

Traducción automática estadística para traductores

Pilar Sánchez Gijón (UAB)

Antoni Oliver (UOC)

Sede de la Comisión Europea

Madrid, 25 de junio de 2018

Objetivo

Entender y gestionar un sistema de traducción automática estadística basado en Moses.

1. ¿En qué consiste un sistema de TAE?
2. ¿Qué opciones tengo?
3. Si decido tener mi propio sistema, ¿cómo puedo hacerlo?

1.¿En qué consiste un sistema de TAE?

A partir de una gran cantidad de texto en la lengua de partida y en la lengua de llegada, alineado y no alineado, el sistema busca la mejor traducción posible de cada uno de los elementos de una oración, partiendo del segmento más grande posible (frase completa) hasta el más pequeño (palabra).

1. ¿En qué consiste un sistema de TAE?

La calidad de la traducción depende de:

1. El corpus de trabajo:
 - Córpora monolingües y corpus bilingüe
2. El tiempo de respuesta:
 - Cantidad de opciones valoradas
 - Cantidad de texto procesado

2. ¿Qué opciones tengo?

- Sistemas de TA (TAE u otros sistemas) ya existentes
- Crear un motor en un sistema público
- Construirse un sistema propio

2. ¿Qué opciones tengo?

Sistemas de TA (TAE u otros sistemas) ya existentes:

VENTAJAS:

- Solución tecnológica integral
- No hay necesidad de buscar/gestionar corpus

2. ¿Qué opciones tengo?

Sistemas de TA (TAE u otros sistemas) ya existentes:

INCONVENIENTES:

- Seguridad
 - ¿Dónde quedan los archivos que traduzco?
- Coste
 - ¿Es un SAAS? ¿Coste por volumen de palabras?
- Interacción con herramientas TAO
 - ¿Tiene API o se integra de cualquier otro modo?
- Adaptabilidad:
 - ¿El corpus prevé los tipos de texto que me interesa traducir?

2. ¿Qué opciones tengo?

Crear un motor en un sistema público:

VENTAJAS:

- Solución tecnológica integral
- Adaptabilidad del corpus
- Seguridad: acceso restringido a archivos y córpora

2. ¿Qué opciones tengo?

Sistemas de TA (TAE u otros sistemas) ya existentes:

INCONVENIENTES:

- Seguridad
 - ¿Dónde quedan los archivos que traduzco?
- Coste
 - ¿Es un SAAS? ¿Coste por volumen de palabras?
- Interacción con herramientas TAO
 - ¿Tiene API o se integra de cualquier otro modo?
- Gestión de archivos:
 - ¿Con qué archivos trabajo? ¿Dónde los obtengo? ¿Cómo los gestiono?

2. ¿Qué opciones tengo?

Construirse un sistema propio

VENTAJAS:

- Adaptabilidad del corpus
- Seguridad: acceso restringido a archivos y córpora

2. ¿Qué opciones tengo?

Construirse un sistema propio:

INCONVENIENTES:

- Tecnología
 - ¿Programación? ¿Linux? ¿Gestión de un servidor?
- Coste
 - ¿Servidor? ¿Aspectos de seguridad?
- Interacción con herramientas TAO
 - ¿Con qué TAO? ¿Cómo integrarlo
- Gestión de archivos:
 - ¿Con qué archivos trabajo? ¿Dónde los obtengo? ¿Cómo los gestiono?

3. Si decido tener mi propio sistema, ¿cómo puedo hacerlo?

MTradumàtica

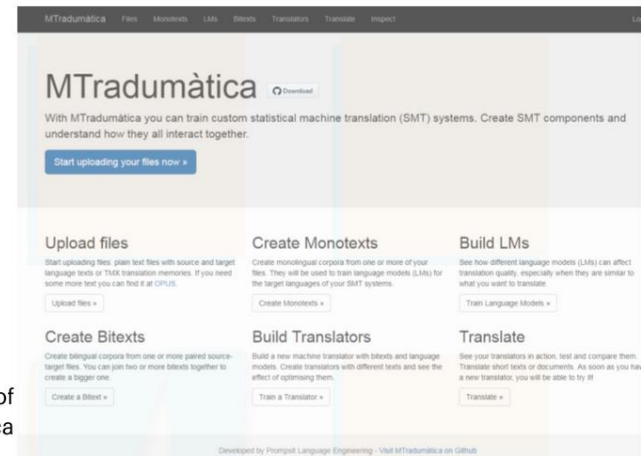
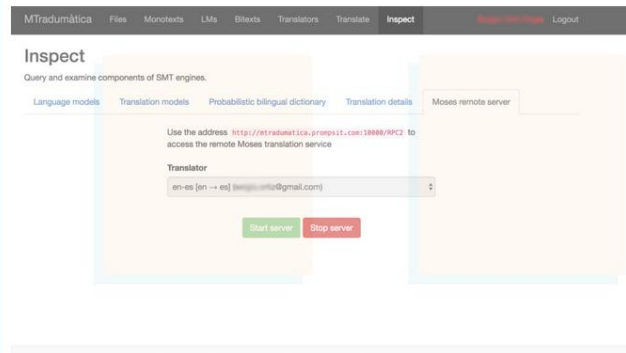
Appropriate to start managing and testing MT technologies.

- ▶ Intuitive creation and management of SMT engines in five steps:
- ▶ (1) Upload files
- ▶ (2) Create and manage monotexts
- ▶ (3) Build language models (LMs)
- ▶ (4) Create and manage bitexts
- ▶ (5) Train SMT translation models.

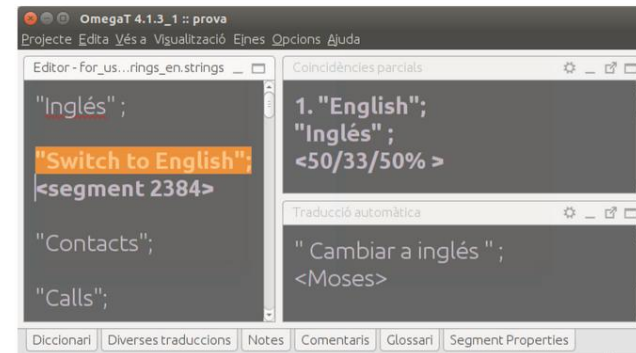
5 Steps

Figure 1. The interface of MTradumàtica

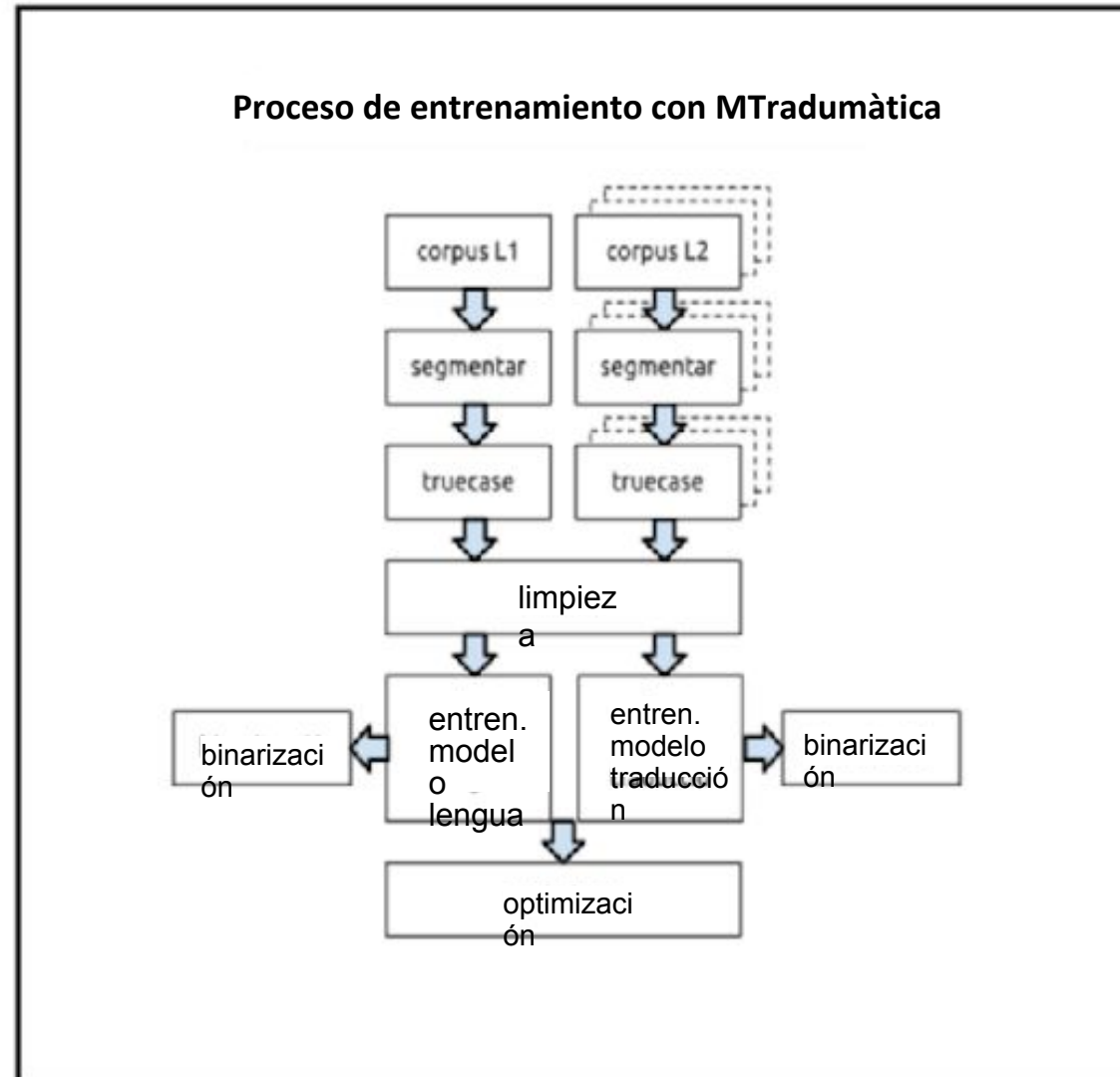
- ▶ Works with parallel, sentence-by-line text files such as Moses.
- ▶ Has a user-friendly web interface.
- ▶ Can be installed in private servers (increased confidentiality).



- ▶ Is able to import TMX files.
- ▶ Includes user management features (increased confidentiality).
- ▶ Integration with TM systems, like OmegaT, through an URL.



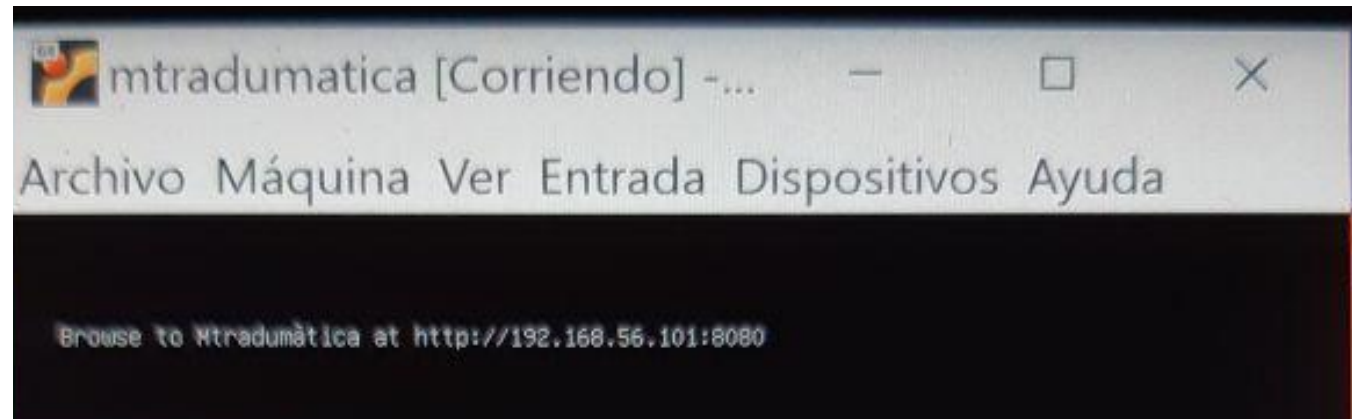
3.1 Proceso de entrenamiento de un motor TAE



3.2. Proceso de instalación

1. Instalar Virtualbox
2. Importar Mtradumatica.ova
3. Cambiar la configuración del motor
4. Iniciar el motor
5. Generar la URL para trabajar a través del navegador

3.2 URL de acceso al servidor en local



Proceso

The screenshot shows the MTradumàtica website interface. At the top, there is a navigation bar with the following items: MTradumàtica, Files, Monotexts, LMs, Bitexts, Translators, Translate, and Inspect. Below the navigation bar, the main content area features the title "MTradumàtica" with a "Download" button. A descriptive paragraph states: "With MTradumàtica you can train custom statistical machine translation (SMT) systems. Create SMT components and understand how they all interact together." Below this is a blue button that says "Start uploading your files now »".

The main content is organized into a grid of six functional blocks, 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.
- Build Translators:** Build a new machine translator with bitexts and language models. Create translators with different texts and see the effect of optimising them.
- 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!

The bottom of the screenshot shows a Windows taskbar with the search bar "Escribe aquí para buscar", various application icons, and the system tray showing the time "18:47" and date "24/06/2018".

1. Subir archivos monolingües o bilingües
2. Crear monotextos
3. Construir modelos de lengua (LM)
4. Crear bitextos
5. Construir traductores (motores)
6. Traducir

Upload files

File manager

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

Show 10 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)

Formatos de archivo:

- Txt (monolingüe)
- Txt (alineado para Moses)
- TMX

Upload files

File manager

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

Show 10 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

Formatos de archivo:

- Txt (monolingüe)
- Txt (alineado para Moses)
- TMX

Upload files

File manager

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

Show 10 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

Formatos de archivo:

- Txt (monolingüe)
- Txt (alineado para Moses)
- TMX

Monotext Manager

The screenshot shows a web browser window with the URL `192.168.56.101:8080/monolingual_corpora`. The page title is "Monolingual corpora" and the navigation menu includes "MTradumàtica", "Files", "Monotexts", "LMs", "Bitexts", "Translators", "Translate", and "Inspect". The main heading is "Monotext manager" with a sub-heading: "Create monolingual corpora to train language models. Add one or more files to each monotext provided that they are all in the same language!". Below this, there is a "Show" dropdown set to "10 entries" and a "Search:" input field. A table with columns "Monotext name", "Language", "Lines", and "Date" is displayed, but it is empty with the message "No data available in table". A red box highlights a trash icon in the table's header area. At the bottom, there are "Previous" and "Next" buttons. The footer text reads "Developed by Prompsit Language Engineering - Visit MTradumàtica on Github". The Windows taskbar at the bottom shows the search bar with "Escribe aquí para buscar", several application icons, and the system tray with the time "19:45" and date "24/06/2018".

Crear corpus monolingües con texto para cada combinación lingüística. Puede ser corpus que después también se alinea.

Language Model trainer

Language model trainer

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

Show 10 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"/>

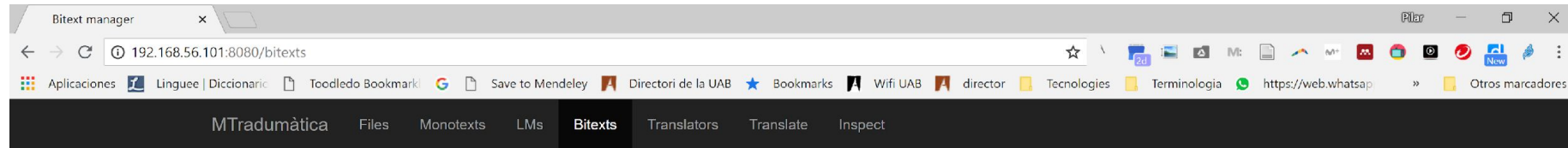
Showing 1 to 2 of 2 entries

Previous 1 Next

Developed by Prompsit Language Engineering - Visit MTradumàtica on Github

- Entrenamiento de cada modelo de lengua a partir de un monotexto. Consiste en binarizar los textos.
- Se puede entrenar un modelo de lengua por monotexto.
- El entrenamiento lleva su tiempo, dependiendo del tamaño del archivo.

Bitext Builder



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

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



Translator trainer

The screenshot shows a web browser window with the URL `192.168.56.101:8080/translators`. The browser's address bar and tabs are visible. The application's navigation bar includes links for `MTradumàtica`, `Files`, `Monotexts`, `LMS`, `Bitexts`, `Translators` (which is active), `Translate`, and `Inspect`.

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 entries Search:

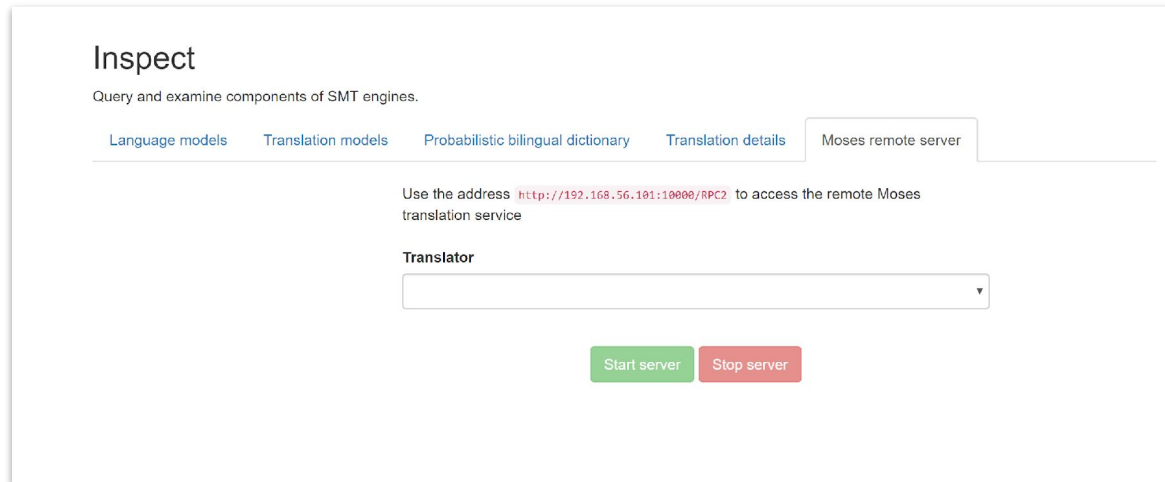
<input type="checkbox"/>	Translator name	Language pair	Bitext	Language model	Date	Training time	Optimization time	
No data available in table								

Showing 0 to 0 of 0 entries

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

The Windows taskbar at the bottom shows the search bar with the text "Escribe aquí para buscar", several application icons (including Edge, File Explorer, and various office tools), and the system tray with the time `22:54` and date `24/06/2018`.

.....y traduce!



Inspect

Query and examine components of SMT engines.

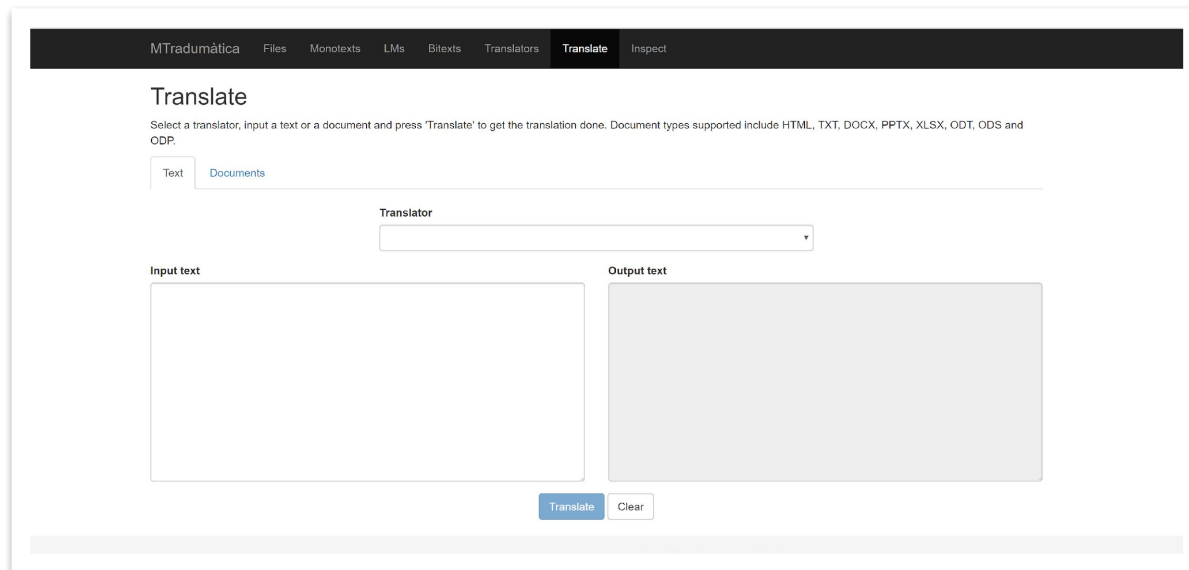
Language models Translation models Probabilistic bilingual dictionary Translation details Moses remote server

Use the address <http://192.168.56.101:10000/RPC2> to access the remote Moses translation service

Translator

Start server Stop server

Vincula el motor con OmegaT mediante esta URL.



MTradumática Files Monotexts LMs Bitexts Translators Translate Inspect

Translate

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.

Text Documents

Translator

Input text Output text

Translate Clear

Traducir textos o frases.

¿Lo intentamos?

- En tu ordenador:
 - Instala Virtualbox
 - Importa mtradumatica.ova
 - Ajusta la configuración (RAM y núcleos)
 - Inicia la máquina
 - y comienza a gestionar tu sistema de motores!!!

Referencias bibliogràfiques

MARTÍN-MOR, A.; PIQUÉ, R. (2017). «MTradumàtica i la formació de traductors en Traducció Automàtica Estadística». Revista Tradumàtica. *Tecnologies de la Traducció*, 15, 97-115.
<https://doi.org/10.5565/rev/tradumatica.199>