 | <input type="checkbox"/> |

Showing 1 to 2 of 2 entries Previous **1** Next

- 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

MTradumàtica Files Monotexts LMs **Bitexts** Translators Translate Inspect

Bitext manager

Create bilingual corpora to train SMT systems. Add as many source-target files as you want to your bitext provided that they are parallel!

Show entries Search:

| <input type="checkbox"/> | Bitext name | ⇕ Languages | ⇕ Lines | ⇕ Date | ⇕ | 🗑️ |
|--------------------------|--------------------------------|-------------|----------|--------------------|----|----|
| <input type="checkbox"/> | Bitexto Genérico-Especializado | en-es | 11607686 | 24/6/2018 21:57:45 | 👁️ | 🔍 |
| <input type="checkbox"/> | Bitexto Genérico-Coloquial | en-es | 9845028 | 24/6/2018 21:41:57 | 👁️ | 🔍 |

Showing 1 to 2 of 2 entries Previous **1** Next

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

3.3.6 Translator trainer

MTradumàtica Data Train Translate Evaluate EN

Translator trainer

Train SMT systems by combining

Show 10 entries

| Translator name | Motor |
|-----------------------|-------|
| Motor_3_3000_NO-Optim | |
| Motor_2_3000 | |
| Motor_1 | |

Showing 1 to 3 of 3 entries

Create a new translator

Name: Give a name to the translator

From bitexts and LMs | From files (no pre-built LM) | Link existing MT

Source language: None

Target language: None

Bitext: [Dropdown]

Language model: [Dropdown]

Close Add

Evaluation: [List of Evaluate buttons with status indicators]

Previous 1 Next

Visit MTradumàtica on Github

Espai d'entrenament i gestió de motors.

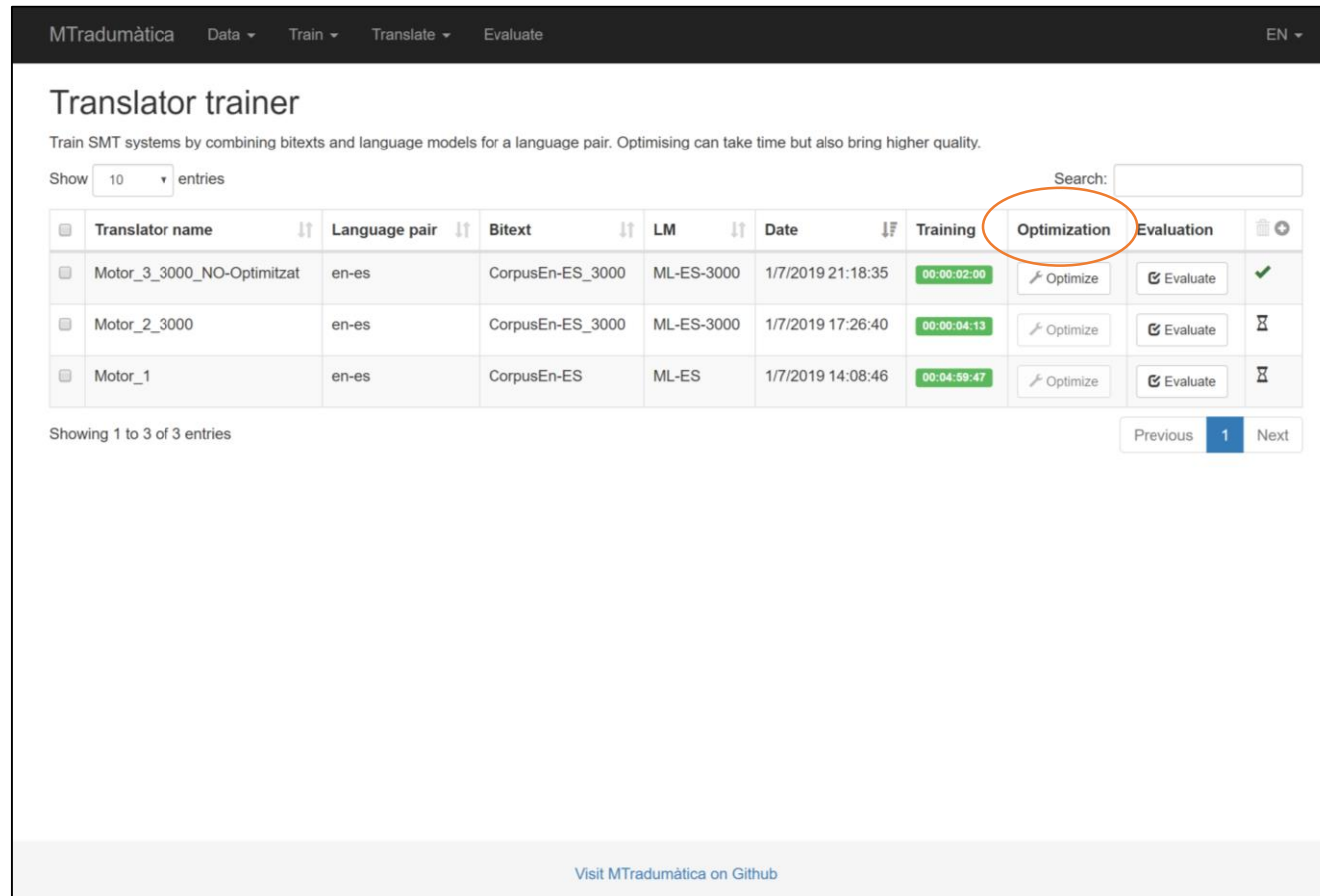
Per entrenar un motor, fem clic sobre el botó +.

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



MTradumàtica Data Train Translate Evaluate EN

Translator trainer

Train SMT systems by combining bitexts and language models for a language pair. Optimising can take time but also bring higher quality.

Show 10 entries Search:

| Translator name | Language pair | Bitext | LM | Date | Training | Optimization | Evaluation |
|---------------------------|---------------|------------------|------------|-------------------|-------------|--------------|------------|
| Motor_3_3000_NO-Optimizat | en-es | CorpusEn-ES_3000 | ML-ES-3000 | 1/7/2019 21:18:35 | 00:00:02:00 | Optimize | Evaluate ✓ |
| Motor_2_3000 | en-es | CorpusEn-ES_3000 | ML-ES-3000 | 1/7/2019 17:26:40 | 00:00:04:13 | Optimize | Evaluate ⌚ |
| Motor_1 | en-es | CorpusEn-ES | ML-ES | 1/7/2019 14:08:46 | 00:04:59:47 | Optimize | Evaluate ⌚ |

Showing 1 to 3 of 3 entries Previous 1 Next

Visit MTradumàtica on Github

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.

MTradumàtica Data Train Translate Evaluate Login CA

MTradumàtica

[Download](#)

With MTradumàtica you can train custom statistical machine translation (SMT) systems. Create SMT components and understand how they all interact together.

[Start uploading your files now »](#)

New: With MTradumàtica you can evaluate the performance of your MT systems and obtain quality metrics such as BLEU, chrF3, TER and WER.

[Evaluate your systems now »](#)

Upload files

Start uploading files: plain text files with source and target language texts or TMX translation memories. If you need some more text you can find it at [OPUS](#).

[Upload files »](#)

Create Monotexts

Create monolingual corpora from one or more of your files. They will be used to train language models (LMs) for the target languages of your SMT systems.

[Create Monotexts »](#)

Build LMs

See how different language models (LMs) can affect translation quality, especially when they are similar to what you want to translate.

[Train Language Models »](#)

3.3.7.1 Comparar MTradumàtica i motors externs

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

Evaluate MT performance

Calculate MT quality metrics using human translation as a reference and as many different MT outputs as you attach to it. The input files need to be sentence-aligned UTF-8 text files in order to get accurate scores. As a practical limit, only the first 3000 lines of the files are taken into account.

Human translation

Browse...

No file selected.

Machine translation (one or more)

Browse...

No files selected.

Evaluate

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:

| | | | | | | | | | |
|--------------------------|------|-------|--|--|------------------------|-------------|----------|--|--|
| <input type="checkbox"/> | enca | en-ca | | | 29/01/2019 16:18:40 | 00:00:01:06 | Optimize | <input checked="" type="checkbox"/> Evaluate BLEU: 73.46; chrF3: 80.44; TER: 19.68; WER: 20.89; BEER: 80.27 | |
|--------------------------|------|-------|--|--|------------------------|-------------|----------|--|--|

.....i tradueix!

The screenshot shows the 'Inspect' page in the MTradumàtica application. The top navigation bar includes 'MTradumàtica', 'Data', 'Train', 'Translate', and 'Evaluate'. The 'Translate' menu is open, showing 'Translate' and 'Inspect' options. The main content area is titled 'Inspect' and has the subtitle 'Query and examine components of SMT engines.' Below this, there are tabs for 'Language models', 'Translation models', 'Probabilistic bilingual dictionary', 'Translation details', and 'Moses remote server'. The 'Moses remote server' tab is active. It contains instructions: 'Use the address <http://192.168.56.101:10000/RPC2> to access the remote Moses translation service'. There is a 'Translator' dropdown menu and two buttons: 'Start server' (green) and 'Stop server' (red).

Vincula el motor amb OmegaT mitjançant aquesta URL.

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

The screenshot shows the 'Translate' page in the MTradumàtica application. The top navigation bar includes 'MTradumàtica', 'Data', 'Train', 'Translate', and 'Evaluate'. The 'Translate' menu is open, showing 'Translate' and 'Inspect' options. The main content area is titled 'Translate' and has the subtitle 'Select a translator, input a text or a document and press 'Translate' to get the translation done. Document types supported include HTML, TXT, DOCX, PPTX, XLSX, ODT, ODS and ODP. You can translate also translation memories in TMX format.' Below this, there are tabs for 'Text', 'Documents', and 'TMX'. The 'Text' tab is active. It contains a 'Select TMX' field with a 'Browse...' button and a 'Translator' dropdown menu set to 'Motor_2_3000 / en-es'. There is a 'Translate' button at the bottom.

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!!!