Machine Translation for Translators

Ruben de la Fuente

ata

American Translators Association

About Me

- 4-year degree in translation
- Worked as translator for 10+ years
- Only working full time in MT for the past year

ata

American Translators Association

Do You Think MT Will Be Part of Your Toolbox in a Future?

Poll

Why should you concern yourself with Machine Translation?

- Produce the **same quality**
- Only **faster**
- Make **more revenue**

American Translators Association

Does MT Make Sense For You?

- **Volumes** are high
- Texts are **consistent** in terms of **terminology** and **style**
- **Domain** does not matter so much if texts are consistent

Customize to suit your needs

- Customize upfront
- Update regularly

MT engines are not finished products, but on-going work

Keys to evaluate MT

Use hard facts and not impressions:

- Run an automatic QA tool (Xbench): flags obvious errors (terminology, capitalization, punctuation)
- See how much you can edit in one hour
- Calculate edit distance and generate change reports (SymEval)

Checklist for MT Tool Selection

- Linguistic quality
- Language pairs
- Ease of integration
- Customizability
- Confidentiality

Rule-based MT

- Dictionaries+Translation rules
- Output is grammatically correct and predictable, but maybe not very natural
- More costly and take longer to develop
- More user-friendly

ProMT

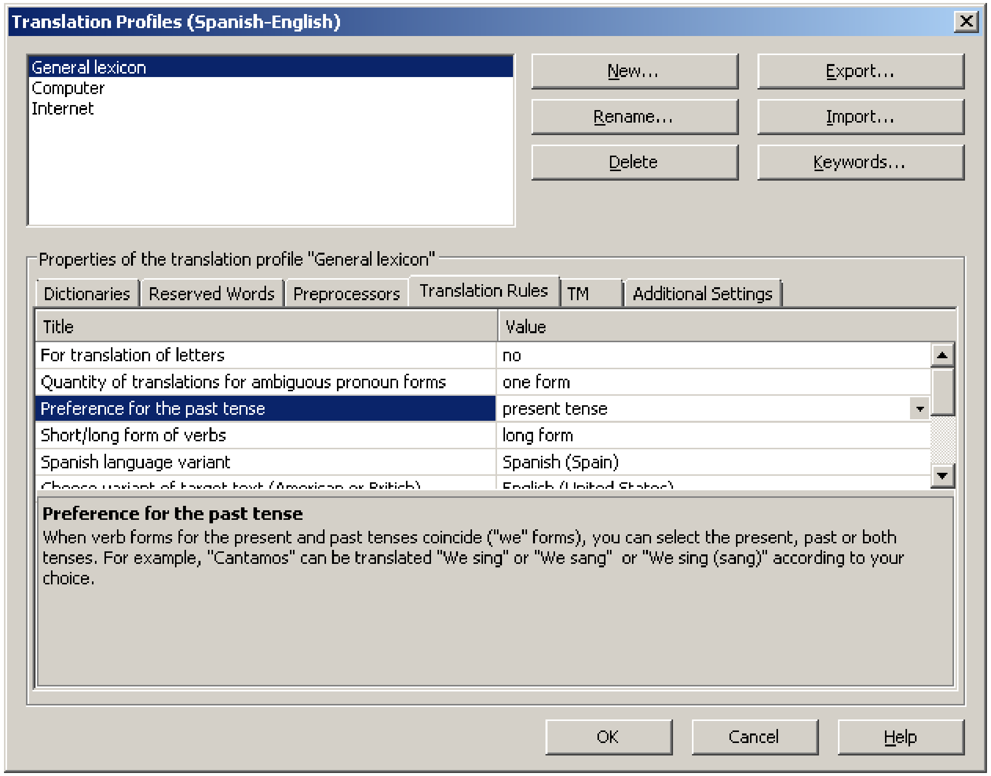
- Demo fully functional for 30 days
- Available languages: EN<>FR, IT, DE, ES, RU, PT
- Download Language Service Provider 9.5

ata

American Translators Association

Profiles

- Dictionary hierarchy
- Translation rules
- Translation memories and thresholds



General lexicon
Computer
Internet

Properties of the translation profile "General lexicon"

Title	Value
For translation of letters	no
Quantity of translations for ambiguous pronoun forms	one form
Preference for the past tense	present tense
Short/long form of verbs	long form
Spanish language variant	Spanish (Spain)
Choose variant of target text (American or British)	English (United States)

Preference for the past tense
When verb forms for the present and past tenses coincide ("we" forms), you can select the present, past or both tenses. For example, "Cantamos" can be translated "We sing" or "We sang" or "We sing (sang)" according to your choice.

ata
American Translators Association

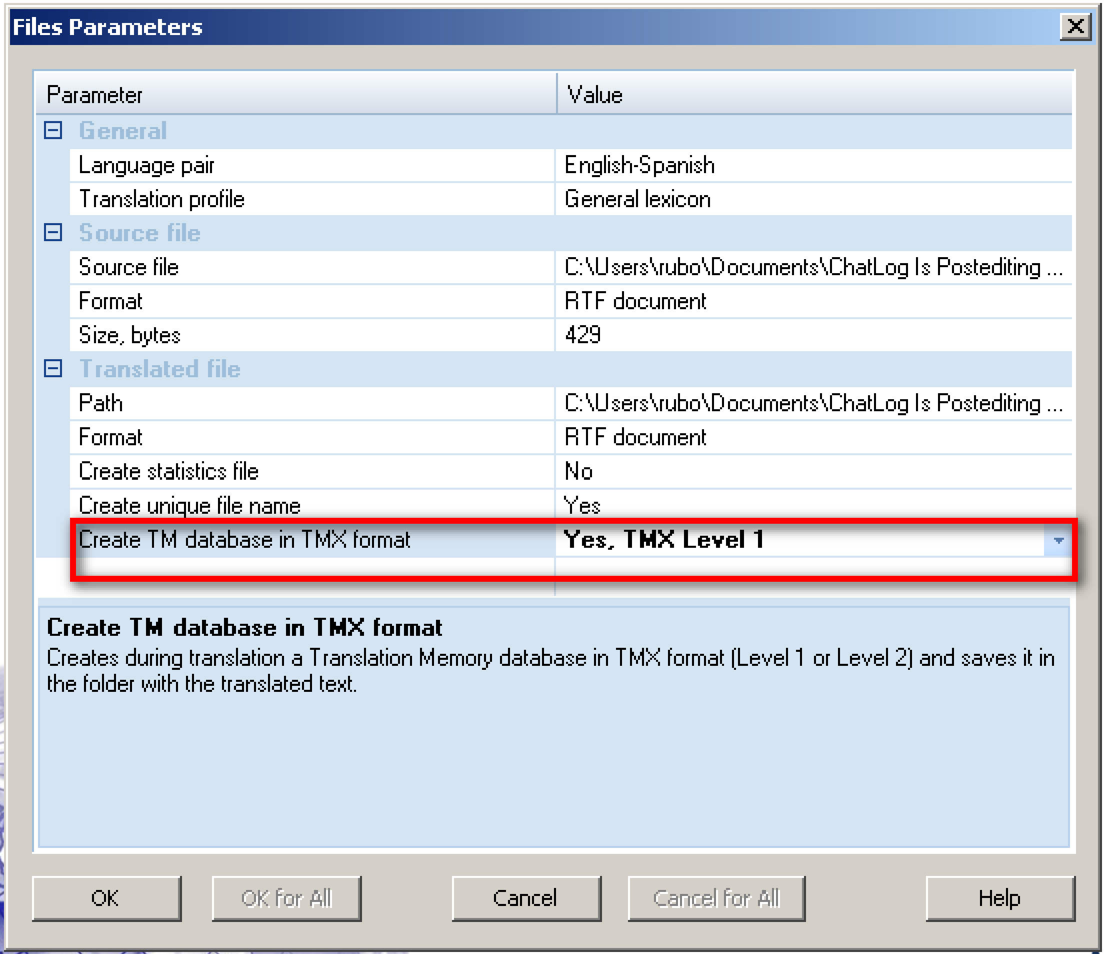
Legacy Glossary Import Demo

Demo

Translating with ProMT

Demo

Always Save Output to TMX



Parameter	Value
⊟ General	
Language pair	English-Spanish
Translation profile	General lexicon
⊟ Source file	
Source file	C:\Users\rubo\Documents\ChatLog Is Postediting ...
Format	RTF document
Size, bytes	429
⊟ Translated file	
Path	C:\Users\rubo\Documents\ChatLog Is Postediting ...
Format	RTF document
Create statistics file	No
Create unique file name	Yes
Create TM database in TMX format	**Yes, TMX Level 1**

Create TM database in TMX format

Creates during translation a Translation Memory database in TMX format (Level 1 or Level 2) and saves it in the folder with the translated text.

ata

American Translators Association

Statistical Machine Translation

- Figure out word/phrase alignment by statistical analysis of bilingual corpora
- Output is **not predictable**, but can be more fluent than RbMT
- **Cheaper** and **quicker** to develop
- Require more technical skills. Very **user-unfriendly** for now.

American Translators Association

Caveat

Garbage in, garbage out principle was never truer

Make sure TMs are in good shape

Do-Moses-Yourself (DoMY)

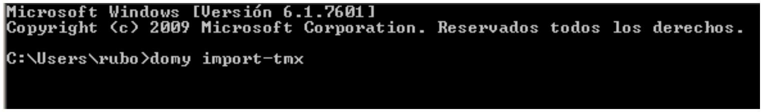
Graphs: import-tmx, clean-LM/TM, build LM/TM, train, translate.

Ini files: configuration (language pairs, paths for input and output).

Folder structure: hierarchy and stages

Running DoMY

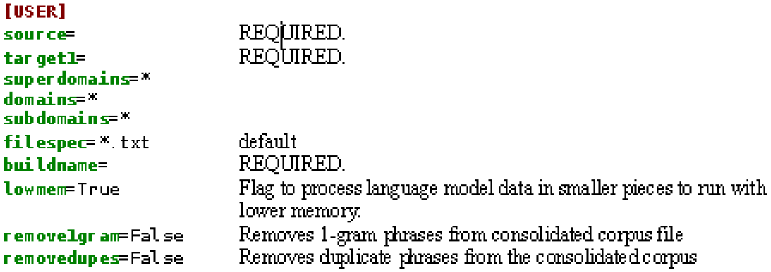
- Command line:



```
Microsoft Windows [Versión 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Reservados todos los derechos.

C:\Users\rubo>domy import-tmx
```

- Ini for configuration



```
[USER]
source=                 REQUIRED.
target1=                REQUIRED.
superdomains=*
domains=*
subdomains=*
filespec=*.txt          default
buildname=              REQUIRED.
lowmem=True             Flag to process language model data in smaller pieces to run with
                        lower memory.
remove1gram=False       Removes 1-gram phrases from consolidated corpus file
removedupes=False       Removes duplicate phrases from the consolidated corpus
```

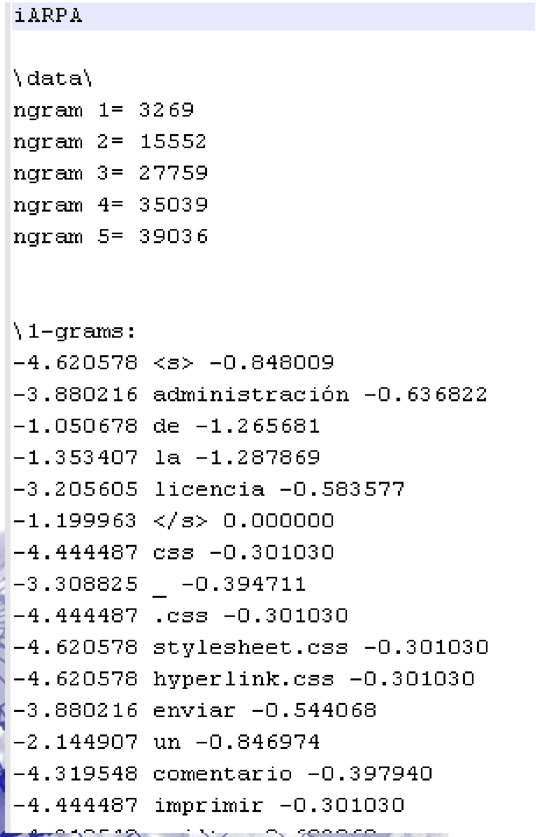
American Translators Association

DoMY Output

- **Translation Model**: table containing source and target phrases, together with a probability score.
- **Language Model**: monolingual corpus the system refers to produce more fluent output (e.g. reordering)

Language and Translation Models

- LM (fluency)

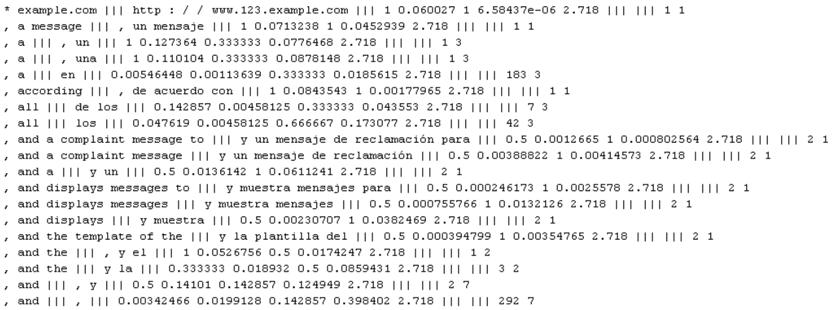


```
iARPA

\data\
ngram 1= 3269
ngram 2= 15552
ngram 3= 27759
ngram 4= 35039
ngram 5= 39036


\1-grams:
-4.620578 <s> -0.848009
-3.880216 administración -0.636822
-1.050678 de -1.265681
-1.353407 la -1.287869
-3.205605 licencia -0.583577
-1.199963 </s> 0.000000
-4.444487 css -0.301030
-3.308825 _ -0.394711
-4.444487 .css -0.301030
-4.620578 stylesheet.css -0.301030
-4.620578 hyperlink.css -0.301030
-3.880216 enviar -0.544068
-2.144907 un -0.846974
-4.319548 comentario -0.397940
-4.444487 imprimir -0.301030
```

- TM (equivalences)



```
* example.com ||| http : / / www.123.example.com ||| 1 0.060027 1 6.58437e-06 2.718 ||| ||| 1 1
, a message ||| , un mensaje ||| 1 0.0713238 1 0.0452939 2.718 ||| ||| 1 1
, a ||| , un ||| 1 0.127364 0.333333 0.0776468 2.718 ||| ||| 1 3
, a ||| , una ||| 1 0.110104 0.333333 0.0878148 2.718 ||| ||| 1 3
, a ||| en ||| 0.00546448 0.00113639 0.333333 0.0185615 2.718 ||| ||| 183 3
, according ||| , de acuerdo con ||| 1 0.0843543 1 0.00177965 2.718 ||| ||| 1 1
, all ||| de los ||| 0.142857 0.00458125 0.333333 0.043553 2.718 ||| ||| 7 3
, all ||| los ||| 0.047619 0.00458125 0.666667 0.173077 2.718 ||| ||| 42 3
, and a complaint message to ||| y un mensaje de reclamación para ||| 0.5 0.0012665 1 0.000802564 2.718 ||| ||| 2 1
, and a complaint message ||| y un mensaje de reclamación ||| 0.5 0.00388822 1 0.00414573 2.718 ||| ||| 2 1
, and a ||| y un ||| 0.5 0.0136142 1 0.0611241 2.718 ||| ||| 2 1
, and displays messages to ||| y muestra mensajes para ||| 0.5 0.000246173 1 0.0025578 2.718 ||| ||| 2 1
, and displays messages ||| y muestra mensajes ||| 0.5 0.000755766 1 0.0132126 2.718 ||| ||| 2 1
, and displays ||| y muestra ||| 0.5 0.00230707 1 0.0382469 2.718 ||| ||| 2 1
, and the template of the ||| y la plantilla del ||| 0.5 0.000394799 1 0.00354765 2.718 ||| ||| 2 1
, and the ||| , y el ||| 1 0.0526756 0.5 0.0174247 2.718 ||| ||| 1 2
, and the ||| y la ||| 0.333333 0.018932 0.5 0.0859431 2.718 ||| ||| 3 2
, and ||| , y ||| 0.5 0.14101 0.142857 0.124949 2.718 ||| ||| 2 7
, and ||| , ||| 0.00342466 0.0199128 0.142857 0.398402 2.718 ||| ||| 292 7
```

American Translators Association

M4Loc

Set of scripts to integrate Moses with L10N tools:

- **Xliff to Moses**: remove mark up
- **Moses to xliff**: reinsert mark up
- Also include **Adobe Moses toolset**

ata

American Translators Association

Where To Get Corpora

- Opus (ECB, EMEA, OpenOffice)
- Acquis Communautaire
- Europarl
- Multilingual websites: Bitextor

Tips for Efficient Postediting

- Embrace the **keyboard**
- Run **automatic QA** tool
- Fix repetitive issues with **global search and replace** (powered with wildcards and regular expressions)
- Store S&R operations in **macros**

ata

American Translators Association

RegEx Basics

- Charsets: [a-z] [0-9]
- Non-printable: \n, \r, \t
- Anchors: ^, $, \b
- Alternation: |
- Grouping: ()

American Translators Association

Search and Replace Macro

Demo

Do You Still Think it Takes Less Time to Translate from Scratch?

Poll

American Translators Association

Questions?

Speak now…

Or reach me at:

www.facebook.com/xlation

www.wordbonds.es

@rubendelafuente

http://www.linkedin.com/in/rubendelafuente

American Translators Association